FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

Etnische Profilering en de Consequenties: een case-study in Rusland

door

Tim VAN ERUM

Studentennummer: 01405190

Promotor: Prof. Dr. Koen Schoors

Co-promotor: Tom Eeckhout

Masterproef voorgedragen tot het bekomen van de graad van:

Master of Science in de Algemene Economie

Academiejaar: 2020-2021

UNIVERSITEIT
GENT

Voorwoord

Het schrijven van een thesis is, zoals altijd, opnieuw een enorme opgave geweest. Een uitdagend onderwerp, uitdagende omstandigheden en uitdagende uitdagingen die gepaard gingen met de situatie. Dergelijke situaties los je niet alleen op, zeker niet door een teamspeler als ik. Ik wil dan ook even deze witruimte gebruiken om een aantal belangrijke mensen te bedanken voor hun ongelofelijke hulp bij dit werk.

Een ongelofelijke merci aan mijn co-promotor Tom Eeckhout, die van in het begin van het jaar met heel veel enthousiasme en zeer inzichtvolle feedback dit project begeleid heeft. Altijd aanspreekbaar en altijd bereid om te helpen.

Daarnaast een onvoorstelbaar grote dankuwel aan Tim Leers, collega bij dataroots. Halftijds werken in een thesisjaar heeft zo zijn nadelen, maar zijn onvoorwaardelijke hulp en ervaring bij het statistisch onderzoek, hebben mij enorm veel bijgeleerd en het onderzoek een stuk beter gemaakt.

Verder nog een speciaal dankwoordje naar mijn ouders en mijn vriendin, die niet alleen mijn karakter en onregelmatig schema hebben moeten verdragen tijdens het schrijven van deze thesis, maar ook nog eens regelmatig gefunctioneerd hebben als klankbord en inspiratiebron.

Tim Van Erum, 1 juni 2021

Copyright Notice

Etnische Profilering en de Consequenties: een case-study in Rusland

door

Tim VAN ERUM

Masterproef voorgedragen tot het bekomen van de graad van:
Master in de algemene economie

Promotor: Prof. Dr. Koen Schoors
Supervisors: Tom Eeckhout

Academiejaar: 2020-2021

Abstract

Etnische profilering is een breed besproken onderwerp in de literatuur. De redenen ervoor zijn gekend, gaande van Terrorism Acts, verschillen in misdaadcijfers, etc. De invloed die etnische profilering, en bij uitbreiding etnische verdeeldheid op een maatschappij kunnen hebben, worden besproken binnen de context van generalized en particularized trust. Rusland is dan ook een ideale onderzoeksbodem voor dergelijke problemen, met een hoge graad van etnische verdeelheid en een gebrekkig etnisch beleid. In dit onderzoek wordt de invloed van etniciteit op verschillende verkeersparameters nagegaan. Door te onderzoeken wat de invloed van verschillende Russische etniciteiten op 1) de hoeveelheid boetes, 2) de kans op een boete en 3) de snelheid van recidivisme is, probeert het onderzoek een patroon vast te stellen van etnische profilering in Russiche verkeersdata. Met behulp van een samengestelde dataset van verkeersdata in en rond Moskou, worden verbanden tussen (al dan niet geaggregeerde) variabelen onderzocht, en worden verschillen naargelang etniciteit bekeken.

Keywords

Ethnic Profiling, Traffic Data, Logistic Regression, Multivariate Linear Regression, Survival Analysis.

Contents

1 Introductie **1**

1.1 Probleembeschrijving 2

1.2 Uitdagingen 2

2 Literatuurstudie **4**

2.1 De context in Rusland 4

2.1.1 Demografie 5

2.1.2 Etnische situatie in Rusland 6

2.1.3 Bepalen etniciteit op basis van naam 8

2.2 Verkeer 10

2.3 Etnische profilering 13

3 Data en Model **17**

3.1 Experimenten 17

3.1.1 Verklaren van de hoeveelheid boetes 17

3.1.2 Kans op boete en afname rijbewijs 18

3.1.3 De rol die etniciteit speelt in recidivisme 19

3.1.4 Assumpties 20

3.2 Dataset en -pipeline 21

3.2.1 Herkomst en beschrijving 21

3.2.2 Exploratie 1 23

3.2.2.1 Algemeen 23

3.2.2.2 ongevallen 28

3.2.2.3 Overtredingen 32

3.2.3 Data Preprocessing 1 40

3.2.3.1 Truncation . . . 40
3.2.3.2 Cleaning . . . 42
3.2.4 Ethnicity Prediction . . . 43
3.2.5 Data Preprocessing 2 . . . 46
3.2.5.1 Verfijnen . . . 46
3.2.5.2 Experimenten voorbereiden . . . 47
3.2.6 Exploratie 2 . . . 49

4 Resultaten **51**
4.1 Aantal boetes . . . 51
4.1.1 Model . . . 51
4.1.2 Colineariteit . . . 53
4.1.3 Resultaten . . . 55
4.2 Kans op een boete/ingetrokken rijbewijs . . . 57
4.2.1 Model . . . 57
4.2.2 Colineariteit . . . 57
4.2.3 Resultaten . . . 58
4.3 Recidivisme . . . 60
4.3.1 Survival Analyse . . . 60
4.3.2 Model . . . 61
4.3.3 Resultaten . . . 62

5 Conclusie **68**
5.1 Bevindingen . . . 68
5.2 Beperkingen en Future Work . . . 69

List of Figures

2.1 Populatiepyramide in Rusland (2021) 6

3.1 Verdeling over (1993-2006) 25
3.2 Aantal registraties per leeftijd 26
3.3 Percentage vrouwelijke registraties doorheen de jaren 27
3.4 Leeftijdsverdeling in respectievelijk de algemene populatie, de populatie aan ongevallen, ongevallen zonder overtreding en ongevallen met overtreding . . . 30
3.5 Verdeling van de registraties, ongevallen, ongevallen met overtreding en ongevallen zonder overtreding overheen de seizoenen - genormaliseerd 31
3.6 Verhouding ongevallen veroorzaakt door overtreding tegenover ongevallen zonder overtreding overheen de seizoenen 32
3.7 Verhouding dure - goedkope wagens in de ongevallen met en zonder overtreding 33
3.8 Verdeling van overtredingen, beboete overtredingen, onbeboete overtredingen en algemene registraties over de leeftijden 34
3.9 Verdeling van overtredingen, beboete overtredingen, onbeboete overtredingen en algemene registraties over de leeftijden 35
3.10 Ratio van beboete overtredingen tegenover totale overtredingen overheen de leeftijd 35
3.11 Verdeling van de gemiddelde boete-grootte overheen de leeftijden 36
3.12 Ratio van beboete overtredingen tegenover totale overtredingen overheen de leeftijd voor 2 genders 37
3.13 Ratio van beboete overtredingen tegenover totale overtredingen vergeleken tussen dure en goedkope wagens 38
3.14 Ratio van beboete overtredingen tegenover totale overtredingen in functie van de leeftijd, voor vanity plates en gewone nummerplaten 38

3.15 Gemiddeld aantal maanden ingetrokken rijbewijs overheen de leeftijden 39
3.16 Gemiddeld aantal maanden ingetrokken rijbewijs overheen de leeftijden 40

List of Tables

3.2 Verdeling van voorspelde etniciteiten 49
3.3 Caption 50

4.1 Verdeling van personen over de groepen 52
4.2 VIF Factor van de verschillende variabelen 53
4.3 Correlatietabel van de variabelen 54
4.4 VIF Factor van de verschillende variabelen 58
4.5 Logit Regression Results 59
4.6 OLS Regression Results 66
4.7 Resultaten van het Cox Proportional Hazard Model 67

Chapter 1

Introductie

Etnische profilering als methodiek bij de politie krijgt steeds vaker te kampen met kritiek. Schandalen zoals de dood van George Floyd en anderen zijn dan ook schering en inslag in de hedendaagse media. De exacte definitie van etnische profilering krijgt echter een verschillende invulling in verschillende contexten. Zo hanteert Meehan & Ponder [32] de definitie: "When police officers stop or cite a disproportionate number of minorities", die duidelijk voornamelijk gefocust is op statistieken. Andere voorbeelden zijn "Stopping minorities who are walking or standing in public space" (Russel, 1999) [42], "The practice of targeting minorities for unwarranted traffic stops" (Tomaskovic-Devey, Mason and Zingraff, 2004) [51] en "The practice of detaining a suspect based on a broad set of criteria which casts suspicion on an entire class of people without any individualized suspicion of the particular person being stopped" (Wetgeving van de staat California). Deze laatste definities zijn duidelijk meer gefocust op het werkelijke proces, dan op de feitelijke statistieken.

De bovengenoemde, sterk gemediatiseerde zaken inzake etnische profilering zijn afkomstig uit de Verenigde Staten van Amerika. De VS is echter zeker niet het enigste land met etnische verdeeldheid. Een claim in het werk van Richard Arnold [2] stelt dat Rusland het gevaarlijkste land in Europa is voor etnische minderheden. Zo argumenteert het dat etnisch geweld in Rusland in veel sterkere mate *systemisch* is dan voor andere ontwikkelde landen. De oorzaak hiervoor is, zoals bij veel zaken in het huidige Rusland, terug te brengen naar economische en sociale veranderingen in het post-Soviet tijdperk.

1.1 Probleembeschrijving

In dit werk focussen we ons op de definitie van Meehan & Ponder [32], waarbij etnische profilering wordt gedefinieerd als een overrepresentatie van minderheden in verkeersstops en -boetes. Op basis van een uitgebreide dataset aan Russische verkeersdata wordt er getracht om etnische profilering, volgens de definitie van Meehan en Ponder, vast te stellen in het verkeer in Rusland. Hiervoor wordt er gekeken naar de invloed die etniciteit heeft op een aantal verkeersparameters, zoals de kans op een boete, kans op afname van het rijbewijs, etc.

1.2 Uitdagingen

In een ideale wereld is het oplossen van bovenstaand probleem een eenvoudige taak van het op zoek gaan naar patronen in perfecte verkeersdata. De verzamelde dataset is echt geen perfecte dataset. De redenen hiervoor zijn uiteenlopend:

- De dataset is niet gepubliceerd door een officiële instantie, zoals de Russische verkeerspolitie, of een Russisch data-agentschap. Deze data is beschikbaar aangezien ze gelekt uit via verschillende bronnen, met behulp van data engineering technieken gecombineerd en verrijkt is, en zo gepubliceerd. Dit bemoeilijkt het bevestigen van de accuraatheid van de data.

- Rusland is een land dat bekend staat vanwege de hoge graad van corruptie. In 2014 scoorde Rusland volgens Transparency International 27/100 voor zijn Corruption Perception Index [25]. Dit is een maatstaf voor het subjectieve niveau van corruptie dat de inwoners van een land ervaren. Slechts 40 landen scoorden op dat moment slechter op deze index wereldwijd (ter illustratie, in 2020 scoort België met een score van 76 een 15e plaats in de wereld). Verder stond volgens Rimskyi, VL [39] de verkeerspolitie binnen Rusland nog eens bekend als één van de, dan niet meest, corrupte overheidsdienst in Rusland. Dit niveau van corruptie plaatst vraagtekens bij de data waarop dit onderzoek zich baseert.

- Een belangrijke parameter in de verklaring van veel ongeval- en boetecijfers is de zogenaamde *exposure* van chauffeurs. Dit verband geeft aan dat hoe meer tijd/kilometers een bepaalde chauffeur op de baan aflegt, hoe hoger de kans dat hij/zij een

accident en/of boete zal hebben. Het ontbreken van deze variabele kan dus een significante invloed hebben op het onderzoek.

- Ten slotte is etniciteit geen eenvoudige variabele om te meten. Aangezien de etniciteit van een chauffeur niet direct gemeten is, moet deze geïnfereerd worden uit de andere beschikbare data. Dit wordt (imperfect) opgelost, doormiddel van machine learning-technieken die aan de hand van neurale netwerken etniciteiten kunnen voorspellen op basis van namen. Hier ziet echter een zekere foutenmarge op, wat opnieuw een uitdaging vormt.

Chapter 2

Literatuurstudie

De literatuurstudie van dit onderzoek kan ruwweg onderverdeeld worden in 3 delen.

In het eerste deel wordt er wat context gegeven bij dit onderzoek in verband met de situatie in Rusland. Het is belangrijk om een duidelijk beeld te krijgen op de omgeving waarbinnen deze dataset vergaard is. In die optiek is er een groot deel van de literatuurstudie ook besteedt aan het beter begrijpen van de sociale en ethnische omgeving in Rusland. Hier komen een aantal thema's aan bod, zoals een demografische beschrijving van het land, en een overzicht van de etnische situatie in Rusland.

Ten tweede wordt de bestaande literatuur besproken in verband met verkeer, boetes, ongevallen en de mogelijke verklaringen voor verbanden tussen deze onderwerpen. In deze stap worden de invloedsfactoren voor ongevallen en boetes besproken, zoals leeftijd, gender, etc., en vervolgens een beschrijving van de literatuur inzake risico-gedrag bij chauffeurs.

Ten slotte eindigen we deze literatuurstudie met een beschrijving van het bestaande onderzoek naar het fenomeen etnische profilering. Hier bespreken we verschillende definities van het fenomeen, en de mogelijke gevolgen van etnische profilering op een bevolking.

2.1 De context in Rusland

Aangezien de dataset afkomstig is uit de politie-databanken van Russische departementen, is het belangrijk om de context van het land beter te begrijpen, alvorens het onderzoek verder te zetten.

In een eerste stap bespreken we de demografische karakteristieken van de bevolking,

hoe de bevolking verdeeld is inzake leeftijd, gender, inkomen en een korte beschrijving van de evolutie daarvan. Vervolgens gaan we speciale aandacht besteden aan de etnische situatie in Rusland. Het land bestaat uit zeer veel verschillende etniciteiten wegens de grootte van het land, zijn deze ook zeer sterk geografisch verspreid. Dit is dus een interessant thema om te bespreken binnen de context van dit onderzoek. Ten slotte wordt er in een stuk ook besproken hoe etniciteit in Rusland verbonden is met de naam. Dit is een belangrijk onderdeel van het onderzoek, en bouwt verder op een vorige thesis.

2.1.1 Demografie

De demografische gegevens hier gepresenteerd zijn deze afkomstig uit de *CIA World Factbook*[14], gecombineerd met een analyse van DaVanzo et al.[9].

In Rusland zijn er in de census van 2010 142,8 miljoen mensen geregistreerd. Hiervan zijn er 77,7% etnische Russich, 3,7% Tataar, 1,4% Oekraïns, 1,1% Bashkir, 1% Chuvash, 1 % Checheen en nog 10,2% andere etniciteiten (er zijn bijna 200 nationale en/of etnische groepen opgenomen in de census van 2010).

In termen van religie heeft het land zeer veel niet-practiserende gelovigen en niet-gelovigen, het gevolg van meer dan 70 jaar aan officiëel atheïsme onder de Soviet-Unie. Voor de rest is 15-20% Russich Orthodox, 10-15% Moslim en nog 2% is Christelijk.

Het land kampt sinds de jaren 1990 (na de val van de Soviet-Unie) met een zeer laag geboortecijfer. Het geboortecijfer ligt zelfs lager dan het sterftecijfer (10/1000 vs. 13,4/1000), en het land heeft 1 van de laagste vruchtbaarheidsratio's ter wereld met 1,6 kinderen per vrouw (lager dan de 2,1 nodig om de autochtone bevolking niet te doen afnemen). De enige reden dat de bevolking in Rusland lange tijd is blijven groeien is de positieve netto migratiegraad (in 2020 geschat op 1,7 migranten/1000 inwoners), maar ook dit is niet meer voldoende om in 2020 een positieve bevolkingsgroei te bekomen (negatief met -0,16%).

De populatiepyramide van het land kan gevonden worden in 2.1

Een analyse van de populatiepyramide 2018 door Vishnevsky en Shcherbakova [55] (populatiepyramides hebben van nature een behoorlijk inert karakter), verklaart waarom de pyramide de golvende structuur volgt die ze heeft. Volgens hen heeft niet enkel het verouderen van de bevolking een sterke invloed gehad, maar sterker nog hebben de zogenaamde *echo's* van historische gebeurtenissen gezorgd voor een dergelijke structuur. Vooral ook het onstabiele karakter van de ratio man-vrouw is hierin zeer opmerkelijk.

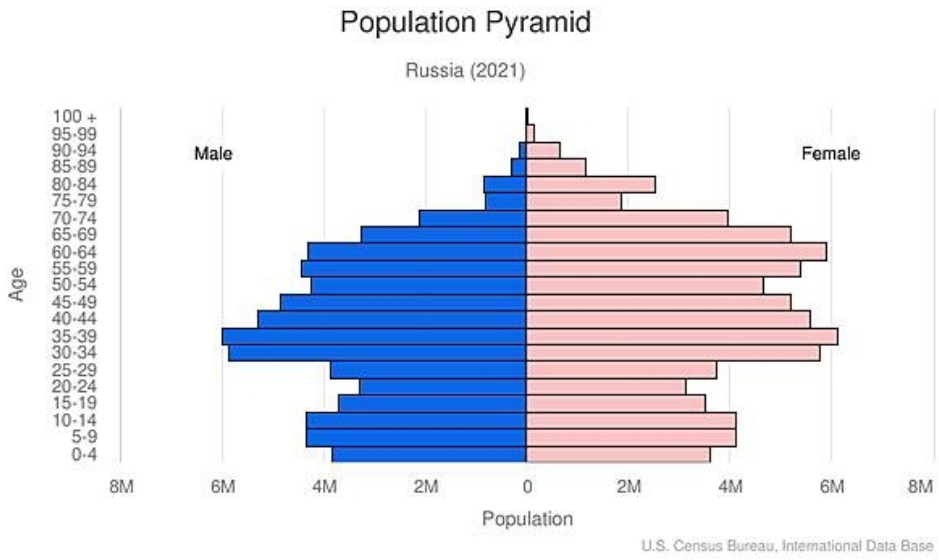


Figure 2.1: Populatiepyramide in Rusland (2021)

2.1.2 Etnische situatie in Rusland

Nadat de Sovjet-Unie uiteengevallen is in 15 onafhankelijke staten in 1991, zijn veel nieuw-onafhankelijke landen hun etniciteit beginnen promoten. Rusland viel echter niet onder die landen, want, zoals gesteld in Peter Rutland [43], er waren veel onduidelijkheden over de Russische nationale identiteit. Dit komt omdat het Russisch nationalisme een duidelijke rol gespeeld had in het opbreken van de Soviet-Unie, maar dat er nooit een duidelijk concept geformuleerd is voor de 'Russische identiteit'.

Tegelijk is er in Rusland een etnisch verdeelde bevolking, waarbij etnische Russen slechts 4/5en van de bevolking uitmaken en komt de etnische verdeeldheid ook duidelijk naar voor in de politieke structuur: een etnisch federalisme, waarbij de rechten van bepaalde etnische groepen erkend werden in sommige gebieden, maar niet in andere. Zoals gesteld door etnograaf Emil Pain:

> "How can a single and positive identity form among the inhabitants of a state that is regarded by both the authorities and the public as an unexpected, illegitimate child, a cripple, the victim of a catastrophe or plot?"

Deze verdeeldheid wordt verder geïllustreerd wanneer er gekeken wordt naar vertrouwen. In de paper van Bahry et al. [3] wordt er verwezen naar het laag vertrouwen in het algemeen (*generalized faith in others*) en het hoge vertrouwen binnen de groep (*particularized trust*). Zo gaven surveys van Dowley en Silver [11], Gibson [20] en Rose [41] aan dat

slechts 25 tot 30% van de mensen een vertrouwen vertonen in het algemeen, terwijl ze een veel sterker vertrouwen vertoonden in kennissen, familie en vrienden (*particularized* of *in-group trust*).

Door de toename aan etnische assertiviteit aan het einde van de 20e eeuw, is er ook een toename gekomen van het gebruik van de titulaire taal, een aangewakkerde interesse in de lokale cultuur en gebruiken en soms zelfs een toename in het aanhangen van de traditionele religies.

Deze etnische verdeeldheid komt ook voor in het onderzoek van Bessudnov en Shcherbak[6]. Zij hebben een experiment in de arbeidsmarkt gedaan waarbij ze 9 000 jobapplicaties hebben gestuurd naar werkgevers in 4 verschillende steden: Moskou en Sint-Petersburg (beiden grote metropolen met een diverse, maar toch voornamelijk Russische bevolking) en Kazan en Ufa (beiden hoofdsteden van een etnische autonomie, waar er meer een verdeling tussen etnisch Russische en inheemse Moslim bevolking was). Hun onderzoek bevestigde de voorafgaande literatuur van Sidanius & Pratto [45] door aan te tonen dat er wel degelijk een consistente, impliciete hiërarchie werd geobserveerd. Minderheden die uiterlijk herkenbaar waren (donkerdere huid, herkenbare fysieke eigenschappen) werden veel vaker gediscrimineerd in Moskou en Sint-Petersburg dan minderheden met een Europese origine. Ze stelden ook vast dat de discriminatie tegen minderheden sterker was tegenover mannen dan ten opzichte van mannen. In dit onderzoek werden slechts 13 etnische groepen opgenomen:

- Etnische Russen
- Armenen
- Azeri
- Tsjetsjenen
- Georgiërs
- Duitsers
- Joden
- Letten
- Litouwers

- Tajiks
- Tataren
- Oekraïners
- Oezbeken

We bespreken binnen de context van dit onderzoek enkel het deel van Bessudnov en Shcherbak's onderzoek dat uitgevoerd wordt in Moskou en Sint-Petersburg, aangezien de dominante locatie van registraties in de dataset de regio in en rond Moskou is, en de grote metropolen dus een betere context schetsen voor deze dataset.

Van bovenstaande etniciteiten werden etnische Russen het vaakst gecontacteerd (41%). Oekraïnse, Joodse en Duitse namen hadden slechts licht lagere contactratio's, en die verschillen zijn niet statistisch significant. Aan de andere kant hebben de etnische groepen van niet-Europese origine wel significant lagere contactratio's, gaande van 26% voor Georgiërs tot 28% voor Tataren. De contactratio's voor Letse en Litouwse namen zaten daar tussen, met een contactratio van 34%. Hier werd dus een duidelijke etnische hiërarchie geobserveerd, waar groepen met een Europese origine een voorkeur kregen op Zuiderse groepen en groepen met een niet-Europese origine. Deze minderheden zijn niet toevallig ook een stuk meer herkenbaar dan de groepen met een Europese achtergrond. Dit is een belangrijk inzicht dat we kunnen meenemen in het onderzoek naar etnische profilering in het verkeer.

De lijst met etnische conflicten in Rusland is dan ook lang, en gaat over allerlei minderheden. Zo worden er etnische conflicten met Tjetsjenen en Noord-Kaukasiërs beschreven in het artikel van Patrick Sewel [44], beschrijft Payin et al. [34] conflicten met Meskhetische Turken en Armenen, en is er een hele waslijst aan conflicten te vinden in het werk van Tishkov [50].

2.1.3 Bepalen etniciteit op basis van naam

In de gebruikte dataset voor dit onderzoek is er geen variabele beschikbaar voor de etniciteit van de chauffeur. 1 mogelijkheid zou zijn om de etniciteit te gaan bepalen op basis van de geboorteregio, en op basis daarvan een assumptie te maken over de etniciteit. Met deze werkwijze zijn er echter 2 problemen:

- De geboorteregio is niet steeds aanwezig. Er is veel ontbrekende data voor deze variabele.

- Etnische afkomst in Rusland is niet zo eenvoudig als simpelweg kijken naar de geboorteregio. Zoals aangegeven in het werk over etniciteitsbeleid in Rusland van Peter Rutland, heeft het land een zeer grote etnische verdeeldheid. In tegenstelling tot andere landen van de voormalige Soviet-Unie, is er in Rusland geen coherente etnische policy, geen duidelijke identiteit, etc.[43].

Om die reden werd er besloten om de etniciteit niet te bepalen op basis van geboorteregio, maar, voortbouwend op het werk van De Palmenaer et al., werd er een voorspelling van de etniciteit gemaakt op basis van de naam (voornaam - familienaam - patroniem)[10]. Deze voorspelling werd gedaan met behulp van een recurrent neuraal netwerk, dat op basis van de voornaam, familienaam en patroniem. Dit neuraal netwerk blijkt in staat te zijn om op basis van subtiele verschillen in de namen, voorspellingen te kunnen genereren welke de meest waarschijnlijke etniciteiten zijn van een bepaalde persoon, zelfs wanneer deze naam tot dat punt niet gezien was door het model.

In hetzelfde werk wordt er vermeld dat het model een accuracy haalt van 72,25%. Wanneer er gekeken wordt naar de 5 meest waarschijnlijke etniciteiten (van de 21 in totaal), dan stijgt de accuracy van het neurale netwerk naar 95,59%, hetgeen aangeeft dat het model typisch geen grote fouten maakt in de voorspelling.

Het is mogelijk om deze resultaten te gaan vergelijken met de resultaten die mensen halen wanneer ze de etniciteit moeten voorspellen op basis van de naam. Zo hebben Alexey Bessudnov en Andrey Shcherbak [6] in 2019 een onderzoek uitgevoerd waarbij ze een lijst van namen toonden aan de deelnemers. Zij moesten op basis van die namen de etniciteit voorspellen als een open antwoord, zonder voorgedefinieerde lijst van etniciteiten. Het is duidelijk dat het neuraal netwerk beter presteert dan de menselijke deelnemers, aangezien verschillende de correctheid op verschillende etniciteiten varieerde van 91% (Georgiërs) tot 12% (Tajik). Zelfs wanneer gekeken wordt naar de klasse 'ongeveer' correct, blijft het variëren tussen 98% (opnieuw Georgiërs) en 62% (German). Dit bevestigt wel de assumptie dat ook menselijke respondenten de neiging hebben om verschillende etniciteiten over dezelfde kam te scheren en dus weinig onderscheid te maken tussen die groepen.

Jack Glazer beargumenteerd in zijn werk waarom dat een imperfecte voorspelling inzake etniciteit geen probleem hoeft te zijn binnen de context van dit onderzoek [21]. Hij haalt in zijn boek zo aan dat politie-agenten bij het maken van een beslissing over het tegenhouden, beboeten, rijbewijs intrekken, enz. van een bestuurder ook geen perfecte

informatie heeft over etniciteit. Zo is er bijvoorbeeld het bekende probleem bij douanecontrole op de luchthaven, waar Sikhs vaker tegengehouden worden omdat ze aanzien worden als moslim. Dit verandert echter niets aan het feit dat zij dus ook de gevolgen dragen van die etnische profilering, of dat nu is als Sikh, of als moslim.

Als we dit koppelen aan de resultaten van het neurale netwerk, waar in meer dan 95% van de gevallen geen ernstige fout gemaakt werd door het model, nemen we de beslissing om de voorspelde etniciteiten door het neuraal netwerk als waarheid aan te nemen.

2.2 Verkeer

In dit stuk van de literatuurstudie wordt de literatuur aan onderzoek inzake verkeer, ongevallen, boetes, etc. gepresenteerd. Aangezien het onderzoek opereert op een dataset van verkeersdata, is het belangrijk om een duidelijk beeld te hebben op de oorzaken (en gevolgen) van ongevallen en boetes.

Het verklaren van waarom een ongeval of overtreding plaatsvind is een zeer complex gebeuren. Er zijn dan ook zeer veel factoren die een invloed kunnen hebben op een ongeval: het weer, de leeftijd van de persoon (bijvoorbeeld vertraagd reactievermogen), de leeftijd van de auto en nog veel meer. In dit deel bekijken we de bestaande literatuur over de invloeden van verschillende factoren op ongevallen. Analoog aan ongevallen, zijn ook boetes een complex fenomeen. Men kan de vraag stellen waarom mensen een overtreding begaan, wat de beweegredenen zijn, of er invloeden zijn van bepaalde persoonseigenschappen op het krijgen van een boete en nog veel meer.

Een ongeval is een onvrijwillige gebeurtenis, terwijl een overtreding eerder gepaard gaat met een keuze. Bij het zoeken naar de oorzaken voor ongevallen, en de beweegredenen van overtredingen, werd het al snel duidelijk dat er een zeer sterk verband tussen beide factoren is.

Volgens een onderzoek van Shinar et al. [31], is het gedrag van een chauffeur een dominant onderdeel in de verkeersveiligheid. De studie geeft aan dat een overgroot deel van de ongevallen veroorzaakt worden door menselijke factoren. Deze studie wordt bevestigd door de studies van Gebers en Peck, Elvik en Christensen, Goldenbeld et al. en Factor [13, 16, 19, 22]. Allen bevestigen deze studies dat de gedragingen die normaalgezien beschouwd worden als overtredingen van de verkeersregels, ook aanleiding zouden moeten geven tot een verhoogd risico op verkeersongevallen en verwondingen wanneer

een ongeval plaatsvindt. De studie van Parker et al. geef zelfs aan dat verkeersovertredingen, los van hun herkomst, bij de meest significante factoren zijn voor een toegenomen risico op ongevallen.

De studies trachten aan de hand van grotendeels dezelfde factoren het risico op een ongeval te gaan voorspellen, zijnde:

- Het aantal boetes
- Demografische variabelen zoals leeftijd, gender, ...
- Socio-economische variabelen, zoals plaats van woonst, inkomen, huishoudelijke klasse (obv. beroep en status op het werk), opleiding en religie.
- Type voertuig

Concrete resultaten kunnen gevonden worden in de studie van R. Factor [15]. Die stelt bijvoorbeeld dat slechts 82% van de ongevallen veroorzaakt worden door chauffeurs die slechts 1 boete of minder hadden per jaar, terwijl de resterende 18% van de ongevallen werden veroorzaakt door de resterende 6% van de chauffeurs die meer dan 1 boete per jaar hadden. Dit betekent dat mensen die veel boetes hebben, ook overgerepresenteerd zijn in het aantal ongevallen. Natuurlijk geeft dit niet direct een causaal verband aan tussen de neiging om overtredingen te begaan en de kans op een ongeval: een eenvoudige verklaring zou zo kunnen zijn dat mensen die meer boetes begaan simpelweg ook meer op de baan zijn dan mensen met minder boetes, en de kans op een ongeval is simpelweg ook hoger als je meer rijdt.

Alle bovenstaande studies geven aan dat het gedrag van de chauffeur een significante invloed heeft op de kans op ongevallen. Risico-gedrag achter het stuur, hetgeen tot expressie komt in het breken van de verkeersregels, te snel rijden, dronken achter het stuur, etc., is dus ook een belangrijke voorspeller voor ongevallen.

Dit lijkt een intuïtieve veronderstelling, maar dit voorspelt dat we in de data een belangrijk verband zouden moeten vinden tussen de hoeveelheid ongevallen en de hoeveelheid boetes die een bepaalde groep personen heeft.

Voorbeelden van concrete resultaten kunnen we halen uit de studie van R. Factor[15]. Die stelt bijvoorbeeld dat slechts 82% van de ongevallen veroorzaakt worden door chauffeurs die slechts 1 boete of minder hadden per jaar. De resterende 18% van de ongevallen werden veroorzaakt door de resterende 6% van de chauffeurs die meer dan 1 boete per jaar

hadden. Verder berekende hij dat een toename van 1 unit in het aantal boetes per jaar een toename van de kans op een fatale of ernstige crash veroorzaakt van 65%. Deze toename wordt zelfs nog sterker hoe hoger het aantal boetes wordt, bijvoorbeeld voor chauffeurs die 6 boetes of meer per jaar ontvangen, blijkt de kans tot 1051% hoger te liggen dan voor chauffeurs met slechts 1 boete per jaar.

Er bestaat een enorm corpus aan literatuur over de oorzaken van ongevallen , die allerlei situationele variabelen zoals leeftijd, gender, weersomstandigheden, opleidingsniveau, en zo verder bespreekt. Het onderzoek van Lourens et al. [30] bijvoorbeeld, neemt bepaalde input variabelen van de rijdende populatie, zoals karakteristieken van de chauffeur (leeftijd, gender, opleidingsachtergrond, etc.) inclusief de jaarlijkse kilometers per persoon, en bekijkt twee outputvariabelen: het rijgedrag (geoperationaliseerd door het aantal boetes) en de betrokkenheid bij ongevallen . Deze variabelen worden bestudeerd aan de hand van een Nederlandse dataset. De conclusie van dit onderzoek is gelaagd:

- Als er niet gekeken wordt naar de jaarlijkse kilometers (of dus de *exposure*) van de chauffeur, dan is het duidelijk dat mannelijke chauffeurs, en bij uitstek jonge mannelijke chauffeurs, het hoogste risico lopen op een ongeval.

- Wanneer de jaarlijkse kilometers wel in rekening gebracht worden, dan wordt dit onderscheid een stuk minder sterk: het verschil tussen jongere en oudere chauffeurs neemt af, en vrouwelijke chauffeurs presteren op vele vlakken zelfs slechter (meer boetes en meer ongevallen) dan mannen.

- Het vele onderzoek in verband met dit onderwerp bevestigend, wordt er ook in dit onderzoek een positieve correlatie gevonden tussen het aantal boetes en de betrokkenheid bij ongevallen . Dit is robuust na controle voor de verschillende klassen van jaarlijkse kilometers en wordt dus geacht onafhankelijk te zijn daarvoor.

In de review paper van Peck [35] wordt een overzicht gemaakt van verschillende onderzoeken naar de parameters die het risico op ongevallen beïnvloeden. Veel verschillende parameters worden naar voren geschoven en onderzocht, maar er wordt geen individuele, dominante variabele gevonden. De meest consistente en krachtige voorspeller op de kans op een ongeval blijft het aantal boetes van een persoon voor het ongeval. Hiermee wordt opnieuw het sterke verband tussen ongevallen en boetes bevestigd.

Het werk van Mercer [33] bevestigd de conclusies van Lourens et al. In zijn werk vindt

hij opnieuw dat mannen, en zeker jonge mannen, overgerepresenteerd worden in zowel ongevallen en boetes, maar dat dit verband (grotendeels) wegvalt na correctie voor het gereden aantal kilometers. Ook Simon en Corbett [46] komen tot deze conclusie. Rajalin [37] kwam tot de conclusie dat bestuurders die betrokken zijn in een fataal ongeval, gemiddeld genomen meer beboet worden in de 3 jaar voor het ongeval dan andere chauffeurs. Dit verband was nog steeds geldig als ook chauffeurs in acht genomen werden die niet in de fout waren bij het ongeval.

Ten slotte hebben Bradzil et al. [8] een analyse gemaakt van fatale ongevallen veroorzaakt door weersomstandigheden. Zo hebben ze de ongevallen gelinkt met 1 van de volgende omstandigheden: overstromingen, windstormen, convectieve stormen, regen, sneeuw, rijm, ijs, vrieskou, hitte en mist. De voornaamste weersgerelateerde oorzaken zijn vrieskou, gevolgd door rijm, ijs, regen en sneeuw. De meeste fatale ongevallen vonden plaats in januari en december, de minste in april en september.

De belangrijkste consusies uit dit deel is dat het aantal boetes de belangrijkste voorspeller zijn van het aantal ongevallen en dat leeftijd en gender in beperkte mate een invloed hebben op deze parameters. Deze bevindingen zijn belangrijk voor het verdere onderzoek.

2.3 Etnische profilering

Het fenomeen etnische profilering kent verschillende definities:

- "Any time an officer uses race in his or her decision to stop a vehicle" - Gaines et al. [17]
- "When police officers stop or cite a disproportionate number of minorities" - Meehan & Ponder. [32]
- "Stopping minorities who are walking or standing in public space" - Russel [42]
- "The practice of targeting minorities for unwarranted traffic stops" - Tomaskovic-Devey, Mason en Zingraff [51]
- "Any police-initiated action that relies on the race, ethnicity or national origin rather than the behavior of an individual who has bene identified as being, or having been, engaged in criminal acitivity." - Ramirez, McDevitt en Farrell [38]

De basis van al deze definities gegrond in de intenties van de agenten en geaggregeerde verkeersdata. In dit onderzoek zullen we voornamelijk de definitie van Meehan & Ponder hanteren, die gebaseerd is op een overrepresentatie van bepaalde minderheden in verkeerscijfers. Gaines et al. [17] plaatst hier echter wat bedenkingen bij. Zij stellen methodologische problemen voor die verbonden zijn aan het gebruiken van ruwe verkeersdata voor het opsporen van etnisch profileren. De conclusie dat etnische profilering bestaat op basis van een niet-evenredige representatie van minderheden in verkeersdata is namelijk gebaseerd op 3 assumpties die mogelijks verkeerd kunnen zijn:

1. Dergelijke redeneringen gaan uit van uniforme graden van het bezit van een wagen en jaarlijks gereden kilometers. Hier kan een significant verschil in zitten.

2. De ratio van verkeersovertredingen en criminele activiteiten word geacht gelijk te zijn overheen te verschillende etniciteiten. Hier kan mogelijks ook een verschil in bestaan (al haalt het onderzoek zelf wel aan dat er geen bewijs voor het tegendeel gevonden is).

3. Ordehandhaving is niet uniform verspreid in tijd en ruimte, wat betekent dat niet iedere mobilist evenveel kans heeft om tegengehouden te worden door een agent.

Deze veronderstellingen gaan verder besproken worden bij de onderzoeksvragen, binnen de context van dit onderzoek.

De consequenties van etnische profilering zijn niet min, zoals opgesomd in het werk van Dukes en Kahn [12]. Zo is er aangetoond dat een hogere graad van waargenomen en ervaren discriminatie gepaard gaan met een zwakkere gezondheid, zoals hyptertensie, verhoogd risico op cardiovasculaire ziekten, en ook meer kans op gezondheidsrisicovolle gedragingen zoals roken en alcoholgebruik [27, 56]. Verder zijn er ook psychologische effecten, zoals een verhoogde staat van psychisch ongemak, verhoogde kans op zware depressie en andere depressieve symptomen en lagere waarden op verschillende geluksmetrieken. [26, 28, 36, 49, 54]. Dezelfde observaties kunnen gemaakt worden wanneer er gekeken wordt naar geobserveerd racisme tegenover de groep waartoe het individu behoort, in plaats van enkel individueel ervaren racisme.

Verder kan de vraag natuurlijk gesteld worden wat de invloed van etnische profilering is op vertrouwen. Volgens het model van [3] is cross-etnisch vertrouwen een functie van 8 verschillende variabelen:

- *Generalized trust*, of algemeen vertrouwen. Dit is gebaseerd op de alombekende vraag of 'de meeste mensen vertrouwd kunnen worden'.
- *Political trust*, wat op zich weer een functie is van vertrouwen in de overheid op zijn verschillende niveaus.
- *Intermediate trust*, wat gaat over vertrouwen in buren, collega's, etc.
- *Outgroup stereotypes*. Dit zijn de gangbare stereotypen die gelden voor een bepaalde etnische groep. Belangrijk hierbij is het sentiment dat gekoppeld is aan deze stereotypen.
- *Intergroup contacts*. Dit is een metriek voor de gemiddelde hoeveelheid interactie die 2 groepen hebben.
- *Ingroup norms*, wat aangeeft hoe sterk de normen en de handhaving ervan zijn binnen een etnische groep.
- *Individual discrimination*. Dit is een maat voor de sterkte van de ervaren discriminatie van een individu.
- *Collective discrimination*. Dit is een maat voor de sterkte van de ervaren discriminatie van de groep waartoe het individu behoort (in termen van discriminatie in de arbeidsmarkt, bijvoorbeeld).

Uit de definitie van deze componenten kan er duidelijk afgeleid worden dat etnische profilering een directe invloed kan hebben op een aantal zaken, die verder besproken zullen worden. Een eerste factor waarom etnische profilering een invloed kan hebben is *political trust*. In een interactie tussen een bestuurder en een politie-agent, representeert de agent steeds meer dan 1 individuele ordehandhaver. Er kan beargumenteerd worden dat op basis van het uniform en wettelijke status, de agent op dat moment een weerspiegeling is van de overheid in het algemeen. Onrechtmatige behandeling door een politie-agent kan dan leiden tot een verlaagd vertrouwen in de politie, en bij uitbreiding de overheid. Dit wordt onderbouwd door het onderzoek van Garofolo, Hindelang en Huang & Vaughn[18, 23, 24]. Op die manier heeft etnische profilering dus een directe negatieve invloed op *political trust*. De invloed op *individual discrimination* en *collective discrimination* daarnaast is duidelijk.

Het mag duidelijk zijn dat etnische profilering ook op vertrouwen een belangrijke invloed kan hebben op vertrouwen, wat ook nog eens direct geïllustreerd wordt in het werk

van Tyler [53], waar het verschil in vertrouwen onderzocht werd tussen blanke New Yorkers en minderheden. De conclusie was dat vertrouwen het sterkst gecorreleerd was met de ervaren eerlijkheid van het gehanteerde beleid van de politie.

Chapter 3

Data en Model

In dit hoofdstuk worden eerst de onderzoeksvragen besproken in 3.1 en vervolgens nemen we een kijkje naar de beschikbare dataset in 3.2. Dit doen we eerst algemeen, met een beschrijving van de dataset en zijn herkomst, en vervolgens gaan we de dataset exploreren en voorbereiden voor gebruik.

3.1 Experimenten

3.1.1 Verklaren van de hoeveelheid boetes

Door de dataset te segmenteren naargelang verschillende dimensies zoals leeftijd, geslacht, dure/goedkope wagen, voertuigklasse, gaan we kijken wat de invloed is van etniciteit in het verband tussen het aantal boetes en het aantal ongevallen voor een bepaalde groep. Door rekening te houden met de leeftijd (zie 3.2.2.2 en 3.2.2.3, het geslacht (zie 3.2.2.3), het aantal ongevallen binnen die bepaalde groep, de gemiddelde kostprijs van de wagen, kunnen we trachten het effect van etniciteit te isoleren van andere effecten zoals risicovol gedrag, sociale klasse, etc. Dit bouwt verder op de bevindingen van 2.2.

De onderzoeksvraag die we hier dus proberen te beantwoorden is:

Wat is de invloed van etniciteit op het aantal boetes die een bestuurder ontvangt?

Hiervoor gaan we gebruik maken van een multipele regressieanalyse, met de volgende formule

$$y = a + b_1 X_1 + b_{2->n} X_{2->n} + c_{i->l} C_{i->l}$$

waarbij y het aantal geregistreerde boetes is voor de groep, X_1 is het aantal ongevallen voor de groep, $X_{2->n}$ zijn dummy variabelen voor de verschillende etnische groeperingen en $C_{i->l}$ is een set aan controlevariabelen. De controlevariabelen controleren voor variaties in gender, leeftijd, kostprijs van de wagen, etc. die niet gecaptured worden door het rechtstreekse verband tussen ongevallen en boetes (zie 2.2). Door te kijken of de $b_{2->n}$ significant zijn, kunnen we kijken of specifieke etniciteiten een rol spelen in de hoeveelheid boetes die de politie uitschrijft.

3.1.2 Kans op boete en afname rijbewijs

Het onderzoek van Experiment 1 kunnen we verder trachten te onderbouwen door te zoeken naar de rol die etniciteit speelt in de kans dat een bepaalde overtreding beboet wordt, hoe zwaar die beboet wordt, de kans dat een voor een bepaalde overtreding het rijbewijs wordt afgenomen en voor hoelang. Op dezelfde manier als bij Experiment 1 wordt er getracht om dat effect zoveel mogelijk te isoleren door te controleren voor confounding variables zoals kostprijs van de wagen, gender, leeftijd, etc.

De onderzoeksvraag die we hier dus proberen te beantwoorden is:

Wat is de invloed van etniciteit op de kans dat een bestuurder een boete ontvangt wanneer er een overtreding wordt vastgesteld?

We zoeken dus een significant verschil in de kans op

- boete
- invordering rijbewijs

bij verschillende etnische groepen. En we gaan op zoek naar de rol van etniciteit in het verklaren van

- de grootte van een boete
- de duur van de invordering van het rijbewijs.

Voor de kans op boete en de kans op invordering van het rijbewijs wordt er gebruik gemaakt van logistische regressiemodel.

$$f(y_i) = ln\frac{P(y_i)}{1-P(y_i)} = \beta_0 + \beta_1 X_1 + \beta_{2->n} * X_{2->n} + c_{i->l} C_{i->l}$$

Hierbij is $P(y_i)$ de kans dat een boete/invordering van rijbewijs plaatsvindt, X_1 is het gemiddeld aantal ongevallen per persoon voor de groep waartoe het individu behoort (deze groepen zijn berekend in het vorige experiment), $X_{2->n}$ zijn de dummy variabelen voor de etniciteit en $c_{i->l}$ zijn opnieuw de controle-variabelen. Door te kijken naar de coëfficiënten $\beta_{2->n}$ kunnen we opnieuw zien of etniciteit een significante rol speelt in de kans op een boete/ingetrokken rijbewijs.

3.1.3 De rol die etniciteit speelt in recidivisme

In het 3e experiment gaan we enkel kijken naar recidivisten, oftewel mensen waarvoor meerdere boetes geregistreerd zijn in de dataset. De redenering is opgebouwd op het idee van legitimiteit van de politie, zoals aangegeven in Tankebe et al. [48]. Legitimiteit refereert naar 'het gezag dat als rechtmatig wordt ervaren bij de relevante partijen, inclusief de machtshebbers, zij die onderworpen zijn aan die macht en derde partijen van wie de erkenning of ondersteuning de macht kan bevestigen' [5]. Dit idee zegt dat mensen de regels zullen volgen wanneer de legitimiteit van de politie hoog is. Dit betekent dat de politie consistent en fair moet zijn, want anders verliest de burger zijn incentive om de wet te volgen.

Neem bijvoorbeeld een hypothetische situatie van een persoon van vreemde etniciteit die tegengehouden worden onder valse voorwendselen door een politie-agent, en merkbaar onfair behandeld wordt door de agent vanwege die etniciteit. Op dat moment is er voor die chauffeur geen reden meer om de verkeersregels te volgen, want de legitimiteit van de agent, en bij uitbreiding de politie, is afgenomen.

Volgend op dit idee zouden mensen van vreemde etniciteit meer en sneller moeten recidiveren, aangezien het corrigerend effect van boetes, zoals beschreven in Becker, 1968 in zijn *basic model on crime and punishment*, wegvalt [4] en dus etnische minderheden volgens de *procedural justice theory* van Tyler de politie als minder legitiem ervaren.

De onderzoeksvraag die we hier dus proberen te beantwoorden is:

Wat is de invloed van etniciteit op de snelheid van recidivisme, gedefinieerd door de tijd tussen de 1e registratie en de 1e boete, of de tijd tussen 2 opeenvolgende boetes?

In dit experiment gaan we dus testen of er een significant verschil gemeten kan worden in de duur tussen 2 overtredingen voor verschillende etniciteiten. Hiervoor maken we gebruik van survival analysis, en specifiek *Cox Proportional Hazard Model*. Dit is een

veelgebruikt model in survival analysis, waar de duur van een event bestudeerd wordt. Dit model modelleert, zoals verwacht, de *hazard* op basis van een aantal voorspellers. In ons geval zijn dat

- Het aantal ongevallen van het individu voor de 1e boete
- Dummy's voor de etniciteit
- Controlevariabelen voor de leeftijd, gender, kostprijs van de wagen.

3.1.4 Assumpties

In de literatuurstudie werden er enkele assumpties aangehaald die gemaakt werden bij de definitie van Etnische profilering van Meehan & Ponder die hier gehanteerd worden, en die mogelijks verkeerd konden blijken te zijn[32]. We sommen deze assumpties hier nog eens op.

1. Dergelijke redeneringen gaan uit van uniforme graden van het bezit van een wagen en jaarlijks gereden kilometers. Hier kan een significant verschil in zitten.
2. De ratio van verkeersovertredingen en criminele activiteiten word geacht gelijk te zijn overheen te verschillende etniciteiten. Hier kan mogelijks ook een verschil in bestaan (al haalt het onderzoek zelf wel aan dat er geen bewijs voor het tegendeel gevonden is).
3. Ordehandhaving is niet uniform verspreid in tijd en ruimte, wat betekent dat niet iedere mobilist evenveel kans heeft om tegengehouden te worden door een agent.

De 1e assumptie is redelijkerwijs de belangrijkste binnen de context van dit onderzoek. Aangezien er in de beschikbare data geen directe informatie beschikbaar is over het aantal gereden kilometers, ontbreekt dit natuurlijk in de modellering. Een tegenargument hiervoor is echter dat dit onderzoek geen ruwe cijfers bestudeerd en op basis van verschillen in deze cijfers conclusies gaat trekken, maar verbanden tussen cijfers. Zo wordt het verband tussen ongevallen en overtredingen en het verband tussen overtredingen en boetes/ingetrokken rijbewijzen bestudeerd. Aangezien er in de literatuur bewijs aanwezig is dat het aantal geregistreerde overtredingen de belangrijkste predictor is van het aantal ongevallen bijvoorbeeld, zelfs wanneer er gecontroleerd wordt voor gereden kilometers, blijft deze assumptie overeind.

De 2e assumptie is een moeilijke assumptie om te weerleggen. Er is geen onderzoek gedaan (gevonden) naar verschillen in criminele activiteiten bij verschillende etniciteiten in Rusland. De auteur zelf haalt echter aan dat er in de literatuur ook geen tegenbewijs gevonden is voor deze assumptie, dus is dit niet echt een probleem.

De laatste aangehaalde assumptie gaat over de verspreiding van de ordehandhaving in ruimte en tijd. Hier zou inderdaad het argument gemaakt kunnen worden dat agenten meer ingezet worden in regio's in de buurt van Moskou waar procentueel gezien meer minderheden wonen. Anderzijds, aangezien er zoals gezegd voor de vorige assumptie geen direct tegenbewijs gevonden is, kan dit ook een vorm van profilering inhouden en wordt het exacte doel van dit onderzoek dus wel bevestigd.

Het weerleggen van de problemen met deze assumpties is duidelijk geen waterdicht betoog, maar valt buiten de scope van dit onderzoek.

3.2 Dataset en -pipeline

In dit onderdeel wordt de dataset en bijbehorende data-pipeline besproken. Dit omvat het volledige proces, gaande van een beschrijving van de dataset in zijn ruwe vorm, over de transformatie- en verrijkingsoperaties die erop uitgevoerd zijn tot het exploreren en cleanen van deze dataset. Alle stappen worden hier beschreven in de volgorde dat ze voorkomen in de werkelijke data pipeline. Dit proces is echter niet lineair verlopen (om een dataset te kunnen cleanen moet er een duidelijk beeld zijn van de data bijvoorbeeld), zoals voorgesteld in dit document.

3.2.1 Herkomst en beschrijving

Een eerst beschrijving van de dataset in zijn ruwe vorm gaat over de externe karakteristieken van de dataset (herkomst, aanwezige variabelen, structuur, etc.). In een volgende stap zal er meer een deep-dive genomen worden in de dataset, waarbij er gekeken wordt naar verdelingen, correlaties, etc.

De dataset is een verzameling van ongeveer 104 verkeersregistratiedatabases die gelekt zijn door 50 verschillende regionale overheden, gecombineerd met één grote samengestelde database (die verschillende geografische regio's bestrijkt). Deze gelekte data is vervolgens verzameld en gecombineerd door Tom Eeckhout en dit heeft geresulteerd in de dataset die

in dit onderzoek gebruikt wordt. Verder is de dataset reeds eerder gebruikt in het werk van Braguinsky et al. [7] om een specifiek type van corruptie te bestuderen in Moskou.

De dataset bestaat uit twee grote categorieën van registraties, waarbij de tweede categorie twee interessante subcategorieën bevat:

- Algemene verkeersregistraties (adreswijziging, inschrijving, uitschrijving, etc.)
- ongevallen en overtredingen

De voornaamste focus van dit onderzoek situeert zich in de categorie van ongevallen en overtredingen. Logischerwijs kan deze groep onderverdeeld worden in 2 overlappende groepen, nl. de ongevallen en de overtredingen. ongevallen kunnen gepaard gaan met een overtreding, en overtredingen kunnen gelinkt worden aan een ongeval. Een ongeval kan echter plaatsvinden zonder dat er een overtreding wordt vastgesteld en omgekeerd kan een overtreding ook plaatsvinden zonder dat dat die een ongeval moet veroorzaken. Beide groepen worden om die reden dus ook afzonderlijk geëxploreerd.

De overtredingen kunnen echter nog verder onderverdeeld worden in meer gedetailleerde subcategorieën. Zo kan bijvoorbeeld een overtreding al dan niet beboet worden (subcategorie 1: beboete overtredingen) en kan bij een vastgestelde overtreding het rijbewijs in sommige gevallen ingetrokken worden (subcategorie 2: overtredingen waarbij het rijbewijs wordt ingetrokken).

De andere groep van registraties, de grootste groep, zal voornamelijk gehanteerd worden als een proxy van de populatie. Op deze manier kan deze gebruikt worden om een indicatie te geven over de verdeling van etniciteiten, gender, kostprijs van wagens, e.d. in het verkeer in Rusland.

In conclusie hebben we dus deze high-level structuur in de data:

- Algemene verkeersregistraties
- ongevallen en overtredingen
  - ongevallen
    - ongevallen veroorzaakt door een overtreding
    - ongevallen zonder overtreding
  - Overtredingen
    - Beboete overtredingen

    * Overtredingen waarbij het rijbewijs ingetrokken wordt

In totaal hebben we 14 715 264 registraties, en voor iedere registratie zijn er, voor het cleanen van de dataset, 102 variabelen beschikbaar. Een overzicht van deze variabelen kan gevonden worden in Appendix 1.

In een eerste fase werden er 35 variabelen uit de dataset gelaten. Het merendeel van deze variabelen werden niet relevant geacht voor het onderzoek (relatieve prijs die betaald werd voor de wagen, nummer van het rijbewijs, etc.), anderen werden geëncapsuleerd in andere variabelen (codering van de naam, regiocode van de chauffeur, etc.)

De resterende 67 variabelen werden vervolgens geëxploreerd.

3.2.2 Exploratie 1

De exploratie beschreven in deze stap is de ruwe exploratie die gedaan is op de initiële dataset, voordat er data cleaning op is uitgevoerd, voor de verrijking met etniciteit, etc.

Deze exploratie van de data gebeurde in 3 stappen. In een eerste stap werd de data op een high-level manier beschouwd, en gekeken naar populatiebrede karakteristieken als referentiekader voor verder onderzoek. Het tweede deel focust zich specifiek op de ongevallen, en bestudeert de populatiekenmerken van deze groep registraties. Ten slotte worden ook de populatiekenmerken van de overtredingen en alle subcategorieën daarvan bestudeerd.

3.2.2.1 Algemeen

In de algemene exploratie werd er gekeken naar de hele dataset. In een poging om een referentiekader op te bouwen waarbinnen de andere registraties bekeken kunnen worden werden een aantal kenmerken bestudeerd.

Zo werd er onder meer gekeken naar :

- Hoeveel unieke individuen zijn er in de dataset aanwezig en hoeveel keer komt ieder individu voor?
- Hoe is de data verspreid in de tijd, wanneer vonden de registraties plaats?
- Hoe verdelen de leeftijden van de chauffeurs zich overheen de registraties?
- Wat is de verdeling van gender binnen de registraties: wat is de ratio mannen/vrouwen?

- Volgens bovenstaande categorieën, hoeveel datapunten hebben we voor iedere categorie afzonderlijk? Hoe verhouden de categorieën zich?

Algemeen - Individu Om individuele personen te gaan identificeren in de dataset, moet er eerst vastgesteld worden wat de unieke identifier is van een individu. Hiervoor baseert men zich binnen deze data op een eenvoudige veronderstelling: er zijn geen 2 verschillende mensen met dezelfde combinatie van voornaam (imya), familienaam (familya), patroniem (otchestvo) en geboortedatum. Op basis van deze veronderstelling kan er een index opgebouwd worden voor de dataset, die bestaat uit de combinatie van familya, imya en otchestvo (de volledige naam van een persoon) en zijn geboortedatum.

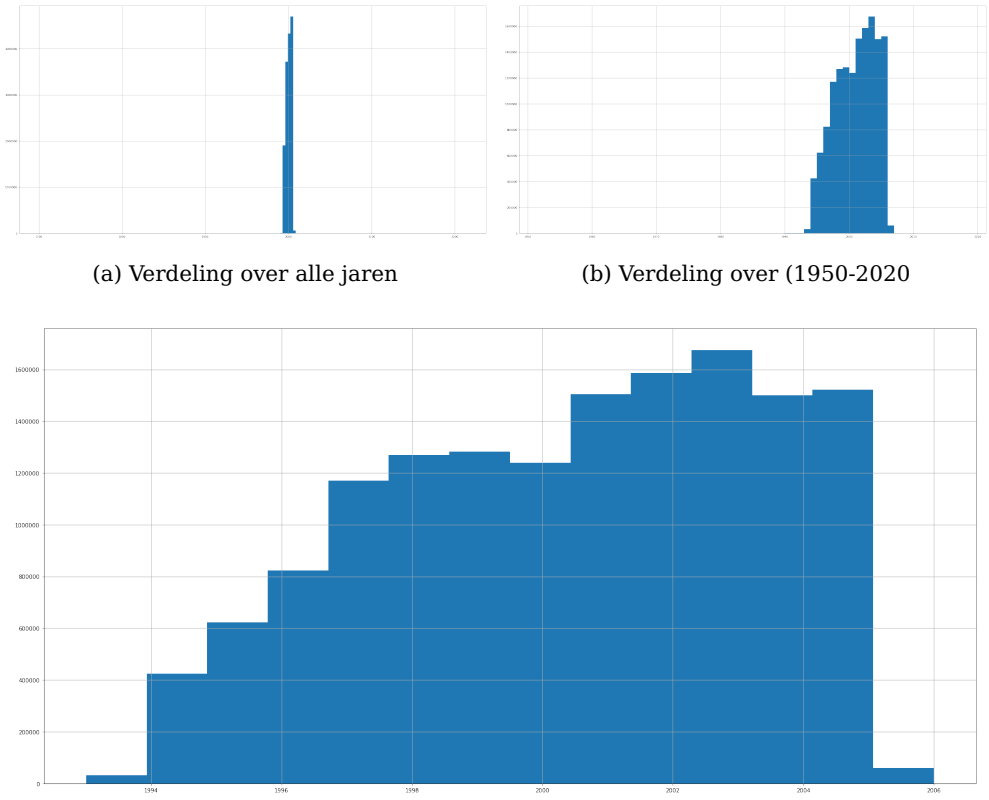
Op basis van deze index vinden we op 14 715 264 registraties 8 332 254 individuen. Dit komt neer op gemiddeld ongeveer 1.77 registraties per persoon.

Meteen wordt duidelijk dat deze dataset behoorlijk sparse is: er zijn veel registraties beschikbaar, maar weinig recurrente individuen. Dit is iets waar later rekening mee gehouden zal moeten worden.

Algemeen - Tijd Een belangrijke parameter in deze dataset is `registration_date`, die de tijdsdimensie voorstelt voor de registraties. `registration_date` is de datum waarop een bepaalde registratie heeft plaatsgevonden (in tegenstelling tot `birth_date`, wat de geboortedatum is van de chauffeur). In eerste instantie wordt er bekeken hoe de registraties verdeeld zijn over de jaren heen. Het vroegste jaar waarvoor een registratie beschikbaar is in de dataset is 1702. Dit is natuurlijk volkomen onrealistisch, aangezien de eerste auto's pas ontwikkeld zijn in 1885. Verder is het laatste jaar waarvoor er data beschikbaar is 2212, wat opnieuw duidelijk verkeerd is. Voorgaande observaties zijn een illustratie van de onzuiverheden in de data die hoogstwaarschijnlijk het gevolg zijn van fouten bij het bijeenvoegen van verschillende databronnen, of bij incorrecte data-invoer in de politiesystemen.

Het is dan ook belangrijk dat er in de data-cleaning hiermee rekening wordt gehouden. Om dit op een correcte manier te doen wordt er gekeken naar de verdeling over alle jaren heen (zie Figuur 3.1a).

Het is duidelijk dat de registraties zeer sterk geconcentreerd zijn rond de jaren 2000. Door de jaren te beperken tussen 1950 (post-WO II) en 2020, kan er duidelijk geobserveerd worden dat de registraties zich in de tijd voornamelijk tussen 1990 en 2010 bevinden. Als een arbitraire drempelwaarde van 30 000 datapunten gehanteerd wordt, dan worden



(a) Verdeling over alle jaren (b) Verdeling over (1950-2020

Figure 3.1: Verdeling over (1993-2006)

enkel die registraties tussen 1993 en 2006 (eindjaren inclusief) geselecteerd. Dit zijn de datapunten die ook verder bestudeerd zullen worden.

Tussen 1993 en 2006 ziet de verdeling van registraties over de jaren heen eruit als getoond in Figuur 3.1

Algemeen - Leeftijd chauffeur Het is duidelijk dat leeftijd een belangrijke rol kan spelen op ongevallen en overtredingen. Enerzijds speelt leeftijd een rol in het stellen van risicovol gedrag, zoals gesteld in Rolison et al. [40], waarbij oudere mensen minder snel risico's gaan nemen. Daarnaast hebben oudere mensen in het algemeen genomen ook meer ervaring achter het stuur. Daartegenover staat dan weer dat ouderdom ook gepaard gaat met het aftakelen van bepaalde fysieke eigenschappen, zoals zicht, reactievermogen en dergelijke.

In deze sectie wordt dus bekeken hoe de dataset verdeeld is inzake leeftijd. Om de leeftijd te berekenen wordt simpelweg het verschil genomen tussen de datum van registratie en de geregistreerde geboortedatum. Al snel werd hier echter een probleem bij ontdekt. In verschillende registraties bleek dat de registratie plaatsvond voor de geboorte-

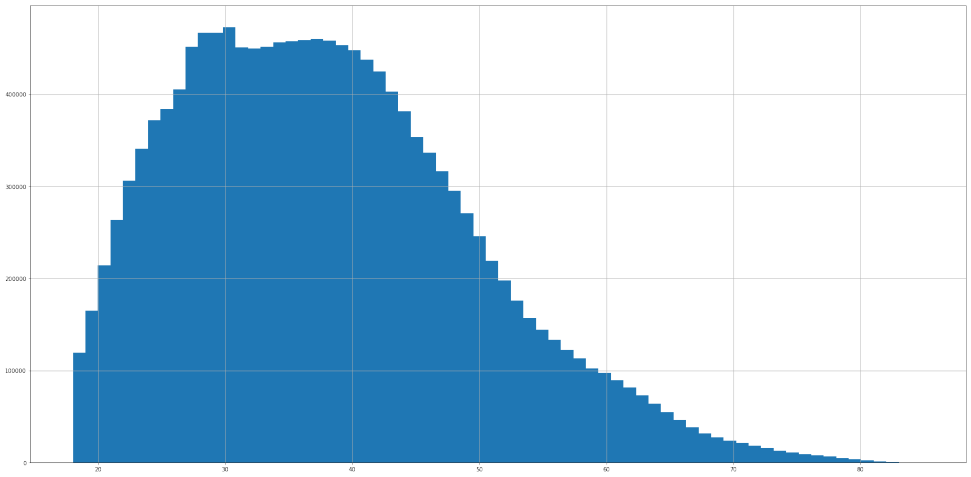


Figure 3.2: Aantal registraties per leeftijd

datum. Opnieuw is dit natuurlijk volkomen onrealistisch, en dergelijke registraties werden dan ook uit de dataset gehaald.

Een eerste filtering was dus die registraties filteren waarvoor `registration_date < birth_date`. Daarna werd de leeftijd berekend als `registration_date - birth_date`. Hierbij werd er geobserveerd dat er zeer onwaarschijnlijke leeftijden tussen de registraties zaten. Zo waren er 940 registraties met een leeftijd jonger dan 10 jaar, 4509 registraties jonger dan 15 jaar en 42860 registraties jonger dan 18 jaar. Voornamelijk voor die eerste categorie lijkt het zeer onwaarschijnlijk en in ieder geval niet representatief dat de chauffeurs jonger zouden zijn dan 10 jaar. De keuze werd dan ook gemaakt om alle registraties met een leeftijd jonger dan 18 jaar (de wettelijke minimumleeftijd om te mogen rijden in Rusland) te weren uit de dataset.

Het resultaat van deze stap was een overzicht van het aantal registraties per leeftijd. Deze kan gevonden worden in Figuur 3.2. De kern van de registraties ligt duidelijk tussen 25-35. Dat kan logisch verklaard worden aangezien deze groep mensen oud genoeg zijn om een eigen auto te bezitten, en toch nog jong genoeg zullen zijn om zeer actief op de baan te zijn. Naargelang de leeftijd toeneemt dalen het aantal registraties weer sterk.

Algemeen - Gender Inzake genderverdeling, is er duidelijk een dominante aanwezigheid van mannen bij de registraties. Zo zijn er 11 851 086 registraties waarvoor het geregistreerde gender 'Man' is, terwijl er maar 2 557 293 registraties zijn die geregistreerd staan als 'Vrouw'. Een verklaring hiervoor kan zijn dat tijdens de periode die deze dataset bestrijkt de gendergelijkheid in Rusland nog niet zo sterk was. Het gender dat geregistreerd

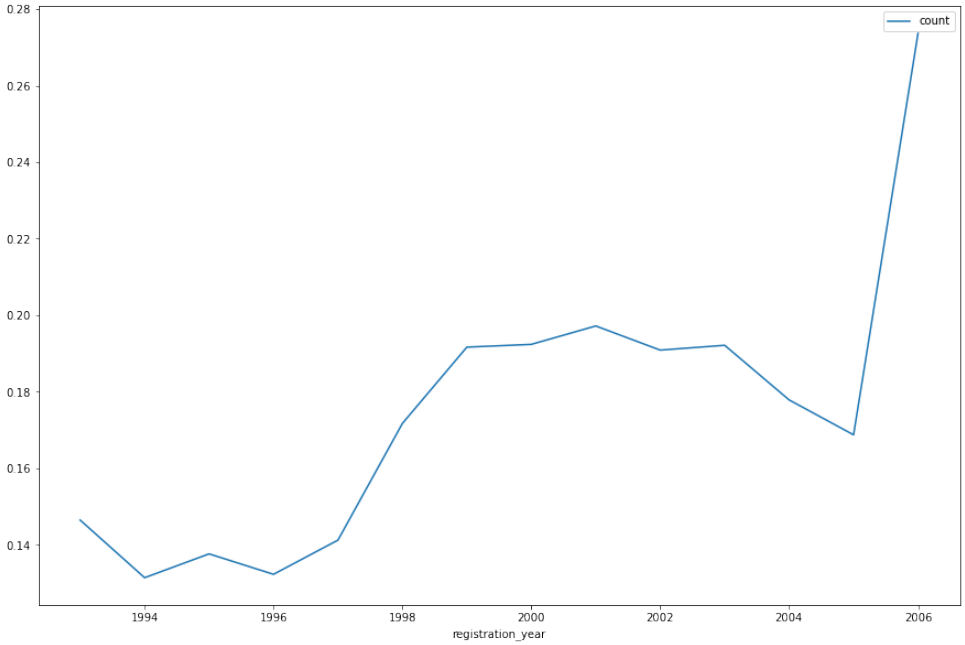


Figure 3.3: Percentage vrouwelijke registraties doorheen de jaren

is, is ook het gender van de persoon op wie de wagen is ingeschreven, en aangezien, in een patriarchale maatschappij, bij gezinnen met 1 wagen deze hoogstwaarschijnlijk ingeschreven zal zijn op de man, verklaart dit gedeeltelijk de overrepresentatie van mannen in de dataset.

Anderzijds als men kijkt naar de evolutie van het percentage registraties met gender 'Vrouw' doorheen de jaren, dan valt er wel een duidelijke algemene stijgende trend op te merken (zie Figuur 3.3).

Algemeen - etniciteit Op dit moment is de variabele etniciteit nog niet aanwezig in de dataset. In een later stadium zal de dataset verrijkt worden met deze gegevens.

Algemeen - Categorieën Zoals reeds aangegeven in sectie 3.2.1 zijn de registraties die de kern vormen van dit onderzoek de ongevallen en de overtredingen. Deze 2 groepen overlappen gedeeltelijk. Verder kunnen de overtredingen meer in detail gespecifieerd worden als overtredingen die beboet zijn en overtredingen waarvoor het rijbewijs is ingetrokken.

In deze sectie wordt er gekeken naar de verhoudingen binnen deze (sub)groepen. Voor ieder type registratie bestaat er in de dataset een dummy-variabele. Dat betekent dat een ongeval wordt aangegeven door `ongeval_dummy == 1`, een overtreding door `violation_dummy`

== 1, etc. Door simpelweg te tellen hoeveel registraties van ieder type er zijn, wordt de volgende verdeling bekomen.

Van de 14 715 264 registraties in totaal zijn er

- 4 088 642 ongevallen en overtredingen
- 10 625 418 andere registraties

Verder zijn er in meer detail

- 599 849 ongevallen waarvan
  - 237 490 gepaard gingen met een overtreding
- 3 727 927 overtredingen waarvan
  - er 2 888 574 beboet zijn
  - er bij 214 320 registraties een rijbewijs werd ingetrokken.

Een rijbewijs werd steeds ingetrokken wanneer er ook een overtreding werd vastgesteld (`license_deprived_dummy` kan dus nooit 1 zijn wanneer `violation_dummy` verschilt van 1). Ook een boete werd uitsluitend vastgesteld wanneer er een overtreding werd vastgesteld. Een rijbewijs kon echter wel ingetrokken worden zonder dat er een boete werd uitgeschreven.

Ongeveer de helft van de ongevallen zijn dus veroorzaakt door een overtreding, en slechts 10% van de overtredingen veroorzaken werkelijk een ongeval. Het is verder ook mogelijk dat er een overtreding werd vastgesteld, maar dat er geen boete werd uitgeschreven of rijbewijs werd afgenomen. Dit gebeurde bij 59 224 registraties.

3.2.2.2 ongevallen

De eerste subgroep die bestudeerd wordt zijn de ongevallen. Zoals vastgesteld in sectie 3.2.2.1, is er een sterke overlap met de categorie overtredingen (ongeveer de helft van de ongevallen wordt veroorzaakt door een overtreding). Bij de ongevallen worden specifieke parameters bestudeerd die mogelijks van belang kunnen zijn bij een ongeval.

- Leeftijd van de chauffeur
- Seizoen waarin het ongeval plaatsvindt

- Kostprijs van de auto
- Kracht van de auto
- Leeftijd van de auto

Verder worden deze parameters ook bekeken in het licht van de onderverdeling in ongevallen gepaard met een overtreding als oorzaak tegenover de ongevallen zonder.

ongevallen - Leeftijd chauffeur In deze sectie worden de leeftijden van chauffeurs onderzocht voor de ongevallen. Daarna worden deze vergeleken voor de verschillende subcategorieën en afgetoetst aan de algemene verdeling van leeftijden uit sectie 3.2.2.1. Een eenvoudige berekening van de gemiddelde leeftijd levert voor alle registraties een gemiddelde leeftijd op van 37,93 jaar. Als er enkel naar ongevallen gekeken wordt, dan verlaagt deze gemiddelde leeftijd meteen al naar 36,04 jaar, bijna 2 jaar jonger. Nog extremer wordt het wanneer er uitsluitend gekeken wordt naar de ca. 50% ongevallen die veroorzaakt werden door een overtreding. Hierbij is de gemiddelde leeftijd 34,31 jaar, oftewel meer dan 3,5 jaar jonger dan tegenover de algemene populatie. Wanneer er daarentegen uitsluitend gekeken wordt naar ongevallen die niet veroorzaakt werden door een overtreding, dan wordt het verschil een stuk minder afgetekend, namelijk een gemiddelde leeftijd van 37,18 jaar, slechts iets meer dan een half jaar jonger dan de algemene populatie.

Op basis van deze resultaten kan er een argumentatie opgebouwd worden dat jongere chauffeurs meer risicovol gedrag vertonen en, gecombineerd met een gebrek aan ervaring op de weg, hierdoor meer ongevallen veroorzaken.

Een visuele voorstelling van deze verdelingen kan gevonden worden in Figuur 3.4. Hierin kan duidelijk gezien worden dat de bulk van de ongevallen met overtreding (groen) op een jongere leeftijd ligt dan die voor de ongevallen zonder overtreding (rood). Beide liggen eerder dan de algemene populatie (blauw), als is het verschil met de ongevallen zonder overtreding niet heel groot.

ongevallen - Seizoen Bij ongevallen speelt de periode van het jaar natuurlijk ook een belangrijke rol. In de winter ligt er bijvoorbeeld misschien sneeuw en/of ijs, waardoor de baan gladder ligt en er meer ongevallen zullen gebeuren. Anderzijds is het ook vroeger donker in de winter, waardoor mensen misschien minder buiten zullen komen en meer thuis zullen blijven. Anderzijds is het in de winter langer klaar, waardoor mensen misschien meer

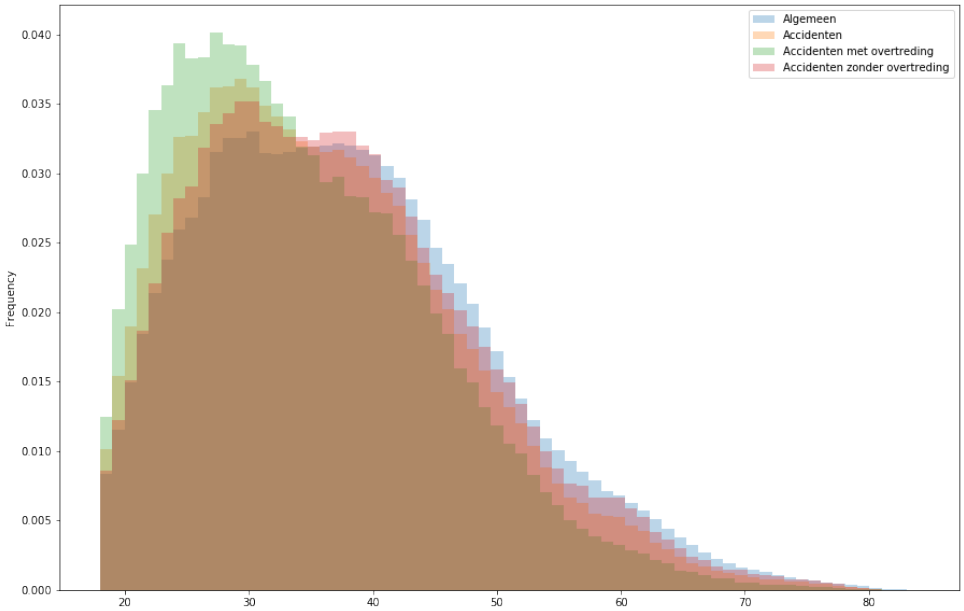


Figure 3.4: Leeftijdsverdeling in respectievelijk de algemene populatie, de populatie aan ongevallen, ongevallen zonder overtreding en ongevallen met overtreding

buiten zullen komen, langer wegblijven etc. De invloed van het seizoen op ongevallen is dus een interessante parameter om eens te exploreren.

Eerst wordt op basis van de registratiedatum het seizoen berekent. Hiervoor worden de volgende data gebruikt:

- Lente: 21 maart tem. 20 juni
- Zomer: 21 juni tem. 22 september
- Herfst: 23 september tem. 20 december
- Winter: 21 december tem. 20 maart

De totale verdeling van registraties over de seizoenen is behoorlijk gelijk verdeeld: 4 102 974 in de lente, 3 920 925 in de zomer, 3 355 472 in de herfst en 3 084 126 in de winter. Er is dus wel een licht dalende trend richting het najaar. Hier is geen specifieke verklaring voor, maar het is wel opmerkelijk. Een mogelijkheid is dus dat tijdens de lente en de zomer er simpelweg meer *exposure* is van chauffeurs op de baan, mensen gaan vaker de auto ergens naartoe nemen, blijven langer op de baan, etc.

Vervolgens worden 4 verdelingen over de seizoenen heen visueel bekeken: de algemene verdeling van registraties, de ongevallen, de ongevallen met overtreding en de ongevallen

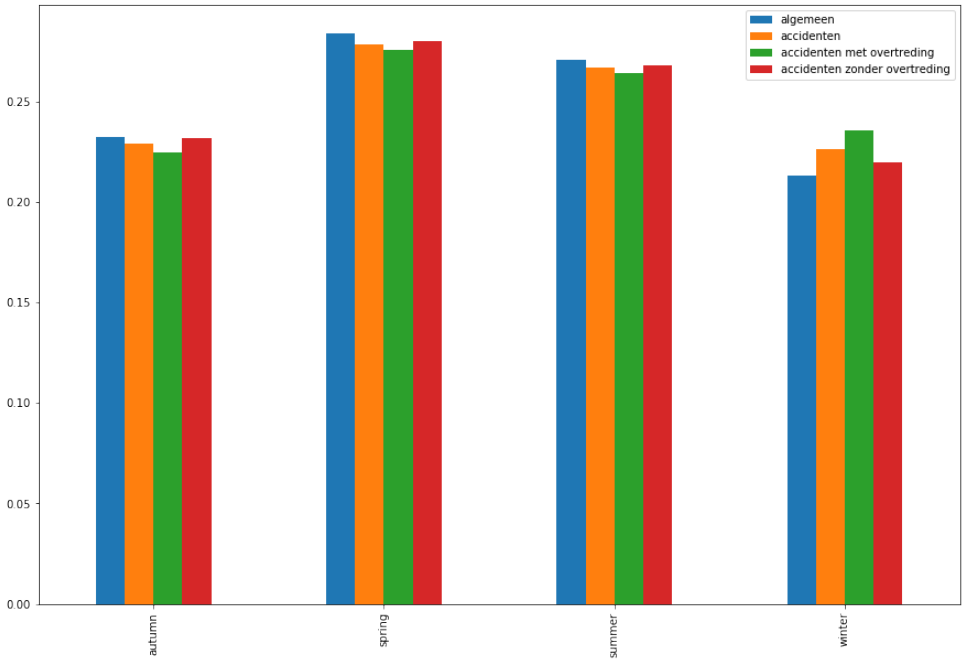


Figure 3.5: Verdeling van de registraties, ongevallen, ongevallen met overtreding en ongevallen zonder overtreding overheen de seizoenen - genormaliseerd

zonder overtreding. Het resultaat van deze (genormaliseerde) visualisatie kan gevonden worden in Figuur 3.5. 1 opmerkelijke vaststelling bij deze verdeling is dat de ratio aan ongevallen die veroorzaakt worden door een overtreding in de winter een stuk hoger ligt dan in de andere seizoenen. Dit wordt sterker duidelijk wanneer deze verhouding gevisualiseerd wordt overheen de seizoenen in Figuur 3.6.

Het mag duidelijk zijn dat het seizoen een invloed speelt op ongevallen, specifiek op de verhouding ongevallen zonder overtreding versus ongevallen met overtreding. Dit zal belangrijk zijn bij het opstellen van modellen later.

ongevallen - Kostprijs wagen Deze sectie gaat over de karakteristieken van de wagen. Vanzelfsprekend spelen de kritieke eigenschappen van een wagen ook een rol in de kans op een ongeval. Zo zullen duurdere wagens betere rijhulpsystemen hebben, meer betrouwbare remmen, etc. Daarentegen zullen mensen die meer op de baan zijn, mogelijks ook nieuwere (bedrijfs)wagens hebben, en zal de exposure voor nieuwere wagens dus ook mogelijks groter zijn.

Wagens in de dataset worden onderverdeeld in dure en goedkope voertuigen. De grens daarvoor ligt tussen 224 768,39 en 224 777,15, en door die grens te hanteren zijn er

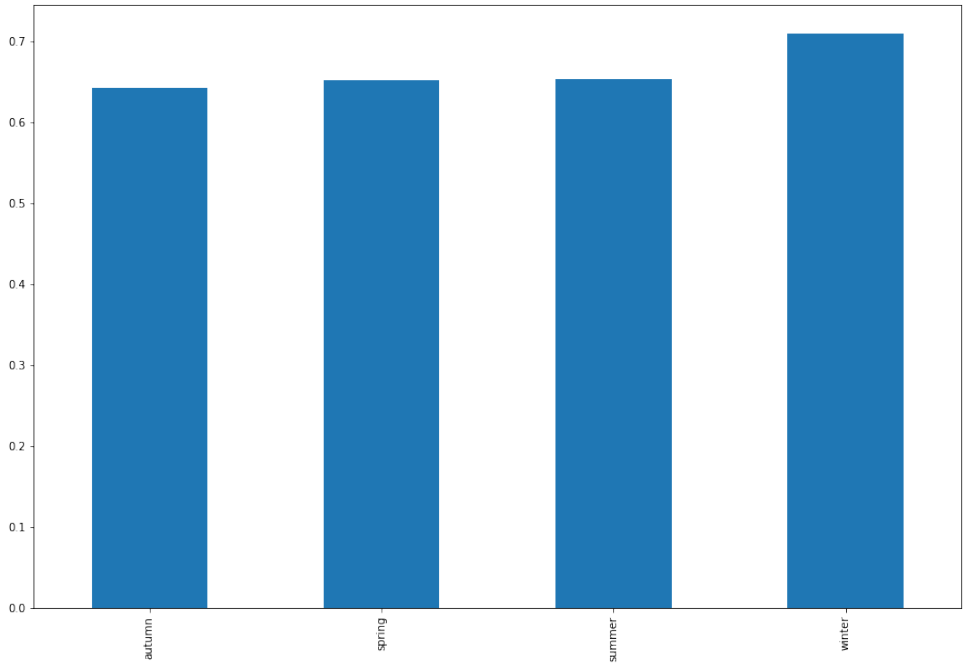


Figure 3.6: Verhouding ongevallen veroorzaakt door overtreding tegenover ongevallen zonder overtreding overheen de seizoenen

337 063 goedkope voertuigen en 92 312 dure voertuigen (ongeveer een 3:1-verdeling). In Figuur 3.7 valt te zien dat goedkope wagens overgerepresenteerd zijn in de ongevallen met overtreding en dat dure wagens overgerepresenteerd zijn in ongevallen zonder overtreding.

Ook de kostprijs van de wagen blijkt interessant te zijn in de verklaring van ongevallen.

3.2.2.3 Overtredingen

De andere voorname interessante klasse van registraties zijn de overtredingen.

Zoals vermeld in sectie 3.2.1 zijn er bij overtredingen 2 verder definiërende variabelen. De eerste variabele is een dummy die aangeeft of de overtreding al dan niet beboet is, terwijl de andere variabele (opnieuw een dummy) aangeeft of er bij een overtreding al dan niet een rijbewijs is ingetrokken. Bij beboete overtredingen is er vervolgens informatie over de grootte van de boete, bij ingetrokken rijbewijzen is er informatie over de duur van de intrekking. In dit deel bekijken we de verdeling van deze variabelen overheen de leeftijd, etniciteit, kostprijs van de wagen (als proxy voor het inkomen), gender en vanity plates.

Deze laatste variabele geeft aan of een bepaald voertuig een hoge kans heeft om een

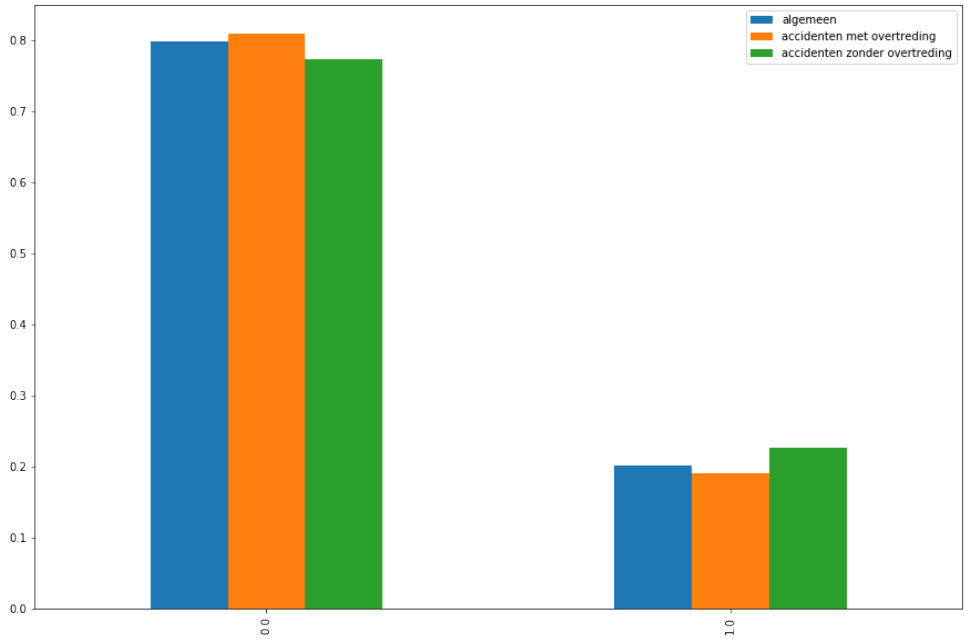


Figure 3.7: Verhouding dure - goedkope wagens in de ongevallen met en zonder overtreding

vanity plate te bezitten. Vanity plates zijn speciale nummerplaten die een speciale connectie, status of iets dergelijks aangeven. Er zou hierbij de vraag kunnen gesteld worden of deze vanity plates aanleiding geven tot gemiddeld lagere boetes, minder kans op een boete of iets dergelijks. Dat is dan ook wat er onderzocht wordt sectie 3.2.2.3.

Overtredingen - Leeftijd chauffeur De eerste stap bij het exploreren van de overtredingen is kijken naar hoe overtredingen zich verhouden in termen van leeftijd van de chauffeur. Hiervoor gaan wordt de verdeling van 4 groepen in termen van leeftijd uitgezet in een grafiek in Figuur 3.8: algemene registraties, overtredingen, beboete overtredingen en onbeboete overtredingen. Ter verduidelijking zijn die laatste 2 ook nog eens in een afzonderlijke grafiek geplaatst in Figuur 3.9. Het is duidelijk dat in vergelijking tot de populatie (de algemene registraties), het zwaartepunt van de andere 3 verdelingen een stuk vroeger ligt. Voor beboete overtredingen ligt het zwaartepunt het jongst, wat zeer duidelijk zichtbaar is in Figuur 3.9.

Het is duidelijk dat jongere mensen relatief meer beboet worden als ze een overtreding maken, dan wanneer oudere mensen een overtreding maken (extra verduidelijkt in Figuur 3.10). Hiervoor kunnen er verschillende verklaringen zijn: jongere chauffeurs kunnen simpelweg sneller beboet worden door minder vertrouwen van de politieagent in kwestie, maar

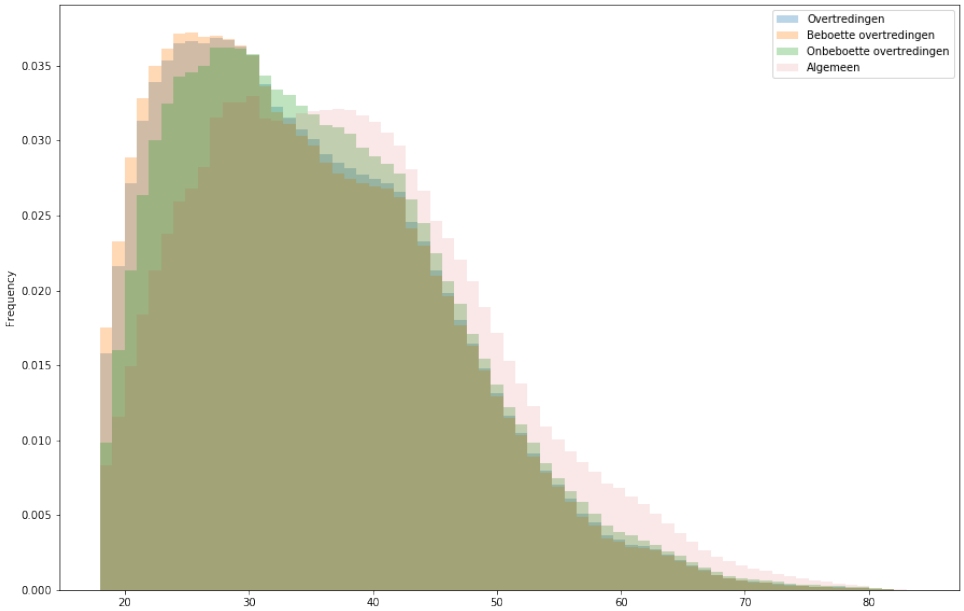


Figure 3.8: Verdeling van overtredingen, beboete overtredingen, onbeboete overtredingen en algemene registraties over de leeftijden

een andere verklaring is dat jongere chauffeurs typisch ook meer risicovol gedrag zouden vertonen en daardoor dus ook vaker in ongevallen zouden betrokken geraken. In sectie 3.2.2.2 is duidelijk geworden dat het inderdaad voornamelijk jongeren zijn die betrokken zijn bij ongevallen veroorzaakt door een overtreding. Die bevinding ondersteunt de argumentatie dat een hogere ratio aan beboete overtredingen een weerspiegeling is van hoger risicovol gedrag.

Bij een deel van de uitgeschreven boetes is er verder ook informatie beschikbaar over de grootte van de boete. Vergelijken we de gemiddelde boete-grootte overheen de verschillende leeftijden, dan resulteert dat in Figuur 3.11. Er zit dus een duidelijke dalende trend, waarbij een toenemende leeftijd overeenkomt met een dalende gemiddelde boete-grootte. Dit kan opnieuw op verschillende manieren verklaard worden, waarbij enerzijds jongere chauffeurs meer risicovol gedrag stellen en dus met zwaardere boetes zullen geconfronteerd worden. Anderzijds kan het ook zijn dat politieagenten bevooroordeeld naar confrontaties met jongere chauffeurs gaan, en dus de neiging zullen hebben om ze zwaardere boetes uit te schrijven, terwijl ze op een mildere manier oudere mensen zullen behandelen.

Overtredingen - Gender Op dezelfde manier als voor de leeftijden in sectie 3.2.2.3 is het mogelijk om te kijken naar de verschillen inzake gender voor de ratio van overtredingen

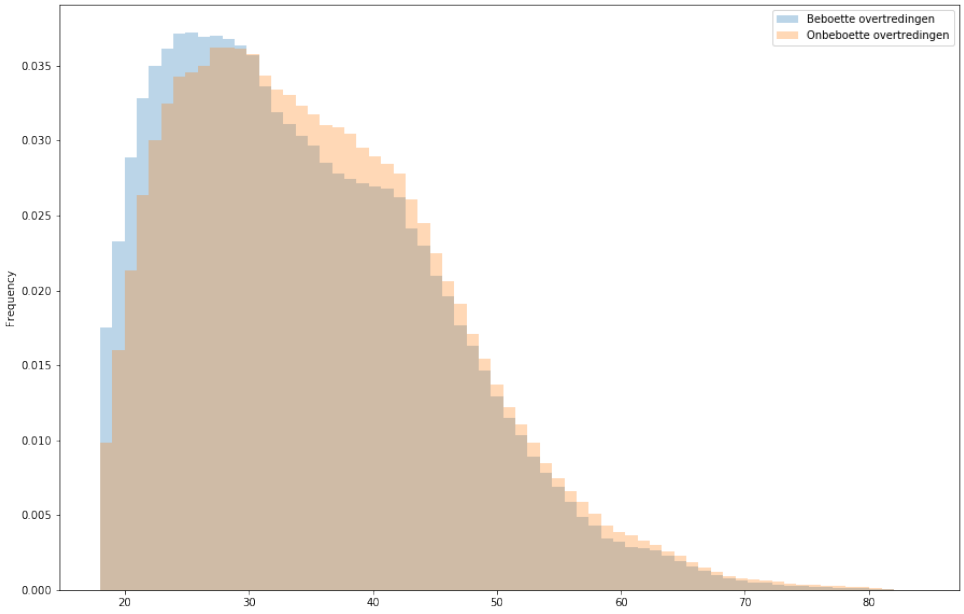


Figure 3.9: Verdeling van overtredingen, beboete overtredingen, onbeboete overtredingen en algemene registraties over de leeftijden

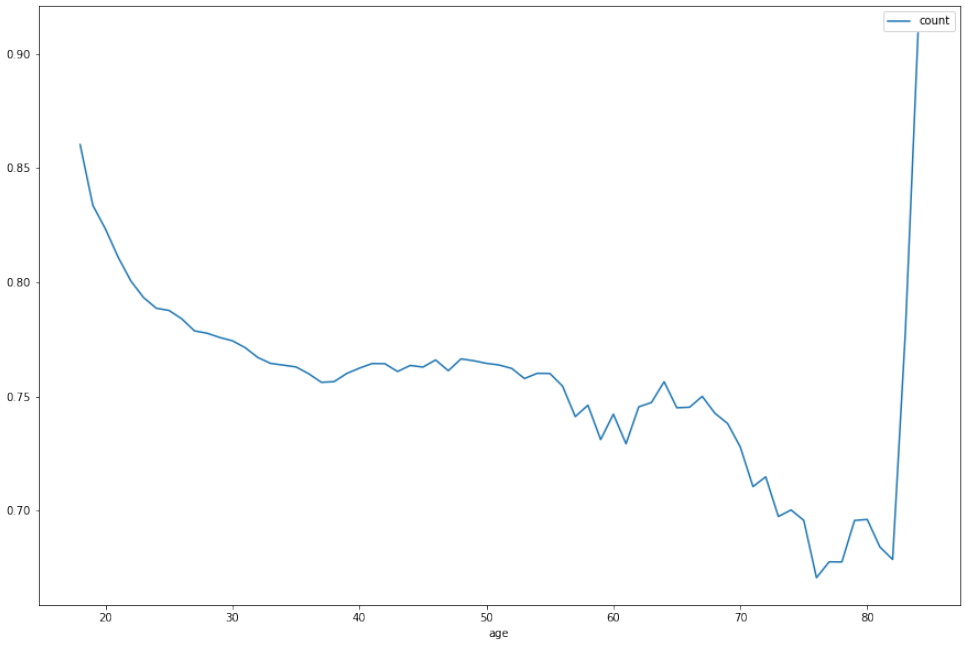


Figure 3.10: Ratio van beboete overtredingen tegenover totale overtredingen overheen de leeftijd

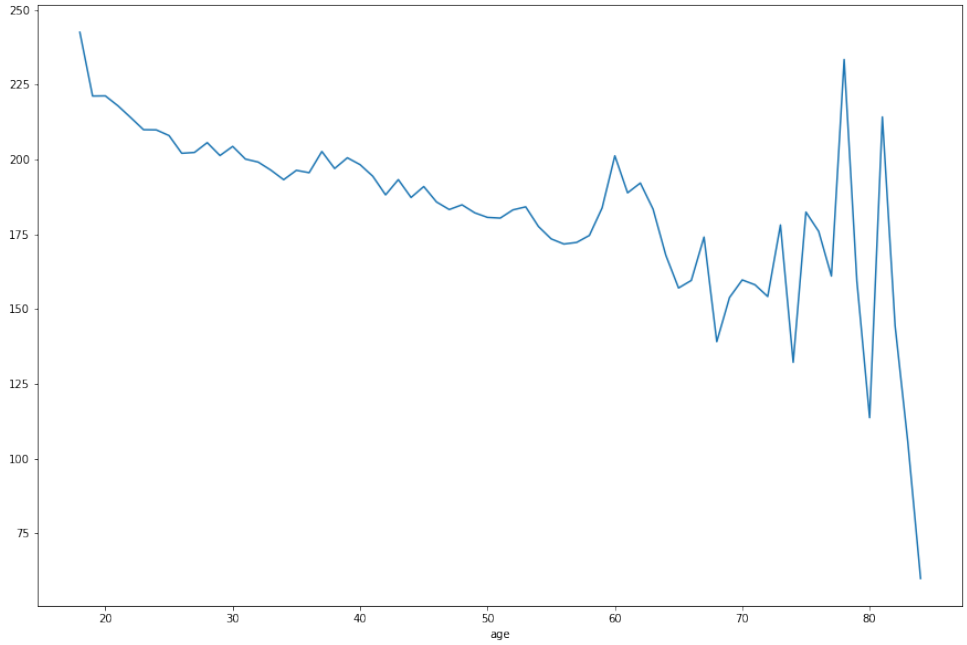


Figure 3.11: Verdeling van de gemiddelde boete-grootte overheen de leeftijden

die beboet worden. Door per leeftijd het aantal boetes te delen door het totaal aantal overtredingen en deze ratio in een grafiek te plotten, kunnen we de verschillen bestuderen tussen mannen en vrouwen in hun behandeling door de politie. Het resultaat hiervan is te vinden in Figuur 3.12. Uit deze grafiek kan afgeleid worden dat mannen, overheen alle leeftijden, ongeveer 10% meer boetes krijgen voor overtredingen dan vrouwen. Enkel in de oudste leeftijden, na 70, is er wat variabiliteit, maar dit is voornamelijk veroorzaakt door een laag aantal samples.

Het mag duidelijk zijn dat gender een relevante voorspeller is voor de kans op een boete. Dit is belangrijk om mee te nemen in verdere stappen waar we modellen ontwikkelen om deze kans te voorspellen.

Overtredingen - Kostprijs wagen Ook voor overtredingen wordt de invloed van een dure wagen bestudeerd op de kans op een overtreding. Hiervoor wordt hetzelfde principe gehanteerd als bij sectie 3.2.2.3: voor iedere leeftijd wordt het percentage beboete overtredingen vergeleken tussen dure en goedkope wagens (opnieuw zoals in 3.2.2.2 aan de hand van de dummy variabele ExpCar). De resultaten van deze exploratie zijn te vinden in Figuur 3.13. In tegenstelling tot bij gender in de vorige sectie, wordt er voor een dure wagen geen voordeel of nadeel gevonden. Door het kleinere aantal registraties met een dure wagen is

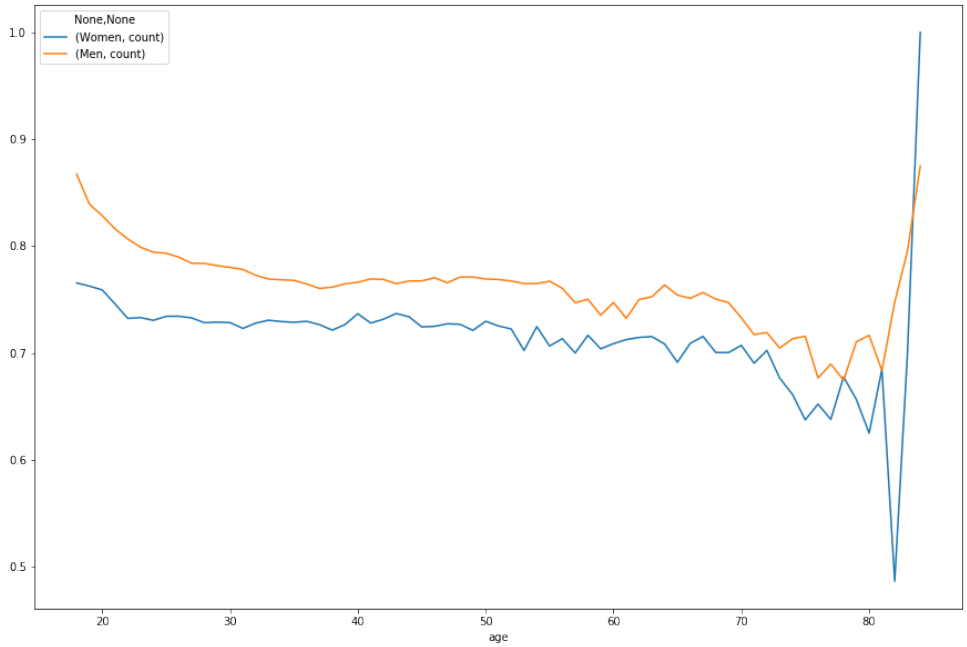


Figure 3.12: Ratio van beboete overtredingen tegenover totale overtredingen overheen de leeftijd voor 2 genders

er wel meer variabiliteit in de data, maar er wordt geen consistent verschil overheen de leeftijden vastgesteld.

Overtredingen - Vanity plate Analoog aan de kostprijs van de wagen, wordt ook de invloed van een vanity plate bekeken op de kans op een boete. Een vanity plate is een niet-officieel type nummerplaat die verworven wordt door middel van geluk, omkoping of connecties bij de juiste personen. Vermoed wordt dat deze vanity plates gebruikt worden als een signaal naar de politie voor een voorkeursbehandeling. Om die reden wordt dan ook verwacht dat bij een vanity plate de kans op een boete bij een overtreding lager zou liggen.

De resultaten van deze analyse kunnen gevonden worden in 3.14. Opnieuw, net zoals in sectie 3.2.2.3 wordt hier geen consistent verband gevonden, en enkel vastgesteld dat er door het lagere aantal samples, een grotere variabiliteit zichtbaar is bij de vanity plates.

Overtredingen - Ingetrokken rijbewijzen De laatste belangrijke klasse van registraties zijn die registraties waarbij het rijbewijs van de chauffeur wordt ingetrokken. Een rijbewijs wordt steeds ingetrokken bij een overtreding, dus deze klasse is een subklasse van

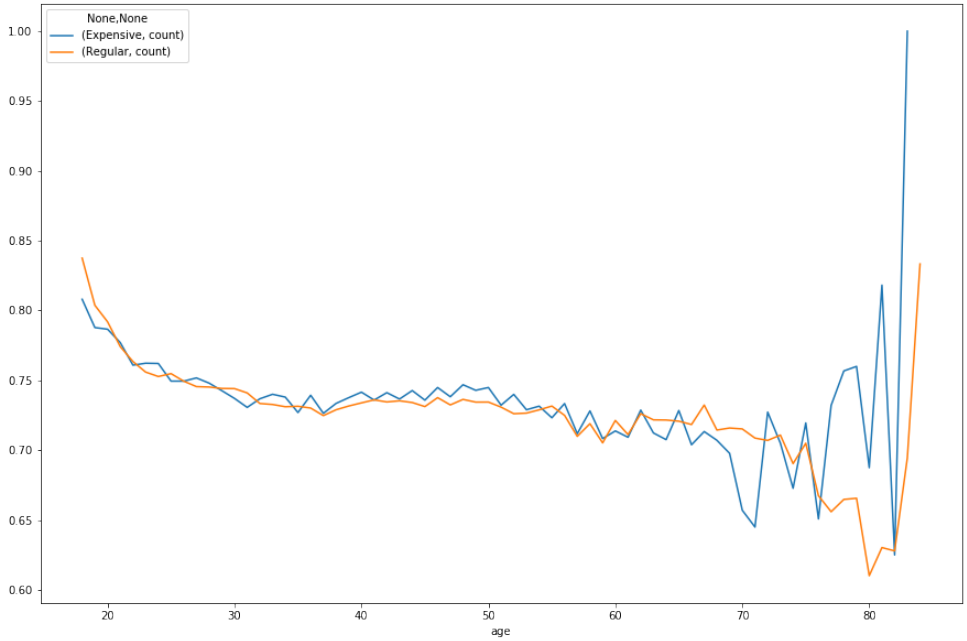


Figure 3.13: Ratio van beboete overtredingen tegenover totale overtredingen vergeleken tussen dure en goedkope wagens

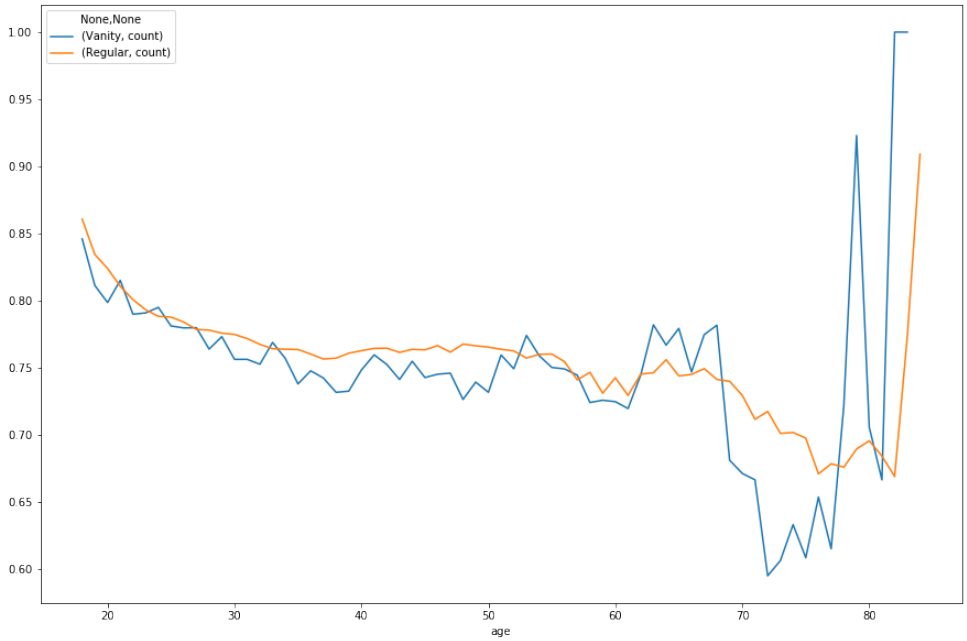


Figure 3.14: Ratio van beboete overtredingen tegenover totale overtredingen in functie van de leeftijd, voor vanity plates en gewone nummerplaten

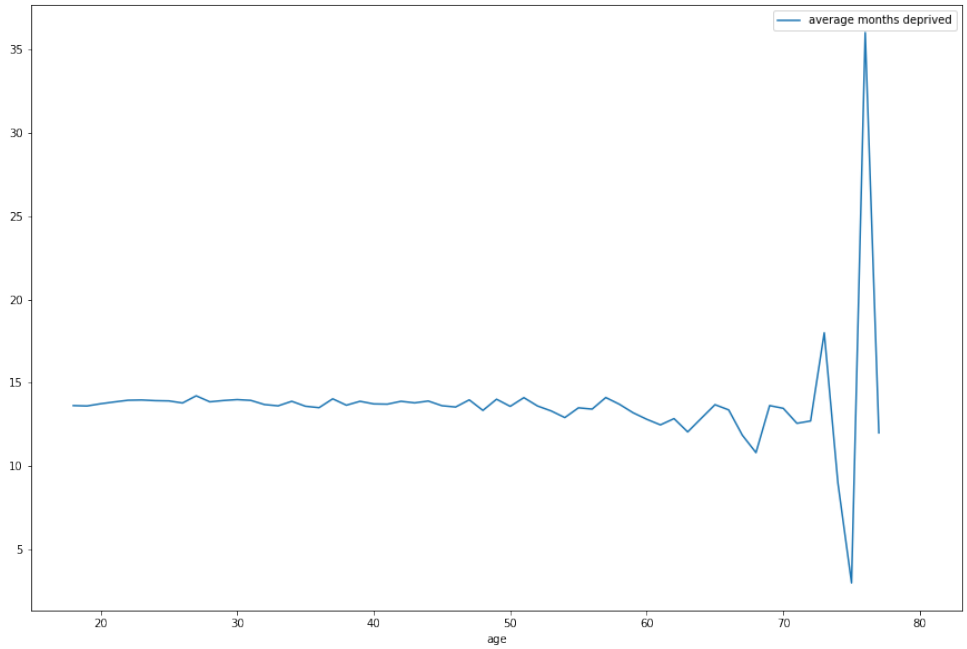


Figure 3.15: Gemiddeld aantal maanden ingetrokken rijbewijs overheen de leeftijden

de overtredingen, maar gaat niet noodzakelijk gepaard met een boete. Er is wel degelijk een overlap tussen de beboette overtredingen en de ingetrokken rijbewijzen, maar die is niet compleet. Zo zijn er 59 224 overtredingen geregistreerd waarbij het rijbewijs is ingetrokken, zonder dat er een boete is uitgeschreven (van de in totaal 211 551 registraties waarbij er een rijbewijs is ingetrokken).

Voor het merendeel van deze registraties is er ook beschikbaar hoeveel maanden het rijbewijs is ingetrokken. Kijkt men vervolgens naar de verdeling van het gemiddeld aantal maanden dat een rijbewijs is ingetrokken over de leeftijden heen, dan vindt men deze keer geen dalend verband (zoals bij de gemiddelde boete-groote), maar een eerder constant verband, geïllustreerd door Figuur 3.11.

Als er een subset genomen wordt van de registraties waarbij er wel een boete is uitgeschreven, dan kan men het verband onderzoeken tussen het aantal maanden dat een rijbewijs wordt afgenomen en de grootte van de boete. Deze variabelen blijken behoorlijk sterk positief gecorreleerd, met een correlatiecoëfficiënt van 0,32. Dit wordt ook weergegeven in Figuur 3.16. In deze figuur stelt de grootte van de bubble het aantal registraties met die combinatie van grootte van boete en aantal maanden ingetrokken rijbewijs. Er is inderdaad een licht stijgende trend op te merken in de data, met uitzondering van een aantal outliers links bovenaan.

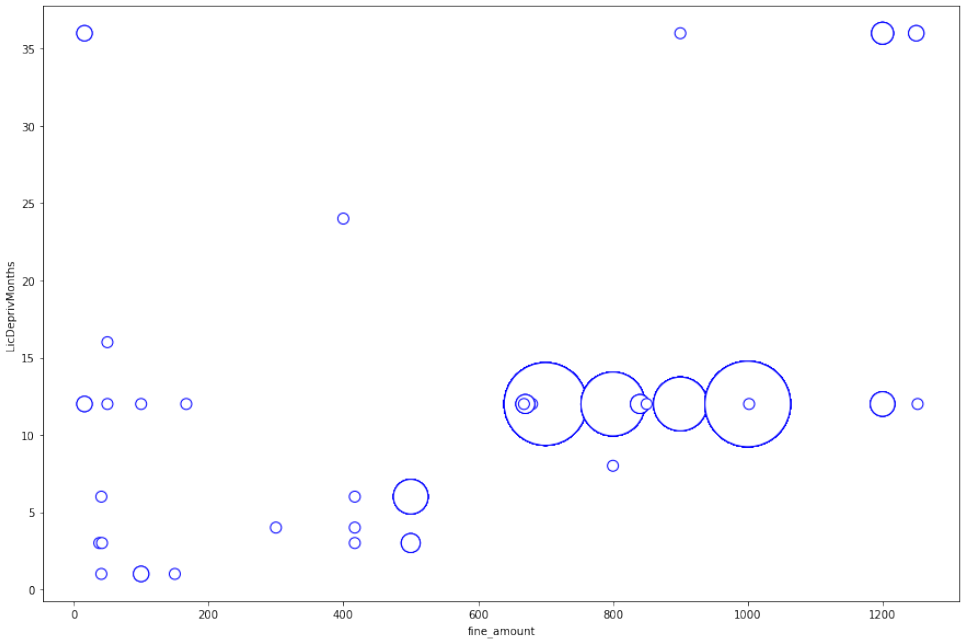


Figure 3.16: Gemiddeld aantal maanden ingetrokken rijbewijs overheen de leeftijden

3.2.3 Data Preprocessing 1

Na het verkrijgen van een betere inzage in de data, was het een kwestie om de dataset voor te bereiden om wiskundige analyses te kunnen uitvoeren. Op basis van de bekomen informatie uit het vorige hoofdstuk, de data-exploratie, wordt de data zo goed mogelijk opgekuist, ontbrekende waarden woorden ingevuld, onbelangrijke kolommen worden uit de dataset gelaten, etc.

3.2.3.1 Truncation

In een eerste stap worden irrelevante kolommen uit de dataset geweerd. Variabelen die niet van belang zijn voor onze analyse, worden genegeerd en weggelaten.

Variabele	**Verklaring**
person_index	(Verondersteld) unieke identifier van een persoon, gebaseerd op de combinatie van voornaam, familienaam en patroniem enerzijds, en zijn geboortedatum anderzijds.
first_date	Eerste datum waarop de persoon geregistreerd staat in de database

last_date	Laatste datum waarop de persoon geregistreerd staat in de database
productionyear_car	Het productiejaar van de wagen.
birth_date	Geboortejaar van de chauffeur
police_department	Politie departement die de registratie gemaakt heeft
gender	Het geregistreerde gender van de chauffeur
violated_dummy	Dummy-variabele die aangeeft of de registratie een overtreding betreft
fined_dummy	Dummy-variabele die aangeeft of de registratie een boete betreft
license_deprived-_dummy	Dummy-variabele die aangeeft of de registratie een ingetrokken rijbewijs betreft
ongeval_dummy	Dummy-variabele die aangeeft of het een ongeval betreft
ongeval_or_violation-_dummy	Dummy-variabele die aangeeft of het een ongeval en/of overtreding betreft
expensive_dummy	Dummy-variabele die aangeeft of de auto geclassificeerd wordt als 'duur'. De grens hiervoor blijkt te liggen tussen 224 777,15 en 224 768,39 roebels.
suspect_issuance-_dummy	Dummy-variabele die aangeeft of er een hogere kans is dat de nummerplaat verkregen is door middel van corruptie, connecties, etc.
passport_below100-_dummy	Dummy-variabele die aangeeft of de cijfers van het paspoort onder de 100 liggen, wat een signaal zou kunnen zijn voor speciale connecties, mogelijks een hogere kans op corruptie, etc.
military_passport-_dummy	Dummy-variabele die aangeeft of het een militair paspoort betreft, wat mogelijks preferentiële behandeling zou kunnen veroorzaken.
chauffeur_class	Klasse van chauffeur (obv. rijbewijs): A, B C, ...
adress_city	Geregistreerde stad van de chauffeur.
issue_date_passport	Uitgavedatum van het passpoort
familya	Familenaam van de chauffeur

imya	Voornaam van de chauffeur
otshestvo	Patroniem, naam die aangeeft hoe de vader van de naamdrager heet.
passport_region	Regio aangegeven op het passport van de chauffeur
passport_country	Land aangegeven op het passport van de chauffeur
passport_type	Er zijn 2 types passporten, namelijk de oude Sovjet-passporten of de nieuwe, Russische passporten.
passport_region_code	Code van de regio die geregistreerd staat op het passport.
horsepower	Paardekracht van de wagen
license_deprived-_months	Aantal maanden dat het rijbewijs wordt ingetrokken, ingeval `license_deprived_dummy == 1`.
vanity_dummy	Dummy-variabele die aangeeft of het een vanity-nummerplaat betreft. Een vanity plate is een gepersonaliseerde nummerplaat, die mogelijks een signaal van corruptie, connecties, e.d. kan aangeven.
car_brand	Merk van de wagen
car_model	Specifiek model van het wagenmerk

Hiermee blijven er dus 38 kolommen over. De variabelenamen zijn voor consistentie allemaal in python-style casing geschrveen (lowercase met underscores) om de leesbaarheid van de code te verbeteren.

3.2.3.2 Cleaning

In een volgende stap proberen we de onzuiverheden uit de data te halen. Tijdens de data-exploratie (zie vorig hoofdstuk) viel het op dat sommige variabelen zeer onwaarschijnlijke, of zelfs onmogelijke waarden vertonen. Voorbeelden zijn registraties uit 1702, chauffeurs van 2 jaar oud, en zo verder. Het doel is dus om deze 'fouten' in de data te filteren, en verder te werken met een dataset met waarden die logisch steek houden.

Registratiedatum De eerste variabele waarop we gaan filteren is de registratiedatum. Hierbij werden er tijdens de data-exploratie een aantal vreemde waarden vastgesteld. Ten

eerste gingen de registratiedata van de jaren 1702 (voor de wagen was uitgevonden) tot 2212. Door het grote aantal registraties in totaal, keken we dus naar de bulk van de registraties en werden deze beperkt tussen de jaren 1993 en 2006 (eindpunten inclusief).

In deze stap ging de dataset van 14 715 264 rijen naar 14 706 921 registraties, wat betekent dat er ongeveer 9 000 registraties wegvielen.

Geboortedatum Hierbij wordt de geboortedatum-variabele (aangeduid door `birth_date`) gebruikt om verder te filteren. Op basis van de geboortedatum en de registratiedatum kan de leeftijd van de chauffeur ten tijde van de registratie berekend worden. Nu, door de interne werking van de gebruikte library (pandas), zorgt een negatieve datum (wanneer de registratiedatum voor de geboortedatum ligt) voor een overflow. Dit gaf meteen een eerste probleem in de dataset aan, waarbij een persoon reeds in de database voorkomt, voordat hij/zij geboren is. In een eerste stap werden deze registraties verwijderd.

Vervolgens werd ook de assumptie gemaakt dat er een onzuiverheid in de data zat bij die registraties waarbij de berekende leeftijd kleiner was dan 18 jaar (de wettelijke minimumleeftijd om te mogen rijden in Rusland). Alle registraties met leeftijd kleiner dan 18 jaar werden dus ook verwijderd uit de dataset.

In deze stap ging de dataset van 14 706 921 naar 14 543 485 rijen. In totaal betekent dat dus een afname van ruwweg 150 000 rijen.

Seizoen In deze sectie voegen we het seizoen toe aan de dataset. Uit de data-exploratie blijkt dat het seizoen toch een invloed heeft op ongevallen, dus gaan we die variabele dan ook toevoegen.

Voor het seizoen worden de volgende data gebruikt, zoals aangegeven in 3.2.2.2:

- Lente: 21 maart tem. 20 juni
- Zomer: 21 juni tem. 22 september
- Herfst: 23 september tem. 20 december
- Winter: 21 december tem. 20 maart

3.2.4 Ethnicity Prediction

In deze stap van de data pipeline wordt de dataset verrijkt met een voorspelling van de etniciteit van een persoon. Dit is een belangrijke variabele aangezien die later in het on-

derzoek een prominente rol gaat spelen.

De dataset werd verrijkt met de voorspelde etniciteit op basis van het Recurrent Neuraal Netwerk (RNN) van Feliciaan De Palmenaer [10]. Hiervoor werd de volgende methode gehanteerd:

1. De voornaam (*imya*), familienaam (*familya*) en patroniem (*otshestvo*) werden als input aan het RNN gegeven.

2. Dit resulteerde in de output van het RNN, namelijk een tensor van 21 elementen tussen 0 en 1, waarbij ieder element 1 voorspelde etniciteit voorstelt.

3. De etniciteit met de hoogste corresponderende waarde werd geselecteerd, en de waarde tussen 0 en 1 werd beschouwd als het vertrouwen dat het netwerk had in zijn voorspelling.

Op die manier bekwamen we voor iedere registratie een voorspelling van de etniciteit. Een eerste snelle kijk geeft echter aan dat er slechts ruwweg 30% etnische Russen aanwezig waren in deze voorspellingen. Dit is nogal onrealistisch, aangezien de census in Moskou rond die periode aangeeft dat etnische Russen aan ongeveer 90% voorkomen. Een mogelijke verklaring hiervoor is dat de trainingsdata waarop het Recurrent Neuraal Netwerk getraind is, nogal anders verdeeld is dan de etniciteiten van de applicatiedata. De trainingsdata voor het RNN was afkomstig uit 2 datasets:

- Een website over slachtoffers van politiek geweld die bijgehouden werd door de Soviet-Unie waar zowel de namen als de etniciteit van een persoon beschikbaar waren.

- Een dataset die de gegevens bevat van de Russische slachtoffers in de Tweede Wereldoorlog

De verdeling van deze datasets is een stuk anders, ook omdat deze datasets verzameld zijn in een volledig andere tijdsperiode dan onze data. Om die reden werd er besloten om een correctie van de data door te voeren. Hiervoor baseren we ons op de census data voor Moskou [1]. Daarin ziet de etnische samenstelling van Moskou er uit als volgt:

- Etnisch Rus: 91,6%

- Oekraïner: 1,42%

- Tataar: 1,38%

- Armeen: 0,98%
- Azeri: 0,5%
- Jews: 0.49%
- Belarusian: 0.4%
- Uzbek: 0.3%
- Tajik: 0.2%
- Moldovan: 0.2%
- Mordvin: 0.2%
- Chechen: 0.1%
- Chuvash: 0.1%
- Ossetians: 0.1%
- Other: 1.6%

Met behulp van deze etniciteitsverdeling en met de gegenereerde voorspellingstensors werd er een correctietensor opgesteld. Dit werd gedaan door de som te nemen van alle gegenereerde voorspelde tensoren (die bestaan uit waarden tussen 0 en 1), deze te normaliseren zodat we een tensor bekwamen van de relatief voorspelde, geaggregeerde kans dat een bepaalde etniciteit voorspeld zou worden. De tensor met de etniciteitsverdeling in Moskou werd vervolgens gedeeld door de bekomen tensor met relatieve geaggregeerde kans.

$$N = [n_1, n_2, ..., n_l] \quad \text{Etniciteitstensor} \tag{3.1a}$$

$$P = [p_1, p_2, ..., p_l] \quad \text{Geaggregeerde voorspellingen} \tag{3.1b}$$

$$C = N/P = [\frac{n_1}{p_1}, \frac{n_2}{p_2}, ..., \frac{n_l}{p_l}] \quad \text{De correctietensor} \tag{3.1c}$$

De bekomen methodologie werd dus de volgende:

1. De voornaam (*imya*), familienaam (*familya*) en patroniem (*otshestvo*) werden als input aan het RNN gegeven.

2. Dit resulteerde in de output van het RNN, namelijk een tensor van 21 elementen tussen 0 en 1, waarbij ieder element 1 voorspelde etniciteit voorstelt.

3. **Corrigeer de bekomen tensor door te vermenigvuldigen met de correctietensor zoals hierboven gedefinieerd**.

4. De etniciteit met de hoogste corresponderende waarde werd geselecteerd, en de waarde tussen 0 en 1 werd beschouwd als het vertrouwen dat het netwerk had in zijn voorspelling.

Belangrijk hierbij te vermelden is dat er geen 100% overeenkomst was tussen de voorspelde etniciteiten en de etniciteiten die geregistreerd staan in de census. Sommige etniciteiten zijn om die reden gedropt.

Aan het eind van deze stap werden de voorspelde etniciteiten terug in de dataset gebracht.

3.2.5 Data Preprocessing 2

Na het invoeren van de etniciteitsvoorspelling in de dataset en het aldus verrijken van de dataset met meer informatie, werd een nieuwe ronde van data preprocessing gedaan. De reden waarom dit in de data pipeline niet samengevoegd is met de vorige preprocessing stap is omdat het genereren van etniciteitsvoorspellingen voor de gehele dataset een zeer kostbaar gebeuren was in termen van tijd. Om die reden werd geopteerd om in plaats van de pipeline voor de *ethnicity prediction* aan te passen, een nieuwe preprocessing en data-exploratie stap toe te voegen na de *ethnicity prediction*.

In de tweede preprocessing stap was het doel om de dataset verder op te kuisen en specifiek ook een afzonderlijke dataset te gaan creëren die klaar was voor consumptie in ieder van de 3 experimenten.

3.2.5.1 Verfijnen

Op dit punt in de data pipeline waren er nog steeds onzuiverheden aanwezig in de dataset. In dit onderdeel wordt er aandacht besteed aan de volgende 3 onzuiverheden:

- Er waren een disproportioneel aantal registraties die geregistreerd waren op éénzelfde persoon, namelijk de persoon geïdentificeerd door de `fio_dob_index` (index opbgebouwd op basis van de naam en geboortedatum) 20 358 361. Deze persoon had zo

bijvoorbeeld meer dan 14 000 boetes begaan. Aangezien dat heel onwaarschijnlijk is in het tijdsbestek waarin de dataset vergaard is, is het vermoeden dat dit een soort van *umbrella*-registratie is, voor specifieke personen of diensten. Alle registraties met deze `fio_dob_index` werden dan ook uit de dataset geweerd

- Er waren ontbrekende genderwaarden aanwezig in de dataset. Voor sommige registraties was er geen gender opgegeven. Wat de reden hiervoor is, is niet helemaal duidelijk, maar de registraties zonder gender werden gedropt.

- Sommige registraties leken dubbel te hebben plaatsgevonden. Zo waren er ongeveer 10 000 registraties waarbij een persoon 2 ongevallen had op 1 dag en ongeveer 100 000 registraties waarbij een persoon 2 overtredingen begin per dag. Vanwege de *sparseness* van de data werd er gekozen om dergelijke dubbele registraties te weren uit de dataset.

Initieel werd er na de vorige stap van data preprocessing vertrokken van 14 543 484 registraties. Na het verwijderen van de registraties met `fio_dob_index == 20358361` bleven er nog 14 411 762 registraties over, wat betekent dat deze "persoon" ongeveer 130 000 registraties zou zijn bekomen in iets meer dan 10 jaar tijd.

Na het verwijderen van de rijen met ontbrekende gender waarden, bleven er nog 14 276 656 registraties over en na het verwijderen van de dubbele overtredingen/boetes is de dataset uiteindelijk gereduceerd tot 14 162 868 registraties.

3.2.5.2 Experimenten voorbereiden

Voor het 1e experiment, dat de invloed van etniciteit onderzoekt op het aantal boetes, werd de totale populatie van de dataset onderverdeeld naargelang leeftijd, gender en etniciteit. Over deze groepen werden vervolgens geaggregeerde statistieken getrokken:

- Het totaal aantal ongevallen van de groep en het gemiddeld aantal per persoon

- Het totaal en gemiddelde aantal overtredingen.

- Het totaal en gemiddelde aantal boetes.

- Het totaal aantal personen.

- De gemiddelde boetegrootte

- De gemiddelde kostprijs van de wagen, bepaald op basis van het merk en model van een wagen
- De ratio aan nummerplaten die geclassificeerd staan als vanity plates
- De ratio aan nummerplaten die geclassificeerd staan als verdachte uitgave
- De gemiddelde paardenkracht van een wagen

Hieraan werden dan etniciteitsdummies toegevoegd om te gebruiken in het model.

Voor het 2e experiment was het eenvoudiger, hiervoor hadden we simpelweg de overtredingen nodig.

Voor het laatste experiment was het een stuk moeilijker. Om aan survival analysis te doen, zijn er events nodig. Events zijn gebeurtenissen die een bepaalde duur hebben, met een beginpunt en een eindpunt. Binnen de context van dit onderzoek zijn events dus de periode's tussen enerzijds de 1e registratie en de eerste boete, en anderzijds de periode tussen 2 opeenvolgende boetes (indien dit het geval is). Om die reden werden de registraties gefilterd die `fined_dummy == 1` hadden, die vervolgens gesorteerd werden op `person_index` en `registratiedatum` (oplopend). Vervolgens werd er geïtereerd over de volledige DataFrame, waarbij het volgende algoritme gebruikt werd.

- Er werd een state bijgehouden, van de huidige `person_index, current_timestamp` en de reeds gegenereerde events.
- Voor iedere rij in de DataFrame werd er gekeken over de `person_index` overeenkwam met die in de state.
  - Als dit het geval was, dan was er een nieuwe boete gevonden, werd het tijdsverschil met de `current_timestamp` (in dagen) berekend en werd er een nieuw event toegevoegd. (1)
  - Als dit niet het geval was, dan waren we bij de boetes van een nieuw individu terechtgekomen. Als eerste event werd een event gecreëerd met als duur de tijd tussen de variabele `first_date`, wat de 1e datum van registratie (eender welke) is in de dataset en `registration_date`, de datum van de huidige rij. Dit werd enkel gedaan indien de huidige registratie niet ook de 1e registratie was. (2)

Op die manier werd er een dataset van events met een duur berekend waarop survival analysis kon toegepast worden. In geval (1) was dit dus de 1e boete voor dat individu, dit werd ook als nieuwe dummy-variabele meegegeven. In geval (2) was het individu dus reeds eerder in tijd beboet.

3.2.6 Exploratie 2

Een laatste stap in de data pipeline voor het uitvoeren van de analyses was het exploreren van de verrijkte dataset. Specifiek de verdeling in termen van etniciteiten was zeer van belang.

	ethnicity
Russian	12213549
Armenian	802923
Tatar	408484
Jewish	236433
Ukrainian	117928
Korean	109500
Kazakh	91048
Polish	86791
Chechen	59635
Bashkir	20238
Chinese	16339

Table 3.2: Verdeling van voorspelde etniciteiten

In tabel 3.2 kan de verdeling van voorspelde etniciteiten teruggevonden worden, terwijl tabel 3.3 de genormaliseerde verdeling toont. Het is duidelijk dat Russische etniciteit, in overeenkomst met de census data, de dominante etniciteit is (met 86,2%, wat iets minder is dan de censusdata weergeeft (90%). Dit is beter dan de originele, ongecorrigeerde voorspelling, waarbij Etnisch Rus ongeveer 30% voorkwam.

Het is duidelijk dat sommige minderheden relatief gezien zeer weinig voorkomen, zoals Chinees en Bashkir. Dit plaatst dan ook een kanttekening bij de conclusies die getrokken kunnen worden uit de data.

	ethnicity
Russian	0.862364
Armenian	0.056692
Tatar	0.028842
Jewish	0.016694
Ukrainian	0.008327
Korean	0.007731
Kazakh	0.006429
Polish	0.006128
Chechen	0.004211
Bashkir	0.001429
Chinese	0.001154

Table 3.3: Caption

Chapter 4

Resultaten

In dit hoofdstuk bespreken we de resultaten van de verschillende experimenten. In het eerste experiment wordt gekeken wat de invloed is van etniciteit op het aantal boetes die een persoon krijgt. Het tweede experiment kijkt naar de invloed die etniciteit heeft op de kans om een boete te ontvangen bij een overtreding. Hierbij wordt er ook gekeken naar de grootte van de boete, en welke rol etniciteit daarin speelt. Als laatste experiment kijken we naar recidivisme, hier wordt er onderzocht welke rol etniciteit speelt in de snelheid van recidiveren.

4.1 Aantal boetes

4.1.1 Model

Om het aantal boetes te berekenen wordt gebruik gemaakt van geaggregeerde data. Doordat er weinig mensen meerdere keren beboet zijn, is het logischer om hiervoor geaggregeerde data over leeftijd, gender en etniciteit te gaan gebruiken.

In een eerste stap worden er dus groepen gevormd:

$$a_i, j, k$$

waarbij i over de leeftijden 18 tot 84 gaat, j over de genders man (=0) en vrouw (=1) en k over de 11 verschillende etniciteiten. De verdeling van het aantal personen per groep ging als getoond in 4.1

Vervolgens werd per groep het gemiddeld aantal ongevallen per persoon (aantal ongevallen / uniek aantal individuen in de groep) en het gemiddelde aantal boetes per persoon berek-

	total_persons			
	sum	mean	max	min
ethnicity				
Armenian	799709	7014.991228	22774	6
Bashkir	19698	317.709677	814	6
Chechen	59027	952.048387	2280	5
Chinese	15658	200.743590	728	3
Jewish	234995	1958.291667	5166	4
Kazakh	90355	982.119565	3799	6
Korean	108420	1013.271028	2997	3
Polish	85887	802.682243	2484	2
Russian	12209330	94645.968992	325682	58
Tatar	407421	3573.868421	11614	51
Ukrainian	116448	1049.081081	2492	22

Table 4.1: Verdeling van personen over de groepen

end. Verder werd, binnen iedere groep, de gemiddelde prijs van een voertuig, de ratio aan vanity plates, de ratio aan verdachte nummerplaten en de gemiddelde paardenkracht van een wagen berekend. Aan de hand hiervan werd het eerste initieel model opgesteld:

$$\begin{aligned} avg\_fine = \alpha + \beta_{accident} \times avg\_accident + \\ \beta_{age} \times age + \\ \beta_{gender} \times gender\_dummy + \\ \beta_{cost\_car} \times avg\_cost\_car + \\ \beta_{horsepower} \times mean\_horsepower + \\ \beta_{suspect_issuance} \times ratio\_suspect\_issuance + \\ \beta_{vanity} \times ratio\_vanity + \\ \epsilon_{1->10} \times ethnicity_{1->10} \end{aligned} \tag{4.1}$$

Hierbij stelt $ethnicity_{1->10}$ een one-hot vector voor van etniciteiten, waarbij etnisch Rus als baseline gebruikt wordt (de bedoeling is om het verschil te onderzoeken in behandeling van andere etniciteiten).

4.1.2 Colineariteit

Na dit model voor een eerste keer te runnen werd de colineariteit van de predictors onderzocht. Hiervoor werd gebruik gemaakt van VIF, oftewel de *variance inflation factor*. Dit is een maatstaf voor de colineariteit tussen twee of meerdere variabelen. Een drempelwaarde van 3 werd geselecteerd als drempel om te spreken van colineariteit. Na het berekenen van de VIF voor iedere variabele werden de volgende resultaten bekomen (zie 4.2).

	VIF Factor	Variables
0	82.2	const
1	1.4	average_accidents
2	1.4	gender
3	1.6	age
4	1.6	average_cost_car
5	1.1	ratio_suspect_issuance
6	1.7	Armenian
7	1.4	Bashkir
8	1.5	Chechen
9	1.6	Chinese
10	1.7	Jewish
11	1.6	Kazakh
12	1.7	Korean
13	1.7	Polish
14	1.7	Tatar
15	1.7	Ukrainian

Table 4.2: VIF Factor van de verschillende variabelen

De variabelen *avg_cost_car* en *mean_horsepower* hebben dus een VIF-factor boven de drempelwaarde van 3. Om te kijken welke predictors met welke andere predictors correleren werd een correlatietabel opgesteld. Deze is te vinden in 4.3.

Op basis van deze tabel werd sowieso de beslissing genomen om *mean_horsepower,* die zeer sterk gecorreleerd was met *avg_cost_car* te verwijderen uit het model.

	avg_accidents	gender	age	avg_cost_car	ratio_si	ratio_van	mean_hp
avg_accidents	1.000000	0.186166	-0.481707	0.192960	-0.034345	0.018004	0.212784
gender	0.186166	1.000000	-0.138837	0.466842	-0.093005	0.199908	0.226735
age	-0.481707	-0.138837	1.000000	-0.390528	0.199264	-0.168319	-0.496913
avg_cost_car	0.192960	0.466842	-0.390528	1.000000	-0.061455	0.179147	0.773973
ratio_si	-0.034345	-0.093005	0.199264	-0.061455	1.000000	-0.059366	-0.213013
ratio_van	0.018004	0.199908	-0.168319	0.179147	-0.059366	1.000000	0.238324
mean_hp	0.212784	0.226735	-0.496913	0.773973	-0.213013	0.238324	1.000000

Table 4.3: Correlatietabel van de variabelen

Het overblijvende model ziet er uit als volgt:

$$\begin{aligned} avg\_fine = \alpha + \beta_{accident} \times avg\_accident + \\ \beta_{age} \times age + \\ \beta_{gender} \times gender\_dummy + \\ \beta_{car\_cost} \times avg\_car\_cost + \\ \beta_{vanity} \times ratio\_vanity + \\ \beta_{suspect\_issuance} \times ratio\_suspect\_issuance + \\ \epsilon_{1->10} \times ethnicity_{1->10} \end{aligned} \tag{4.2}$$

4.1.3 Resultaten

De resultaten van het model zijn te vinden in

Het eerste, duidelijke verband dat op te merken valt, werd voorspeld door de literatuur. Er blijkt een positief verband te zijn tussen het gemiddeld aantal boetes en het gemiddeld aantal ongevallen die een bepaalde groep heeft. Dit verband kan op verschillende manier verklaard worden, die allemaal besproken zijn in de literatuurstudie: mensen die vaker op de baan zijn hebben een hogere kans op ongevallen als boetes, het aantal overtredingen is positief gecorreleerd met, en een belangrijke voorspeller van het aantal ongevallen, risicogedrag, etc. Dit verband wordt dus bevestigd door onze resultaten.

Vervolgens zien we dat leeftijd en gender significant negatief gecorreleerd zijn met het gemiddelde aantal boetes. Dit betekent dat voor éénzelfde gemiddelde aantal ongevallen, oudere mensen en vrouwelijke mensen, minder boetes krijgen dan jongeren en mannen. Aangezien dit model geen rekening houdt met de ernst van de ongevallen en boetes, is er een mogelijkheid dat dit simpelweg het resultaat is van een gebrek aan informatie. Vervolgens is *average_cost_car* ook negatief gecorreleerd met boetes, hetgeen aangeeft dat hoe duurder de auto gemiddeld genomen is, hoe minder boetes. Ten slotte is de variabele *ratio_vanity* negatief gecorreleerd met het aantal boetes. Dit geeft aan dat bevolkingsgroepen waar *vanity plates* relatief gezien meer voorkomen, gemiddeld genomen minder boetes krijgen dan bevolkingsgroepen waar dergelijke nummerplaten minder voorkomen. Dit zou mogelijks een argument kunnen zijn voor de theorie dat deze nummerplaten een signalisatiefunctie vertonen voor connecties, geld, invloed, etc.

Als we kijken naar de invloed van etniciteiten, dan is het resultaat minder éénduidig. Bij sommige etniciteiten, zoals Tsjetsjenen en Chinezen wordt er een positieve correlatie

Dep. Variable:	average_fines	**R-squared:**	0.696
Model:	OLS	**Adj. R-squared:**	0.692
Method:	Least Squares	**F-statistic:**	154.5
Date:	Mon, 31 May 2021	**Prob (F-statistic):**	1.69e-265
Time:	23:49:57	**Log-Likelihood:**	1758.4
No. Observations:	1096	**AIC:**	-3483.
Df Residuals:	1079	**BIC:**	-3398.
Df Model:	16		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.3881	0.013	28.783	0.000	0.362	0.415
average_accidents	0.2750	0.087	3.147	0.002	0.104	0.446
gender	-0.1005	0.004	-28.109	0.000	-0.107	-0.093
age	-0.0033	0.000	-30.040	0.000	-0.003	-0.003
average_cost_car	-2.714e-07	4.9e-08	-5.541	0.000	-3.67e-07	-1.75e-07
ratio_suspect_issuance	0.0032	0.048	0.066	0.948	-0.092	0.098
ratio_vanity	-0.3603	0.102	-3.540	0.000	-0.560	-0.161
Armenian	0.0029	0.006	0.459	0.646	-0.010	0.015
Bashkir	-0.0398	0.008	-5.160	0.000	-0.055	-0.025
Chechen	0.0268	0.008	3.433	0.001	0.011	0.042
Chinese	0.0256	0.007	3.554	0.000	0.011	0.040
Jewish	-0.0079	0.006	-1.258	0.209	-0.020	0.004
Kazakh	0.0126	0.007	1.864	0.063	-0.001	0.026
Korean	0.0058	0.006	0.893	0.372	-0.007	0.018
Polish	0.0005	0.007	0.074	0.941	-0.012	0.013
Tatar	-0.0086	0.006	-1.356	0.175	-0.021	0.004
Ukrainian	-0.0069	0.006	-1.085	0.278	-0.019	0.006

Omnibus:	554.602	**Durbin-Watson:**	1.257
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	7557.942
Skew:	1.990	**Prob(JB):**	0.00
Kurtosis:	15.233	**Cond. No.**	1.24e+07

vastgesteld, wat aangeeft dat deze etniciteiten relatief gezien meer boetes krijgen in verhouding tot de ongevallen waarbij ze betrokken zijn. Voor Tsjetsjenen is dit verband niet onlogisch, het wantrouwen van de Russische bevolking is goed gedocumenteerd. Dit verband is niet ook niet min: zo betekent tsjetsjeen zijn - ceteris paribus - dat het gemiddeld aantal boetes per persoon stijgt met 20,76% en chinees zijn houdt een stijging in van 20,98%.

Voor een andere etniciteit (Bashkir) is er een negatieve correlatie. Dit geeft aan dat deze etniciteit ondergerepresenteerd is in de boetes in verhouding tot de ongevallen.

Over de resterende etniciteiten is er geen éénduidig verband. Dit kan het resultaat zijn van de methodiek, waar door het gebruik van geaggregeerde data (een deel van) de informatie mogelijks verloren gaat.

4.2 Kans op een boete/ingetrokken rijbewijs

4.2.1 Model

Voor dit experiment worden uit de dataset alle registraties gefilterd die een overtreding voorstelden. In deze registraties werd er onderzocht wat de kans was op een boete/ingetrokken rijbewijs wanneer er een overtreding werd vastgesteld, en wat de invloed van de etniciteit hierop was. Deze kans werd gemodelleerd als volgt:

$$\begin{aligned} P\_fine = \alpha + \beta_{accident} \times accident\_dummy + \\ \beta_{age} \times age + \\ \beta_{gender} \times gender\_dummy + \\ \beta_{vanity} \times ratio\_vanity + \\ \epsilon_{1->10} \times ethnicity_{1->10} \end{aligned} \tag{4.3}$$

Opnieuw stellen $ethnicity_{1->10}$ hier de 10 verschillende etniciteiten voor en wordt *Etnisch Russich* als baseline etniciteiten genomen.

Voor de technische implementatie werd een Logit-model van *statsmodels* gebruikt in python.

4.2.2 Colineariteit

Dezelfde procedure werd genomen voor colineariteit als bij het 1e experiment. Hierbij werden dezelfde variabelen geselecteerd als in het vorige experiment. De VIF-tabel kan

gevonden worden in 4.4. Merk op dat alle VIF-factoren kleiner zijn dan 10 en dus voldoen aan de drempelwaarde voor non-multicolineariteit.

	VIF Factor	Variables
1	1.0	accident_dummy
2	1.0	chauffeur_age
3	1.0	gender
4	1.0	vanity_dummy
5	1.0	Armenian
6	1.0	Bashkir
7	1.0	Chechen
8	1.0	Chinese
9	1.0	Jewish
10	1.0	Kazakh
11	1.0	Korean
12	1.0	Polish
13	1.0	Tatar
14	1.0	Ukrainian

Table 4.4: VIF Factor van de verschillende variabelen

4.2.3 Resultaten

De resultaten van het logistische regressiemodel zijn te vinden in 4.5

Opmerkelijk genoeg stellen we vast dat betrokkenheid bij een ongeval de kans op een boete verlaagt, overtredingen die gepaard gingen met een ongeval zijn maar 0.58 keer zo waarschijnlijk om beboet te worden dan overtredingen zonder ongeval. Voor dit opmerkelijke resultaat is er niet meteen een verklaring. Verder stellen we vast, in lijn met de verwachtingen dat een hogere leeftijd en vrouw zijn de kans op een boete ook verlaagd. Ieder jaar ouder dat een persoon wordt verlaagt de kans op een boete met ongeveer 1% en vrouwen zijn 23% minder likely om een boete te krijgen dan mannen. Ook het hebben van een vanity plate verlaagt de kans met ongeveer 6%. Voor de etniciteiten zijn de resultaten niet eenvoudig te interpreteren, aangezien de significantie moeilijk te beschrijven is door

Dep. Variable:	fined_dummy	**No. Observations:**	3528665
Model:	Logit	**Df Residuals:**	3528650
Method:	MLE	**Df Model:**	14
Date:	Mon, 31 May 2021	**Pseudo R-squ.:**	0.005491
Time:	23:54:45	**Log-Likelihood:**	-1.8778e+06
converged:	True	**LL-Null:**	-1.8881e+06

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	1.5399	0.004	353.765	0.000	1.531	1.548
accident_dummy	-0.5405	0.005	-115.627	0.000	-0.550	-0.531
chauffeur_age	-0.0067	0.000	-59.214	0.000	-0.007	-0.007
gender	-0.2563	0.004	-62.873	0.000	-0.264	-0.248
vanity_dummy	-0.0595	0.007	-8.342	0.000	-0.073	-0.045
Armenian	-0.0101	0.005	-1.944	0.052	-0.020	8.52e-05
Bashkir	-0.1474	0.035	-4.267	0.000	-0.215	-0.080
Chechen	-0.1017	0.017	-6.126	0.000	-0.134	-0.069
Chinese	-0.0359	0.044	-0.815	0.415	-0.122	0.050
Jewish	-0.0524	0.011	-4.967	0.000	-0.073	-0.032
Kazakh	0.0774	0.014	5.458	0.000	0.050	0.105
Korean	0.0121	0.017	0.730	0.466	-0.020	0.044
Polish	-0.0395	0.016	-2.467	0.014	-0.071	-0.008
Tatar	-0.0525	0.008	-6.994	0.000	-0.067	-0.038
Ukrainian	-0.0684	0.015	-4.672	0.000	-0.097	-0.040

Table 4.5: Logit Regression Results

het grote aantal samples (p-value probleem, beschreven in Lin et al. [29]), en de kleine z-waarden. Hier is dus geen duidelijk effect te observeren, zoals bij het vorige experiment, voor Tsjetsjenen en Chinezen, of andere etniciteiten.

We stellen dus opmerkelijk vast dat overtredingen die gepaard gingen met een ongeval maar 0.6 keer zo waarschijnlijk zijn om beboet te worden dan overtredingen zonder ongeval. Wanneer we echter gaan kijken naar de grootte van de boete, door deze lineair te regresseren over dezelfde predictors, dan zien we wel een positief verband voor de accident_dummy. Alhoewel er minder kans is op een boete (mogelijks wegens administratieve redenen), is de gemiddelde boete dus wel groter. Er moet echter wel opgemerkt worden dat dit model een zeer lage verklarende kracht heeft, met een R^2 van 0.001.

4.3 Recidivisme

In het laatste experiment doen we onderzoek naar de invloed die etniciteit heeft op recidivisme. De theorie hiervoor is gebaseerd op de *procedural justice theory* van Tyler [53] die aangeeft dat autoriteit aan de politie enkel gegeven wordt in die situaties waar de politie als *legitiem* ervaren wordt. Legitimiteit wordt op zijn beurt dan weer bepaald door een aantal factoren, onder dewelke effectiviteit en eerlijkheid. Zeker die laatste is van belang, aangezien bij etnische profilering bestuurders de politie niet als eerlijk zullen ervaren, waardoor, volgens de theorie, etnische minderheden meer recidivisme zouden moeten vertonen.

4.3.1 Survival Analyse

Om dit te onderzoeken wordt gebruik gemaakt van een statistische techniek genaamd *Survival Analysis*, meer specifiek *Cox proportional-hazards model*. Deze techniek, die voornamelijk gebruikt wordt in medisch onderzoek, onderzoekt het verband tussen de duur van events en verschillende predictors. Het doel van het model is om tegelijkertijd (multivariabele analyse) het effect van verschillende factoren op *survival* te onderzoeken. Het geeft aan hoe de gespecifieerde factoren de ratio van het plaatsvinden van een bepaald event beïnvloeden.

In deze context is het event gedefinieerd als de 2e keer dat een bepaalde bestuurder een boete ontvangt, de duur is de verstreken tijd sinds het ontvangen van de eerste boete. Analoog dus aan patiënten die geïntroduceerd worden in een medische studie en mogelijks

komen te overlijden, worden bestuurders dus in deze studie geïntroduceerd op het moment dat ze hun 1e boete ontvangen. Het overleven van de patiënt komt vervolgens overeen met het niet ontvangen van een boete voor de bestuurder. Wanneer de bestuurder ten slotte een boete ontvangt, komt dit overeen met de patiënt die overlijdt. Op deze manier kan survival analysis toegepast worden op dit onderzoek, en kan het verband van etniciteit op recidiviteit onderzocht worden.

4.3.2 Model

De algemene vorm van het Cox model wordt uitgedrukt door de *hazard function*, aangegeven door h(t). De *hazard function* h(t) geeft voor ieder tijdstip t het risico aan om te "sterven" (in deze beschrijving komt "sterven" overeen met "een 2e boete ontvangen"). De functie kan geschat worden met behulp van de volgende formule:

$$h(t) = h_0(t) \times e^{b_1 x_1 + b_2 x_2 + ... + b_p x_p} \tag{4.4}$$

Hierbij stelt h(t) de *hazard function*, die bepaald wordt door p predictors (x_1, x_2, ..., x_p). De coëfficiënten $b_{1->p}$ stellen de impact voor van de predictors op de *hazard* en de term h_0 stelt de baseline *hazard* voor. Dit komt overeen met de *hazard* wanneer alle x_i gelijk zijn aan 0 (variërend intercept over de tijd).

De interpretatie van dit model gaat als volgt: de waarden van e^{b_i} stellen de *hazard ratio's* voor. Een waarde van b_i groter dan 0, en dus een *hazard ratio* groter dan 1, geeft aan dat als de i-de predictor toeneemt, dat de *event hazard time* toeneemt, en dus de *survival lengte* afneemt (sneller een 2e boete).

Als we dit algemene model toepassen op de situatie voorhanden, dan bekomen we hier-

mee het volgende model:

$$\begin{aligned} h(t) = h_0(t) \times \exp(&\beta_{accident} \times accident\_dummy + \\ &\beta_{age} \times age + \\ &\beta_{gender} \times gender\_dummy + \\ &\beta_{accident} \times accident\_dummy + \\ &\beta_{vanity} \times ratio\_vanity + \\ &\epsilon_{1->10} \times ethnicity_{1->10}) + \\ &\beta_{fined} \times prior\_dummy + \\ &\delta_{1->10} \times ethnicity_{1->10}\_fined \end{aligned} \tag{4.5}$$

Hierbij zijn er 11 nieuwe variabelen, naast de gebruikelijke variabelen:

- *prior_dummy*: dit is een variabele die aangeeft voor een bepaald event, of de bestuurder in kwestie al eerder een boete heeft ontvangen.
- $ethnicity_{1->10}\_fined$: dit zijn interactievariabelen tussen bovenstaande dummy en de etniciteiten.

In dit model stellen de variabelen $ethnicity_{1->10}$ de algemene verschillen per etniciteit voor, los van de invloed die "beboet geweest zijn" daarop heeft. De interactievariabele $ethnicity_{1->10}\_fined$ geeft aan hoe anders verschillende etniciteiten reageren op "beboet geweest zijn".

Een belangrijke kanttekening die geplaatst moet worden bij dit model is dat de zogenaamde *proportional hazards condition*. Deze voorwaarde stelt dat predictors multiplicatief verbonden zijn aan de *hazard*. Dit houdt in dat de *hazard curves*, de curves die aangeven wat de *hazard* is die verbonden is met een bepaalde patiënt (bestuurder in ons geval), altijd evenredig moeten zijn en nooit mogen kruisen. Dit impliceert dat de *hazard ratio* van 2 bestuurders onafhankelijk moet zijn van het tijdstip t. De *hazard* van een gebeurtenis (2e boete voor een bepaald individu) is een constant veelvoud van de *hazard* van een ander individu.

Deze veronderstelling moet dus gecontroleerd worden voor het model.

4.3.3 Resultaten

De resultaten van het model zijn te vinden in 4.7

Uit de resultaten kunnen een aantal algemene vaststellingen afgeleid worden, volgens de interpretatie in 4.3.2:

- De leeftijd van de chauffeur, het gender, de *accident_dummy* en de `vanity_dummy` hebben een negatieve hazard ratio. Dit betekent dat als de bestuurder ouder is, de bestuurder een vrouw is of de auto een *vanity plate heeft*, de *event hazard time* afneemt, en dus de *survival lengte* toeneemt. Dit betekent dat een oudere persoon, een vrouwelijke persoon of een chauffeur met een *vanity plate* - ceteris paribus - minder snel een 2e boete zal krijgen.

In verband met de etniciteiten zien we dat de volgende etniciteiten een significante waarden hebben:

- Armeens: 0.03
- Chinees: 0.29
- Joods: 0.06
- Koreaan: 0.12

Alle andere etniciteiten hebben een niet-significante coëfficiënt. Dit betekent dat bij de hierboven opgelijste groepen hun etniciteit een positieve *hazard ratio* hebben en dus gepaard gaan met sneller boetes krijgen. Wanneer we kijken naar de prior, dan zien we dat chauffeurs die eerder een boete gehad hebben, veel sneller tegen een 2e boete aanlopen. Dit toont aan dat er inderdaad een belangrijke gedragsfactor aanwezig is bij het verklaren van hoe snel een persoon een boete krijgt. Er kan echter ook gezegd worden dat mensen die vaker op de baan zijn meer kans hebben om reeds een boete gehad te hebben, en dat meer op de baan zijn een oorzaak is om sneller een nieuwe boete te krijgen.

Ten slotte zijn de interactievariabelen van grote interesse. Deze interactievariabelen modelleren de gedragswijziging die het krijgen van een boete met zich mee zou moeten brengen. We zien bij de interactievariabelen dat er een aantal (Chinezen, Koreanen en Joden) significant negatief zijn. Dit betekent dat voor deze etniciteiten het krijgen van een boete ervoor zorgt dat ze in de toekomst minder snel een boete krijgen. Dit effect is het sterkst bij de Chinezen, bij wie er echter ook vastgesteld wordt dat zij sneller een boete krijgen. Chinese bestuurder krijgen dus sneller een boete, maar ingrijpen (in de vorm van een boete) door een agent, zorgt wel voor het gewenste effect. Enkel bij Tsjetsjenen is de

coëfficiënt noemenswaardig positief, wat dus betekent dat het krijgen van een boete ervoor zorgt dat ze sneller een nieuwe boete zullen krijgen. Dit verband is echter niet significant. Er kan dus een onderscheid gemaakt worden tussen die etniciteiten op wie een boete wel een effect heeft, en die etniciteiten waarbij het effect van een boete variabel is.

Zoals eerder vermeld gaat dit model uit van de assumptie van *proportional hazard*. Deze veronderstelling moet achteraf wel afgetoetst worden om de validiteit van de resultaten na te gaan. Dit doen we hieronder.

Het testen van de *proportional hazard assumption* wordt gedaan door de `check_assumptions`-methode van `lifelines`. Deze methode berekent de statistieken om de *proportional hazard assumption* na te gaan, plots te maken om die visueel te checken etc. Daarnaast genereert de methode ook rechtstreeks advies om dit te verbeteren.

Na het uitvoeren van deze test, bekomen we dat de variabelen `chauffeur_age`, `gender`, `vanity_dummy` en `Chechen` (de dummy voor Tsjetsjeense etniciteit) de niet-proportionele test niet slagen. Een manier om dit op te lossen is door de variabelen te stratifiëren. Dit betekent dat we de dataset gaan onderverdelen in subgroepen en het model te runnen voor iedere subgroep, om er zo voor te zorgen dat er wel voldaan is aan de *proportional hazard assumption*. Er werd echter gekozen om dit niet te doen, en wel om de reden dat het justifiëren van de *proportional hazard assumption* niet volledig wetenschappelijk onderbouwd is:

- Als het doel *survival prediction* is, dan is het niet nodig om ons zorgen te maken over *proportional hazards*, de manier van voorspellingen generen is dan niet belangrijk. Dit is duidelijk niet het geval hier, aangezien we willen de juistheid van het model bepalen.

- Als de sample size groot genoeg is, dan zullen zelfs kleine overtredingen van de veronderstellingen zich tonen. Dit is relevant voor onze dataset, aangezien we 432 490 observaties is, wat een groot aantal is.

- Er zijn onderbouwde redenen waarom iedere dataset de veronderstelling gaan overtreden. Dit is onderbouwd in het werk van Stensrud en Hern [47].

- Zelfs wanneer de *hazards* niet evenredig zijn, het veranderen van het model om te voldoen aan de veronderstelling, niet volledig wetenschappelijk is. Hiermee zouden

we bijvoorbeeld het effect van leeftijd, Tsjetsjeense afkomst, geslacht en vanity plates verliezen, wat niet de bedoeling is. Dit wordt geïllustreerd door de quote van Tukey:

> ”Better an approximate answer to the exact question, rather than an exact answer to the approximate question.”

[52]

Om die redenen is er gekozen om het model niet te wijzigen naar aanleiding van het overtreden van de *proportional hazard assumption*. We accepteren het feit dat dit de bevindingen van dit experiment verzwakt.

Dep. Variable:	fine_amount	**R-squared:**	0.001
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	150.0
Date:	Mon, 31 May 2021	**Prob (F-statistic):**	0.00
Time:	23:56:22	**Log-Likelihood:**	-1.2578e+07
No. Observations:	1742694	**AIC:**	2.516e+07
Df Residuals:	1742679	**BIC:**	2.516e+07
Df Model:	14		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	230.7326	0.849	271.843	0.000	229.069	232.396
accident_dummy	12.5122	1.177	10.627	0.000	10.204	14.820
chauffeur_age	-0.9996	0.023	-43.181	0.000	-1.045	-0.954
gender	-10.7341	1.224	-8.771	0.000	-13.133	-8.335
vanity_dummy	6.1419	1.439	4.269	0.000	3.322	8.961
Armenian	-3.8933	1.000	-3.893	0.000	-5.853	-1.933
Bashkir	-8.9923	7.047	-1.276	0.202	-22.804	4.819
Chechen	-15.0603	3.188	-4.724	0.000	-21.309	-8.811
Chinese	-19.1966	11.937	-1.608	0.108	-42.594	4.200
Jewish	-1.6018	2.209	-0.725	0.468	-5.931	2.727
Kazakh	-6.6619	2.533	-2.630	0.009	-11.626	-1.698
Korean	5.2166	3.968	1.315	0.189	-2.560	12.994
Polish	-0.7672	3.200	-0.240	0.810	-7.038	5.504
Tatar	-3.3654	1.493	-2.254	0.024	-6.292	-0.439
Ukrainian	3.4169	3.134	1.090	0.276	-2.727	9.560

Omnibus:	1226207.893	**Durbin-Watson:**	1.830
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15180420.462
Skew:	3.401	**Prob(JB):**	0.00
Kurtosis:	15.759	**Cond. No.**	1.72e+03

Table 4.6: OLS Regression Results

covariate	coef	$\overline{c}$	se(c)	c < 95%	c > 95%	$\overline{c}$ < 95%	$\overline{c}$ > 95%	z	p	-log2(p)
age	-0.03	0.97	0.00	-0.03	-0.03	0.97	0.97	-326.18	0.00	inf
gender	-0.24	0.79	0.00	-0.25	-0.23	0.78	0.79	-53.73	0.00	inf
acc_dummy	-0.05	0.95	0.00	-0.06	-0.04	0.94	0.96	-11.54	0.00	99.94
vanity_dummy	-0.06	0.94	0.01	-0.07	-0.05	0.93	0.95	-11.16	0.00	93.72
Armenian	0.03	1.03	0.01	0.01	0.05	1.01	1.05	2.82	0.00	7.70
Bashkir	0.14	1.15	0.08	-0.02	0.30	0.98	1.35	1.77	0.08	3.69
Chechen	0.01	1.01	0.04	-0.06	0.08	0.94	1.08	0.27	0.79	0.35
Chinese	0.29	1.34	0.09	0.13	0.46	1.13	1.59	3.44	0.00	10.74
Jewish	0.06	1.06	0.02	0.02	0.10	1.02	1.11	3.18	0.00	9.42
Kazakh	0.02	1.02	0.03	-0.04	0.07	0.97	1.07	0.63	0.53	0.92
Korean	0.12	1.13	0.03	0.06	0.18	1.06	1.19	3.93	0.00	13.49
Polish	0.01	1.01	0.03	-0.05	0.07	0.95	1.07	0.23	0.82	0.29
Tatar	0.02	1.02	0.01	-0.01	0.05	0.99	1.05	1.29	0.20	2.33
Ukrainian	0.07	1.07	0.03	0.01	0.13	1.01	1.13	2.47	0.01	6.19
prior_dummy	0.44	1.55	0.00	0.44	0.45	1.55	1.56	165.92	0.00	inf
arm_fined	0.01	1.01	0.01	-0.01	0.03	0.99	1.03	0.59	0.55	0.85
bash_fined	-0.13	0.87	0.09	-0.31	0.04	0.74	1.04	-1.54	0.12	3.01
che_fined	0.06	1.06	0.04	-0.02	0.13	0.98	1.14	1.45	0.15	2.77
chi_fined	-0.39	0.68	0.10	-0.59	-0.20	0.55	0.82	-3.91	0.00	13.41
jew_fined	-0.06	0.94	0.02	-0.11	-0.02	0.90	0.98	-2.95	0.00	8.28
kaz_fined	0.00	1.00	0.03	-0.05	0.06	0.95	1.06	0.06	0.96	0.07
kor_fined	-0.15	0.86	0.03	-0.22	-0.09	0.80	0.92	-4.44	0.00	16.76
pol_fined	0.01	1.01	0.03	-0.05	0.08	0.95	1.08	0.41	0.68	0.55
tat_fined	0.01	1.01	0.02	-0.02	0.04	0.98	1.04	0.63	0.53	0.92
ukr_fined	-0.03	0.97	0.03	-0.10	0.03	0.91	1.03	-1.09	0.28	1.85

Table 4.7: Resultaten van het Cox Proportional Hazard Model

Chapter 5

Conclusie

5.1 Bevindingen

Tot slot worden de bevindingen van de verschillende experimenten besproken. In het algemeen zijn de resultaten consistent met wat men zou verwachten, maar helemaal niet éénduidig genoeg om sluitende conclusies te kunnen trekken over het al dan niet voorkomen van etnische profilering in het Russische verkeer in en rond Moskou.

In het eerste experiment werd de invloed van etniciteit onderzocht op het gemiddelde aantal boetes die een persoon ontving. Hier kwam naar voor dat bepaalde etniciteiten, zoals Tjetsjenen en Chinezen, er gemiddeld meer boetes worden vastgesteld per persoon dan voor andere etniciteiten, en dit gecontroleerd voor aantal ongevallen waarbij betrokken, leeftijd, geslacht, kostprijs van de wagen, vanity plates, en zo verder. Ook na controle voor deze *confounding variables* bleef het verband bestaan, hetgeen aangeeft dat etniciteit toch een rol speelt in het aantal boetes.

Over het exacte mechanisme waarop etniciteit dergelijke variabelen beïnvloedt, kan er geen sluitende conclusie getrokken worden, al zal experiment 3 wel iets meer informatie verschaffen.

In het tweede experiment werd er onderzocht of etniciteit een rol speelt in het bepalen van de kans op een boete, wanneer er een overtreding wordt vastgesteld. De resultaten van dit experiment waren op zijn minst gezegd merkwaardig te noemen. Zo kwam er naar boven dat overtredingen die gepaard gaan met een ongeval minder snel beboet worden, maar wel zwaarder. Zo werd duidelijk dat zowel oudere mensen als vrouwen minder kans hadden op een boete bij een ongeval dan jongeren en mannen. Dit verband is consistent

met wat we vinden in de literatuur, al kan er opnieuw geen conclusie getrokken worden over het exact causaal verband.

In dit experiment is de invloed van etniciteit een stuk minder duidelijk. Vanwege de grote dataset en een gebrek aan technieken om hiermee om te gaan, kon er niet op een effectieve manier nagegaan worden of er een significante rol was voor etniciteit in het verklaren van de kans op een boete.

Ten slotte was er het laatste experiment, waar de technieken van survival analyse (oorspronkelijk uit de medische sector) toegepast werden op verkeersboetes. Zo werd er onderzocht wat de invloed is van verschillende variabelen op de frequentie waaraan verschillende bestuurders een boete krijgen. De conclusie hierbij was dat er een heleboel factoren waren die ervoor zorgen dat bestuurders minder snel een boete gaan krijgen, zoals leeftijd die toeneemt, geslacht, het meemaken van een ongeval, het bezit van een vanity plate, etc. Al deze verbanden komen overeen met wat er gevonden wordt in de literatuur, waar bijvoorbeeld oudere mensen en vrouwen minder risicovol gedrag vertonen op de baan dan jongeren en mannen. Intuïtief kan ook beredeneerd worden dat een ongeval een zeker afschrikeffect kan hebben op een bestuurder voor het breken van de verkeersregels.

In termen van etniciteit konden hier 2 grote categorieën vastgesteld worden: enerzijds zijn er die etniciteiten waar het krijgen van een boete een significant positief effect heeft op de survival time (ofwel de tijd tussen 2 boetes dus), en anderzijds die etniciteiten waar het krijgen van een boete geen duidelijk corrigerend effect heeft op de frequentie waarmee een bestuurder boetes krijgt. Uit dit experiment kan dus ook afgeleid worden dat etniciteit en cultuur een zekere rol spelen in rijgedrag.

5.2 Beperkingen en Future Work

Doorheen dit onderzoek zijn een aantal beperkingen aangehaald, we sommen de belangrijkste daarvan hier nog een keer op:

- **Corruptie** Er is geen correctie uitgevoerd voor mogelijke corruptie. In een land als Rusland, met een zeer hoge graad van (geobserveerde) corruptie, is dat iets waar in verder onderzoek misschien wel rekening mee moet gehouden worden.

- **Imperfecte etniciteitsvoorspelling** Doordat etniciteit niet aanwezig is als expliciete variabele in de dataset, moet er gebruik gemaakt worden van voorspellingen van

etniciteit. Dit introduceert natuurlijk mogelijks fouten. Doordat de etniciteitvoorspelling getraind is op data waarvan de verdeling niet heel goed overeenkomt met de verdeling van onze data, kan men de betrouwbaarheid van het etniciteitsmodel in vraag stellen. Een training van het model op hedendaagse, representatieve data of een finetuning op deze data kan de voorspellingen ten goede komen.

- **Grote dataset** Er is in dit onderzoek nog ruimte om technieken toe te passen om betere statistische analyse te doen op deze grote dataset (zie bijvoorbeeld experiment 2). Zoals aangehaald kampen we bij dat experiment met het p-value problem als gevolg van de grote dataset en is dat iets waar in verder onderzoek meer aandacht aan besteed kan worden.

- **Ontbrekende data** In de gebruikte dataset is er in vele gevallen ontbrekende data. Zo zijn er bijvoorbeeld veel boetes geregistreerd zonder een `fine_amount`, rijbewijzen ingetrokken zonder te weten hoeveel maanden, etc. Ook is de data zeer sparse. Er is ontbrekende data over de kostprijs van de wagen, er is ontbrekende data over de locatie van registraties, enz. Indien er meer en volledigere data beschikbaar is in de toekomst, kan dit onderzoek verder verfijnd worden.

Bibliography

[1] Moscow population 2021.

[2] Richard Arnold. Systematic racist violence in russia between 'hate crime' and 'ethnic conflict'. *Theoretical Criminology*, 19(2):239–256, 2015.

[3] Donna Bahry, Mikhail Kosolapov, Polina Kozyreva, and Rick K Wilson. Ethnicity and trust: Evidence from russia. *American Political Science Review*, pages 521–532, 2005.

[4] Gary S Becker. Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer, 1968.

[5] David Beetham. *The legitimation of power*. Macmillan International Higher Education, 2013.

[6] Alexey Bessudnov and Andrey Shcherbak. Ethnic discrimination in multi-ethnic societies: Evidence from russia. *European Sociological Review*, 36(1):104–120, 2020.

[7] Serguey Braguinsky, Sergey Mityakov, and Andrey Liscovich. Direct estimation of hidden earnings: Evidence from russian administrative data. *The Journal of Law and Economics*, 57(2):281–319, 2014.

[8] Rudolf Brázdil, Kateřina Chromá, Lukáš Dolák, Jan Řehoř, Ladislava Řezníčková, Pavel Zahradníček, and Petr Dobrovolnỳ. Fatalities associated with the weather in the czech republic, 2000–2019. *Natural Hazards and Earth System Sciences Discussions*, pages 1–47, 2021.

[9] Julie DaVanzo, Julie S DaVanzo, Clifford Anthony Grammich, and Davanzo Grammich. *Dire demographics: Population trends in the Russian Federation*. Number 1273. RAND corporation, 2001.

[10] Feliciaan De Palmenaer, Tom Eeckhout, and Koen Schoors. ETHNIC CLUSTERING IN FORMER SOVIET RUSSIAN PERSONAL NAMING NETWORKS.

[11] Kathleen M Dowley and Brian D Silver. Social capital, ethnicity and support for democracy in the post-communist states. *Europe-Asia Studies*, 54(4):505–527, 2002.

[12] Kristin Nicole Dukes and Kimberly Barsamian Kahn. What social science research says about police violence against racial and ethnic minorities: Understanding the antecedents and consequences—an introduction. *Journal of Social Issues*, 73(4):690–700, 2017.

[13] Rune Elvik and Peter Christensen. The deterrent effect of increasing fixed penalties for traffic offences: the norwegian experience. *Journal of Safety Research*, 38(6):689–695, 2007.

[14] CIA Factbook. The world factbook. *See also: https://www. cia. gov/library/publications/the-world-factbook*, 2010.

[15] Roni Factor. The effect of traffic tickets on road traffic crashes. *Accident Analysis & Prevention*, 64:86–91, 2014.

[16] Roni Factor, Joseph N Prashker, and David Mahalel. The flashing green light paradox. *Transportation research part F: traffic psychology and behaviour*, 15(3):279–288, 2012.

[17] Larry K Gaines. An analysis of traffic stop data in riverside, california. *Police Quarterly*, 9(2):210–233, 2006.

[18] James Garofalo. *Public opinion about crime: The attitudes of victims and nonvictims in selected cities*, volume 1. US Department of Justice, Law Enforcement Assistance Administration . . . , 1977.

[19] Michael A Gebers and Raymond C Peck. Using traffic conviction correlates to identify high accident-risk drivers. *Accident Analysis & Prevention*, 35(6):903–912, 2003.

[20] James L Gibson. Social networks, civil society, and the prospects for consolidating russia's democratic transition. *American Journal of Political Science*, pages 51–68, 2001.

[21] Jack Glaser. *Suspect race: Causes and consequences of racial profiling*. Oxford University Press, USA, 2015.

[22] C Reurings Goldenbeld, MCB Norden, and HL Y van & Stipdonk. Relatie tussen verkeersovertredingen en verkeersongevallen: verkennend onderzoek op basis van cjib-gegevens. 2011.

[23] Michael J Hindelang. Public opinion regarding crime, criminal justice, and related topics. *Journal of Research in Crime and Delinquency*, 11(2):101–116, 1974.

[24] Wilson Huang and Michael S Vaughn. Support and confidence: Public attitudes toward the police. *Americans view crime and justice: A national public opinion survey*, pages 31–45, 1996.

[25] Transparence International. Corruption perceptions index 2020 for russia.

[26] Ronald C Kessler, Kristin D Mickelson, and David R Williams. The prevalence, distribution, and mental health correlates of perceived discrimination in the united states. *Journal of health and social behavior*, pages 208–230, 1999.

[27] Nancy Krieger and Stephen Sidney. Racial discrimination and blood pressure: the cardia study of young black and white adults. *American journal of public health*, 86(10):1370–1378, 1996.

[28] Hope Landrine and Elizabeth A Klonoff. The schedule of racist events: A measure of racial discrimination and a study of its negative physical and mental health consequences. *Journal of Black Psychology*, 22(2):144–168, 1996.

[29] Mingfeng Lin, Henry C Lucas Jr, and Galit Shmueli. Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, 2013.

[30] Peter F Lourens, Jan AMM Vissers, and Maaike Jessurun. Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. *Accident Analysis & Prevention*, 31(5):593–597, 1999.

[31] Masha Maltz and David Shinar. Imperfect in-vehicle collision avoidance warning systems can aid distracted drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(4):345–357, 2007.

[32] Albert J. Meehan and Michael C. Ponder. Race and place: The ecology of racial profiling african american motorists. *Justice Quarterly*, 19(3):399–430, 2002.

[33] G William Mercer. Traffic accidents and convictions: group totals versus rate per kilometer driven. *Risk Analysis*, 9(1):71–77, 1989.

[34] Emil Payin, K Rupesinghe, and V Tishkov. Settlement of ethnic conflicts in post-soviet society. *Valery Tishkov and Kumar Rupesinghe, Ethnicity and Power in the Contemporary World*, pages 69–82, 1996.

[35] Raymond C Peck. The identification of multiple accident correlates in high risk drivers with specific emphasis on the role of age, experience and prior traffic violation frequency. *Alcohol, Drugs & Driving*, 1993.

[36] David R. Williams and Ruth Williams-Morris. Racism and mental health: The african american experience. *Ethnicity & health*, 5(3-4):243–268, 2000.

[37] Sirpa Rajalin. The connection between risky driving and involvement in fatal accidents. *Accident Analysis & Prevention*, 26(5):555–562, 1994.

[38] Deborah Ramirez, Jack McDevitt, and Amy Farrell. *A resource guide on racial profiling data collection systems: Promising practices and lessons learned*. US Department of Justice, 2000.

[39] VL Rimskyi. Rezultaty sotsiologicheskih issledovanii ispolneniya gosudratvom pravookhranitelnoi funktsii (results of sociological surveys of law enforcement functions of the state). *ReSoIsIspGoPF. pdf*, 2012.

[40] Jonathan J Rolison, Yaniv Hanoch, Stacey Wood, and Pi-Ju Liu. Risk-taking differences across the adult life span: a question of age and domain. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 69(6):870–880, 2014.

[41] Richard Rose. Russia as an hour-glass society: A constitution without citizens. *E. Eur. Const. Rev.*, 4:34, 1995.

[42] Katheryn K Russell. Driving while black: Corollary phenomena and collateral consequences. *BCL Rev.*, 40:717, 1998.

[43] Peter Rutland. The presence of absence: Ethnicity policy in russia. In *Institutions, ideas and leadership in Russian politics*, pages 116–136. Springer, 2010.

[44] Patrick Sewell and special to RBTH. Why russia fails in ethnic conflict resolution, Aug 2016.

[45] Jim Sidanius and Felicia Pratto. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press, 2001.

[46] Frances Simon and Claire Corbett. Road traffic offending, stress, age, and accident history among male and female drivers. *Ergonomics*, 39(5):757–780, 1996.

[47] Mats J Stensrud and Miguel A Hernán. Why test for proportional hazards? *Jama*, 323(14):1401–1402, 2020.

[48] Justice Tankebe, Kofi E Boakye, and Moses Agaawena Amagnya. Traffic violations and cooperative intentions among drivers: the role of corruption and fairness. *Policing and society*, 30(9):1081–1096, 2020.

[49] Vetta L Sanders Thompson. Perceived experiences of racism as stressful life events. *Community mental health journal*, 32(3):223–233, 1996.

[50] Valerii Aleksandrovich Tishkov. *Nationalities and conflicting ethnicity in post-communist Russia*, volume 50. United Nations Research Institute for Social Development, 1994.

[51] Donald Tomaskovic-Devey, Marcinda Mason, and Matthew Zingraff. Looking for the driving while black phenomena: Conceptualizing racial bias processes and their associated distributions. *Police Quarterly*, 7(1):3–29, 2004.

[52] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[53] Tom R Tyler. Policing in black and white: Ethnic group differences in trust and confidence in the police. *Police quarterly*, 8(3):322–342, 2005.

[54] Shawn O Utsey, Joseph G Ponterotto, Amy L Reynolds, and Anthony A Cancelli. Racial discrimination, coping, life satisfaction, and self-esteem among african americans. *Journal of Counseling & Development*, 78(1):72–80, 2000.

[55] Anatoly Vishnevsky and Ekaterina Shcherbakova. A new stage of demographic change: A warning for economists. *Russian Journal of Economics*, 4:229, 2018.

[56] David R Williams, Harold W Neighbors, and James S Jackson. Racial/ethnic discrimination and health: Findings from community studies. *American journal of public health*, 93(2):200–208, 2003.