A.I. & Data Science Challenge 1:

Red life: Predicting Casualty Severity in Traffic Accidents

University of Applied Sciences

Mateusz Mierzejek

2775409

12/20/2021

Introduction

In the world of artificial intelligence there are endless possibilities. Revealing untold secrets, or even foreseeing the future. But what sometimes gets forgotten is working towards a better and brighter future. Continuous technological improvement is a daily standard nowadays. But is it safe?

One of the ongoing highlights of society is automation. In every way, shape, and form. Ranging from fridges that update your shopping list, to self-driving cars. Simple life tasks get replaced with the press of a button. But does this benefit society?

This can be seen as entering a different state of mind. Instead of working towards your objective, you rely on assistance. And getting this assistance, requires your attention. Our focus gets devoured each day. Messages, notifications, adverts, and an overall overflow of information. From day-to-day life it is already noticeable how big of an effect these systems have on our lives. And especially on others.

One area that experiences a magnificent impact from technology that is a grasp away. Namely, traffic. Vehicles and road networks have evolved massively over the past few decades. But so have the incidents. In the worst cases resulting in death.

Would it be possible to predict a scenario in which a death is bound to occur during a traffic incident? And how can we prevent it from happening?

Contents

1. Domain Understanding

To envision a goal, a deeper understanding of the subject in necessary. While keeping focus on traffic incidents and related deadly occurrences, what other outside factors might play a role in the situations?

1. Context understanding

There can be many causes of traffic incidents. Not paying attention for a few seconds and hitting someone or always being careful and getting rear-ended is only an example of the dangers that we encounter daily. This happens all over the world under different circumstances. Let's take a deeper look into some of the topics of interest:

1. Frequency

In 2018 WHO[1] released their *Global status report on road safety 2018*[2]. With a reported 1.35million deaths it is a staggering number. It is now the leading death cause of people aged 5-29. The rate of death is disproportionally borne by pedestrians, cyclists and motorcyclists. With an even higher rate in developing countries.

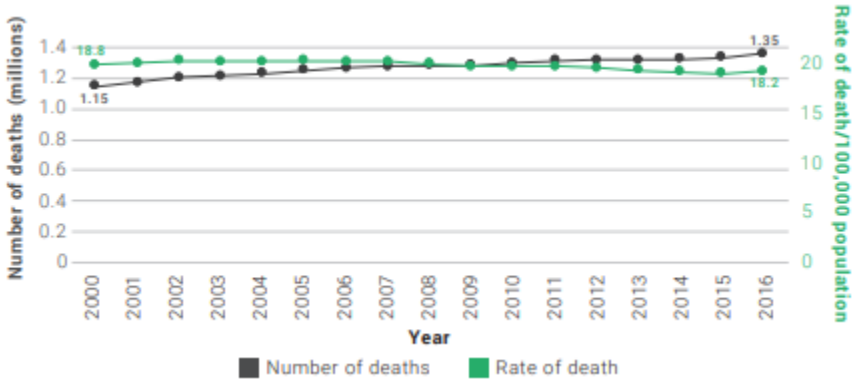

FIGURE 1 NUMBER AND RATE OF ROAD TRAFFIC DEATH PER 100K 2000-2016

Looking at the rate one might say luckily it has decreased. But in fact, the number of deaths has increased. This is because of the influx of population, and even more so in developing countries which have the highest rate of death.

In 2019 Romania was top scoring in European traffic related death rates. A staggering 96 killed per million inhabitants[3]. The average death rate in the EU for that year was 51.2 per million. Still there are quite some deviations, for instance Sweden has a rate of 21.6 per million. And there is also Bulgaria reaching 86.7 per million. The Netherlands scores in the middle with 33.9 per million.

What could be the cause of these fluctuations? As seen above WHO discovered that there is a correlation between developing countries and higher rates of traffic related deaths. Taking Sweden as a well-developed country, it is obvious the rates are low. And in the case of Romania (not so heavily developed) the rates are high.

[1] World Health Organization

[2] https://www.who.int/publications/i/item/9789241565684

[3] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road_accident_fatalities_-_statistics_by_type_of_vehicle#In_2019.2C_Romania_had_96_persons_killed_per_million_inhabitants_in_road_accidents.2C_the_highest_in_the_EU

2. Infrastructure

Another interesting topic to look at is the infrastructure. This is usually heavily correlated to the development of the country. For instance, Luxembourg and Iceland have been leaders in least number of cyclists killed in traffic incidents with an amazing number of 0[4].

In the case of Iceland cyclists are allowed on normal roads. But this does not happen frequently. The harsh weather conditions and rugged nature make it a though ride[5]. Many cyclists in Iceland do not do it for transport, but as a leisure. This could be one of the cases for such low numbers.

Luxembourg is part of the Benelux. Cycling is second nature here. In the last few years massive amount of care has been but into converting it from a car-centric country to a heaven for bikers[6]. By taking care and prioritizing cyclists in traffic they managed to reduce the numbers to zero! Their key concept *Multimodal* has been their cornerstone. Giving the possibility of flawless switching of transport between various locations.

While on the other hand you also have the Netherlands which has a strong cycling culture. Here cyclists account for 25.3% of all road accident deaths. Why is the number so high compared to Luxembourg? And is there a way to prevent it?

3. Vehicles

The infrastructure might play a big role in some cases. But often it comes down to the vehicle. Cases where some participants of traffic are more vulnerable than other occur often. In 2019 44.2% of people killed were passenger car occupants. This might come to no surprise as this is the main way of travel for most people. But next are the most vulnerable participants. Pedestrians, motorcycles and bicycles.

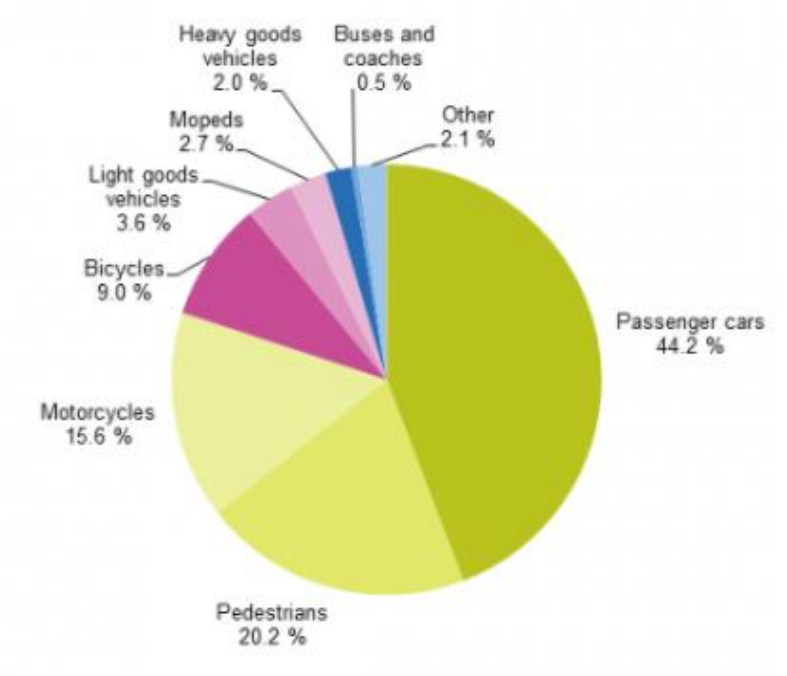

FIGURE 2 ROAD ACCIDENT FATALITIES BY CATEGORY, EU 2019[7]

[4] https://ec.europa.eu/eurostat/databrowser/view/TRAN_SF_ROADVE__custom_1647646/default/table?lang=en
[5] https://cyclingiceland.is/en/all-you-need-to-know-2/
[6] https://luxembourg.public.lu/en/vivre/mobility/by-bicycle.html
[7] https://ec.europa.eu/eurostat/databrowser/view/tran_sf_roadve/default/table?lang=en

Other relevant information about the drivers may also be of use in establish the cause of these cases. As well as other specifics about the vehicle might be of use like the length, height, weight, build year and condition etc.

4. Other factors

In the end a lot of it comes down to the circumstances during the incident. What was the speed like, what were the weather conditions and what was the current traffic situation like?

Many of these questions often go unanswered because of lack of witnesses or persons of interest fleeing the scene. A national AAA U.S. study shows that severity of hit-and-runs[8]:

- Over 5% of traffic fatalities from hit-and-runs
- Average increase of 7.2% per year
- Hit-and-runs account for 20% of pedestrian fatalities

In these cases, it is very hard to fully explain all circumstances that led to the occurrence. And without the help of external sources, it is hard to distinguish a solution.

[8] https://aaafoundation.org/wp-content/uploads/2018/04/18-0058_Hit-and-Run-Brief_FINALv2.pdf

2. Project goal

The scope of requirements for this project would be predicting situations in which traffic incidents are deadly. Where the key point of interest is to find solutions for a safer and more stable network of travel.

Suppose individual *A* drives a car and wants to know what his chances are of getting involved in a traffic accident which might even result in death?

This seems like an odd use case for a regular person. But think about insurance companies. This could be a very helpful tool for them to check the security of their clients. By filling out a set of features a probability rating can be predicted.

According to the statements made above, the main point of interest will be predicting death given a certain traffic related incident. A scope of possible interesting features sets in a dataset might be:

- Vehicle
- Person of interest
- Infrastructure

With these goals we can define a key point of interest. This will be the guideline when performing the exploratory data analysis and further analytic approach.

3. KPI

TARGET	MEASURE	GOAL	APPROACH
Derive the situation in which a traffic incident becomes deadly	Collection of target variables related to occurrences of deadly and non-deadly incidents	Based on the derived features a prediction can be made on the outcome of the incident	Combining datasets involving deadly and non-deadly situations and finding correlated features

4. Base Target Goals

In order to find usable datasets target variables, must be established. There are 3 forms we will take a further look at:

1. Categorical

This type of data will be used for clustering types together and finding new pieces of information. i.e., combining the type of vehicle to an area of interest. This type of data can later be used for classification.

2. Temporal

Features that regard a specific moment in time can be considered temporal. The date of the crash, and the time of death could be useful in predicting time of death after being involved in an incident.

3. Spatial

These features could be useful for creating locational maps with clustered points of interest based on the nature and outcome of the incident. These locations can be easily combined with external data sources for cross referencing.

2. Exploratory Data Analysis

During the EDA we will investigate several datasets covering extensive features for each case. It is important to check the data thoroughly for invalid, missing or even misleading datapoints. We will take a closer look at each dataset and display the data in a statistical and analytical way.

1. Deaths per year in EU countries by vehicle category (2011-2019)

With this spatial dataset we can observe the rate of vehicle deaths per country. This dataset contains many vehicle categories with the number of deaths each year.

TIME	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
GEO (Labels)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
European Union - 27 countries (from 2020)	54680	53410	51282	50343	47331	44466	42552	40365	40038	36880	32978	29576	28671	26457.0	24182.0	24131.0	24358.0	23812.0	23394.0
European Union - 28 countries (2013-2020)	58244	56990	54880	53924	50989	47834	45888	43663	43097	39525	35315	31481	30668	28231.0	25983.0	25987.0	26162.0	25672.0	25250.0
Belgium	1397	1470	1486	1306	1213	1162	1089	1089	1071	944	944	850	884	827.0	764.0	745.0	762.0	670.0	609.0
Bulgaria	1047	1012	1011	959	960	943	957	1043	1006	1061	901	776	656	601.0	601.0	661.0	708.0	708.0	682.0

FIGURE 3 DATASET 1, SHEET 1: TOTAL DEATHS ALL CATEGORIES EU

TIME	GEO (Labels)	European Union - 27 countries (from 2020)	European Union - 28 countries (2013-2020)	Belgium	Bulgaria	Czechia	Denmark	Germany (until 1990 former territory of the FRG)	Estonia	Ireland	Greece	Spain	France
1999	NaN	:	:	851	:	775	271	:	:	236	886	3186	5454
2000	NaN	:	:	922	:	784	235	4396	:	262	922	3285	5291
2001	NaN	:	:	899	:	715	242	4023	:	231	803	3140	5284
2002	NaN	:	:	779	:	759	246	4005	:	202	793	3102	4862
2003	NaN	:	:	688	:	798	236	3774	:	174	761	3212	3685
2004	NaN	:	:	623	:	779	186	3238	:	205	775	2691	3365
2005	NaN	:	:	624	:	679	169	2833	88	222	816	2390	3065
2006	NaN	:	:	589	:	567	138	2683	106	226	722	2095	2627
2007	NaN	:	:	550	:	661	168	2625	122	171	771	1817	2466
2008	NaN	:	:	479	623	573	196	2368	69	160	708	1493	2205

FIGURE 4 DATASET 1, SHEET 14: TRANSPOSED PASSENGER CAR DEATHS EU

From the second dataset we can see that there are quite some unknown values. But this can often be the case with datasets. Some countries that are for instance not in the EU did not always track the statistics. Comparing Belgium (EU member 1958) to Bulgaria (EU member 2007)[9].

The dataset has been transposed for easier use with plotting and the GEO labels can be cross referenced with another dataset containing the geolocations of each country.

[9] https://www.schengenvisainfo.com/countries-in-europe/eu-countries/

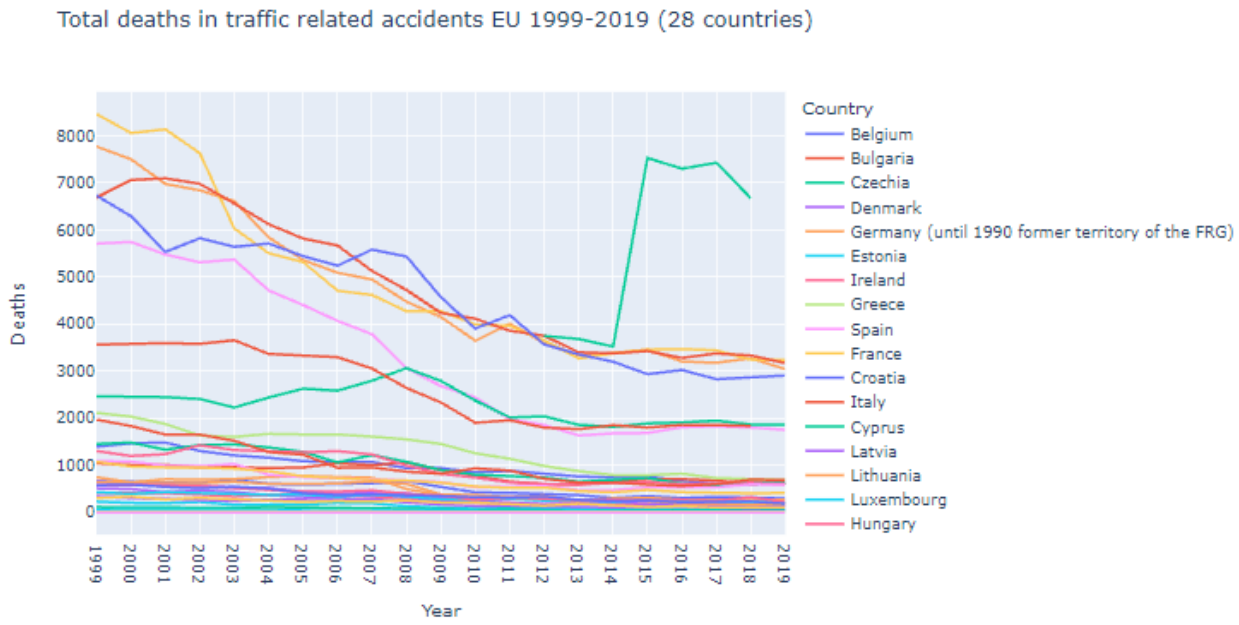

FIGURE 5 DATASET1, SHEET1: DEATHS PER COUNTRY ALL CATEGORIES EU

Here we can observe the total deaths for all vehicle categories in the EU. We observe that there is a downwards decline over the years. This means choosing a dataset that's relatively shorter ago will have less datapoints. But the data might be of higher quality. This data and all other categorical spatial data can be used alongside with time series forecasting in predicting future composition of accidents across the EU.

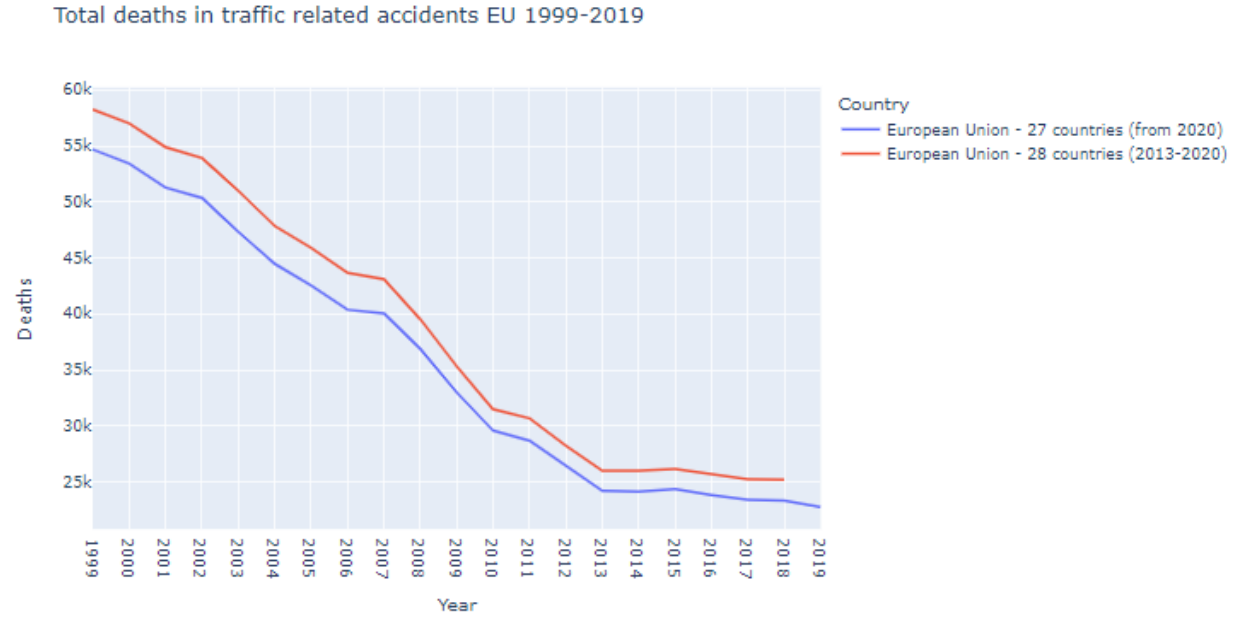

FIGURE 6 DATASET1, SHEET1: TOTAL DEATHS ALL CATEGORIES EU

Here the steady decline is heavily visible. The difference between the two is the exclusion of the United Kingdom because of Brexit. We observe that England by itself had quite a high rate of traffic incidents.

2. Deadly traffic incidents in the Netherlands 2006-2012

This dataset is mostly comprised of categorical data. There is a lot of opportunity here for data combination and extraction. A few points of interest:

- **Year**
  Influence of time on rate of death
- **Age**
  Distribution and concentration
- **Gender**
  Reoccurring incidents and frequency/ distribution
- **Vehicle**
  Type and frequency
- **Province**
  Distribution and concentration
- **Longitude** & **Latitude**
  GEO location for mapping

```
#   Column
---  ------
0   Jaar
1   LEEFTIJD
2   GESLACHT
3   vervoerwijze
4   vervoerwijzefilter
5   wegvak/kruispunt
6   wegbeheerder
7   wegnummer
8   HECTOMETER
9   STRAAT1
10  STRAAT2
11  STRAAT3
12  GEMEENTE
13  GEMEENTE 2013
14  PROVINCIE
15  type locatie
16  X_COORD
17  Y_COORD
18  latitude
19  longitude
```

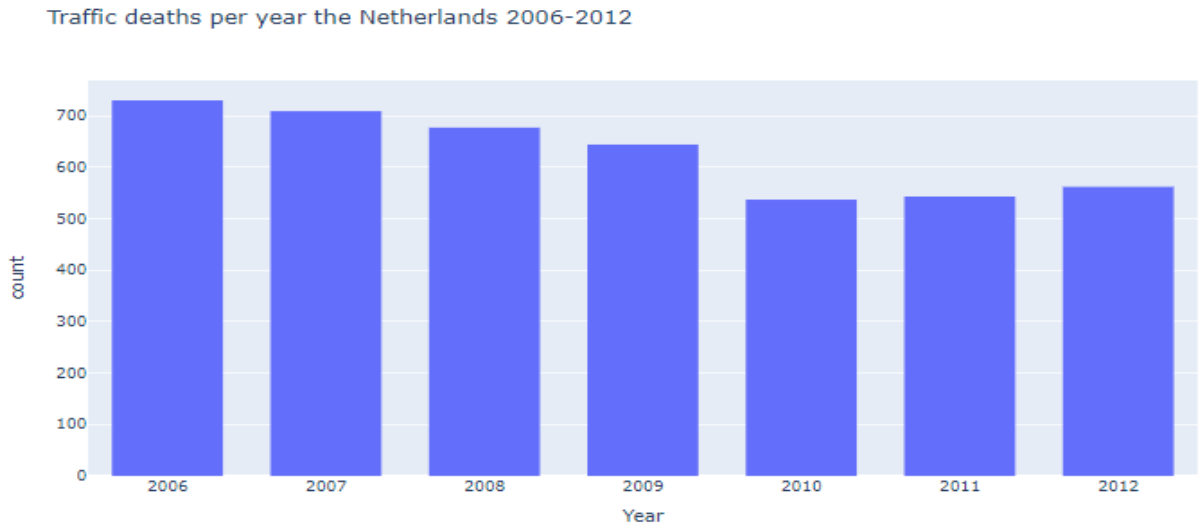

FIGURE 7 DATASET 2: TRAFFIC DEATHS PER YEAR

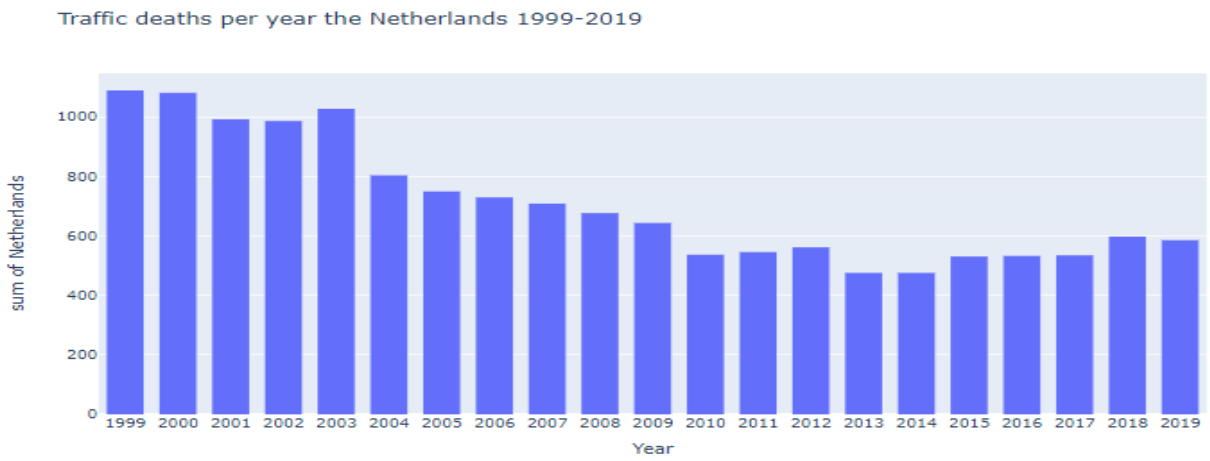

FIGURE 8 DATASET1, SHEET1: TOTAL TRAFFIC DEATHS NL

From the graphs above we can see that the two separate datasets are in line with each other. All values from *dataset2* between 2006-2012 correlate to the values from *dataset1*.

Taking a further look at age distribution we can see something interesting.

FIGURE 9 DATASET 2: DEATHS BY AGE

There is a great spike in with young individuals starting from the age of 18. Right when you get a driver's license. Let's compare vehicle type to age.

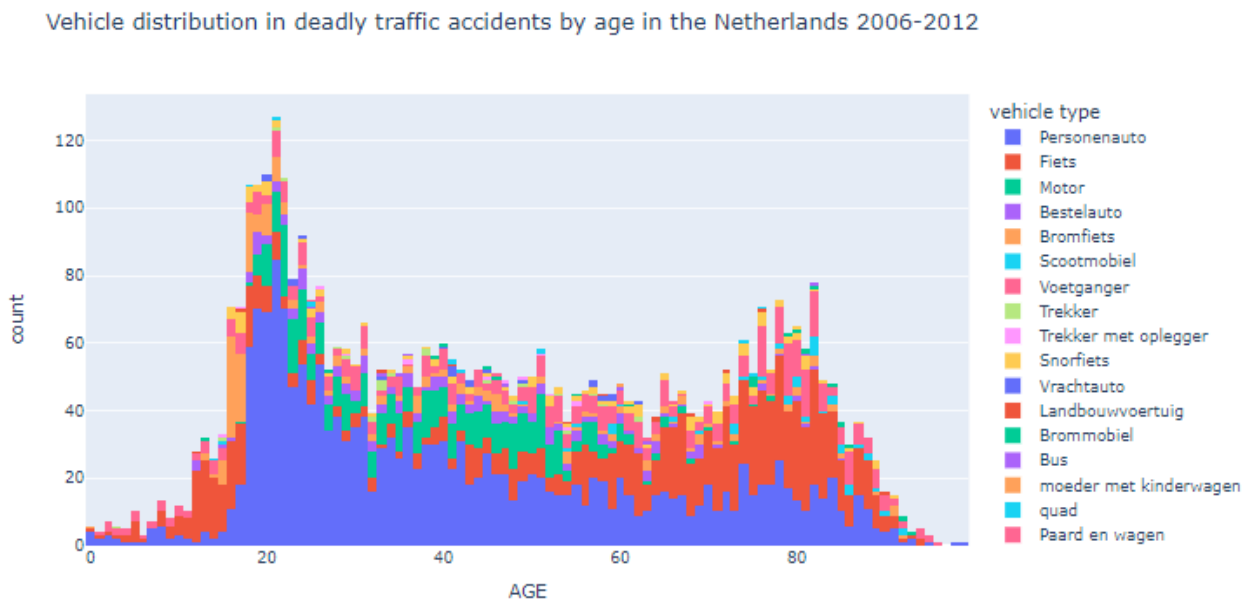

FIGURE 10 DATASET2: DEATH BY VEHICLE TYPE PER AGE GROUP

Here we can see a perfect example of a use case where the deaths per age group are in some way correlated to a type of vehicle which would most often be used in that age category. For instance:

- Below the age of **18** there is a significantly higher amount of **bicycle** related deaths, this can be to the fact that children use the bike from a young age for i.e., school.
- At age **16** there is an influx of deaths occurring from **scooter** accidents. This because teens in the Netherlands are allowed to obtain a scooter driver's license at the age of 16.
- At age **18** the rate of scooter deaths decreases, but **passenger car** death rates increase. This is to the fact teens can obtain their driver's license at 18.
- At age **18** there is also a spike in **motorcycle** deathrates. This is because the age limit of motorcycle licenses is 18+.
- After age **50** a steady increase in **bicycle** related deaths starts appearing. This can be to the fact that the elderly become more vulnerable in traffic.
- After age **40** the **passenger car** death rates decrease. This might be because individuals become more self-aware and protective of relatives.
- After age **50** more pedestrian death start occurring. This can also be to the fact that they are more vulnerable and less attentive.

Another insight is the occurrences per province:

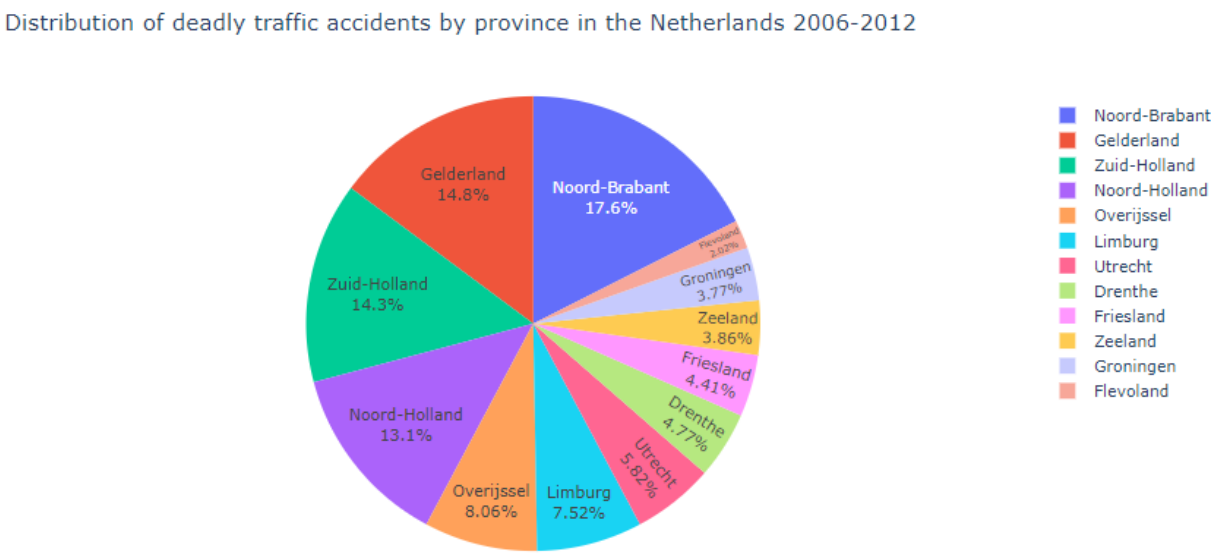

FIGURE 11 DATASET2: DISTRIBUTION OF DEADLY INCIDENTS

Here we can see that the leader is **Noord-Brabant** and **Gelderland** in second. One might think that **Zuid-Holland** or **Noord-Holland** would have most cases because of the density. But this is not the case. An interesting comparison would be population density and death rates per province. Here's a general idea:

We can see that Noord-Holland and Zuid-Holland have the highest population. But Noord-Brabant and Gelderland have relatively high population in addition to a massive ground area.[10]

This would be interesting to investigate further.

Groningen	586.813	2.324
Fryslân / Friesland	651.459	3.336
Drenthe	494.760	2.633
Overijssel	1.166.478	3.319
Flevoland	428.264	1.412
Gelderland	2.096.620	4.964
Utrecht	1.361.093	1.485
Noord-Holland	2.887.906	2.665
Zuid-Holland	3.726.173	2.700
Zeeland	385.379	1.782
Noord-Brabant	2.573.853	4.905
Limburg	1.115.895	2.147

FIGURE 12 POPULATION PER PROVINCE IN THE NETHERLANDS 2021

3. Traffic incidents the Netherlands (various timeframes)

This upcoming dataset is an even bigger collection. It as well contains categorical, temporal and spatial data.

```
RangeIndex: 122051 entries, 0 to 122050
Data columns (total 116 columns):
 #    Column                  Non-Null Count   Dtype
---   ------                  --------------   -----
 0    Longitude               122051 non-null  float64
 1    Latitude                122051 non-null  float64
 2    OngevalID               122051 non-null  int64
 3    Communicatie_Ref        122049 non-null  object
 4    ProcesverbaalOpgem      23414 non-null   object
 5    Afloop3                 122051 non-null  object
 6    AantalPartijen          122051 non-null  int64
 7    Aard                    122051 non-null  object
 8    GekoppledNiveau         122051 non-null  object
```

FIGURE 13 SOME FEATURES OF DATASET 3: TRAFFIC INCIDENTS IN THE NETHERLANDS 2006

[10] https://www.metatopos.eu/provincies_eu.php

Here we can see the total number of **122051** datapoints and **116** features. This is just a small sample of the available features.

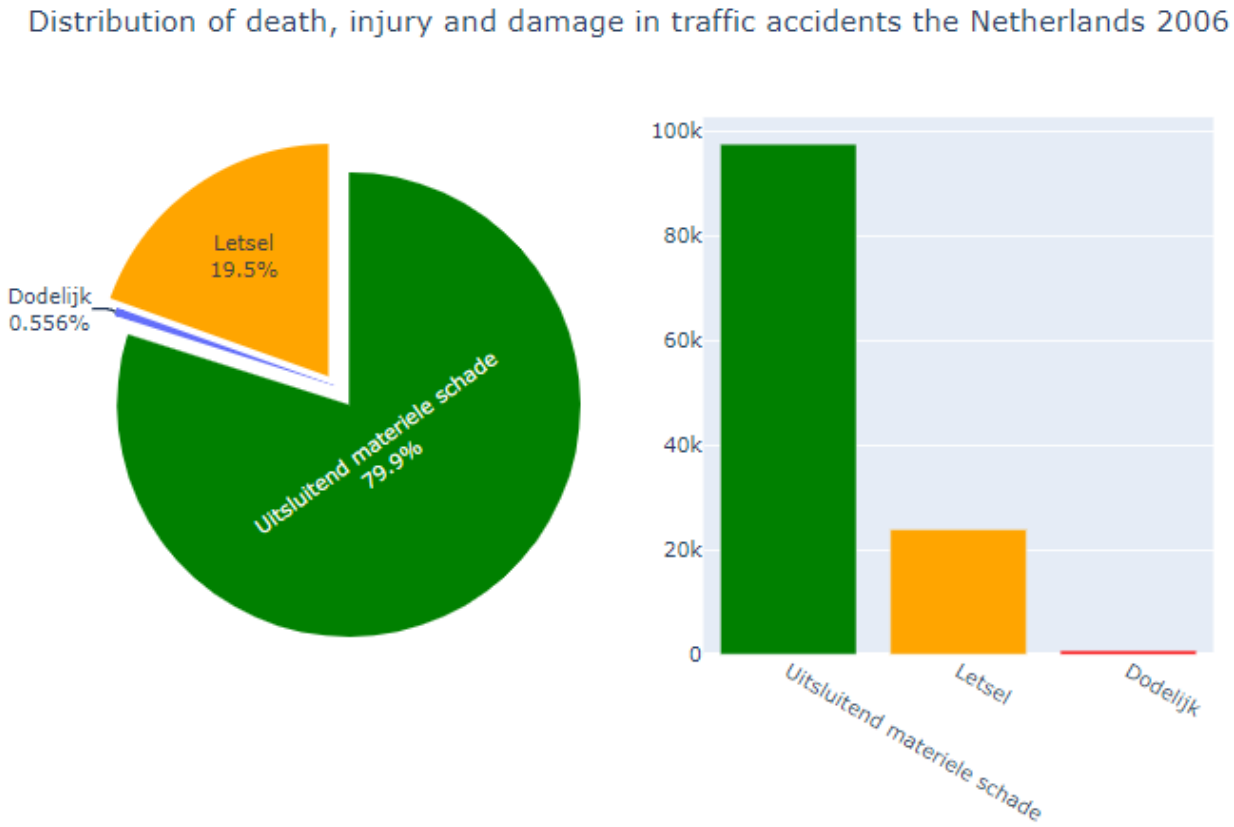

FIGURE 14 DATASET 3: DISTRIBUTION OF DEATH, INJURY AND DAMAGE

We can see that the death rates are relatively low at about 0.5%. On a total of 120k datapoints this still results in **678** traffic related deaths in 2006. This is lower compared to the **730** found in *Dataset1 and Dataset2*. The relatively low number of death cases also means this data needs to be satisfied with more datapoints. We can combine the datasets of multiple years for this.

In order to select usable features, we need to filter through the data. Some of the feature groups can be categorized as follows:

1. Location

This spatial category can give insights into the density and positioning of accidents in a map. General points of interest could be established where the rate of death is higher. Another interesting point is the infrastructure:

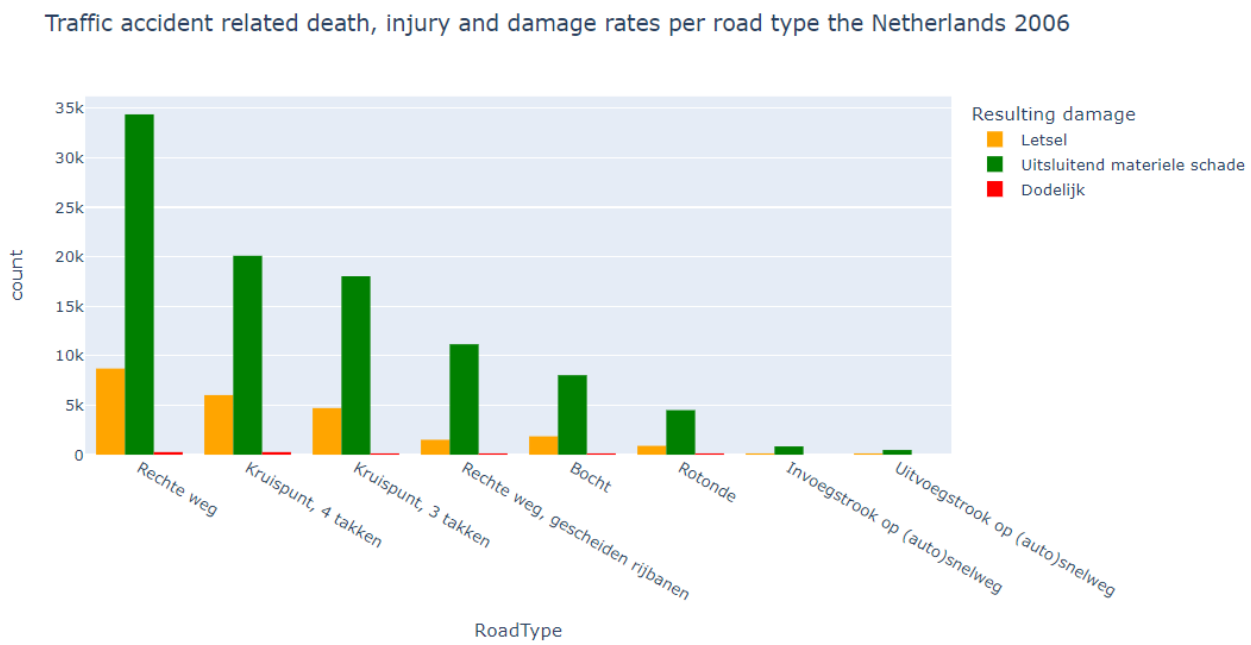

FIGURE 15 DATASET 3: INJURY, DAMAGE AND DEATH BY ROAD TYPE

In the graph above we can see that most deaths, just like accidents occur on the straight road. With the use of geo data, we can identify the saturation per area. Also, a link can be established between the categorical types of i.e., roads, vehicles, etc. A general overview of this data can be applied on a scale ranging from city street to province.

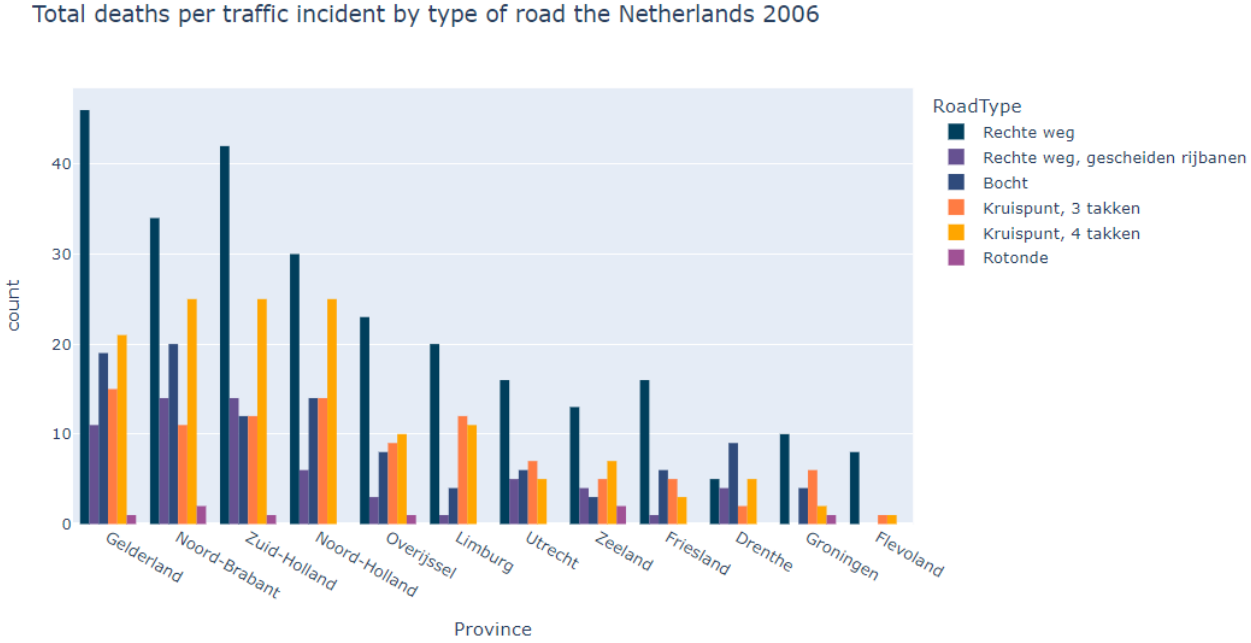

FIGURE 16 DATASET 3: TOTAL DEATHS PER PROVINCE BY ROAD TYPE

FIGURE 17 DATASET 3: DEATHS BY ROAD TYPE

We can see the concentration of traffic incidents in this map. As we saw before in *figure 11* Noord-Brabant has the highest percentual density. But we can see on the map that the highest concentration of points is found in Zuid-Holland.

Also as seen in *figure 16* Gelderland is on top with most total deaths. Yet on the map the concentration is not clearly distinguishable.

2. Vehicles / Objects

The objects in this dataset are represented by their involvement during the accident. The object can range anywhere from cars to light posts.

FIGURE 18 DATASET 3: TOTAL ACCIDENTS PER OBJECT TYPE

As we can see here the **passenger car** clearly takes the top spot. One might think that this would be a safe option of transport. But what is the actual distribution of death rates?

FIGURE 19 DATASET 3: DEATHS PER OBJECT TYPE

We can observe an interesting phenomenon. Even though the **bicycle** accidents occur less frequently, the rate of death in them is much higher.

3. Person(s) of interest

Does gender play a role in dangerous behavior? One might arrive at that conclusion relatively quickly when browsing YouTube. But does this translate over into traffic? The dataset contains some information about the parties involved in the accident.

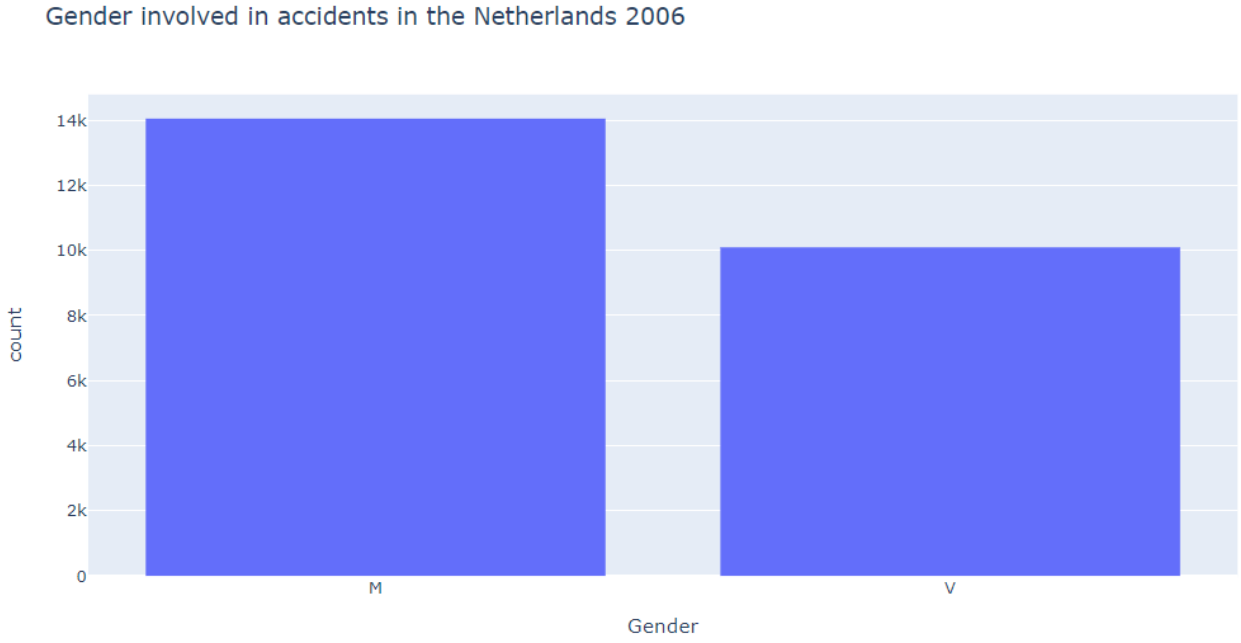

FIGURE 20 DATASET 3: INCIDENTS BY GENDER

There is a visible difference between males and females. With males getting involved 40% more frequently than females.

A combination that could give insight into the behavior of individuals is they drivers license.

FIGURE 21 DATASET 3: LICENSE TYPES PER GENDER

The dataset contains a lot of unknown entries. This could be because of the big number of hit-and-runs.

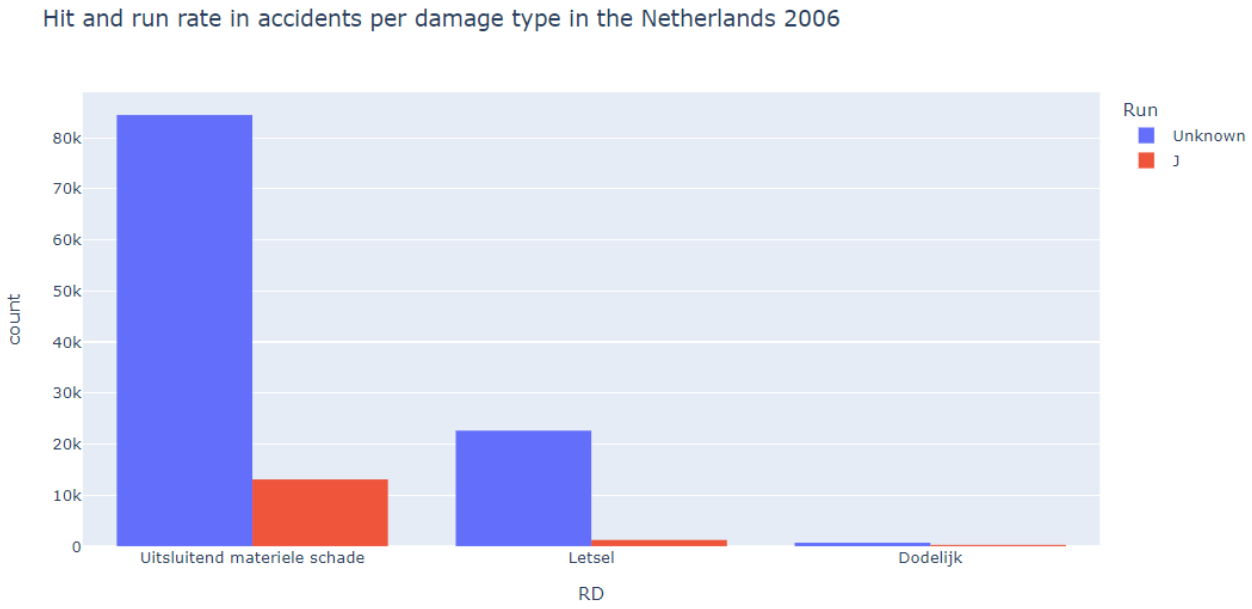

FIGURE 22 DATASET 3: HIT AND RUN RATE

As we can see only a small part of the accidents are hit-and-runs. Sadly, the available information per case might now be enough. A possible solution is excluding the known cases. And cross referencing these with *Dataset1*. The downside being a possible underfitted model.

4. Situational

Besides the actions of involved parties. Infrastructure and traffic code can result in some unpredictable behavior. Think about an intersection with road works. A few points of interest are:

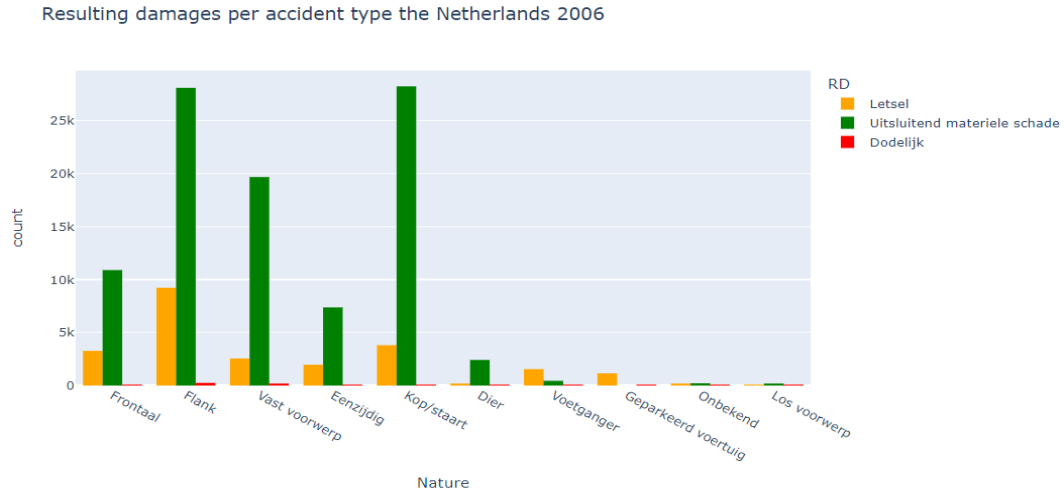

FIGURE 23 DATASET 3: DAMAGES PER ACCIDENT TYPE

Here we can observe that most deaths occur with accidents from the **Flank (Side)** and **Vast voorwerp (solid object)**. Comparing it to **Kop/staart (Frontal/rear),** the total count of cases for **solid objects** is relatively low, but the death rate is much greater. This can be to the fact that the incidents with **solid objects** and collisions from the **side** can be very unpredictable and hard hitting.

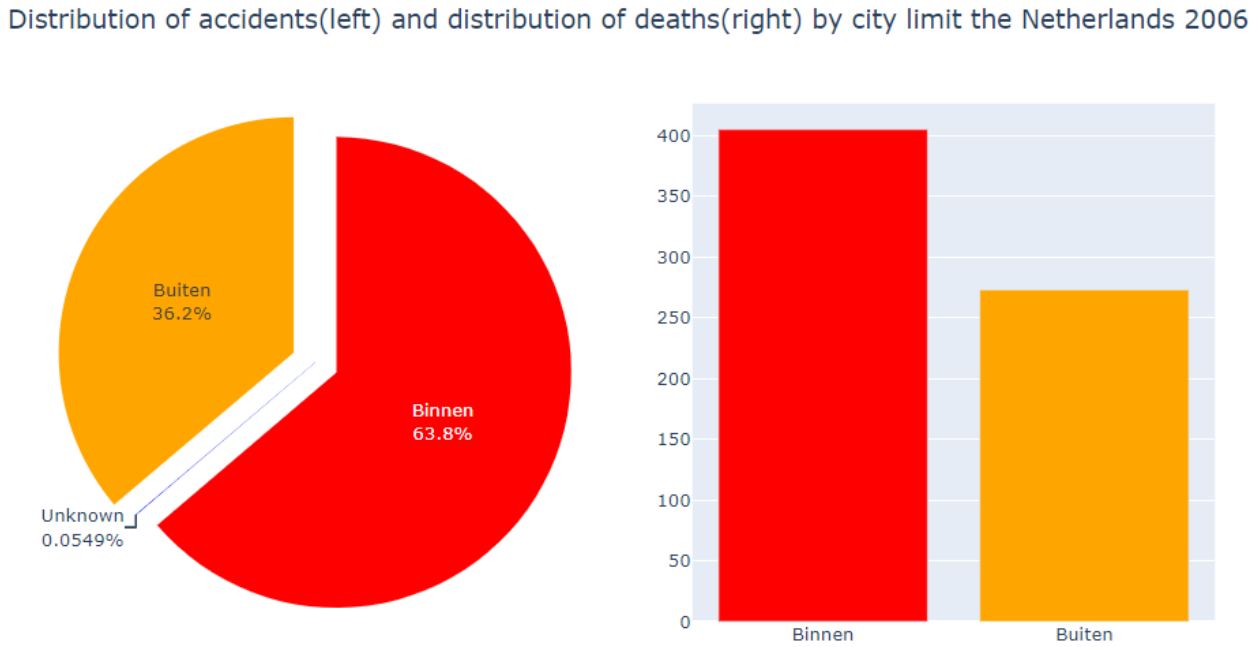

FIGURE 24 DATASET 3: TOTAL ACCIDENTS (LEFT) DEATHS ONLY(RIGHT)

Here we can see that most incidents occur within city limits. This is usually where the highest concentration of traffic is located, including pedestrians and cyclists. Also, public traffic has a frequent appearance in cities. From the graph on the right, we can see that the deaths within city limits are higher. But the rate is lower comparatively lower.

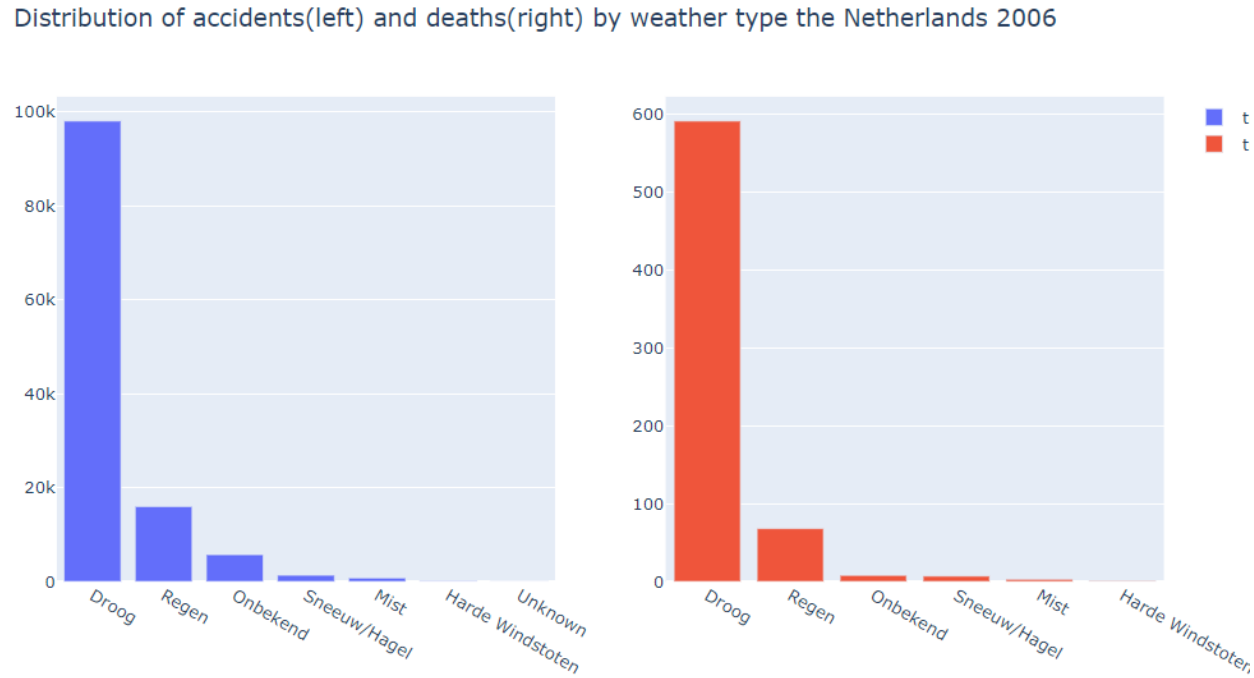

FIGURE 25 DATASET 3: RESULTING DAMAGES BY WEATHER TYPE

Here we can further see that there is no clear connection between the type of accident and weather. With most accidents occurring during dry weather. This could be because participants of traffic ensure better safety with harsher weather conditions.

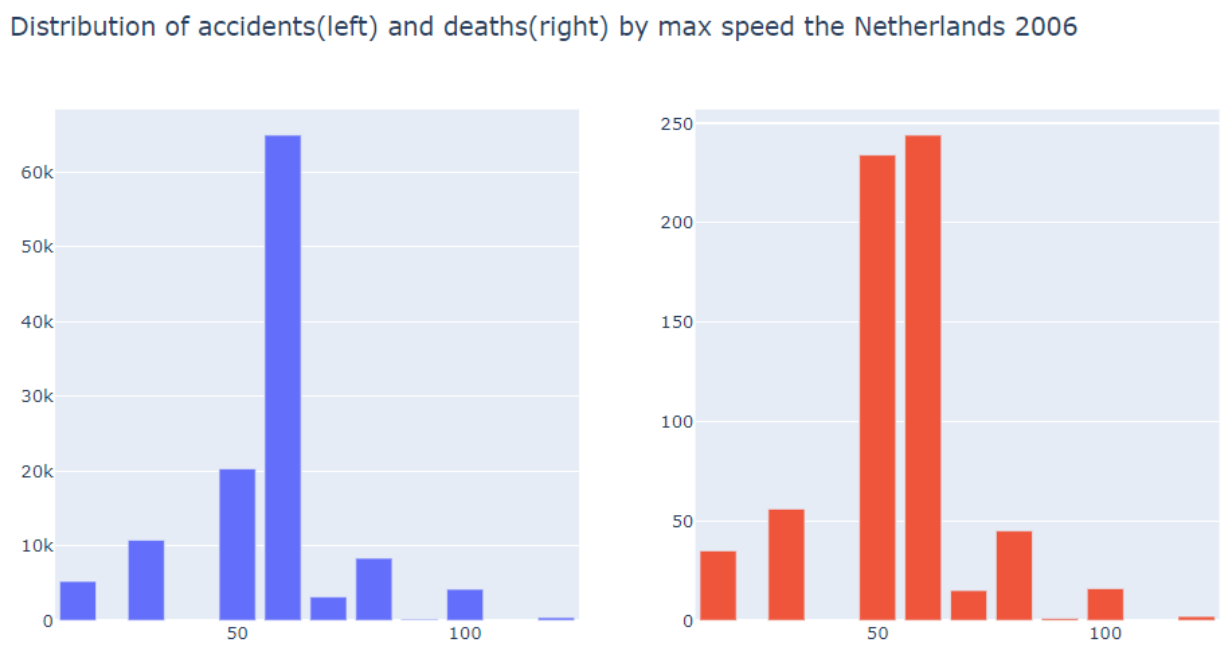

FIGURE 26 DATASET 3: ACCIDENTS AND DEATHS BY SPEED TYPE

As seen in *Figure 24* most incidents and deaths occur inside city limits. This correlates to the speed seen above (city limits max speed in certain circumstances 70km/h). We observe the deaths occur most frequently around the 50-60 mark. These can be fast moving center roads through the cities with traffic coming in and out on all sides. The speed of travel plays a big role in an incident. Less time to react, and more damage done on impact results in high cases of death rates.

4. Final Dataset: Accidents in the UK 2020

In the end the datasets that have been examined don't fully live up to the expectations. That is why use has been made of a final dataset containing all traffic incidents in the UK. The set is split between 3 categories. *Characteristics*, *vehicles* and *casualties*.

Each of these sets has been examined comparatively to the previous sets. These are available in the notebook [1].

3. Analytical Approach

To ensure unbiased and secure data use has been made of several datasets, comparing and tweaking the features so that they match with our requirements. Coming down to one dataset. It is crucial before any decision can be made what is going to predicted?

Target Variable

In our case we are interested in the severity of the incident. By combined use of features predictions can be made using algorithms. The nature of the target variable determines what machine learning algorithms should be applied. We observe three categories

1. Fatal
   Dead on the scene of the incident or at hospital
2. Serious
   Heavy injuries, in need of hospital
3. Slight
   No injury, free to continue journey

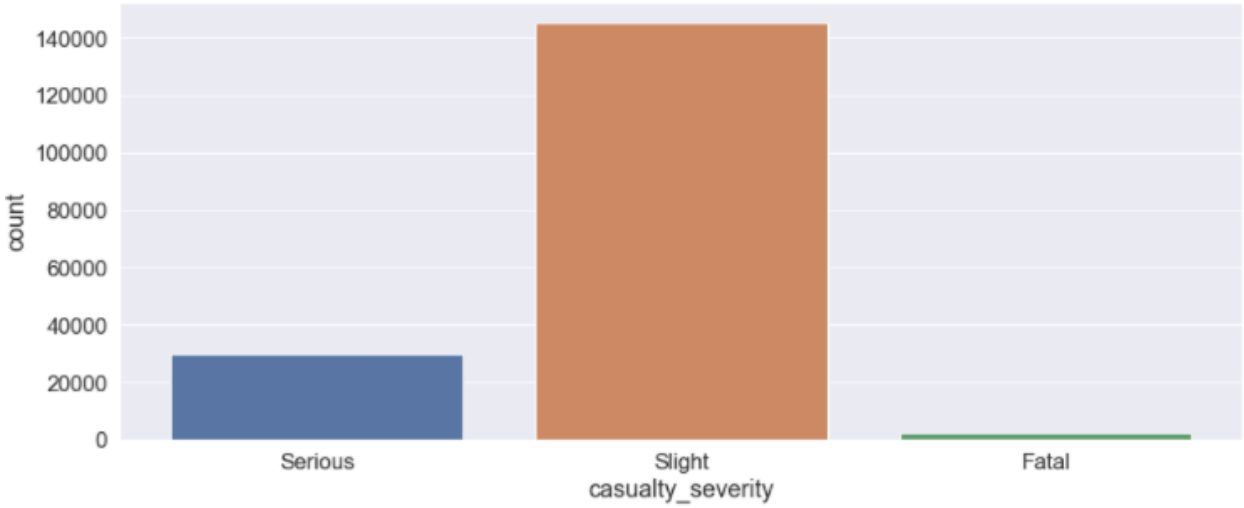

FIGURE 27 DATASET 4: TARGET VARIABLE DISPLACEMENT

In our case we have a big displacement. With most casualties being slightly injured. For the 'machine' to learn data is required. Tactics like over and under sampling should be considered before modelling.

One thing that is certain is the fact that this is a case of classification. Based on the input features, a decision tree algorithm like Random Forest could be applied to assert the severity of this incident. And another interesting challenge would be the chance for each of the classes with regressive algorithms.

For now, the focus will be put to classifying the severity of each case. Regarding this fact use of multiple algorithms will be made to test the perfect fit. Returning to the point of under and over sampling, both sets will be put through the modelling process so that any bias can be overseen and prevented.

Verification

A model can assert a high accuracy. But this does not assert its true nature. In order to have insight into the steps, decision tree classifiers will be applied so that each decision can be overseen. In the end the testing of the results will be based on the comparing actual cases to predicted cases. Accuracy of the models will be based on the r2 score and application of cross validation.

Another tactic that will be applied is a 'Dummy classifier'. This will answer the question, what would the accuracy be if I would guess at random? This will be the final baseline for the success of the model.

References

[1] Jupyter notebook *Traffic Incident Casualty Severity Phase 2 + 3.html*