Task 2. Identifying and Define goals

This is not a business project. The project is a freelance project using data from the Estonian Road Administration

Background:

Over the years the number of accidents in traffic has risen. In the year 2022 over 51 people died and 1562 were injured due to an accident.

The government has changed the speed limit on many roads but still, there are a lot of misdemeanours- speeding, driving while drunk etc.

Goals:

- The main goal is to loIr the misdemeanours percentage and that way prevent major accidents and deaths.
- Provide a heatmap with the misdemeanours so that the government can take action (change the speed limit, put up traffic signs, awareness campaign etc.)

Assessing situation

Inventory of resources

- Dataset 1 (62.5 MB): Misdemeanors for this and last year
- Dataset 2 (177 MB): Misdemeanors for the last five years
- Dataset 3 (92.9 MB): Misdemeanors from the last five to ten years
- Dataset 4 (321 KB): Estonian Iather (2017-2019)
- The Internet
- Jupyter Notebook
- PyData libraries

1 https://www.mnt.ee/et/ametist/statistika/inimkannatanutega-liiklusonnetuste-statistika

Terminology

Misdemeanour- a crime less serious than a felony. Eg: " defacing school property is a *misdemeanour*"

Cost: 0 €

Data-mining goals

Goal 1: Find out and visualize different misdemeanours (what are the most likely places to have serious accidents)

Goal 2: Find out if speed limit changes and speed cameras have affected the number of misdemeanours

Goal 3: Find the correlation between lather and misdemeanours

Data-mining success criteria

I found a correlation between lather and misdemeanours. I can predict the hot spots for misdemeanours or I can predict what kind of lather (day) is more likely to cause a misdemeanour.

Task 3. Data understanding

During this project, I analysed traffic control misdemeanours in Estonia during the last 10 years (2009-2019). The dataset is from europeanddataprotal.eu Ibpage and is originally from opendata.riik.ee Ibpage.

The original size of the dataset is 333MB with 796763 unique rows and 26 columns and the dataset is in Estonian.

Most of the columns are useful with unique and interesting information but there are also some unnecessary columns that I needed to remove or that I needed to modify.

Columns:

JuhtumId - the ID of the incident

ToimKpv - the date of the incident

ToimKell - the time of the incident

ToimNadalapaev - the date of the incident

Seadus - under which law this misdemeanor is occurring

Paragrahv - the law's paragraph

ParagrahvTais - the paragraph's header

Loige - the paragraph's section

Punkt - the section's clause

RikutudOigusnorm - law, paragraph and section taken together in a compact way

MaakondNimetus - the name of the county where this misdemeanor took place

ValdLinnNimetus - the name of the parish or city where this misdemeanor took place

KohtNimetus - the name of the exact place

MntVoiTanav - either highway or street (values: TNV - street, MNT - highway)

MntTanavNimetus -the name of the highway or the street

KM - on which kilometre this occurs when the MntVoiTanav **column is MNT Lest_X** - X coordinates of the misdemeanor in L-EST97 format

Lest_Y - Y coordinates of the misdemeanor in L-EST97 format

SoidukLiik - the type of the vehicle

SoidukRegRiik - the vehicle's location of registration

SoidukMark - the model of the vehicle

SoidukVIAasta - the year of the vehicle

RikkujaSugu - the gender of the person who did the misdemeanour

RikkujaVanus - the age of the person

RIkkujaElukoht - the location where this person lives

SyyteoLiik - the type of the misdemeanour

Keeping my goals in mind, the most important information comes from columns that do not go too in-depth with their information. For example, I needed to know under which paragraph the person got fined, but I don't need to know the precise section of the law. As such, I can drop the "loige" and "punkt" columns.

I have had a brief look at our dataset and I have slightly modified it already. For example, I merged our data together because I had three different datasets and our data was overlapping. By that I mean that I had misdemeanors from 1) this and last year

2) Last five years 3) last five to ten years. By analysing data from dataset 1 and dataset 2 I can see that the last five years and last year are overlapping. I have also found out that there is another irrelevant column that I need to drop. That is "RikkujaElukoht", the location where the person who committed the misdemeanor lives. I am going to drop it because most of the fields are empty

and by having our goals in mind I don't actually need it. There are still modifications that I need to do for our dataset. For example, I need to decide what I am going to do with Nan values in the "RikkujaSugu" column that represents the gender of the person.

Task 4

1. Organizing data - Cristian: 1-2h
    a. Put the misdemeanor data into one collective file
    b. Drop unnecessary data as described in task 3
    c. Replace NaN values with most frequent values where applicable
2. Goal 1: Find out and visualise different misdemeanors (what are the most likely

    places to have serious accidents) - Cristian: 10h
    a. Group the data by misdemeanors
    b. Disregard the misdemeanor groups which don't affect the chance of an accident happening, e.g. "§ 239. Turvavarustuse nõuetekohaselt kinnitamata jätmine"
    c. Visualise the data in a point map using the coordinates from out data. Find a method that is easy to use and looks good.
3. Goal 2: Find out if speed cameras have affected the number of misdemeanours - Marten
    a. Search for the location and install date of the speed cameras in Estonia

b. Group the data by the highways on which the speed cameras are installed, drop other rows
c. Compare the misdemeanor data from before and after the speed cameras Ire installed for the whole highway
d. Compare the misdemeanor data from before and after the speed cameras Ire installed near every speed cameras location

4. Goal 3: Find the correlation betIen Iather and misdemeanors - Maria

    a. Filter out the misdemeanors that are not affected by the Iather

    b. Add the Iather data to our misdemeanor dataframe by closest location/coordinates
    c. Group the data by different days, and check if the day had at least x hours of rain/snow/fog etc, is there a correlation in the number of misdemeanors. The x hours should be chosen depending on what value gives us better information.

    a.