Task 2.Identifying business goals

This is not a business project. The project is for the Estonian Road Administration

Background:

Over the years the number of accidents in traffic has risen. Only this year over 51 people have died and 1562 have injured due to an accident.[1]

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

Business goals:

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

Business success criteria:

The success can be measured by comparing the misdemeanours number, if it decreases, we have succeeded. If it does not decrease, we have failed.

Our success can be judged by the Estonian Statistical Office.

Assessing situation

Inventory of resources

- Dataset 1 (62.5 MB): Misdemeanors for this and last year
- Dataset 2 (177 MB): Misdemeanors for last five years
- Dataset 3 (92.9 MB): Misdemeanors from last five to ten years
- Dataset 4 (321 KB): Estonian weather (2017-2019)
- Team of future data miners (Cristian, Maria Pibilota, Marten)
- Data mining expert Meelis Kull
- The Internet
- Jupyter Notebook
- PyData libraries
- Estonian legislation document

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

Requirements, assumptions, and constraints
We verify that we have access to appropriate data.

Risks and contingencies

- No internet or electricity -> work at a coffee shop or in our institution
- No motivation -> write to the team, go for a walk, take a break
- Not enough time (other schoolwork) -> make a plan, ask for help, less sleep
- Not enough knowledge -> write to team members or Meelis Kull

Terminology
Misdemeanor- noun for when you break the law. Ex: „I went over the speed limit and the police said this is a misdemeanor.“
Paragraph – a chapter of a legalisation document
Section – paragraphs sub-point. Ex: „Paragraph b section a states that misdemeanors must have consequences.“
Case - an instance of a particular situation
Costs and benefits

Cost: 0 €
Benefits: over millions of €
The benefit is that we save lives (misdemeanors cause accidents) and the penal rate is quite high

Defining your data-mining goals
Data-mining goals
Goal 1: Find out and visualise different misdemeanors (what are the most likely places to have serious accidents)
Goal 2: Find out if speed limit change and speed cameras have affected the number of misdemeanours
Goal 3: Find the correlation between weather and misdemeanors

Data-mining success criteria

We find a correlation between weather and misdemeanors. We can predict the hot spots for misdemeanors or we can predict what kind of weather (day) is more likely to cause a misdemeanor.

Task 3. Data understanding (2 points)

During this project we are analysing traffic control misdemeanors in Estonia during the last 10 years (2009-2019). The dataset is from europeanddataprotal.eu webpage and is originally from opendata.riik.ee webpage. The whole dataset comes from Estonian Police and Border Guard Board agency. 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 we need to remove or that we need 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 occuring
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 kilometer this occured 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 misdemeanor
RikkujaVanus - the age of the person
RIkkujaElukoht - the location where this person lives
SyyteoLiik - the type of the misdemeanor

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

We have had a brief look at our dataset and we have slightly modified it already. For example, we merged our data together because we had three different datasets and our data was overlapping. By that we mean that we 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 we can see that last five years and last year are overlapping. We have also found out that there is another irrelevant column that we need to drop. That is "RikkujaElukoht", the location where this person who conducted the misdemeanor lives. We are going to drop it because most of the fields are empty and by having our goals in mind we don't actually need it. There are still modifications that we need to do for our dataset. For example, we need do decide what we are going to do with Nan values in "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 were installed for the whole highway
    d. Compare the misdemeanor data from before and after the speed cameras were installed near every speed cameras location
4. Goal 3: Find the correlation between weather and misdemeanors - Maria
    a. Filter out the misdemeanors that are not affected by the weather
    b. Add the weather 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.