Project D16: Traffic accidents with sustained injuries in Estonia (2011-2021)

Team members:	Anna Maria Tammin,
Robert Raul Matsar,
Martin Hans Keskküla

Project repository: https://github.com/gCoreByte/car-accidents-in-estonia

Dataset 1 (7.66 MB): A collection of traffic accidents in Estonia, with case number, participants, location, date and time.

Task 2. Business understanding

There are lots of statistics about traffic accidents in Estonia, yet there has been very little studies into the effects of location and time. There are also a lot of myths surrounding traffic accidents, such as that holidays are more dangerous as well as locations that locals fear. Our goal is to give road users and road planners an overview of the most dangerous dates and places in Estonia as well as a tool to predict the amount of casualties in an accident from its circumstances.

The exact business goal of this study would be to reduce the amount of casualties from traffic accidents in Estonia. The success criteria for this goal would be a 15% decrease in casualties, however achieving this is beyond the scope of this work and would require a significant investment of time and money.

The resources for this study include a publicly available dataset, which is available at https://avaandmed.eesti.ee/datasets/inimkannatanutega-liiklusonnetuste-andmed, Jupyter Notebook and various data processing libraries, such as scipy, numpy, pandas.

In order for the study to be a success, there are a number of requirements to be fulfilled:

1. Present a poster by the due date
2. Create an interactive heatmap to be able to explore the data freely
3. Have clear and understandable results
4. Be able to draw conclusions from the results

Due to the nature of the study, there is no special terminology used nor is there any additional costs associated. Most, if not all terminology used needs only basic knowledge of traffic and statistics.

To fulfill the business goals, we have set the following data mining goals:

1. Create a heatmap for accidents regarding time and location
2. Create a model to predict the amount of casualties based on various features such as speed, location

The first data mining goal can be fulfilled by using libraries such as hvPlot and Bokeh. This heatmap will be useful in providing a quick overview of the accident frequency by time and location. The second data mining goal can be used as both a cautionary tool and to prove that higher speeds do result in an exponentially growing casualty rate.

Task 3. Data understanding

In order for us to address our data mining goals, we require a dataset of traffic accidents which have taken place in Estonia. The dataset has to at least include the accident's exact location, date, time and some information about the accident's consequences, such that the severity of the accident can be concluded from the data. In addition, we require a dataset of all Estonian holidays. This dataset must include the date and name of the holiday.

We found a suitable dataset of car accidents with sustained injuries in Estonia which is available to the public. It includes all the required information about traffic accidents. The dataset is accessible here. This dataset includes detailed information about 15 708 accidents which have all taken place between 2011 and 2021. But the dataset is not perfect. It includes some empty values, especially regarding the location of the accident. These shortcomings will be addressed, either by discarding some of the data points if they are incomplete, or substituted for default values if such substitution would not change the outcome of the analysis. Lack of accident location will only affect the analysis of the data for our second task of mapping the data, and thus if the location can not be determined, these data points can not be used. If however the data points location can be determined in a less precise way from the other variables (the address of the accident can be used to place the location of the accident on the map in a less precise way than direct coordinates), it can still be used in the generation of the heatmap. Other parts of our analysis are not affected by the lack of location data.

Regarding the data of Estonian holidays, we can easily create the dataset ourselves using the information found on the following wikipedia page: Estonia's national holidays. From the information found on this page, we will put together a simple dataset including just the name and the date of each holiday. This can then be compared to the dataset of the car accidents and used for our first task of understanding when it is the most dangerous to drive in Estonia. One thing that must be understood in this task is that the lack of data for some dates is in itself a datapoint, as it means that on those dates, there were less accidents. For figuring out what constitutes a dangerous day, accidents lethality, severity, count and grouping can be taken as variables to be used when generating the danger levels. These levels can then be assigned to days and the days can be compared to each other.

Task 4. Planning your project

The project will be completed in multiple steps. In regards to the data, first the data must be prepared for analysis. This is an important first step. After data is understood and prepared for processing, we will perform the analysis, in this case, we will have to come up with methods that would perform the needed calculations on the data, such that we can solve the questions we set out to answer. As such, we will split the tasks between the members of the team, develop these methodologies, test them and then analyze the results.

1. Data preparation: splitting variables to be machine readable
2. Data preparation: fixing missing or incomplete/error data points
3. Data analysis:
    a. Figure out what conditions are dangerous for driving, find links to them in the data
    b. Development of methodology that would return the dates of a year when it is most dangerous to drive
    c. Validate the results
4. Data analysis:
    a. Set up and prepare tools and mapping software that would visualize accidents over Estonia
    b. Develop methodology to visualize data on a map
    c. Develop methodology to group data into points on the map for heatmapping
    d. Develop methodology to visualize data in different manners, including in a heatmap
    e. Validate the results
5. Data analysis:
    a. Split the data in a way that its possible to use it for training and testing
    b. Develop the logic that would predict the amount of deaths based on the circumstances of the accident.
    c. Test and validate the prediction algorithm.

It is difficult to predict the hours that each task will take, but it's expected that tasks 1 and 2 are around 5 hours, while task 4 is around 30 hours and tasks 3 and 5 are around 25 hours.

In addition to these tasks, we will all work together to test and analyze each others results, this will help us deliver better results and increase productivity.

The tasks will be split among us in a manner that the workload is similar.

Task	Anna	Robert	Martin
1	Yes		
2			Yes
3	a,b,c		
4		a,b,c,d	
5			a,b,c