Project D6: Road accidents of Estonia - Analyze data on road accidents with human casualties in Estonia 2011-2020

Merlin Raud, Philipp Alexander Hölscher, Kalmer Kaurson

Project repository: https://github.com/kalmerkaurson/IDS-2021-D6

Task 1: Business Understanding

General opinion on traffic safety in Estonia is that the situation is not safe and too many accidents happen. A lot of effort has been put into traffic control and road safety to decrease the number of accidents. Certain periods like holidays (midsummer day, christmas 1st of september etc.) or seasons (summer) are especially controlled due to increased traffic. Construction works are conducted non-stop to make roads safer by making them wider, redesigning intersections, building passing roads etc. In spite of that, many roads and intersections have a reputation of being especially dangerous for drivers and pedestrians.

Business goals

Goal of the project is to find possible correlations between different types of variables such as road parameters, time of year and locations of accidents. This information can be used to find roads where accidents occur more likely or time periods where traffic control should be increased to improve road safety.

Current situation

- Inventory and resources

The project is based on a database containing information about road accidents, with casualties in Estonia between the years 2011 and 2020. 12633 rows of data or 3.61 MB. Each team member has a seperate computer to process the data. For software we can use Python.

- Requirements

The deadline for the project is 13. December and this will be presented on 17.december. The interim report must be submitted 29. November and by this time initial data analysis must be conducted and further project plans must be drafted.

- Risks and contingencies

The highest risk is not being able to follow a plan and not finish the project in time by fulfilling all goals.
When it comes to other risks there aren't many to worry about. If there was an internet outage it would not prevent the work because data can be processed offline. If there were to be a power outage then the best course of action would be to find an alternative location with electricity to finish the work. If some of us were to be unable to contribute to the project then it would be best to start the work early to prevent any time related problems.

- Terminology

The terminology is mostly based on data where different parts of roads like intersections and different accident types are defined.

- Costs and benefits

Costs of the project are time spent on analysing the data and writing reports. The data was gained free from Estonian open data portal.
The most important benefit in this project is the educational aspect for project authors. But also this project intends to find correlations between variables in road accidents that can be used to improve road safety.

Data-mining goals

Our data-mining goal of the project is to find possible correlations between different types of variables such as road parameters, time of year and locations of accidents. A final goal is to prepare a poster and present the results in final seminar

Data-mining success and criteria

The project is considered to be successful when the time and location patterns in Estonian road accidents have been found and interesting information can be visualised.

Task 2: Data Understanding

Gathering data

- Outline data requirements

To achieve tasks, data containing information about road accidents are required. Minimal requirements are: time and date of the accident, location of accident, number of casualties, road and weather conditions, road type and element, type of the accident, type of the driver and vehicle.

- Verify data availability

The data containing information about accidents was available from Estonian open data portal free of charge. The data contains all required information.

- Define selection criteria

Data contains all required information and it is suitable for this project.

Describing data

The data originates from a website called avandmed.eesti.ee. Data is saved as an excel file. In the file there are a total of 12633 cases and 55 variables. The data was in estonian.

Exploring data

The first outlook of the data showed that there are a lot of duplicating variables (complete address as one variable and address separated into multiple variables, many ways to describe weather or road conditions, road elements etc in different columns). Some of the duplicating columns were deleted but keeping in mind that columns containing most information would be kept. In order to use this data, the data was translated into english. First overview of the data and translation was done using EasyMorph software

Verifying data quality

Generally, the data quality is good - not many missing data fields and the values of variables are mostly numeric or categorical.

Task 3: Planning the Project

Task	**Tools**	**Merlin**	**Philipp**	**Kalmer**
Homework 10	Google Docs, Jupiter notebook,	3	3	3
Collecting background info	Internet browser	5	5	5
Goal 1	Jupiter notebook, cartopy, pandas	3	3	3
Goal 2	Jupiter notebook, cartopy, pandas	3	3	3
Goal 3	Jupiter notebook, cartopy, pandas	4	4	4
Goal 4	Jupiter notebook, cartopy, pandas	4	4	4
Goal 5	Jupiter notebook, cartopy, pandas	5	5	5
Preparing poster and presentation	paint.net, Google Drawings, PowerPoint	3	3	3
Total		30	30	30

Goal 1: Find patterns depending on the time of year

Goal 2: Find out what are the main causes for accidents.

Goal 3: Find out what kind of locations have the most accidents.

Goal 4: Find out what kind of locations have the severest accidents.

Goal 5: Predict what locations will most likely have accidents in the future.

List of methods and tools:

We need to use a programming language that is easy to use and has the capability of reaching the goals of the project. Python has many packages that can help out finding correlations between time of year and locations of accidents.