Dataset: https://opendata.riik.ee/et/dataset/avaliku-korra-vastased-ja-avalikus-kohas-toime-pandud-syyteod

Business understanding

1.1. Identifying business goals

Background:
Crime is an everyday part of life. The single criminal events have been documented, however looking at all of the data, it is impossible for a human to make meaningful conclusions. The people in Law Enforcement wish to reduce crime rates to make living in Estonia safer. With the aid of Data Mining, it is possible to find more connections between the single events, and with these conclusions even come up with solutions to decrease the crime rate.

Business goals:

- Identify areas with the highest crime rate.
- Classify crimes with their severity level and identify where and at what time the most severe crimes happen.
- Identify locations where new security cameras should be implemented to discover crimes.
- Information for law enforcement where to show more presence to reduce crime rate.
- Make Estonia safer for its residents and visitors.
- Find out the days with the biggest number of recorded crime cases.

Business success criteria:

Results will be measured with plots and a heat map of Estonia. On the map there would be areas with a hot color where crime rate is high and areas with a cold color (blueish) where crime rate is low. Another map would be a heat map but this time instead of using amount of crimes to describe an area we would use a crime classification. With the help of these maps, it will be possible to organize crime prevention to the areas, with high probability of crime. The real success criteria will be visible after a while. As time passes, crime rates in the defined area should start to drop.

For a measure, if a similar study will be carried out after two years, there should be a drop in crime rates in the defined areas if action has been taken.
The knowledge about times with higher criminal activity can be taken into account, when making the work schedules of law enforcers, the heatmaps, when planning patrol routes.

1.2. Assessing situation

1.2.1. Inventory of resources

Available resources for this project:

- Dataset on criminal offence cases against property in public space (2016-2017);
- Dataset on criminal offence cases against property in public space (2011-2015);
- Explanation letter for the datasets;
- Police and Border Guard Board website;
- Penal Code of Estonia;
- Course materials (lectures, practical sessions) for the Data Mining (MTAT.03.183) course;
- RStudio - a free and open-source integrated development environment for R, a programming language for statistical computing and graphics;
- Human resources (the team, supervisors).

1.2.2. Requirements, assumptions, and constraints

The due date for this project is Jan. 8, when the final poster session will be held.
There are no legal or security obligations since both of the datasets are open to the public and free to use. It has to be noted, that the datasets consist of data from operational level and might change after the proceedings. The datasets are updated once a week on Tuesday.
Our team has access to the datasets.

1.2.3. Risks and contingencies

Contingency plan:

Risk	Solution
Power outage in our workspace.	Find someplace else with access to power (school, library).
Network connectivity problems.	Relocate to either school, library or another place with internet access.
PBGB (Police and Border Guard Board) decides to make the datasets private and/or applies legal restrictions to the data.	We will have to contact them and enquire about the usage of their data. In worse case scenario, we will have to change the topic of this project.
One or more members of the team is unable to work for a period of time.	The other member(s) will do their work and if the situation demands, then we will contact the supervisors of this course to work out a solution.
Underestimating the effort and time that is required to finish the project.	Either make compromises within the project and reschedule or just work hard and do tasks at the first possible opportunity.

In all other scenarios, where the situation is not under our control, we will contact the course supervisors and find the solution.

1.2.4. Terminology

- PBGB - Police and Border Guard Board;
- Open data - data that has been made public and therefore is available for everybody to use;
- Public space - a piece of land, a building, a room or public transport that has no restrictions on who can use them;
- Penal code - document which compiles all, or a significant portion, of a particular jurisdiction's criminal law;
- L-EST - planar rectangular coordinate system, geographical coordinates X and Y are represented as an interval, where the difference of start and end is in meters;
- Data classification - process of organizing data into relevant categories for its most efficient use;
- Sequential patterns - method for identifying trends or regular occurrences of similar events.

1.2.5. Costs and benefits

Costs

- Disk storage - data sets, working data, images, plots, resulting files
- Man hours spent on the project
- Electricity spent during the project
- Psychological trauma caused by discovering crime

Benefits

- Saved monetary value from prevented crimes
- Secure community
- Better management of resources needed to organize public supervision

1.3. Defining data-mining goals

Data-mining goals:

- Group the data according to geographical location.
- Classify the data according to the crime severity.
- Visualise the data to identify the highest crime rate areas.

- Find out abnormalities (e.g there are many crimes happening on the first day of the year)

Data-mining success criteria:

- A heatmap describing areas with higher crime rates (the quality of the criteria will be assessed by the collective of our team)
- Subsets of data according to the classes of crime severity (decided during work, assessed by the team)
- Heatmaps for defined crime types (crime types will be decided during the work - the quality of the criteria will be assessed by the collective of our team)
- Other visualisations (decided on during work, assessed by the collective of our team)
    - Comparisons of different crimes according to some attribute such as lost resources or crime severity
    - Plot of crime occurrences by daytime
- Report of results containing the conclusion and abnormalities (assessed by our team)

Data understanding

Dataset source:

https://www.politsei.ee/et/organisatsioon/analuus-ja-statistika/avaandmed.dot

Brief about the dataset:

https://www.politsei.ee/dotAsset/800283.pdf

Previous studies:

https://www.politsei.ee/et/organisatsioon/analuus-ja-statistika

2.1 Gathering data

We have already gathered two datasets where one datafile covers the years 2016-2017 and the other covering years 2011-2015. The data falls under the following category:

- Public crimes and crimes against property committed in public

2.1.1 Outline data requirements

Type	Format
• Public crimes and crimes against property committed in public in years 2016-2017	CSV

• Public crimes and crimes against property committed in public in years 2011-2015	CSV

2.1.2 Verify data availability

The data is available on the Police and Border Guard Board official website. Links for the datasets are given above and the data has been already saved to our disks incase the data gets broken or taken down.

2.1.3 Define selection criteria

We will be requiring the Public crimes and crimes against property committed in public dataset.

Firstly, the data is given in a single table. Fields of most importance are 'Lest_X' and 'Lest_Y' providing us with the location of the crime. The type of crime (field SyndmusLiik), the law (field Seadus) and kind of crime (field SyyteoLiik) can be used for further classification of instances. Fields like day of week of event (field ToimNadalapaev), time of event (field ToimKell) could be used for comparing instances by times of day or days of week, the date of event (field ToimKpv) can be used to further divide crime by the year, month, season and many more (Holidays).

The field Paragraph (field Paragrahv) should be grouped together with the Law and gives detailed information about the paragraph of the previously mentioned law, about the offense. As we will be looking at crime against property, the expenses (field Kahjusumma) might be of interest - this is given as a range and could be used as a classifier for instances. The various fields about the place of the offense, as well as the type of the place can be used to classify and create comparisons between different locations.

2.2 Describing data

All the detailed explanations about the datasets are in a document[1] provided by the Police and Border Guard Board. The document is in Estonian.

The data in the CSV files is in machine readable format.

The dimensionality of both datasets is described below:

	Dataset 1	Dataset 2
Rows of data	18,498	62,890
Columns	18	18

[1] https://www.politsei.ee/dotAsset/800283.pdf

2.3 Exploring data

Number	Problem	Possible solution
1	Coordinates are in L-EST format	There is a possible API that understands the coordinates in that way.
2	There are columns/rows with no information and it doesn't have N/A there	It will be solved during data cleaning
3	PĆ¼hapĆ¤ev should be "pühapäev" etc	Data cleaning.
4	Some of the robberies don't have the amount of damages	Data cleaning
5	At some points category "SyndmusTaiendavStatLiik" is the same but the main category is different which is odd. This should be thoroughly investigated.	Data cleaning

Summary:
Base on the first exploration of the data, multiple columns with missing data have been found. The number of empty cells is insignificant when compared to the whole number of cells, and often times, the data can be used for other comparison. Our guess for the reason of missing data, is:

- missing time - the event could not be identified in the time;
- missing loss - either there was no loss or the value could not be be estimated;
- missing coordinates - the area was broad or unknown.

Less important missing fields:

- additional information about the kind of event.

At first glance on the distribution of data, we can see, that all weekdays are almost equally covered. In case of loss due to the crime, the number of high value crimes was a lot smaller than the number of petty crime. Also the ranges are rather broad. The reason for this could be the anonymity of cases, given that the events are timestamped. The distribution of

Initial hypotheses:

1. There are more criminal offence cases in high population areas (bigger cities, city centres, malls etc).
2. There are less criminal offence cases related to high value properties.

3. Saturdays seem to have slightly more recorded criminal offence cases.

2.4 Verifying data quality

The data we have is enough to support our goals. There are some quality issues but they are mainly not affecting our goals rather than just some side issues. The issues can be dealt with easily in our opinion and it is just part of data cleaning which needs to be done anyway.

The possible problems are brought out above with the possible solution to this problem. We do not know how much time will be consumed to solve these problems but we have planned to put a lot of time to make the data perfect for us. This data should be good enough to reach our goals if we put in effort.

Setting up and planning the project

- https://github.com/martinuiga/ut_dataminingPosterProject

Task number	Task	Author	Expected time
1	Set goals	ALL	4h
2	Acquire dataset(s)	Leiger	1h
3	First look at the dataset(s)	ALL	1,5h
4	Exploration of data and describing	Karl-Martin	2h
5	Clean dataset(s)	Mart	4h
6	Basic statistics	Leiger	3h
7	Filter data based on business goals	Karl-Martin	5h
8	First data visualization	Mart	4h
9	Heatmap displaying crime density	Leiger	4h
10	Heatamps by crime types (on map)	Karl-Martin	5h
11	Comparison by classes: ● part of land ● Monetary loss	Mart	5h

12	Comparison by time: ● Hourly ● Day of week ● By season ● By year	Leiger	5h
13	More specific analysis (grouping by multiple columns and either comparing or visualising)	Karl-Martin, Mart	5.5h
14	Analysis and summary of results	ALL	6h
15	Visualisation by cities	ALL	5h
16	Writing a report	ALL	5h
17	Designing and finalizing a poster	ALL	10h

SUM 75h