Homework 10 - First steps of the project

Team C3: Regita Luukas, Kadi-Riin Toomoja
Repository: https://github.com/regitaluukas/IDS_C3
Project: Offences against public order and offences committed in public places in Estonia

Task 2

Identifying your business goals

Background

Estonian police and border guard keeps order in Estonia and is a vital part of making our country safer. Every day, the police receive an average of 400 cases to which a patrol goes out. About 26,000 crimes are investigated and about 130,000 misdemeanors are prosecuted each year. Crimes against public order and offenses committed in a public place form a part of these statistics.

We believe that a sense of security is a very important part of our daily lives but we often pay attention to it only when it gets disturbed. Although the biggest responsibility for violating the feeling of security lies with the offender, knowing the patterns and trends of offenses, we can also do a lot ourselves to be more protected and feel safer. It would also help law enforcement to carry out prevention and information work accordingly.

Currently there are not any public easily understandable statistics about how often and where offences against public order and offences committed in public places take place in Estonia (although we assume that such an analysis of the data is already being carried out by the police, but it is not public). Estonian Police and Border Guard Board register every offence, but the big dataset for public of different features about every offence is hard to read and difficult to make assumptions about.

Business goals

Our main goal is to provide people with information that will help them better protect themselves from potential offences and thereby increase their sense of security. We intend to do this by finding different occurring patterns in the offences and visualizing them.

Business success criteria

The project is counted successful when all our goals are met, which means that we manage to analyze and visualize the data so that other students in this course can easily understand the patterns of offences.

Assessing your situation

Inventory of resources

We have a team of two, both team members are new to data science and intend to broaden their skills through the project. The program language used will be Python and also its modules as required. For the data analysis part we have an open dataset with about 39981 offences committed in Estonia from 01.01.2016 to 31.12.2020. We also have a file with explanations about different attributes.

Requirements, assumptions, and constraints

The final deadline for the project on the 13th of December is the deadline for a video and a poster about the project. The dataset that we are using is public and accessible for everyone and the source of the data is the police procedural information system POLIS. The data is also already in csv-format, which is what we need. Our finished work has to include our code, a video and a poster and has to meet the 3 goals that we have set in place.

Risks and contingencies

Main risk is that we don’t manage to achieve all goals or collect any interesting patterns before the deadline. We believe that there can be two main things that could cause this. Here are the problems and solutions for them:

- We don't know what would be the best methods to solve our goals because of our limited knowledge in data science.
- We will get stuck because we have some kind of problem when writing the code (syntax error, logic error etc).
- unforeseen problems

Solution: all the materials (lectures, slides, homeworks and practice session exercises) that we have worked through in the course are available in course wiki and we can probably find answers there, a lot of things can be found through the internet and also for extra help if needed we can contact our lab supervisors.

Terminology

In data-mining terms we will be using only the ones that have been introduced in this course, so we believe that everyone is aware of them.

Offences against public order- Crimes that affect the smooth running of orderly society. Term “offence” includes both crimes and misdemeanors.

Costs and benefits

There is no financial cost because it is just a student project.

Defining your data-mining goals

Data-mining goals

Our data-mining goals are to find interesting relations of attributes in the 5 year timeframe that we have the data from, to form a report on how the amount of reported offences in Estonia varies by county and month and visualize the found data through different diagrams and charts.

Data-mining success criteria

We consider the targets to be met if we have analyzed whether there are any links to the number of offences per month and/or county and also worked on the data to find other interesting links. And then we have been able to visualize them and present them in a form that students and teachers can understand.

Task 3

Gathering data

Outline data requirements

One of our goals is to find patterns about offences throughout the 5 years. We based this requirement on the fact that we have the data from a 5 year period. In another goal we will analyse how the amount of reported offences in Estonia varies by county and month, for that we will definitely need the date of the offence and the county, both are represented in our dataset with other attributes. As mentioned above, our data is already in a suitable csv format, except that the field separator is a tab element. The text in the dataset is in UTF-8 format.

Verify data availability

The data is publicly available at https://avaandmed.eesti.ee/datasets/avaliku-korra-vastased-ja-avalikus-kohas-toime-pandud-suuteod

In our project we use the dataset that is in the “avalik_2.csv” file.

Define selection criteria

We will be using the dataset in the “avalik_2.csv” file that consists of one table.

The irrelevant part of the data are the case id’s, which can be replaced by usual simpler indexes. Also we will not use the X-EST and Y-EST coordinates, paragraph’s text, section and additional events statistical type in our project, so they are irrelevant for us. We can't use the additional events statistical type because this is only required to be marked for misdemeanors, but the data we use covers both misdemeanors and criminal offenses, so some of the rows don’t have it.

Describing data

The data is from Estonian Police and Border Guard Board collected from 01.01.2016 to 12.12.2020. As stated before, the dataset comes as Microsoft Excel's CSV-file. There are in total 39981 offences.

There are 17 features: Case id, Offence’s date, Offence’s time, Offence’s day, Offence’s type, Offence’s statistical type, Law, Paragraph number, Paragraph’s text, Section, Damage, Type of place, County, City, Place, X-EST, Y-EST, Offence’s type.

Based on the task and field values given in the previous sections, we can confirm that all the necessary fields are in place to complete our project and start working to achieve the set goals.

Exploring data

All the attributes in our dataset are described in the explanatory file that is available on the webpage that we got the data from, so we feel it would be unnecessary to just repeat what is already told there, but we will still reports some of the more important things and findings here:

- Damage attribute (in euros) is the only value that is presented in ranges (4 different ranges)
- Even Though it is stated that the text in table is in UTF-8 format we can’t seem to read the letters Ä,Ö,Ü,Õ.

- Right now we have a problem with the fact that some values have the offence's statistical type property and others don't, but the property is also in the middle of the dataset (so not in the beginning or end) and we can't figure out how to remove it from the rows where it is given.

Link to webpage where you can find the explanatory file under the tag "Andmete kasutamise seletuskiri":
https://www.politsei.ee/et/juhend/politseitoeoega-seotud-avaandmed/avaliku-korra-vastased-ja-avalikus-kohas-toime-pandud-varavastased-suuteod

Verifying data quality

The data seems to be high-quality as there are no nan-values and fulfills all our needed requirements for this project.

Task 4

We made our project plan as a table that can be found below. We have divided our work into 6 bigger subtasks. For the work with data we will be using Jupyter Notebook and Python library Pandas, visualization part will be made using Pandas and Numpy (another Python library), for video we will use Google Slides presentation and poster will be made using Photoshop software. Last two columns show how much time each team member will be contributing for this task.

Task	**Methods, tools**	**Regita**	**Kadi-Riin**
Data cleaning	Jupyter Notebook, Pandas	2h	2h
Finding interesting patterns	Jupyter Notebook, Pandas	7h	8h
Analysing variation by month and county	Jupyter Notebook, Pandas	1h	3h
Visualizing data	Jupyter Notebook, Pandas, Numpy (maybe also Tableau)	7h	8h
Video	Google Slides	9h	5h
Poster	Photoshop	4h	4h