Business understanding

Project D11: OFFENCES AGAINST PUBLIC ORDER

Kaarel Rhede and Mirjam Jesmin

https://github.com/kaarelrh/ids2022

https://avaandmed.eesti.ee/datasets/avaliku-korra-vastased-ja-avalikus-kohas-toime-pandud-suuteod

Terminology:

- Breach of public order - Unlawful action that occurs in a public spot.
- Misdemeanour – A lesser offence against public order
- Crime – More significant offence, or a repeated lesser offence

The overall goal of this project is to get an overview of offences against public order in Estonia. The data used for this project is from the Estonian Police and Border Guard Board's website. The three main goals are to get to know where most of these violations happen, second – is there a day or a time of day when these offences happen more often and the final goal is to create a model to predict whether the offence was a misdemeanour or a crime based on the time and location of a given offence. The project is considered successful when these questions have an answer and the model predicts with at least 69% accuracy. The project could be interesting for the Estonian Police and Border Guard Board, but is in no way affiliated with any organisation.

The resources we have available are 2 students, with limited time and motivation. As we do not have university provided hardware, we will at least make a point of using university electricity to charge them.

We assume that the Police and Border Guard have released accurate information, that does not leave out any offences and does not manufacture additional ones to get more funding. However we are constrained by the quality and quantity of data, most of which is caused by some officers not being able to operate a watch and a pen simultaneously. Biggest constraints we have are the time and willingness to concentrate on this project.

The biggest risks arise from commuting to and from Delta Centre of the University of Tartu by car. To minimise the probability of getting into any accidents on the way, we will be travelling during hours with minimal traffic.

Costs	**Benefits**
Time (estimated 60h in total)	Up to 20 points
Fuel for commutes	Personal development
Electricity for laptops	
Wear on laptop keyboards	
Personal development	

Data-mining goals are to have a model that can predict whether an offence is a misdemeanour or a crime, based on the location and the time of the offence and visualise the data.
We will consider our model a success, when we achieve an AUC greater than 0.69.

Data understanding

Project D11: OFFENCES AGAINST PUBLIC ORDER

Kaarel Rhede and Mirjam Jesmin

We need our data to include the time, date, location and type of offence that took place. Estonian Police and Border Guard website has data fulfilling exactly these requirements available under the Creative Commons 3.0 licence. It has a total of 97474 entries (from 2012 until 2021, ten years worth of data). In 237 cases the location is not reported and in 3900 cases the exact time is not reported – depending on the model these entries will be excluded.

The data is from 2012 up to 2022 and is updated weekly. Every entry includes the time, date, location (L-EST coordinates + description e.g. store), type of offence and monetary damage that was suffered. As pointed out earlier, the all the entries not include all the information, but even when we exclude them, we are left with #TODO entries.

Most important attributes for our project are ToimKell (time of the offence), ToimNadalapaev (weekday of the offence), KohtLiik (type of place where the offence took place, for example a store or a street), MaakondNimetus (countie), ValdLinnNimetus (city or municipality), KohtNimetus (district) and SyyteoLiik (type of offence).

46951 reported crimes an
d 50523 reported misdemeanours. Most offences happen in Harjumaa which is logical considering population density. Next are Ida-Virumaa and Tartumaa. Highest percentage of crimes (out of all offences) happen in Tartu (~54%), then Harjumaa (51%), Jõgevamaa (45%) and Võrumaa (44%), lowest percentage of crimes happen in Saaremaa (17%), Lääne-Virumaa (18%) and Hiiumaa (19%).
Most offences happen on Saturdays and least on Mondays (16690 cases vs 12674 cases). Proportion of crimes to misdemeanours is highest on saturdays and lowest on mondays, although the difference is not very significant (46,7% vs 49,4%)

The quality of data is good, all values seem to make sense and do not need extra work.

Project Plan

Project D11: OFFENCES AGAINST PUBLIC ORDER

Kaarel Rhede and Mirjam Jesmin

Task	Time
Procrastinate - anything goes	As much as possible
Investigate data - eyes	~2x 2h 22m 48s
Clean up data - mouse and keyboard + Python	Until it looks fine ~ 2x 2h 22m 48s
Analyse the data - Python and R	~2x 4h 20m 00s
Prepare data for training and testing models - Python	~2x 1h 20m 37s
Try out different models - Python, google, lecture slides, prayer	~2x 11h 52m 18s
Visualise - Python	~2x 1h 41m 19s
Making the poster - canva.com	~2x 3h 52m 29s
Preparing and presenting the project - natural talents	~2x 2h 7m 41s

Task 1. Setting up (0.5 points)

- ~~Create a project repository, either in GitHub or Bitbucket, if You have not done this already.~~
- ~~Make sure that instructors have access to the repository. Invite us by using our usernames, if that does not work, invite us by our e-mails.~~
- Add reports from the following subtasks (business understanding, data understanding and planning) to the repository as a single separate PDF file named GROUP_NR_report.pdf (e.g "A0_report.pdf").
- ~~Add the link of the repository also to the report.~~

Task 2. Business understanding (1 point)

NB! Don't forget to mention your project title and team members at the beginning of the report.

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan.

- For this exercise, please, develop a business understanding of your project. According to CRISP-DM, you should report the following:

- ~~Identifying your business goals~~
  - ~~Background~~
  - ~~Business goals~~
  - ~~Business success criteria~~
- ~~Assessing your situation~~
  - ~~Inventory of resources~~
  - ~~Requirements, assumptions, and constraints~~
  - ~~Risks and contingencies~~
  - ~~Terminology~~
  - ~~Costs and benefits~~
- ~~Defining your data-mining goals~~
  - ~~Data-mining goals~~
  - ~~Data-mining success criteria~~

Please, follow this given structure and cover all these aspects in your report. Consult this PDF-file with a chapter on Embracing the Data-Mining Process for more information on each

of the deliverables. Keep the report concise and **feel free to state that some aspect is not relevant in your project**. If your project is not meant to benefit a 'business', then please specify who will benefit from the project and perform business understanding from their perspective. For instance, this could be either one or multiple individuals, organizations, or societies. **Please focus on the goals that you plan to directly contribute to**, not on the generic goals (like making the world a better place).
The report of task 2 should be **400-800 words**.

Task 3. Data understanding (2 points)
Data understanding within CRISP-DM consists of performing four tasks: gathering data, describing data, exploring data and verifying data quality.

- For this exercise, please develop a data understanding of your project. Report the results of the tasks according to the following structure:
    - Gathering data
        - Outline data requirements
        - Verify data availability
        - Define selection criteria
    - Describing data
    - Exploring data
    - Verifying data quality

Consult the above-given book chapter to understand what is expected under all these deliverables. Take inspiration from when describing and exploring the data. As a result of this exercise, you should have gathered and understood the data. You should have decided which parts of the data you are potentially going to use and understood the meaning of all fields within these parts. Note that data cleaning is part of the data preparation step in CRISP-DM but you might choose to do some of it already during this task.

The report of task 3 should be **400-800** words.

	MaakondNimetus	Vanus.kokku	n	osakaal
1	Harju maakond	614567	77485	0.12608064
2	Ida-Viru maakond	132741	5729	0.04315923
3	Valga maakond	27651	906	0.03276554
4	Tartu maakond	157760	4427	0.02806161
5	Jõgeva maakond	27858	697	0.02501974
6	Põlva maakond	23991	579	0.02413405
7	Saare maakond	31292	752	0.02403170
8	Lääne-Viru maakond	58709	1386	0.02360796
9	Hiiu maakond	8497	197	0.02318465
10	Pärnu maakond	85710	1768	0.02062770
11	Rapla maakond	33529	603	0.01798443
12	Järva maakond	29697	515	0.01734182
13	Lääne maakond	20229	313	0.01547284
14	Viljandi maakond	45413	697	0.01534803
15	Võru maakond	34180	499	0.01459918
16		*NA*	236	*NA*

Juhtumite arv jagatud rahvaarvuga (2021) maakonnas

Task 4. Planning your project (0.5 points)

Please perform the following tasks:

- ~~Make a detailed plan of your project with a list of tasks. There should be at least 5 tasks.~~
- ~~Specify how many hours each team member is going to contribute to each task.~~
- List the methods and tools that you plan to use. ~~Add any comments about the tasks that you think are important to clarify.~~

The report of task 4 should be **100-300** words.

https://andmed.stat.ee/et/stat/rahvaloendus__rel2021__rahvastiku_paiknemine__elukoht-ja-soo-vanusjaotus/RL21001/table/tableViewLayout2