Homework 10: First steps of the project

Satisfaction with life environment and local government in Tallinn and traffic accidents with human casualties

Project repository: https://github.com/liisa-j/IDS-project

TEAM: Liisa Jullinen (Group 7)

Business understanding

Business goals

Satisfaction with life environment plays an important role in overall well-being, both physical and psychological. Since resources are finite, it's important to find the right focus in order to improve both the local environments and the satisfaction with them. Local governments try to achieve this partly by collecting data about satisfaction with life environment and the local governments yearly and analyze it as customer satisfaction data. The local development strategies are partly based on this data and it definitely feeds political decisions.

The motivation of this project was to look at this data a little bit differently, without the political implications, and see, what it is that people are genuinely unsatisfied about and if these attitudes are precise enough to base political decisions upon (or how could this data be combined with other available data). It was also of interest, if demographic factors play a part in how people voice their concerns and attitudes.

At first, the life environment satisfaction dataset was modified (data selected, cleaned, questions arranged into new variables according to correlations and different goals) and some exploratory analyses was conducted. Next, data about traffic accidents was cleaned, summarized (and new variables calculated) and added to the satisfaction dataset (the datasets were combined on municipalities). Due to clarity, only data about Tallinn and its municipalities was analysed.

Questions:

- How are municipalities in Tallinn perceived by their inhabitants? Which demographic variables have an effect on how people perceive their environment?
- How is perceived danger correlated to objective danger (are there more accident / casualties in areas where traffic is seen as dangerous and do the perceived problems in traffic translate to accident causes)?

Assessing the situation

Data about the satisfaction with life environment and local municipal governments is publicly available, as well as data about for example traffic accidents with human casualties. Both are commissioned by governmental organizations and can be used under Creative commons licences (data source has to be cited).

Data will be analysed with relevant Python libraries and results with relevant code will be presented within a Jupyter notebook.

The deadline for finishing the analyses is 9.12.

The most important risks revolve around the large volume of data and trying to find meaningful connections from all the possible information and keeping a reasonable and focused scope.

Costs and benefits: The possible benefits of this analyses could be insights into truly problematic areas of local life.

Data-mining goals

The first goal is to preprocess the data into a form that could be reasonably analysed. This includes combining the separate questions into more reasonable themes (based on both intuition / logic and their intercorrelations).

The second large goal is to find a common feature to be able to merge the datasets meaningfully (including summarizing and feature engineering the traffic accident information).

The third goal is data exploration and finding correlations between features of the two datasets, that would help explain the relationship between people's opinions and the objective documentation of accidents.

In addition to that, it would be interesting to see, which features could be used to create a meaningful linear model for predicting an overall satisfaction score.

Data understanding

This work was based on two main datasets:

1) Satisfaction with life environment and public services in local governments *(Rahulolu elukeskkonna ja avalike teenustega kohalikes omavalitsustes)*: https://avaandmed.eesti.ee/datasets/rahulolu-elukeskkonna-ja-avalike-teenustega-kohalikes-omavalitsustes.
   This is a publicly available dataset, the data was gathered by Turu-uuringute AS in 2022 and is under the creative commons licence 4.0 (author has to be cited).
   Dataset comes in an xlsx format and consists of 218 columns and 10416 rows. The features can be divided into two main groups: demographic variables (nationality, sex, housing

conditions, municipality, family size, salary etc. of the respondent) and questions about the satisfaction in several different areas (including housing conditions, traffic safety and roads, accessibility and quality of public services and governments, immediate environment in terms of safety, pollution, architecture etc.). Data about satisfaction is either in binary format or on a likert scale (1-10).
Main attention will fall on the opinion / satisfaction related questions, but also demographic data will be analysed. In order to keep the analyses more simple, then in the beginning, only rows about Tallinn inhabitants will be kept (ca 10% of the data).

2) Traffic accidents with human injuries/casualties *(Inimkannatanutega liiklusõnnetuste andmed):* https://avaandmed.eesti.ee/datasets/inimkannatanutega-liiklusonnetuste-andmed

   This dataset is also publicly available and is under the creative commons licence 3.0 (author has to be cited). The data has been collected by the Police and Border Guard and is analysed by Estonian Transport Administration. The data covers a period of 2011 to 2022.
   This dataset comes in a csv format and consists of 54 columns and 17412 rows and contains detailed information about traffic accidents with human casualties (including the data and time, place, number of people injured or died, information about the nature of the accident, road conditions, types of vehicles or people involved etc.).
   In this dataset the fields regarding the number of people injured or deceased, the municipality where the accident took place, who participated (either pedestrians or motor vehicle passengers) and what the conditions were like (road and light conditions) will be used. In this dataset also, for purpose of simplicity, only data about Tallinn will be used.

In addition to that data about Tallinn demographics (https://www.tallinn.ee/et/media/505586) and Tallinn municipalities (https://et.wikipedia.org/wiki/Tallinna_linnaosad) will be added to dataframes, with the goal to calculate the average number of people injured/deceased per km^2 or per 10 000 inhabitants in the municipalities of Tallinn within the 11-year period (2011-2022).

The data described is readily available and does not have any major quality issues. There are missing values, but these can be addressed and missing values will be imputed by column means (missing values have different representations within these datasets, for example ‘nan’, ‘99’ and in some cases other numeric representations).

Planning your project

1. **Data preparation.** Most of the effort in this project will go into data preparation, including cleaning, constructing features, integrating and formatting data. In this phase features will be selected, renamed where necessary (mostly in the satisfaction of environment dataset), necessary data transformations conducted. Features concerning respondents satisfaction with living environment will be combined into new features and possibly also clustering of the participants will be conducted. Missing values will be dealt with.

The traffic data will have to be cleaned, summarized by a common features with the previous dataset, new features based on added information about the size and population of municipalities engineered.

2. **Data exploration** – including constructing figures that represent either some interesting facts found inside the data or an overall estimation of the satisfaction will be presented. In case of overall estimations, also the mean satisfaction in the whole country will be used as a reference point.
3. **The relationship** of inhabitants' satisfaction in a specific field (in this case traffic security) with the actual accident rates will be analysed.
4. The overall satisfaction score will be tried to be **modelled** using demographic factors (an extra).
5. **Evaluating the success** of the process, possibly returning to some previous steps.

All of the relevant parts of the whole process will be documented and the workload will be approximately 30 hours.