Report WIP

https://github.com/hanneskolk/Project-D2

Business understanding

We are Hannes Kolk and Erik Kippus and our data science project D2 is aimed at analysing Estonian driving school exams. We aim to provide a better understanding of driving exams in Estonia and which factors have significance. Students who are currently planning on getting a driver's license or are already in the process may come across difficulties and questions surrounding driving exams. Information on the subject is scarce, which is the primary driving factor for this project.

We believe the answers people seek the most concern the following: driving school selection, which is the most important choice in a driver's license process; driving exam location, as many decide to do their driving exam in a different city in order to have the highest chance of success; frequent mistakes during an exam (also applicable in driving lessons), to have the best overview of possible mistakes to avoid. Answering these questions is difficult, as people have different skills, preferences and possibilities. Therefore, the analysis is based purely on statistics and data, to allow individuals to make their own choices with sufficient knowledge.

The final goal of the project is to improve the success rate of driving exams. By providing people with statistical tips on driving exams, schools and locations, we hope to increase the percentage of people who pass their exams. This leads to better drivers in traffic, safer roads and fewer accidents.

For this project, we have decided to use Estonia's Transportation Authority's publicly available datasets, 'Transpordiameti avaandmed'. These datasets contain data from 2021 to 2024. Data from the latter year is still underway, but as of writing this report, it contains exam results until October 2024. For analysis and visualisation, we will use libraries and algorithms available to us in the Jupyter Notebook environment. This may include elements such as pandas and numpy for analysis, matplotlib and seaborn for visualisation, and scipy for more advanced statistical processing.

The data we use is publicly available, meaning there are few legal limits or obligations in its usage. However, due to personal data policies, we don't have access nor do we need individuals’ names or other personal data. The information we use is purely on driving exams. Driving school names are available and will be used for the purposes of this project.

There are few risks, however preparation is still necessary. One possible risk is loss of data, should the website hosting the datasets encounter issues. This requires little planning and preparing, as we have the datasets saved and ready for use. Another risk is data integrity. Our first visual analysis showed that the data has issues, such as poor formatting, missing contents, etc. This may slow down the data analysis process, should the cleaning and reformatting prove to be more difficult than anticipated. This however would be purely a delay and does not require contingencies.

Next is a description of terminology used in the project:

- Driving instructor - Individual responsible for teaching a student, guiding them through the process and providing information on correct and incorrect behaviour.
- Examiner - Individual responsible for grading a student on their exam. Usually the examiner does not provide information on mistakes during the exam, however they will do so once the exam is over.
- Cancelling an exam and failing an exam - It’s important to bring out the distinction between these two. An exam is cancelled, if the exam has not started for a reason, such as the student not showing up. An exam is failed if the people involved are in the process of the exam already and a critical mistake is made.
- Right of way - Term used in traffic for someone who has the legal right to move first in a situation. An example of this is a crossroads, where two people wish to move in a manner that their paths cross. In this situation, the person “with the right of way” moves first.

The costs and benefits section of this report is insignificant, as the project requires no costs other than time. Therefore, the benefits exceed the costs, as they provide informational benefits.

For this project, our goal regarding data mining involves analysis and visualisation of the aforementioned data. Our first goal is to process the datasets, clean them up and refactor them into data that is ready to be used. Next, we plan to begin analysing the data. We have 4 primary goals we wish to achieve for this:

- Analyse the success rates of driving exams per year
- Compare the effects of driving schools on exam success rates
- Discover the highest impact mistakes, in regards to both severity and occurrence
- Analyse rates of success by location

Next, we wish to visualise the data to simplify the information, assist in identifying patterns and communicate our results to others efficiently. The goal is to make our data easily readable, explain any irregularities and trends, and to provide a visual description of our analysis.

Our data mining process is successful, if there are no issues accessing the data, analysis shows no unexplained irregularities or extremes, and visual models display our results in a logical manner, where the information is easily interpretable and any irregular spikes are explainable.

Data understanding

Gathering data:

- Data Requirements: The dataset must include the following:
  - **Student Information**: Anonymous ID of the student.
  - **Exam Details**: Date, city, category of license (A, B, C, etc.), special conditions, and whether the driving instructor was present.
  - **Examiner Information**: Anonymous ID of the examiner.
  - **Exam Outcome**: Pass/fail status, duration, and reasons for failure or interruption.
  - **Violation Data**: Traffic rules violated, along with detailed failure reasons (if available).
- Data Availability: The datasets are sourced from Estonia’s Transportation Authority’s publicly available resources (2021–2024). All necessary data has been downloaded and is accessible for analysis.
- Selection Criteria:
  - Date of exam, location, license category, last driving school, examiner ID,

exam results, length of the exam, and reasons for failure (if applicable).
    - Irrelevant fields, such as unrelated personal information, will be excluded.

Describing Data:

The dataset includes the following columns:

- ID of the student taking the exam (“EKSAMI_SOORITAJA”)
- Date of the exam (“KUUPAEV”)
- City where the exam took place (“BYROO”)
- License category (A, B, C, etc.) (“KATEGOORIA”)
- Special conditions (“ERITINGIMUSED”)
- Last driving school before the exam (“VIIMANE_AUTOKOOL”)
- Whether or not the driving instructor was with the student on the exam (“SOIDUOPETAJA_KAASAS”)
- ID of the examiner (“EKSAMINEERIJA”)
- Whether or not the student passed the exam (“SEISUND”)
- Exam duration (“KESTUS”)
- Reason for interruption (“KATK_POHJUS”)
- Not passed (“MITTEARVESTATUD”)
- Traffic rules violated and detailed reasons for failure (if available) (“VEAD”)

Initial inspection revealed 40000-50000 rows each year. The dataset contains fields with varying data types: categorical (e.g., “SEISUND”), numerical (e.g., “KESTUS”), textual (e.g., “VIIMANE_AUTOKOOL”), and date fields (e.g., ”KUUPAEV”). Certain fields, such as reasons for interruption, may contain missing or null values requiring preprocessing. Some textual fields, such as reasons for failure, may include multiple values separated by delimiters, requiring parsing for analysis.

Exploring Data

Key exploratory steps:

- **Basic Statistics**: Checked distribution of pass rates, failure reasons, and geographic trends.
- **Visual Patterns**: Preliminary bar and pie charts show differences in success rates by year and location.

- **Missing Values**: Observed incomplete records for "Failure Reason" and inconsistencies in city names.

Verifying Data Quality

- **Completeness**: Most of the rows are missing the field “MITTESOORITATUD”, but the result of the exam is already described in the field “SEISUND”.
- **Accuracy**: Some fields (e.g., “BYROO”) require standardization to avoid duplicate entries.
- **Consistency**: Numeric fields (e.g.,”SEISUND”) align with expected ranges.
- **Validity**: Data formatting issues include non-uniform capitalization and extra whitespace.
- **Action Plan**: Exclude rows with critical missing values (e.g., no exam result or location) during analysis. Standardize textual fields (e.g., driving school names) to avoid duplication. Analyze fields like “KATK_POHJUS” and “VEAD” separately for interrupted and failed exams.

Project planning

Our project is divided into 3 primary tasks. The tasks are as follows:

- Data gathering, cleaning and understanding
  - First we need to understand what information we can gather from the datasets, what they describe and how they will be specifically used for the analysis.
  - Next we need to gather the data, reformat it into data we can directly process and work with.
  - Finally we need to make the data clean, remove any unusable or unnecessary elements and make it accessible for software.
    - This will require at most 10 hours from both of us. This is a more pessimistic estimate, considering the format of the data and the possible time it takes to clean and reformat any strange fields.
    - We will most likely be using the ‘pandas’ library for this, as it is the best pick for reading and cleaning csv files.

- Data analysis
    - In this section we will analyse the data in regards to the 4 goals mentioned before.
    - Analyse the success rates of driving exams per year.
    - Compare the effects of driving schools on exam success rates.
    - Discover the highest impact mistakes, in regards to both severity and occurrence.
    - Analyse rates of success by location.
        - This is the most critical part of the project, current estimation is around 20 hours from both of us, meaning around 10 hours in total for reaching each goal.
        - Here we will be using many libraries, such as pandas, numpy, scipy and others to assist us in analysis, comparisons and possible correlations.
- Data visualisation
    - This section is linked with the analysis section, as we will provide visual models and representations for each goal separately.
    - The primary goal here is to make the models understandable and provide explanations for possible irregularities. A lesser goal is to make the models visually appealing and easily readable.
        - Time estimation for this section is around 25-30 hours in total.
        - Currently we plan on using a mix of matplotlib and seaborn libraries, as they provide the most versatility.