Project title: Analysing Tartu Smart Bike usage data 2019-2021

Aleksander Tamm, Martin Vendelin, Otto Jõeleht

Task 1. Setting up

https://github.com/OttoJoeleht/IDS-2021-C2

Task 2. Business understanding

Background

A bicycle-sharing system is a service in which a fleet of shared bicycles is made available to the public on a short-term basis through self-served docking stations. The bicycles may be rented and returned to any of the docking stations included in the system. The main goal is to provide a sustainable and flexible alternative to commuting in urban areas to reduce congestion and pollution caused by personal cars. To ensure the success of the business the whole city is covered with docking stations that hold nearly 20 bicycles each. Each station has an information board with detailed instructions on how to use the bicycle.

Business goals

- Improve overall usage of bike-sharing in the city.
- Locate docking stations that are used very often and those that aren't.
- The number of cases where a bike was used for more than five hours on a year basis.

Business success criteria

After successful data mining we can rank the docking stations by usage frequency. Based on the ranking we can locate stations that are used very often and those that aren't. In addition, we can consider data mining to be a success if we find at least one customer who has used a bike for more than five hours straight.

Situation

We have been provided access to sensitive data about Tartu city bikes, for the duration of 2019 June - 2021 April. The dataset contains 1.7 million rides, and for each ride, there is the following information: bike ID; start station and timestamp; end station and timestamp; length of the ride in km; gender and year of birth of the rider.

Our main tool to use for analysing the data is Pandas. Pandas is a software library written for the Python programming language for data manipulation and analysis. To give a better overview of the results visually we are going to use Matplotlib, which is a plotting library for the Python programming language.

There are a couple possible risks that can occur. Firstly, we only have two weeks left to complete the whole project, but this time may not be enough to achieve the goals we have set for ourselves. Secondly, the data we have provided may be incomplete, which can lead to a problem where we don't have enough data to find solid patterns between data points. Thirdly, the results we achieve are below our set expectations.

The main cost that accompanies this project is mainly time. Each team member must work at least 30 hours. On the other hand, the benefit far outweighs the cost. We acquire important knowledge about data mining and also get a better understanding of bike-sharing systems.

Data-mining

- Predict the number of rentals on a daily and hourly basis.
- Find the type of customers using a certain station based on its location factors.
- Find out which stations are most actively used.
- Find out between which stations the bikes are most frequently used.
- Classify which age groups and sex use the bikes most frequently.

Data-mining success criteria

We want to achieve an accuracy of 75% to predict the number of rentals on a daily and hourly basis. To achieve better accuracy we can train the model for each year separately and also for all years combined, if necessary. We aim to divide people into age groups as follows: 0-18; 19-35; 36-50; 51-65; 66 and older.

Task 3. Data understanding

- Gathering data
    - Outline data requirements

We need data of Tartu Smart Bike usage from the years 2019 to 2021. Since we are analysing data we use given data and do not set requirements to it but base our goals on it.

- Verify data availability

Required sensitive data exists and we have granted access to it by Tartu City Council.

- Define selection criteria

We have 3 csv files (2019.csv, 2020.csv, 2021.csv) which contain data from those years. Each file consists of table with columns: cycle number, unlocked at(date), unlocked at time(time), locked at(date), locked at time(time), startstation serial number, start station name, endstation serial number, endstation name, length(in kilometers), year Of Birth, first 3 Id Number(first 3 digits of Estonian ID- code). And rattaringlus_koordinadid.xls which contains data of bike stations name, opening year, status, no of bike slots, x and y coordinates. We will use data in which sex can be made sure, eg by 1 digit of ID. Also If the end station is missing the data for this ride can't be used to assess behaviour, although it can be used to find out how many people did not dock their bike.

- Describing data

Data is from Tartu city administration, who manages those bikes. Data is contained in 3 .csv files containing data of their respective years. Those three files contain 1 701 595 entries consisting of columns: cycle number, unlocked at(date), unlocked at time(time), locked at(date), locked at time(time), startstation serial number, start station name, endstation serial number, endstation name, length(in kilometers), year Of Birth, first 3 Id Number(first 3 digits of Estonian ID- code) . In addition one file contains information of 93 bike docking stations which contains data of bike stations name, opening year, status, no of bike slots, x and y coordinates. Data contains necessary fields and 1.7million cases should be sufficient for analysis. Since our goals are quite broad: to find habits and maybe other interesting conclusions - lack of problem, our data-mining goals will be suited for data, thus data is suitable for our project.

- Exploring data

Quite a big proportion of data is missing, nearly 40% of year of birth's from 2019 data. Also rides with faulty distances exist in numbers, eg ride with 161 kilometers in 7 minutes. And the station serial number and name don't match. Also much of the rides are 0 km and 0 min and those also need to be excluded, about 70 000 rides in the whole data. In conclusion data needs to be cleaned thoroughly and checks need to be placed(eg average speed over 25 km/h is suspicious. Good example of broken data:

```
cyclenumber                    2721.0
unlockedat                  2020-12-20
unlockedattime                21:53:02
lockedat                    2020-12-20
lockedattime                  22:00:12
startstationserialnumber            29
startstationname           Raudteejaam
endstationserialnumber              29
endstationname                   Tulbi
length                          161.83
yearOfBirth                        NaN
first3IdNumber                     NaN
```

- Verifying data quality

  Data is with defects, but some of it can be repaired and since original data consists of over 1.7 million entries excluding in the worst case ~30% of it will leave quite much to work with. Also if we have to discard more data we can do separate analysis by excluding data with some defects but leaving others in, missing year of birth doesn't disturb finding out most popular stations etc.

Task 4. Planning your project

1. Cleaning data
    a. finding missing data and trying to restore it, eg first 3 digits of id gives date of birth
    b. excluding data with missing elements (for example some of the bikes are not docked in the end of ride and we don't have information about where it was picked up. But we still can get some information from it about general usage (gender, age, picking up location to understand which docks are more popular for starting points). Also we could try to find a profile of a person who is most likely to not dock his/her bike after the ride.)
    c. tools: jupyter notebook
    d. hours: 2-5
2. first statistics
    a. How much data is left after cleaning? How much data is unusable? Which data can be fully used and which data can be partly used?
    b. Some simple general statistics (averages, means and so on) which will be base for further data mining (based on age, gender, length)
    c. tools: jupyter notebook
    d. hours: 1-3
3. Constructing the map about rides
    a. Major benefit for the client is that they can understand where to provide the better solutions regarding bike roads and traffic. When we understand which parts of city is being more used by the bikes, then for these parts they have the opportunity to improve the traffic situation.
    b. Since we only have starting points and ending points, we don't know the specific road exactly, but we can assume that the average rider only rides from point A directly to point B. When we draw the map considering that, then the most busy places should come out.
    c. tools: jupyter notebook and GeoPandas.

d. hours: 5-30 (hard to predict, haven't used this tool before)

4. Trying to construct a model to predict future instances:

    a. The main goal of this task for the client is so they can provide a better experience for the riders.

    b. Some cases: when it comes to starting rides (for there to be sufficient amount of bikes), finishing rides (to be available docks), trying to guess the person who don't dock their bike (maybe a little popup warning already ahead in the app for these people?), and when we look at the actual data we can probably get more ideas. We can even separate (if we have enough time) the weather data for the winter period and try to see the usage then when it comes to bad weathers, snow height and so on.

    c. tools: jupyter notebook, keras

    d. hours: 15-30

5. Conclusions

    a. The conclusion will be done in a readable way for average people who isn't so interested to see data only in the charts (so for the client in that case). Of course there will be some simple charts too because charts are an excellent way of showing data.

    b. Project report for example will give information about (but not only!):

        - from which stations are started the shortest routes (benefit for the client is the fact that to these stations they can bring less electric bikes)
        - from which stations are started the longest routes (benefit for the client is the fact that to these stations they can bring more electric bikes)
        - which is the average profile of a person who don't dock their bike
        - from which stations averagely the oldest riders start from (benefit for the client is the fact that to these stations they can bring more electric bikes)
        - different statistics based on gender, age, routes.
        - the map about where the bikes ride the most.

    c. Majorly on Google slides

    d. hours: 2-4