Analysis of smart bike data in correspondence to weather and coronavirus test data

Team members: Agnes Annilo, Kerdo Kurs, Cardo Tisler
Link to repository: https://github.com/kerdokurs/ids_project

Business Understanding

Identifying Business Goals

Background: The aim of this project is to study the effects of the corona pandemic on the biking traffic in Tartu. It is difficult to guess if the pandemic caused the traffic to increase or decrease because on one hand there were less events and less reasons to go out, which means there were probably less reasons to use the bikes, however there were probably more people preferring the bikes as a means of transportation rather than public transport since public transport tends to get quite crowded and that increases your chances of getting infected with the virus.

Business goals: Our goal is to find out how the coronavirus pandemic affected the Tartu Citybike traffic. We think that the answers to these questions might be of use when deciding how to act in this pandemic. For example if the pandemic increased the overall traffic of people using the bikes, that might be a good reason to further increase investing into the Tartu Citybike system, because if there are more bikes and more people preferring this method of transport, we can decrease the amount of people sharing public transport and potentially infecting each other, therefore reducing the virus-related strain that has currently been placed on our healthcare system. The same logic applies to public transport, if the pandemic decreased the amount of people using Citybikes, perhaps it would be smart to analyze public transport data to see if more people started using other means of transport. In that case perhaps increasing funding to those sectors could also give us the same benefits as listed here.

Business Success Criteria: Our project does not have a success/failure threshold, since we think that the knowledge gained from this project will be useful either way. We will measure our success by the feedback gained from our course supervisor and the teaching assistants.

Assessing our situation

Inventory of resources: Our main help contact would probably be Meelis Kull, as he has knowledge of the bike data. We are using three datasets: Tartu city bike data, Tartu city weather data and Coronavirus test result data. We are doing our analysis in Python on Google Colab.

Requirements, assumptions, and constraints: We can not publish the data to anyone and our results can be published as a summary. The summary must be assessed by Meelis Kull before publishing. We must complete the project by the 16. December.

Risks and contingencies: Our main risks are related to the datasets being quite huge. We might need to reserve extra time to take this into account. All of our team members work and study at the same time, so finding time which suits us all to complete tasks might be difficult, however to avoid this, we decided to split up tasks between the team.

We do not have terminology for our project yet, and we do not need to assess cost and benefits. The only cost would be time, but since we need to finish this project to get a grade, we will need to get some sort of results by the 16th of December.

Defining your data-mining goals

Data-Mining goals: Main goal is to identify the effects of the pandemic on Tartu Citybike traffic. Our plan is to find spikes and drops in the traffic and compare it to the reported amount of infected people. We must also take into account the periods where the pandemic-related restrictions were active. For a deeper understanding, we will create a model to predict whether a person would ride a bike on a given day (based on weather, coronavirus and personal data).

Data-Mining Success Criteria: We will measure the success of our data-mining process by the feedback gained from the course supervisor and the teaching assistants. We would want our model to have an accuracy of at least 75%, however if our model does not perform well, we might have to stick to different types of analysis.

Data Understanding

Gathering data

For this project we will be using three datasets. Tartu city bike data, Tartu city weather data and Coronavirus test result data. We received the Tartu city bike data from Tartu city with the assistance of Meelis Kull. The coronavirus test result data was gained from Terviseamet (https://www.terviseamet.ee/et/koroonaviirus/koroonaviiruse-andmestik). The weather data was acquired from the Physicum website (https://meteo.physic.ut.ee/?lang=en).

Data requirements

Outline data requirements: We decided that for all datasets we will analyze 6 months of data for each year from the year 2019, 1st of May to 31st October.

Verify data availability: The data availability has been verified and the data has been acquired.

Selection criteria:
Weather data: We will use data in the columns [Aeg, Temperatuur, Niiskus, Tuule kiirus, Sademed] aka [Time, Temperature, Humidity, Wind speed, Precipitation].

City bike data: We will use data in the columns [unlockedat, unlockedattime, lockedat, lockedattime, startstationserialnumber, startstationname, endstationserialnumber, endstationname, length, yearOfBirth].

Coronavirus test result data: We will use all data in this dataset.

Describing data:

We will have 3 different sources of data in csv format. We will have to do a bit of data cleanup and grouping, so we do not yet have a clear understanding of the sizes of our datasets. The fields are not described, but the field names are understandable and do not need any deeper descriptions. We have all the fields we wish to use in the datasets. The fields do not seem to have any missing values or errors and we have a lot of data.

Exploring data:
In our weather dataset, we have the values: aeg (datetime), temperatuur (numeric) in celsius, niiskus (numeric) in percentage, tuule kiirus (numeric) meters per second, sademed (numeric) millimeters.
City bike data values are: unlockedat (date), unlockedattime (datetime), lockedat (date), lockedattime (datetime), startstationserialnumber (numeric), startstationname (nominal), endstationserialnumber (numeric), endstationname (nominal), length (numeric) the length of the bike ride, yearOfBirth (numeric).

Data quality issues:
Our data did not have any serious issues. The data is correct and values are logical.

Planning our project

Task	Description of task	Member	Time
Cleaning the dataset	Getting needed columns, finding errors in data, removing excess data outside specified date range, selecting Coronavirus result data only from Tartu city	Kerdo	~3 hours
Initial analysis of data	Creating an initial overview of the data	Agnes	~4 hours
Grouping data/ writing function to group data	We will need to group coronavirus results to mirror the number of results ~7 days before a given date. This would give us an overview of how the last ~7 day result data affected city bike usage	Cardo	~3 hours
Creating visualisation for at least 3 different relations	To visualise the analysis for the end result. The initial analysis will give some sort of understanding what	Every team member	1 to 2 hours per team member.

	might have a correlation in the data, and what might be interesting to visualise.		
Creating a model	This includes assessing which data columns to use, based on the initial analysis. What grouping of time to use, etc. This means we might need to reword our main goals.	Every team member will do something to help	at least 5 hours per team member