Hannaliina Rakovitš
Emma Belinda Semilarski
Kati Sõer

Tartu Smart Bike Share

Project number: D8

https://github.com/emmasemilarski/Tartu_smart_bike_share

Business understanding

Background

Tartu Smart Bike Share is a business that has been around since the year 2019. The business idea is that there are bike stations all over the city and there are multiple bikes in a station, both classic and electric assist. Bikes in the stations are available for a ride from 5AM to 1AM to people with a membership or a bus ticket. Tartu Smart Bike Share offers people a more environmentally friendly alternative to a car or a bus. In the last two years some problems with the service have arisen - such as some bike stations are empty whilst other stations are full. This not only allows less people to use the service but makes using the bikes less convenient as the desired start station might be empty. More people using this service would be in the best interest of the city of Tartu. It would make a great environmental impact and improve the overall health of the people in this town. There would also be less traffic on the roads if more people used bicycles as their primary means of transport. This begs the question of how to get more people to use this service – who the people are already using cycling as their main way of commuting and to whom should this service be advertised.

Business goals

This project intends to make Tartu Smart Bike Share a better experience for its customers. The goal is to make Tartu Smart Bike Share an easy everyday way of commuting. To ensure the best possible experience for its users, the plan is to analyse how the logistics of this business can be organized in a way where there are always bikes nearby for the customers. This entails identifying the most popular locations at a given date and time by analysing the data of millions of rides so that

the bikes could possibly be relocated. Our wish is to optimise the occupancy rates of the stations and to get more people to use this service.

Business success criteria

The project is considered successful when

- we have identified a way to optimise the occupancy rates of the stations,
- we have profiled the existing customer base,
- gained concrete insights on to whom to advertise this service,
- Tartu Smart Bike Share team gains ideas from our project.

Inventory of resources

We have data from over a million rides done since the year 2019 and information about all the bike stations in the city.

The first dataset contains information about specific rides done - such as:

- the ID of the bicycle,
- when the bicycle was unlocked (date and time),
- when the bicycle was locked (date and time),
- the start station's name and ID,
- the end station's name and ID,
- the length of the ride in kilometres,
- the user's year of birth and the first three digits of their social security number.

The second dataset contains information about all the bike stations – such as:

- the name of the station,
- the station's maximum capacity,
- the status of the station,
- the station's year of installation,
- the station's x and y coordinates.

Requirements, assumptions, and constraints

The project deadline is 16th December 2021. The report of this project must be submitted by 29th November 2021. The report must not only include the initial analysis of the business and given data, but the project plan as well. We are required

not to share data given to us with anybody that's not related to this project, and we must delete it from our computers by 1st February 2022.

Risks and contingencies

We face the following risks:

- we might not be able to achieve all goals set,
- there could be corrupt and/or contradictory data,
- we might not be able to come to a clear conclusion,
- one of our computers stops working.

Our contingency plan if those things happen contains the following:

- we might have to realise that our knowledge in this field is limited, and due to that might not be able to accomplish everything,
- we might have to change some goals along the way,
- we might have to exclude some data from our analysis,
- we must always save the changes made to our project.

Terminology

- Electric assist bicycle – an electric assist bicycle is a bicycle with an integrated electric motor used to assist propulsion.
- Dock - a dock is what holds each individual bicycle. The bicycles are locked into the docks and must be unlocked using the mobile app or a bus card.
- Bike station – a bike station is a structure for secure bicycle parking.

Costs and benefits

The costs are the following:

- time spent on gathering the data,
- time spent on getting valuable information from the data,
- time spent on making clear conclusions from information found.

The benefits could be the following:

- increased usage of environmentally friendly transport alternatives,
- improved health of service users,
- optimised station capacity,
- actionable insights of the client base.

Data-mining goals

Our goals are the following:

- to train models that predict how many bikes there are in a station at a certain time,
- to find out information about user behaviour to enhance customer experience and to improve the marketing program.

Data-mining success criteria

The data mining is considered successful when

- we can report on user behaviour,
- we can make accurate predictions on future user behaviour,
- we can profile potential new users.

Data understanding

Data requirements outline

To address our first data mining goal, we need to know when and where the ride was started and ended, a specific station's maximum capacity and its popularity.
To address our second data mining goal, we need to know about a specific ride:

- the gender of the user,
- the age of the user,
- the length of the ride.

Data availability verification

The data being used for this project is confidential. It was provided to us by the Tartu Smart Bike Share team.

Selection criteria

Data provided to us is in both CSV file format and XLS file format.
We have three CSV files that contain information about rides done between 02/06/2019 and 30/04/2021. All these files have the same format and contain the following information:

- cyclenumber – the ID of the bike that was used for a ride,
- unlockedat – the date of unlocking the bike for a ride (dd/mm/yyyy),
- unlockedattime – the time of unlocking the bike for a ride (hh:mm:ss),
- lockedat – the date of locking the bike after a ride (dd/mm/yyyy),
- lockedattime – the time of locking the bike after a ride (hh:mm:ss),
- startstationserialnumber- the ID of the station where the ride started,
- startstationname – the name of the station where the ride started,
- endstationserialnumber – the ID of the station where the ride ended,
- endstationname – the name of the station where the ride ended,
- length – the length of the ride in kilometres,
- yearOfBirth – the year of birth of the user who made the ride (yyyy),
- first3IdNumber – the first three numbers of the user's social security number.

We also have one XLS file that contains the following information:

- asukoha aadress – the location of the bike station,
- kinnitavate rataste arv – the station's maximum capacity,
- staatus – the status of the bike station,
- paigalduse aasta – the station's year of installation,
- POINT_X – the station's x coordinate,
- POINT_Y – the station's y coordinate.

Irrelevant information for us is possibly the following:

- cyclenumber – we will not be looking at any specific bicycles,
- startstationserialnumber or startstationname – both give us the same information,
- endstationserialnumber or endstationname – both give us the same information,
- yearOfBirth – we can get that information from first3IdNumber,
- staatus – bike stations that are still in development probably won't show up in our findings,
- paigalduse aasta – the station's year of installation is irrelevant to the rides made,
- POINT_X and POINT_Y – the x and y coordinates of the station are irrelevant to the rides made.

Data description

The data was provided to us by the Tartu Smart Bike Share team. They have been collecting data since this service started – we have data from 02/06/2019 to 30/04/2021. About 1.7 million rides have been made since the summer of 2019, each ride containing 12 different pieces of information, such as when the ride took place, how long the ride was, the age and gender of the user, etc.
The data seems to be suitable for analysing which stations are the most/least popular to start/end a ride, which seasons are the most/least popular among the users to use this service, what days of the week do people use this service the most/least, what time of the day is the most/least popular among users to use this service, in which age range is the service the most/least popular, what is the sex of the average customer, etc. Answering these questions (and combinations of them) helps us to tackle our data mining goals.

Data exploration

- CSV files:

Name	Format
cyclenumber	xxxx (where x is a digit)
unlockedat	dd/mm/yyyy (where dd is the day, mm is the month and yyyy is the year)
unlockedattime	hh:mm:ss (where hh is the hour, mm is the minute and ss is the second)
lockedat	dd/mm/yyyy (where dd is the day, mm is the month and yyyy is the year)
lockedattime	hh:mm:ss (where hh is the hour, mm is the minute and ss is the second)
startstationserialnumber	xx (where x is a digit)
startstationname	Given in a string format
endstationserialnumber	xx (where x is a digit)
endstationname	Given in a string format
length	x (where x is a floating-point number)

yearOfBirth	yyyy (where yyyy is the year)
first3IdNumber	gyy (where g is the sex and yy is the year)

- XLS file:

Name	Format
Asukoha aadress	Given in a string format
Kinnitavate rataste arv	x or xx (where x is an integer)
Staatus	Given in a string format
Paigalduse aasta	yyyy (where yyyy is a year)
POINT_X	x (where x is a floating-point number)
POINT_Y	x (where x is a floating-point number)

For more convenient data processing we should convert the user's social security number to the person's age and gender, we should use the end and start times of the ride to calculate the length of the ride, we should use a station's serial number instead of its name, etc.

Data quality verification

The data given to us seems to be suitable for our project. We need to modify some columns for more convenient use and leave out some contradictory data (such as when the last two digits of yearOfBirth doesn't match the last two digits of first3IdNumber), but other than that we have sufficient information to reach our goals.

Project plan

Task	Tools	Hannaliina	Emma Belinda	Kati
Homework 10 – first steps of project	Google Docs, Microsoft Excel	9h	1h	1h
Research about previous work done in that area	Google, research papers	2h	2h	2h
Data	Microsoft	2h	2h	2h

exploration	Excel, Jupyter Notebook			
Data preparation	Microsoft Excel, Jupyter Notebook	5h	0h	5h
Analysis	Jupyter Notebook	5h	0h	5h
Building models using supervised machine learning methods	Jupyter Notebook	0h	15h	0h
Evaluation of built models	Jupyter Notebook	0h	4h	6h
Poster	Google Slides, Canva	6h	1h	4h
Video	iMovie, OBS Studio	1h	5h	5h