# MINI PROJECT CANVAS Title: City Bikes Usage in Helsinki Group members: Quy Anh Nguyen, Aleksi Pikka, and Otto Uusipaikka Workshop # : 4 .

| MOTIVATION 🎯***Which is the target group of our mini-project? Who is the end-user?***Our mini-project aims to provide a data-driven solution for the Helsinki Regional Transport Authority (HSL) to better manage the Helsinki City Bikes system.***What are their objectives? What needs do we need to address with our work? How will they benefit from this proposed solution?***We aim to study the usage of city bikes in Helsinki. It is a rather difficult mathematical problem to estimate the number of city bikes in demand at each station. We aim to provide a prediction for this statistic so that HSL can pre-emptively allocate sufficient bikes and that there is no bike shortage at any station at any time. To study this problem we plan to use time-series methods to study how the demand for city bikes at each station is influenced by various factors, e.g. the weather and time of day. | DATA COLLECTION🧩***Which data sources are we planning to use?***HSL provides free and open data about every journey made by city bikes in Helsinki and Espoo from 2016 to 2023.We downloaded the data from [https://dev.hsl.fi/citybikes/od-trips-[year]/od-trips-[year].zip](https://dev.hsl.fi/citybikes/od-trips-%5Byear%5D/od-trips-%5Byear%5D.zip), where [year] is a value from 2016 to 2023. We extracted the ZIP files to get the CSV files.We also fetch historical weather data in Helsinki using the following free API: <https://open-meteo.com/en/docs/historical-weather-api>. We specifically fetched weather data from April 11th, 2016 to September 12th, 2024.***Which is the data management plan?***The dataset is not huge in size and we do not need to support real-time updates to the data, so there is no need for a database and we will just store a few CSV files in the repository. | PREPROCESSING🛠***What are the goals of the preprocessing pipeline?***The ultimate goal of preprocessing is to combine both traffic data and weather data into one Pandas Dataframe.***Give some examples of data preprocessing steps.***First, we will combine the bike data from all the different CSVs into one DataFrame. This can be done without issue, since the format is consistent across all the CSVs from HSL.Our next task is to convert categorical variables, like bike stations, into numerical variables..Additionally, if there are any missing observations in the data, we will decide on the imputation approach. ***What are some possible data cleaning/wrangling methods you’re planning to use?***Our data is rather clean to begin with, so not much cleaning is needed. Perhaps we might perform label encoding for the station names.***What are some possible data transformations that could be useful?***Logarithm Transformation and Square Root Transformation ***Any feature engineering necessary?***Our dataset is not multi-dimensional, so feature engineering techniques like principal component analysis are probably not required. | EXPLORATORY DATA ANALYSIS (EDA)🔎***Look at the data! What steps are you planning to take towards exploring and understanding better the data you have?***First, we will take a look at high-level descriptive statistics, like the mean and variance of the number of bike trips at different times during the day.We will then attempt to choose an appropriate statistical model. We might transform our data in some way, e.g. by taking logarithms, to get a better understanding of the dataset. After transforming the data, we can perform some simple visualisations, e.g. creating scatter plots and bar charts, to see how the variables correlate to each other. These visualisations might suggest that certain statistical models are more appropriate, and we will proceed accordingly.***What properties would be meaningful to summarise/visualise in this step?***- Average number of bikes in demand at each station- Most frequent destination station (mode)- Average time it takes to travel from one station to another… | VISUALISATIONS📊 ***Are you planning to produce any interactive visualisations?***We do not strictly plan to, but perhaps some simple interactivity is possible. ***If so, which types of interactivity might be useful to the end user?***Line chart is the obvious starting point to find trends in our time-series data. We could perhaps use D3.js to create a line chart where the x-axis is the time, the y-axis is the number of bikes in demand, and the different lines represent different stations. When hovering above a line, we would display a tooltip containing information about our prediction for the number of bikes in demand at that point in time. |
| --- | --- | --- | --- | --- |
| LEARNING TASK 🐭 (focus on problem definition)***Define the problem setting. Is this supervised / unsupervised / other…?***Supervised, as we’re using explicit input variables to predict a correct output***Classification / regression / other…?****C*lassification and regression ***What are we planning to learn? E.g. What is the target variable / learning outcome?***Our dependent variable is the number of bikes in demand at any given time at a specific station.***What variables are we using as input?*** - How many bikes have left a certain station at any given time  - How many bikes are heading towards a certain station at any given time  - Weather status at any given time, like the temperature and humidity level, in addition to whether it’s currently sunny, raining, or something else  - …  Since we’re using the above independent variables to predict one dependent variable, we have a regression problem.  We are also going to classify a station at any given point in time as either lacking bikes, having just enough bikes, or having more than enough bikes, so we have a classification problem as well. | LEARNING APPROACH🐭 (focus on solution implementation)***Which ML/statistical methods seem more relevant for the defined problem setting and why?***We will first employ classic methods to process time-series data like vector autoregressive models. Later on, we will consider applying more complex methods and/or models, if and when necessary.***Which evaluation metrics could be relevant?***How accurate our prediction model is, i.e., how many bikes we predict will be in demand at a specific station versus how many bikes will actually leave the station.***Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?***No special treatment seems to be required. |  | COMMUNICATION OF RESULTS***Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?***- Project Technical Report - Back-end REST API - Front-end User Interface (optional):+ Web Application+ Mobile App (PWA)+ Desktop App***How do we best communicate our results to the predefined target group?******Short description of your interface/workflow (if applicable).***- First, we will analyse the data and build the Machine Learning model locally on our computers using Python.- Second, we will write a technical report for our model.- Third, we might build a REST API backend that provides access to our model using FastAPI. HSL will be able to make use of this API. - For regular travellers, we might build a Graphical User Interface that communicates with the backend for them to try out the model. This GUI can come in the form of: + A Next.js Web Application  + A cross-platform mobile app (PWA)  + An Electron.js cross-platform desktop app | DATA PRIVACY AND ETHICAL CONSIDERATIONS 🔐 (if applicable)***Are there any fairness constraints that apply to our proposed pipeline?***No, there are not.***Is there a need to ask for consent during the data collection process?***No, there is not.***Is there a need for data pseudonymisation/anonymisation?***No, there is not.***Any other privacy considerations that come to mind?***Our data is collected from publicly available sources and has been completely anonymised, so there are no direct privacy concerns. Both the bike data and the weather data are licensed under CC BY 4.0, so we’re free to use and redistribute the data provided that we give City Bike Finland and Open-Meteo appropriate credits. |
|  | ADDED VALUE 🎁***Is there a possibility for added value from the data we’re planning to use?******What is the added value?******How are predictions turned into added value for the end-user?*** | Yes, there is. More revenue from ticket sales for HSL. Our prediction helps HSL optimise the bike supply to different stations, thereby eliminating bike shortages. This increases people's satisfaction with the Helsinki City Bikes system and makes them more likely to pay or more willing to pay a higher price for the service. |  | LEGENDWEEK 1: Data collection/preprocessing  WEEK 2: EDA & visualisations   WEEKS 3-4: Machine/deep learning  WEEK 5: Fairness & data privacy |