Project Proposal

Group:
Pairs-proposal 1

Zahra Moradi
Student number: 2690281

Alexander van der Linden
Student number: 2508637

Research question and Business Understanding

Anyone who has traveled through Amsterdam on a bicycle during rush hour has experienced large amounts of traffic causing big queues at junctions. The municipality of Amsterdam even advises to avoid peak hours if possible[1]. Similarly, cities such as Utrecht[2] and Groningen contain some of the busiest bicycle lanes in the Netherlands[3].

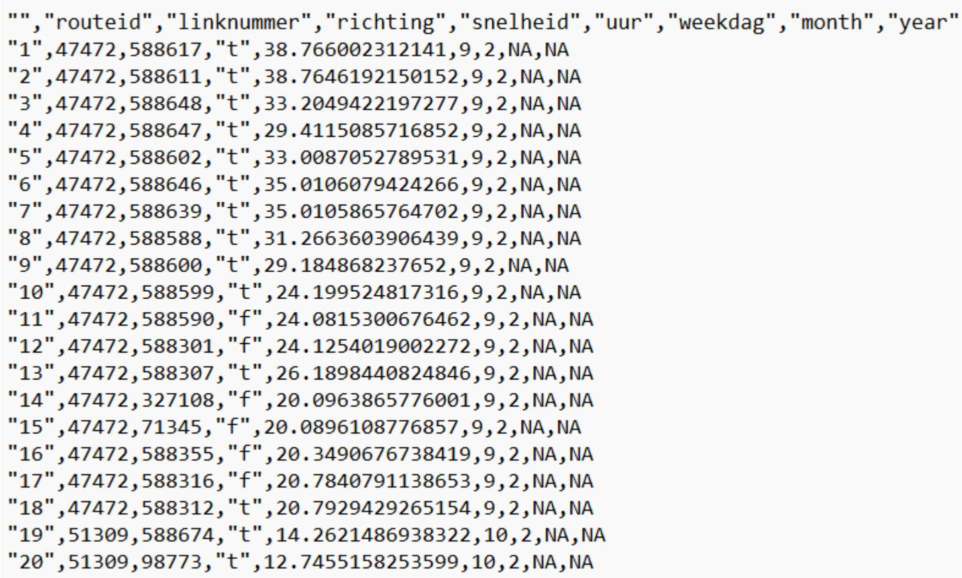
The research question that we will focus on is “How to efficiently predict bicycle traffic density during morning and evening rush hour in the city of Amsterdam”.

To create a potential solution for people traveling on bicycle at peak hours we propose a prediction model for mapping bicycle traffic density. By predicting the traffic density, the model enables travelers to identify potential alternative routes to their destination. As well as allowing municipalities to gain insight in the flow of traffic in their city. The objectives here are to allow for understanding of bicycle traffic flow by identifying bottlenecks in infrastructure as well as reducing high traffic by offering alternative routes for commuters. Close cooperation is needed with city planning of the selected cities by means of exchanging information on bicycle lanes layout and identification of known large and/or problematic traffic junctions. On top of this, data has to be collected on traffic numbers during peak hours. The data to be collected needs to be anonymous to prevent privacy infringement. Furthermore, the focus of this project excludes usability of prediction outside of peak hours, as well as during extraordinary situations that are not covered by the initial data collection (i.e. major events, extreme weather). As a result, the output is dependent on these features.

Data Understanding

To get an accurate picture of bicycle traffic, data needs to be collected on how many bicycles are passing through a large urban environment, with a focus on how many, at what time, and at which location. Subsequently, the velocity and direction of travel are also helpful features in predicting future traffic.

It is optimal to attempt to solve the problem of predicting bicycle traffic by applying data science techniques on existing datasets. For instance, the Fietstelweek (FTW) data for 2015 and 2016 [7] contains bicycle traffic data for the municipalities of Smallingerland and Texel in the Netherlands. However, this dataset is constrained by location and could be considered outdated since it was retrieved around six years ago. We believe that a reasonable solution is to collect new data from the municipality of Amsterdam to be able to conduct a better analysis. This data would ideally follow the structure of the FTW dataset and we would extract this new data from The Trajan dashboard “Amsterdam bikeroutes usage” interactive tool [8]. This tool was created for the The Ping if you Care project and gives insightful information about cyclists in the city of Amsterdam. The Ping if you Care project aimed to obtain data from the cycle count week (FTW) and information about the routes cyclists chose in 2019 [4]. The municipality of Amsterdam used the results from Ping if you Care to identify cyclists' experiences with the infrastructure, speed, air quality and set priorities in their cycling policies and actions [5]. An overview of the structure of the raw data is displayed in Figure 1. The data contains information about the routeid, weekday and the hour of the cyclist activity, and whether they are riding. It is evident from Figure 1 that the data also contains NA values. However, since we aim to predict the traffic during the rush hour, the month and the year of the data is not of importance to our business objective and the last two columns can be deleted when cleaning the dataset. We plan to use predictive models, such as linear regression together with statistics approaches to predict the outcomes for future cycling route traffic. In addition, we will use visual support by means of graphs and plots to study the data and inspect our assumptions. The data collected from the Trajan Dashboard originates from cyclists sharing their experiences and could possibly contain bias. An analysis of the data will need to be performed to look at distribution of geographical location as well as time and the other features to identify this bias.

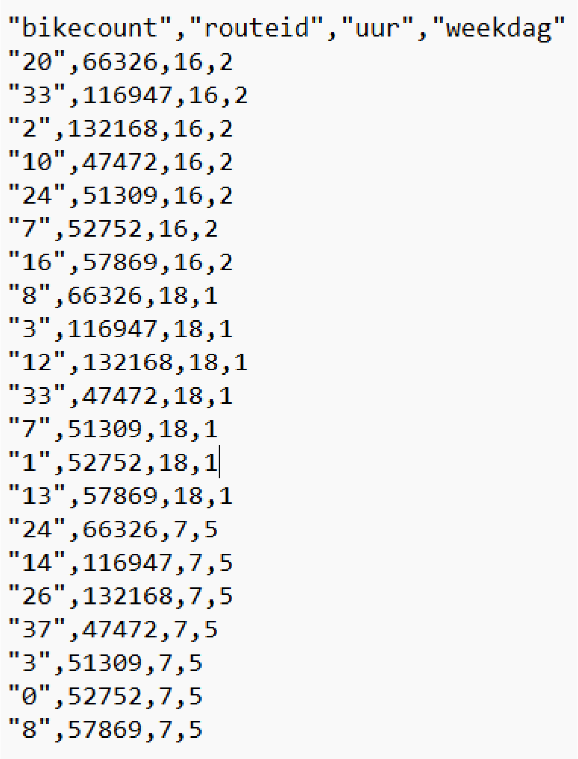


```
"","routeid","linknummer","richting","snelheid","uur","weekdag","month","year"
"1",47472,588617,"t",38.766002312141,9,2,NA,NA
"2",47472,588611,"t",38.7646192150152,9,2,NA,NA
"3",47472,588648,"t",33.2049422197277,9,2,NA,NA
"4",47472,588647,"t",29.4115085716852,9,2,NA,NA
"5",47472,588602,"t",33.0087052789531,9,2,NA,NA
"6",47472,588646,"t",35.0106079424266,9,2,NA,NA
"7",47472,588639,"t",35.0105865764702,9,2,NA,NA
"8",47472,588588,"t",31.2663603906439,9,2,NA,NA
"9",47472,588600,"t",29.184868237652,9,2,NA,NA
"10",47472,588599,"t",24.199524817316,9,2,NA,NA
"11",47472,588590,"f",24.0815300676462,9,2,NA,NA
"12",47472,588301,"f",24.1254019002272,9,2,NA,NA
"13",47472,588307,"t",26.1898440824846,9,2,NA,NA
"14",47472,327108,"f",20.0963865776001,9,2,NA,NA
"15",47472,71345,"f",20.0896108776857,9,2,NA,NA
"16",47472,588355,"f",20.3490676738419,9,2,NA,NA
"17",47472,588316,"f",20.7840791138653,9,2,NA,NA
"18",47472,588312,"t",20.7929429265154,9,2,NA,NA
"19",51309,588674,"t",14.2621486938322,10,2,NA,NA
"20",51309,98773,"t",12.7455158253599,10,2,NA,NA
```

Figure 1: This figure displays the first twenty rows of the FTW dataset for the municipality of Texel in the Netherlands.

Data Preparation

The data contains the time and route id for each bike that is registered but we are only interested in analyzing the traffic during the rush hours. For this purpose, we only need to consider rows whose “uur” value is in the range of [6:30,9] and [16,18:30] as well as only keep the “weekdag” (weekday) that refers to Monday till Friday. However, the data has information at every hour interval and does not take into account half-hours. Hence, we decided to expand the ranges to [6,9] and [16,19] to ensure that we do not miss important information regarding the rush hours as well as making the cleaning of the data easier. In addition, we are interested to know how many bikes are riding in a specific route in the rush hours. Therefore, we will need to group the data by their “routeid” and consider only the values that have “t” for the “richting” column to calculate the traffic. It is also important to remember that our data also contains missing data which are the “NA” values, and these values need to be deleted from the dataset in order to proceed. These issues indicate that the data indeed needs cleaning and structuring before it can be used in our prediction model. To ensure the data quality the desired structure of the data should resemble Figure 2 and give more insight about our findings.



```
"bikecount","routeid","uur","weekdag"
"20",66326,16,2
"33",116947,16,2
"2",132168,16,2
"10",47472,16,2
"24",51309,16,2
"7",52752,16,2
"16",57869,16,2
"8",66326,18,1
"3",116947,18,1
"12",132168,18,1
"33",47472,18,1
"7",51309,18,1
"1",52752,18,1
"13",57869,18,1
"24",66326,7,5
"14",116947,7,5
"26",132168,7,5
"37",47472,7,5
"3",51309,7,5
"0",52752,7,5
"8",57869,7,5
```

Figure 2: This figure displays an example of how the data would look like after taking proper data preparation steps.

The road network in Amsterdam contains almost 5.000 streets[12] and as such the dataset would contain a huge amount of datalines. One approach is to use python together with the pandas library to create the desired dataframe out of our original data. This takes more time, but it prevents third-party application mistakes (costs accompanied by our choices). Considering the potential size of the dataset, another approach to consider is using more professional software such as Microsoft Power BI to streamline the process of cleaning and transforming the data [9].

Furthermore, we will use Microsoft Power BI to help visualize the final data to detect any inconsistencies, biases, and identify the outliers. This is an important step in data preparation and it ensures the reliability of the data (the kind of tooling we use).

Lastly, to adhere to the protocols of the General Data Protection Regulation we need to make sure to delete the location data of each individual bike since it can be considered a privacy violation. This is also the reason that the final data structure only contains “bikecount” in a route and not the “linknummer” of all the bikes as the original data (adjustments to be made in the data for legal or ethical reasons).

Modelling

We will be using the ‘bikecount’ value as our dependent variable which we want to predict, and the values of ‘hour’, ‘weekday’, and ‘routeid’ as our independent variables since they are all key factors in determining the traffic. In addition, due to our dataset of labeled data (we have clear output of the traffic estimate for each input of our independent variables), we will determine a supervised machine learning algorithm to model our framework. There exists several models that we can proceed with such as Classification, Regression and Ensemble models. However, we have opted for a regression model specifically since our values are numeric, and categorizing our large data set is not time efficient. Ensemble models (e.g. decision trees) or Convolutional Neural Network (CNN) could also be considered to predict the traffic estimate. However, training CNNs and decision trees is expensive in terms of computational power and time when the input dataset is very large [10]. Furthermore, decision trees lack accuracy when dealing with continuous values and their predictions need

to be divided into discrete categories which could cause a loss of information for continuous values [11].
As a result, we will use a regression predictive model to estimate future traffic density. The desired outcome for our model is to be able to output accurate predictions about the number of bicycles at every specified location based on a given time and day. Due to multiple independent variables determining the amount of bicycles at any time, a Multiple Linear Regression model (MLR) will be used. With the 'hour', 'weekday', and 'routeid' as features and the 'bikecount' as label, the model can be trained to predict an estimate of 'bikecount' for unlabeled data. The outliers in the dataset should not affect the performance of the model too much. Instead of using the mean squared error as a loss function, the mean absolute error (MAE) will be used to prevent amplifying the larger errors made by outliers in the regression model.

Evaluation
In order to be able to evaluate the performance of the regression model, the dataset will need to be split into a training and testing set. Eighty percent of the data will be used for training the model. The remaining twenty percent will be used for testing the model when it is done training. The performance can then be analyzed through the use of mean absolute error, which gives a scoring value. The closer this value is to zero the better and as a criteria should not exceed a range of ten percent from the true value. By resampling the data multiple times and repeating the above process, we can start to analyze the bias and variance that the model has based on multiple performances. The goal here would be to see how well the model fits the distribution of the dataset and identify overfitting with high bias or underfitting with high variance. To see if the business criteria is met, an extra tool needs to be developed to visualize the model's output. With an interactive map of the bicycle network, the predicted number of bicycles for every 'routeid' can be plotted on the matching bicycle roads. The experienced usability of this interactive tool determines the success for further use by travelers and the municipality. To see how well the developed tool performs, a testing phase will be implemented with a group of approximately one hundred people to get feedback on how they experience the tool. The focus lies on people participating in rush hour commutes with a bicycle.

Deployment
Once the testing phase returns positive results and no further adjustments are necessary to both the model and the interactive tool, the product can be launched in an online environment available to the public. By hosting the tool on a website it allows for easy access for anyone with an internet connection. In deployment to the municipality, the observed performance metrics for the prediction model will be presented in a report. Including visualization of the prediction with a scatterplot containing the sample data and the prediction of the regression model. In addition, we will present five plots that focus on the average bicycle traffic of the routes during the rush hour on different days of the week (excluding the weekends). Figure 3 displays the ideal plot that will be presented to the municipality of Amsterdam based on the model's predictions for each day. The underlying data of these plots will further be used when the users are suggested with different routes in the interactive tool on the website. Subsequently, the collected data on potential bias and variance that are present need to be detailed upon. When considering the implications of the results that the end product gives, an unknown factor is the level of accuracy the model has compared to the real life situation. Based on this true performance the usability of the model can be determined. When the model's performance is mediocre, the effect of this could at best be giving bad travel advice to someone looking to evade dense traffic. Much worse could be if the municipality makes structural decisions, thus altering the public space, based on poor model prediction. Thereby negatively impacting the problem that the model sets out to alleviate. However, with strong and accurate performance, the opposite might be true.

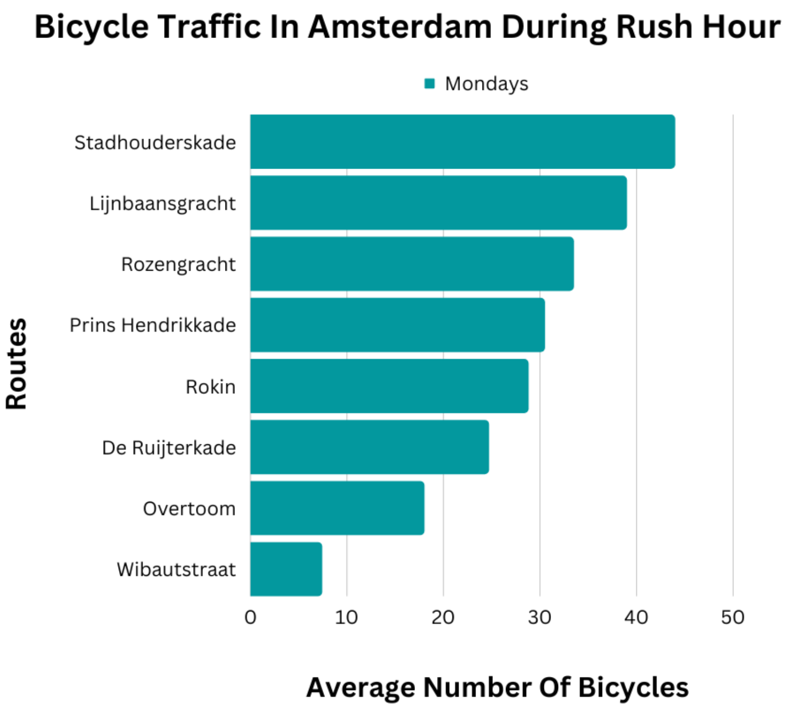


Figure 3: This figure displays an example of how the plot would look like for each day of the week.

References

1. I amsterdam. (2019, August 28). *Cycling safely in Amsterdam*. iamsterdam. Retrieved 22 September 2022, from https://www.iamsterdam.com/en/plan-your-trip/getting-around/cycling/cycling-safely#:%7E:text=Avoid%20rush%20hour%3A%20between%2008,until%20the%20rush%20calms%20down
2. Buiting, J. (2020, April 29). *Dit zijn de drie drukste fietspaden van Nederland*. De Openbare Ruimte. Retrieved 22 September 2022, from https://deopenbareruimte.nu/dit-zijn-de-drie-drukste-fietspaden-van-nederland/
3. Es, M. van, & Slütter, M. (2019, April 9). Hoeveel wordt er gefietst in Nederland? Alle cijfers op een rijtje. Fietsersbond. Retrieved September 23, 2022, from https://www.fietsersbond.nl/ons-werk/mobiliteit/fietsen-cijfers/
4. Amsterdam Bike City. (2021, May 12). *Amsterdam Bike City | Experiment: Ping if you Care Amsterdam 2019*. Retrieved 22 September 2022, from https://bikecity.amsterdam.nl/en/inspiration/ping-if-you-care-2019/
5. Franchoise, E. (2019). *PING Amsterdam*. PING if You Care! Retrieved 22 September 2022, from https://pingifyoucare.eu/amsterdam/
6. Den Haag verbetert doorstroom fietsers met applicatie FietsViewer. (2019, December 2). VNO-NCW. Retrieved September 23, 2022, from https://www.vno-ncw.nl/weekbulletin/den-haag-verbetert-doorstroom-fietsers-met-applicatie-fietsviewer
7. Crespo, L. A. (2019, February 19). *Fietstelweek Data for 2015 and 2016*. GitHub. Retrieved 23 September 2022, from https://github.com/loreabad6/ftw
8. Amsterdam Bike City. (2021a, May 12). *A*. TrajanDashboard. Retrieved 26 September 2022, from https://trajan-dashboard.maphive.net/permalink/@trajan?lng=en#00000000000000000000999999900001.bc9d6567f3b24bba8c6dbc81751ae095
9. Rad, R. (2021, June 2). *How to Reduce the Size of Power BI file in a few Steps*. RADACAD. Retrieved 27 September 2022, from https://radacad.com/how-to-reduce-the-size-of-power-bi-file-in-a-few-steps#:%7E:text=Steps%20to%20reduce%20the%20size%201%20Step%201%3A,Date%2FTime%20. . .%203%20Step%203%3A%20Remove%20un-used%20columns
10. GeeksforGeeks. (2022, August 22). *Decision Tree*. Retrieved 27 September 2022, from https://www.geeksforgeeks.org/decision-tree/#:%7E:text=Decision%20Tree%20%3A%20Decision%20tree%20is%20the%20most,leaf%20node%20%28terminal%20node%29%20holds%20a%20class%20label.
11. Editorial. (2020, July 30). *When to consider Decision Tree Algorithm - Pros and Cons*. RoboticsBiz. Retrieved 27 September 2022, from https://roboticsbiz.com/when-to-consider-decision-tree-algorithm-pros-and-cons/#:%7E:text=Disadvantages%20of%20Decision%20Tree%201%20Overfitting%3A%20A%20common,that%20it%20calls%20for%20heavy%20feature%20engineering.%20