Project Bike-Sharing Chicago 2018

Group: Team 10

Subject: Analytics and Applications

University: University of Cologne

Instructors:
Prof. Dr. Wolfgang Ketter
Nastaran Naseri

Members:
Ercan Tazegül
Simon Knecht
Diego Longhitano
Mathias Werwie
Vincent Sedlacek

Date: 31.01.2023

Here you can find our Git repository: Github Link

Contents

List of Tables

List of Figures

1 Summary

The objective of this project is to study the 2018 Divvy Bikes Chicago bike ride dataset, which comprises two datasets: one containing data on Chicago bike rentals in 2018, and the other containing hourly weather data for 2018 obtained through the weather.com API. In order to understand and optimize the performance of the bike fleet, we have defined key performance indicators (KPIs) and analyzed the datasets for temporal and spatial demand patterns. Cluster analysis was used to identify recurring patterns and inform business decision-making. Furthermore, we have applied predictive analysis techniques, such as scientific forecasting models, to forecast future demand and optimize operations. The important results are as follows: The descriptive analysis shows that the number of trips in the summer months from June to September is about 450,000 trips. In addition, the number of trips in the winter months of January, February, March, November and December decreases significantly and is about 130,000 trips. Also, it can be noted that there is a high amount of borrowing activity on weekdays between 7 am and 8 am and between 4 pm and 6 pm. This changes on weekends, where the greatest activity occurs between 11 am and 5 pm. In the Geographic Demand Pattern, it was found that the Streeter Dr and Grand Ave station is the most popular station, likely because it is located near Ohio Street Beach, a popular tourist destination. Therefore, this station is also popular only during the summer months. The KPI "average available bikes for hours of a day" also notes that occupancy is highest at 5 pm. It was also found that on average, about 20 percent customers and 80 percent subscribers use bicycles annually. Cluster analysis showed that most stations are located in downtown Chicago and near train stations. It can be seen that many people use bicycles for a short distance home after taking the train. Likewise, the cluster analysis shows that the peak hours are 8 am and 5 pm. In the prognostic analysis, we were able to evaluate four different models. The results of the four algorithms : Ridge Regression, Decision Tree, Random Forest and Nearest Neighbor showed that all algorithms performed well on the metrics. The Random Forest algorithm is the best for our dataset and can be used for good demand prediction. The Nearest Neighbor algorithm is the worst for our dataset. Based on the analysis results, the company can optimize their marketing strategy by targeting different user segments such as tourists, customer and subscribers. In addition, it may be beneficial to introduce a trial period to increase the likelihood of converting casual users into subscribers. Considering the lower user numbers in the winter months, the company could also suggest reducing the availability of bicycles as a cost-saving measure.

2 Problem Description

Transport-related greenhouse gas emissions account for a large share of total emissions in the EU, and it is widely recognized that our approach to mobility needs to change in order to achieve our decarbonization goals (Umweltbundesamt, 2022). Traditional urban mobility is mainly based on internal combustion engine vehicles, which have four negative impacts: Contribution to global greenhouse gas emissions, pollution with serious health risks for urban populations, high accident rate with nearly 1.3 million fatal accidents annually worldwide, and inefficient use of motor vehicles with low occupancy and high space requirements for roads and parking, and traffic congestion (Statistisches Bundesamt, 2022). The need for a major transformation of the mobility system has been recognized, and the mobility landscape is changing rapidly, with the important trend of Mobility-as-a-Service (MaaS) and On-Demand (MoD), as well as the use of bikesharing platforms and similar platforms for other modes such as cars, mopeds, and e-scooters. "Faster than walking, cheaper than rideshare, and more fun than the train.". That is the tagline of DivyBikes, a fleet rental company in Chicago. In this project, we explore how DivyBikes can leverage increasingly ubiquitous real-time data streams to monitor and optimize their fleet operations, increase profitability, and improve service levels. Here we focus on system monitoring to understand the operational performance of the fleet as well as demand prediction to predict future demand. Thus, the provider can improve its existing rental service.

3 Data Collection and Preparation

3.1 Bikesharing Data

The dataset contains bike sharing data from Divvy Bikes Chicago from 2018. The overview of the variables in the dataset can be seen in Table 1.

Table 1: Description of bikeshare dataset columns

Variable name	**Format**	**Description**
start time	datetime	Day and time trip started
end time	datetime	Day and time trip ended
start station id	int	Unique ID of station where trip originated
end station id	int	Unique ID of station where trip terminated
start station name	str	Name of station where trip originated
end station name	str	Name of station where trip terminated
bike id	int	Unique ID attached to each bike
user type	User	membership type

The data preparation process to construct and clean the data set includes multiple steps. Started by removing the duplicates to avoid redundant data, continued by dropping null values and also checking for consistency. With consistency we look that putting the data in a context makes sense. In this case we compared the starttime attribute with the endtime attribute and dropped every row in which the starttime is greater or equal to the endtime. Next, we ensured that every bike trip with the unique bikeid is happening just for once at the same time. We also created two new columns, one of them displays the trip time in hours and the other the difference between endtime and starttime. Lastly, we set an upper limit for the duration time to drop further outliers. Every bike trip with a length not longer than 10 hours will be kept. Calculating the 0.999 quantile gave the best restriction of the data set. This assumption we take as the most reasonable time range and finish the data preparation process by exporting the cleaned dataset for further processing.

3.2 Geological Data

Geolocation data was imported from the Chicago website to create location-based analytics and heat maps

3.3 Wheater Data

The weather data is provided by the wheater.com API and contains the variables listed in Table 2.

Table 2: Description of weather dataset columns

Variable name	**Format**	**Description**
date time	datetime	Day and time of measurement
max temp	float	Maximum temperature recorded in degC
min temp	float	Minimum temperature recorded in degC
precip	int	Binary indicator for precipitation (1=yes,0=no)

In order to improve the quality of the weather data, we removed all rows containing NaN values. The hottest and coolest temperatures recorded in the dataset appeared to be reasonable, so no further removal was necessary. We then identified 1328 duplicates in the dataset and decided to retain only the last recorded entry for duplicates, as this is generally considered the most reliable in such situations. There were also several rows with data for the same time, so we chose to take the average and remove the duplicates. The earliest recorded date in 2018 was January 1 at midnight, while the latest was December 31 at 11pm. During this time period, 623 hours of data were missing, which is almost 26 days. Ultimately, we decided to estimate the missing data and found that there are missing data every month, distributed throughout the year. There are only a few sequences that are longer than 1, with a maximum length of 6. In the worst case, we therefore do not have data for a period of 6 hours. Taking into account the above arguments, we have decided that it should be possible to estimate the weather for the missing data without making overly inaccurate estimates.

4 Descriptive Analysis

4.1 Temporal Demand Patterns and Seasonality

Demand is highest in the summer months and lowest in the winter and autumn months (Figure 5). Especially in July, where the average demand is the highest which could be due to the warm weather (Figure 5). Between 6 a.m. till 8 am and 4 pm till 5 pm, the demand of renting bikes starts to increase (Figure 3, Figure 4). Especially at 5 p.m. which is the rush hour and therefore the highest hourly peak demand on average (Figure 4). The preference to rent a bike is greater within the week instead of weekends (Figure 1). Within the week, Wednesday is the day where most bike trips are made. Between Friday to Sunday, the number of trips decreases (Figure 1). The important aspect is that the bike rental system in Chicago is popular for users within the week and are used at times to get to and from work/school.

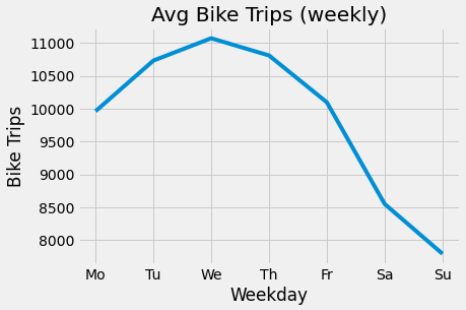

Figure 1: Average Bike Trips weekly

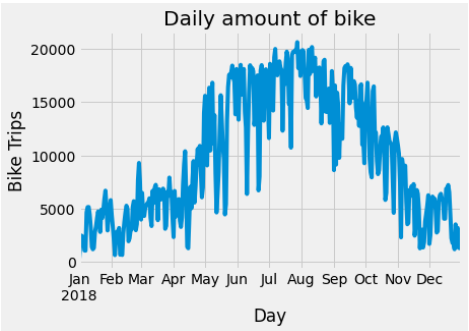

Figure 2: Daily Amount of Bike Trips

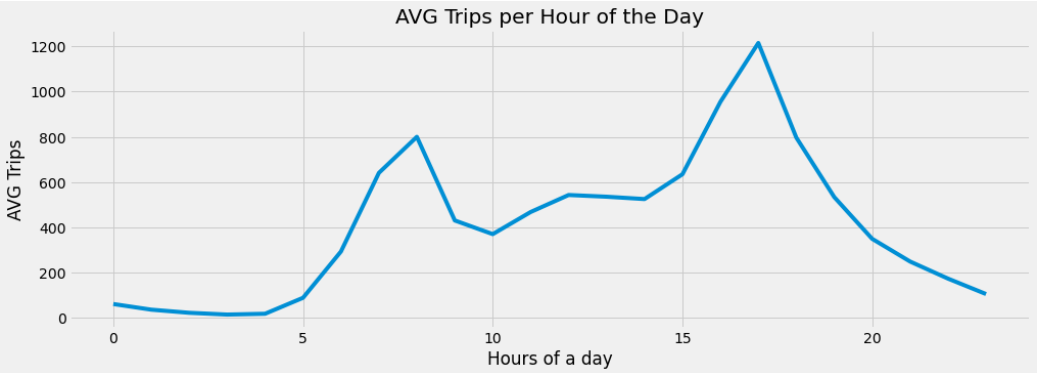

Figure 3: Average Trips per hour of the Day

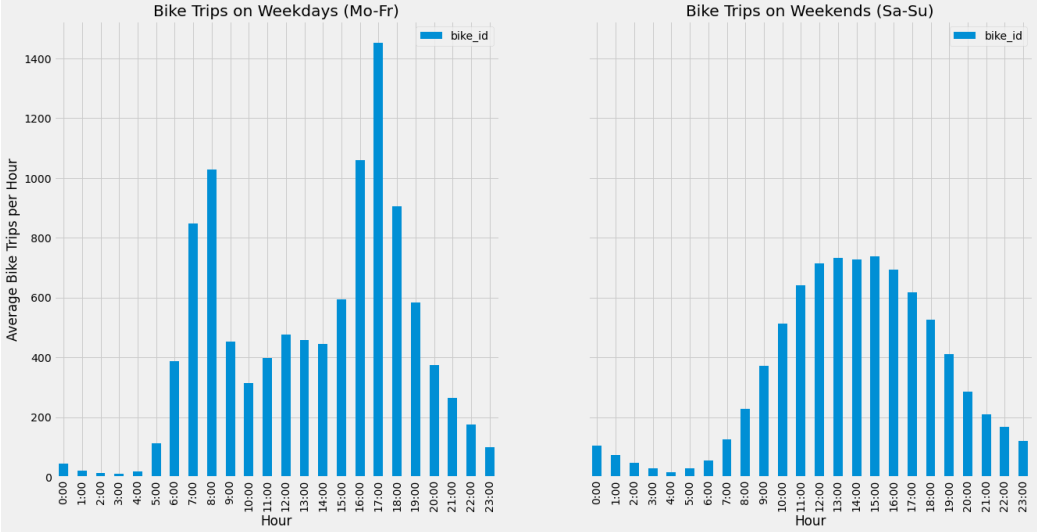

Figure 4: Bike Trips on Weekdays and Bike Trips on Weekends

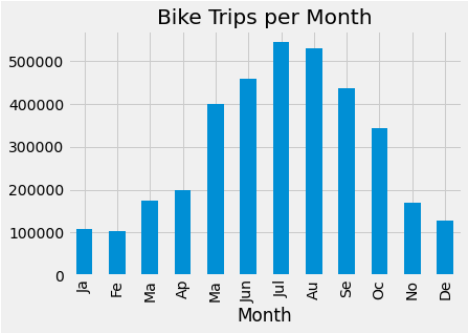

Figure 5: Bike Trips per Month

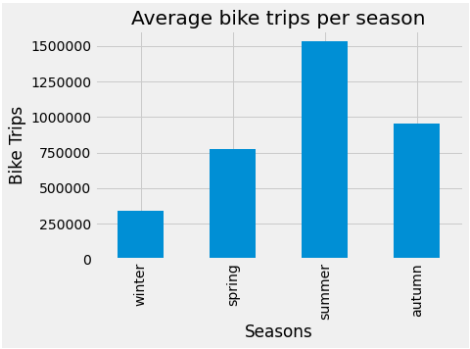

Figure 6: Average Bike Trips per season

4.2 Geographical Demand Patterns

Station popularity, defined as the count of bike rides involving the particular station, depends on multiple factors including the season, the geographic position of the station and weather features. The most popular stations throughout the whole year, which are depicted in (Fig. 07), include 'Streeter Dr and Grand Ave' (1) and 'Canal St and Adams St' (2), both being critical stations to the bike sharing network. While station (1) is located near the ohio street beach, providing access to one of the most popular destinations throughout the summer, station (2) is adjacent to the Chicago Union Station connecting commuters to the entire city. As we can see in (Fig. 08) most of the top popular stations are located near lake Michigan towards the center of the city. While this is not the geographic center of Chicago it coincides with the most high traffic street in Chicago, being 'Michigan Ave and Washington St'. Generally, a higher distance from the center correlates with a lower popularity see (Fig. 11). The least popular stations are all located towards the edge of the network's scope as illustrated in (Fig 09). A further factor influencing the popularity of a station is the weather. While station (1) is highly popular in the summer it is barely used in winter months. This relationship is further visualized in (Fig. 10).

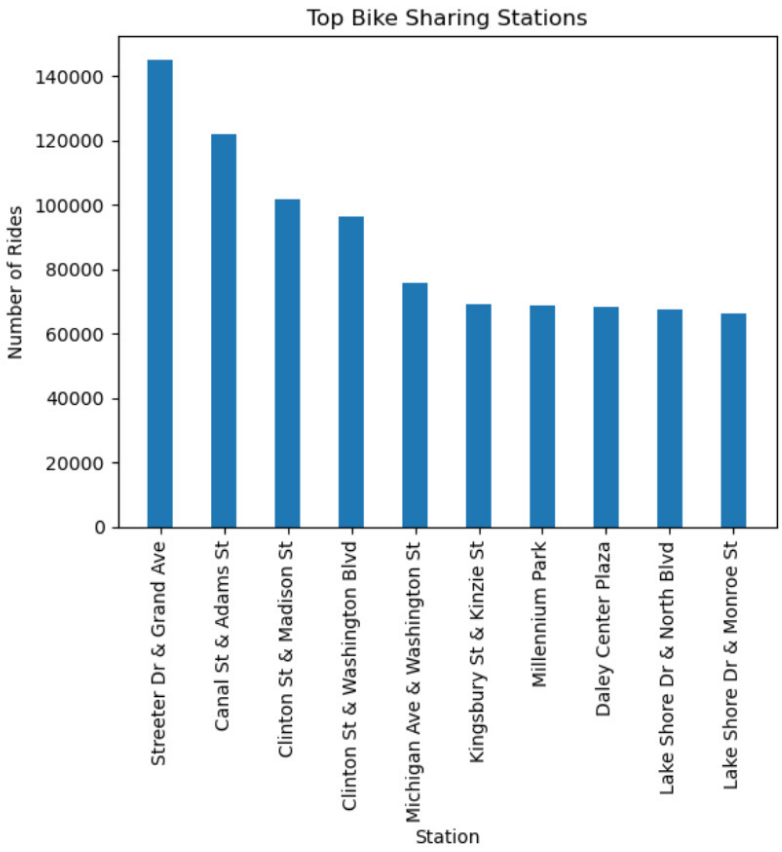

Figure 7: Top Bike Sharing Stations throughout the year

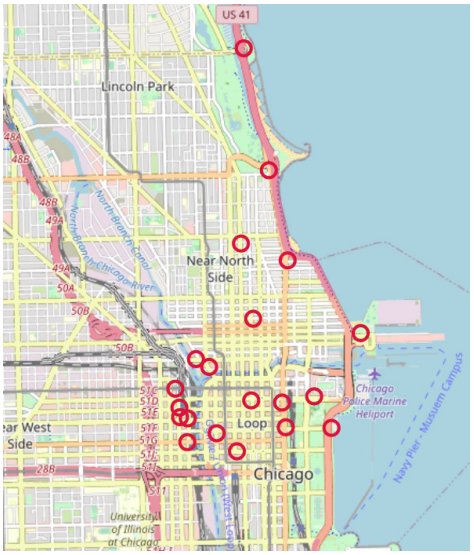

Figure 8: Location of most popular stations

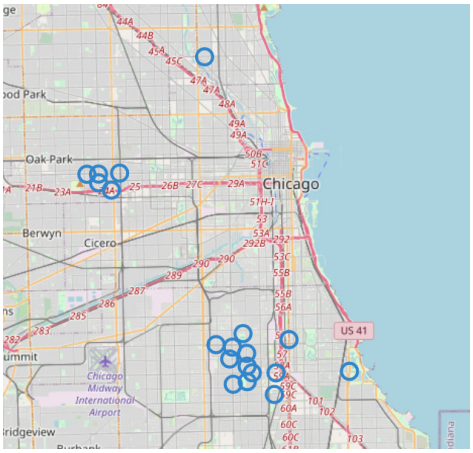

Figure 9: Location of least popular stations

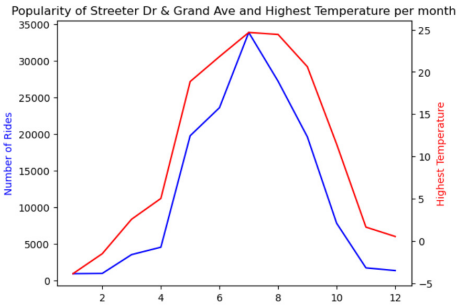

Figure 10: Popularity and Highest Temperature per month

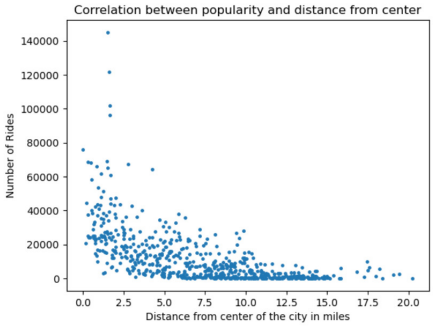

Figure 11: Popularity and distance from center

4.3 Key Performance Indicators (KPIs)

The utilization of key performance indicators (KPIs) has been implemented in order to conduct a detailed analysis of the bikeshare business results. These KPIs have been specifically designed to identify crucial service indicators, providing bikeshare providers with a comprehensive overview of the business and enabling informed decision-making for future endeavors.

4.3.1 KPI: Average available bikes for hours of a day

To cover the demand of bikes, we look at the utilization of bike rental patterns evolves throughout the day. We look at the number of available bikes per hour to avoid potential bottlenecks and thus have an indicator that represents the peak times of bikes during a day on average. This KPI represents as a fundamental base the temporal utilization and is again concretized with the building up KPI's, at which localities at which time the demand is highest, so that DivyBikes Chicago receives an overview of the utilization for the day for different locations. From midnight to 5 am bikes are available in large quantities. After 6 am to 8 am, many bikes are used, so the availability is quite low. However, between 9 am to 3 pm more bicycles are available. The rush hour is at 5 pm where the number of available bikes is the lowest and starts to increase from 6 pm to 11 pm.

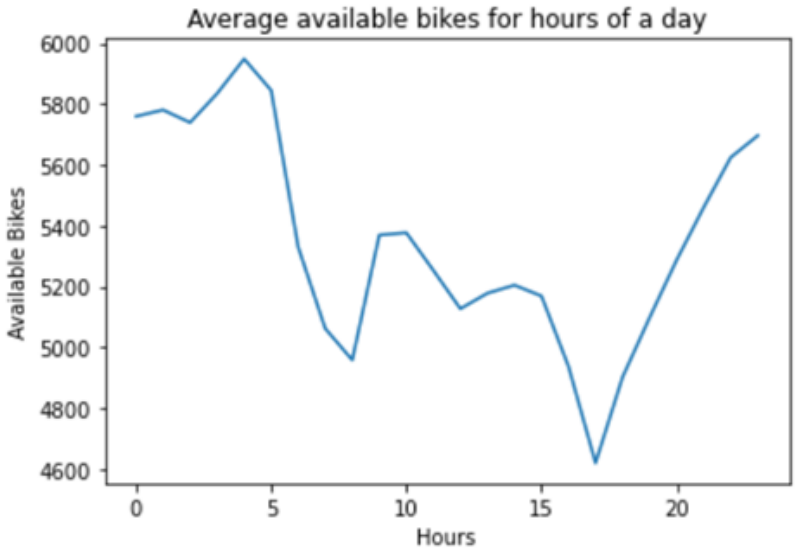

Figure 12: Average available bikes for hours of a day

4.3.2 KPI: Bike Availability per Station

Building on top of the previous KPI, we define the bike availability per station to be the count of available bikes positioned at a particular station. While our previous analysis determined when and where demand spikes occur, this KPI focuses on how to supply the given demand such that the distribution of bikes across the network remains balanced. More precisely, it allows one to determine which stations have excess bikes available at a given time in order to use them as a source to supply a given demand at a later time. Furthermore, the KPI illustrates the distribution of bikes and their activity, allowing an immediate overview of the performance of the fleet. The key findings regarding this KPI are as follows: in order to fuel the high demand at the station 'Canal St and Adams St' during morning hours (see Fig. 13 and 14) bikes can be sourced from the stations 'Lake Shore' and 'Streeter Dr and Grand Ave' (see Fig. 15) during summer months and from the station 'Millenium Park' (see Fig. 16) during winter months.

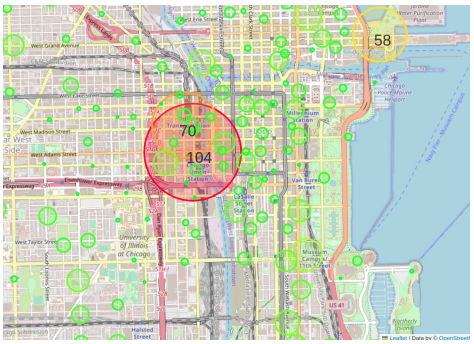

Figure 13: Summer Wednesday 7am

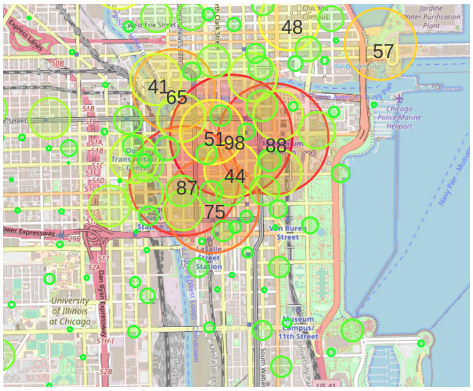

Figure 14: Summer Wednesday 9am

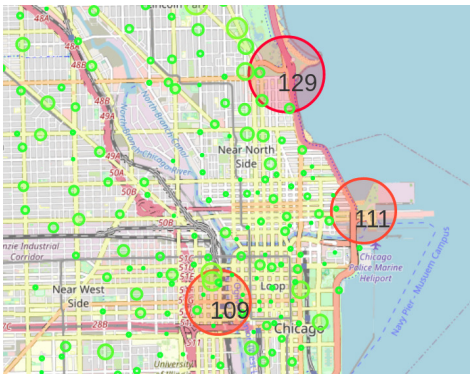

Figure 15: Summer Wednesday 7pm

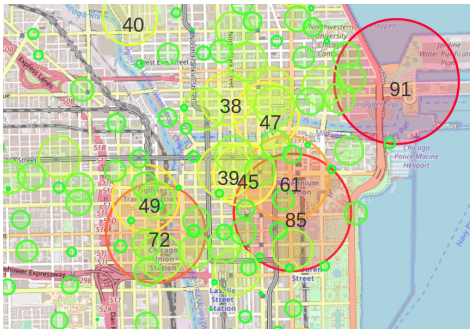

Figure 16: Winter Saturday 11pm

4.3.3 KPI: Utilization of user types Customer and Subscriber

Key performance indicators (KPIs) were introduced for the user types customers and subscribers to analyze the usage behavior of these groups. The KPIs provide information on the hourly, monthly, weekday, and general average usage of these user types. Analysis of these KPIs shows variations in usage among customers, with a higher usage rate among subscribers in the morning and after hours, as well as during the winter months. Specifically, it is noted that usage by subscribers during these periods is significantly higher than that of customers. In summary, the majority of subscribers use the Bikeshare service 80 percent of the time, while only a minority of customers do so 20 percent of the time.

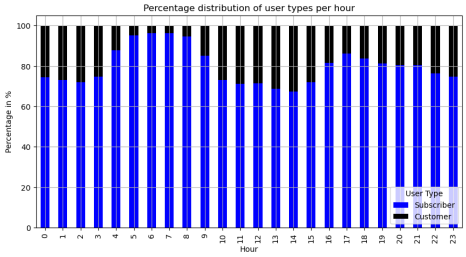

Figure 17: Percentage distribution of user types per hour

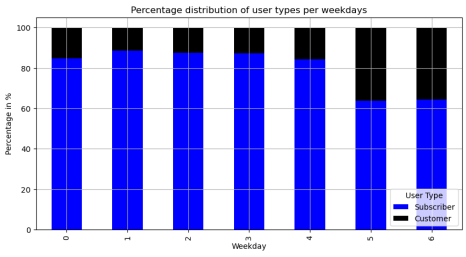

Figure 18: Percentage distribution of user types per weekdays

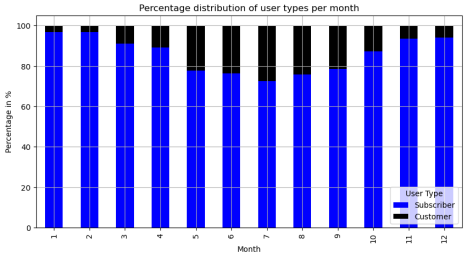

Figure 19: Percentage distribution of user types per month

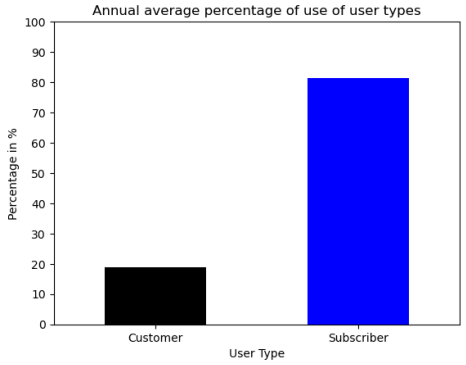

Figure 20: Average percentage of use of user types

5 Cluster Analysis

To identify clusters of trip and customer types, as well as for stations, we perform a cluster analysis based on the bike rental demand patterns. We decided to use the k-means++ algorithm, as it has an efficient computation and is simple to implement. For every investigation we chose varying fitting features to get the best cluster results.

5.1 Trip Clustering

In this analysis we want to find clusters of trip and customer types. Therefore we scan the trip-dataframe based on the features: duration, user-type, hour and day of week. Six cluster can be observed:

Table 3: Description Trip Clustering

Cluster	Cluster Description
1.	Employees bicycle to work in the morning at weekdays.
2.	Medium long trips of non-subscribers with tendency on weekends.
3.	Short trips in the afternoon at weekdays of people going home or out for the evening.
4.	Short trips in the night of people returning home after going out.
5.	Short trips in the daytime on mostly Friday to Sunday, people make trips in their free time.
6.	Long trips of non-subscribers with tendency on weekends, small distribution.

5.2 Weather Clustering

To take the weather with temperature and precipitation into account, we group the trips to their hour and cluster based on the features: temperature, precipitation, hourly demand, day of week, hour and quarter of year. Seven clusters can be observed.

Table 4: Description Weather Clustering

Cluster	Cluster Description
1.	A lot of people want to bike in the afternoon because of good and warm weather, to get back home after work or make a trip in their free time.
2.	Many people want to bike in the morning because of good and warm weather, to get to work. Like cluster 1, trips are primarily in the second and third quarter of the year.
3.	Trips in the last quarter of the year, low hourly demand due to colder temperatures.
4.	Trips at daytime primarily in the first quarter of the year, low hourly demand due to cold temperatures.
5.	Trips at nighttime primarily in the first quarter of the year, very low hourly demand due to cold temperatures and late hours.
6.	Trips at nighttime primarily in the second and third quarter of the year, low hourly demand due to late hours but compared to cluster 5 more trips due to warm temperatures.
7.	This cluster contains all trips at hours with precipitation, not many people want to bike when it rains.

5.3 Station Clustering

For the third analysis we want to check if there exist clusters for the stations regarding homes and workplaces. For this we calculate the trip demand for every station and every hour. We don't consider weekends because these don't include trips to and from work. Two clusters can be observed.

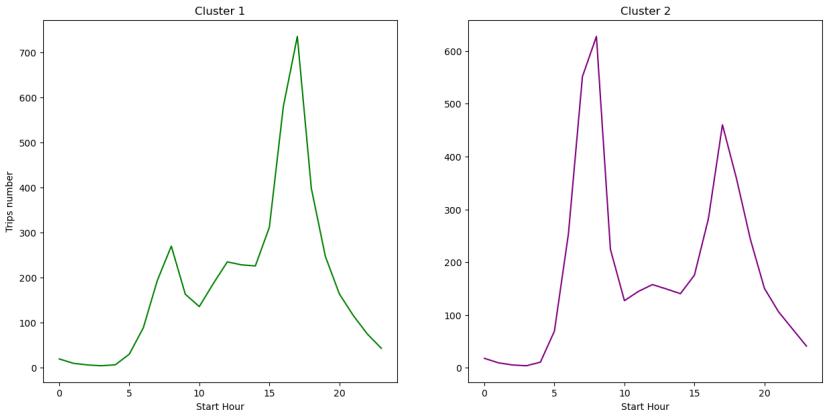

Figure 21: Trips number vs start hour

Cluster 1: the majority of the trips start at around 17 hours. Most of the stations are in the center of Chicago, especially in an area called Loop, where a lot of shops and offices are located. People rent bicycles to get home, get to the next train station if they live further away, or make a trip for going out in the evening.

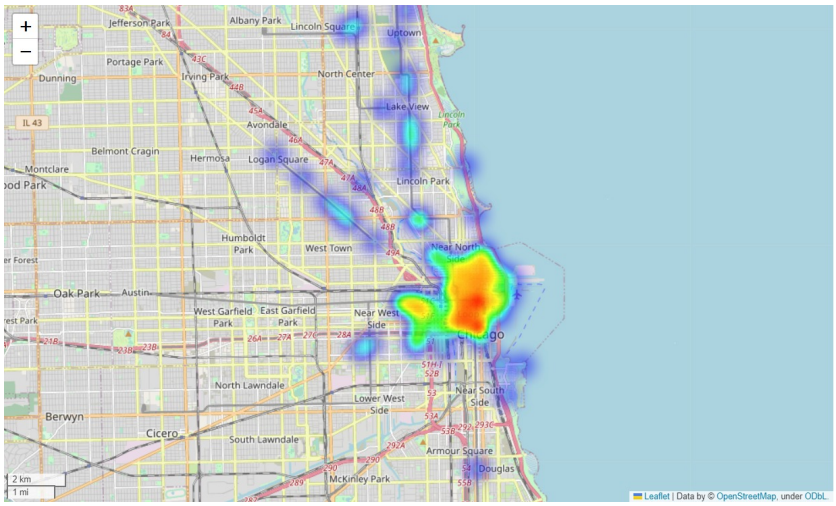

Figure 22: Station location of cluster 1 at 17 hours

Cluster 2: most of the trips start at around 8 hours, which another smaller peak at 17 hours. This cluster contains more stations outside the center where more people live. The areas where most of the stations are located are around train stations. People rent a bike after taking a train to either get to work in the morning or in the afternoon bicycle home or do evening activities.

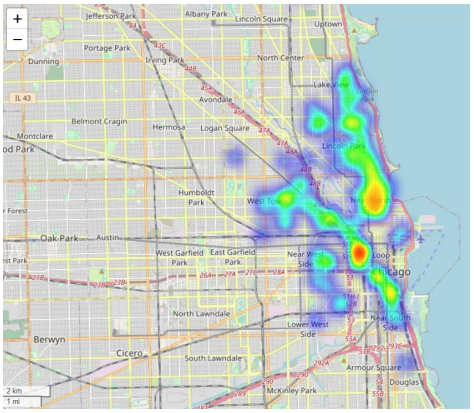

Figure 23: Station location of cluster 2 at 8 hours

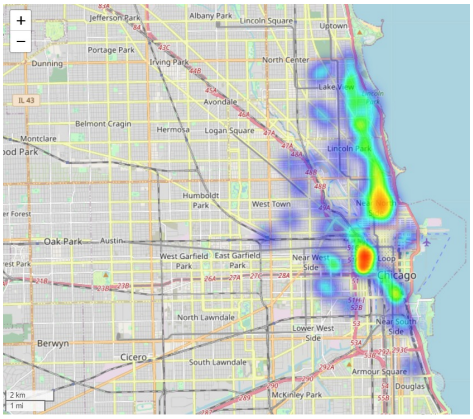

Figure 24: Station location of cluster 2 at 17 hours

6 Predictive Analysis

The reliable prediction of future demand can be a powerful tool for managing a bike-sharing network. It allows one to anticipate high demand spikes and detect low traffic hours in which operational tasks can be conducted. During our analysis, we aimed to predict total system-level demand in the next hour, which we defined as the count of bike rides undertaken in the next hour. Before doing any type of prediction, we needed to start with Feature Engineering. This process describes the engineering of multiple features, which we assumed, could be helpful for our prediction models.
First, we took a look at the columns we already had, coming from our cleaned data sets and thought about what we could be done with them. Out of this inital data, we managed to engineer 15 possible features and, of course our label the hourly demand. Those features were as follows: precip, avgtemp, starthour, sinhour, coshour, startmonth, sinmonth, cosmonth, isHoliday, dayofWeek, sindayofWeek, cosdayofWeek, season, inrushHour and prevDemand.
We had a couple of cyclical features, like the start hour of day of Week, which needed transformation. Since, our prediction models would not know that those feature are of cyclical nature, it would just assume that, for example, the days 1 and 7 were vastly apart. Thus, by utilizing sinus and cosinus functions, we made created a more continuous and thus fitting representation of this data. Now that we had all possible features together, we had to figure out which of them were actually of benefit for our prediction. This narrowing of features, would allow us to avoid overfitting and optimize the performance of our chosen models. Our preferred methods for this purpose were one, a Lasso

(L1) regression, two a correlation matrix. We scaled our features and executed the Lasso regression, which quickly tuned its hyperparamter λ to 0.5. Then, it returned us each feature's coefficient, representing its influence and thus importance on the model. With this we could already perform an intial feature selection, resulting in the following feature set: coshour, sinmonth, avgtemp, inrushHour, prevDemand. With our inital feature set now determined, we want right into model building. In order to make sure that our models produce optimal results, we need to do things before actually applying them to our data. First, we needed to split our data into a training and testing section, allowing cross-validation later on. And second, we needed to tune our hyperparameters. For finding the optimal polynomial degree for our ridge regression later, we utilized a polynomial regression, looping through multiple degrees and calculating the respective MSEs. With this technique, we determined that our initial optimal polynomial degree lies at four. Next, we needed to find the optimal λ. Similarly to the discovery of the polynomial degree, we looped through various values for λ and determined the respective MSEs. After more thorough investigation of the λ from 10 to 16, we found the optimal λ to be at 13. Hence, we successfully tuned both hyperparameters.
After the hyperparameter tuning, we actually started applying the models. Since, we already used ridge regresion to tune the hyperparameters, we already had the performance benchmarks we needed for this model. We continued with a decision tree model. . . .). Lastly, we conducted a model evalutation. More specifically, we calculated the performances benchmarks for different feature sets of or all of our chosen prediction models. This way, we wanted to determine, which of our models has the best overall performance on our dataset.

Table 5: Feature coefficient

	Data Set	**Prediction Model**	**MSE**	R^2
0	Training Data	Ridge Regression	20660.728749	0.897352
1	Test Data	Ridge Regression	21508.868680	0.900934
2	Training Data	Decision Tree	15933.624822	0.928543
3	Test Data	Decision Tree	19691.169923	0.919737
4	Training Data	Random Forest	14194.545441	0.936342
5	Test Data	Random Forest	17032.140476	0.930576

6.1 Feature Engineering and Selection

We started our predictive analysis with the step of feature engineering. This process included engineering multiple features, which we assumed could be helpful for our prediction models. First, we looked at the columns we already had, coming from our cleaned data sets. Out of this initial data, we managed to engineer 15 possible features and our target, the hourly demand. The features can be inferred from Table X.
A a couple of cyclical features were included, like the start hour or the weekday, which needed transformation. Since our prediction models would not know that those features are cyclical, it would just assume that, for example, weekdays 1 and 7 were vastly apart. Thus, by utilizing sinus and cosinus functions, we created a more continuous and therefore fitting representation of this data. After collecting the features, we analyzed which were actually beneficial for our prediction. By narrowing them down, we avoided overfitting and optimized the performance of our chosen models. Our preferred methods for this purpose were one, a Lasso (L1) regression and two a correlation matrix.
We scaled our features and executed the Lasso regression, which quickly tuned its hyperparameter λ to 0.5. Then, it returned to us each feature's coefficient, representing its influence and thus importance on the model (see Table X). Choosing the highest scoring features, we performed an initial feature selection, resulting in the following feature set: coshour, sinmonth, avgtemp, inrushHour, prevDemand. The results of the correlation matrix further underlined the findings of the lasso regression.

Table 6: Performance Benchmarks

Feature	**L1 coefficient**
precip	6.507
starthour	22.266
sinhour	36.445
coshour	86.587
startmonth	7.412
sinmonth	40.014
cosmonth	37.223
isHoliday	4.732
dayofWeek	11.096
sindayofWeek	0.707
cosdayofWeek	9.53
season	3.167
inrushHour	115.553
prevDemand	293.585

6.2 Model Building and Evaluation

With our initial feature set determined, we continued with model building. First, we split our data into a training and testing section, allowing cross-validation later. And second, we conducted hyperparameter tuning. The following three regression models were subject to our analysis: ridge regression, decision tree and random forest. For finding the optimal polynomial degree for our ridge regression later, we utilized a polynomial regression, looping through multiple degrees and calculating the respective MSEs. With this technique, we determined that our initial optimal polynomial degree lies at four. Next, we needed to find the optimal λ . Similarly to the discovery of the polynomial degree, we looped through various values for λ and determined the respective MSEs. After more thorough investigation of the λ from 10 to 16, we found the optimal λ to be at 13. Hence, we successfully tuned both hyperparameters for the ridge regression. For our decision tree regressor, tuning the tree depth was necessary.
After iterating over a range of tree depths we noticed overfitting to occur after a tree depth of seven, as the performance on the training set started to worsen. For this reason we chose a tree depth of six for our decision tree, which coincidentally was also the optimal maximum depth for our random forest regressor, in terms of minimizing the MSE. Lastly, for our random forest regressor, the number of trees was relevant. A considerably low number of trees seemed to achieve similar performance to higher numbers, which led us to choose a count of seven trees for our random forest.
Lastly, we conducted the model evaluation. For evaluating the models we considered the MSE and R^2metrics. The results can be viewed in table X. The random forest resulted in the best performance, with an MSE of approximately 17032 and a R^2 of approximately 93

6.3 Conclusion of Predictive Analysis

For deployment we would recommend the random forest regressor, as it achieved the best performance on our training data. The application of ensemble methods which combine the outputs of multiple models to generate a combined prediction could also be considered as an option for deployment. In our case the combination of the random forest regressor with the ridge regression could help form a more robust model for prediction. This option however, was not further explored and will be left to future research.

7 Conclusions

Our analysis of rental, weather, and geological data revealed that all observed trends and outliers can be attributed to logical and natural events, as detailed in our report. The insights derived from our analysis enable us to identify the most popular bike stations, their peak usage periods, and user preferences. Furthermore, our demand forecasting capabilities provide valuable information to inform future bike operations and support informed decision making by management. It is important to note that the analysis was limited to only one year and did not take into account the impact of competing companies that could significantly impact demand. The distinction between recreational and commuting users was made possible through our analysis, and it is recommended to tailor marketing strategies to reach these different user segments. Additionally, it is crucial to align the supply of bikes with the identified demand to offer an optimal service. In this way, it is possible to reduce the demand for available bikes in the winter and increase it in the summer that costs are kept as low as possible while still maintaining ease of use.

8 Responsibilities

Ercan Tazegül:

- Data Collection and Preparation
- Temporal Demand Patterns and Seasonality
- KPI: Average available bikes for hours of a day
- Report: 3,4

Simon Knecht:

- KPI: Utilization of user types Customer and Subscriber
- Report creation
- Report: 1, 2, 3, 7

Diego Longhitano:

- Data Collection and Preparation
- Geographical Demand Patterns
- KPI: KPI: Bike Availability per Station
- Report: 3, 4, 7

Mathias Werwie:

- Data Collection and Preparation
- Cluster Analysis
- Report: 5

Vincent Sedlacek:

- Feature engineering
- Hyperparameter tuning for ridge regression (polynomial degree, lambda)
- model evaluation (ridge regression)
- Report: 7

9 References

1. Umweltbundesamt (2022): Treibhausgas-Emissionen in der Europäischen Union, https://www.umweltbundesamt.de/daten/klima/treibhausgas-emissionen-in-der-europaeischen-unionhauptverursacher [am 28.01.2023]

2. Statistisches Bundesamt (2022): Strassenverkehr: Dominanz des Autos ungebrochen, https://www.destatis.de/Europa/DE/Thema/Verkehr/Auto.html [am 28.01.2023]