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

You can find the Github link here: 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.

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. 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.

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

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 a.m. and 4 p.m. till 5 p.m., 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.

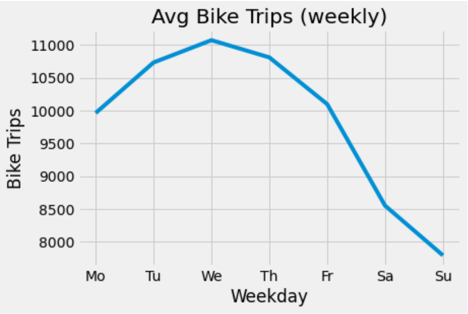


Figure 1: Average Bike Trips weekly

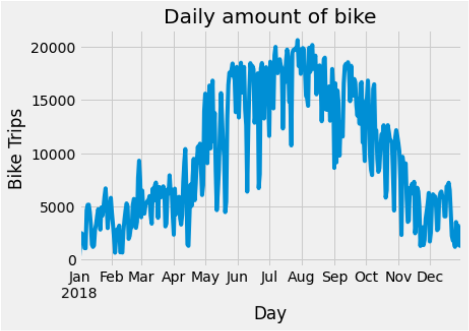


Figure 2: Daily Amount of Bike Trips

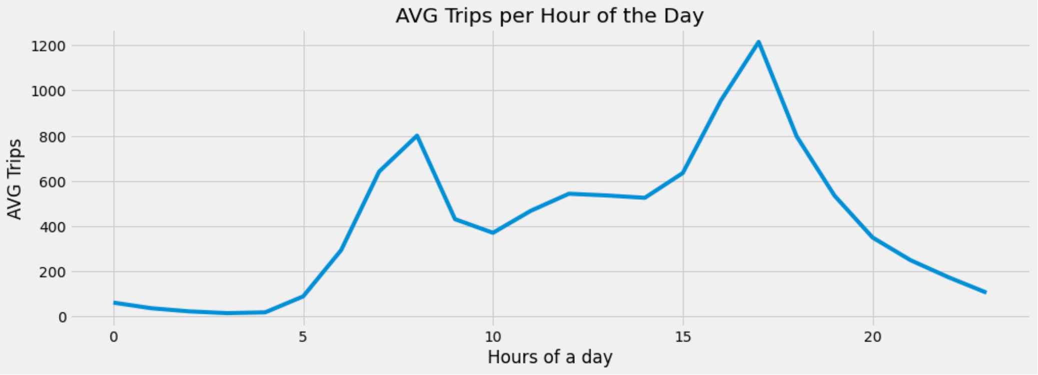


Figure 3: Average Trips per hour of the Day

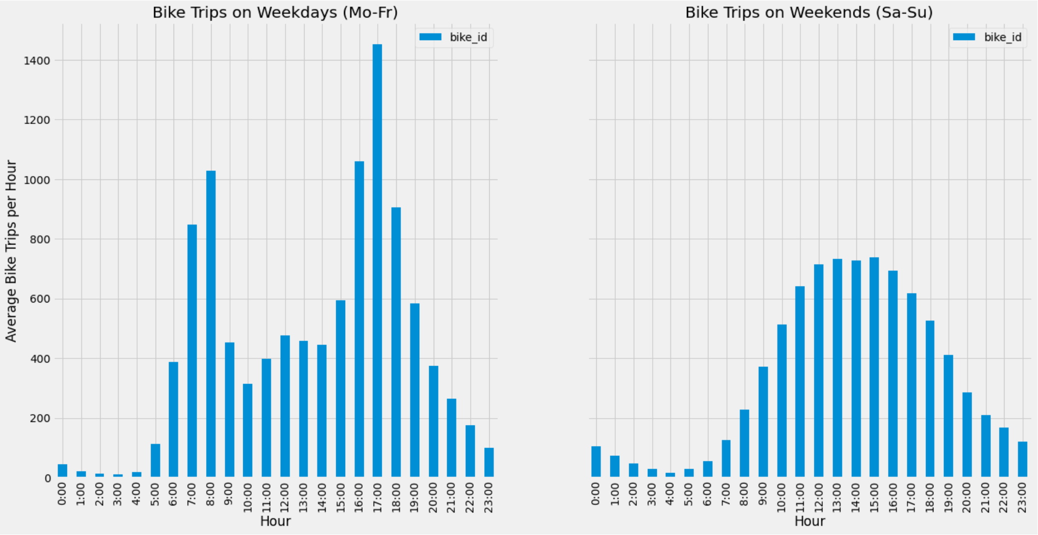


Figure 4: Bike Trips on Weekdays and Bike Trips on Weekends

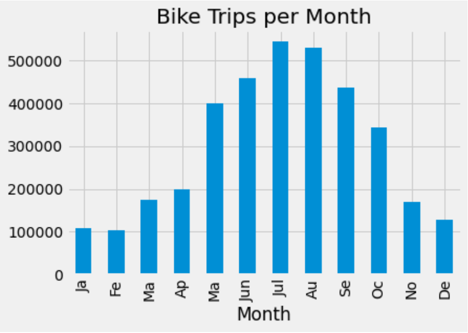


Figure 5: Bike Trips per Month

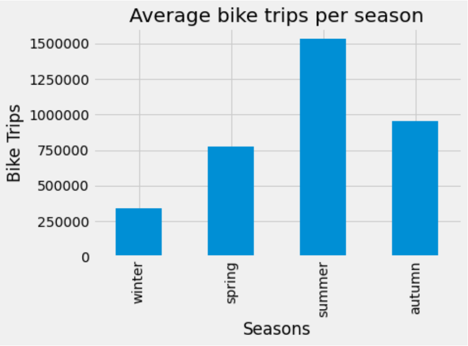


Figure 6: Average Bike Trips per season

4.2 Geographical Demand Patterns

4.2.1 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.

4.2.2 KPI:

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 5am bikes are available in large quantities. After 6 a.m. to 8 a.m., many bikes are used, so the availability is quite low. However, between 9 a.m. to 3 p.m. more bicycles are available. The rush hour is at 5 p.m. where the number of available bikes is the lowest and starts to increase from 6 p.m. to 11 p.m..

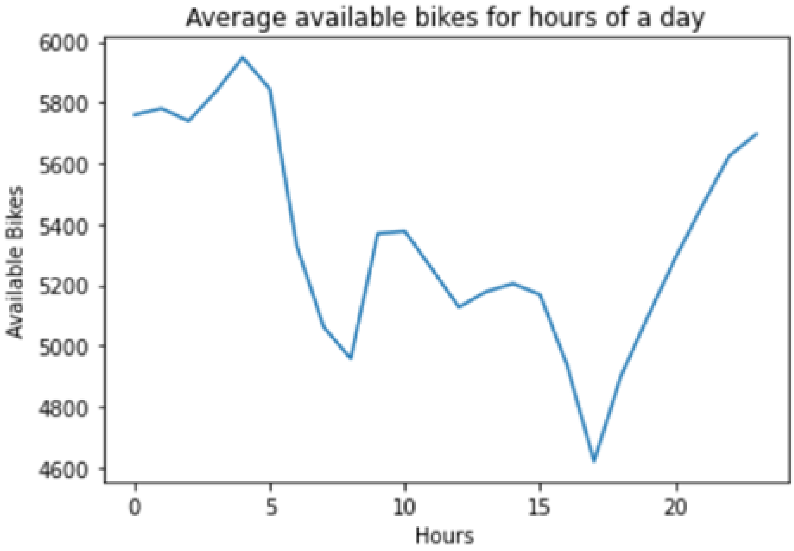
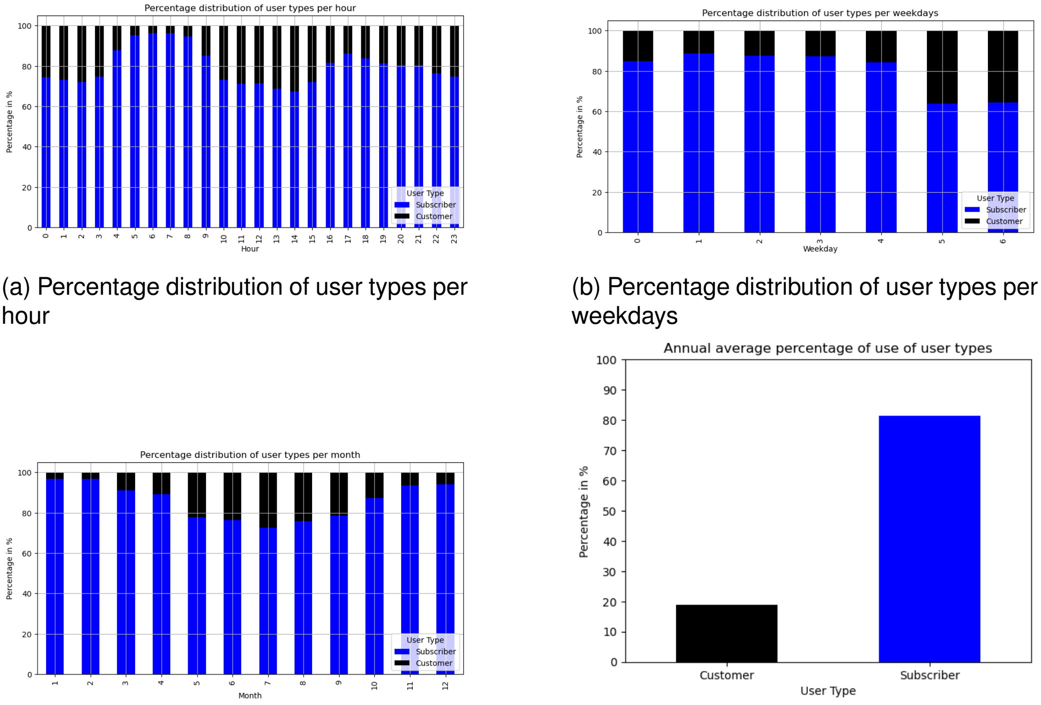


Figure 7: Average available bikes for hours of a day

4.2.3 KPI:

4.2.4 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.



(a) Percentage distribution of user types per hour

(b) Percentage distribution of user types per weekdays

(c) Percentage distribution of user types per month

(d) Annual average percentage of use of user types

Figure 8: KPI: Utilization of user types Customer and Subscriber

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

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

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.

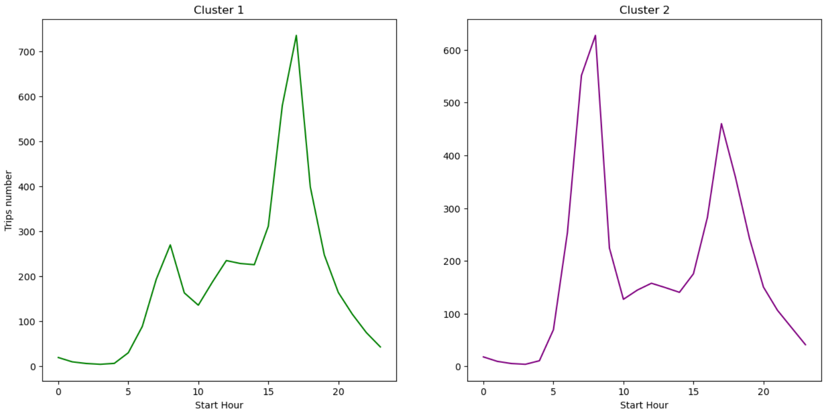


Figure 9: trips number vs start hour

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.

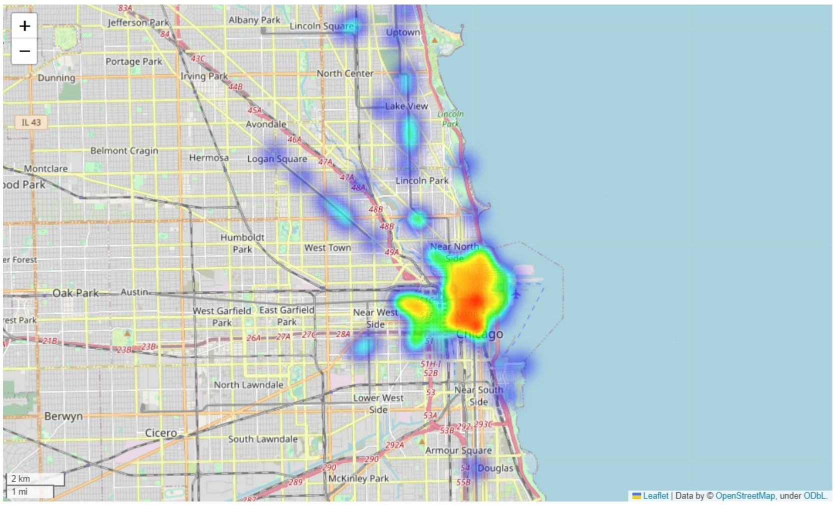


Figure 10: Cluster 1

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.

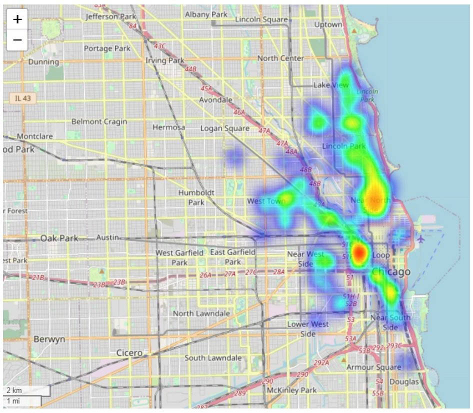


Figure 11: Cluster 2

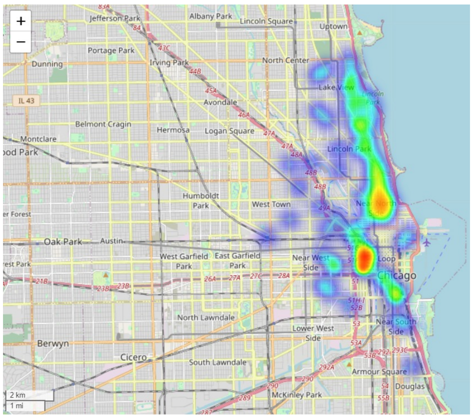


Figure 12: Cluster 2

6 Predictive Analytics

Z

7 Conclusions

Z

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:

Mathias Werwie:

Vincent Sedlacek:

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]
[0.2cm] 2. Statistisches Bundesamt (2022): Strassenverkehr: Dominanz des Autos ungebrochen, https://www.destatis.de/Europa/DE/Thema/Verkehr/Auto.html [am 28.01.2023]