

Using real-time data to monitor and optimize fleet operations

Group: TEAM ZEHN

Members: ERCAN TAZEGÜL, SIMON KNECHT, DIEGO
LONGHITANO, MATHIAS WERWIE , VICENT SEDLACEK

Instructor: PROF. DR. WOLFGANG KETTER, NASTARAN
NASERI

Subject: ANALYTICS AND APPLICATION



UNIVERSITY OF COLOGNE

31.01.2022

Contents

1	Summary	2
2	Problem Description	3
3	Data Collection and Preparation	4
3.1	Bikesharing Data	4
3.2	Geological Data	4
3.3	Wheater Data	4
4	Descriptive Analysis	6
4.1	Temporal Demand Patterns and Seasonality	6
4.2	Geographical Demand Patterns	8
4.2.1	Key Performance Indicators (KPIs)	8
4.2.2	KPI:	8
4.2.3	KPI:	9
4.2.4	KPI: Utilization of user types Customer and Subscriber	9
5	Cluster Analysis	10
5.1	Bikesharing Data	10
5.2	Bikesharing Data	10
6	Predictive Analytics	11
7	Conclusions	12

List of Tables

1	Description of bikeshare dataset columns	4
2	Description of weather dataset columns	5

List of Figures

1	Average Bike Trips weekly	6
2	Daily Amount of Bike Trips	6
3	Average Trips per hour of the Day	6
4	Bike Trips on Weekdays and Bike Trips on Weekends	7
5	Bike Trips per Month	7
6	Average Bike Trips per season	7
7	Average available bikes for hours of a day	8
8	KPI: Utilization of user types Customer and Subscriber	9

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. 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. 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. In the

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	name	str Name of station where trip originated
end station name	name	str Name of station where trip terminated
bike id	int	Unique ID attached to each bike
user type	User	membership type

preprocessing phase of the data, we first eliminated duplicate entries by keeping only the last occurrence and discarding the first. Then we filtered out cases where the starttime was later than the endtime. We also ensured that each stationname was uniquely associated with a single stationid and vice versa, and that the maximum number of stationid associated with a single stationname and vice versa did not exceed 1. To merge cases where a single starting station name was associated with multiple starting station IDs (and vice versa), we implemented a merging procedure. We then calculated the average duration and the total duration of all trips, including the 0.999 quantile of trip times to mitigate the effects of unrealistic outliers. Finally, we added two new columns to the data: 'nextride' and 'nextbike', which indicate whether a particular bike ID is used during a given period, and the identity of the next ride for each bike, respectively. After that, we still included the trips that occurred before and after.

3.2 Geological Data

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

3.3 Wheater Data

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

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)

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.

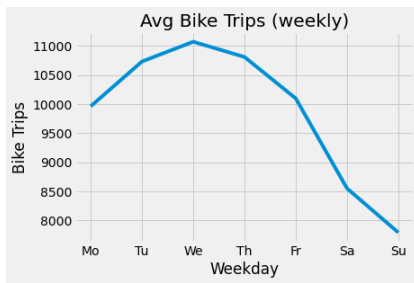


Figure 1: Average Bike Trips weekly

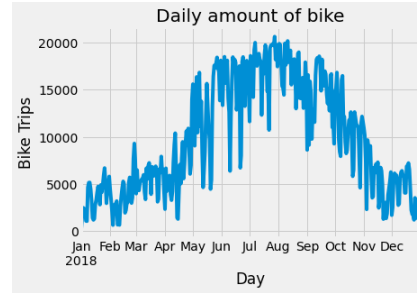


Figure 2: Daily Amount of Bike Trips

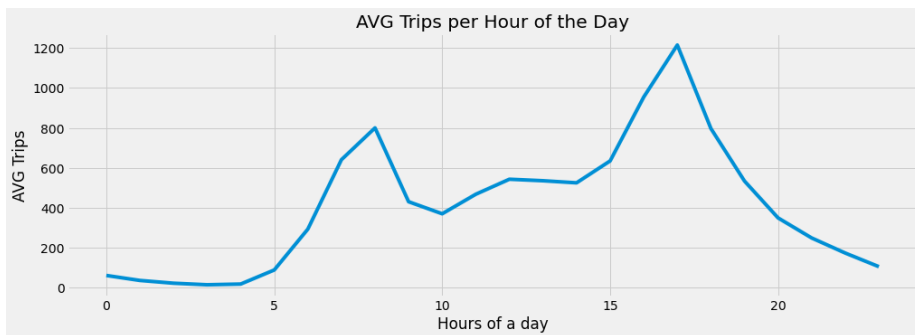


Figure 3: Average Trips per hour of the Day

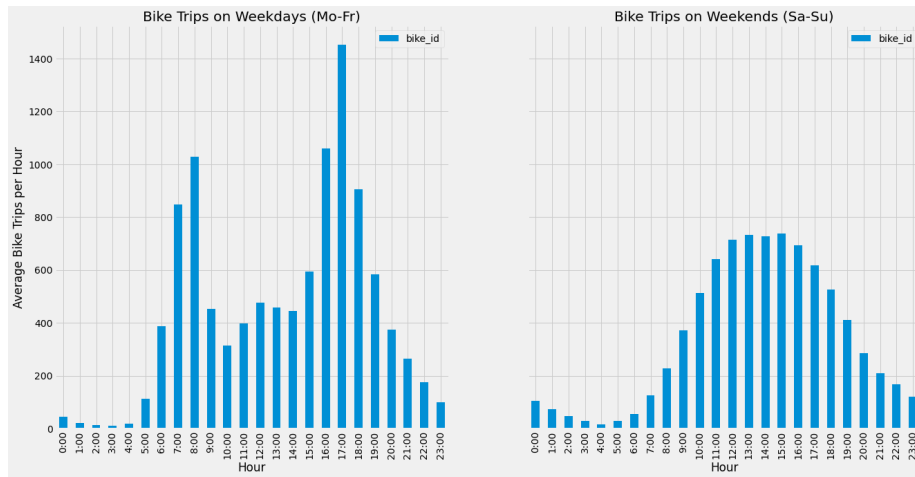


Figure 4: Bike Trips on Weekdays and Bike Trips on Weekends

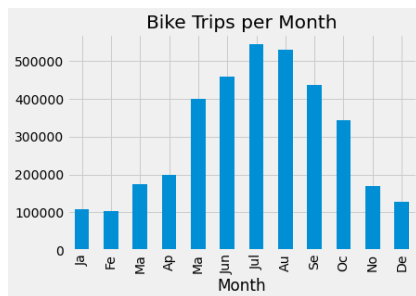


Figure 5: Bike Trips per Month

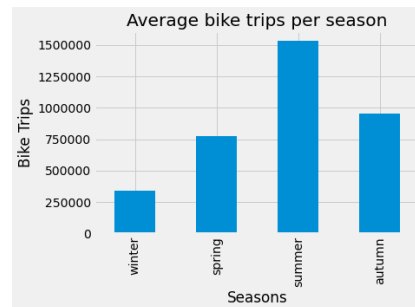


Figure 6: Average Bike Trips per season

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..

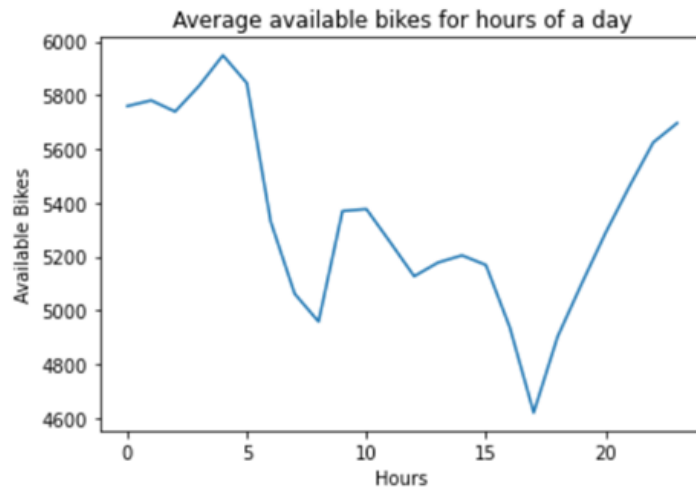
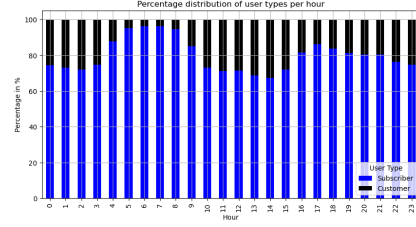


Figure 7: Average available bikes for hours of a day

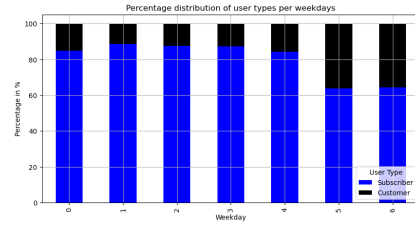
4.2.3 KPI:

4.2.4 KPI: Utilization of user types Customer and Subscriber

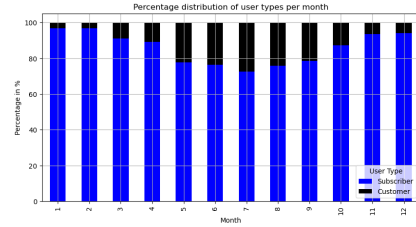
Key performance indicators (KPIs) were introduced for the user types customers and subscribers to analyze the usage behavior of these groups. The KPIs provide information on the hourly, monthly, weekday, and general average usage of these user types. Analysis of these KPIs shows variations in usage among 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.



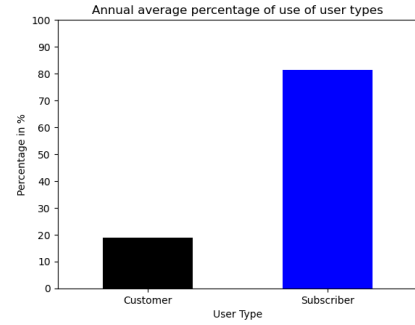
(a) Percentage distribution of user types per hour



(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

5.1 Bikesharing Data

5.2 Bikesharing Data

Z

6 Predictive Analytics

Z

7 Conclusions

Z