Bike-Share Usage in London Network Analysis

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-jiaotonguniv@a3ab6d0 on November 19, 2022.

Authors

- **Mulin Wan**
  · mulin-wan
  CEE, University of Illinois Urbana-Champaign

- **Jingwen Yao**
  · jingwenyaooo · Yaojune
  CEE, University of Illinois Urbana-Champaign

- **Yunze Guo**
  · cyfcx2
  CEE, University of Illinois Urbana-Champaign

- **Bo-Yang Wang**
  · byw-5
  CEE, University of Illinois Urbana-Champaign

Abstract

1 Introduction

Description

1.1 Dataset:

In this project, our goal is to understand how various conditions affect the usage of public bicycle sharing system. We picked London area as the observing site. The main data came from two datasets on Kaggle, titled "London and Taipei Bike-Share Data" and "London bike sharing dataset."

1.1.1 London and Taipei Bike-Share Data

This dataset contains every single bike rental transaction in a total of 802 bike-sharing stops in the London area from 2017 until the Covid outbreak. Each transaction provides the following information:

Table 1: Description of London.csv

Object	Description
rental_id	id of people who rent the bike
duration	duration of rental
bike_id	id of bike
end_rental_date_time	date and time of end rental
end_station_id	id of end station
end_station_name	name of end station
start_rental_date_time	date and time of start rental
start_station_id	id of end station
start_station_name	name of start station
start_rental_date_time	date and time of start rental

1.1.2 London bike sharing dataset

This dataset shows how many bike-sharing transactions took place in each hour in 2015 to 2016. Comparing to the first dataset, this one is more compact since it does not contain individual information. However, it helped providing information on weather conditions. Although the timespan doesn't overlap with the first dataset, it encourages us to find timespan-matching weather data to help with further analysis.

Table 2: Description of London_merged.csv

Object	Description
timestamp	timestamp field for grouping the data
cnt	the count of a new bike shares
t1	real temperature in Celsius
t2	apparent temperature in Celsius
hum	humidity in percentage
windspeed	wind speed in km/h
isholiday	boolean field - 1 holiday / 0 non holiday - refers to bank holidays

Object	Description
isweekend	boolean field - 1 if the day is weekend / 0 if a working day
season	category (0-spring; 1-summer; 2-autumn; 3-winter)
weathercode	different weather condition

Table 3: Description of weathercode

weathercode	Description
1	clear; mostly clear but have some values with haze/fog
2	scattered clouds / few clouds
3	broken clouds
4	clear; cloudy
7	clear; light rain shower / rain / light rain
10	clear; rain with thunderstorm
26	snowfall
90	freezing fog

In addition to season and isweekend, from the timestamp feature we can extract many separate time features - day of the week (as one scaled column or as seven columns of ismonday, istuesday etc.), month number, day of the month, week number, hour, minute. In combination with external data, we could add is_light for after dawn times and is_schoolholiday to match London school holiday times.

1.1.3 Link of dataset:

London and Taipei bikeshare

London bikeshare

1.2 Proposal

Recently, bike-sharing in big cities has become an important part of residents' daily life, and its role in urban transportation system has never been more significant. Looking at the spatiotemporal bike-sharing data in London, we could explore patterns, describe variations, or model the data in many different ways. From the two datasets, we may have a chance to take a peek at the residents' bike-renting behavior through many angles.

Previous work has shown that weather is a key driver for variation in usage. [1] [2] Aside from weather, We believe there are a lot more important factors such as peak/off-peak hours, weekday/weekend, bike-stop location etc. By utilizing these datasets, we hope to find as many correlations between the users behavior and various factors.

We plan to start by looking at the trends. How does weather or other factors affect the London area overall? Although the answer could be found in both datasets, the structure of the second dataset(see 1.1.2 London_merged.csv) would make the job easier if we were only looking at big trends. Then we would look at the microscopic data provided by the first dataset(1.1.1 London.csv), and hope it would support our claims.

Lastly, after each correlation is explored, we will try to formulate a model that would help us predict the hourly bike-sharing usage in the stops. Our objective is to give a usage forecast in order to help users manage their travel time, and for service providers to better dispatch bikes to maintain service quality.

2 Exploratory Data Analysis

In this section, we look at different factors affecting the usage of the bike sharing system in London. Each factors that we are interested is plotted along with the average usage per hour. Then we look for micro trends in specific bike-stop that contradicts with the big trends we found.

2.1 Data Wrangling

2.1.1 Data cleaning process

Both of the csv files needed to be restructured in some ways before the analyzing process. For instance, the exact time is stored as strings:

"2015/1/4 12:00:00 AM"

Information such as date and time could be extracted from within. The somehow trickier part is the day-of-the-week. We add a certain number to the date and take the remainder after divided by 7 to get the day-of-the-week.

2.2 Analysis and Visualization

2.2.1 Large trends

In this section, we look at how different factors affect the average usage per hour in 2015-2016.

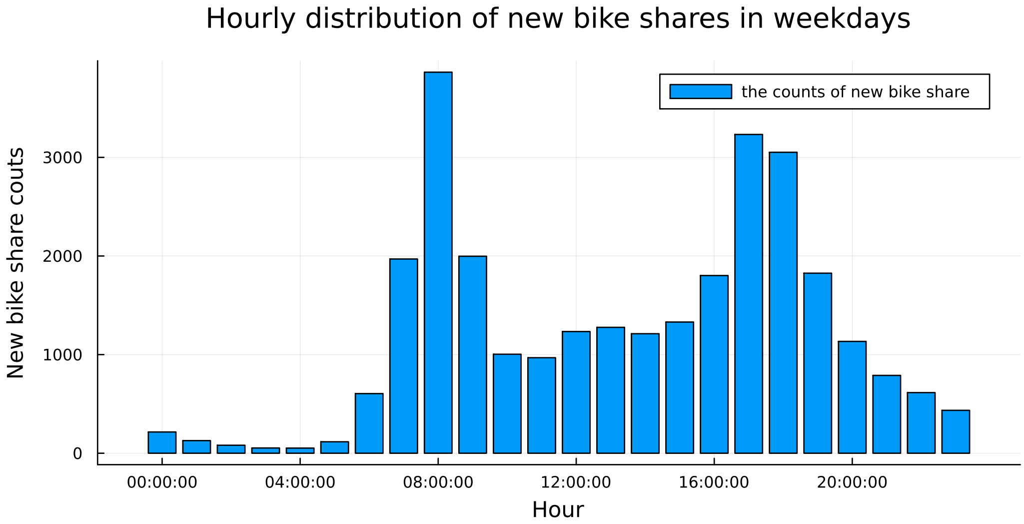

Figure1: Hourly average usage on weekdays

Figure1 shows how average bike-sharing usage distribute in different hours in a weekday. In the image, one could easily spot a double-peaked distribution. This comes with no surprise - the rush hour in weekdays generates a lot of commuting demands, and apparently people turn to bike-sharing in these hours. On average, over three thousand people rented a bike at 08:30 everyday, the busiest time in terms of bike-sharing usage.

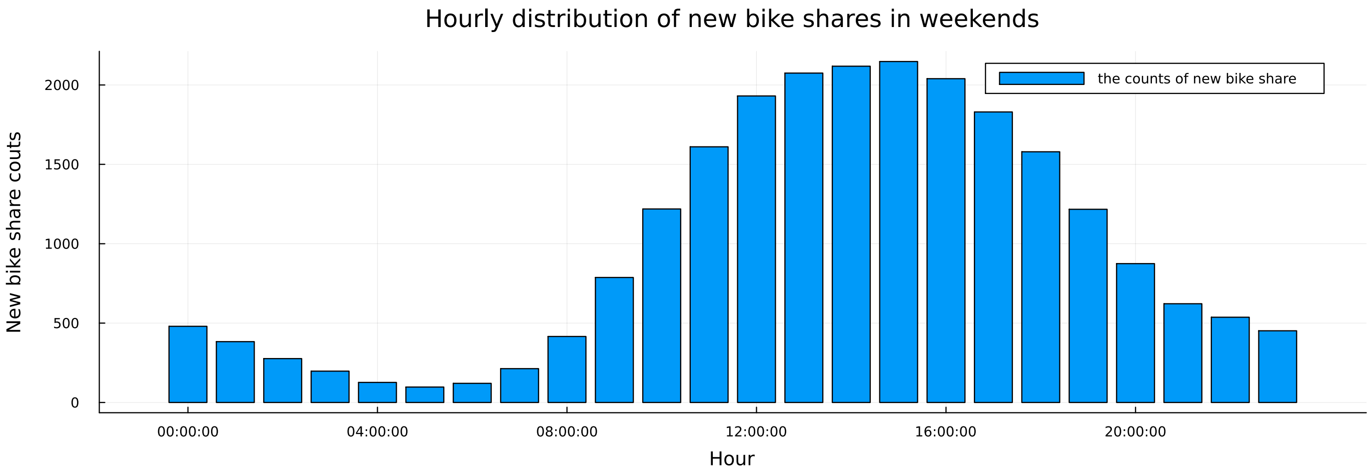

Figure2: Hourly average usage on weekends

Figure2 shows how the new bike share demand distribute in different hours in weekends. Base on the image, we can speculate that Londoners are most active between 11:00 and 19:00 on weekends.

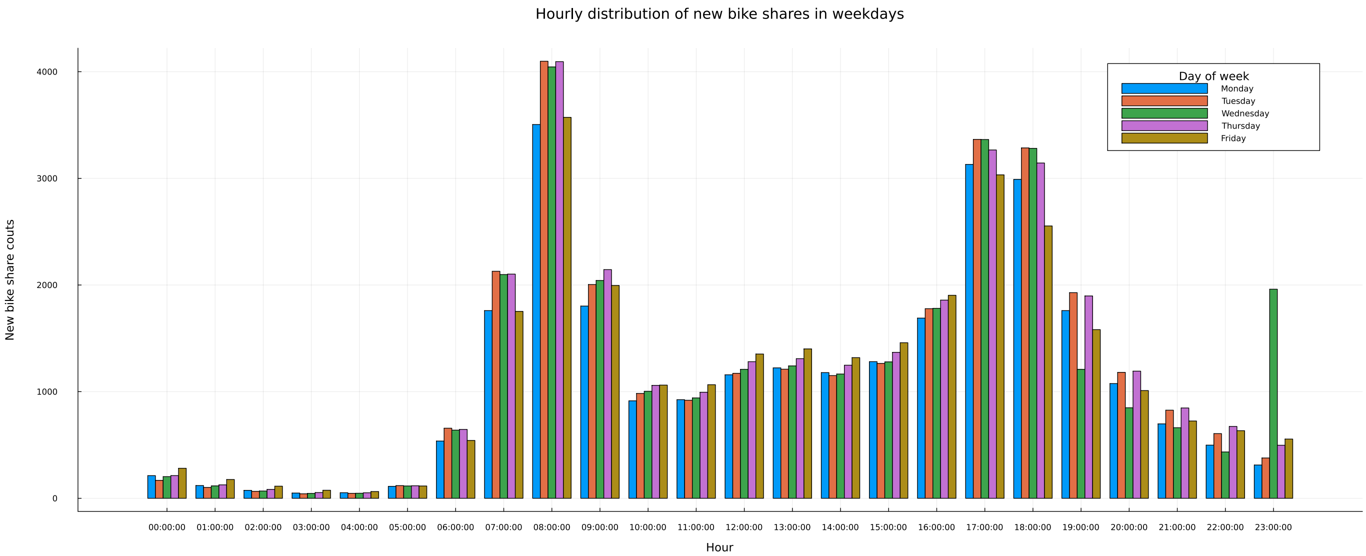

Figure3: Hourly average usage between different weekdays

Figure3 generally agrees with Figure1. During rush hours, bike-sharing usage climates. There are not many conclusions to make according to this figure, except that usage characteristics are mostly the same during Tuesday to Thursday. Consider a two-working-day span that lies in Tuesday to Thursday, with nearly identical weather conditions, we could speculate that these two days would have similar bike-sharing usage. Mondays and Fridays on the other hand, are seen to have slight difference to their weekday counterparts.

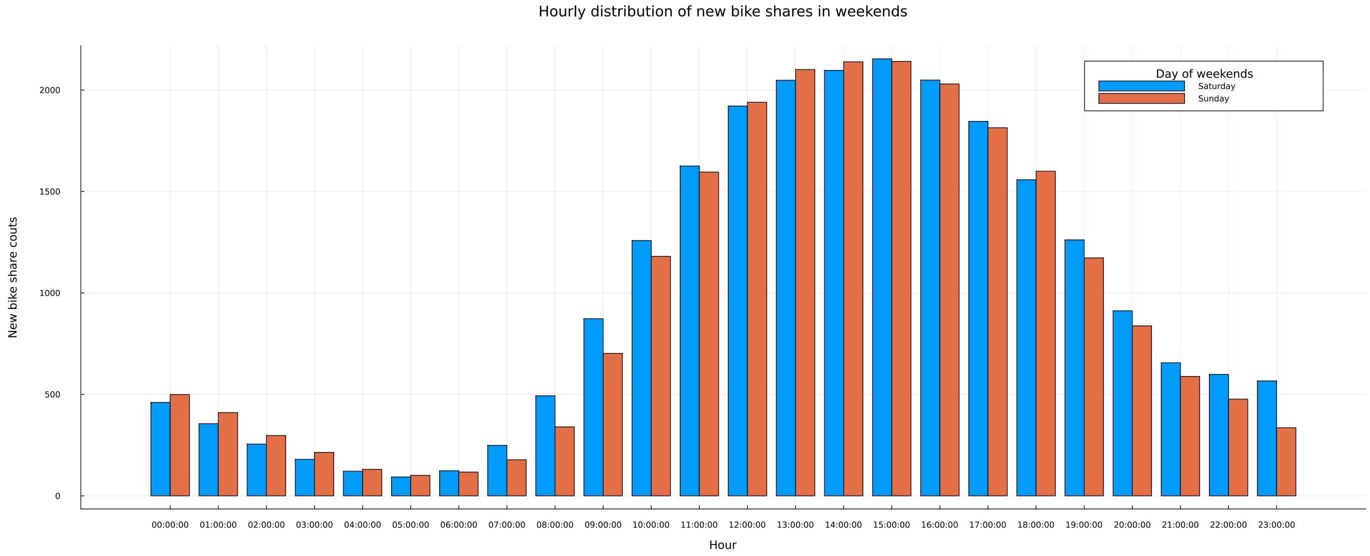

Figure4: Hourly average usage between different days in the weekend

Figure4 shows that the overall difference of new bike share between the two days of the weekend is not big except one logical difference: since Monday is a working day, Sunday's usage at night can be seen to be smaller than Saturday.

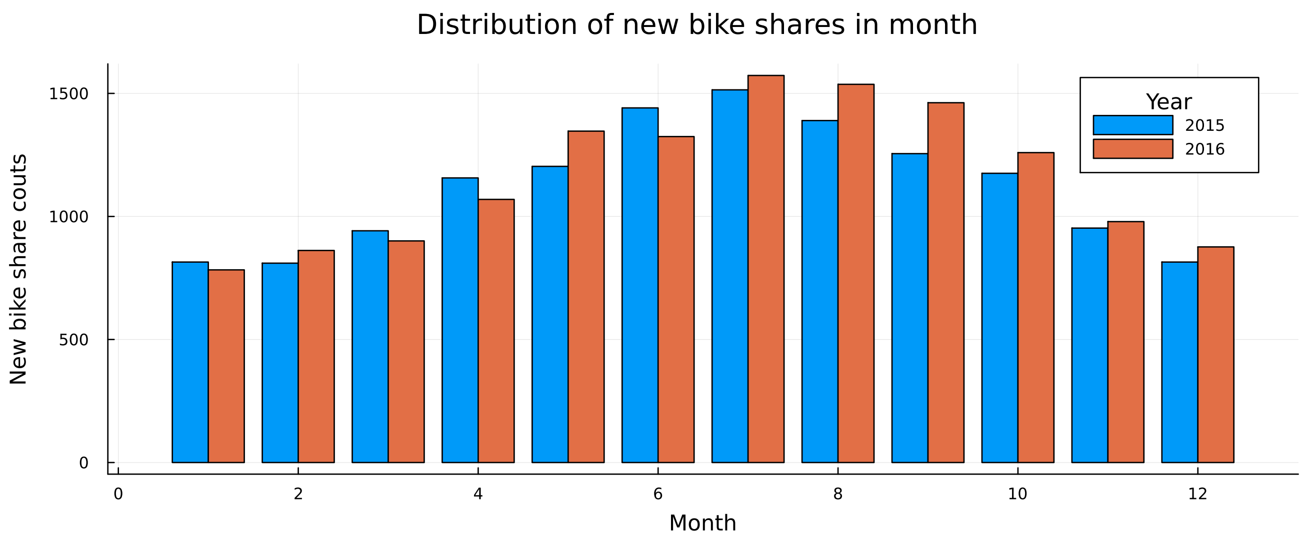

Figure5: Average usage/hour between different months

In figure 5, the hourly trend is still similar to that in figure1. Meaning in a given day regardless of the month, rush hour still generates the most bike-share usage. However, large differences between months could be spotted, especially between the April to October period and the November to March period. We can easily come to a conclusion that users are less willing to ride a bike in the cold.

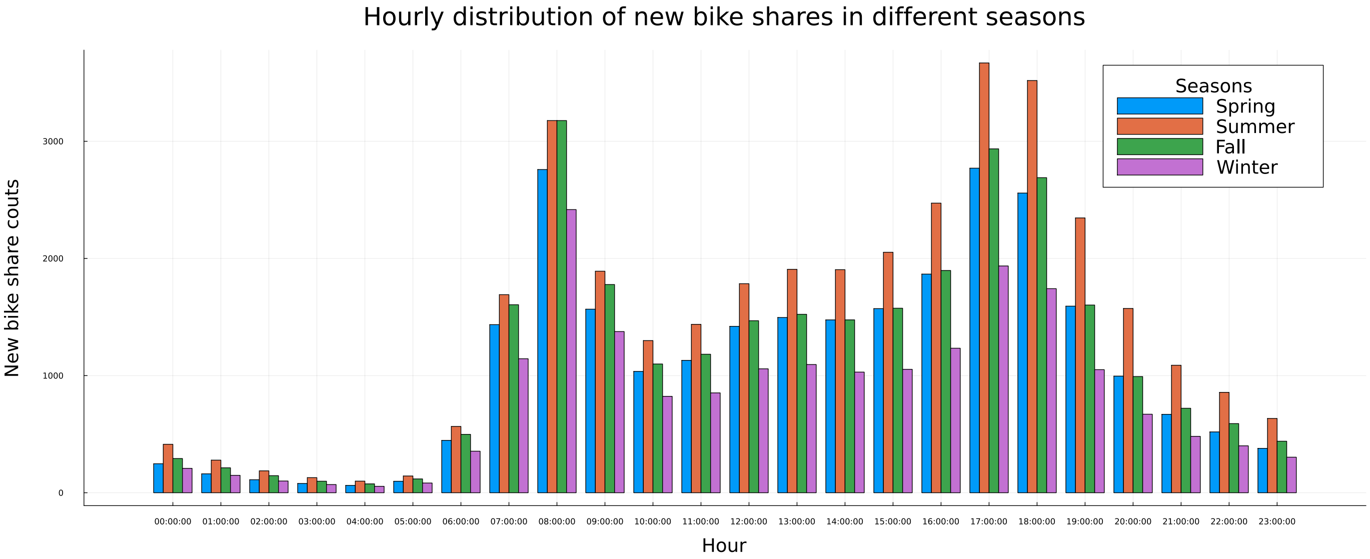

Figure6: Hourly average usage between seasons

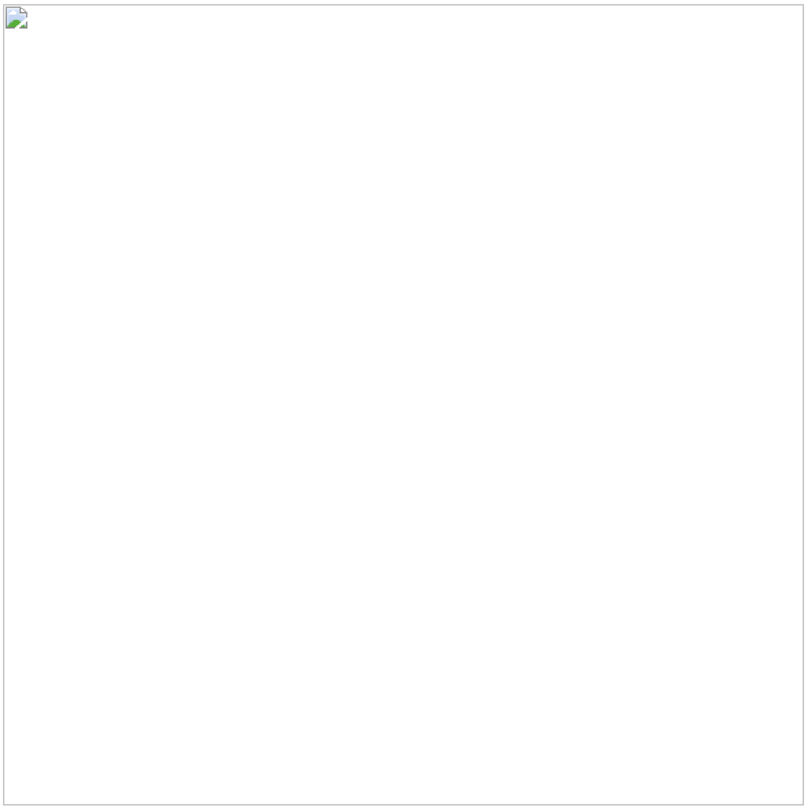

Figure7: Average usage/hour between seasons

According to figure6 and figure7, it can be seen that the demand for new bike share in London is relatively higher in summer and autumn overall, especially in summer. Winter is undoubtedly the lowest, but in this image it can be seen that the demand for new bike share is lower in spring than in autumn. We can speculate that people are more willing to rent a bike in the season of relatively higher temperature, and the weather in autumn is more suitable for bike share than spring in the London area.

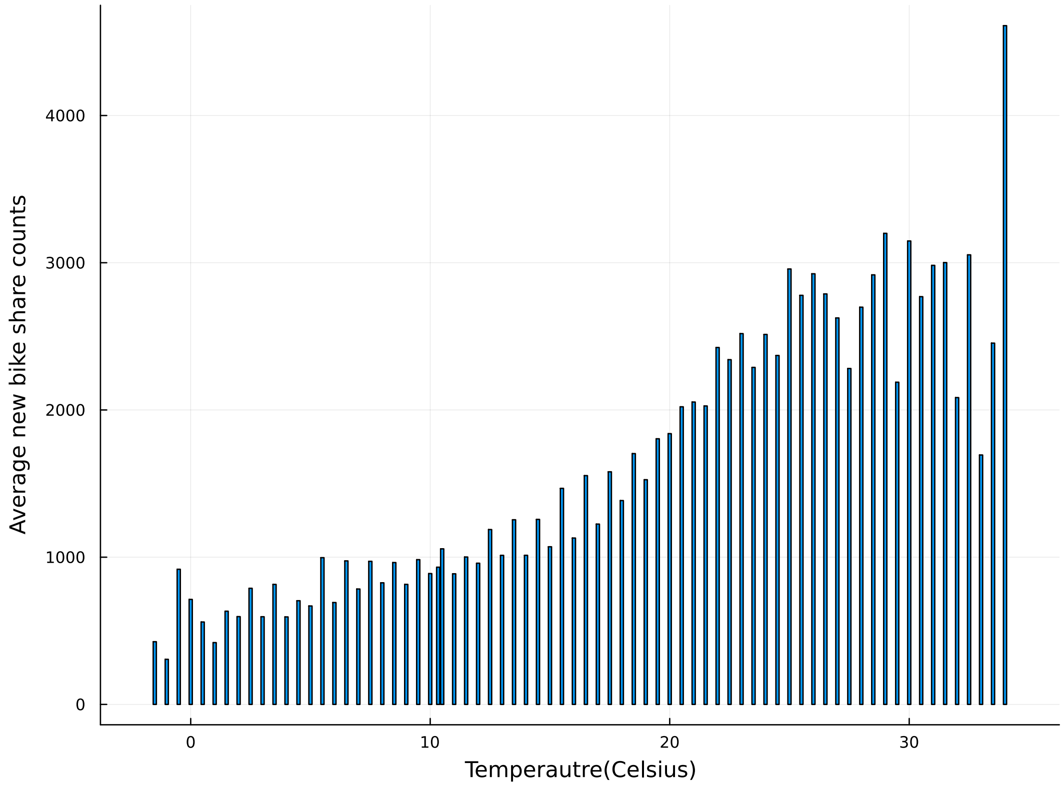

Figure8: Average usage/hour in different temperatures

In figure8, temperatures over 34 degree Celsius are all recorded as 34. If we neglect the last bar, we can see that bike-sharing usage gradually increases until the temperature reaches 30, and then went downwards.

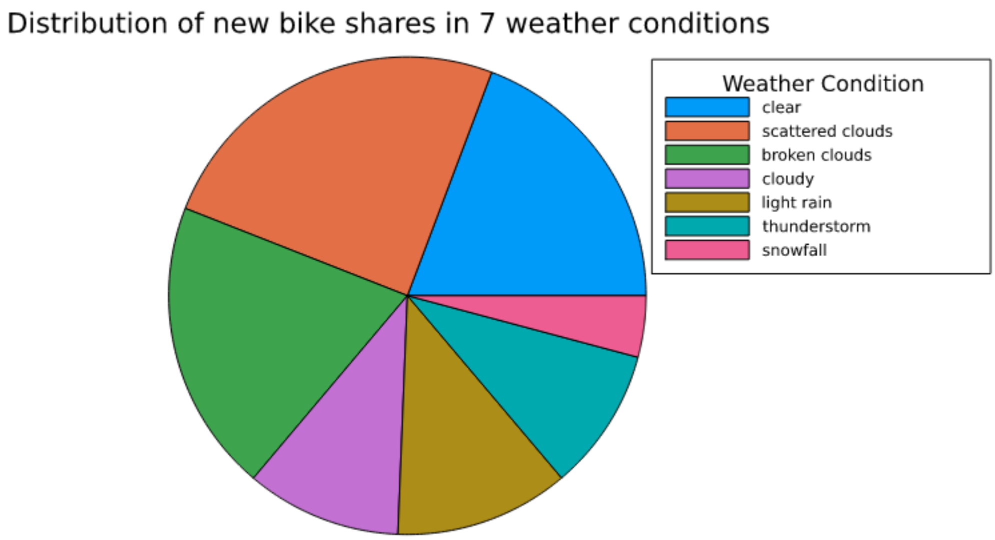

Figure9: Bike-share usage under 7 weather conditions

This pie chart shows the how users utilize the bike sharing system in different weather conditions. Basically, when the weather is good, people tend to utilize the bike sharing system more often, which is easy to understand. In London, raining doesn't bother this city that much since Londoners have developed a certain life style or fashion to accommodate their unique weather condition. This phenomenon can also be spotted right here,since there is not a huge difference in usage between "cloudy", "light rain" and "thunderstorm".

2.3 Predictive Modeling

2.3.1 Spotting micro trends (Individual behaviors varying with bike stop locations)

In the previous section, we have come up with some speculations, such as:

During weekdays, usage during rush hours are often higher than non-peak hours. Usage in weekdays are often higher than weekends. Usage in warmer days are often higher than colder days.

But as we move closer the the actual stop-by-stop prediction, we need to understand how the location and the characteristc of each stop changes how the large trends' impact on those stops. The main dataset (1.1.1 London.csv) provides a chance to look extremely closely to certain stops in certain time spans, for us to verify out speculations, or to discover new revealations.

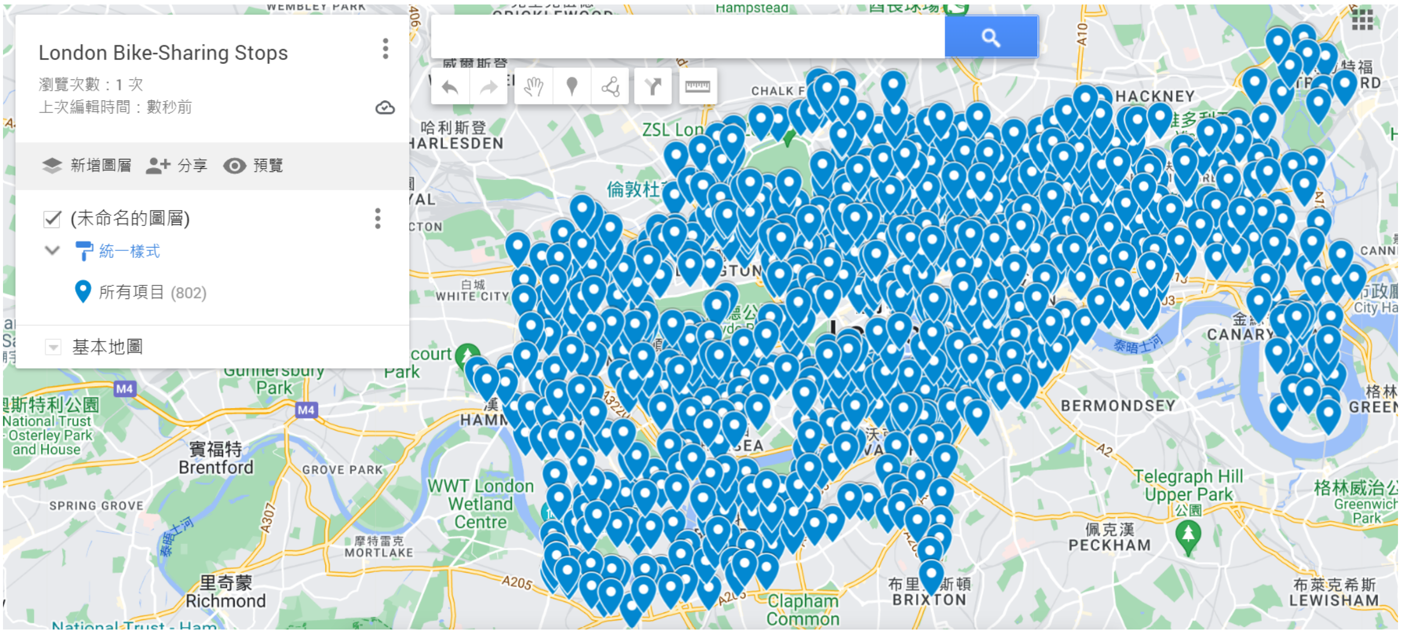

Figure10: Bike-share stops in London area

In the dataset there are 802 stops, as shown in the figure above. We will be looking at two stops:

Triangle Car Park, Hyde Park : Located right in the middle of the famous tourist attraction Hyde Park. (Will be later denoted as Hyde) Queen Street 1, Bank : Located in the central of business districts . (Will be later denoted as CBD)

And we will see how different conditions affect them respectively.

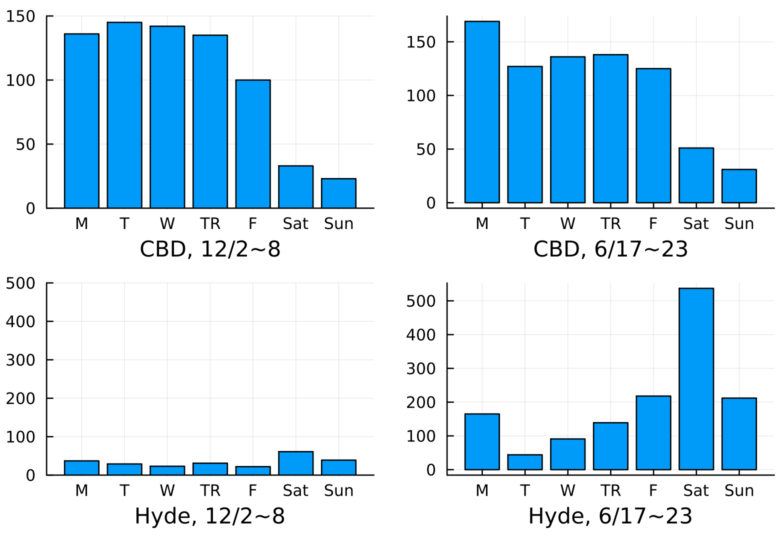

Figure11: CBD and Hyde Park Comparison, Winter versus Summer

In figure11, we can see the date is set on 12/02 ~ 12/08 and 06/17 ~ 06/23 (2019). They are both regular non-holiday weeks with little to none precipitation. Although we almost came a conclusion that usage in winter is almost always lower than that of summer, in the busy business district, we can hardly tell the impact from weather.On the other hand, bike-share usage took a great hit from summer to winter in Hyde Park. Meaning tourist activities are significantly lower in cold times.

Bike stops near tourist attractions can have another trait different than the speculations we made from observing large trends. We can see in June, Hyde Park attracts large amount of bike usage in weekends. This serves as a reminder that weekdays do not always have larger usage than weekends when predicting.

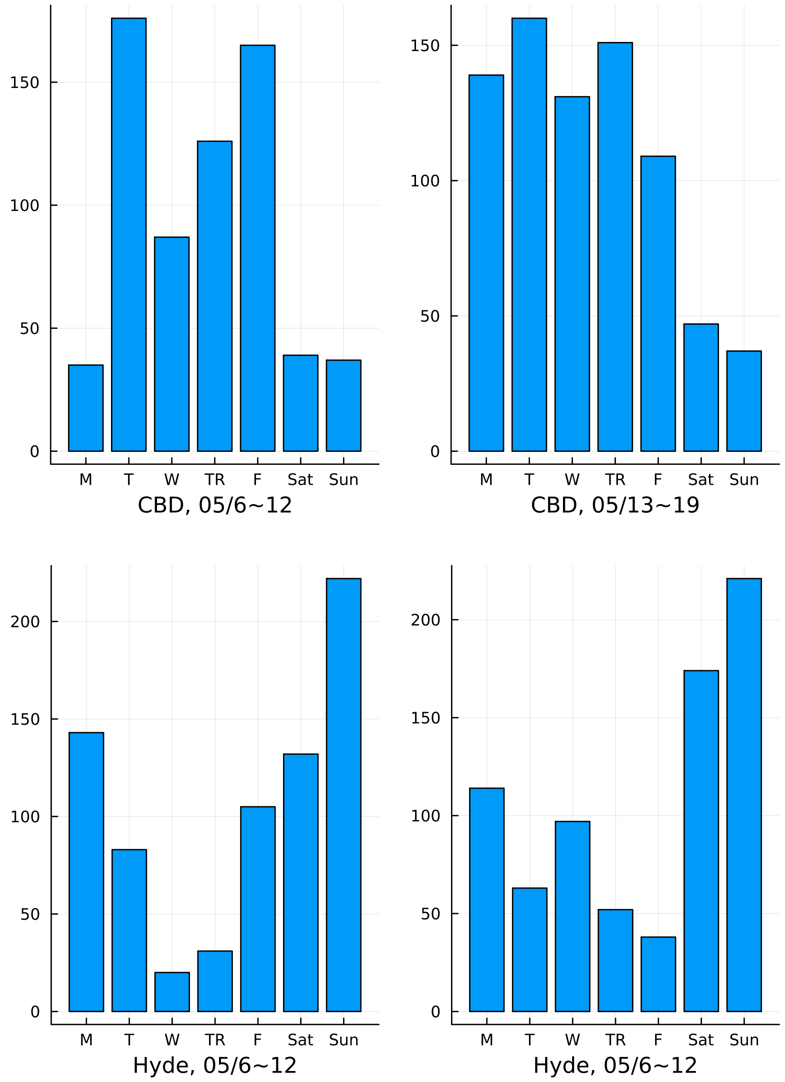

Figure12: The effect of Rain and National Holidays

In figure12, we can see the date is set on 05/06 ~ 05/12 and 05/13 ~ 05/29 (2019). 05/06 (Mon) is a national holiday in UK. Also, there are heavy rainfall during 2019/05/08 ~ 09. We can see the national holiday having drastic on usage in CBD, causing a giant difference between the two Mondays, 05/06 and 05/13. However Hyde Park was not that severely affected. From the figure we can also see less bike usgae on 2019/05/08 ~ 09, regardless the location. Comparing to the result in Figure11, it is safe to say that precipitation affects bike users in CBD more than low temperature. But is this phenomenon universal across London? Or is this a business district thing? We may need to look for other proofs.

2.3.2 Variables for predicting

The main dataset (1.1.1 London.csv) although contains rich content, is too time-consuming to perform a thorough exploratory data analysis right now. But until the next step, it would be neccesary to look for deeper connections between the dots. For now, combining large trends and micro trends, we have thought of the following variables for predicting bike-share usage:

Variables	Description
Time	What time of the day
Day	What day of the week
Holiday	Is it a holiday or not
Temp	Temperature in Celsius
Light	Whether there is still daylight
Location	Characteristics of the bike stop location
Surrounding	Renting availability in nearby stops
Transport	Other means of transportation available
Crime	Level of safety in the neighborhood

Due to the complexity of the problem, we would then narrow the observing area from London entirely to a certain area, hopefully containing schools, tourist attractions, business areas and residential area in order to give diversity to the problem.

3 Preliminary Predictive Modeling

3.1 Train a Regression Model

3.1.1 Splitting the data

As the data has been explored, the next step is to train a regression model and predict the bike sharing number(cnt):

'''python X = df.drop("cnt", axis=1) y = df["cnt"] print('Parameters:',X[:10]) '''

```
Parameters:
    t1   t2    hum  wind_speed  is_holiday  is_weekend  month  da
season_0.0  season_1.0  season_2.0  season_3.0
0  3.0  2.0   93.0         6.0         0.0         1.0      1
0           0           0           1
1  3.0  2.5   93.0         5.0         0.0         1.0      1
0           0           0           1
2  2.5  2.5   96.5         0.0         0.0         1.0      1
0           0           0           1
3  2.0  2.0  100.0         0.0         0.0         1.0      1
0           0           0           1
4  2.0  0.0   93.0         6.5         0.0         1.0      1
0           0           0           1
5  2.0  2.0   93.0         4.0         0.0         1.0      1
0           0           0           1
6  1.0 -1.0  100.0         7.0         0.0         1.0      1
0           0           0           1
7  1.0 -1.0  100.0         7.0         0.0         1.0      1
0           0           0           1
8  1.5 -1.0   96.5         8.0         0.0         1.0      1
0           0           0           1
9  2.0 -0.5  100.0         9.0         0.0         1.0      1
0           0           0           1
```

Figure13: Splitting The Bike Sharing Number From Other Parameters

To validate the training model, we split the dataset into two subsets; the first set is used to train the model, and the second(and smaller) one is used to validate the model by comparing the predicted labels to the known labels. The data is randomly split to about 7:3. We realize it by the `train_test_split` function in the 'scikitlearn' library in python. And the result is four datasets:

- **X_train**: The feature values we'll use to train the model
- **y_train**: The corresponding labels we'll use to train the model
- **X_test**: The feature values we'll use to validate the model
- **y_test**: The corresponding labels we'll use to validate the model

The next step is to train the model with a proper regression method. The group used a linear regression algorithm, which is basic and commonly used, to find a linear relationship between X and y.

3.1.2 Scaling & Training

Scaling is also a preparation step for machine learning. By scaling the numeric columns in the dataset to a common scale with the standardization method, the distribution could have a unit standard deviation. 'sklearn.preprocessing' and 'sklearn.linear_model' package provide a convenient algorithm to realize the model:

```
# Fit a linear regression model on the training set
model = LinearRegression().fit(X_train, y_train)
```

3.2 Evaluate Trained Model

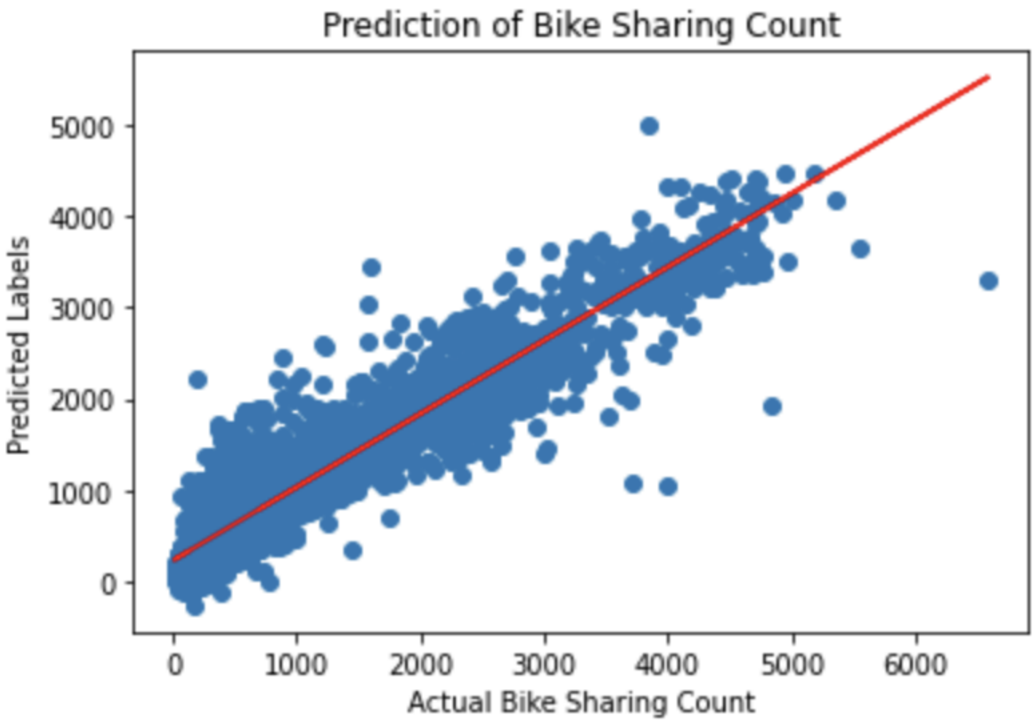

Figure14: Predicted and Actual Bike Sharing Number of Linear Regression

It is a generally diagonal trend with several deviation values. The group uses mean square error to identify our model's error level, and the result is 814512. One reasonable explanation is linear regression can only clearly show the data trend, but it cannot cover too much data in a dataset. To improve the power of our model, the group also uses the decision tree method, and the result of MSE is 97082:

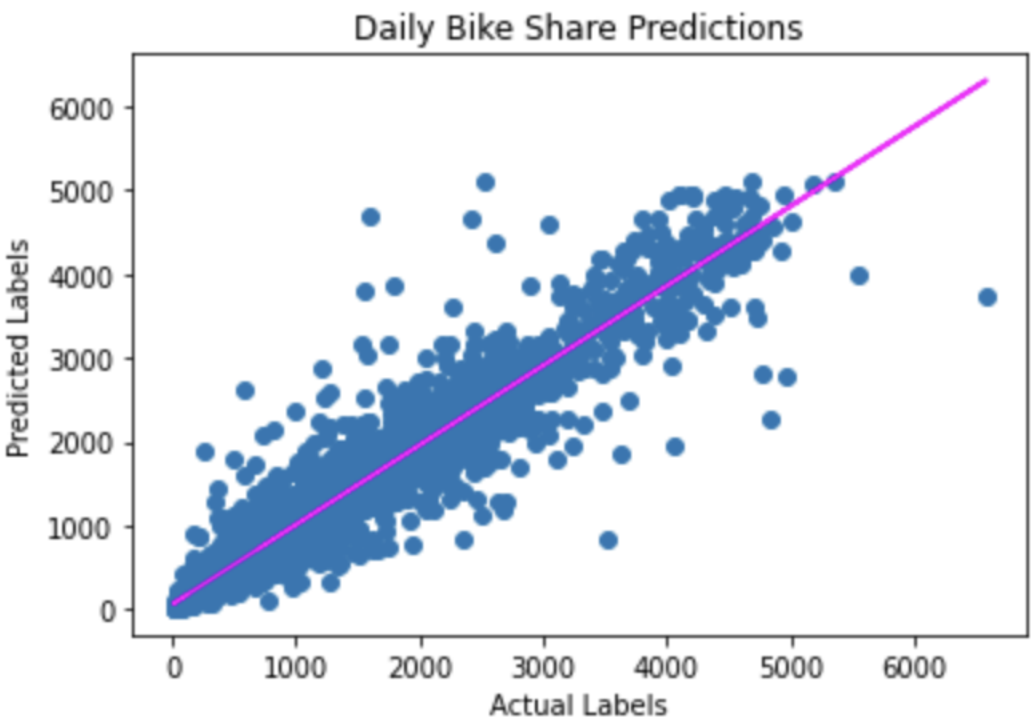

Figure15: Predicted and Actual Bike Sharing Number of Decision Tree

A more intuitive way to compare the improvement of the model is using the coefficient of determination(R-squared).

Evaluation Metrics	Linear Regression	Decision Tree
MSE	814512	97082
R^2	0.31	0.91

As the R^2 is more than 90% now, the improvement of the model is obvious.

References

[1] AndersOhrn (2020) Bike-share usage in London and Taipei Network, Kaggle. Available at: https://www.kaggle.com/datasets/ajohrn/bikeshare-usage-in-london-and-taipei-network (Accessed: October 24, 2022).

[2] Mavrodiev, H. (2019) London Bike Sharing Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset/discussion?resource=download (Accessed: October 24, 2022).