Prediction modeling for bicycle-sharing system
- A case study on NYC Citi Bike

1. Introduction

A bicycle-sharing system, public bicycle system, or bike-share scheme is a service in which bicycles are made available for shared use to individuals on a very short-term basis for a price. Since past 5 years there is an exponential growth in the companies providing bicycle sharing services. Bike share schemes allow people to borrow a bike from point A and return it at point B. Many bike-share systems offer subscriptions that make the first 30–45 minutes of use either free or very inexpensive, encouraging use as transportation. This allows each bike to serve several users per day. In most bike-share cities, casual riding over several hours or days is better served by bicycle rental than by bike-share. For many systems, smartphone mapping apps show nearby stations with available bikes and open docks.

Bicycle Sharing Systems are popular worldwide. OfO company in China is one of the pioneers of this system. Other famous companies include Santander in London and Citi Bike in New York City, USA. Most recently, Ola, cab service provider in India secured more than a billion-dollar capital to launch Ola-Pedal, the bicycle sharing system.

Ola-Pedal: India

Ofo: China

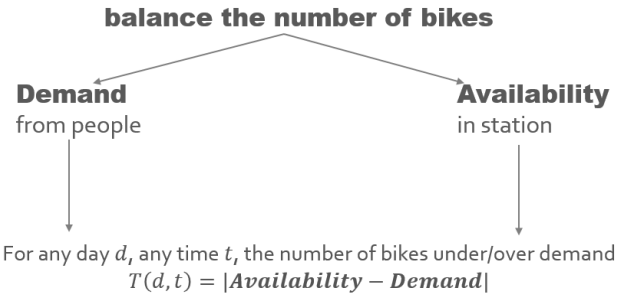
Santander: London

Citi Bikes-NYC

2. Problem description

For this project we have selected Citi Bike in New York City. Citi Bike is the nation's largest bike share program, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City. It was designed for quick trips with convenience in mind, and it’s a fun and affordable way to get around town. Bicycle-sharing system (BSS) has attracted much attention worldwide due to its great success in providing a low-cost and environment-friendly public transportation method. However, if we look at Citi Bike NYC, on every weekday morning Penn Station docks run out of bikes. This problem arises because of unbalanced bike availability at stations, to balance the number of bikes among stations is a problem that worth investigation to better satisfying customers' demand in different time period. The goal of the project is to

predict the availability and demand of bikes in each station by using statistical analysis techniques. The prediction is based on the daily (24hours) and weekly behavior observed from the historical data provided by Citibike in New York City.



3. Data Description

The availability data can be obtained from a real-time json file provided in the Citibike website (https://feeds.citibikenyc.com/stations/stations.json). Fortunately, some organizations have extracted the station status data for the past years (https://drive.google.com/drive/u/0/folders/0B6H9nKo1G98uS3kxQ1VrNGt5SjA) . The data structure is demonstrated in Table 1. We only use the one-year data from July 2016 to June 2017.

Since we only care about the number of bikes/docks of a given station at certain time, the data structure of the station status is re-formatted as Table 2. The visualization of NYC map can be done based on the latitude and longitude of the popular places besides venders provided by Foursquare location data.

Table 1. Raw Data Structure of Station Status

Variable	Format
id	Number
stationName	String
availableDocks	Number
totalDocks	Number
latitude	Number
longitude	Number
statusValue	String
statusKey	Number
availableBikes	Number
stAddress1	String
lastCommunicationTime	Timestamp

Table 2. Final Data Structure of Bike/Dock Availability

Variate	Format
Station_id	Number
Date	String
Hour	Number
Minute	Number
Avail_bikes	Number
Avail_docks	Number
Tot_docks	Number
In_service	String
Status_key	Number

The demand data is indicated in the historical trip data, whose structure is shown in Table 3. The demand of bikes can be obtained by counting the number of records showing the same "Start Station ID" at a given hour in a given day. In addition, a new variable named "weekday" is added. It is ranged from 0-6 indicating Monday to Sunday. The final data structure for demand analysis is shown in Table 4.

Table 3. Raw Data Structure of Historical Trip

Variable	Format
Trip Duration	In hour, minute and second format
Start Time	Timestamp
Stop Time	Timestamp
Start Station ID	Number
Start Station Name	String
Start Station Latitude	Number
Start Station Longitude	Number
End Station ID	Number
End Station Name	String
End Station Latitude	Number
End Station Longitude	Number
Bike ID	Number
User Type	Casual or Registered
Birth Year	Number
Gender	String

Table 4. Final Data Structure of bike/dock Demand

Variable	Format
Station ID	String
Date	String
Hour	Number
Weekday	Number
Dock_demand/Bike_demand	Number

4. Data Exploration

Station Overview

The Figure 4. shown above is the total demand of all station. Each bar refers to total demand in a station. We can easily see the popularity of stations and the demand distribution in an area of stations. The lengths of bars have a periodic tendency which means the stations has contiguous station id may has same demand of bike. Thus, the station could be settled by areas. Station id in range 0 to 350 has a steady high average demand which are more than 20000 totally, otherwise the number after 350 demand much lower continuously which is close to 4000.

Dock demand by station is presented as well as a comparison to demand of bikes. A few insights can be developed from these two figures. The distribution of dock demand is almost same as bike demand. It could be one of explanation that the users of bike employed the bike for round way which means the bikes are more likely to be returned to the station where they be lent out.

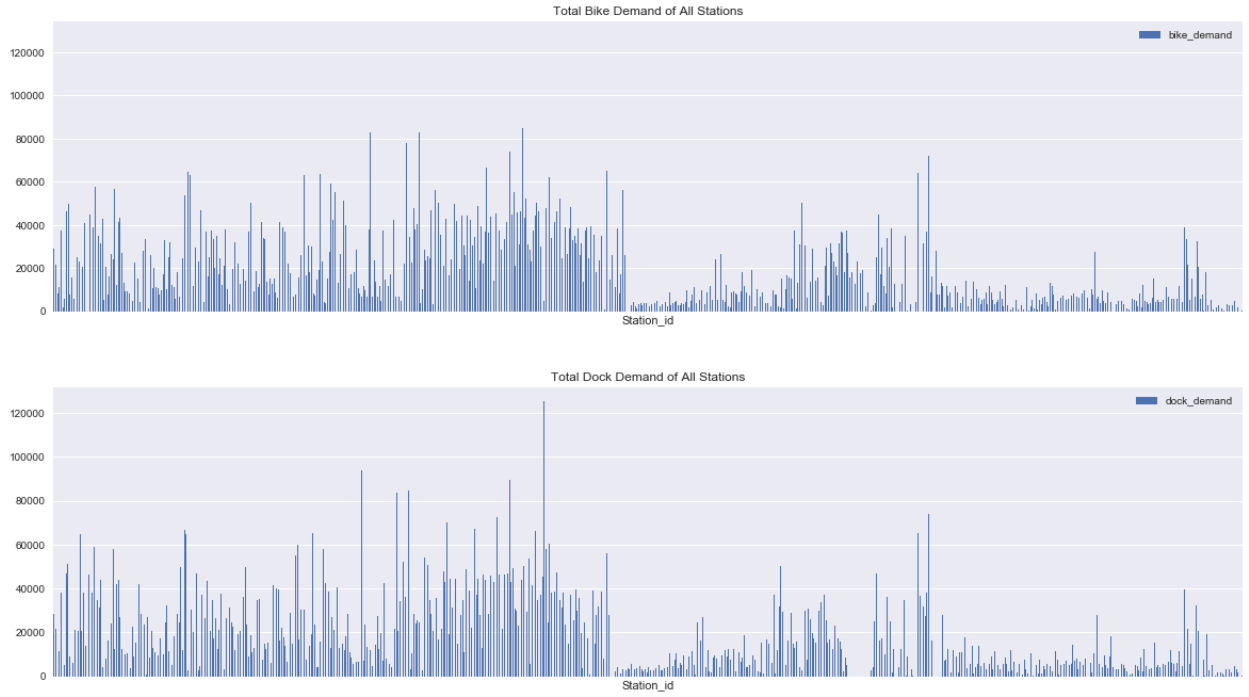
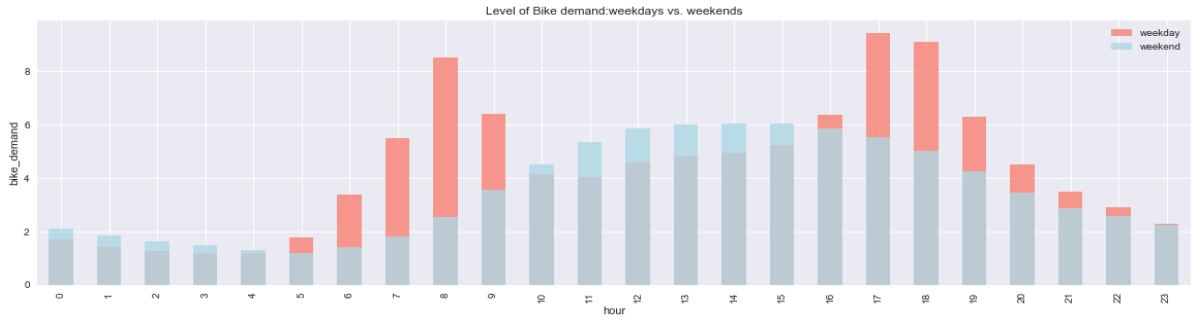


Figure 4. Total Bike/Dock Demand of All Stations

Demand Overview

The three graphs in Figure 5 summaries the level of bike demand for 24 hours in different stations. The first graph, a stacked bar chart is plotted for 24 hours against bike demand. The stacked bar chart shows the distinction for the bike demand between weekday and weekday. It can be seen for weekdays the demand is high between 7.00 am and 10.00 am, later during evening between 4.00 pm and 7.00 pm. But, for the weekends the demand for bikes can be seen increasing between 10.00 am and 12.00 pm, during later noon the demand is constant but, high compared to the evenings. The line graph in second graph show the all the days in the week in relevance to the hourly bicycle demand. It can be seen all the weekdays (0-Monday to 4-Friday) have same pattern and (5-Saturday and 6-Sunday) have the same pattern. The third chart in Figure 5 shows the line graph is plotted between 24 hours and the bike demand, each line represents a day it can be seen the demand for bike is uncommonly high for few days but, during the rest normal days the demand for bike seems to be constantly below 20. During the early morning hours between 12.00 am - 6.00 am the demand for bikes can be found lowest.



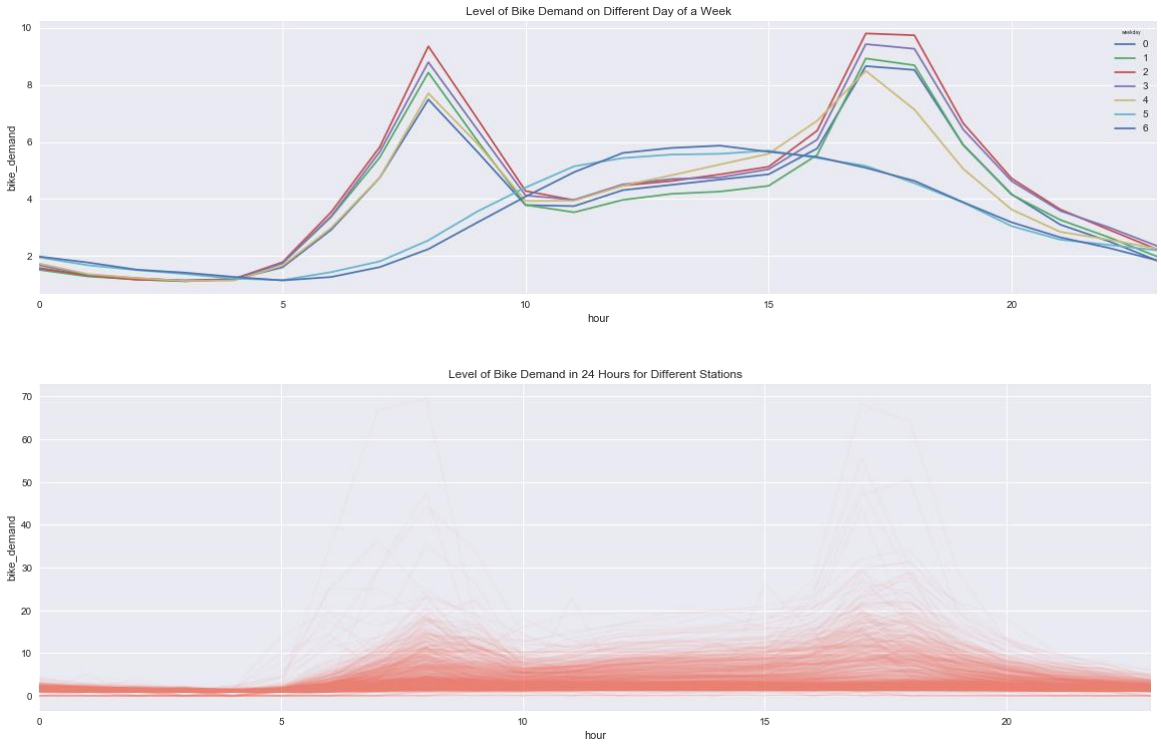
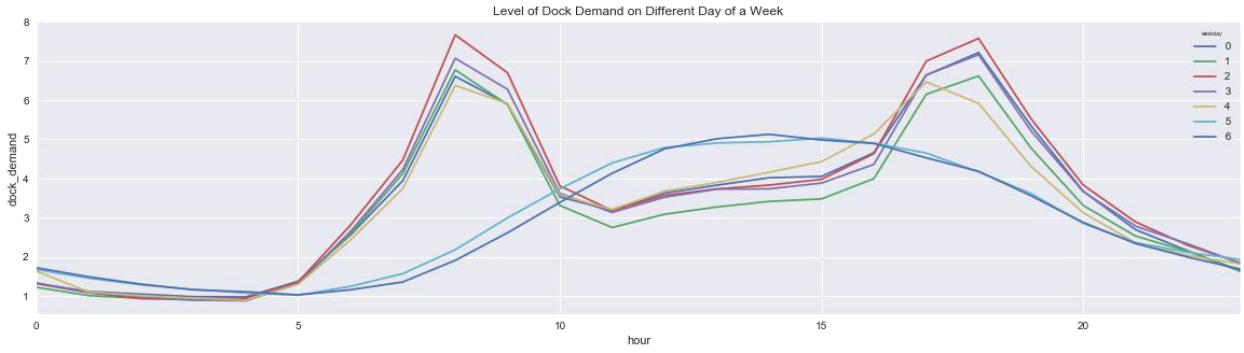


Figure 5. Level of Bike Demand in 24 hours for Different Stations

Figure 6 expresses the dock demand in 24 hours group by weekday and weekend in different charts. The first line chart shows the demand in each day in one week. Each line refers to one day in week. It is obviously that the lines mainly have two shapes classified by weekday and weekend. To be more clear, a bar chart is shown to indicate the difference between weekday and weekend. Grey part stands for the demand where weekday and weekend overlap. Red and blue parts refer to weekday surpass weekend and weekend surpass weekday respectively. In 5am-10am and 17pm-21pm, the two rush hours in weekday, the demand for dock are reasonable much higher because of the travel to work and home. The third chart in Figure 6 shows the line graph is plotted between 24 hours and the dock demand, it can be seen the demand for dock is rapidly increasing during the peak time periods i.e. between 6.00 am-10 am and 4.00pm-7.00pm but, during the normal hours the demand for docks seems to be constantly below 20. During the early morning hours between 12.00 am - 6.00 am the demand for docks can be found lowest.



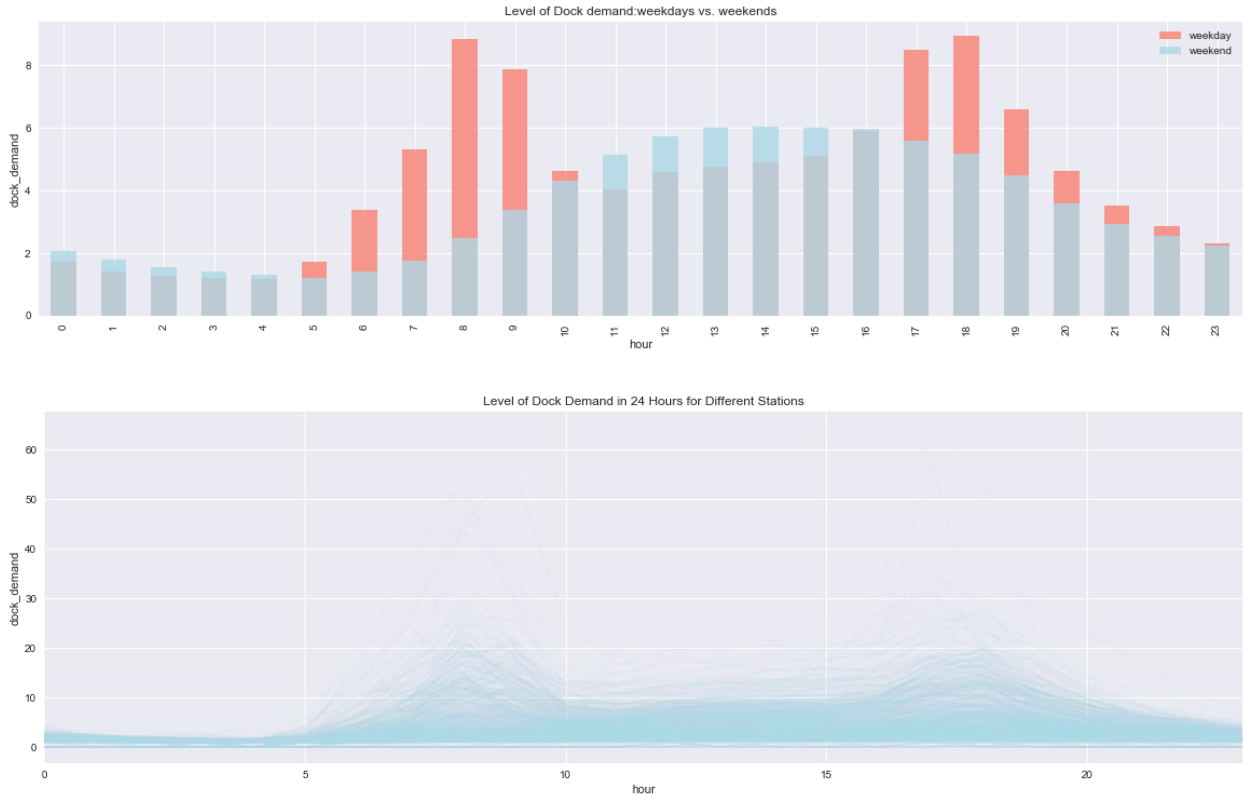


Figure 6. Level of Dock Demand in 24 hours for Different Stations

Availability

The availability of bikes/docks is plotted in 24 hours. Each line in the Figure 7 and Figure 8 indicates the daily changing pattern of the available bikes and docks in each station. The x-axis shows the hour from 0 to 23 and the y-axis shows the number of available bikes/docks in a given station.

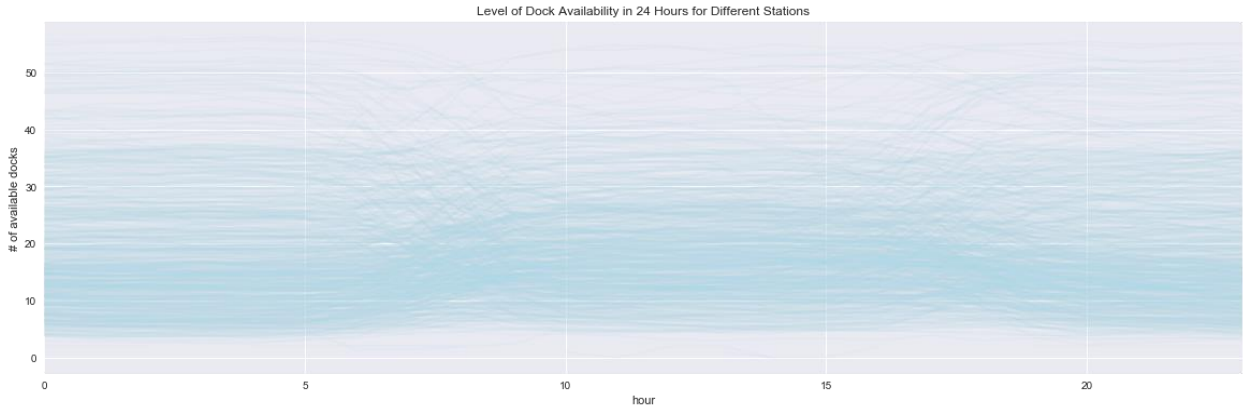


Figure 7. Level of Bike Availability in 24 hours for Different Stations

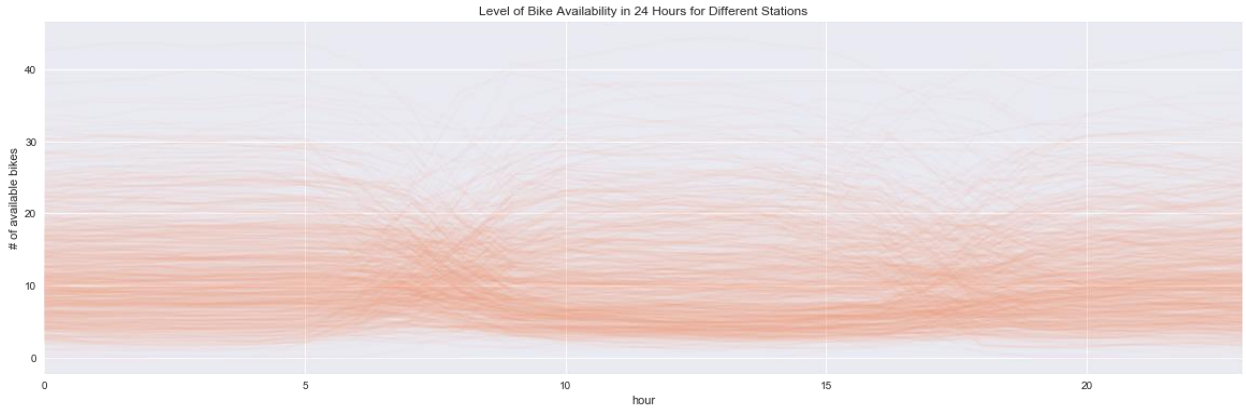


Figure 8. Level of Dock Availability in 24 hours for Different Stations

5. Method Description

Based on daily changing pattern of different stations (shown in Figure 5, Figure 6, Figure 7 and Figure 8), it is high likely that some stations might have strong correlation with each other for showing similar or reversed changing patterns. Therefore, correlation and clustering analysis is conducted before the demand and availability prediction. The followings in this section would mainly focus on analysis of bikes availability. Same procedures could be applied on dock availability and demand data.

Correlation Analysis

The correlation between each pair of stations is measured by correlation coefficient. And the result is visualized with heatmap (shown in Figure 9). The darker the color is, the stronger the correlation is. The figure does not present the correlation for all the station; however, it is observed that a large amount of stations are either positively correlated or negatively correlated. For each station, we consider those having correlation coefficient larger than 0.9 as strongly correlated, all of which is recorded in a form of dictionary (shown in Figure 10). The data of all stations that are in this dictionary would be involved in the future prediction.

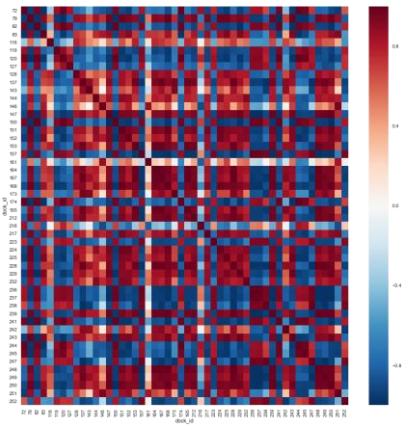


Figure 9. Correlation Heatmap among Stations. (due to the size, not all stations are presented)

```
{72: Int64Index([  72,   79,   82,   83,  120,  144,  150,  152,  174,  216,
             ...
            3411, 3412, 3413, 3421, 3422, 3423, 3430, 3445, 3449, 3454],
           dtype='int64', name='dock_id', length=226),
 79: Int64Index([  72,   79,   82,   83,  119,  120,  144,  147,  150,  151,
             ...
            3430, 3434, 3438, 3440, 3445, 3449, 3452, 3454, 3461, 3462],
           dtype='int64', name='dock_id', length=416),
 82: Int64Index([  72,   79,   82,   83,  120,  127,  144,  147,  150,  152,
             ...
            3423, 3424, 3427, 3430, 3434, 3440, 3445, 3449, 3454, 3461],
           dtype='int64', name='dock_id', length=361),
 83: Int64Index([  72,   79,   82,   83,  120,  144,  150,  152,  174,  216,
             ...
            3411, 3412, 3413, 3421, 3422, 3423, 3430, 3445, 3449, 3454],
           dtype='int64', name='dock_id', length=235),
```

Figure 10. Dictionary of Strong Correlated Stations.

`Given the result of correlation analysis, we plot the 24-hour availability of bikes that have high correlation with the given station. For example, the daily changing pattern of all stations that have high correlation with station #72 and #79 is demonstrated in Figure 11.

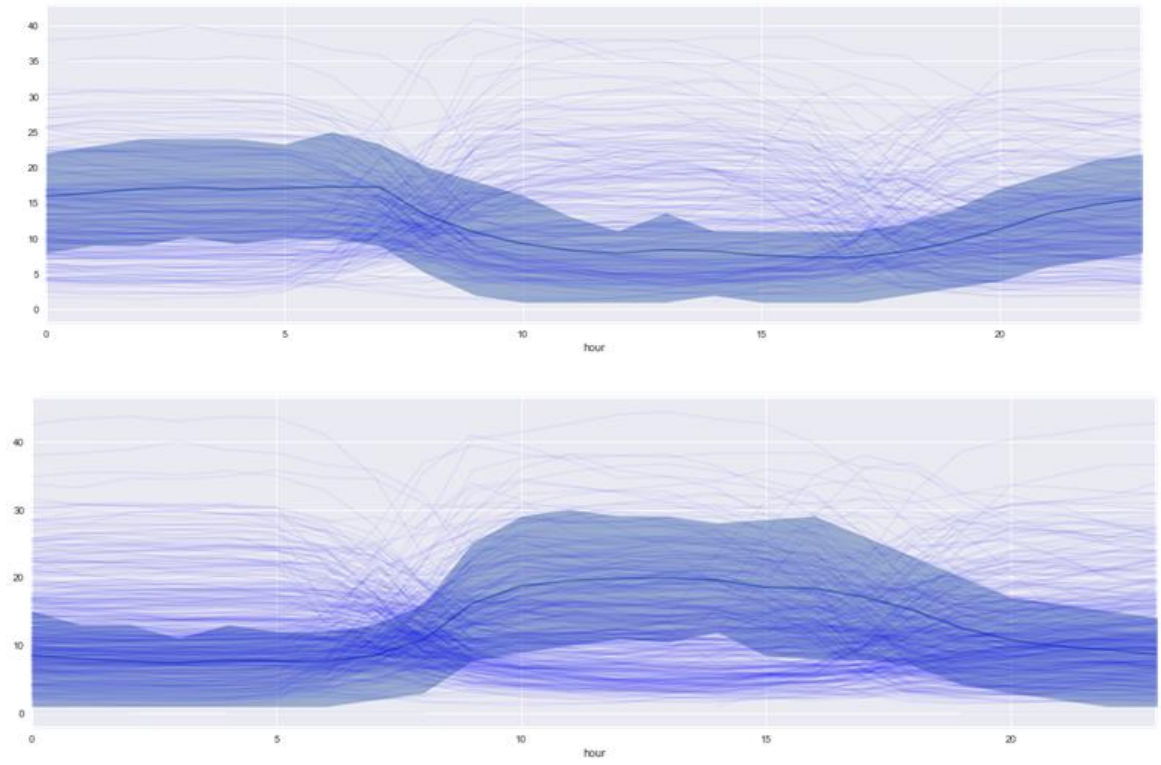


Figure 11. Daily changing pattern of all high correlated station of Station #72 and Station #79. (The figure on the top is for station #72, and the other is for station #79. The darker line in the middle shows the mean of all these correlated stations and the shadow indicate the band between the 25% and 75% confidence interval)

Clustering Analysis

Based on Figure 11, it is observed that not all correlated stations have the same changing pattern. It is worth investigation that what the difference will be if we only select out those stations that have same changing patterns with the given station.

In this project, k-means clustering is applied. However, instead of using Euclidean distance, Dynamic Time Warping (DTW) is prefered for time series data clustering. For time series data, the Euclidean distances between alignments are then much less susceptible to pessimistic similarity measurements due to distortion in the time axis. Dynamic time warping finds the optimal non-linear alignment between two-time series and it is quadratic in the length of the time series used. A python code for DTW is provided by Alex Minnaar (http://alexminnaar.com/time-series-classification-and-clustering-with-python.html). A few modifications is made to fit his code to python coding environment. The result of clustering is recorded as dictionary as what we did in correlation analysis. All stations that in the same cluster are demonstrated in Figure 12.

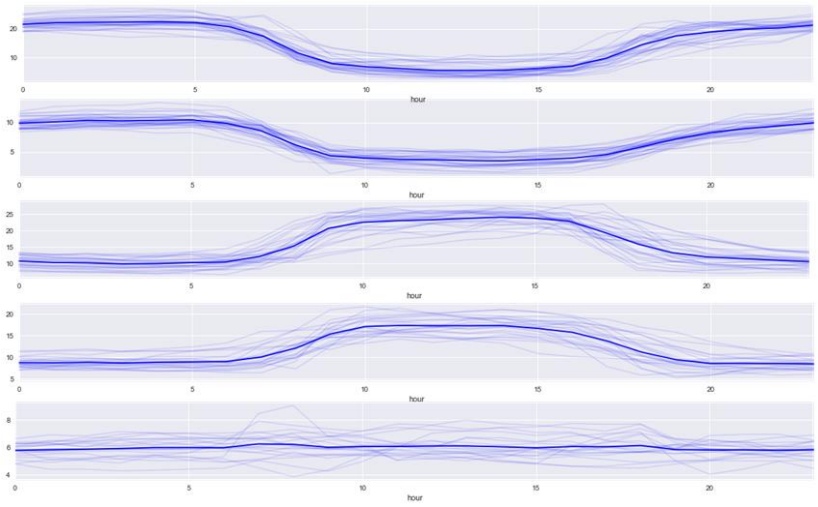


Figure 12. Clustering result (due to pages, not all clusters are shown)

Prediction

For a given time point (hour in 24), it is possible to predict the number of available bikes in a given station by using the availability level of previous hours. Thus, the feature data **X** and response data **y** is defined in the Table 5-Table 7 below. It is expected that the prediction accuracy could be improved by using the information from the high correlated stations or those in the same cluster.

Table 5. Structure of Data for Prediction Using the Information of the Target Station

Index	X			y
date	Hour	A(t-1)	A(t-2)	A(t)
		Available bikes in the given station at time (t-1) and (t-2)		

Table 6. Structure of Data for Prediction Using the Information of Both the Target Station and the High Correlated Stations

Index	X							y
date	Hour	A(t-1)	A(t-2)	$A_1(t-1)$	$A_2(t-1)$	.....	$A_n(t-1)$	A(t)
				Available bikes in the correlated stations at time (t-1)				

Table 7. Structure of Data for Prediction Using the Information of Both the Target Station and the Stations in the Same Cluster

Index	X							y
date	Hour	A(t-1)	A(t-2)	B1(t-1)	B2(t-1)	.....	Bn(t-1)	A(t)
				Available bikes in the stations that are in the same cluster at time (t-1)				

Different machine learning methods, including linear regression (LR), support vector regression (SVR), decision tree regression, random forest regression and neural network, were applied in analysis for prediction.

Linear regression is a basic and most commonly used statistical analysis method for predictive analysis. There is a linear approach for modeling the relationship between the dependent variable i.e. y also known as response variable and one or more explanatory variables represented by X also as regressor variables. SVR, works similar to SVM's as described in class, but is adapted to handle regression. It attempts to approximate the value of a continuous variable by using a loss function that is insensitive to the error. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-sample of the dataset. To improve the predictive accuracy and control over-fitting, averaging is used.

6. Prediction Result

Based on the methods illustrated in the previous section, the result by using different prediction method is shown as follows. It is worth mentioning that the prediction of demand did not include the information of correlated stations or stations in the same cluster. The feature is the given day of the week and the hour on that day. As a result, the prediction error is comparatively larger than those in availability prediction.

Demand Prediction

For bike and dock demand prediction, the best fitting method are accord which is the Decision Tree Regressor. It has the best performance with largest R-square 0.9352/0.9489 and smallest MSE 2076.1104/1677.8532. Otherwise the method Linear Regression has worst performance both for bike and dock with lowest R-square 0.03/0.0309 and largest MSE 90135.8693/80647.5477. Thus, Decision Tree is most recommended method for similar problem prediction.

Table 8. Bike Demand Prediction Comparison

	Linear Regression	SVR	Decision Tree Regressor	Random Forest Regressor	Neural Network
R-square	0.0300	-0.0720	0.0399	0.9352	0.2053
MSE	90135.8693	34333.0470	31729.5360	2076.1104	25451.6440

Table 9. Dock Demand Prediction Comparison

	Linear Regression	SVR	Decision Tree Regressor	Random Forest Regressor	Neural Network
R-square	0.0309	-0.0784	0.0414	0.9489	0.2196
MSE	80647.5477	35425.9728	31392.8499	1677.8532	25637.6504

Availability Prediction

Figure 13. shows the comparison between observed demand and predicted demand. The red line and blue line refer to predicted and observed respectively. Six single days in Station 72 are drawn to indicate the differences. We can see that the period which has very obvious differences are mostly concentrated on 10-20 hours. For the prediction results, correlation and cluster are considered to combine with basic and apply to methods separately. By finding the lowest MSE and highest R-square value, the basic and correlation by Random Forest Regressor is proved to be best fitting with MSE 15.67 and R-square value 0.6152.

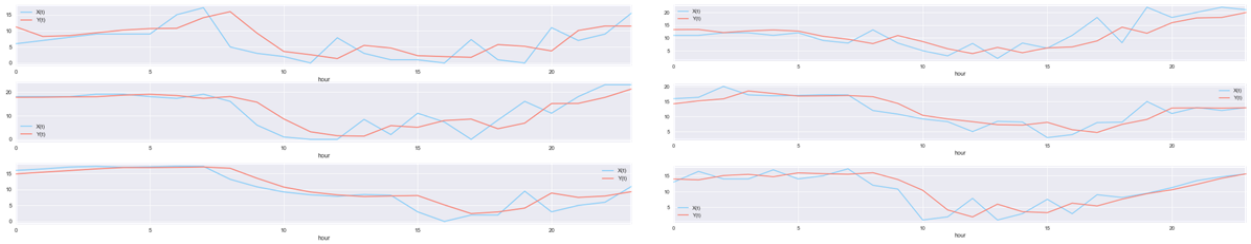


Figure 13. Observed Availability vs. Predicted Availability (The plot shows the prediction result of six random days for station #72. The blue line is the observed value and the red line the predicted value)

Table 10. Availability Prediction Comparison

		Linear Regression	SVR	Decision Tree Regressor	Random Forest Regressor	Neural Network
Basic	MSE	18.45	19.58	20.96	16.73	16.59
	R-square	0.5468	0.5191	0.4852	0.5892	0.5923
Basic+corr:	MSE	17.77	18.90	27.00	15.67	18.25
	R-square	0.5634	0.5357	0.3368	0.6152	0.5517
Basic+cluster	MSE	17.40	18.40	27.39	17.85	19.69
	R-square	0.5724	0.5480	0.3274	0.5615	0.5165

7. Conclusion

This project is aimed at predicting the bike/dock availability and demand for better balancing the number of bicycles among the stations. By using descriptive visualization, different daily changing patterns were observed for both bicycle demand and availability. In order to improve the prediction accuracy, for each station, the stations corresponding high mutual correlation or the stations with the same cluster pattern were investigated using different statistical methods including decision tree regression, neural network,

random forest regressor, SVR and linear regression. Through analysis in applying various machine learning methods, it resulted that random forest regression is the most effective way in predicting the bike demand and availability.

However, the prediction accuracy for demand data is not satisfying due to the lack of information (e.g. weather, traffic, holiday, etc) that might have significant influence on it. Also, the information of correlated stations is not included because of the limited time. In the future, these factors, as well as the correlated stations, should be involved in the prediction. In addition, the correlation analysis does not include the spatial information of stations. So spatial-temporal analysis would be another direction in the future research to investigate the correlation among the stations.