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

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

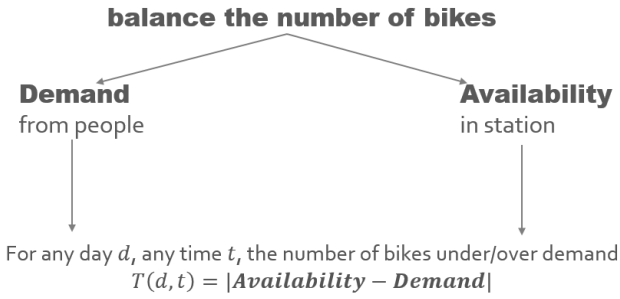
Santander: London

Citi Bikes-NYC

Problem description

For our 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 our

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.



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