Infer NYC Citibike usage

Contents

1 Abstract

NYC Citibike launched in September 2013 and becomes new city symbol of New York City similar to Empire State Building, since it becomes popular choice for not only daily commuting but also for tourists to walk around New York. Therefore the usage of Citibike possibly reflects hidden social pattern of how people move inside city. To address the issue, we utilize modern visualization methods to give an overview of Citibike usage. To provide Citibike company a potential guide on how to manually balance available bikes in stations, we perform unsupervised machine learning to find out the distinctive type of Citibike station in terms of usage in different time frame. Furthermore, we run time-series analysis to predict the usage of several hot-spot of stations, e.g. Citibike stations around Penn Station, Central Park, etc.

2 Data

In light of data format consistency, 21 months of Citibike trip data are chosen, i.e. from 2015-01 to most recent 2016-09.

Data source: https://s3.amazonaws.com/tripdata/index.html

Citibike data-set has following attributes in terms of trip:

- Trip duration time (seconds)
- Trip start and end time
- Trip start and end station, together with station ID and longitude/latitude information
- Bike ID

In terms of the customer information of trip:

- User type (one-time customer / annual member)
- Gender (0=unknown, 1=male, 2=female)
- Year of birth

3 Planned methods and expected results

3.1 Data I/O

Since the analysis is involved in large volume of csv files, `readr` (https://github.com/hadley/readr) and `dplyr` (https://github.com/hadley/dplyr) are used to achieve efficient data I/O.

3.2 Overivew of Citibike usage within 24 hours

Playing 'God-view' game to identify possible 'heart-beat' of New York City. This part is data visualization to show the temporal and dynamic pattern of citibike trip within 24 hours. In particular there might be some hot-spot emerge.

There are following 2 aspects to furthermore dig the data to provide possible interesting findings:

3.2.1 24-hr trip of annual member v.s. one-time custormer

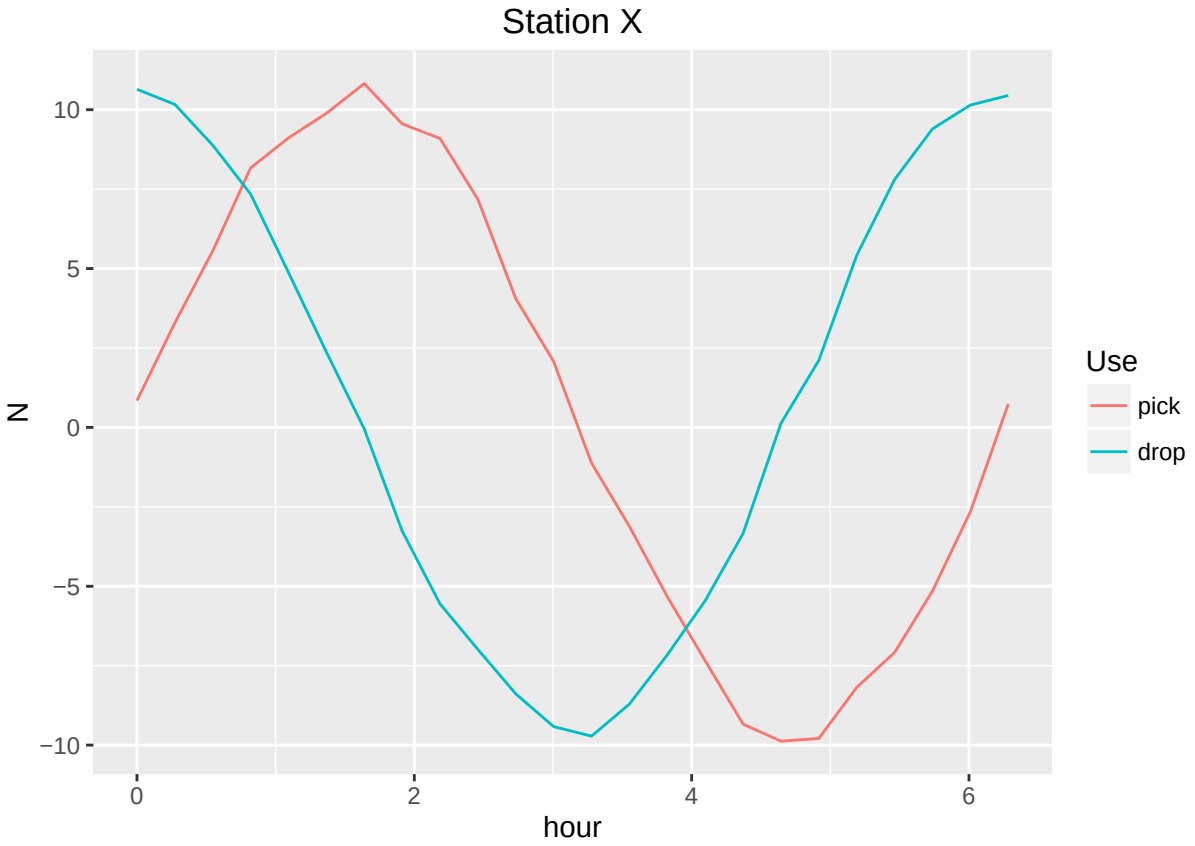
As annual members are more likely to choose Citibike for daily commuting, with this **hypothesis** in mind, is there any possible difference exist in preference of the two types of users?

3.2.2 24-hr trip in weekdays v.s. weekends for annual member

How about annual member use Citibike in weekends? Would it be silent, or more likely as tourist to travel around city?

3.3 Identify subpopulation of Citibike stations

Station would vary on bike usage in different hours per day, shown as the below figure generated by dummy data. With this **hypothesis** in mind, it is possible to identify sub-population of Citibike stations by performing unsupervised **clustering** methods, e.g. some stations are active during rush hour and silent during working hour, while some other stations have opposite pattern.



These outcomes would be helpful for identifying the hot-spot of stations and thus Citibike companies can take care of it, for example, manually balancing the specific stations before their own 'rush-hour' come.

3.4 Predict bike usage of specific station

Once the activity patterns of stations are learnt, the next question is how to take care of stations before their own 'rush-hour' come, e.g. how many bikes needed to be manually put? Therefore we run **time-series analysis** to quantitatively predict the possible needs in next few hours of specific station.

3.5 Optional: network analysis of trip flow between stations

As the Citibike trip data provides the start station and end station information, it is not surprising we can run network analysis to have an idea of how trip flows between stations. It depends on the content of coming lecture, thus this part is optional.

4 Summary

We first give overview of citibike trip around New York City by modern data visualization methods. Then we cluster the temporal usage of stations to identify subgroups with different 'active-hour'. Finally we run time-series analysis to quantitatively predict the activity of certain station. In sum, we provide a real-time strategy to keep the stations balanced.