Classifying New York City Subway Stations

Ryan Albertson
Coursera – Applied Data Science Capstone
December 2019

INTRODUCTION

Every day, roughly 5 million people ride the New York City subway system, which expands 665 miles across the 5 boroughs of the city. The system is currently comprised of 473 subway stations. I will be employing data from the *Foursquare API* to learn more about each of these stations. Specifically, I will pull venue data to understand which types of establishments are prominent around each station. From this information, the traffic of each station can be predicted. For example, a station with primarily university venues nearby will be busier on weekdays and not so much on the weekends. Whereas an area that has many entertainment venues will likely experience a traffic spike during the weekends. I will also use k-means clustering to see if the clusters have any correlation to the venue category types of each station.

Data

The dataset was taken from *NYC Open Data* (https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49). The only features needed from the dataset are the subway station names, and the matching geographical coordinates. There is a total of 473 subway stations. From the Foursquare API, the names of the 10 main categories of venues were first obtained. Then all venues within a given radius of each subway station were retrieved. With this data, the venues surrounding each station will be ranked from 1st most common to 5th most common based on their count. These classifications will be saved to the location of every station.