IBM Data Science Professional Certificate

Applied Data Science Capstone

New York City Metropolitan Transportation Authority (MTA):

Find best venues categories near metro entrances

WEEK 1

Eleonora Picca

02 - 2019

Contents

Chapter 1

Business Problem and stakeholders

1.1 Business Problem and Stakeholders

New York City is the third most congested city in the world in terms of traffic and the second worst in the United States. In 2017, New York drivers averaged 91 peak hours stuck in traffic and spent 13 percent of their time sitting in congestion, with 11 percent of that being attributed to daytime traffic.

The reasons behind traffic problems are various, but one of them is that lots of New Yorkers choose to go around city by car.

NYC Metropolitan Transportation Authority (MTA) needs a strategy to alleviating traffic. Furthermore, if more people use public transportation, gain would raise and it could be spent to pay for a major overhaul of the aging subway.

One way to do this is to make New Yorkers use the subway instead of other road-kind transportation.

One solution could be to make agreements with venues near metro entrances in order to give special offers to New Yorkers who use the subway. What is proposed and presented in this report is how to choose the venues categories:

- Venues has to be near the metro entrance, in this case we chose a radius of 500 m.
- Best venues are the one more popular, because they have more appeal.

Chapter 2

Data Sets

In this chapter will be described data sets used for the analysis.

2.1 Kaggle Metro Entrances

The first data set used is the list of metro entrance in NYC (that can be found here: https://www.kaggle.com/parichartpanichpol/nyc-subway-entrances).

Data set is composed by the following columns in tab. (2.1):

ColumnName	*Description*	*Example*
OBJECTID	Unique ID	1734
URL	URL to the subway line information	http://web.mta.info/nyct/service
NAME	Entrance name	Birchall Ave and Sagamore St at NW corner
the geom	Longitude and Latitude	POINT (-73.86835600032798 40.84916900104506)
LINE	Line number	2-5

Table 2.1: Original Kaggle data set.

These information will be used to identify the location of the entrances. Latitudes and longitudes will be used for calling the Foursquare APIs.

- For limited budget and computational resources, only the first 400 entrances will be used for the analysis.
- Data transformation will be performed and latitudes and longitudes will have dedicated columns. Some rows don't have coordinates information, so they will be dropped.

The columns used will be in table (2.2):

Original Column	*Final Column*
NAME	Subway Entrance
the geom	Latitude
the geom	Longitude

Table 2.2: Initial and final columns of Kaggle data set.

2.2 Foursquare APIs

In addition to the subway entrances locations, we will use the Foursquare explore API in order to extract a list of recommended venues near the given location. The relevant parameters of the API are:

- Radius: Radius to search within, in meters. We will use 500 m.
- Limit: Number of results to return. We will use 100.
- Latitude and Longitude: We will used the ones in the Kaggle dataset.

The API returns a list of venues with its category. In this way we will have a list of 100 recommended venues within 500 meters for each of the 400 subway entrances and we will find the 10 most common categories per each entrance. We will use these venues to create clusters of entrances in order to identify the most common venues categories.