1.Introduction:

1.1 Background

New York City also called as NYC is the most populous city in the United states with an estimated population of 8,398,748 in 2018.The Government has provided free public WIFI access across the city but there is a huge traffic experiencing due to the usage, Link NYC – Citybridge is one of the popular internet service provider has been asked to resolve the issue by installing more hotspot in the NYC.

Before making an investment for this, the Service Provider wanted to know the areas which are experiencing huge network traffic so that more hotspot can be installed in those Areas. The Objective of this Project is to cluster the locality based on the Internet Usage so that Link NYC – Citybridge can use the data for the installation purpose.

1.2 Problem

Data that might contribute in determining the Network Traffic include the Geospatial Data with the network provide Information. This Project aim in providing the areas where more hotspot need to be installed to control the network traffic.

1.3 Interest

Obviously Link NYC – Citybridge in interested in the Data because they are going to install the Hotspot by management the cost effectively ensuring that appropriate area has been identified experiencing huge network and thus reducing resource wastage.

We are trying to find an answer for below questions:

Does free public WiFi tend to cluster around certain (more affluent) areas?
Who are the free WiFi providers, and where do they do it?

2. Data acquisition and cleaning

2.1 Data Source

Data is provided by the NYC Department of Information Technology and Telecommunications. The dataset consists of records for every public WIFI hotspot (ones provided by or in partnership with the city) in New York City. It contains over 2500 records overall.

ColumnName
OBJECTID
Borough
Type
Provider
Name
Location
Latitude
Longitude

X
Y
Location_T
Remarks
City
SSID
SourceID
Activated
BoroCode
BoroName
NTACode
NTAName
CounDist
Postcode
BoroCD
CT2010
BCTCB2010
BIN
BBL
DOITT_ID
Location (Lat, Long)

2.2 Data cleaning

All the above field might not be required for the Analysis. EDA will be performed on the Entire Dataset and will come up with the required variable which are important and relevant. The Dataset available will be used with the Foursquare location Data.

First issue was the missing value in the Dataset, all the missing value record was dropped from the Dataset also all the exact duplicate records was removed from the Dataset. All these steps gave us a clean Data.

2.2 Feature Selection

Total Records: 2500

7 Features which were important for us are the below:

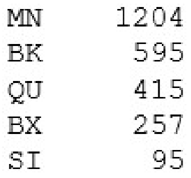
- Borough
- Type
- Provider
- Name
- Location
- Latitude
- Longitude

A new Dataset was Build by taking only the required Features and named as public_wifi_cleaned.

3. Exploratory Data Analysis

Looking at the Neighborhood Tabular Areas (a type of neighborhood area designation used by the City of New York) assigned to the various WiFi nodes, we need to identify which Borough having the most WIFI hotspot Installed

Table 1.1



```
MN    1204
BK     595
QU     415
BX     257
SI      95
```

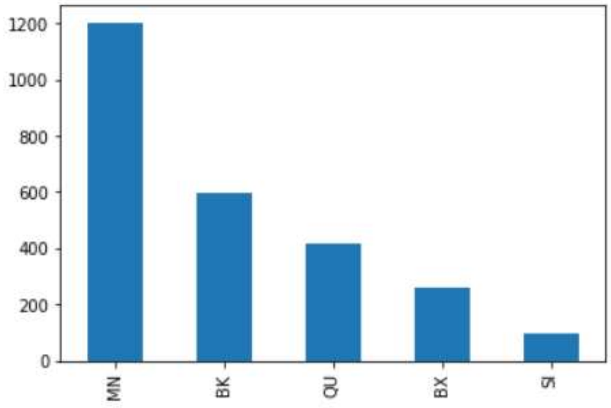
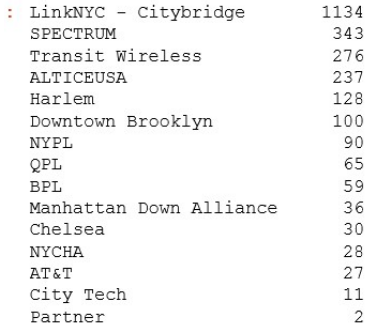


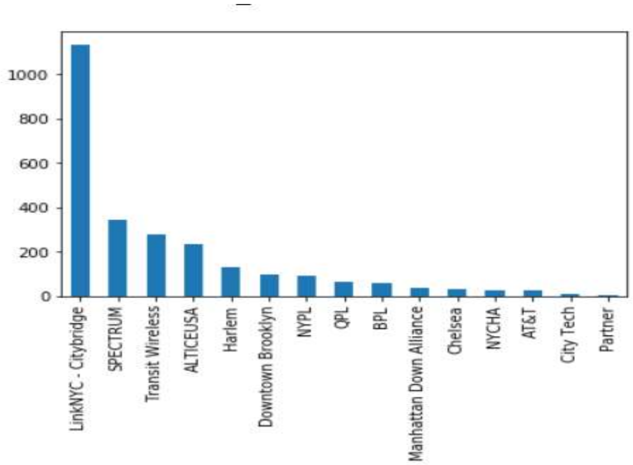
Table above show that Manhattan is having the Highest Number of WIFI Hotspot Installed.

Table 1.2

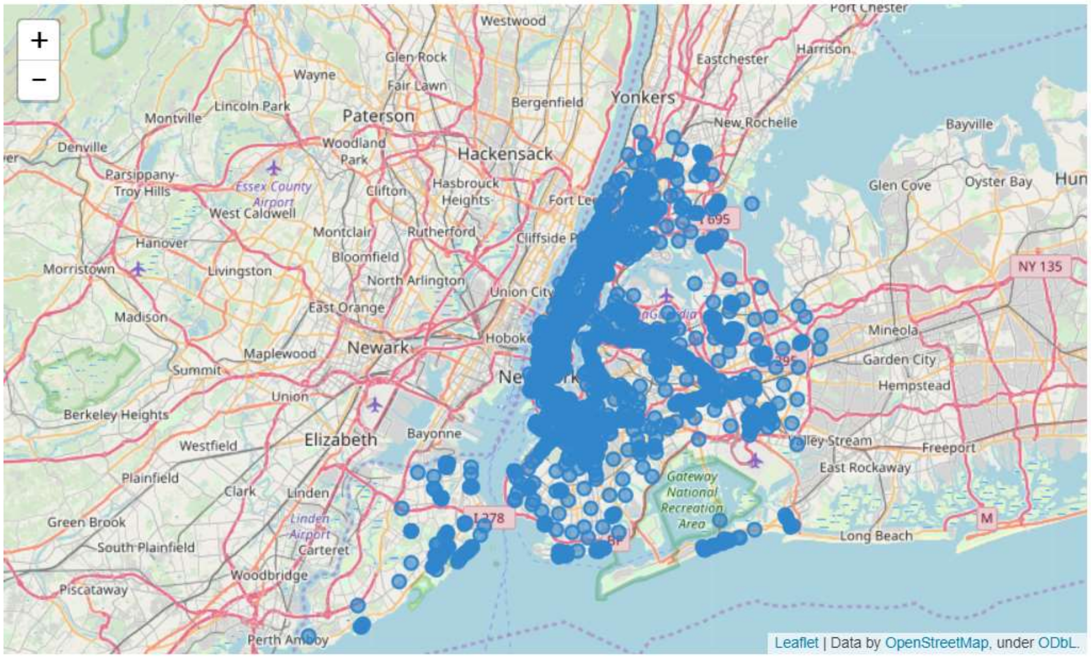


```
: LinkNYC - Citybridge       1134
  SPECTRUM                    343
  Transit Wireless            276
  ALTICEUSA                   237
  Harlem                      128
  Downtown Brooklyn           100
  NYPL                         90
  QPL                          65
  BPL                          59
  Manhattan Down Alliance      36
  Chelsea                      30
  NYCHA                        28
  AT&T                         27
  City Tech                    11
  Partner                       2
```

If we are looking for the largest service provider in NYC ,above Table clearly show that Link NYC – Citybridge is in the number one position followed by SPECTRUM and Transit Position which maintain the number 2 and 3 position.

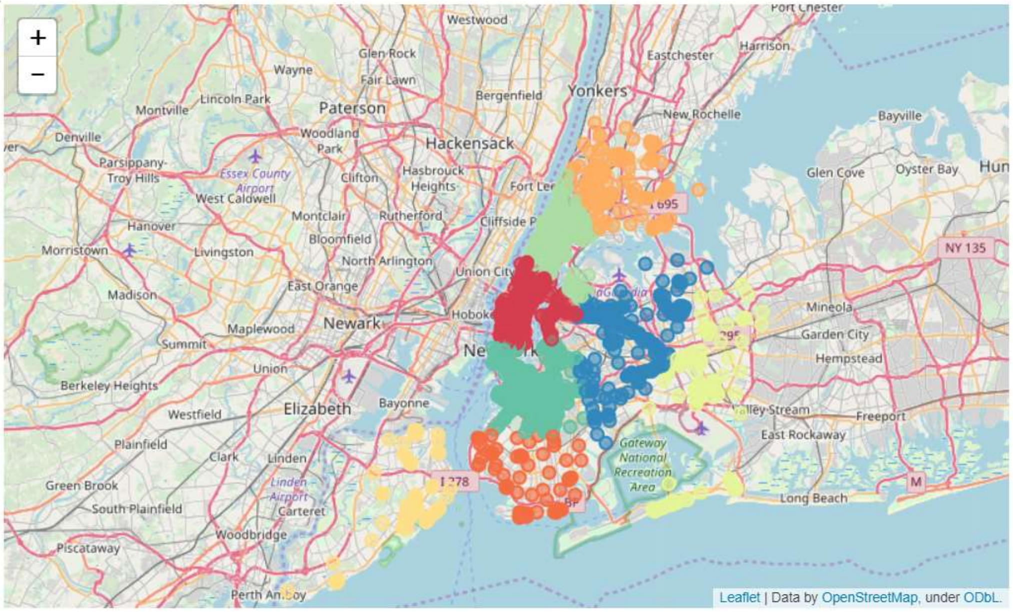


We can get a better feel for what the distribution is with a map:



The WiFi hotspots are (logically) concentrated alongside certain arterial roadways. What stands out is how strong that effect is. 3rd Avenue and 8th Avenue in Manhattan both "stand out" in terms of their WiFi offerings, as do certain other roadways, like Queens Boulevard.

Interestingly enough, KMeans clustering (a decent default) doesn't stop on the borough boundaries quite like I thought it would, instead pushing into "uptownsy" (and popular!) parts of Brooklyn and Queens close to Manhattan as well. This is evidence that unlike certain other projects, public WiFi "pushes out" past Manhattan and into the outer boroughs as well, at least a little.



Leaflet | Data by OpenStreetMap, under ODbL.

4. Cluster Model

There are many models for **clustering** out there. In this we will be using the model that is considered one of the simplest models amongst them. Despite its simplicity, the **K-means** is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from **unlabeled data**. In this notebook, you will learn how to use k-Means for customer segmentation.

4.1 K mean Clustering

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1. Data assignment step
2. Centroid Update step

5.Observation

This type of clustering gives great insights into the structure of a WiFi network in a city. For instance, there are 650 separate points in cluster 1, whereas 100 points exist in cluster 6.

This indicates that the geographic region marked by cluster 1 shows heavy WiFi traffic. On the other hand, a lower number of connections in cluster 6 indicates low WiFi traffic.K-Means Clustering in itself does not tell us why traffic for a specific cluster is high or low. For instance, it could be the case that cluster 6 has a high population density, but poor internet speeds result in fewer connections. However, this clustering algorithm provides a great starting point for further analysis—and makes it easier to gather additional information to determine why traffic density for one geographic cluster might be higher than another.

6.Conclusion

This example demonstrated how k-means clustering can be used with geographical data in order to visualize WiFi access points across New York City. In addition, we have also seen how k-means clustering can also indicate high- and low-density zones for WiFi access, and the potential insights that can be extracted from this regarding population, WiFi speed, among other factors.

Notebook:
https://github.com/kichu1990/Coursera-Data-Science /blob/master/NYC%2BPublic%2BWifi%2B.ipynb

Presentation :
https://github.com/kichu1990/Coursera-Data-Science-
/blob/master/NYC%20Public%20WIFI%20Upgradation.pptx