**Machine Learning helps me in my Venue hopping Tour Venture**

**Gurpratap Singh (**[**gpsingh07331@yahoo.com**](mailto:gpsingh07331@yahoo.com)**)**

**Date Published: 15th June 2019**

**Table of Contents:**

1. Introductory section

Problem Description

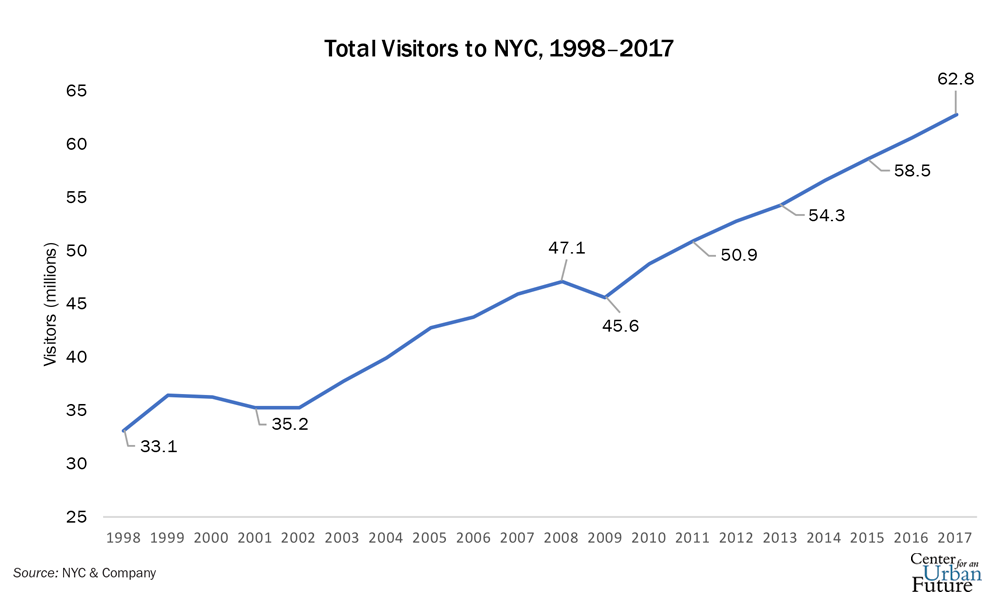
1. Methodology section

Data and Research and Approach

1. Exploratory Data Analysis
2. Discussion section
3. Result & Conclusion section
4. References
5. Acknowledgment
6. **Introduction: Problem Description.**

New York City comprises 5 boroughs sitting where the Hudson River meets the Atlantic Ocean. It is one of world’s top tourist spots due to its diversity in Arts, Architecture, Shopping and Dining.

New York City received an eighth consecutive annual record of approximately 62.8 million tourists in 2017, counting not just overnighters but anyone visiting for the day from over 50 miles away, including commuters. Overall the city welcomed 38 million visitors who stayed overnight of which 13.3 million were international in 2018

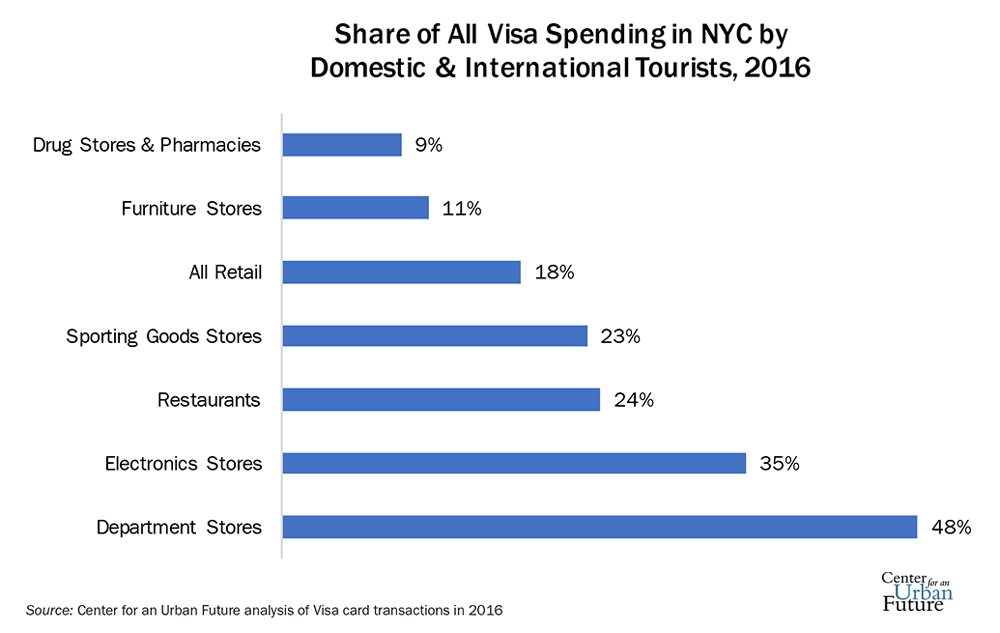


**Fig 1.**

[**https://www.google.com/search?client=firefox-b-1-d&q=New+York+City#**](https://www.google.com/search?client=firefox-b-1-d&q=New+York+City)

[**https://nycfuture.org/research/destination-new-york**](https://nycfuture.org/research/destination-new-york)

And the expenditure by these tourists have been mostly in Departmental stores and also restaurants. See Fig 2.



**Fig 2.**

Now imagine you are one of those tourists and you want to spend your time in NYC wisely. With size of NYC and so many choices available and gems of them hidden around and you want to cover the best of everything, how would you do it. How can you optimize your day so that you can experience the best of everything?

Here I offer you an optimized tour so you can spend your time wisely and get the best experience of whole of NYC in least amount of time.

Well at least let me start small and give you best of Eating, departmental shop and electronics experience across NYC in smallest possible time and allow you to select the section of NYC as a choice.

**Stakeholders:**

* **I, as the owner of Travel related online business.**
* **Tourists who have come to get relief from daily stress.**

Questions I need answer for so as to provide the above experience.

* How can divide the NYC into sections so as to get the location data for each of these sections?
* Where can I find the trusted data source for the best rated and liked Eating, Departmental and electronics shop location in NYC.
* How can I optimize the time taken to visit each of the top venue categories in each section of NYC?
* Which section should I visit first or should be the only section to visit?

1. **Methodology**

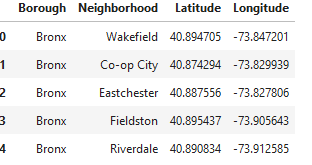
**Data**

For this project we need the following data:

* **New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.**

Data source: <https://geo.nyu.edu/catalog/nyu_2451_34572>

Description: This data set contains the required information. And we will use this data set to explore various neighborhoods of New York city.



* **Venues in each neighborhood of New York city with category, ratings and likes.**

Data source: Fousquare API

From the categories available,

<https://developer.foursquare.com/docs/resources/categories>

I will be looking for Food, Electronics Store and Department Store only.

To get Listing of venues in NYC

<https://developer.foursquare.com/docs/api/venues/search>

To get ratings and likes for the venues.

<https://developer.foursquare.com/docs/api/venues/details>

Description: By using this api we will get all the venues in each neighborhood. To keep things simple, if there is a tie, I will pick up one randomly.

* **GeoSpace data**

Data source: <https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm>

#### Or Use geopy library to get the latitude and longitude values of New York City

address = 'New York City, NY'

geolocator = Nominatim(user\_agent="ny\_explorer")

location = geolocator.geocode(address)

latitude = location.latitude

longitude = location.longitude

print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, longitude))

Description: By using this geo space data we will get the New York Borough boundaries that will help us visualize choropleth map.

* **Route Time Travel Data**

Data source: <https://docs.microsoft.com/en-us/bingmaps/rest-services/routes/calculate-a-route>

*Or* [*https://github.com/Project-OSRM/osrm-backend/blob/master/docs/http.md#responses*](https://github.com/Project-OSRM/osrm-backend/blob/master/docs/http.md#responses)

Description: A waypoint is a specified geographical location defined by longitude and latitude that is used for navigational purposes. The route includes information such as route instructions, travel duration, travel distance or transit information.

### Approach

### EDA – Collect Information, revise expectation or collect more data.

### Formal Modeling – Primal model answers question.

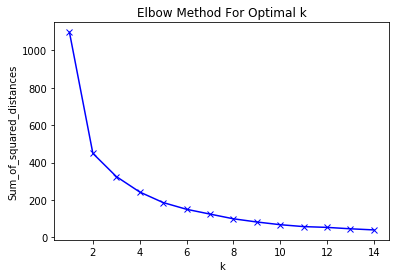
### Interpretation of analysis provides a specific & meaningful answer to the hypothesis.

### Sub-Steps:

* Collect the New York city data from <https://geo.nyu.edu/catalog/nyu_2451_34572>
* Validate the data obtained, look for any nulls or inconsistent data type.
* Using FourSquare API, we will find all venues for each neighborhood.
* Filter out all venues that are of type Indian Restaurant, Shopping and Electronics.
* Using rating=5 and above and likes > 50 for each above venue, we will sort that data.
* Get their Lat and Long coordinates.
* Use K-Means Clustering to create our own sections with top of each type in each section.

<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

* Elbow method was used to obtain optimal k value



* Folium map was used to visualize the spread of clusters.
* Using above Lat and Long and using Time Travel API, get the optimized route in each cluster.
* Bing API will be used to optimize the travel time

[http://dev.virtualearth.net/REST/v1/Routes?wayPoint.1={wayPpoint1}&viaWaypoint.2={viaWaypoint2}&waypoint.3={waypoint3}&wayPoint.n={waypointN}&heading={heading}&optimize={optimize}&avoid={avoid}&distanceBeforeFirstTurn={distanceBeforeFirstTurn}&routeAttributes={routeAttributes}&timeType={timeType}&dateTime={dateTime}&maxSolutions={maxSolutions}&tolerances={tolerances}&distanceUnit={distanceUnit}&key={BingMapsKey}](http://dev.virtualearth.net/REST/v1/Routes?wayPoint.1=%7BwayPpoint1%7D&viaWaypoint.2=%7BviaWaypoint2%7D&waypoint.3=%7Bwaypoint3%7D&wayPoint.n=%7BwaypointN%7D&heading=%7Bheading%7D&optimize=%7Boptimize%7D&avoid=%7Bavoid%7D&distanceBeforeFirstTurn=%7BdistanceBeforeFirstTurn%7D&routeAttributes=%7BrouteAttributes%7D&timeType=%7BtimeType%7D&dateTime=%7BdateTime%7D&maxSolutions=%7BmaxSolutions%7D&tolerances=%7Btolerances%7D&distanceUnit=%7BdistanceUnit%7D&key=%7BBingMapsKey%7D)

* Confirm we have plausible tour covering all 3 venue categories and have couple of options within NYC.
* We also use one hot encoding for clustering

**Feature Selection**

* For NYC dataset, we only need following

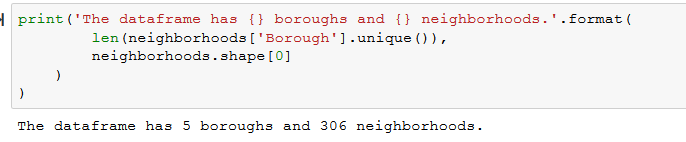
column\_names = ['Borough', 'Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude']

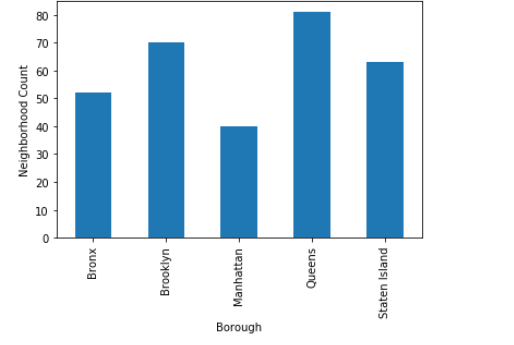
* For data obtained from Four square we need

columns = ['Borough', 'Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude', 'Category ID ', 'Venue', 'Venue Id', 'Venue Latitude', 'Venue Longitude', 'Venue Category']

1. **Exploratory Data Analysis**
   * Read NYC data from csv file.
   * Check for shape, distribution and top and tail of data.

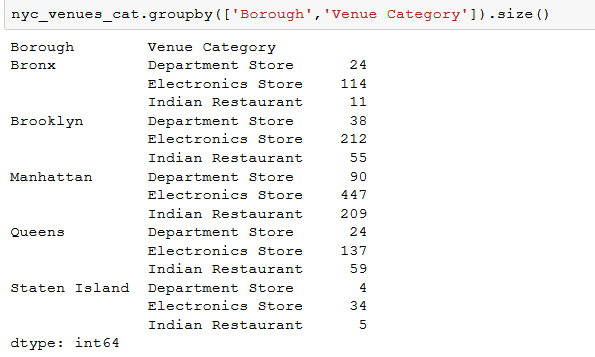
**NYC Data Set**





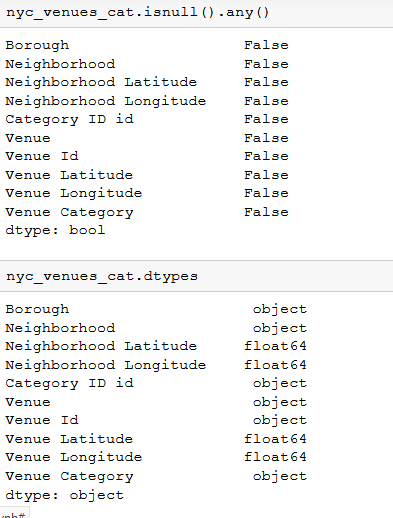
**Foursquare API output Summary**

* + Read Foursquare data using API call.
  + Check for shape, distribution and top and tail of data.

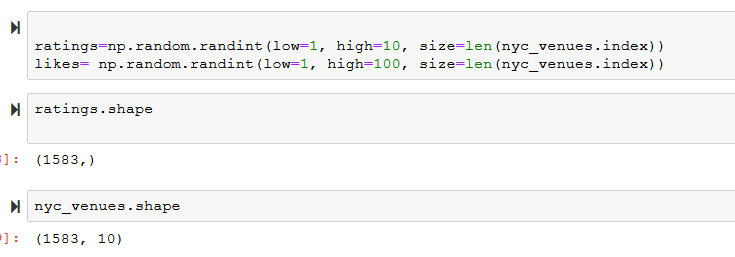


So, we have unbalanced distribution but good enough to create our section of NYC to create clusters.

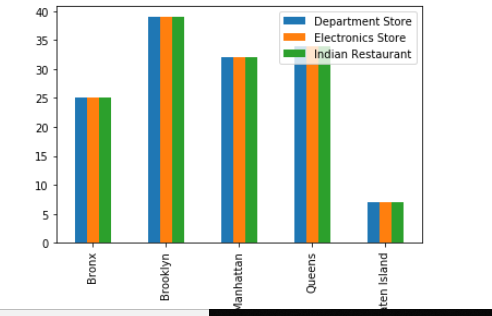
Simple data quality validation



**To give best experience, we are only interested in venues with good ratings and good number of users likes.**

**Had data missing issue for ratings and likes from Foursquare output so I am using mock up data**.

Distribution of Venue data with rating and likes.



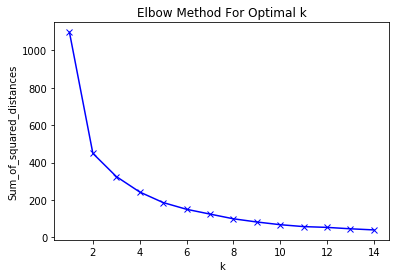
We have good data to form our clusters.

**Primary Model Creation**

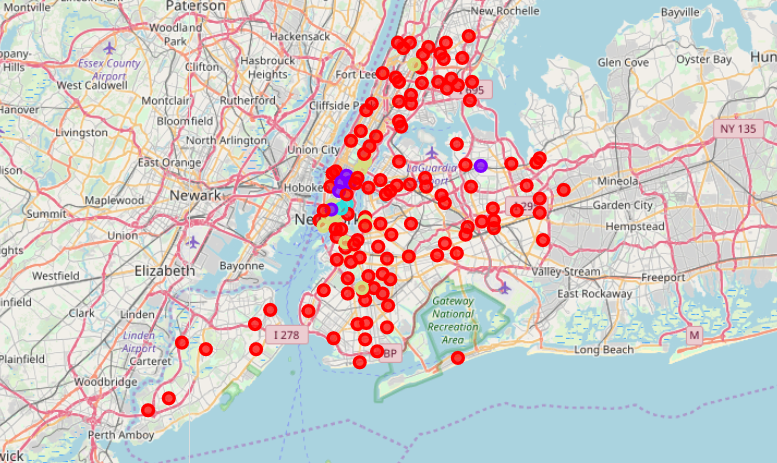
* We used only 3 categories to create our model. Inference is that is we are able to create good enough clusters with these 3 categories then we can add ore categories to create different cluster groups to provide unique experience based on the tourist’s taste.
* A cluster refers to a collection of data points aggregated together because of certain similarities. Since we want to classify the NYC into clusters having set of interested venue categories and we do not care about the boundaries of boroughs being crossed, we will use unsupervised clustering.
* Use K-Means Clustering to create our own sections with top of each type in each section.

<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

* Elbow method was used to obtain optimal k value

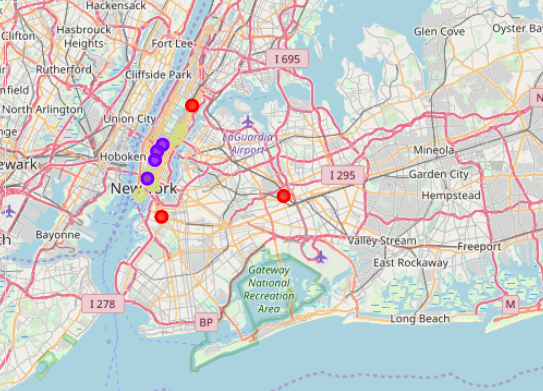


**We used unsupervised KMeans clustering onto our prepared data set to view the distribution of clusters across NYC.**



**Cluster 0 has the biggest spread. But we are interested in clusters which have at least one of each venue category.**

**We will retrieve the data based on above criteria and see the spread of the cluster in NYC.**



**We have few options but spread across NYC. We have good enough cluster spread to validate our concept of offering tourist venue hops spread across NYC.**

**We can add more categories to form new clusters at a later stage.**



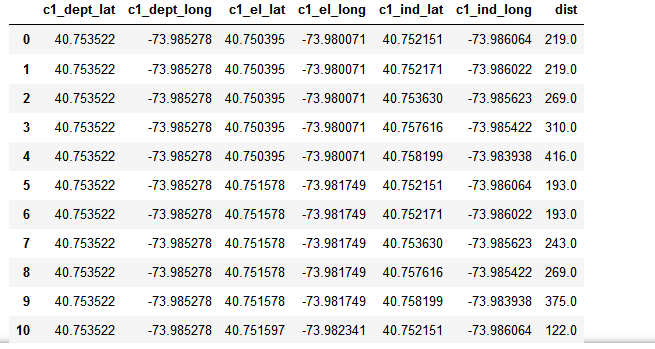
**Bing API for time travel optimization**

#### Now we will use Bing API to get the best time trave option for each cluster.

[http://dev.virtualearth.net/REST/v1/Routes?wayPoint.1={wayPpoint1}&viaWaypoint.2={viaWaypoint2}&waypoint.3={waypoint3}&wayPoint.n={waypointN}&heading={heading}&optimize={optimize}&avoid={avoid}&distanceBeforeFirstTurn={distanceBeforeFirstTurn}&routeAttributes={routeAttributes}&timeType={timeType}&dateTime={dateTime}&maxSolutions={maxSolutions}&tolerances={tolerances}&distanceUnit={distanceUnit}&key={BingMapsKey}](http://dev.virtualearth.net/REST/v1/Routes?wayPoint.1=%7BwayPpoint1%7D&viaWaypoint.2=%7BviaWaypoint2%7D&waypoint.3=%7Bwaypoint3%7D&wayPoint.n=%7BwaypointN%7D&heading=%7Bheading%7D&optimize=%7Boptimize%7D&avoid=%7Bavoid%7D&distanceBeforeFirstTurn=%7BdistanceBeforeFirstTurn%7D&routeAttributes=%7BrouteAttributes%7D&timeType=%7BtimeType%7D&dateTime=%7BdateTime%7D&maxSolutions=%7BmaxSolutions%7D&tolerances=%7Btolerances%7D&distanceUnit=%7BdistanceUnit%7D&key=%7BBingMapsKey%7D)

**We tested the API using one section of cluster 1 and as we can see, it validates our hypothesis that we can provide users distance optimization for each tour hop.**

**Sample output**



1. **Discussion section**

So, what all we did. We hypothesized that I can combine NYC data set with Foursquare and Bing API to create a tour covering different similar sections of NYC using 3 categories. Department Store, Electronics Store and Indian Restaurant. The first two category selections were based on the Tourist information wherein it stated that Tourists top destination are Department and electronics Store. The third one, well they got to it eat somewhere.

We used Foursquare API to get venues for those categories and tried to get ratings and likes. Unfortunately, the rating and like information does not exist for all venues and hence cannot be used. For now, I have mocked the data but we have to think how to do it. either we can use some other API like google or facebook or Yelp or asks for interests and based on interests from user and create the database. Anyway, the concept can be used as we got interesting clusters of NYC.

We then used Bing API to get the optimized distance for each neighborhood of each cluster. We only did for one of them but it can be repeated for others.

We got interesting and positive results back.

Google API now charges so we have to use alternative options, folium is good option to draw the locations and path. Overall, I am satisfied that we are on right track and the concept is feasible to be built into our tourist website.

1. **Result & Conclusion section**

We did clustering based on that the rating is > 5 and likes is greater than 50. We got clusters with some neighborhoods having either one or two of the interested categories. This was disappointing. We were only interested for neighborhood having all 3 categories with rating > 5 and likes > 50.

So, we removed all neighborhoods not matching our requirement. The biggest cluster, cluster 0 was totally removed because of this. So, now we have 3 clusters matching our requirement.

The number of neighborhoods were also considerably reduced but still promising. Maybe we should have applied Kmeans on the data with above check applied.

Bing was a great addition, we can add further categories like airport, hotel and give better experience to the tourist persona.

My hypothesis that I can use NYC data, Foursquare API and Bing API for a Virtual tour is correct. NYC itself has a potential for multiple tour locations. I can use similar method for other cities. This will value add to my Travel website. But I need to add more Venue categories. I also need to add Hotel location. I can also use the above clustering and distance optimization technique to make it interactive with the user.

1. **References**

[https://www.google.com/search?client=firefox-b-1-d&q=New+York+City#](https://www.google.com/search?client=firefox-b-1-d&q=New+York+City)

<https://nycfuture.org/research/destination-new-york>

1. **Acknowledgements**

I am thankful to IBM for this awesome course, it has been a great experience. I also thank to the members of the discussion group and IBM Watson support group. And in the last, towardsdatascience online blog and stackexchange to help me whenever I got stuck.