**Problem Statement and a discussion of the background**

New York City is the world’s liveliest citiy where the world flocks to for the best food, theatre, museum and business activity. New York city is home to a large working population with a huge disposable income and a large number of tourists. This makes New York one of the ideal locations to open a new restaurant. Having said that New York city also has one of the largest densities for restaurants. With its high real estate prices and stiff competition, its quite a challenge to decide the best neighborhood to open a restaurant.

I would like to use data science to guide the decision making. I would like to take Manhattan which is the borough with the highest GDP. I would like to identify the top ten neighborhoods. Once I have identified the top ten neighborhoods in Manhattan, I would like to perform a search for the different types of restaurants in each neighborhood. I would like to identify the most popular category of restaurants and also the neighborhood in the top ten neighborhoods with the least density of restaurants.

**People who will be interested in this project**

This would be a huge opportunity for any potential investors looking to open a new restaurant. This analysis would be useful not just for investors interested in opening restaurant but also other consumer-related businesses. The data analysis will also be useful to realtors who can leverage the data to attract potential home buyers or renters.

**Data description and analysis**

I will use the table from the following database <https://infogram.com/2019q2priciestneighborhoods-1h984wgnwp0z2p3> to identify the most priciest neighborhoods in New York. This would show me the neighborhoods with the wealthiest people. The wealthiest neighborhoods would have the highest spending power. I want to filter this data for Manhattan and then pick the top 5 wealthiest neighborhoods in Manhattan. Once the top 5 neighborhoods in Manhattan are identified, I will add latitude and longitude information to the data using geocoder package.

With the data for the top 5 Manhattan Neighborhoods, I would like to use the foursquare API to look for venues or popular spots. I will convert the JSON file into a dataframe using pandas. In venues I want to filter by restaurants to look for number of restaurants. I will plot that data and visualize it using Folium library. We would like to then sort the restaurant data to show the top categories of restaurant. Once we have the top category of restaurants sorted out, I would like to sort by neighborhood to identify which has the lowest number of restaurants and which category is least represented out of the most popular trend. This would be a huge opportunity for any potential investors looking to open a new restaurant, or a consumer-related business.

**Examples of Data to be used:**

Neighborhoods of New York city with latitude and longitude data from <https://geo.nyu.edu/catalog/nyu_2451_34572>

You can download the JSON file from the link above and download it in the Jupyter notebook using wget command. The downloaded data will look like

{'type': 'FeatureCollection',

'totalFeatures': 306,

'features': [{'type': 'Feature',

'id': 'nyu\_2451\_34572.1',

'geometry': {'type': 'Point',

'coordinates': [-73.84720052054902, 40.89470517661]},

'geometry\_name': 'geom',

'properties': {'name': 'Wakefield',

'stacked': 1,

'annoline1': 'Wakefield',

'annoline2': None,

'annoline3': None,

'annoangle': 0.0,

'borough': 'Bronx',

'bbox': [-73.84720052054902,

40.89470517661,

-73.84720052054902,

40.89470517661]}},

I will then get the top wealthiest neighborhoods from this source <https://www.propertyshark.com/Real-Estate-Reports/priciest-NYC-neighborhoods> . I will scrape the data from this table and filter the top ten wealthiest neighborhoods. The table looks like



I will then use the Foursquare API to get the popular spots in those neighborhoods. Then get the venues around those popular spots. I will use the K-mean clustering algorithm to cluster the venues in the top five neighborhoods. Define the clusters based on the number of restaurants. It might look like something below



I will use Folium package to visualize the clusters on the map. I will look for clusters where a particular cuisine is less represented and figure out the best place to open a new restaurant.

