1. Introduction

1.1 Background

Currently I live and work in the Manhattan of New York City and I am amazed by the fact that I have so many different choices of food.
I'm always curious which community is best community for a food lover like me? Is there a scientific methodology for me to rank the communities by restaurant they have? Can I classify the communities by the restaurants they have?
I believe there are several factors we can consider: Diversity (How many kinds of restaurants can this community offer?), Quality (Are the food and service good?), and Price (Is the food expensive or not?).

1.2 Problem

I will rank the communities by based on my preference on the diversity, quality, price and density of the restaurants in each community.
Then I will classify these communities based on these factors, and show the classification in map.

1.3 Interests

Many people in NYC, including my friends here and I, are willing to explore the restaurants. We are curious to know which community has the best restaurants, and how do the restaurants distribute by communities.

2. Data

2.1 Data Source

For New York City neighborhood names and their locations, we still get them from the NYU Spatial Data Repository.
Then I will utilize the Foursquare API to get the restaurants' information in these neighborhoods.

2.2 Data Cleaning

After implementing our framework and testing it with the Foursquare API, I ensured the data quality is good.
Then I only kept data relevant to neighborhoods in Manhattan.

2.3 Feature Selection

Density, Diversity, Quality, and price of restaurants in each neighborhood.

3. Methodology

3.1 Data Analysis

I first generated all the metrics we need.
Density: I first fetch all the restaurants in each neighborhood within 500m of their location; then we use 100 as the ceiling and set the total number of restaurants as density.
Diversity: we set the number of different types of restaurants in each neighborhood as diversity.
Price: our API has data quota limit, so I chose the restaurant closest to the neighborhood as sample and use its price as the price level of whole neighborhood.
Quality: our API has data quota limit, so I chose the restaurant closest to the neighborhood as sample and use its number of total likes as the quality metric of whole neighborhood.

Then I scale these metrics and get the overall score:
Score = Density + Diversity + Quality – Price.

In the end, I sort neighborhoods by their metrics:

Density:

```
df_summary.sort_values('Density', ascending = False).head()
```

index	Density	Diversity	Price	Quality	Score
Chelsea	1.0	0.914894	0.75	0.004638	1.169532
Midtown	1.0	0.851064	0.50	0.204082	1.555145
Financial District	1.0	0.851064	0.25	0.000928	1.601991
Midtown South	1.0	0.702128	0.25	0.004638	1.456766
West Village	1.0	0.851064	0.75	0.650278	1.751342

Diversity:

```
df_summary.sort_values('Diversity', ascending = False).head()
```

index	Density	Diversity	Price	Quality	Score
East Village	1.0	1.000000	0.50	1.000000	2.500000
Murray Hill	1.0	0.957447	1.00	0.050093	1.007540
Chelsea	1.0	0.914894	0.75	0.004638	1.169532
Clinton	1.0	0.893617	0.25	0.000928	1.644545
Flatiron	1.0	0.893617	0.75	0.349722	1.493339

We can see that neighborhood which has high diversity, must first have high density.

Price:

```
df_summary.sort_values('Price', ascending = True).head()
```

index	Density	Diversity	Price	Quality	Score
Marble Hill	0.15	0.234043	0.25	0.000000	0.134043
Stuyvesant Town	0.05	0.106383	0.25	0.029685	-0.063932
Tudor City	0.81	0.659574	0.25	0.009276	1.228851
Turtle Bay	0.86	0.659574	0.25	0.022263	1.291838
Midtown South	1.00	0.702128	0.25	0.004638	1.456766

The ones with cheapest price.

Quality:

```
df_summary.sort_values('Quality', ascending = False).head()
```

index	Density	Diversity	Price	Quality	Score
East Village	1.0	1.000000	0.50	1.000000	2.500000
Noho	1.0	0.893617	0.50	0.803340	2.196957
Little Italy	1.0	0.808511	0.50	0.712430	2.020941
West Village	1.0	0.851064	0.75	0.650278	1.751342
Greenwich Village	1.0	0.851064	0.75	0.469388	1.570452

The ones with highest quality.

Score:

```
df_summary.sort_values('Score', ascending = False).head()
```

index	Density	Diversity	Price	Quality	Score
East Village	1.0	1.000000	0.50	1.000000	2.500000
Noho	1.0	0.893617	0.50	0.803340	2.196957
Little Italy	1.0	0.808511	0.50	0.712430	2.020941
West Village	1.0	0.851064	0.75	0.650278	1.751342
Soho	1.0	0.765957	0.50	0.456401	1.722358

East village, Noho and Little Italy are top 3.

3.1 Statistic Inference & Machine Learning

```
df_final
```

index	Cluster_Labels	Density	Diversity	Price	Quality	Score
East Village	0	1.00	1.000000	0.50	1.000000	2.500000
Noho	0	1.00	0.893617	0.50	0.803340	2.196957
Little Italy	0	1.00	0.808511	0.50	0.712430	2.020941
West Village	5	1.00	0.851064	0.75	0.650278	1.751342
Soho	5	1.00	0.765957	0.50	0.456401	1.722358
Chinatown	3	1.00	0.787234	0.25	0.179963	1.717197
Clinton	3	1.00	0.893617	0.25	0.000928	1.644545
Financial District	3	1.00	0.851064	0.25	0.000928	1.601991
Greenwich Village	5	1.00	0.851064	0.75	0.469388	1.570452
Midtown	3	1.00	0.851064	0.50	0.204082	1.555145
Yorkville	3	0.89	0.744681	0.25	0.126160	1.510840
Flatiron	5	1.00	0.893617	0.75	0.349722	1.493339
Midtown South	3	1.00	0.702128	0.25	0.004638	1.456766
Civic Center	3	0.87	0.765957	0.25	0.058442	1.444399
Turtle Bay	3	0.86	0.659574	0.25	0.022263	1.291838
Tudor City	3	0.81	0.659574	0.25	0.009276	1.228851
Chelsea	2	1.00	0.914894	0.75	0.004638	1.169532
Upper East Side	5	0.80	0.723404	0.75	0.370130	1.143534
Carnegie Hill	1	0.65	0.659574	0.25	0.039889	1.099463
Sutton Place	3	0.80	0.702128	0.50	0.017625	1.019753
Murray Hill	2	1.00	0.957447	1.00	0.050093	1.007540
Lenox Hill	2	1.00	0.702128	0.75	0.028757	0.980885
Hamilton Heights	1	0.63	0.531915	0.25	0.006494	0.918408
Manhattan Valley	1	0.44	0.617021	0.25	0.022263	0.829285
Upper West Side	1	0.60	0.702128	0.50	0.019481	0.821608
Hudson Yards	1	0.46	0.510638	0.25	0.000000	0.720638
Lower East Side	1	0.43	0.531915	0.25	0.000000	0.711915
East Harlem	1	0.52	0.425532	0.25	0.003711	0.699242
Lincoln Square	1	0.53	0.510638	0.50	0.127087	0.667725
Morningside Heights	1	0.40	0.425532	0.25	0.039889	0.615421
Tribeca	2	0.66	0.659574	1.00	0.217069	0.536643
Inwood	1	0.49	0.510638	0.50	0.005566	0.506204
Gramercy	1	0.49	0.468085	0.50	0.007421	0.465506
Battery Park City	1	0.35	0.446809	0.50	0.103896	0.400705
Central Harlem	1	0.42	0.468085	0.50	0.004638	0.392723
Marble Hill	4	0.15	0.234043	0.25	0.000000	0.134043
Stuyvesant Town	4	0.05	0.106383	0.25	0.029685	-0.063932

The we use K-means to classify the neighborhoods based on the density, diversity, price and quality metrics; the result is very good.
We choose k-Means because this is an unsupervised learning problem: we only have the metrics while we don't know neighborhood's nature.

Then we compare the results of K-Means, we find the top 3 based on my methodology was classified into one unique cluster, while the bottom 2 were classified into another unique cluster!
The result of k-means is the same as my intuitive understanding of their ranking, which further supports my conclusion and help me gain a deeper understanding.

4. Discussion

4.1 Observation

The K-means clustering result on top and bottom neighborhood is the same as my intuitive methodology.

4.2 Suggestion

East Village, Noho and Little Italy are the top neighborhoods for eating, while Marble Hill and Stuyvesant Town are at the bottom.

5. Conclusion

We use the Foursquare API to get the metrics on restaurant in each neighborhood of Manhattan and rank them with an intuitive methodology; then we cluster them with K-means and find the machining learning result on top and bottom neighborhood for eating is consistent with our intuition on data.
We believe East Village, Noho and Little Italy are the top neighborhoods for eating, while Marble Hill and Stuyvesant Town are at the bottom.