Median Income and Venue Analysis of New York City

Zachary Avant

June 3, 2021

1. Introduction

New York City is famously competitive, and any new business has to be aware not only of what businesses are already in place, but also the population density and median income of a neighborhood's residents. By clustering New York zip codes based on nearby venues, median income, and density, we can find regions that share distinct characteristics. This information can be useful not only for people looking to start a business, but to residents or immigrants looking to move, or tourists trying to make the most of their visit.

2. Data

Data has been gathered from Citizens' Commission for Children of New York's median income dataset, as well as BetaNYC's neighborhood dataset. These were joined along zip codes, and together provide population densities, median income, and other useful features. Missing numeric data was replaced with the mean value of its column. Zip codes were utilized to find geospatial data, which was put into Foursquare's developer API to find neighboring venues.

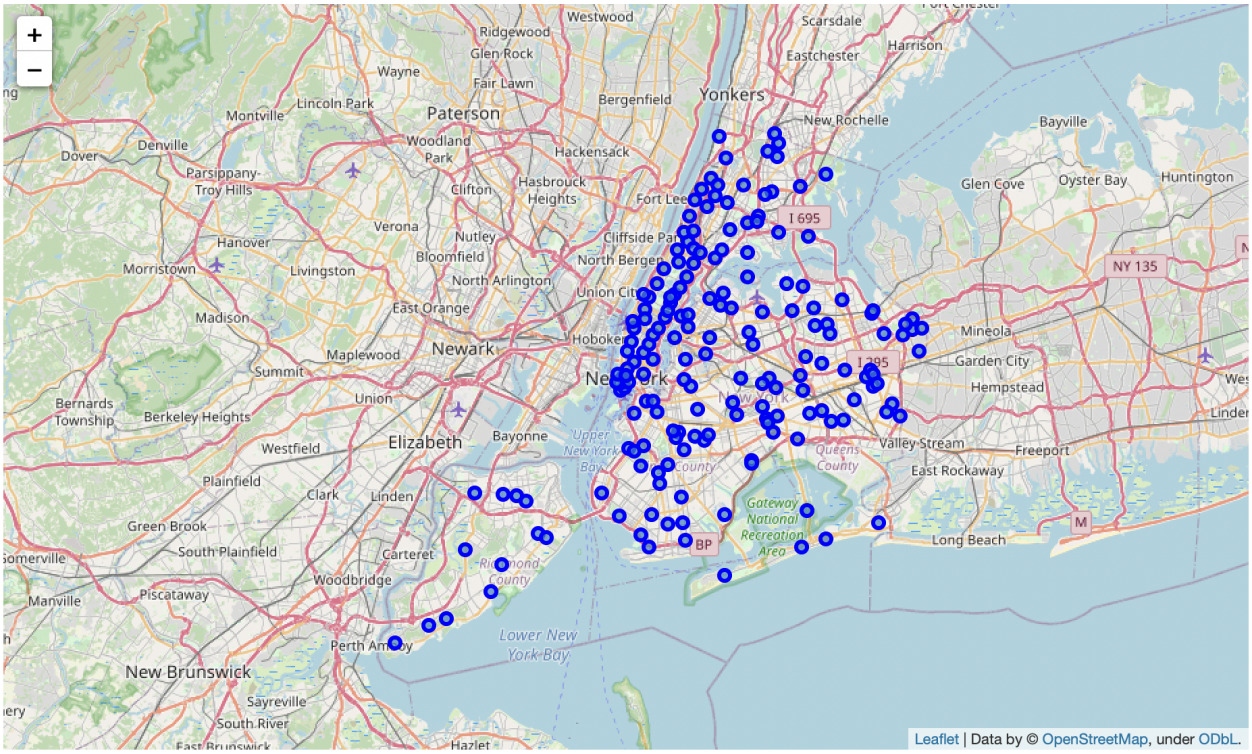
Merged Dataset

	Zip Code	Median Income	Borough	Neighborhood	Population	Density
0	10001	92840	Manhattan	Chelsea and Clinton	21102.0	33959.0
1	10002	36982	Manhattan	Lower East Side	81410.0	92573.0
2	10003	118161	Manhattan	Lower East Side	56024.0	97188.0
3	10004	190223	Manhattan	Lower Manhattan	3089.0	5519.0
4	10005	189702	Manhattan	Lower Manhattan	7135.0	97048.0

3. Methodology

Using zip code data from a merger of the two datasets, I found geospatial coordinates using the ArcGIS API. This was then passed to Folium to generate a map of zip codes in NYC.

Map of NYC by Zip Code

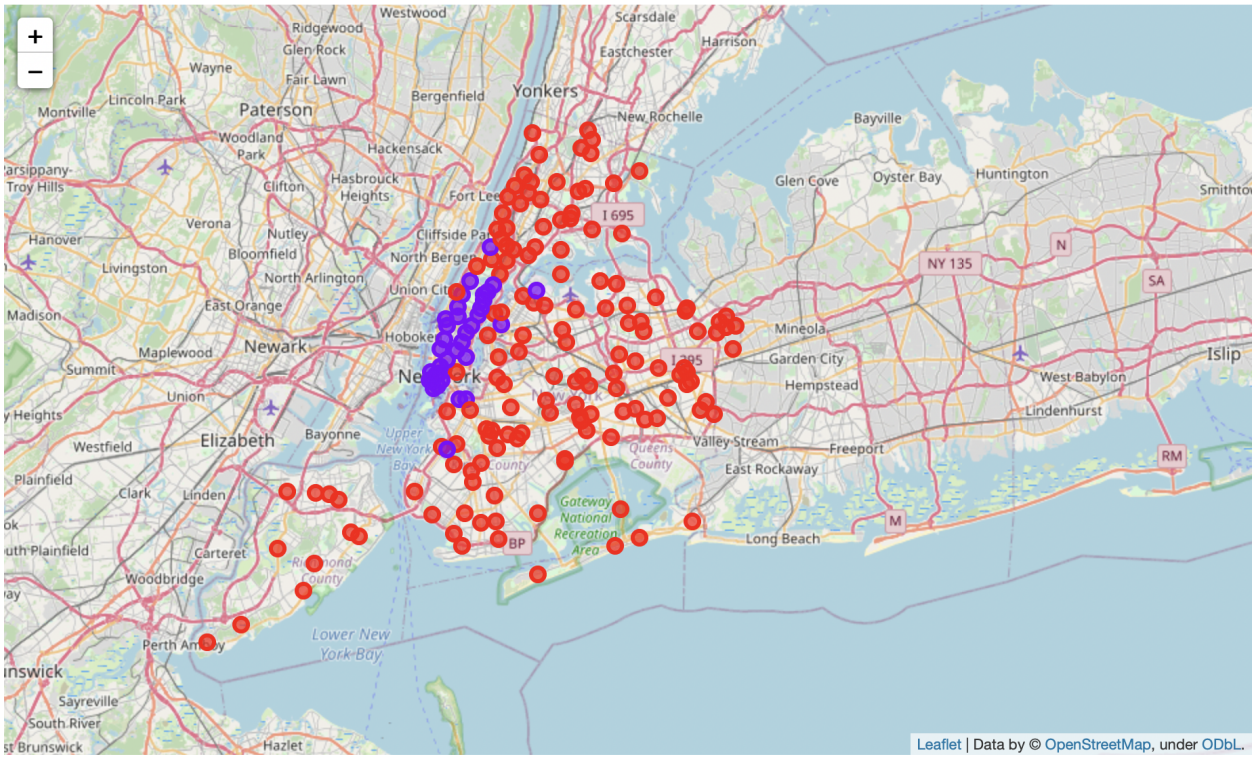


Foursquare was then used to find up to 100 venues within 100 meters of each zip code. From there, I calculated the 10 most common venues for each zip code. This dataset was subjected to one-hot coding and mean frequency calculations, merged into the income/density dataset, then normalized to create a final set for the KMeans clustering algorithm and for fitting.

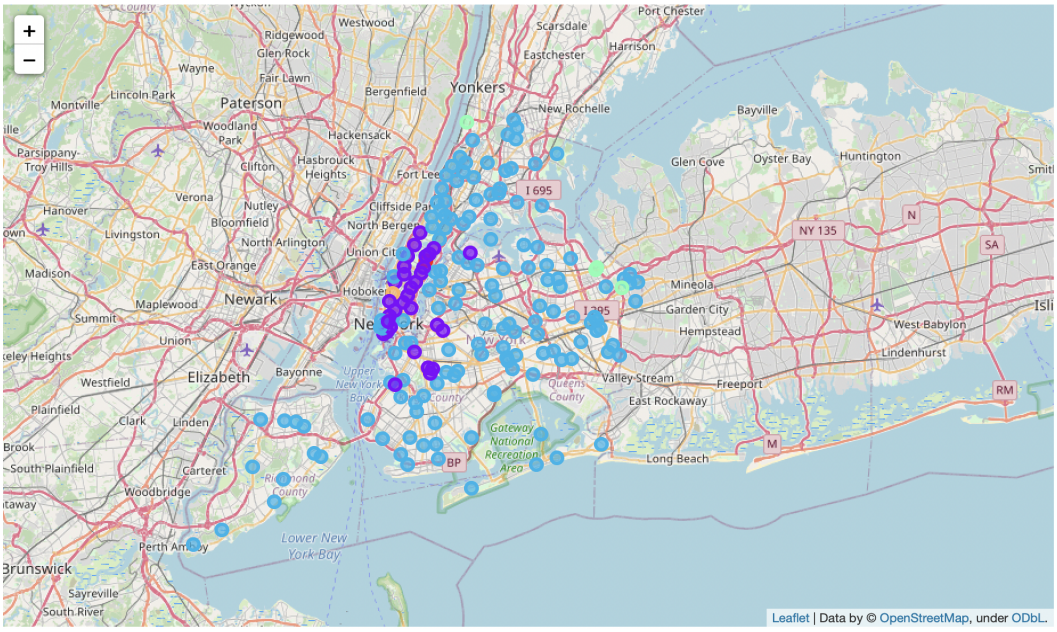
4. Results

Cluster labels from the KMeans algorithm were merged with the rest of the dataset, cleaned and finally mapped. Folium was employed again to create a final, clustered map of NYC zip codes at varying values of K.

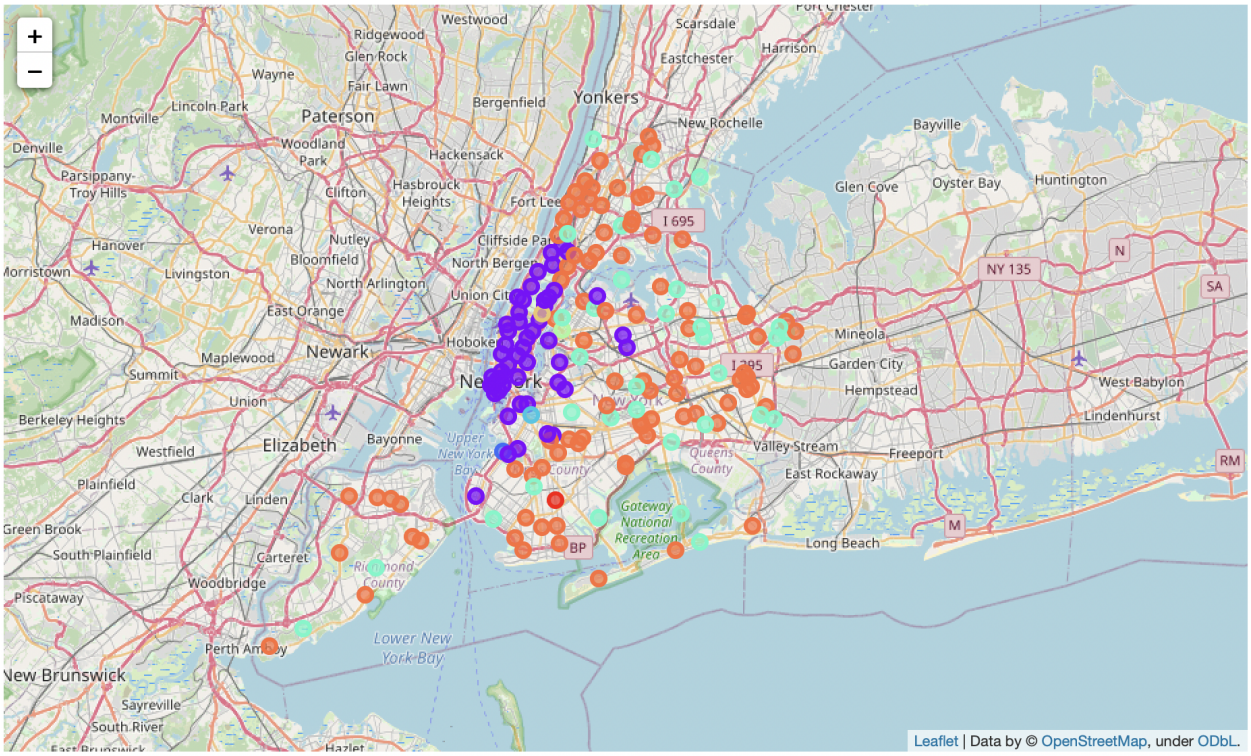
Map of NYC where K=2



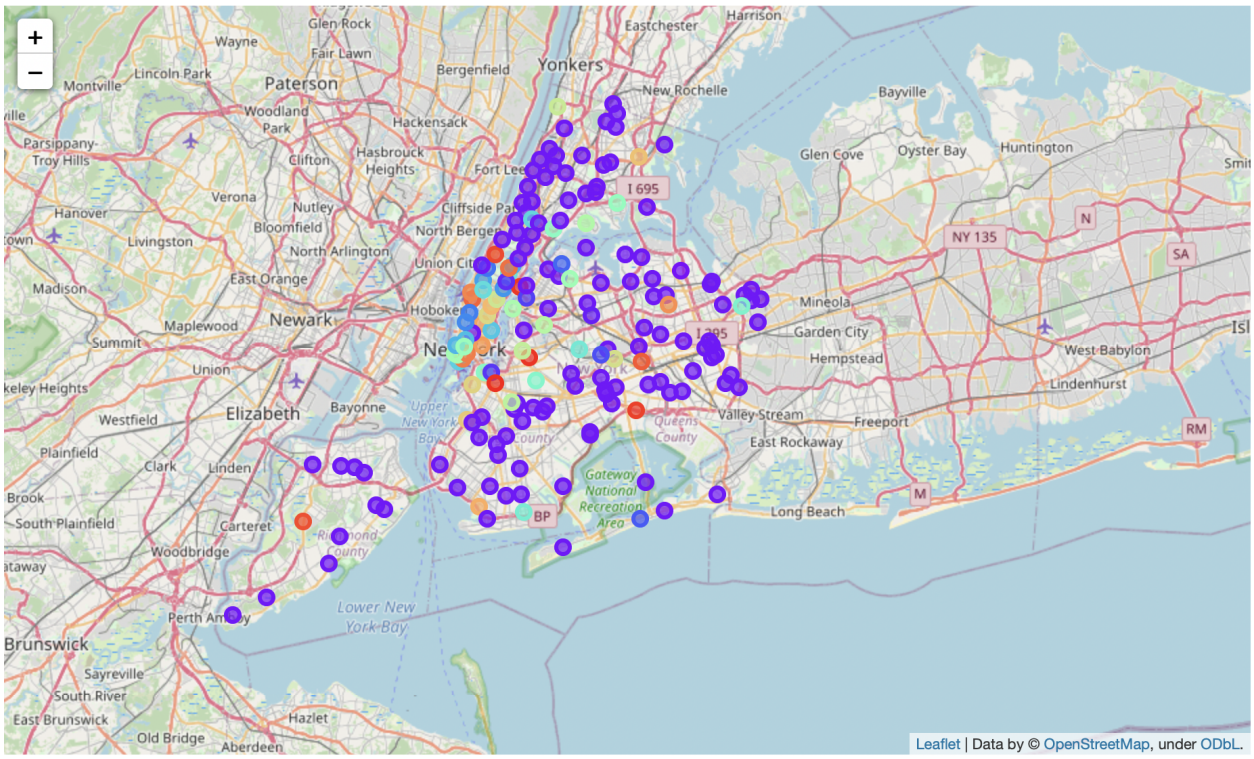
Map of NYC where K=5



Map of NYC where K=8



Map of NYC where K=60



5. Discussion

The KMeans algorithm ran into a number of problems which are beyond the scope of this project, but which would be a useful starting point for future study. Firstly, the differences between Manhattan and the other four boroughs of New York City overwhelm all other factors. Almost all of Manhattan is consistently placed within the same cluster, while much of the remaining boroughs are likewise clustered together in one or two dominant groups. This difference is (obviously) starkest where k=2 (as pictured in the Results section of this report).

As the value for K increases, this trend becomes more pronounced. Lower regions of Manhattan begin to form their own sets of unique clusters, migrating upward as K grows larger. At very high values of K, the Bronx and Queens begin to form the only remaining strong cluster, while Manhattan (except in the north) and neighboring parts of Brooklyn fragment.

Secondly, as a consequence of the peculiar groupings that form, zip codes often form their own clusters, even at smaller values of K. This frustrates the purpose of clustering algorithms and makes it difficult to find distinct clusters within NYC, apart from the tremendous divide between Manhattan, especially in the south, and the rest of the city.

6. Conclusions

Zip codes in lower parts of Manhattan are the most divergent among New York City zip codes. They, and much of Manhattan, share common popular venues like cafes, coffee shops, bars, and the like, but as K increases and venues outside the top three become more popular, Lower Manhattan quickly atomizes.

The strong division between Manhattan and every other borough may nevertheless provide some useful business knowledge, especially when paired with a further study. What works in Manhattan, for example, may not work especially well in other parts of New York City. Density, median income, and common venues radically diverge, and business and city planners must be aware of this.

A future study should be more targeted; New York should be iteratively split into different groups of boroughs to try and mitigate the strong divide between Manhattan and the rest. Doing so will allow one to find more distinct business and social climates within the different boroughs, especially outside of Manhattan. Lastly, a minimum cluster size would

mitigate the tendency of individual zip codes to form their own atomistic clusters. Such an algorithm, however, would go beyond the scope of this project.