Coronavirus Outbreak Analysis using Clustering

Luis Porras

June 7, 2020

1. Introduction

1.1 Background

The coronavirus has drastically changed the way we live in just a matter of weeks. Masks are necessary to enter most shops and restaurants, people are asked to keep 6-feet away from each other, and even a light cough in public draws overcritical stares from everyone around. 3 months into quarantine, many states' outbreak numbers have not gone down by much, and in some cases, have actually increased. The coronavirus also continues to wreak havoc on our nation's economy. Millions of people are filing for unemployment every month, and thousands of small businesses are permanently closing all over the country.

1.2 Interest

Of all the cities affected by the pandemic, New York has remained at the forefront of news headlines for its shocking amount of infected cases and climbing death toll. New York Governor Cuomo stated in his coronavirus briefing on May 20th, "You tell me the zip codes that have the predominantly minority community, lower income community. I will tell you the communities where you're going to have a higher positive, and you're going to have increased spread, and you're going to have increased hospitalization". This project aims to see if there is truly a relationship between New York's neighborhood demographics and coronavirus outbreaks.

2. Data

2.1 Data Sources

This project requires detailed information about the demographics of each New York neighborhood, and for this reason, a New York Census dataset from Kaggle will be used. I will be using a table of coronavirus statistics organized by borough and neighborhood from The New York Times. The Foursquare API will be used to cluster neighborhoods using a variety of different features regarding neighborhood venues for a supplementary analysis. Using these 3 sources of data will give me enough information to engineer new features, create exploratory visualizations, and hopefully come to a conclusion which relates neighborhood demographics to outbreak numbers.

Foursquare API: https://developer.foursquare.com/

New York Census: https://www.kaggle.com/muonneutrino/new-york-city-census-data?select=nyc_census_tracts.csv

New York Coronavirus Outbreaks: https://www.nytimes.com/interactive/2020/nyregion/new-york-city-coronavirus-cases.html

2.2 Data Cleaning

2.21 New York Times Coronavirus Table

Originally, I planned to create a web scraper that would collect data from a table on the New York Times's website. It was soon apparent that using the BeautifulSoup web-scraping library to try and extract the information from the table would not be a simple task, due to the many JavaScript elements on the page. For this reason, I decided to try and copy the table directly from the site, paste it into a string, and use string cleaning methods to build the desired table from scratch. As expected, this was not a simple task. There were letters out of place, numbers incorrectly formatted, and multiple null values and whitespaces characters littered throughout the string. After writing multiple regular expressions using Python's 're' library, the data was correctly engineered into a format similar to a csv file, and was ready to be fit into a dataframe. The resulting dataframe still contained multiple null values and incorrectly formatted data types. Using dataframe methods from the Pandas library, I removed rows containing null values (only 2 null rows of 176 total rows) and changed the data types of some numerical columns to integers, as needed.

2.22 FourSquare API

The FourSquare API has many different endpoints regarding location-based venue data. The API response is a json structure with multiple venue details such as the amount of venues in a given neighborhood, the type of venue, and user reviews for the venues. For this project, we only needed the venue types and their corresponding neighborhoods, explained in detail in section 4: Clustering. With a simple API request for New York data, I was able to collect the information needed for machine learning clustering and visualizations. Due to the structured format of the API response, the data required minimal cleaning. The API response did not include any null values, spelling errors, misplaced values, or other noticeable formatting issues that needed to be addressed before engineering features or creating exploratory visualizations.

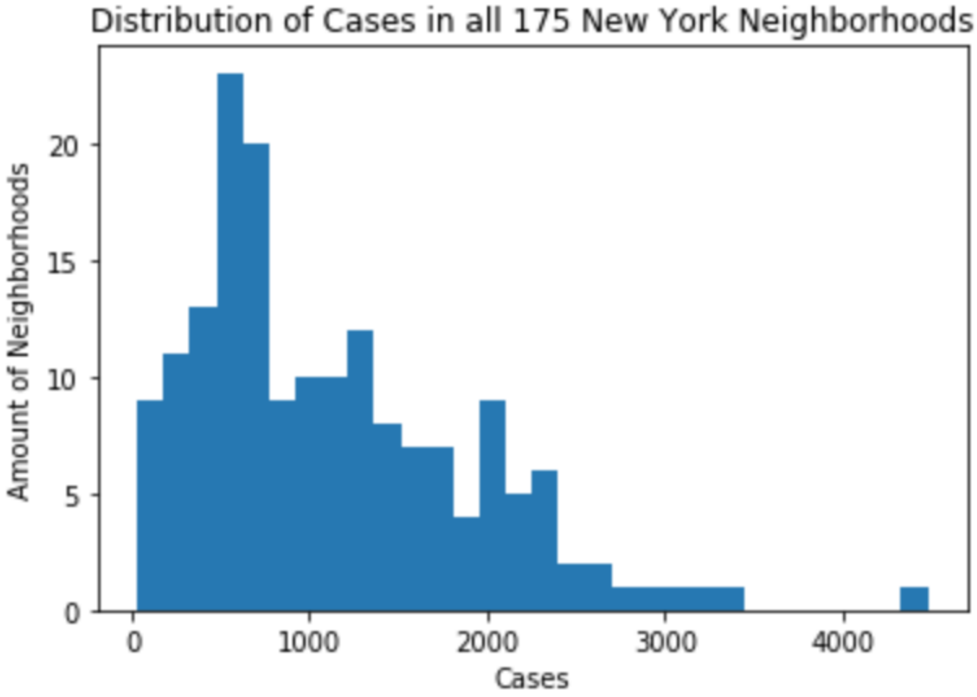
2.23 New York Census Dataset

This demographics dataset had some null values which were removed. The decision to remove these values completely was because these rows made up an insignificantly small fraction of the whole dataset, similar to the decision to remove the null values from the New York Times coronavirus table. Aside from this minor issue, the dataset contained 2 separate csv files which had to be joined in order to merge coordinate information to its corresponding neighborhood and demographic information.

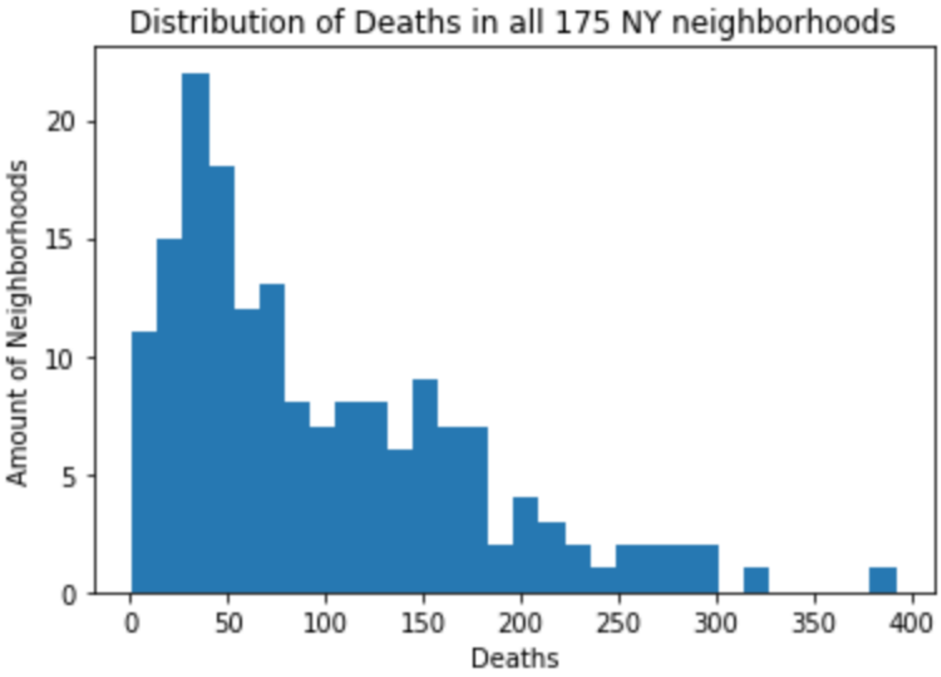
3. Exploratory Analysis

3.1 Distribution Analysis

3.11. Cases and Deaths of all Neighborhoods

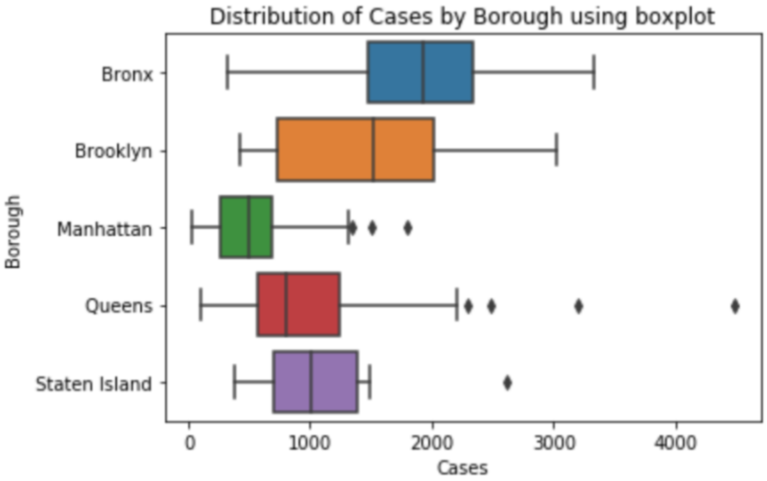


The distribution of the cases in all New York neighborhoods shows that the data is heavily right skewed, with most neighborhoods having somewhere between 500 and 1500 cases, but a handful of neighborhoods having more than 3000 cases. Corona, Queens is a neighborhood predominantly populated by minorities, with the racial make-up of the population being 63% Hispanic and 13.76% African American. Corona, Queens is our far-right outlier with the most coronavirus cases (right-most blue box, 4479 cases). The image below shows a similar distribution for the amount of deaths in all neighborhoods, as expected.

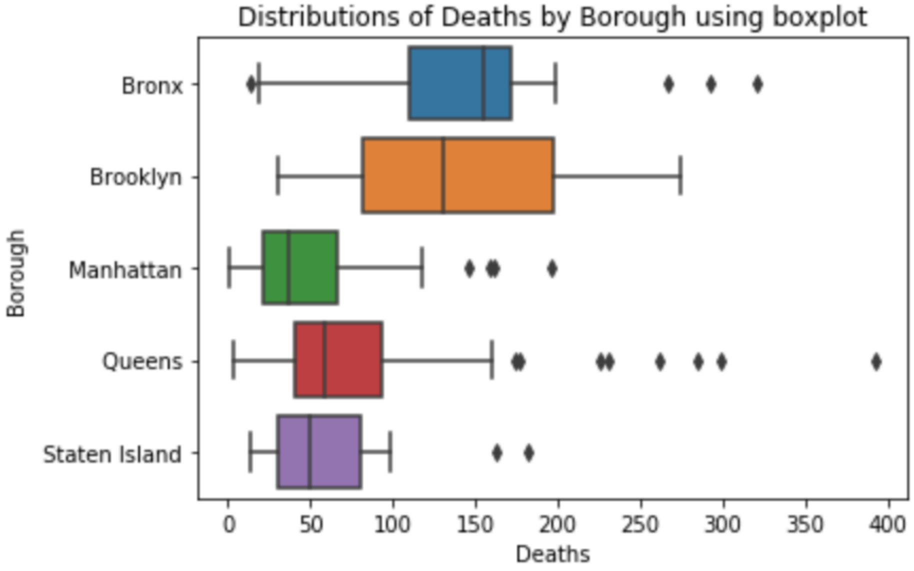


3.12 Cases and Deaths by Borough

A more granular view of the data shows that even though Corona, Queens had the maximum case count of all neighborhoods, the Queens borough (red boxplot) is populated with neighborhoods harboring some of the least amount of cases, between 10-100 cases per neighborhood. The boroughs with the highest median amount of cases per neighborhood were the Bronx, and Brooklyn, which are traditionally low income communities.



The distributions of deaths in each neighborhood organized by borough show some differences from the last image of case distributions. The Bronx and Brooklyn remain at the top of the death toll, just as they had the most cases, but Staten Island's distribution position moved lower than other boroughs' distributions. With this, we can see that the proportion of deaths to cases in Staten Island is lower than those in Queens and Manhattan. Staten Island's population make-up is around 25% minorities, and 75% White New Yorkers. Staten Island is also the "wealthiest" borough, having the highest median annual income of $72,156.



Borough	Median Deaths/ Median Cases
Bronx	8%
Brooklyn	8.3%
Manhattan	7%
Queens	7.5%
Staten Island	5%

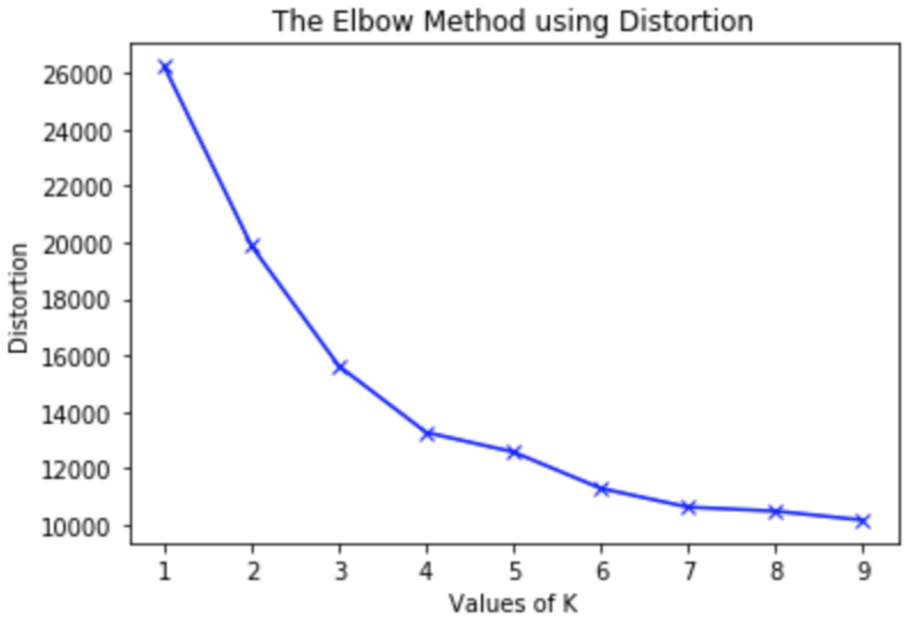
4. Clustering

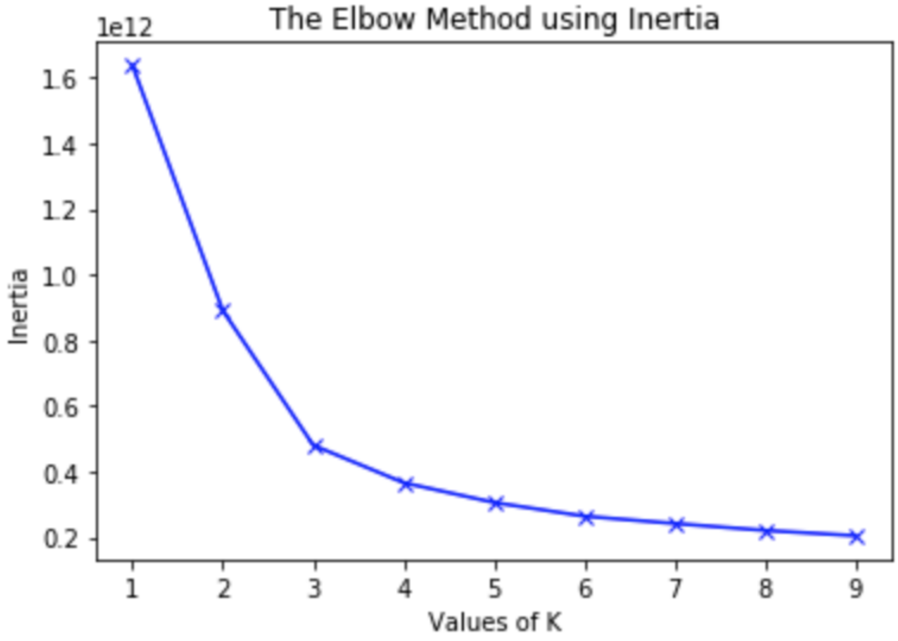
4.1 Cluster Model 1: Demographics data

The approach for this clustering model is to group our neighborhoods based on each neighborhood's demographics described in the Kaggle New York Census dataset. The clustering model being used is a K-Means clustering algorithm, which uses unsupervised learning methods to aggregate certain samples of data according to their similarities. The grouping criteria, or features, for this clustering model were the Child Poverty count, citizen count, county1 ,county2, Hispanic percentage, White percentage, Asian percentage, Black percentage, Income per Capita, Total Population, Men Count, Women Count, and Unemployment count. There were many other features in this dataset that were omitted in the clustering model, such as commute times and type of work. This was because the model was fit with data which focuses on Governor Cuomo's original statement saying 'predominantly minority community, lower income community' groups will tend to have higher spread and hospitalizations. Some of the omitted features were also creating redundancies in the data for the model.

Kept Features		Dropped Features
• Asian • Black • Borough • ChildPoverty • Citizen • County_x • County_y • Employed, • Hispanic • Income • IncomePerCap	• MeanCommute • Men • Native • Poverty • PrivateWork • Professional • PublicWork • SelfEmployed • TotalPop • Unemployment • White • Women	• BlockCode • Borough • Carpool • CensusTract • Construction • Drive • FamilyWork • OtherTransp, • Professional • PublicWork • Walk

To choose the optimal number of clusters, distortion and inertia values were calculated for different values of K from 1-9. Distortion gives us the average of the squared distances from each point to the point's center. Inertia is the sum of the squared distances to their closest cluster center. Two graphs were created to compare each value of K to its corresponding Inertia and Distortion.

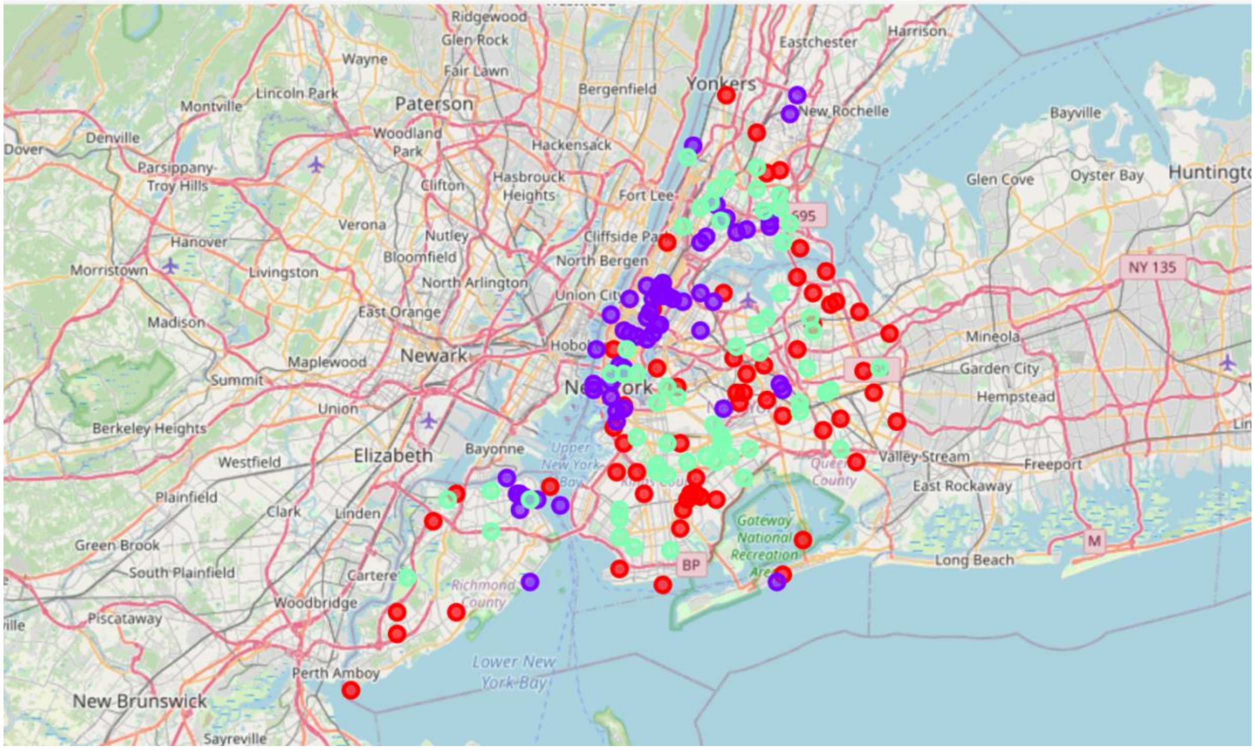




The Elbow method is commonly used to determine the optimal number of clusters, based on the distortion and inertia graphs. The sharpest turning point of the curve, also known as the "elbow", represents the best choice for the value of K in our clustering model. The elbow in our distortion plot seems to be at the K-Value of either 3 or 4, while the elbow in our inertia plot is looking much more like 3. For this reason, the defined number of clusters for the model was 3.

The demographics dataset did not include the 175 neighborhoods from New York, but instead, had over 1000 instances of locations all over New York, based on what is called a Census Tract number. Each Tract number represents a smaller location inside each one of the major 175 neighborhoods, but in aggregation, they represent the same space. Originally, there were 633 points in cluster1, 610 points in cluster 2, and 77 points in cluster 3. Due to the large amount of points on the map visualization (1326 locations), many points were cluttered on top of each other, leaving some points unable to be seen. For the visualization only, I decided to remove samples, also known as “downsampling”, from each of the 3 cluster groups to only 77 samples, which was the amount of the minority group.

Map of Tract Locations based on Demographics Clusters



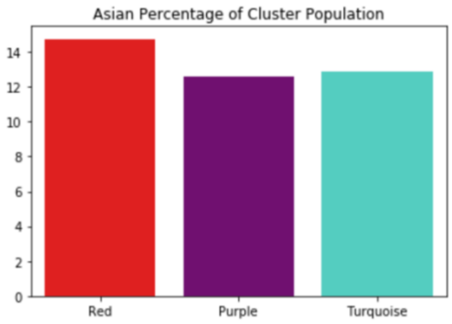
Based on the map, we can see that there are 3 different clusters using the neighborhood demographics: purple, turquoise, and red. Each color represents a group with similar characteristics based on the descriptive features in the “Kept Features” column of the features table above. Most features are somewhat related to ethnicity and income.

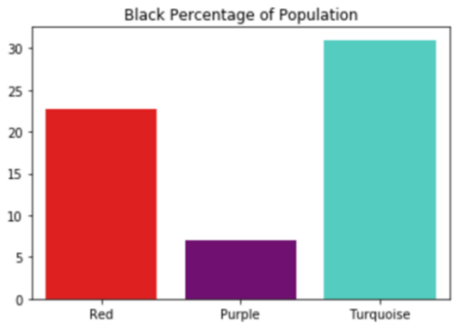
Income per Cluster

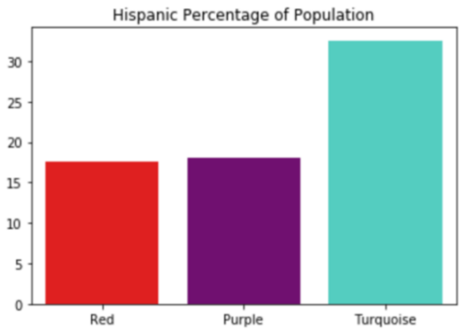
Cluster	Mean	Standard Dev	Maximum	Minimum
Red	$70,654	$18,126	$150,833	$11011
Purple	$105,863	$44,969.6	$222,222	$20,849
Turquoise	$41,678.9	$16,384	$144,293	$9829

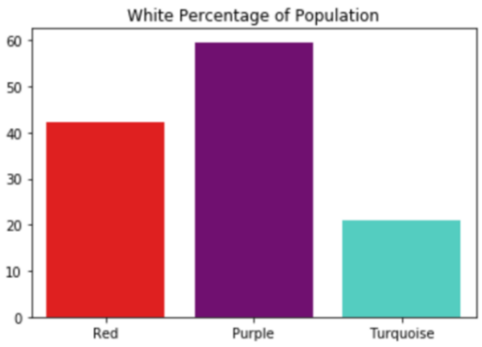
This table shows that the neighborhoods in the purple cluster are much wealthier on average. The purple cluster’s minimum income was also 89% higher than the minimum income in the red cluster, and more than twice the minimum income from the turquoise cluster.

Due to limitations in the original data, the exact ratios of the ethnic makeup of each cluster could not be calculated, but instead, these bar plots show the average percentage of for each neighborhood in each colored cluster.









The racial make-up of the neighborhoods shows that the wealthier purple cluster is likely predominantly White, averaging a 63% White population in all purple cluster neighborhoods. The cluster with the highest proportion of Black and Hispanic people on average is the turquoise cluster. The red cluster's bar plot shows that it is slightly more diverse than the other clusters, but it's largest average proportion is still a massive 40% White people for its neighborhoods.

4.2 Cluster Model 2: Venue Data

While this project is not a business analysis of New York venues, I have used the Foursquare API and the neighborhood clustering lab from the course as data for a supplementary analysis of Manhattan. The clustering lab from the course was a walkthrough of the Foursquare API, feature engineering, and clustering machine learning methods to superimpose neighborhoods onto a map of New York. Comparing the clusters created from venue information to the clusters using neighborhood demographics and coronavirus outbreaks may help us understand which businesses in Manhattan were likely impacted the most by the pandemic.

The clusters in this map were based off of the amount of venues, and types of venues in Manhattan. Looking at the cluster map, we can see which types of businesses might have been affected the worst by the virus by comparing them to worst affected areas from the New York Coronavirus dataframe.

5. Discussion

Using the demographics data, it seems that the wealthy neighborhoods with the smaller minority make-up were a part of the purple cluster. The poorest neighborhoods had much larger minority parts, as described by the neighborhoods in the turquoise cluster. The red cluster was similar to the turquoise cluster, but was slightly wealthier and had a more even balance of races than the other parts. When taking a look at the map of the clustered neighborhoods, the wealthier purple neighborhoods with less minorities than the other clusters seem to pile in the Manhattan area of New York. Manhattan is known for its massive skyscrapers, neon-lit Times Square, and thriving financial district. This borough also has one of the lowest shares of population with a reported case, at around 1 coronavirus case in every 50 people.

Looking at the polar opposite turquoise cluster, which has the lowest average income and highest minority proportion out of all the clusters, the turquoise points are spread all throughout the map, but seem to pile in The Bronx and some parts of the lower Queens and Brooklyn boroughs. The Bronx has the highest amount of reported cases per capita, at around 1 in every 30 people. Brooklyn and Queens also have large amounts of reported cases per capita, but the highest amounts are also in the lower ends, where the clustering model reported multiple low income, high-minority count neighborhoods.

While correlation is not causation, Governor Cuomo was not wrong when he stated there is a clear relationship between demographics and coronavirus outbreaks and deaths. The wealthiest neighborhoods had the lowest amount of cases per capita, and lowest amount of deaths. The poorest areas with the highest proportion of minority population had the worst impact from the pandemic.

6. Future Directions

This project was supported by multiple data sources, with one source being a table of current coronavirus outbreak numbers. Moving forward, I would hope to find a dataset of historical coronavirus outbreak numbers. Using a dataset with historical coronavirus data would give me the information needed to train a regression model, and make predictions on the future trends of outbreaks in New York neighborhoods. This regression model would be similar to the many coronavirus "curves" we see on the news, which make predictions on cities' outbreak potential and death toll. Currently, the historical datasets seem to be behind the closed doors of local governments and researchers, but I hope these datasets will be more accessible to the public in the near future. Publicly available coronavirus data would invite analysts of all kinds to create projects that could aid future pandemic prevention efforts, and advance research in epidemiology.

With more information regarding the performance of local businesses, we can also analyze the economic impact of the virus on a neighborhood level. The Foursquare API can be used to map certain venues to their locations, but metrics such as revenue, profit, and traffic are important measures of the virus's influence on New York local businesses.