Coursera Capstone Project:

Descriptive Analysis on the relationship between key demographic indicators and Covid19 contraction in New York City

Son Le

April 20th, 2020

Introduction

The 2019 novel coronavirus (Covid-19) is an infectious disease caused by the "Severe Acute Respiratory Syndrome Coronavirus 2" (SARS-CoV-2) [3]. As of April 14th, 2020, there have been 1.91 million confirmed cases of infection globally [3]. It is easily contracted from travelling and airborne exposure to the infected, with symptoms developing typically between 2-14 days [4]. The World Health Organization (WHO) declared SARS-CoV-2 a "Public Emergency of International Concern" on January 30th, 2020, and a pandemic on March 11th, 2020 [3]. The virus has caused a massive economic and social recession on a global scale [8].

The United States, officially, has the highest number of confirmed cases of Covid-19 (467,184) as of April 10th, 2020 [4]. At least 161,807 cases were reported in New York State, and 87,028 alone in New York City [5]. This has resulted in the complete shutdown of the entire state, suspending most economic activity and causing massive financial losses. There are reports of empty streets [4], the hoarding of essential goods [4], and even mass burials of those who have succumbed to the disease [6]. Several hospitals within the state have become fully dedicated to treating the infected, focusing on those in critical conditions who are, typically, children and the elderly [5]. While, the novel coronavirus has been speculated to originate from China, it was quickly spread to every country in the world through social interactions with those infected.

Business Problem

New York City (NYC) is, currently, experiencing an unprecedented pandemic that could have major lasting impact on its residents. What started as a small outbreak caused by a super-spreader has resulted in entire city declared to be in a state of emergency [4]. While the city

government's poor leadership and response were cited as factors contributing to the virus's continued spread, it was the initial lack of social distancing by its resident that exponentially increased it [9]. Being one of the world's foremost financial and cultural hub, NYC is the home of many communities spanning different ethnicities and socioeconomic classes. Perhaps, these groups, because of their demographic differences, experience the effects of Covid19 in different levels. According to Cynthia Carr, an epidemiologist in Winnipeg, Canada, "In New York City, high population density and social determinants of health such as income and housing are factors" [7]. Also, with a population of over 8 million, NYC is a densely populated city which "tend to be more affected by the virus" [7]. Also, given the socially driven nature of this virus, the number of nearby venues available to each resident is speculated to be a contributing factor to their possible contraction of Covid19 as well.

This study is an attempt to explore the veracity of these claims by studying the relationship between the number of positive tests of Covid19 and several key demographic indicators in each geographically defined community, or zip code, in New York City such as:

- **Population density**: measured in the number people per square mile.
- **Median household income**: measured in US dollars or USD.
- **Number of nearby venues** within a 500-meter radius of each zip code's coordinates

The nature, or category, of each zip code's most popular venues will also be studied. This is to verify the assumption that large gatherings increase the chance of contraction of the novel coronavirus.

Data Acquisition

Based on the Business Problem defined above, the data needed for this project will consist of:

1. Number of recorded total and positive tests for Covid19 in New York City in each zip code using web-scraping from NYC Health. [2]
2. Current, or estimated demographic data for population density and median income in each NYC zip code using web-scraping from city-data.org [1]
3. Each found zip code's coordinates using the **OpenCage API**
4. Venues within each zip code's 500-meter radius using the **FourSquare API**.

Covid19 data was found for 177 NYC zip codes. Below is a sample of the data

Table 1: Covid19 data for each NYC zip code

	Zip Code	Positive Tests	Total Tests	Percentage of Total
0	10001	260	571	45.53
1	10002	712	1358	52.43
2	10003	347	830	41.81
3	10004	24	64	37.5
4	10005	44	137	32.12
5	10006	14	54	25.93
6	10007	40	130	30.77
7	10009	518	1180	43.9
8	10010	201	561	35.83
9	10011	394	852	46.24

Demographic data was found for 214 NYC zip codes. Below is a sample of the data

Table 2: Relevant demographic data for each NYC zip code

	Zip Code	Population Density	Median Income
0	10001	38085.0	88701.0
1	10002	90078.0	37071.0
2	10003	99889.0	104972.0
3	10004	5515.0	128161.0
4	10005	120158.0	135514.0

All data is updated as of April 19th, 2020.

Data Cleaning

For demographic data from Table 2 was several zip codes, and data from other zip codes was also missing. Therefore, they were removed from their dataframes. The cleaned Tables 1 and 2 were then merged. Using the OpenCage API, the latitude and longitude for each zip code was obtained and added to the merged table. The number of venues were then added to the table using the FourSquare API. The compiled table is shown below with data presented for 123 zip codes.

Table 3: Consolidated Information

	Zip Code	Latitude	Longitude	Population Density	Median Income	Positive Tests	Total Tests	Percentage of Total	Nearby Venues
0	10001	40.729825	-73.960752	38085.0	88701.0	260	571	45.53	67
1	10002	40.722313	-73.987709	90078.0	37071.0	712	1358	52.43	85
2	10003	40.731609	-73.988484	99889.0	104972.0	347	830	41.81	78
3	10004	40.700732	-74.013475	5515.0	128161.0	24	64	37.50	25
4	10005	40.705636	-74.008900	120158.0	135514.0	44	137	32.12	100

The zip codes are mapped below, displaying the population density, median income, and the number of positive tests. It should be noted that after removing missing/messy data from Table 2 and merging it with Table 1, a large portion of Queens borough ended up missing from the map.

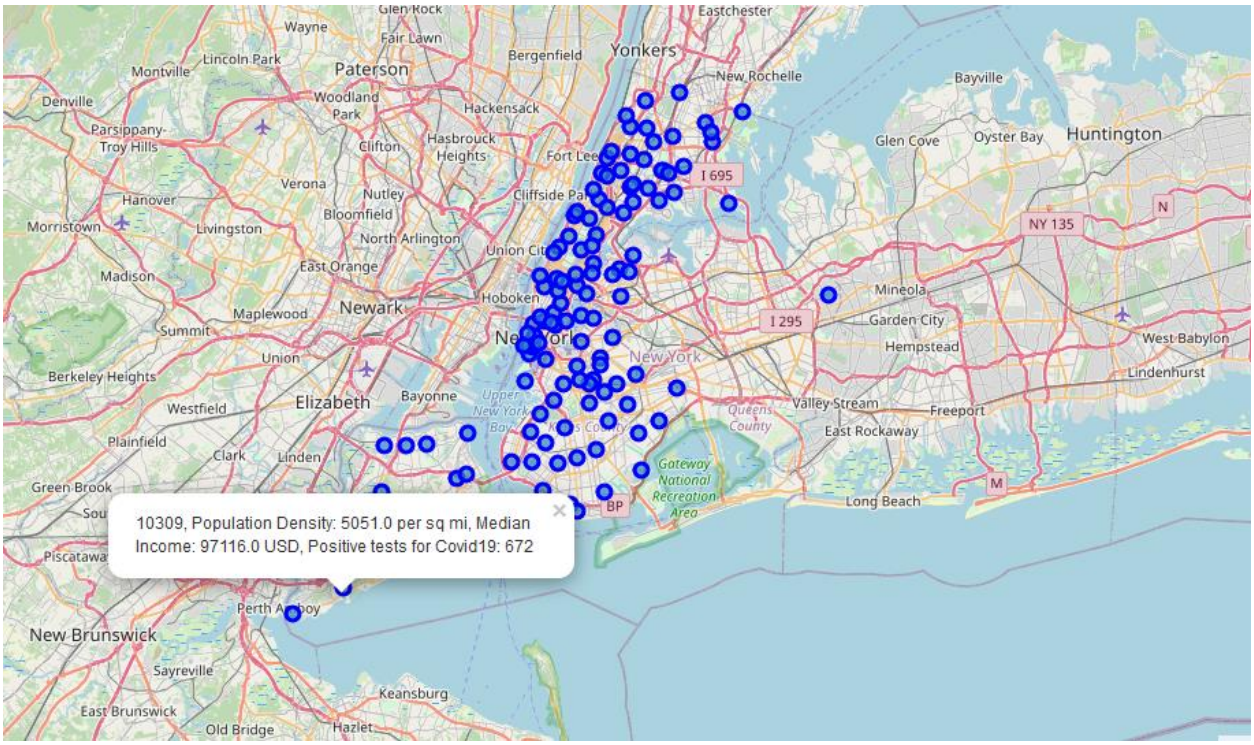

Figure 1: NYC zip code map with relevant information displayed in each pop-up

Lastly, each zip code's list of venues and their relevant information were queried using the FourSquare API then stored in a separate dataframe. A sample of it is shown below.

Table 4: List of each zip code's venues are their relevant information

	Zip Code	Latitude	Longitude	Venue	Venue Category	Venue Latitude	Venue Longitude
0	10001	40.729825	-73.960752	WNYC Transmitter Park	Park	40.729958	-73.960733
1	10001	40.729825	-73.960752	Paulie Gee's	Pizza Place	40.729801	-73.958520
2	10001	40.729825	-73.960752	Bellocq	Tea Room	40.730372	-73.959213
3	10001	40.729825	-73.960752	Ovenly	Bakery	40.729708	-73.959544
4	10001	40.729825	-73.960752	New Love City	Yoga Studio	40.729760	-73.958247

Analysis

Each metric (Population Density, Median Income, Number of Nearby Venues, and Number of Positive Tests) are first explored through Exploratory Analysis. A histogram of the data distribution of each metric is shown below. The y-axis shows the number of zip codes.

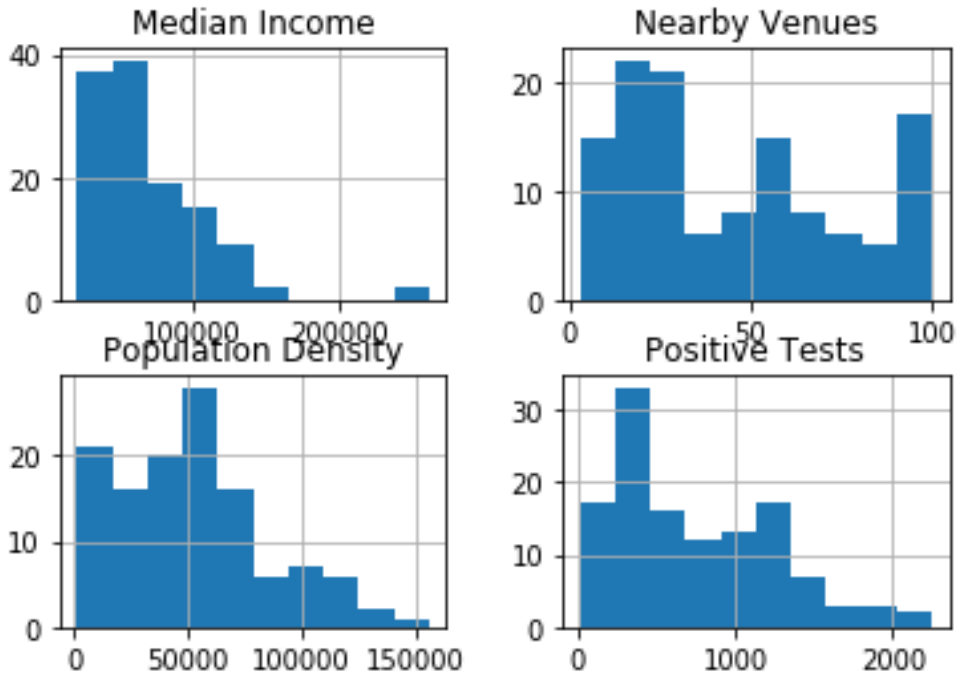

Figure 2: Distribution of Median Income (USD), Population Density (people per square mi), Nearby venues, and Positive tests.

First, before building a model for the number of positive tests, it is necessary to study the correlation between each indicator or feature (population density, median income, and nearby venues) and the target (positive tests).

Population Density

Table 5: Important metrics for Population Density vs Positive Tests

Correlation coefficient	-0.084131
p-value	0.354874
R-square	0.007078
MSE	256692

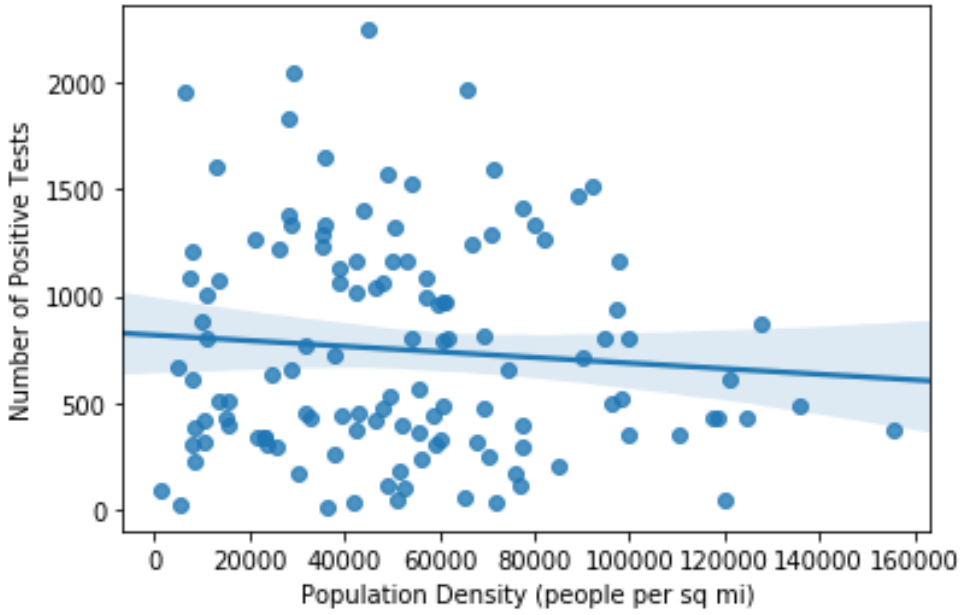

Figure 3: Population Density vs Number of Positive Covid19 Tests

From Figure 3, the data is quite scattered, presenting a very weak negative correlation between Population Density and the number of Positive Tests for Covid19. Pearson's correlation metrics indicate no correlation between these two variables. However, the high p-value shows a little to no certainty in the result. This is reflected in the extremely low R-square score, and MSE which may indicate that the data collected is highly varied or noisy, contributing to a poor model.

Median Income

Table 6: Important metrics for Median Income vs Positive Tests

Correlation coefficient	- 0.540149

p-value	1.134520e-10
R-square	0.291761
MSE	183096

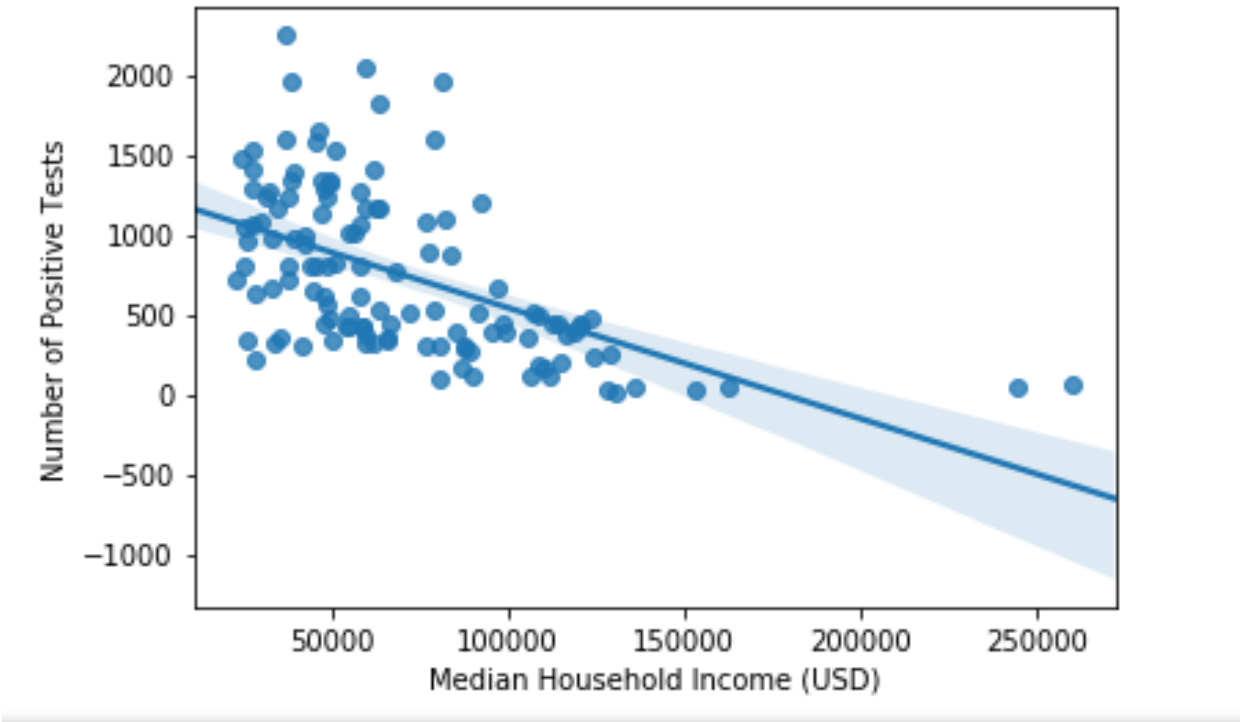

Figure 4: Median Household Income vs Number of Positive Covid19 Tests

According to Figure 4, there is a noticeable negative correlation between the Median Household Income and Number of Positive Tests among zip codes in NYC. This is supported with a moderately negative correlation coefficient found in Table 6. The p-value suggests a strong certainty in the data when building this model. The R-squared score and MSE are higher and lower, respectively, when compared to Population Density, which indicates that the relationship between Median Household Income and Number of Positive Covid19 Tests is much more defined. However, the model still does not do a great job at explaining the high variance in the data.

Number of Nearby Venues

Table 7: Important metrics for Nearby Venues vs Positive Tests

Correlation coefficient	-0.446545
p-value	2.255087e-07
R-square	0.199402
MSE	206972

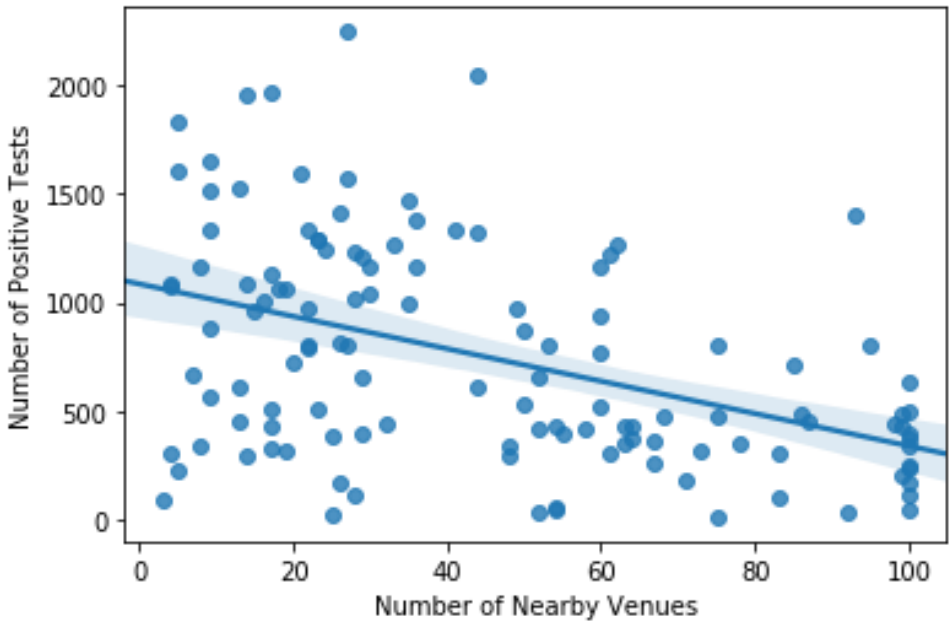

Figure 5: Number of Nearby Venues vs Number of Positive Covid19 Tests

Like Median Household Income, there is a moderate, if slightly weaker, negative correlation between the Number of Nearby Venues and the Number of Positive Covid19 Tests among the zip codes. A p-value of ~2.26e-07 suggests that there's a strong certainty of this result. However, a lot R-squared and high MSE suggest that there is high variance in the data.

Positive Covid19 Tests Model

From the analysis above, in order to build a descriptive model of positive Covid19 tests for all zip codes, Population Density needs to be dropped from the list of features. After performing Multiple Linear Regression using Median Household Income and Nearby Venues as the features, the following was obtained.

Table 7: Important metrics for Multiple Linear Regression Model

Correlation coefficient (Median Income)	-217.8975
Correlation coefficient (Nearby Venues)	-136.4045
R-square	0.351279
MSE	167708

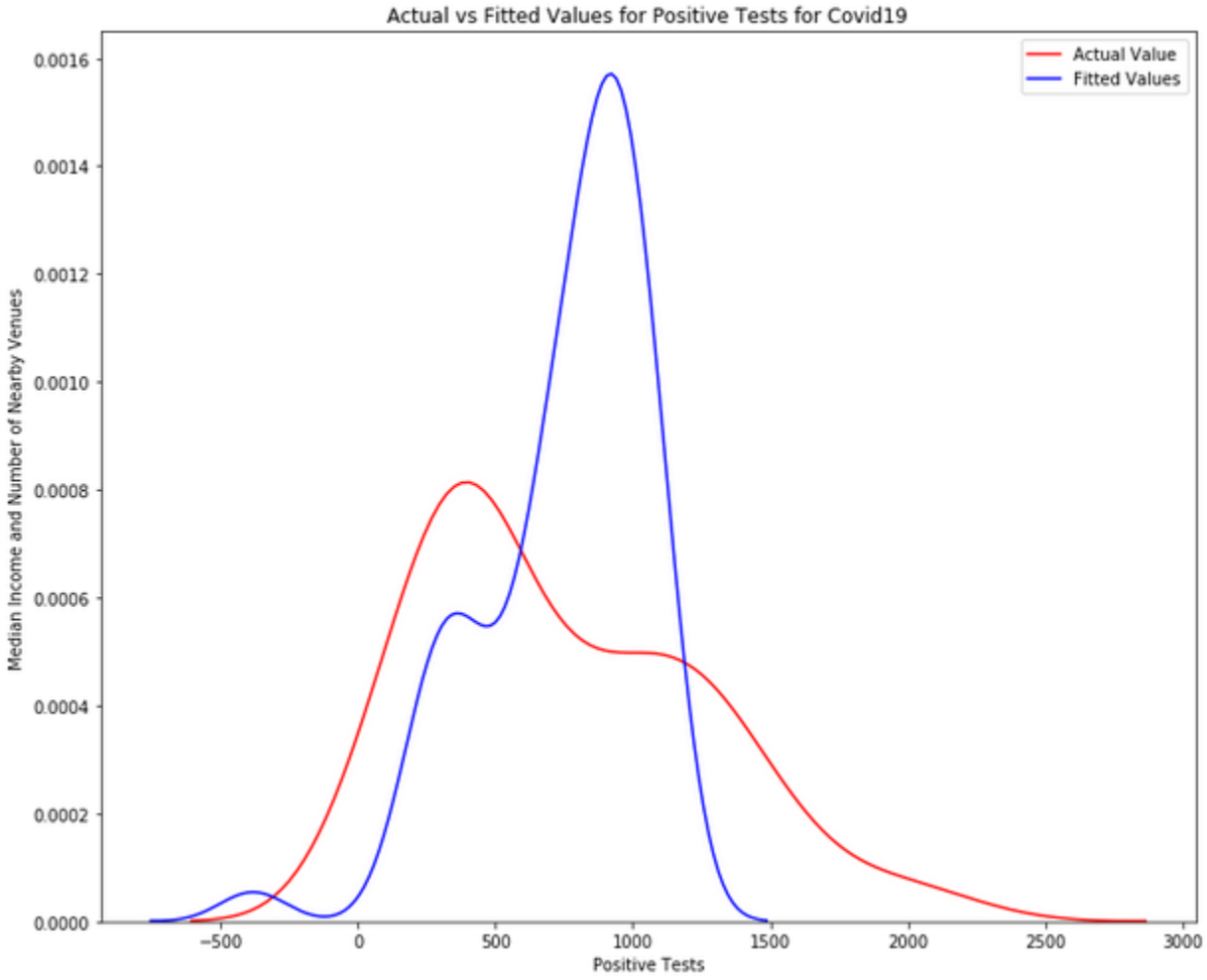

Figure 6: Positive Covid19 Tests Linear Regression Model

From Table 7, similar to the singular Linear Regression Models, both Median Income and Nearby Venues contribute negatively to the Number of Positive Covid19 Test. However, Medium Income has the larger influence. Figure 6 shows the Multiple Linear Regression model's performance compared to the actual results.

Analysis of Nearby Venues

Using the FourSquare API, a list of nearby venues of each zip code was generated, and its10 most common/popular venues were determined and grouped in a new datafram. Using the KMeans machine learning algorithm, the zip codes were clustered and shown in the map below.

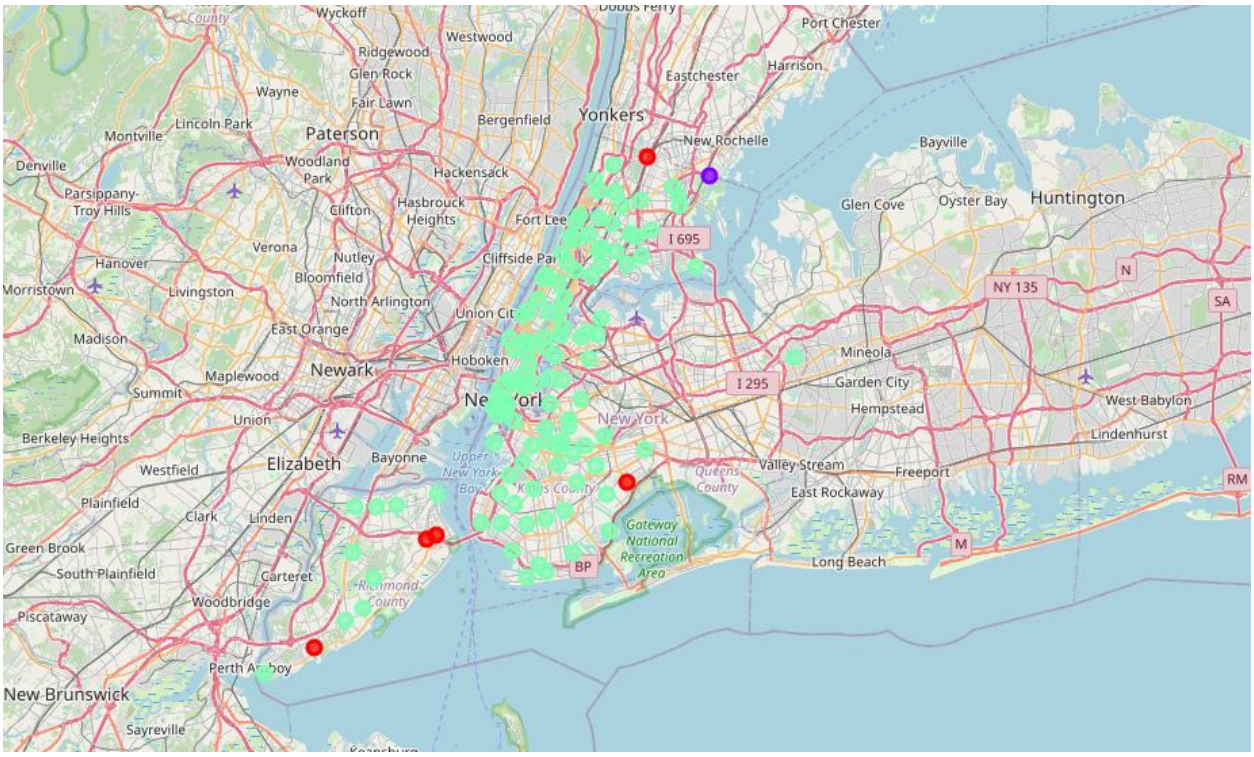

Figure 7: Clustered Zip Codes in New York City

From Figure 7, most zip codes belong in the green cluster (cluster 2) with 6 outliers belonging in cluster 0 and 1. These outliers are shown below.

Table 8: Outlier zip codes

	Cluster Labels	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Population Density	Median Income	Positive Tests	Total Tests	Percentage of Total	Nearby Venues
0	0	10304	Deli / Bodega	Pizza Place	Athletics & Sports	Grocery Store	Park	11009.0	56077.0	1001	2003	49.98	16
1	0	10305	Bus Stop	Hotel	Athletics & Sports	Coffee Shop	Construction & Landscaping	9908.0	77135.0	883	2001	44.13	9
2	0	10309	Pizza Place	Pharmacy	Italian Restaurant	Sushi Restaurant	Bagel Shop	5051.0	97116.0	672	1626	41.33	7
3	1	10464	Border Crossing	Construction & Landscaping	Zoo	Food	Falafel Restaurant	1245.0	80108.0	95	247	38.46	3
4	0	10470	Deli / Bodega	Pizza Place	Pub	Rental Car Location	Donut Shop	10376.0	61280.0	321	617	52.03	19
5	0	11239	Moving Target	American Restaurant	Bus Stop	Business Service	Pizza Place	21639.0	25631.0	339	578	58.65	8

Results and Discussion

From the three selected features “Population Density,” “Median Income,” and “Nearby Venues,” only Median Income (- 0.540149) and Nearby Venues (-0.446545) showed significant correlations to the target “Positive Tests.” Their extremely small p-values of 1.134520e-10 and 2.255087e-07 confirmed the confidence of this correlation. It is interesting to note that these correlations suggest that having more nearby venues within 500 meters of a zip code contributes to a lower number of people who tested positive for Covid19. Perhaps, having more nearby venues may cause residents to become more cautious and more careful with social interactions due to the possibility of contracting the coronavirus. However, having low R-squared scores (0.291761 and 0.199402) and high MSEs (183096 and 206972) suggest that the data used to draw these correlations are highly varied. This makes sense since the tracking of coronavirus in New York City only begun in February, making the current data set very limited. Secondly, the demographic indicators were estimated for the year of 2016. Since the NYC government’s census data is collected every ten years, the data selected may not accurately reflect the current demographic statuses in each zip code. Additionally, the data source for these indicators still have missing data for approximately 25% of the available zip codes. This is reflected in Figure 1, where a large portion of Queens borough’s zip codes are missing. Therefore, additional data in the future may, perhaps, help increase the reliability and accuracy of these correlations. Nevertheless, it is safe to suggest that Median Income and Nearby Venues are factors in determining the number of positive tests for Covid19 for every zip code in NYC.

Median Income and Nearby Venues were then chosen to construct a model for Positive Tests using Multiple Linear Regression. The coefficients for the two features are -217.8975 and -136.4045, respectively, which were large since the data was standardized before constructing the model. Still, these values are consistent with the features’ correlation coefficients when they were analyzed with the target individually. When the model was compared to the actual data in Table 7 and Figure 6, there was a large difference between them. However, this could be contributed to the dataset’s high variance and limitability. Nevertheless, the model’s higher R-squared score of 0.351279 and lower MSE of 167708 suggest that the Multiple Linear Regression model is more accurate and reliable than comparing the features to the target individually.

After performing clustering on the zip codes based on their venues, the average number of positive Covid19 tests, percentage of positive tests, and nearby venues were calculated and shown in the following table.

Table 9: Averaged positive tests and percentage of positive tests for Covid19

	All Zip Codes	Outlier Zip Codes
Avg. Positive Tests	748	552
Avg. Pct. of Positive Tests	51.28 %	47.43 %
Avg. Number of Venues	45	10

From Table 9, the outlier zip codes, on average, have similar percentages of positive tests as the rest of the zip codes. However, the outlier zip codes have significantly lower average numbers of positive tests and nearby venues. This means that these zip codes were put into different clusters because they have very low numbers of nearby venues. KMeans did not Positive Tests or Median Incomes, therefore a better clustering algorithm may help draw better insights in the relationships between these variables.

Conclusion

The late 2019's novel coronavirus has devastated society on a global scale, with New York City being one of the most affected cities. After analyzing a majority of the city's zip codes, it can be concluded that zip codes with higher median household incomes and number of nearby venues have lower cases of residents testing positive for the virus's resultant Covid19 disease. Future updates of the existing Covid19 and census data will increase the accuracy and reliability of future analyses and help produce more useful insights for the global recovery from this pandemic.

References

[1] https://github.com/nychealth/coronavirus-data/blob/master/tests-by-zcta.csv

[2] http://www.city-data.com/zipmaps/New-York-New-York.html#11239

[3] https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic

[4] https://www.nytimes.com/2020/04/10/nyregion/coronavirus-nyc.html

[5] https://nymag.com/intelligencer/article/new-york-coronavirus-cases-updates.html

[6] https://globalnews.ca/news/6804468/coronavirus-new-york-mass-graves/

[7] https://globalnews.ca/news/6737474/coronavirus-new-york-canada-responses/

[8] https://www.theatlantic.com/ideas/archive/2020/03/america-faces-social-recession/608548/

[9] https://www.livescience.com/why-covid19-coronavirus-deaths-high-new-york.html