COVID-19 IN NEW YORK CITY

1. INTRODUCTION

1.1 BACKGROUND

As the coronavirus spreads across the globe, it appears that each country, each state, each city, is not affected in the same way. Of course, each of them are applying different preventive actions to fight the virus, as it became a political and an economic issue. However, the thing that is very striking about that virus is that, even within the same city, inequalities are huge.

1.2 PROBLEM

New York City has rapidly become the epicenter of the U.S. coronavirus outbreak, paralyzing a city famous for never standing still. It offers an interesting case study to better understand why this disease is unequally affecting the population.

Studies found that inequalities and low income relative to the rest of society is associated with higher rates of chronic health conditions such as diabetes or heart disease. At the same time, the decline of labor unions and the rise of part-time work means that low-income workers have fewer protections. As a result, crises like coronavirus can deepen the gap between the haves and have-nots.

1.3 INTEREST

This report will study the way coronavirus is affecting the different neighborhoods of New York City, related to some demographic statistics and to the disparity of services provided in these neighborhoods.

2. DATA

2.1 DATA SOURCES

Several data sources were used in order to carry out this study. The chosen level of detail is the neighborhood: there are 5 boroughs in New York City, divided in 178 neighborhoods. Then, each datset needs to use the same level of granularity.

2.1.1 Geographic data

Several geographic data sources were used:

- *Zip Codes, latitudes and longitudes and square miles of each neighbourhood*

A Github project was used to collect a CSV file which includes USA zipcode programmable database and geometry information. Link : https://github.com/MacHu-GWU/uszipcode-project

- *Neighbourhood and boroughs names*

A table published by the Department of Health of New York State was used in order to define the name of boroughs and neighborhoods. It was converted in an Excel file.
Link: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

- *New-York City Demographic statistics*

We used web scraping in order to collect demographic statistics. The website provides detailed, informative profiles for every city in the United States by zip code. It allowed us to collect information as the density, the median house value, the median age, the average household size, the average Adjusted Gross Income (AGI) or the average wage, for each neighborhood.
Link: http://www.city-data.com/

- *Map of the neighbourhoods*

A GeoJSON file was collected on the website of the New York University, representing ZIP code Tabulation Areas (ZCTAs) for the entire state, i.e. the geometry of each neighbourhood.
Link: https://geo.nyu.edu/catalog/harvard-tg00nyzcta

2.1.2 COVID-19 data

A Github project was used to collect a CSV file which counts the number of New York City residents by ZIP code of residence who were tested positive for COVID-19 (SARS-CoV-2). The cumulative counts are as of the date of extraction from the NYC Health Department's disease surveillance database. Based on that dataset, we can calculate the percentage of population who were tested positive in each neighborhood.
Link: https://github.com/nychealth/coronavirus-data

2.1.3 Provided services in the neighborhoods

The Foursquare API was used in order to find the number of provided services in each neighbourhood. We gathered the number of venues of different categories within a distance of 500 meters of the center of the neighborhood. The information about the following venue categories were collected:
Medical center, Arts & Entertainment, Nightlife spot, Outdoors & Recreation, Shop & Service, and Food.
Link: https://developer.foursquare.com/docs/

2.2 DATA PREPARATION

We merge all the different dataset on one single dataset based on the ZIP code, using *pandas* Python library, in order to ease the analysis and the data preparation.

Rows containing blank data were dropped. The final dataset contains 174 rows out of the 178 total number of ZIP codes. It represents 97.8% of the initial dataset.

Normalization was applied in order to help mathematical-based algorithms to interpret features with different magnitudes and distributions equally. We used the *StandardScaler()* function to normalize our dataset, from *sklearn* Python library.

3. METHODOLOGY

3.1 EXPLORATORY DATA ANALYSIS

3.1.1 Feature selection

Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For instance, there was a feature giving the population density and another feature giving the total population.

Moreover the correlation of independent variables was inspected. However, we found 2 pairs of highly correlated features (Pearson correlation coefficient > 0.85):

- Average adjusted Gross Income and average wage
- Number of food venues and number of shop services
- Number of art and entertainment venues and number of nightlife spot venues

We finally selected **10 features** for each ZIP code:

- Percentage of population tested positive for COVID-19 [%]
- Density [per square mile]
- Median house value [$]
- Median age [years]
- Average household size [people]
- Average annual Wage [$]
- Number of medical center venues [-]
- Number of art and entertainment venues [-]
- Number of outdoor and recreation venues [-]
- Number of shop and service venues [-]

3.1.2 Exploratory visualization

At a preliminary stage, we examined our data to confirm some of the facts related in the introduction of this paper.
Note: The plots in this section are not displaying normalized data in order to ease the understanding.

First of all, we checked that the different neighborhoods of New York City are not equally impacted by the virus. We used *folium* Python library to display the percentage of population tested positive for COVID-19 within each ZIP code (Figure 1). It allowed us to confirm that, within the city, the spread of the virus is not homogeneous. The variability of the infection rate is about 10-fold, going from 0.003% to 0.03% of the residents.

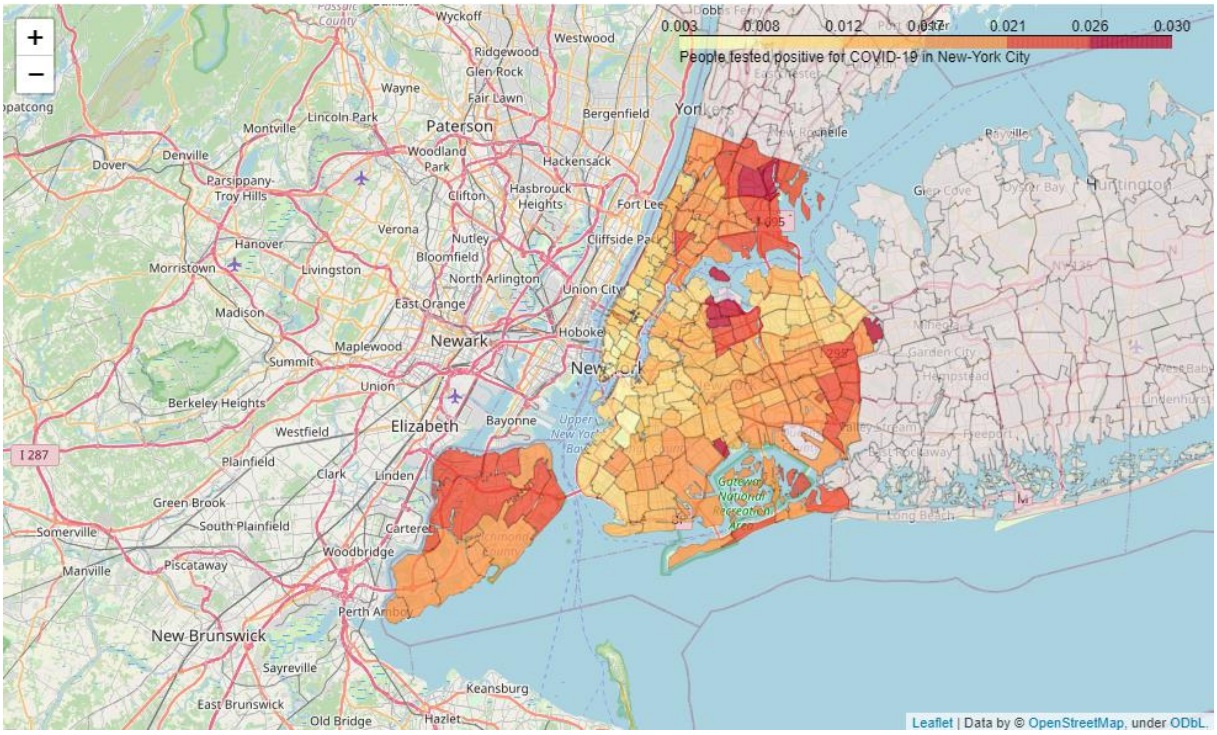

Figure 1 - Percentage of population tested positive for COVID-19 within each ZIP code

Furthermore, we checked that inequalities and low income relative to the rest of society is associated with higher rates of infections. At this stage, a scatter plot with a simple linear regression was generated in order to visualize the correlation between the percentage of population tested positive for COVID-19 and the average annual wage (Figure 2). We notice a negative relationship between the percentage of population infected and the average annual wage. It confirms our hypothesis that the variability of the infection rate is partly explained by demographic features.

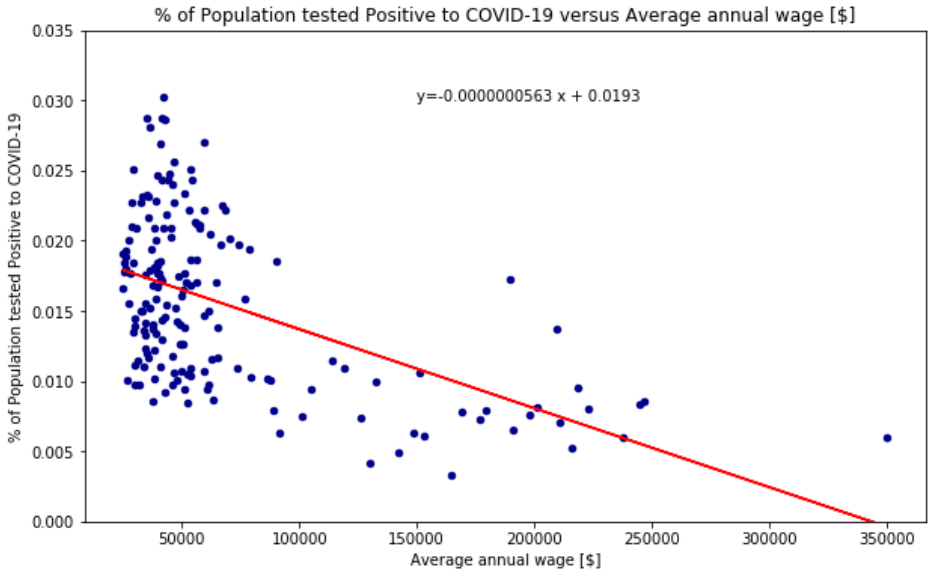

Figure 2- Percentage of population tested positive for COVID-19 versus the average annual wage within ZIP code

The plot clearly shows that the correlation is probably note linear and we can guess that the other features are also playing a role to explain the variability of the infection rate. Therefore, we will undertake more in-depth study using machine learning algorithms.

3.2 PROBLEM MODELING

3.2.1 Unsupervised learning: Clustering

The first model we will deploy is an unsupervised learning algorithm. We will use the k-means clustering algorithm in order to group the neighborhoods into clusters using the selected features.

We complete this task by following these steps:

- Create the model using *KMeans()* function from the *sklearn.cluster* Python library
- Iterate the model for different numbers of clusters (from 1 to 10)
- Chose the right number of clusters looking at the Elbow method (using the *inertia_* method) and Silhouette score (using the *silhouette_score* function from *sklearn.metrics* Python library)
- Run the model with the selected number of clusters
- Visualize the neighbourhoods in New York City and their emerging clusters using the Folium library.

3.2.2 Supervised learning: Regression models

The second model we will deploy is a supervised learning algorithm. We will use the multi-linear regression algorithm in order to try to find a correlation between the percentages of population tested positive for COVID-19 (dependent feature) versus the selected independent features.

We complete this task by following these steps:

- Split the dataset into training and testing sets
- Create the model using *LinearRegression()* function from the *sklearn.linear_model* Python library
- Fit the model with the training dataset
- Predict the test values
- Calculate the Variance score and the R2-score using *sklearn.metrics* Python library

4. RESULTS AND DISCUSSION

4.1 UNSUPERVISED LEARNING: CLUSTERING

After creating and iterating the k-means model, we need to chose the right number of clusters looking at the Elbow method. **K=4** seems to be a good comprise, as shown in Figure 3.

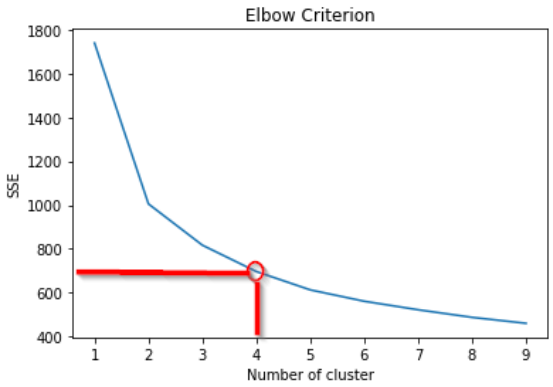

Figure 3 - Chosing the number of clusters with Elbow criterion

Then we run the model with k=4 and we observe the values of the centroids of the clusters, as shown in Table 1. We also can display cluster assignment of each New-York City neighbourhood with a color, as shown in Figure 4.

Table 1 - Centroid values of each cluster

N° cluster	Percentage of population tested positive for COVID-19 [%]	Density [per square mile]	Median house value [$]	Median age [years]	Average household size [people]	Average annual Wage [$]	Number of medical center venues [-]	Number of art and entertainment venues [-]	Number of outdoor and recreation venues [-]	Number of shop and service venues [-]
0	1.9%	37’106	472’218	37.9	2.92	43’482	19.7	7.9	17	37.6
1	0.9%	77’568	1’122’036	37.7	2.00	146’254	40.2	43.7	46	49.2
2	0.7%	63’878	1’801’414	37.0	2.04	234’016	43.7	42.9	45	49.7
3	1.3%	45’736	765’555	37.2	2.70	57’736	27.2	21.0	28	41.3

Firstly, the median age appears not be a discerning feature, as its value is nearly the same between each cluster.

Secondly, we notice that clusters which are the most affected by COVID-19 are the one with:

- the lowest average annual wage ;
- the lowest median house value ;
- the lowest number of medical center venues ;
- the lowest number of art, entertainment, recreation, outdoor, shops, and services venues ;
- the highest average household size.

Then, looking at these characteristics, clusters 0 and 3, which are the most affected by COVID-19, appears to be linked to the poorest neighborhoods, with the lowest access to medical facilities.

On the contrary, clusters 1 and 2, which are the most affected by COVID-19, appears to be linked to the wealthiest neighborhoods, with the highest access to medical facilities.

This study confirms that inequalities and low income relative to the rest of society is associated with higher rates of health conditions, including COVID-19.

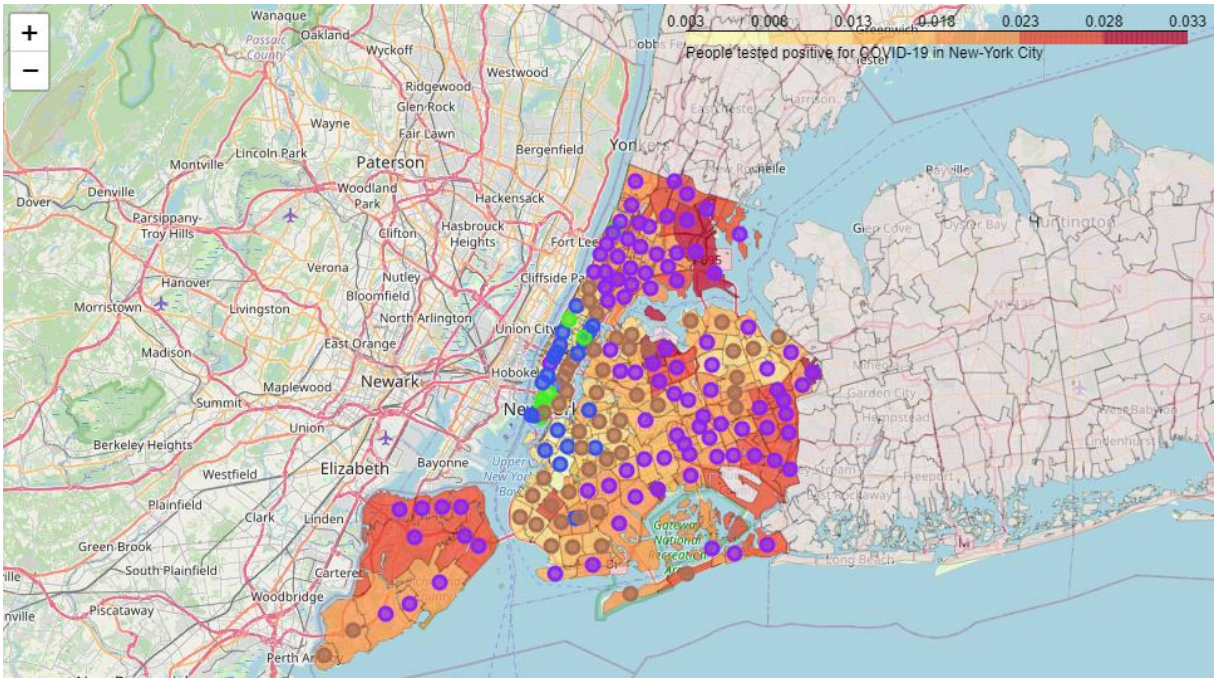

Figure 4 - Cluster assignments of each New-York City neighbourhood

4.2 SUPERVISED LEARNING: REGRESSION MODELS

We wanted to confirm those results using a supervised learning algorithm. We performed a multi-linear regression algorithm in order to try to find a correlation between the percentages of population tested positive for COVID-19 (dependent feature) versus the selected independent features. The results are shown below in Table 2 and Table 3.

Intercept	Density [per square mile]	Median house value [$]	Average household size [people]	Average annual Wage [$]	Number of medical center venues [-]	Number of art and entertainment venues [-]	Number of outdoor and recreation venues [-]	Number of shop and service venues [-]
0.00	-0.04	-0.39	0.15	-0.17	-0.12	-0.18	-0.14	-0.06

Table 2 - Fitted coefficients of the multi-linear model

Variance score	0.66
R2-score	0.43

Table 3- Evaluation of the multi-linear model

As expected, the variance and R2-score are not really good. Indeed, it means that we need other features in order to perfectly predict the percentages of population tested positive for COVID-19 in each neighborhood. The COVID-19 rate is not only linked to demographic features but also to medical features like the rate of comorbid medical conditions in the population, features related to jobs, etc.

However, we find the same trends in the correlation between the rates of infected population, against the poverty and the access to medical facilities.

5. CONCLUSION

This report tried to find an explanation why COVID-19 is not equally affecting the different neighborhoods of New York City.

We demonstrated that some of the features which are playing an important role are the poverty and the access to medical facilities. Indeed, people with low income are living in less favorable conditions (more people in each home) and have higher rates of chronic health conditions such as diabetes or heart disease which are considered as comorbid medical conditions. Moreover, in general, they held less-skilled jobs which are more exposed to the virus (cashier, cleaning person, caregiver, delivery man, etc.). Then, this study could be developed by exploring further features linked to the health condition and the jobs occupied by these populations.