IBM

Data Science Capstone Project Week 4

Ali Ahmad, MD

1- Introduction

The coronavirus 2019 (COVID-19), caused by SARS-COV2 virus, is a pandemic that wreaked havoc on multiple countries, including Italy, Spain and the United States of America. It was surprising to see how many states and cities progressed quickly, while others are still relatively better off. Despite Stay-at-home orders and other laws passed to mitigate the spread of the disease, it is impossible to chase down all people and stores. New York City is one of those examples, where the disease struck hard and it was really tough to contain.

In this project, we aim to determine the characteristics of the zip codes with most cases.

2- Data

The data to this project has been retrieved and processed through multiple sources.

2.1 Data sources

The following were used to extract relevant data

1- Type of venues within a certain radius of each borough (using Foursquare API)

2- Extract all NYC Zip codes by Borough and Neighborhood from (https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm)

3- Extract current corona cases per zip code from (https://github.com/nychealth/coronavirus-data/blob/master/tests-by-zcta.csv)

4- Population density per borough (https://en.wikipedia.org/wiki/Template:NYC_boroughs)

5- Coordinates of every zip code from: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?refine.state=NY&location=7,42.79,-75.84997&basemap=jawg.streets

2.2 Methods of extraction: We will utilize requests, pandas and BeautifulSoup to scrape the pages for the tables. We will then use dataframe manipulation to merge the different datasets to have one master dataset (df).

3- Data preparation

After joining all relevant data as shown in the coding file, we used foursquare API to extract venue data for NYC. After that, we created a dataframe regarding the most common venue type in every zip code. Finally, we added the number of positive cases and population density information and sorted by positive cases.

4- Results and discussion

Our analysis shows that population density is not the ultimate factor determining risk of transmission as none of the top 10 zip codes with cases were in Manhattan, the Borough with the highest population density. Furthermore, we showed that there is no strict pattern in common venues related to number of corona virus cases. In general, there seems to be a trend to having more restaurants/bakery/coffee shops in the top 5 venues in the borough with most cases.

Our analysis is limited because there was no formal statistical interrogation of data. Furthermore, some zipcodes did not report corona virus cases, but we doubt this would affect out data (since they probably do not have a lot of cases). Finally, other risk factors that determine corona virus spreak should be added to the model.

5- Conclusion

The corona virus pandemic origin is yet to be understood, and ways to suppress spread should be investigated further. The origin of cases in the states and the amount of local transmission versus importation of diseases is not known. In this report we show that population density is not the most important factor of transmission, and that the amount of food venues might increase transmission.