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 diseases 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 and to use unsupervised machine learning to understand if there is a certain pattern other than the known risk factors of spread (population density, tourism...).

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

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).