Coursera Capstone Project:

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

Son Le

sonle.h96@gmail.com

April 20th, 2020

Content

- Introduction
- Business Problem
- Data
  - Acquisition
  - Cleaning
  - Visualization
- Analysis
  - Individual Indicators
  - Multiple Linear Regression
  - Nearby Venues
- Conclusion
- References

Introduction

- The 2019-2020 novel Coronavirus has become a global pandemic, causing devastating economic and social damage to virtually every country on Earth.
- The virus leads to the Covid19 disease which is dangerous especially towards young children, the elderly, and those with compromised immune systems.
- The virus is highly contagious and spread through social interactions and contact, with symptoms, akin to the common cold and flu, arising after 2-14 days of those infected.
- New York City has been one of the most negatively affected cities in the world.

Business Problem

- This study is an attempt to explore the relationship between each New York City zip code's, a geographically defined community, demographic indicators and the area's positive Covid19 tests.
- The indicators being analyzed are:
  - **Population density**: measured in number of people per square mile
  - **Median household income**: measured in USD
  - **Number of nearby venues** within the zip code's 500-meter radius

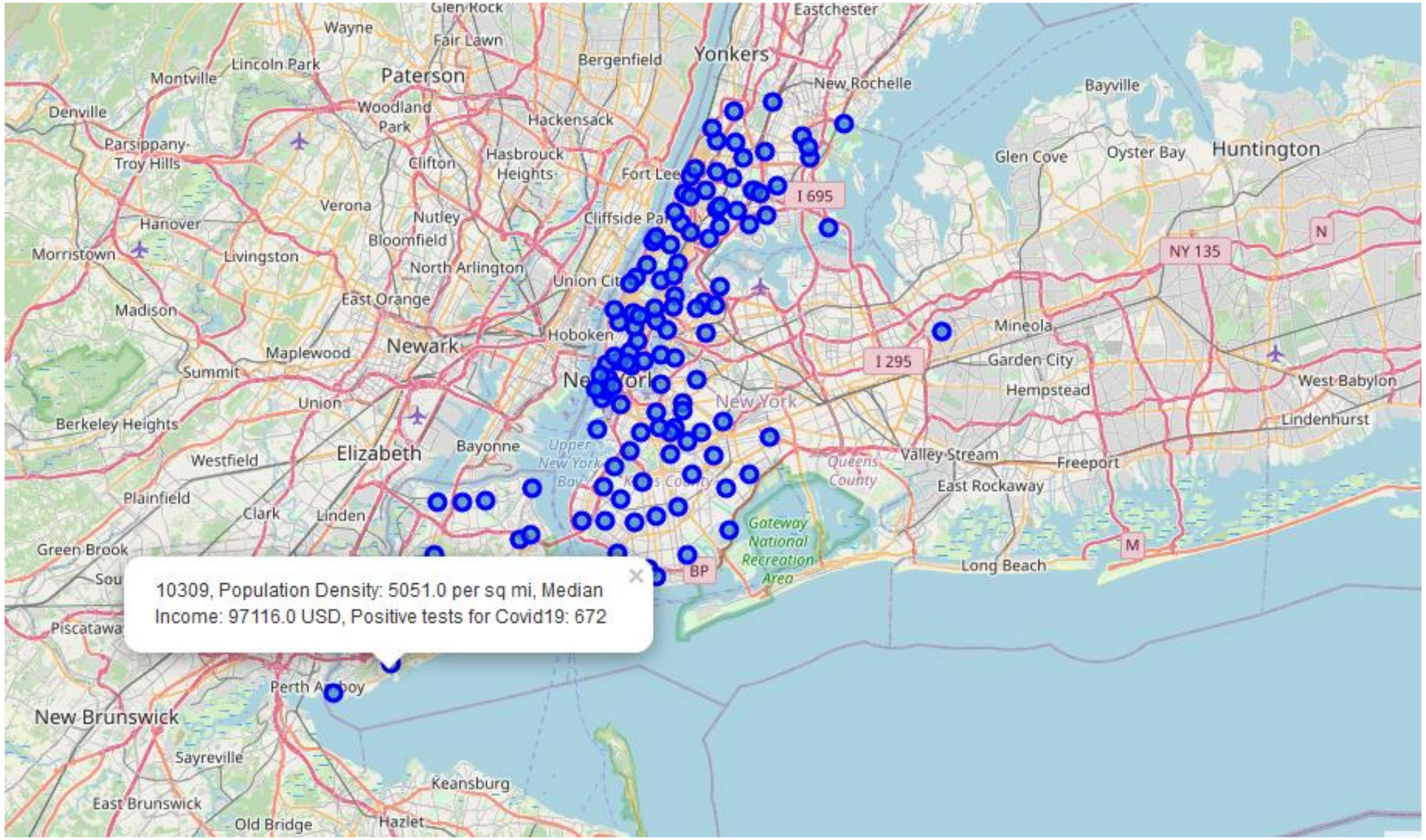
Data Acquisition

- The data needed for this project will consist of:
  - Number of recorded total and positive tests for Covid19 in New York City in each zip code using web-scraping from NYC Health.
  - Current, or estimated demographic data for population density and median income in each NYC zip code using web-scraping from city-data.org
  - Each found zip code’s coordinates using the **OpenCage API**
  - Venues within each zip code’s 500-meter radius using the **FourSquare API**.

Data Cleaning

- Demographic data was derived from a 2016 estimation
  - The NYC government takes a census every 10 years, with the last census taken in 2010.
- Zip codes with missing or incomplete data were removed.
  - A large portion of Queens borough will not be considered.

Data Visualization



10309, Population Density: 5051.0 per sq mi, Median Income: 97116.0 USD, Positive tests for Covid19: 672

Analysis - Individual Indicators

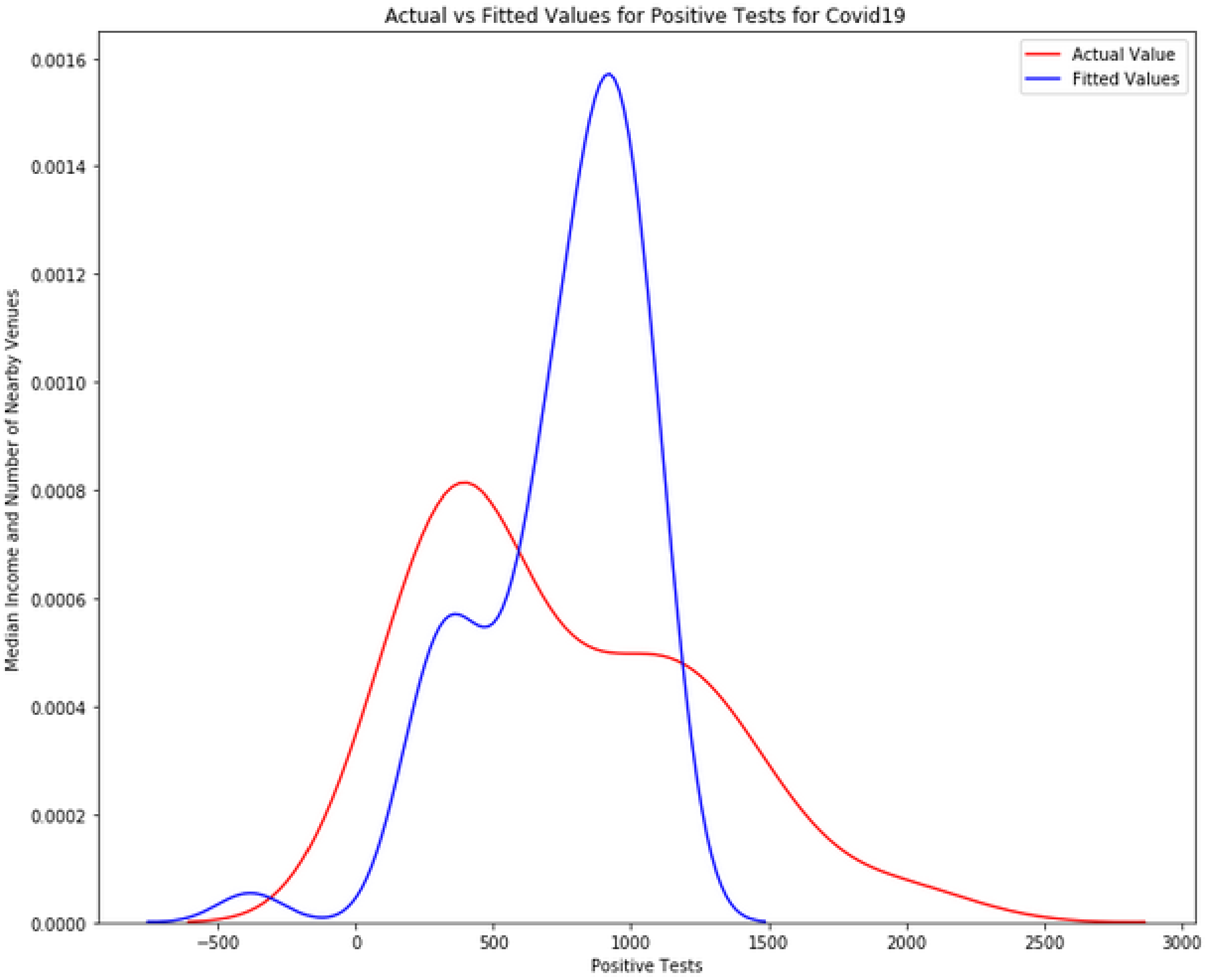
	Population Density	**Median Income**	**Nearby Venues**
Correlation coefficient	-0.084131	-0.540149	-0.446545
p-value	0.354874	1.13E-10	2.26E-07
R-square	0.007078	0.291761	0.199402
MSE	256692	183096	206972

- Population Density was dropped from further analysis due to no correlation, low p-value and R-squared score, and a high MSE.

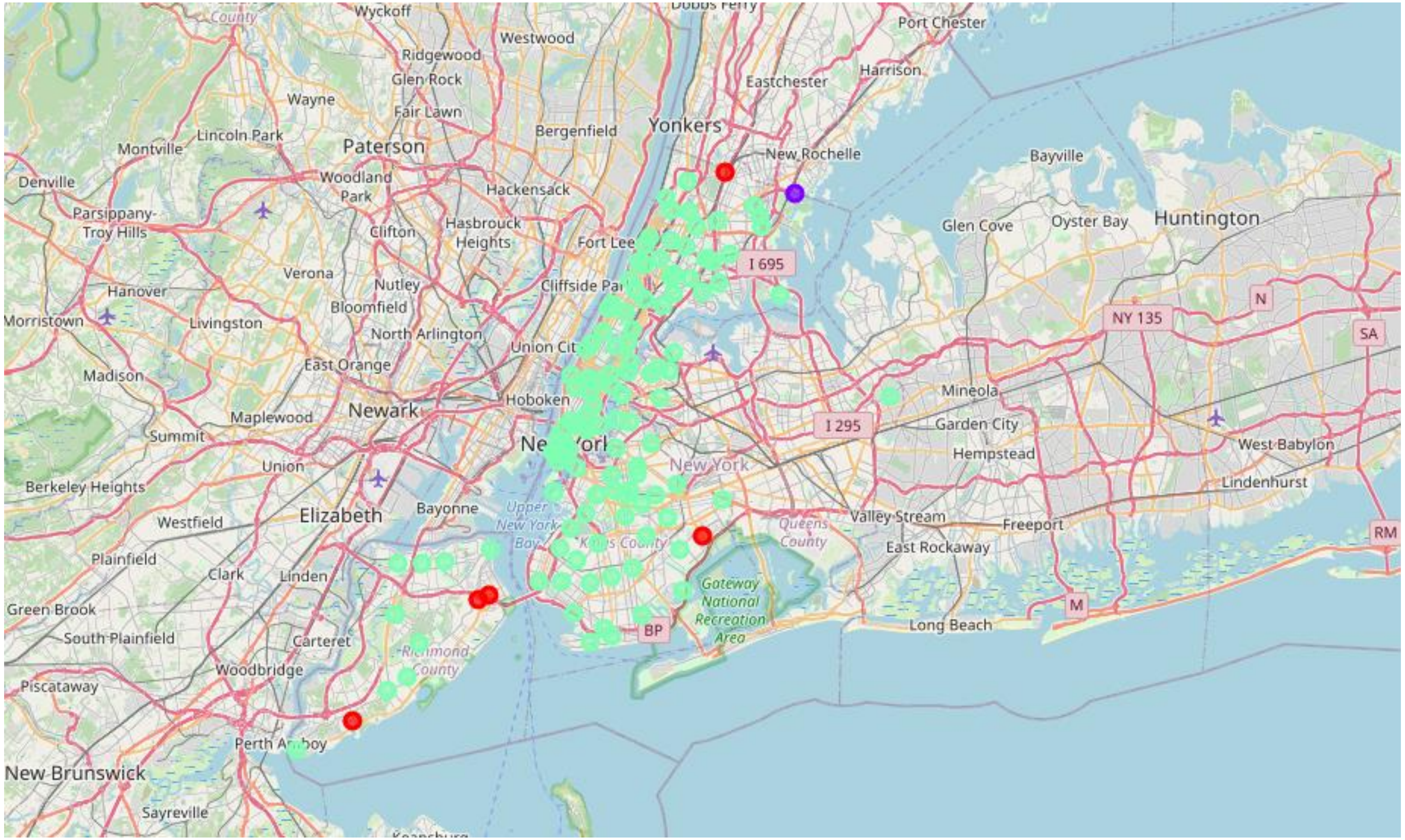
Analysis – Multiple Linear Regression

Median Income correlation	-217.8975
Nearby Venue correlation	-136.4045
R-square	0.351279
MSE	167708

- Population was not part of the feature variables.
- Poor model results from data's high variance and limitability.
- Model was very inaccurate from 600 positive tests and beyond.



Analysis – Nearby Venues



- Two outlier clusters formed because those zip codes have very small numbers of nearby venues

- KMeans probably was not the best clustering method for this case.

Conclusion

- 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