Estimating Covid-19 Spread

Michael Waldmann

June 29th, 2020

1. Introduction

1.1 Background

As of June 29th, 2020, Covid-19 has killed approximately 500,000 people worldwide. The pandemic has been controlled in some areas while in others it is accelerating. Currently in the United States we are seeing a resurgence in the number of reported cases in certain portions of the country. A primary reason for the resurgence in cases in the United States in due to different state governments deciding to re-open their economies. Many state governments have decided to re-open in stages, e.g. stage 1 could be the re-opening of hair salons and parks, stage 2 could be outdoor dining and places of worship, while stage 3 could be full resumption of all business and places. In this report we will try to gain some insight into which establishments lead to higher levels of Covid-19 spread.

1.2 Problem

What types of business venues lead to the highest levels of Covid-19 infection? We will need to seek out geographic information that provides the rate of Covid-19 infection specific to certain areas as well as the types of venues that are present within that same area.

1.3 Interest

Government officials would be interested in knowing which types of establishments are related to higher levels of Covid-19 being spread. These decision makers are torn between re-opening the economy to provide a better standard of living for their constituents and keeping the economy shut to protect those same constituents from contracting the potentially fatal disease. If certain venues could be marked as being higher risk, government officials would be better able to balance which types of venues to re-open while keeping more dangerous ones shut.

2. Data Acquisition and Preparation

2.1 Data Sources

The following link was used to source Covid-19 death rates:

https://github.com/nychealth/coronavirus-data/blob/master/data-by-modzcta.csv

	NEIGHBORHOOD_NAME	MODIFIED_ZCTA	COVID_DEATH_RATE
0	Chelsea/NoMad/West Chelsea	10001	97.61
1	Chinatown/Lower East Side	10002	200.64
2	East Village/Gramercy/Greenwich Village	10003	61.34
3	Financial District	10004	27.39
4	Financial District	10005	23.82

Fig1. Example of data from GitHub source

The data provided lists all the zip codes for New York City as well as the corresponding death rate for those zip codes and which borough the zip code belongs to (Manhattan, Bronx, etc.). We will use death rate instead of case rate because it is much more likely that a case goes unreported (asymptomatic, lack of testing availability, personal choice to not get tested, etc.) than it is for a death to go unreported.

In addition to the death rate data will we also need to collect information on the types of venues which are present within each zip code area. To do this we will use Foursquare.com's 'explore' end point API. The API accepts the zip code as a location as well as a radius parameter set by the user. The API returns a large data set containing many fields of information on all the different venues located near that area.

2.2 Data Filtering, Cleansing, and Refining

Geographically we have decided to only look at the borough of Manhattan which contained 44 different zip codes. The reason for this decision is to try and remove other factors that could influence the Covid-19 infection rate. For example, one borough may have been hit harder than another borough which would cause all its corresponding zip code areas to be inflated relative to other boroughs. After the data for the non-Manhattan boroughs was dropped, we filtered through all the venue data obtained from the Foursquare API.

We specifically care about the type/category which a venue falls into, not the specific name of the venue itself. We were able to append the category (bar, grocery store, gas station, etc.) of each venue located within a zip code area to the dataset. Because we need quantitative data, we one-hot-encoded this categorical data. In the end there were 249 different categories of venue types associated with all the 44 zip codes. We normalized the

data such that each row of data sums to 1.00. This means that each entry of the 249 fields represents the prevalence of that type of venue for that specific zip-code area as a percentage.

Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Assisted Living	Athletics & Sports	Australian Restaurant
0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.027778	0.0	0.0	0.0	0.027778	0.0	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.043478	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.043478
0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.056604	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000

Fig2. Example of data after being normalized; column names are venue categories listed in alphabetical order

Later on, we decided that 249 fields were a lot of data and that were perhaps better served by narrowing down the number of fields. To do this removed all fields whose maximum prevalence was below 2% across the entire data set. This reduced the number of fields from 249 to 28.

3. Modeling

We will attempt to tackle this problem with two different types of models. The first will be a predictive model which is fed supervised data. For this we will use multiple linear regression. For the second approach we will use the K-means classification model which is fed unsupervised data.

3.1 Multiple Linear Regression

For this model, the target variable will be the Covid-19 death rate whereas the input variables are the 249 different categories of venue type. Our hope is that the fitted model will be able to predict the Covid-19 death rate. If we observe a high r^2, this will mean that our input variables explain some of variation in covid-19 death rate. This will let us know whether we are on the right track in terms of identifying a valuable correlation between venue prevalence and Covid-19 death rate.

We used 75% of the 44 data points to fit the model and then tested the model on the remaining 25% of the data points. The following are the results of shape of the input data used to fit the model as well as the performance of the model on the testing data:

```
The shape of the input data is (33, 249)
The R^2 value for the MLR is -0.22
```

A negative r^2 is not intuitive. How could a squared value be negative? All that this means is that our model does a worse job of predicting the target variable than if we were to use the mean of the target variable for the predicted value. This is a very bad outcome.

We decided to reduce the number of fields to 28 as explained in section 2.2 above, and rerun the model. The following were the results:

```
The shape of the input data is (33, 28)
The R^2 value for the MLR is -2.08
```

This is not a good and could be our first sign that venue category prevalence is not good predictor for Covid-19 death rate. However, we will still proceed with K-means in the next section.

3.2 K-Means

This test will help us label different clusters of zip code areas. After labeling the zip codes, we will see if the different clusters have different Covid-19 death rates. We will use ANOVA to see if the differences are statistically significant.

For this model there is no target variable as the data is unsupervised. The purpose of K-means is to produce clusters of similar data. The 'k' in k-means is the number of clusters to be produced by the model. We decide how many k clusters to produce by seeing how the error changes different numbers of clusters. Error is measured as the within cluster sum of squares. This means that each cluster has a mean value and each cluster is individually scored based on the sum of squares against itself, the error of all the clusters is then added together. Adding more clusters will always reduce the error so we are looking for a k value where error reduction drops off. Below is the result of running different values for k:

```
Number of clusters = 1 and score = 10955.999999999993
Number of clusters = 2 and score = 10447.548979960602
Number of clusters = 3 and score = 10068.781435850431
Number of clusters = 4 and score = 9736.763396396651
Number of clusters = 5 and score = 9271.781034171338
Number of clusters = 6 and score = 8917.298556253632
Number of clusters = 7 and score = 8608.862212499216
Number of clusters = 8 and score = 8182.487171712671
Number of clusters = 9 and score = 7864.066357895817
```

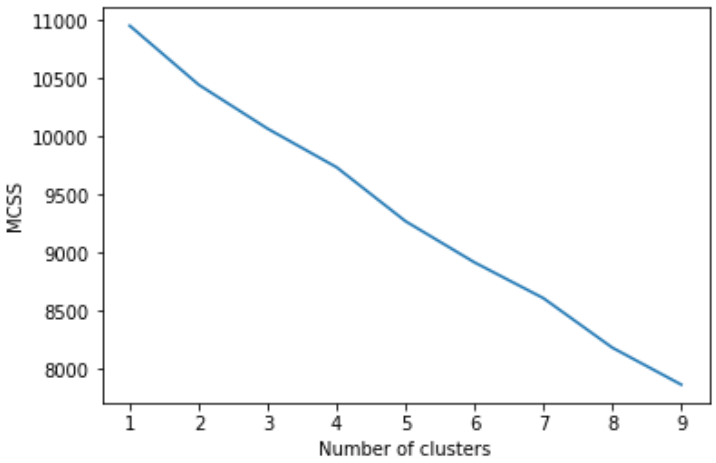

Fig3. Error versus cluster number for our k-means model

The fact that we do not see a kink in the graph is already a bad sign however we proceeded with using k=6. After running the model, we labeled our data. Below is a boxplot of Covid-19 death rate amongst the different clusters:

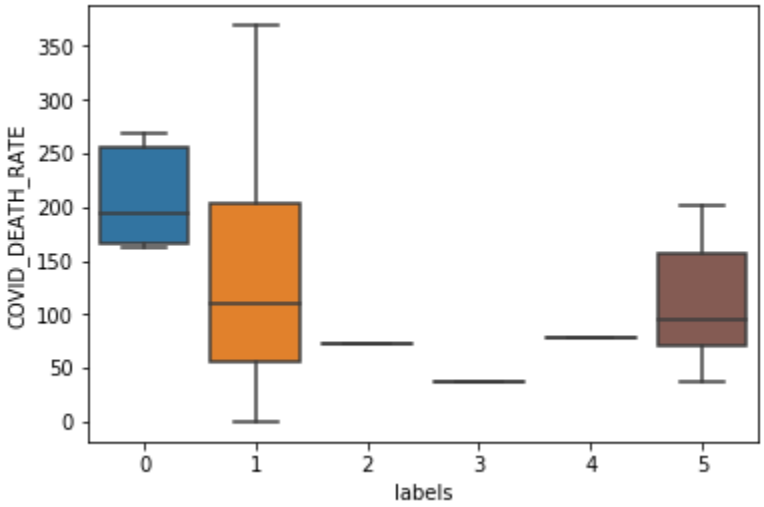

Fig4. Boxplot of Covid-19 death rate amongst the different clusters. Note that labels 2, 3, and 4 each only have one data point.

The clusters do seem to have different values which may be statistically significant however we must remember how small the data set is. To test for a difference of means we will use ANOVA:

```
F_onewayResult(statistic=1.1407122796748182e-45, pvalue=1.0)
```

The p-value is 1 which indicates there is no statistical difference amongst means.

4. Conclusion

The negative value for r^2 from our linear model tells us that our input variable provides no value in predicting our target variable. The p-value from our ANOVA test can be any more conclusive. Based on our data set there is absolutely zero evidence there is difference in covid-19 death rate based off the types of venues found within a geographical area. This does mean that covid-19 is not spread at different rates at different establishments. It is certainly spread at a higher rate where proximity and shared surfaces are common, e.g. at a bar.

5. Issues with Methodology

- Covid-19 death rate data is taken as of June even though lockdowns occurred much sooner as venues were closed in March. This means that we were looking for a link between the spread occurring at these venues pre-lock down and seeing how it affected death rates months later.
- Many people may contract Covid-19 at a venue outside of their zip code area.
- We did not consider venue density or count, only the 'relative' prevalence of a venue type to one's area. For example, an area with only 2 venues one of which is bar would have a 'bar' value of 0.50 whereas an area with 100 venues and a 'bar' value of 0.05 would have 5 times as many bars.
- Demographics are not considered. Certain areas of Manhattan are home to people of different types of incomes, age, etc. One area may have a higher proportion of nurses who still go to work and risk contracting the disease. Another area may have many young or old people who disproportionally choose to ignore social distancing measures.