STEM Fellowship Big Data Challenge 2020

Correlation Between Population Density and Covid-19 Impact

Raymond Romaniuk[1]

[1]Brock University

June 1, 2020

Abstract

The Coronavirus Disease 2019 (COVID-19) has significantly impacted individuals across the planet. This paper aims to determine which factors within our society had an impact on the mortality rate in areas like New York City and Italy, and why these two locations were dramatically affected. COVID-19 data was extracted from the Johns Hopkins Coronavirus Map along with provincial/state level data from Statistics Canada, the US Bureau of Economic Analysis, the US Census Bureau, Stats America and the Centers for Disease Control and Prevention. Python was used to manipulate the data and position it in an optimal format for analysis. Linear regression models were then employed, in SAS, to determine how smoking[1][2], obesity rate[3][4], age[5][6], household size[5][7], population density[8][9] and GDP per capita[10][11] affected the COVID-19 outbreak in a given area. The models were tested with different data sets to determine the optimal model for predicting the ramifications of COVID-19 outbreaks in certain areas. Principal Component Analysis was also used in SAS to explore the correlation between the different variables. Aside from statistical analysis SAS was used to plot the linear regression model and display 95% conifidence and prediction limits. Power BI was then used for supplementary visualizations. Upon analysis of the results it was aparent that population density had the greatest impact on the state of a pandemic within a given city. This finding helps explain why a city like New York was overwhelmed with cases. Another interesting discovery was the relationships between population density and obesity rate. Intuitively, one would think that a higher obesity rate in an area would cause for a higher percentage of the population to fall ill, however this is not the case. It turns out that the states with higher population density actually had some of the lowest obesity rates and these states had a higher percentage of mortality than those with high obesity rates and lower population density. The population density issue has come to the forefront during the current pandemic, but with that issue a solution has also been created and that is the surge of remote work.

Keywords
Coronavirus Disease 2019, COVID-19, Pandemic, New York

1 Introduction

It is very rare that something comes along and has a significant impact on a global scale, however COVID-19 has done just that. From Asia, to Europe, the Middle East and all the way to the Americas, almost every inhabitant on our planet has felt its impact. Some have been affected by the disease much more than others and that is what this paper will be exploring.

This virus began as an afterthought to most people when it was initially reported that, "There has been no evidence to date that this illness, whatever it's caused by, is spread easily from person to person" by Theresa Tam, the Chief Public Health Officer of Canada [12]. The World Health Organization later echoed the same sentiment on January 14th based on preliminary details it had received from China [13]. Fast forward two weeks and these initial reports are proven false by the first person to person transmission of the virus in the United States [14]. Now, at the end of May, we find ourselves with a downward trend in cases [15] and the potential to return back to our daily lives on the horizon, after a two and a half month quarantine.

With the worst behind us this begs the question, why were locations like New York City, which would be ranked sixth in the confirmed cases by country rankings (excluding the United States), and Italy, who ranks second in total deaths among European countries based on the Johns Hopkins Coronavirus Map, impacted so severely? How can one try to prevent themselves from being in harms way in a future global pandemic?

With the use of data from the Johns Hopkins Coronavirus Map, along with data from Statistics Canada, the US Bureau of Economic Analysis, the US Census Bureau, Stats America and the Centers for Disease Control and Prevention, I was able to explore what made a specific location more likely to be heavily impacted by COVID-19. To perform this analysis I utilized the power of Python to manipulate the data, SAS to perform statistical analysis and Power BI to help visualize the data.

Upon review of the linear regression models and correlation coefficients from the principal components analysis, two things stood out, more densely populated areas were hit harder by the virus and lower state obesity rates did not mean that inhabitants of those areas were safe from infection. A possible solution to this population density problem in New York City is the opportunity for remote work. This pandemic has demonstrated to employers how valuable remote work can be with the potential for savings on their current office space. Employees also would have the ability, while working remote, to move out of crowded cities and into a safer environment in case of another pandemic-like situation.

2 Materials & Methods

The main COVID-19 data set used for this project is the data feeding into the Johns Hopkins Coronavirus Map and found at the Johns Hopkins GitHub. This data set contains daily updates on Confirmed Cases, Active Cases and Deaths. The data is broken down at a provincial level for Canada and at a slighty more granular city level for the United States, with data beginning on January 22nd.

Python was used to pull in Johns Hopkins daily data refreshes and merge it with the province/state level data sets about smoking, obesity rate, age, household size, population density and GDP per capita. Finally Python was used to manipulate the data and perform necessary calculations before it was saved as CSV files and ready for analysis in SAS.

Once saved as SAS libraries the data was incorporated into linear regression models, with 95% confidence and prediction limits. The coefficients were analyzed to better understand how each variable impacted the predicted mortality rate, infection rate and infected mortality rate. A principal components analysis produced the correlation coefficients and illustrated the correlation between the variables.

Finally Power BI was used to visualize trends that were noticeable within the statistical analysis performed in SAS.

3 Results

Figures 1 through 4 each display a linear regression model with 95% confidence and prediction limits. Figure 1 and Figure 2 demonstrate the difference between the affect of population density on the population mortality rate in both North America and Canada.

Figure 1: North America affect of population density on population mortality rate.

Figure 3 and Figure 4 are similar to Figure 1 and 2, however they illustrate the impact of obesity rate on North American and Canadian population mortality rate.

Figure 5 displays the correlation coefficient matrix of the variables in the data set that was obtained through principal component analysis.

4 Discussion

It's not everyday that a virus has the type of impact that COVID-19 has had, and at a global scale no less. Simply trying to comprehend the number of viruses that exist on our planet is a challenge. National Geographic reported an estimate of 10 nonillion viruses on Earth [16], as National Geographic explained that's, "enough

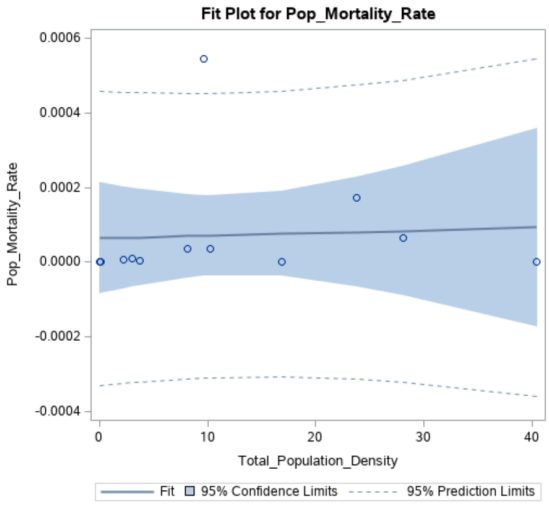

Figure 2: Canada affect of population density on population mortality rate.

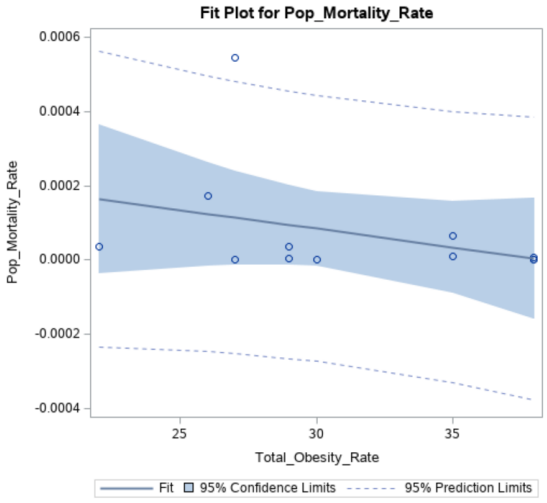

Figure 4: Canada affect of obesity rate on population mortality rate.

Figure 3: North America affect of obesity rate on population mortality rate.

Correlation Matrix

-	Average Age	Average Household Size	GDP Per Capita	Obesity Rate	Percent of Population Infected	Population Density	Smoking Rate
Average Age	1.00	-0.57	-0.40	0.00	-0.06	-0.22	-0.13
Average Household Size	-0.57	1.00	0.06	-0.16	0.03	-0.13	-0.09
Gdp Per Capita	-0.40	0.06	1.00	-0.33	0.51	0.86	-0.09
Obesity Rate	0.00	-0.16	-0.33	1.00	-0.25	-0.24	0.68
Percent of Population Infected	-0.06	0.03	0.51	-0.25	1.00	0.41	-0.10
Population Density	-0.22	-0.13	0.86	-0.24	0.41	1.00	-0.10
Smoking Rate	-0.13	-0.09	-0.09	0.68	-0.10	-0.10	1.00

Figure 5: Principal component analysis correlation coefficient matrix for average age, average household size, GDP per capita, obesity rate, percentage of population infected, population density and smoking rate.

to assign one to every star in the universe 100 million times over". That is a staggering number and it puts into perspective how lucky the human race is, as a whole, that not all viruses are capable of causing such a significant amount damage.

One way an individual could try to protect themselves from future pandemics is by relocating to an area with a lower population density. In both North America and Canada's case an increase in population density will increase the population mortality rate, even if that is very slightly for Canada in Figure 2. This makes sense conceptually as the more densely packed the population, the more contact an individual will have with others and this creates a greater opportunity for a virus, like COVID-19, to spread.

Figure 3 and Figure 4 demonstrate a scenario that is not as intuitive as the population density case. Obesity rate actually has a negative affect on an areas population mortality rate. One would expect that a population with a lower obesity rate would fair better with a virus than one with a higher obesity rate, this is not the case.

Figures 6 through 9 begin to explain why obesity rate has a negative impact, as it increases, on population mortality rate. In Figure 6 the top 5 most dense states, population wise, have some of the lowest obesity rates, with not one of them being within the 40 highest obesity rates. Looking at Figure 7, although not as extreme, states with a higher obesity rate have a lower population density rank.

Looking at the correlation coefficients in Figure 5, it appears that obesity rate has a weak negative correlation (-0.33) with GDP per capita, meaning that as GDP per capita increases it is expected that obesity rate will decrease. This could be due to a plethora of factors that are not in the model, for example a higher educated person earning a higher GDP.

Figure 5 also shows that GDP per capita has a strong positive correlation (0.86) with population density, which brings everything full circle and explains how obesity rate and population density are negatively correlated.

State	Population Density	Population Density Rank	Obesity Rate Rank
District of Columbia	11011	1	50
New Jersey	1218	2	47
Rhode Island	1021	3	41
Massachusetts	871	4	48
Connecticut	741	5	44

City	Total Population Density
New York City	27013

Figure 6: Top 5 states by population density and their corresponding obesity rate rank, with the inclusion of New York City.

State	Obesity Rate (%)	Obesity Rate Rank	Population Density Rank
West Virginia	39.50	1	30
Mississippi	39.50	2	33
Arkansas	37.10	3	35
Louisiana	36.80	4	24
Kentucky	36.60	5	23

City	Total Obesity Rate (%)
New York City	27.60

Figure 7: Top 5 states by obesity rate and their corresponding population density rank, with the inclusion of New York City.

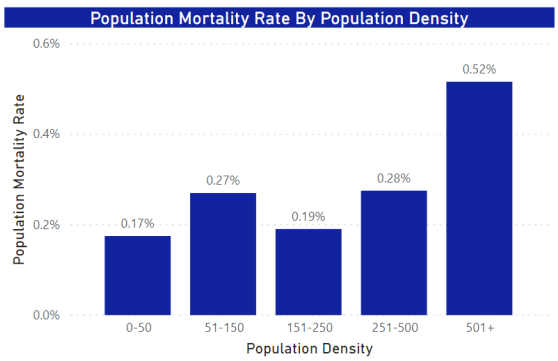

Figure 8: Population mortality rate for each population density group.

To confirm that population density is one of the large forces causing a large amount of

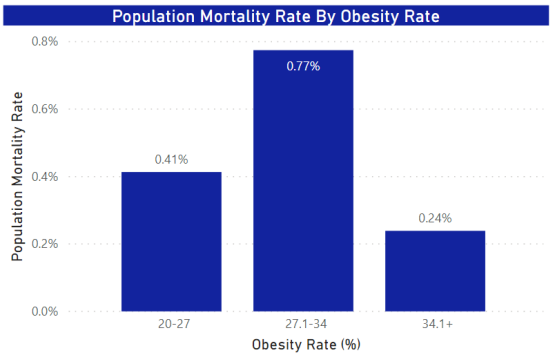

Figure 9: Population mortality rate for each obesity rate group.

COVID-19 cases Figure 10 displays New York City's ranking in each category compared to every state, province and territory. For the five New York City attributes that contribute to its high percentage of population infected, population mortality rate and infected mortality rate, only population density and GDP per capita are ranked significantly high. This confirms that population density had the greatest impact as to why New York City was a COVID-19 hot spot.

New York City Rankings Among States, Provinces & Territories

-	Population Infected (%)	Population Mortality Rate	Infected Mortality Rate	Population Density	GDP Per Capita
New York City	1	1	1	1	3

-	Average Age	Household Size	Obesity Rate
New York City	58	23	50

Figure 10: Ranking New York City's attributes against all states, provinces and territories.

Another interesting case is the outbreak of COVID-19 in Italy. Could population density also be a significant factor in the high population mortality rate of Italy?

Using the full North American Covid-19 data set I created a linear regression model with population density (Figure 11) to test this hypothesis.

Variable	DF	Parameter Estimate
Intercept	1	0.00020213
Total_Population_Density	1	7.260845E-8

Figure 11: Linear regression model used to predict the population mortality rate in Italy.

After exploring the Italian COVID-19 case data I made the decision to attempt to predict the population mortality rate for the region of Lombardy, the Italian epicenter. Using the

model in Figure 10 and the population density of the region of Lombardy it was predicted that the population mortality rate would be 0.0282%, equating to a total of 2845 deaths. In actuality the number of deaths was almost seven times that, at 16079.

As a second effort I tried using Milan, the largest city in the Lombardy region's, population density to determine whether it was possible that the region being spread out more than the main city was causing a number far too low to be estimated. After running the model for the population density of Milan, 0.1622% was the estimated population mortality rate, equating to a total of 16360 deaths, almost exactly the amount currently reported. This could have all come down to the numbers luckily adding up, however without more granular data available it's hard to tell in this case.

The linear regression model's in Figure 1 through 4 are by no means perfect and potentially could not be predicting outcomes accurately with quite a bit of error. None of the R-squared values are greater than 2.2, meaning that the data does not entirely fit the linear model and there is room for error in its predictions. For this study, data was pieced together from quite a few different data sets and merging all this data that was collected and calculated by different entities could potentially cause error in the analysis.

Conclusions

It is still unknown how the world will bounce back from the COVID-19 pandemic and what effects it will have on our society and each of us who live in it.

This paper demonstrated that population density seemed to be the most influential factor in the severity of the pandemic on a given location. In the future it may be wise, for individuals who can, to move out of densely populated cities to curb the spread of potential infectious diseases. This could cause a substantial increase in remote workers which is safer for employees under these circumstances and has the potential to save an employer money on large office spaces and costs that go with them.

The current data also showed an interesting case of population density and obesity rate being negatively correlated, meaning that as population density increases obesity rate is expected to decrease and vice versa.

There will be no shortage of topics to dig into regarding COVID-19 data and it will be exciting to see the possible revelations that could come out of that analysis. Most importantly how might those discoveries impact the world as we know it?

Acknowledgements

I would like to thank the Brock Math Department for sending the initial email about getting involved in this competition. I'd also like to STEM Fellowship for the opportunity to compete and learn along the way. Also thank you to my mentor Mohamad El-Hajj for our initial conversation.

References

[1] Statistics Canada. Canada health characteristics, 2017-2018.

[2] Centers for Disease Control and Prevention. Map of cigarette use among adults, Dec 2018.

[3] Statistics Canada. This infographic looks at obesity in canadian adults for 2016 and 2017. it details how obesity varies by age, education, diet, landed immigrant status and province using data from the 2016 and 2017 canadian health measures survey and the 2017 canadian community health survey., Oct 2018.

[4] Centers for Disease Control and Prevention. Adult obesity prevalence maps, Oct 2019.

[5] Statistics Canada. Census profile, 2016 census ontario [province] and canada [country], Aug 2019.

[6] Stats America. Usa states in profile: Median age, 2018.

[7] US Census Bureau. Social characteristics in the united states, 2018.

[8] Statistics Canada. Population and dwelling count highlight tables, 2016 census, Feb 2019.

[9] US Census Bureau. State population totals: 2010-2019, Dec 2019.

[10] Statistics Canada. Gross domestic product, expenditure-based, provincial and territorial, annual, Nov 2019.

[11] US Bureau Of Economic Analysis. Gdp by state, 2019.

[12] Avis Favaro, Elizabeth St. Philip, and Graham Slaughter. Canadian health authority warns travellers over mysterious illness sickening dozens in china, Jan 2020.

[13] Nick Givas. Who haunted by january tweet saying china found no human transmission of coronavirus, Mar 2020.

[14] Berkeley Lovelace Jr. Cdc confirms first human-to-human transmission of coronavirus in us, Jan 2020.

[15] Katherine DeClerq. Ontario reports lowest number of new covid-19 cases since end of march, May 2020.

[16] Lynn Johnson. There are more viruses than stars in the universe. why do only some infect us?, Apr 2020.