Covid Business is our Business: Predicting Covid Footprint on NYC Businesses

Bing Chen, Jenny Shaojun Jiang, Yiwen Shen

December 5, 2020

I. Abstract

Covid Business is Our Business is a research project aimed at developing a case study on the effects of Covid-19 on New York City businesses. This application analyzed Covid-19 cases in New York City and correlated it to both existing operating businesses and applications for new businesses. We first found that the effects of Covid-19 were not limited to any one location defined by ZIP code. Instead, Covid-19 was found to have a general effect on the number of legally operating businesses and business license applications. Since business growth is also indicative of the current state of the economy, this information may be able to provide one perspective of how the local economy has been affected by Covid-19. Such perspective is: when Covid-19 first hit our country, our government's slow response led to many business closures, indicated by the decrease in the number of legally operating businesses. This led to a huge unemployment wave, suggesting that with better preventive measures, it could have been mostly, if not completely, avoided.

II. Introduction

We live in unusual times at the moment. With Covid-19 comes many questions of its impact with many researchers developing new projects to analyze Covid-19 data. This is one of such projects. Like many other similar studies, we studied the trend of legally operating businesses and business license applications both before and during Covid-19, enabling us to see if the presence of Covid-19 is reflected in the number of legally operating businesses and license applications. If we were able to show that Covid-19 had a significant impact on New York City businesses, we can argue that these effects we see could have been avoided if the government was quicker to react to the virus. In addition, this could be a lesson for the future. We predicted that ZIP codes where there was a higher percentage of people who tested positive would show a bigger change in the number of New York City businesses. But from our research, we found that the impact of Covid-19 on businesses is not confined to any particular area. Knowing what we know now, we can use this case study to better prepare for any national catastrophes and protect our local businesses by providing some sort of insurance to prevent future business closures.

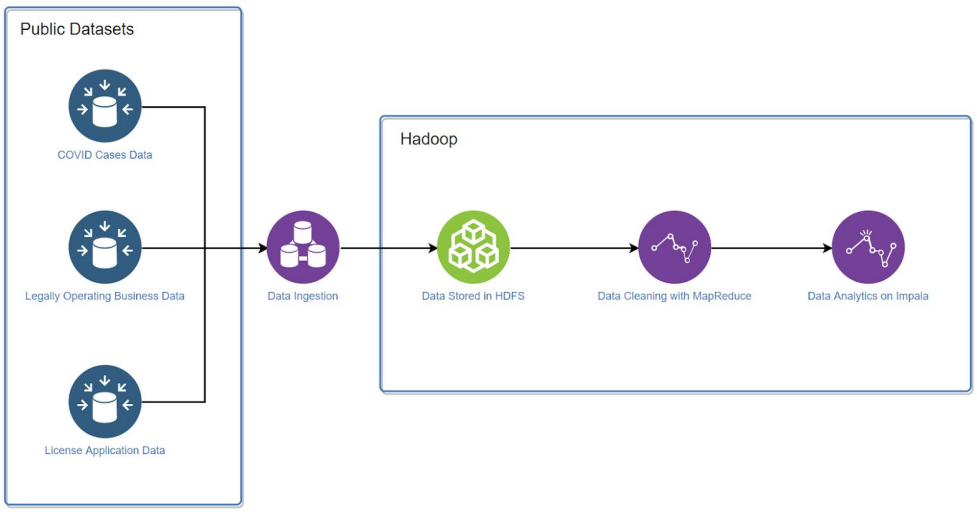

Figure 1: Design Diagram

Figure 1 shows our design diagram, where we put each public dataset through data ingestion and moved them into HDFS located in the New York University Dumbo cluster, where we used various MapReduce programs to clean and profile our data. Cleaning included dropping unnecessary columns and rows not within the date ranges we were interested in. Then we moved the data into Impala to be converted into a more structured table so we can run our analytic. Our analytics included counting businesses and applications over ZIP codes, taking average rates per ZIP code, and combining current year data with previous years.

III. Motivation

Covid-19 has created unprecedented impact in all aspects of our lives, from the health of our citizens to the environment to the economy of the world. During this special time, a surge in unemployment and closure of businesses is extensively reported. Many researchers have studied the effects of Covid-19 across nations, states, and cities. As New York City is a diverse melting pot and home to our college lives, we limited our research to a narrower scope by examining the impact of Covid-19 on businesses in NYC, particularly small businesses. We hoped to find some clear impact of Covid-19 on NYC businesses to learn from how the handling of this pandemic can be examined to prevent or lessen the impact if there was ever another public health crisis in the future. Specifically, we would want to not only confirm that Covid-19 indeed has a negative impact on the overall commercial environment in NYC, but also examine whether there is a correlation between the severity of Covid-19 and the number of businesses in different geographical areas. Knowing the impact of Covid-19 on businesses, especially when it is distinct between regions, business owners will be more informed when making business decisions, and the government will have more insights to better prepare for and provide aid so less businesses will have to close down in future pandemics.

IV. Related Work

In [1], Chernick et. al took a statistical approach to analyze and predict the fiscal effects of Covid-19 across 150 large cities in the U.S. The study concluded that there is an average decrease of 10.25% in earnings with little relative variation across cities, which indicated that the negative economic impact of Covid-19 is evident and universal, and that many of the cities have similar industry composition. Chernick et. al also mentioned that the future economic trend under the effects of Covid-19 depends on future federal, state and local public policies, aid, and interventions.

In [2], Sharifi et. al reviewed and analyzed the impact of Covid-19 on the urban planning of major cities in the world. The work agrees with Chernick et al. that the pandemic has had a significant negative impact on the economy. It suggests that long-term visioning, pre-event planning, and adequate investments from authorities are key factors determining the resilience to such a disruptive event. There is an inconclusive relationship between population density, connectivity, and city size and the spread of the virus, thus it is hard to conclude the severity of the pandemic and its impact on different regions merely from those factors. What is noteworthy, however, is that marginalized social groups are disproportionately affected by the economic impacts of the pandemic and that the homogeneous economic structure increases vulnerability. The pandemic not only intensifies the existing social inequalities but also induces new ones on people who are living in epicenters or have family members infected with

Covid-19. It is imperative to realize the issue and take measures against them. It prompted us to organize our findings into distinct regions defined by ZIP code since demographics vary from location to location.

From these sources, we found evidence on Covid-19's significant impact on the economy and the urban planning and design of different regions, which inspired us to further explore the idea of regional impacts on a more concentrated scope with the hope that our findings will facilitate better government and business decisions.

In [3], a bibliometric study published about Covid-19 research trends in the field of business, "decision making", "risk assessment", and "big data analytics" appeared as keywords. Since the study sampled 107 articles for analysis, this indicates that there is a substantial amount of research concerning such topics. This informs us that our analytic is relevant and timely to the current research being conducted on Covid-19 and business.

In [4], a case study of business closers and launchings in Turkey in 2018 (unrelated to the Covid-19 crisis) informed us that we must factor in seasonal patterns when performing our analysis. This source showed that the number of closers and launchings changed significantly over the course of a year.

In [5], Fernandes compared the prominent Covid-19 to past outbreaks. Significantly, the biggest difference between Covid-19 and viruses with no vaccines like the "SARs, HIV/AIDs, and pandemic influenza," is that there is no economic impact of mortality rates . In the past, the traditional approach to researching outbreaks is to "use information on deaths and illness to estimate the loss of future income due to death and disability." We used the information we learned from this study to decide on the specific Covid-19 dataset we needed. We decided that we would have a much higher chance of finding a correlation between Covid-19 and its business impact if we looked at the percentage of people who tested positive versus those who died from the virus, especially because a low percentage of those who contracted the virus actually die compared to previous viruses.

V. Datasets

The first dataset is the "Percent Positive and Test Rate of Molecular Testing by ZIP Code" dataset from NYCHealth [6], accessed on November 13, 2020. We renamed this to Covid Positive Rates (CPR) for easier handling. This dataset contains 93 columns and 184 rows. Except for the first column which is the label column, each column represents an "End Date", ranging from Aug 9 to November 8, 2020. Each row represents a ZIP, a NYC borough, or the citywide rates. Inside the table are the corresponding Covid Positive Rates - the percentage of tested people who tested positive in the 7 days preceding the indicated "End Date". From those, we removed the boroughs and the citywide rates so only the data for the 177 different ZIP codes and all the dates were left. The schema for the End Date is string, for ZIP is integer, and for Positive Rates is float.

The second dataset is the Legally Operating Businesses (LOB) dataset from NYCOpenData [7], accessed on October 30, 2020. This dataset contained 266,571 records and 27 columns. From those, we cleaned it so there are only the columns for ZIP, Industry, License Creation Date, and License Expiration Date, among the numerous cleaned versions of this dataset. The schema for each column was, respectively, integer, string, string, string, and string. We also modified the dataset so the only records remaining are those with a License Creation Date between August 9 and November 8 of years from 2015 to 2020.

The third dataset is the License Applications (LA) dataset, also from NYCOpenData [8] and accessed on October 30, 2020. This dataset included 375,366 records and 25 columns. We cleaned the

dataset so only the columns for ZIP, Application or Renewal, Status, Start Date, End Date, and License Category remained, using schemas integer for the ZIP and string for the remaining columns. The start date and end date indicated when the application was first made and when action was taken. This dataset was also modified to only contain records between August 9 and November 8 of years from 2015 to 2020.

All of the datasets are in CSV format. They are updated regularly on the websites.

VI. Analytics Stages

A. Data Ingestion

After obtaining our datasets from their respective sources, we moved them into Dumbo using Fugu and then into the '/user/bc2486/project/Inputs' HDFS folder. The permissions were set using the general format 'hdfs dfs -setfacl -R -m default:user:<netid>:rwx /user/bc2486/project' and 'hdfs dfs -setfacl -R -m user:<netid>:rwx /user/bc2486/project' for everyone on our team as well as the graders to enable read, write, and execute privileges. Additionally, we also set the mask to rwx so that the permissions would not be overridden by the default mask. As we created new folders under 'user/bc2486/project', we added permissions to those as well. The permissions can be checked with 'hdfs dfs -getfacl /user/bc2486/project/<path>'.

B. Data Cleaning and Profiling

For our cleaning and profiling, we began by dropping unnecessary columns from the LOB and LA datasets, so that only columns pertaining to the ZIP code, End Date, and Industry were left. For our CPR dataset, we modified it so that instead of each column representing a separate date, the date became a separate column of its own, which resulted in 16284 rows (177 ZIPs * 92 dates). The final form of CPR was then ZIP, Positive Rate, and End Date, which aligned nicely with our business-related datasets. Finally, we ran 2 separate cases of row-dropping for LOB and LA that resulted in:

(1) Rows in the range 8/9/20-11/8/20, which is the range of the CPR dataset
(2) Rows in the range 8/9-11/8 from 2015 to 2019 (i.e. 8/9/15-11/8/15, 8/9/16-11/8/16, …), adding the year as separate column for easier processing for the analytic

We counted the number of records for each of these instances, which resulted in the values seen in **Table 1**. Keeping the specific time range for the years other than 2020 was important because we did not want influence from changes caused by seasonal patterns.

Dataset	Original Rows	Rows in range 8/9/20-11/8/20 **(1)**	Rows in range 8/9-11/8 from 2015-2019 **(2)**
Legally Operating Businesses	266571	651	13247
License Applications	375366	4321	54977

Table 1: Comparison of records in range

In an effort to obtain more varied results, we also created a version of **(1)** for both LA and LOB where neighborhood was indicated instead of ZIP. To change the ZIPs into neighborhoods, we used NY Department of Health's ZIP Code definitions of NYC Neighborhoods [9].

C. Gathering Analytics

We began our analytic by averaging the Covid-19 positive rate by ZIP as a measure of the average rate in the area from 8/9/20-11/8/20. Then, we counted legally operating businesses and license applications per ZIP. We made sure to check the license expiration dates for the legally operating businesses so that the ones that had expired during the time frame were removed. After creating separate tables in Impala, they were joined by corresponding ZIP. Since we used inner join, any ZIPs that did not belong to both the CPR and the business datasets were dropped. We exported our final tables and plotted them using Excel, which resulted in **Figure 2** and **Figure 3**.

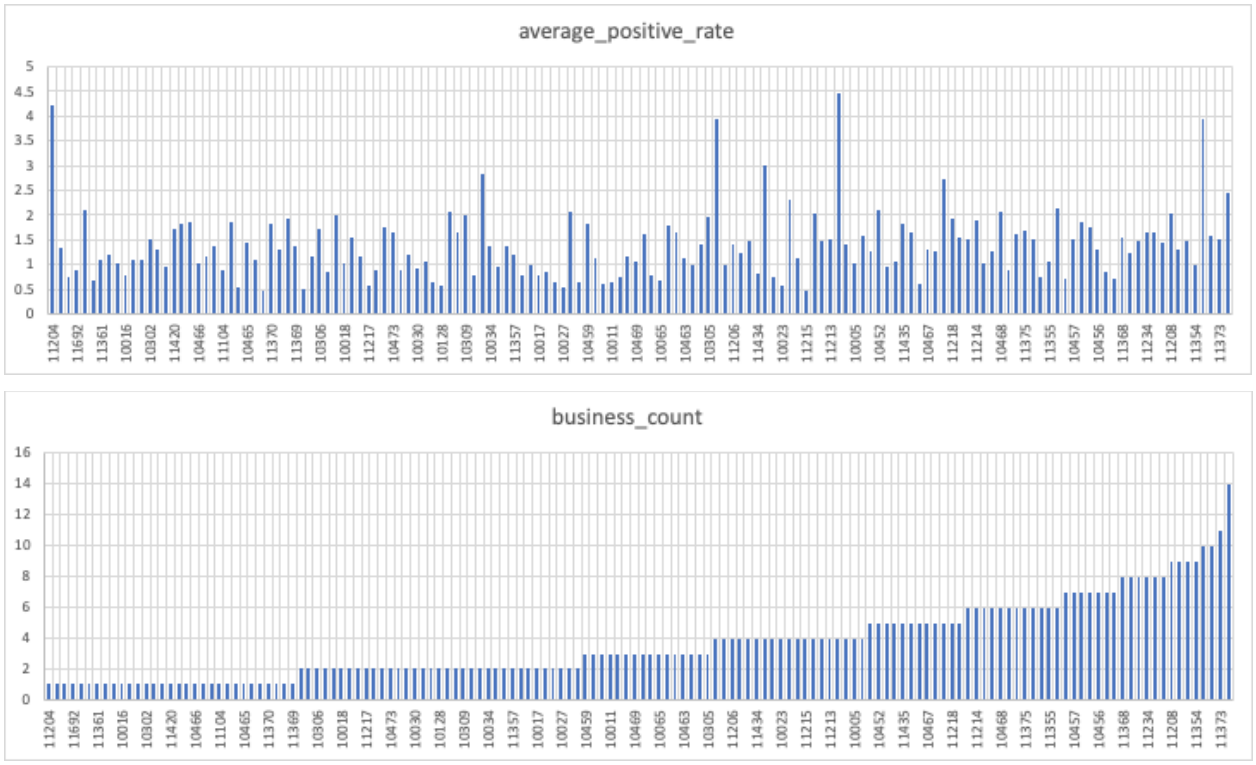

Figure 2: Average Covid-19 Positive Rate per ZIP compared to Number of Legally Operating Businesses per ZIP

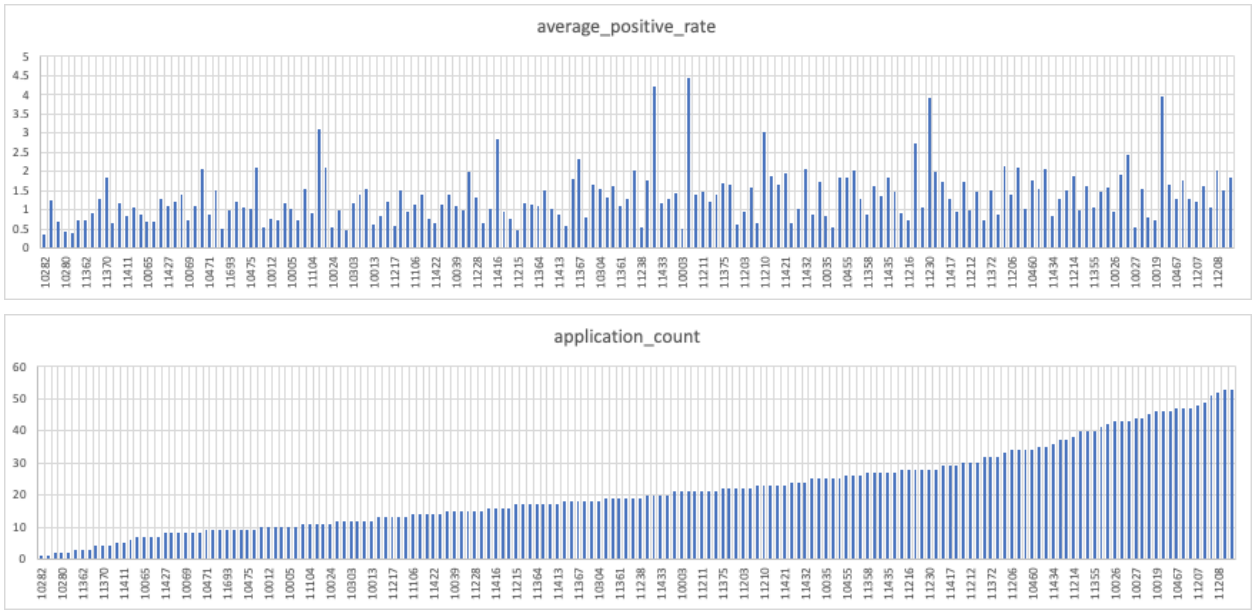

Figure 3: Average Covid-19 Positive Rate per ZIP compared to Number of License Applications per ZIP

The graphs are ordered by increasing business count and application count respectively, and the ZIP codes align with the ZIPs for the average Covid-19 positive rate. From looking at the graphs, one can see that there does not appear to be an obvious correlation.

In an attempt to see if we can glean more information, we also attempted to do the aforementioned actions, but with the ZIPs grouped into neighborhoods. The results appear relatively similar to **Figures 2** and **3**, and can be seen in **Figures 9** and **10** of the Appendix.

To obtain a clearer picture of whether or not correlation existed between the Covid-19 positive rate and businesses by location, we used Python to plot a linear regression to fit the values, as seen in **Figures 5** and **6**, which plot operating business count and license application count against Covid-19 positive rates on the neighborhood and ZIP, respectively. We calculated the R^2 coefficient to measure the fit, and its value is defined by **Figure 4**, where y_{true} represents the actual value, y_{pred} is the model's prediction, and $\bar{y}$ is the mean value for the dependent variable (y-axis). The value measures the percentage of dependent variable (in this case the count) variation that has been accounted for. In other words, a higher value would mean more predictability, since more variance has been accounted for. The R^2 values were 0.125 and 0.072 respectively for **Figure 5**, and 0.042 and 0.083 respectively for **Figure 6**. Since the best possible R^2 coefficient is 1 and a score of 0.125 is still only 12.5% of variation accounted for, we can conclude that the Covid-19 rate is not a significant indicator of where businesses tend to grow or decline in NYC.

$$R^2 = 1 - u/v$$

$$u = \sum (y_{true} - y_{pred})^2$$

$$v = \sum (y_{true} - \bar{y})^2$$

Figure 4: Score function

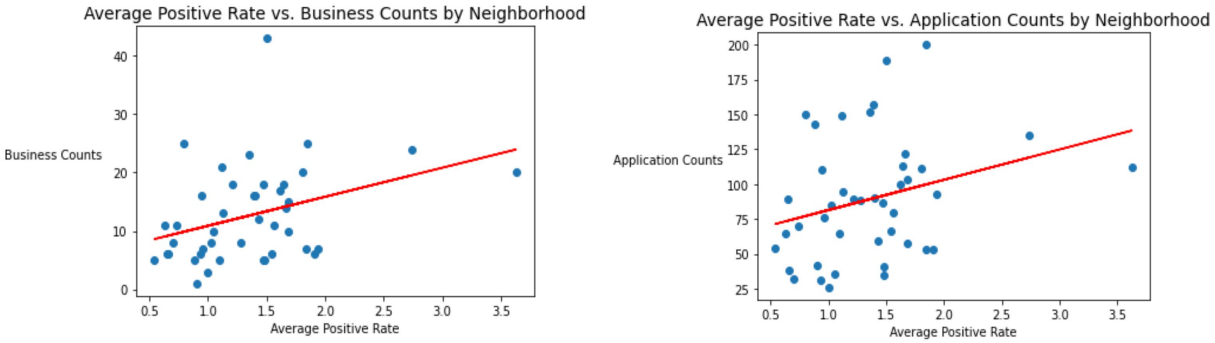

Figure 5: Linear regression by Neighborhood

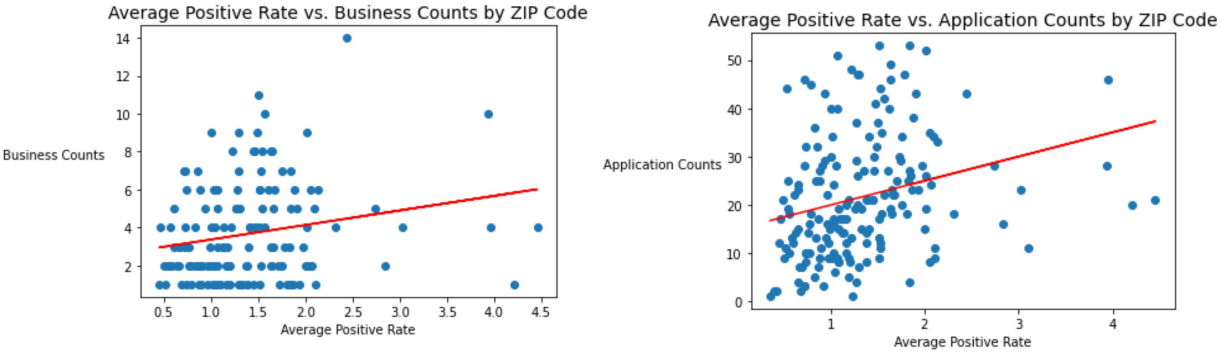

Figure 6: Linear regression by ZIP

Finally, after obtaining unsatisfactory results when we attempted to see Covid-19 impact by location, we decided to perform an analysis by time. To do so, we compared current year business data with data from the past 5 years. We had already performed a count on 2020, so we next performed counts on 2015, 2016, 2017, 2018, and 2019 separately for both legally operating businesses and license applications. The counts (all 6) were then joined into one table and plotted, again with Excel, seen in **Figures 7** and **8**. The ZIP code axis is organized in an ascending order. Each line indicates a different year, and the red line in both graphs represents 2020, i.e. during the Covid-19 crisis. Since it can be clearly seen at the bottom, this means that both the number of legally operating businesses and license applications declined significantly in 2020 compared to 2015-2019. The decline in license applications is less obvious, though this may be due to many license applications coming from sources that regularly renew their licenses and/or are less impacted by Covid-19, such as tow truck companies.

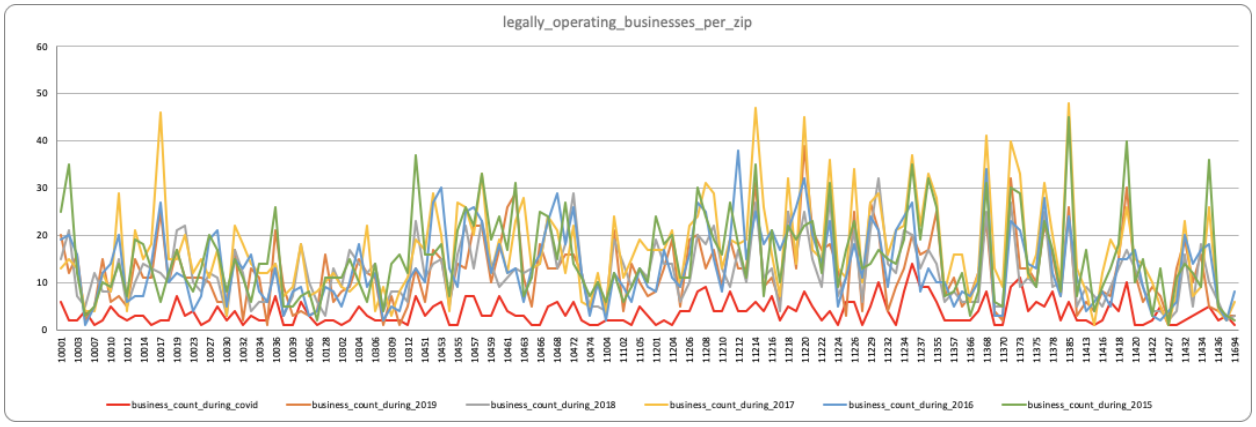

Figure 7: Number of Legally Operating Businesses per ZIP from 2015-2020

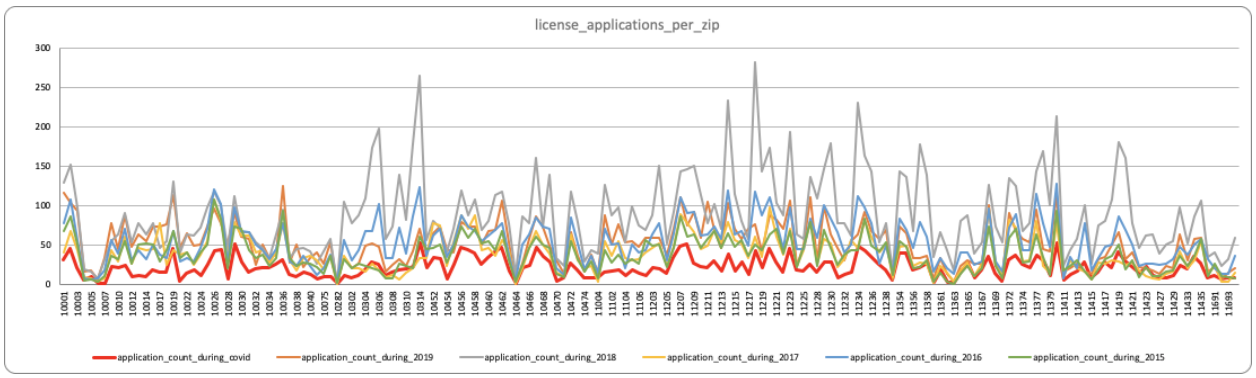

Figure 8: Number of License Applications per ZIP from 2015-2020

Although many trends remained similar over the course of the years (i.e. what the peaks and troughs of each line roughly correspond to), there were a few specific locations that did not follow the trend in 2020, for example, ZIP codes 10017, 10455, 10466, and 11204 in the LOB set. While additional information would be needed to make clearer claims about these locations, these types of patterns could be an indication of locations harder hit by the Covid-19 crisis.

VII. Conclusion

The business environment is fast-changing and highly sensitive to public crises such as the Covid-19 pandemic. It is pivotal for both business owners and authorities to gain comprehensive understanding about the impact of such events so that they can plan ahead and minimize the shock and the loss. Our research quantitatively confirms that on the scope of New York City, Covid-19 caused a significant decrease in the number of businesses and licence applications. We discovered that Covid-19 positive rates is not the most dominant factor in affecting the absolute number of businesses and applications for each area, as we found low correlation between the nominal values of those data. Indeed, some ZIP code areas can be more prosperous than others due to demographics, policies, and historical reasons. We observed, however, some ZIP code areas tend to be more seriously affected by the pandemic than others, such as the ones mentioned in Section VI. With our research findings, business owners and the government can gain insights into the scale of Covid-19's impact on businesses in NYC and the areas

that tend to undergo the strongest influence. Further investigation into why the business environment in those areas tend to be more susceptible to Covid-19 will be beneficial for us to see a more complete picture of the pandemic's impacts. We hope the findings can contribute to the betterment of individual business plannings and city-wide aid to help alleviate the disturbance created by Covid-19 and prepare for future crises.

VIII. Acknowledgements

We would like to thank the NYU High Performance Computing team for providing us with the Dumbo Cluster and Impala table sharing. We would also like to thank NYCHealth and NYCOpenData for providing us with our datasets. Finally, we would like to thank Professor Ann Malavet for her guidance in the project.

IX. Appendix

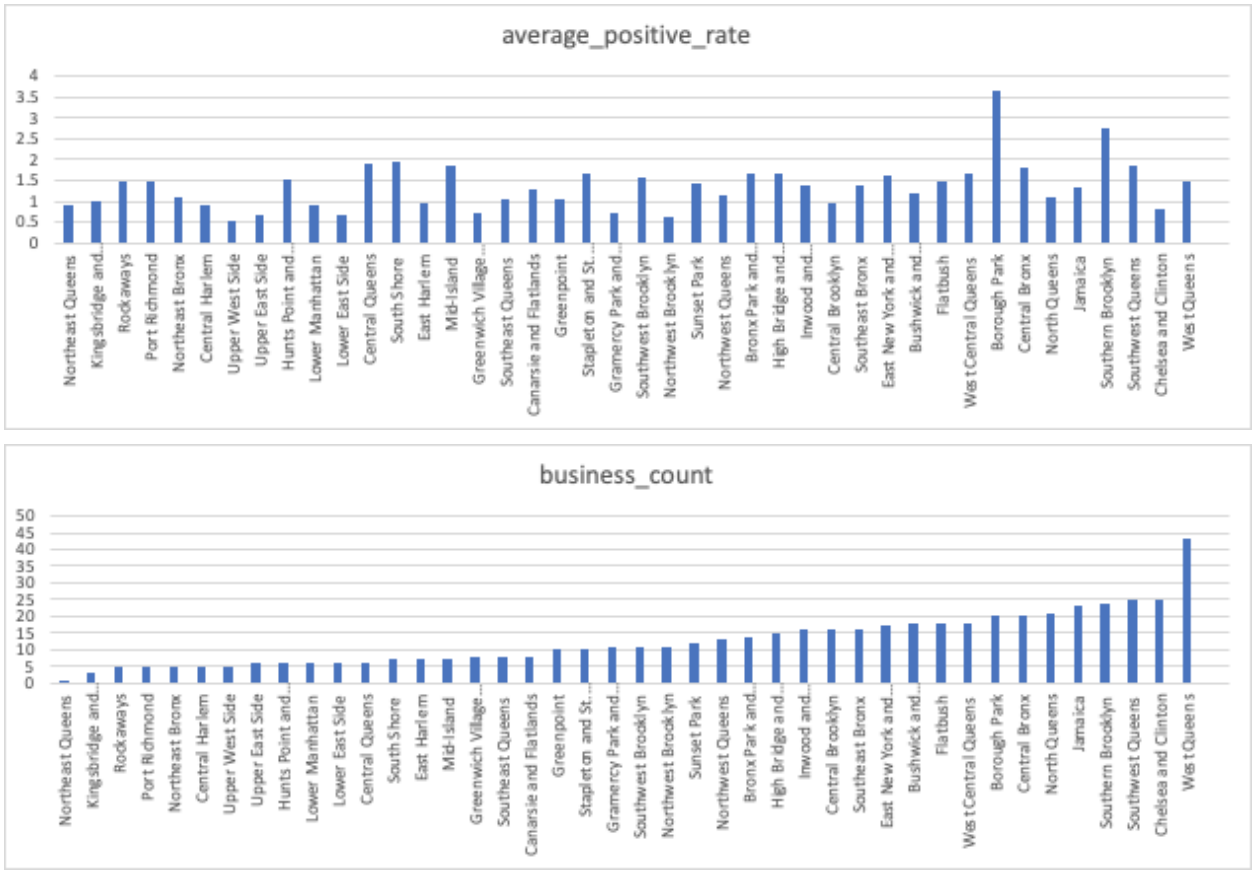

Figure 9: Average Covid-19 Positive Rate vs Legally Operating Businesses by Neighborhood

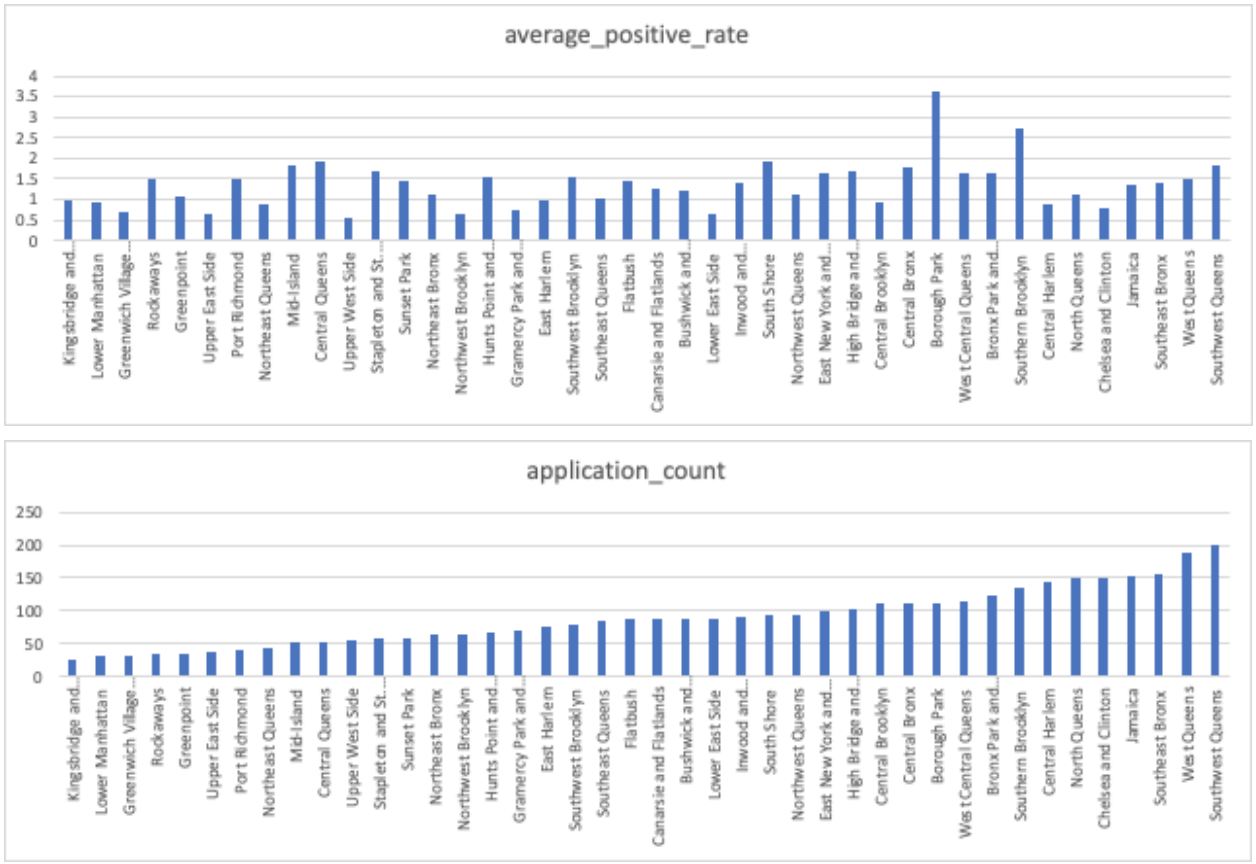

Figure 10: Average Covid-19 Positive Rate vs License Applications by Neighborhood

X. References

[1] H. Chernick, D. Copeland, and A. Reschovsky, "The Fiscal Effects of the COVID-19 Pandemic on Cities: an Initial Assessment", 2020.

[2] A. Sharifi, A. R. Khavarian-Garmsir, "The COVID-19 pandemic: Impacts on cities and major lessons for urban planning, design, and management", 2020.

[3] S. Verma and A. Gustafsson, "Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach", Journal of Business Research, 2020.

[4] Z. D. U. Durmuşoğlu and A. Durmuşoğlu, "An Analysis on the Business Closers and Newly Launched Businesses in Turkey", ICBIM '18: Proceedings of the 2nd International Conference on Business and Information Management, 2018.

[5] N. Fernandes, "Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy", 2020.

[6] NYC Health, "Percent Positive and Test Rate of Molecular Testing by ZIP Code", 2020.

[7] NYCOpenData, "Legally Operating Businesses", 2020.

[8] NYCOpenData, "License Applications", 2020.

[9] New York State Department of Health, "ZIP Code Definitions of New York City Neighborhoods", 2006.