Study of Small Medium Business and Crime During COVID-19

(Final Report)

Team160: Antonio De Los Heros, Muhammad Ibraheem, Muhammad Kaleem Khan
Chandra Narasimhan, Raghvendra Trivedi, John Walker
lbeunza3, mibraheem6, mkhan374, cnarasimhan3, rtrivedi33, jwalker352
*@gatech.edu

1. Introduction – Motivation

Global pandemics are rare events in recent times. There is limited research done to study the interdependencies of economic shift and crime during global pandemics. The study will show deep analysis on the impact of COVID-19 to small-medium businesses (SMB) and crime, as well as correlations between those factors in the following cities: Baltimore, Boston, Denver, and Phoenix. Utilizing publicly available COVID-19, crime, and SMB data for a given time period, this study will not only uncover many insights, but it will also help users understand those insights using interactive data visualizations. This study will help better plan for future pandemics and also help many small-medium business owners in future challenges. Residents of these areas may also find the study informative and allow them to see benefits by the uncovered truths and realities of the COVID-19 situation for a given period of time.

2. Problem Definition

While there have been many articles and studies related to the effects of COVID-19 on either crime or small-medium businesses independently, none of them provide a combined view of the impact of COVID-19 on these factors together. The problem with these independent studies is that they may influence users' decisions negatively by failing to provide a comprehensive view of these factors. Users like government agencies, SMB owners, and residents of these cities may find it difficult to establish relationships (positive, negative, or neutral) among COVID-19, SMB performance and crime and are therefore unable to make appropriate decisions. Furthermore, complicated data visualizations in those independent studies make it difficult to implement their findings in decision making.

3. Survey

3.1. Crime Impact

As outlined by Bowman [6], there has been a reduction in crime measured by calls for service during the COVID-19 pandemic. Related analysis by Mohler et al. [1] confirms these effects using regression analysis, showing that social distancing policies have had a statistically significant impact on certain types of crime in Los Angeles and Indianapolis. Rosenfeld and Lopez [5] performed time-series analysis with a structural breakdown of 11 different types of crime in 27 US cities, showing static scatter plot visualizations of each offense. Feng et al. [7] show the effectiveness of several types of visualization applied to crime data in three US cities, and further explored using the Prophet model and LSTM models to predict time-series crime data. Yoo's work on crime [4] highlighted the benefits of the ETL process and dimensional modeling in analyzing crime data.

3.2. Small/Medium Business Impact

According to Humphries et al. [17], disruptions related to the COVID-19 pandemic had a severe impact on small-medium sized businesses, forcing many to lay off employees. Morawska et al. [16] studied methods of limiting indoor airborne transmission of COVID-19 such as requiring facemasks, enforcing social distancing, implementing enhanced air filtration, and reducing air recirculation that would eventually help many businesses reopen. Analysis of survey data conducted by Bartik et al. in [8] will help to complement our studies for COVID-19's impact on small businesses, as the article outlines business closure patterns, timelines, and strategies of business owners during COVID-19. We intend to use unofficial sources of data to measure the impact of COVID-19 on small-medium businesses. In a working paper by the National Bureau of Economic Research, Glaeser et al. [14] demonstrate evidence that data from unconventional

sources can help predict the figures from official government statistics, which often have a multi-year lag. A study by Kurmann et al. [9] also highlights the robust nature of unofficial data sources in measuring the economic impact on small business employment during COVID-19 by using data from a scheduling software called "Homebase". Studies conducted by Han et al. [2] and Tayeen et al. [3] concluded that location plays a vital role in a small-medium businesses' success. These two studies are useful because we intend for our visualization to help answer how COVID-19 has impacted location-based success of small-medium businesses.

3.3. Visual/Analytic Methods

Data cleaning integration techniques used by Chetty et al. [10] and implemented in [15] will be useful in our endeavor to clean and integrate data from multiple sources into our visualization. The paper published by Kim et al. [13] describes the visual analytics process as an interactive mechanism with various stages such as data preparation, visualization, and user interaction. They also discuss how geo-data is used in mapping to make the data more presentable and meaningful, which we will employ in our approach. Aigner et al. [19] investigate using parameter, clustering, PCA, and event-based methods for interactive data visualization and better pattern recognition in time-oriented data. Time-oriented data analysis is further categorized by Biswas et al. [18] and they recommend phase-wise visualizations for data with exponential trends (such as COVID-19 case data), with each phase created via a knee detection algorithm. We may also find the need to model certain events in our visualization. Work by Chakraborty et al. [11] describes how real-world events impact socio-economic indicators using statistical models (ARIMA). The paper published by Rozsnyai et al. [12] describes methods of capturing, storing, and creating a database structure for time-critical events. We will be using similar methods for gaining access to data, data preparation, and mapping it to various event categories.

4. Proposed Method

4.1. Approach

Crime, COVID-19, and SMB are generally analyzed or observed individually due to the amount of information that can be derived from each subject. The benefit of this project is the ability to not only gain insights from each individual subject, however also looking at the interactions and correlations that exist amongst them; this is believed to be a true differentiator comparing to other in-depth analysis that have been done on crime, COVID-19, and small-medium businesses. Our proposed method ranges from data source collection, data clean up and integration, to performing a comprehensive analysis that provides us the insights and correlations between the data sources; we also provide an exploratory data analysis that will look to guide the broader dashboard concepts containing the most useful views.

4.2. Data Sources

The Project consists of three data sources that have been gathered from different sites/APIs, crime, COVID-19, and small-medium businesses:

4.2.1. Crime Data Source

Our crime data consists of daily crime counts across 16 categories (aggravated assault, arson, etc.) for the four cities. We gathered raw (CSV format) data published by each city's website [23-26] and used python scripts to load the CSV files and perform data cleaning. One challenge in cleaning the data was mapping latitude and longitude values to zip codes - which was necessary for integration with the other data sources in this project. We ended up using a python zip code database [22] to search for nearby zip codes and find the closest based on the lat and long values. Additionally, since each city reported crimes differently, we had to create a mapping file to map each city's crime descriptions to one of the sixteen categories we had chosen to report for this analysis, with higher level grouping done to put each of the sixteen categories into "Violent" or "Non-Violent" categories. After this cleaning, these data files all together were 57.6 MB. The size of the data in raw MB was reduced significantly (down to 2.5 MB) after performing several aggregations and pivoting the data, which resulted in a CSV file with 45,693 rows and 21 columns.

4.2.2. COVID-19 Data Source

Our COVID-19 data is sourced from The New York Times [20]. The New York Times is compiling COVID-19 cases and death data from state and local governments and health departments to provide a complete record of COVID-19 virus outbreak on a daily basis. The New York Times data is available in CSV format and contains Date, COVID-19 cases and deaths at county and state level only. Since our analysis requires COVID-19 data at zip code level, the New York Times county data is extrapolated to zip code level proportionately using census zip code population estimates

from a zip code database [21]. Data integration was performed by standardizing field lengths, data types, and formats. Zip level COVID-19 data has approximately 6.5 million rows and the dataset size is 280 MB. City level COVID-19 data has 25,000 rows and dataset size is 1 MB.

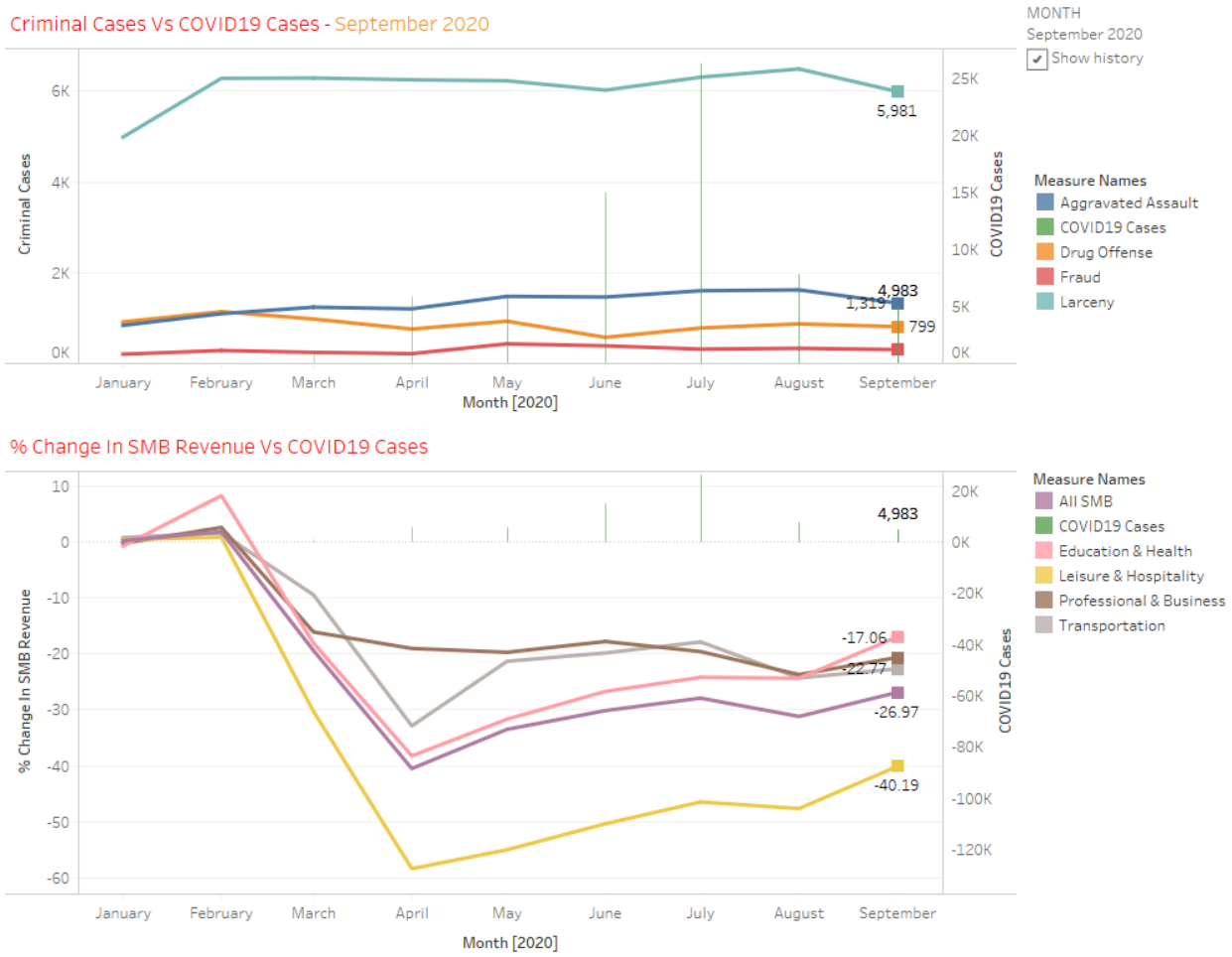
4.2.3. Small-Medium Businesses Data Source

Our initial target was to source data using an API from Yelp. We collected data from Yelp via API for Boston, but it didn't have historical information attached to it and data only provided restaurant status as of today. Due to these limitations, we ended up using another data source [27] compiled by Harvard University. It consists of SMB data sourced from Womply for small-medium business transactions and revenue data aggregated from several credit card processors. Transactions and revenue are reported based on the county in which the business is located. Data is indexed in 2019 and 2020 as the change relative to the January index period in education and health services, leisure and hospitality, and the retail and transportation sectors. The raw dataset was 122 MB with 55 files, and after processing it was reduced to 23,289 rows and 20 columns.

4.3. Data Analysis Visualization

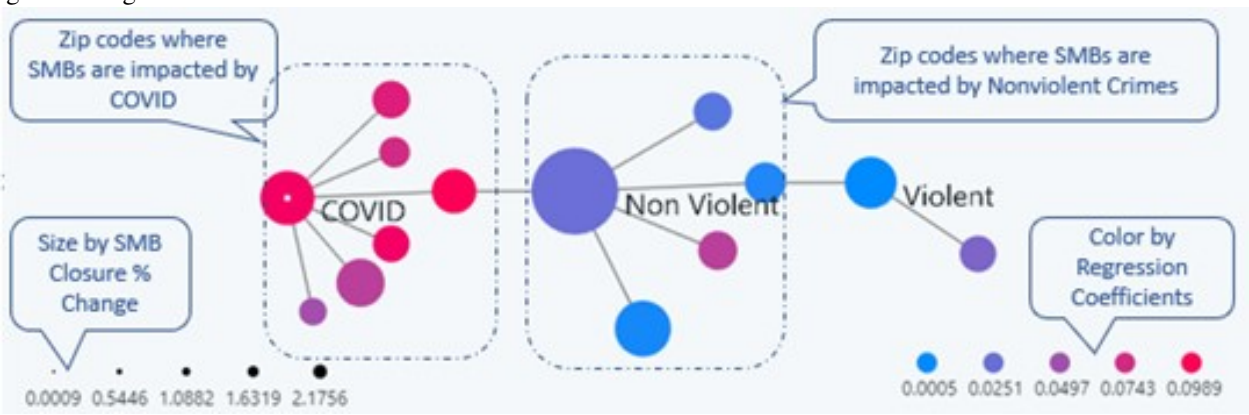
4.3.1. Time-Series Visualizations

To study and analyze trends between COVID-19, crime, and SMB performance we created a Tableau based time-series visualization which displays measure of each factor for all four cities combined. The top chart shows various types of crime against the trend of COVID-19 cases, while the bottom chart shows COVID-19 cases plotted against measures of SMB revenue for various types of industries. Both charts allow the user to watch these trends unfold over time by pressing a play button on the top right.



4.3.2. Linear Regression for Graph Analysis

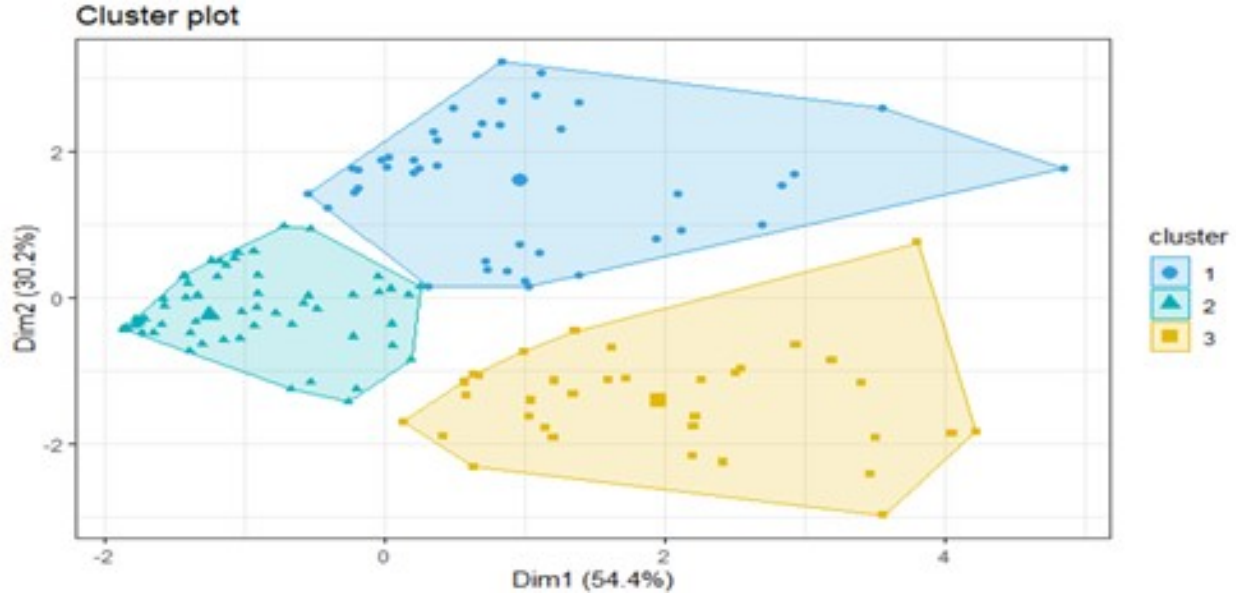
To show relationships between COVID-19, SMB performance and crime at the zip code level, we created a node-edge diagram in Argo Lite:



To test whether each factor had an impact on SMB performance, we ran a linear regression model by taking the SMB performance measure as the response and measures of COVID-19 and crime as the attributes. If the regression coefficients are significant ($p < 0.1$) we keep the result and create an edge between nodes in the node-edge diagram. Each node represents a zip code, with the size of the node representing the avg percentage change in SMB closure. The color of each node indicates the regression coefficient value in the above graph.

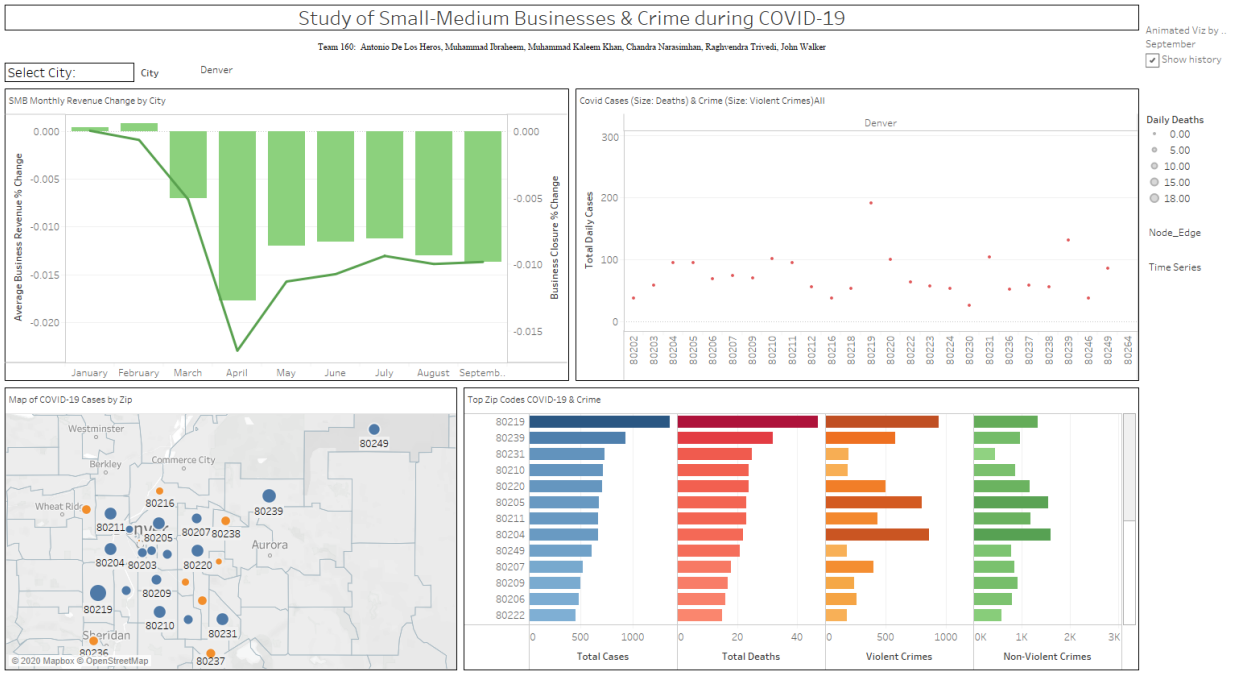
4.3.3. Cluster Analysis for Choropleth Map

In order to identify similarly impacted zip codes across several factors in Boston, Denver, Baltimore and Phoenix, a k-means unsupervised clustering algorithm was applied on the combined dataset of four cities under consideration. We used the "kmeans()" function in R with COVID-19 cases and deaths, violent and non-violent crime counts and percentage change in SMB revenue as inputs. Three clusters (k=3 chosen by elbow chart method) were used in the model, and the "nstart" parameter was set to 25 for stable results - ensuring random starting assignments do not impact clustering. Each zip code was associated with a cluster number (1-3) and geographically mapped in Tableau using a Choropleth map feature – each cluster being visually identified with a unique color. This resulting chart is built into the Tableau dashboard shown in the section below (bottom-left). Below, a scatter plot of the clusters is shown:



4.3.4. Tableau Dashboard

A Tableau dashboard was created to link the visualizations mentioned above, incorporate additional visualizations that help in understanding these factors, and provide the user with one central location to access all of the information. The dashboard is based on the combined data for all four cities via a direct connection to the CSV file. The main tab of the dashboard consists of four visualizations which can be manipulated via a "Select City:" filter at the top. In the top right corner, there is a bubble chart that displays the total number of COVID-19 cases by zip code by month, which can be animated to show how cases evolved throughout the pandemic. In the top left corner, we are able to show the relationship between different measures of SMB performance (business closures and change in revenue) over time via a trend chart. In the bottom right corner, we show the top ranked zip codes by COVID-19 cases, along with measures of violent and non-violent crime to help the user understand if there are similar trends with COVID-19 cases and crime. Finally, the bottom left corner shows the map referenced in the section directly above, not only helping the user understand where the zip codes are geographically, but also showing which zip codes share the same cluster via the color of the bubble. On the far right of the dashboard, we include a link to the node-edge diagram in Argo Lite "Node_Edge", and a separate link to the time series visualizations "Time Series".



5. Experiments/Evaluation

As part of the experimentation and evaluation we sought to answer five questions with the results of our analysis and visualizations. Those questions are listed below, with the results found directly underneath each question.

1. What are the correlations due to the interactions of COVID-19, crime, and SMB? To answer this question, we turned to our linear regression analysis of significant variables and resulting node-edge diagram. Based on this analysis, we found that out of 161 zip codes:

   (a) 59 zip codes show no relationship between COVID-19, SMB or crime

   (b) 22 zip codes have less than 0.1% change in SMB and regression results are insignificant

   (c) 44 zip codes show a strong relationship between SMB % change and COVID-19

   (d) 37 zip codes show a strong relationship between SMB % change and crime

   (e) 26 zip codes show a strong relationship between COVID-19 and crime

2. What type of impact does COVID-19 have on crime, specifically violent crimes vs non-violent crimes?

   As per available data, we have found four types of crime which have shown movement during COVID-19 period from January 2020 to September 2020. From June to August, we see increasing trends in larceny, drug offenses, and a slight increase in

aggravated assault. As the COVID-19 cases peaked, so did the counts of three crimes (larceny, drug offense, and aggravated assault), however we noticed a decreasing trend in fraud in the same time period. During COVID-19 there were many unique events like protests related to the George Floyd case, etc. so it is difficult with available data to establish direct relationships between the correlations of crime with COVID-19 cases.

3. What has been the overall impact on SMB, whether it is due to COVID-19 or increase/decrease of crime for the mentioned cities?

   The chart titled "SMB Monthly Revenue Change by City" in the Tableau dashboard shows a strong relationship between SMB economic performance and COVID-19. SMB activity bottomed out in April, which coincided with COVID-19 related lockdown orders. As lockdowns eased, SMB economic activities resumed at various rates. The chart shows that SMB activities for all four cities are below their pre-COVID-19 levels. Investigating city-level SMB impacts further, Denver's SMB activities showed the deepest contraction when compared to other cities. Among the four cities we studied, Phoenix's SMB activity contracted at a slower pace than other cities and recovered more quickly than other cities. In our time-series charts, we can also see when COVID-19 cases rose in July, SMB revenue started declining again. However, after August revenue started to increase across all SMB revenue services.

4. Are there patterns or insights that emerge from the k-means clustering analysis?

   Through our cluster analysis and visualizations, we were able to identify the following patterns and insights:

   (a) Boston and Phoenix core areas appear to have similarities mapped to COVID-19, crime and change in SMB dimensions.

   (b) Denver and Baltimore core areas appear to have similarities mapped to COVID-19, crime and change in SMB dimensions.

   (c) Regardless of total daily COVID-19 cases and deaths, crime and change in SMB business revenue dominated similarity patterns.

5. Does the number of violent crimes correlate with the number of COVID-19 cases in a given ZIP code?

   Analyzing the "Top Zip Codes COVID-19 and Crime" chart in the Tableau dashboard, we found an interesting observation in terms of COVID-19 and crime. Zip codes with the highest number of cases typically have the most amount of deaths, however, when we compare COVID-19 and incidents of violent and non-violent crime, we can identify that the zip codes containing the most COVID-19 cases are not the zip codes with the highest crimes.

6. Conclusions and Discussion

In this project, we have combined large datasets from multiple sources, leveraged visualization tools such as Tableau and Argo Lite, and applied machine learning algorithms such as linear regression and k-means clustering to help us derive insights from how COVID-19 has impacted crime and SMB performance. We believe future improvements of this project should relate to the acquisition of data for more cities and areas of interest, as this project was limited primarily by the difficulty in obtaining up-to-date crime data. Additionally, the application of more advanced machine learning algorithms could aid in both forecasting these factors and uncovering additional trends and insights.

7. Distribution of Team Member Effort

All team members have contributed a similar amount of effort over the course of the project.

8. References

1. Mohler, G., Bertozzi, A. L., Carter, J., Short, M. B., Sledge, D., Tita, G. E., ... Brantingham, P. J. (2020). Impact of social distancing during COVID-19 pandemic on crime in Los Angeles and Indianapolis. Journal of Criminal Justice, 101692.

2. Tayeen, A. S. M., Mtibaa, A., Misra, S. (2019, August). Location, location, location! quantifying the true impact of location on business reviews using a Yelp dataset. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1081-1088).

3. Han, X., Xu, S., Xiong, Y., Zhao, S. (2020, January). The Relationship between Customer Satisfaction and Location of Restaurant. In Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning (pp. 385-395).

4. Yoo, J. S. (2019, December). Crime data warehousing and crime pattern discovery. In Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems (pp. 1-6).

5. Rosenfeld, R., Lopez, E. (2020). Pandemic, social unrest, and crime in US Cities.

6. Boman, J. H., Gallupe, O. (2020). Has COVID-19 changed crime? Crime Rates in the United States during the pandemic. American journal of criminal justice, 45(4), 537-545.

7. Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., Liu, Q. (2019). Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access, 7, 106111-106123.

8. Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., Stanton, C. (2020). The impact of COVID-19 on small business outcomes and expectations. Proceedings of the National Academy of Sciences, 117(30), 17656-17666.

9. Dearman, D., Truong, K. N. (2010, September). Identifying the activities supported by locations with community-authored content. In Proceedings of the 12th ACM international conference on Ubiquitous computing (pp. 23-32).

10. Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. (2020). Real-time economics: A new platform to track the impacts of COVID-19 on people, businesses, and communities using private sector data. NBER Working Paper, 27431.

11. Chakraborty, S., Venkataraman, A., Jagabathula, S., Subramanian, L. (2016, August). Predicting socio-economic indicators using news events. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1455-1464).

12. . Rozsnyai, S., Schiefer, J., Schatten, A. (2007, June). Concepts and models for typing events for event-based systems. In Proceedings of the 2007 inaugural international conference on Distributed event-based systems (pp. 62-70).

13. . Kim, K. S., Bica, M., Kojima, I., Ogawa, H. (2015, November). RendezView: Look at Meanings of an Encounter Region over Local Social Flocks. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming (pp. 87-96).

14. Glaeser, E. L., Kim, H., Luca, M. (2017). Nowcasting the local economy: Using yelp data to measure economic activity (No. w24010). National Bureau of Economic Research.

15. Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. (2020). The Economic Tracker. Retrieved October 06, 2020, from https://tracktherecovery.org/

16. Morawska, L., Tang, J. W., Bahnfleth, W., Bluyssen, P. M., Boerstra, A., Buonanno, G., ... Haworth, C. (2020). How can airborne transmission of COVID-19 indoors be minimised?. Environment international, 142, 105832.

17. Humphries, J. E., Neilson, C., Ulyssea, G. (2020). The evolving impacts of COVID-19 on small businesses since the CARES Act.

18. Biswas, P., Saluja, K. S., Arjun, S., Murthy, L. R. D. Prabhakar, G., Sharma, V. K., DV, J. S. (2020). COVID 19 Data Visualization through Automatic Phase Detection.

19. Aigner, W., Miksch, S., Müller, W., Schumann, H., Tominski, C. (2007). Visual methods for analyzing time-oriented data. IEEE transactions on visualization and computer graphics, 14(1), 47-60

20. US Zip Codes Database - https://simplemaps.com/data/us-zips

21. Python-based Zip Code Database - https://pypi.org/project/uszipcode/

22. Baltimore Crime Data - https://data.baltimorecity.gov/Public-Safety/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij

23. Boston Crime Data - https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/12cb3883-56f5-47de-afa5-3b1cf61b257b

24. Denver Crime Data - https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime

25. Phoenix Crime Data - https://www.phoenixopendata.com/dataset/crime-data/resource/0ce3411a-2fc6-4302-a33f-167f68608a20

26. SMB Data - https://github.com/OpportunityInsights/EconomicTracker