University of Michigan School of Information
Master of Applied Data Science
Milestone I Project

Descriptive and Geographic Analysis of COVID-19-associated Economic and Public Health Intervention and Relationship to Essential Services and Consumer Behaviorism

Jenna Mekled
Tony Purkal

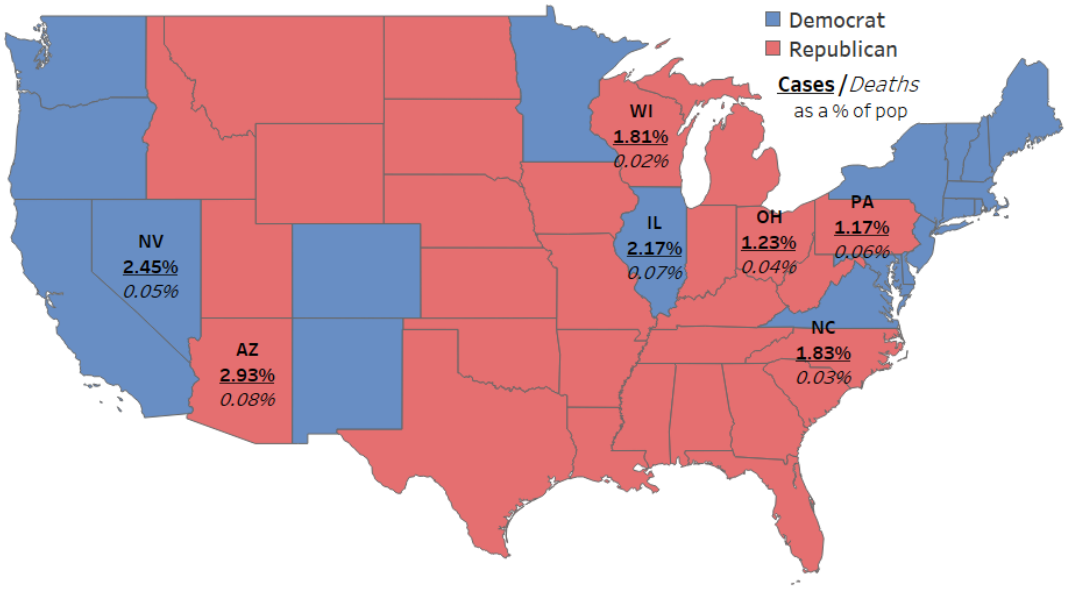
Fall 2020

MOTIVATION

2020 will forever be a year marked by the severe acute respiratory syndrome coronavirus two, more commonly known by the disease it inflicts: COVID-19. With the spread of this virus, another crisis has manifested in the form of panic and fear with a lack of understanding about the virus as well as a lack of policy decisions implemented to prevent the spread. In the United States COVID-19 has even become a political issue, one that will likely throw the upcoming political landscape into pandemonium. In addition to the stalling of in-person campaigning and voting, politicians in this upcoming election will be judged heavily on their responses to the pandemic. In addition to a nation of people reeling from sickness, so are businesses, especially small businesses, because of both the altered behaviors of the people as well as the various restrictions put in place by individual state governments.

The purpose of this project is to investigate the survival of essential services during COVID-19 based upon various economic factors, public health measures, and crowdsourced consumer sentiment. The project will highlight the implications of localized political leaning, as well as the consequences of political polarity in public health decision-making and its relation to survival of businesses and the highlights of adapted services they provide. The nature of this project was decidedly important to pursue given its prominent and pervasive nature and current relevance.

This political map of the Case Count and Death Counts of victims of COVID-19 as of September, 20, 2020* sets the context for the 7 states we analyzed within this study.



*an interactive HTML visual with all Case/Death counts is available for download here

Specific questions within this context include:

- Does consumer behavior vary geographically, and in what ways?
- Does the political climate in a state impact business survival?
- What are the impacts of public health measures like stay-at-home orders and social distancing advice?
- How is consumer sentiment related to public policy?
- Did PPP Loans do anything to help businesses survive?
- Which business categories are impacted most and how can they pivot and market?
- Can we characterize any consistent consumer behaviors?

DATA SOURCES

Yelp COVID-19-related Business Highlights Addendum Dataset
Business operations updates due to COVID-19

Size	64.8 MB / 209,795 records
Format	JSON
Access method	Download via web (publicly available)
Location	https://www.yelp.com/dataset
Variables	business_id, Temporary Closed Until

business_id

Yelp Open Dataset (business subset)
Businesses with category, location, and operations data

Size	152.9 MB / 209,393 records
Format	JSON
Access method	Download via web (publicly available)
Location	https://www.yelp.com/dataset
Variables	business_id, categories, city, is_open, latitude, longitude, state

state

State/US Abbreviation Data Dictionary
List of abbreviations for US states

Size	1.5 KB
Format	Text
Access method	Read from web
Location	https://www2.census.gov/geo/docs/reference/state.txt
Variables	STUSAB, STATE_NAME

STUSAB

State-level Government COVID-19 Responses
State-level COVID-19 activity restrictions and timing Wiki

Size	2.9 MB
Format	HTML
Access method	Scraped from web with BeautifulSoup
Location	https://w.wiki/cws
Variables	State/territory, State of emergency declared, Stay at home ordered, Out-of-state quarantine

State/territory

Small Business Administration Paycheck Protection Program Loan Data
Distribution data for the program designed to provide direct incentives for small businesses to keep workers on the payroll

Size	879 KB
Format	PDF
Access method	Extracted from PDF using tabula
Location	https://home.treasury.gov/system/files/136/SBA-Paycheck-Protection-Program-Loan-Report-Round2.pdf
Variables	State, Loan Count, Net Dollars

State

Web-scraped COVID-19 Tweets containing COVID-19 related hashtags
Tweets Mar-Apr 2020 with hashtags #coronavius #coranavirusoutbreak #coronavirusPandemic #covid19 #covid_19

Size	5.83 GB
Format	CSV
Access method	Downloaded from web, prcoessed in Alteryx
Location	https://www.kaggle.com/smid80/
Variables	text, place_full_name

US Cartographic Boundary File
States

Size	191 KB
Format	XML shapefile
Access method	Download from web
Location	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

Insight

Visualizations

DATA MANIPULATION

Overarching Methodology

Data from individual sources was manipulated into pySpark DataFrames, then registered as tables within a SparkSession for compilation of all data into a master DataFrame by executing SQL join clauses on geographic location over the individual source tables. Further analysis was performed on temporary views of this master pySpark DataFrame using groupBy aggregation methods and conditional Column filter expressions and Boolean filters using the INSTR substring column function.

Yelp COVID-19-related Business Highlights Addendum Dataset

Though the more comprehensive Yelp Open Dataset contains information related to business closures, this specific dataset includes information related to temporary business closures due to COVID-19 in addition to other highlights of special offerings during COVID-19 such as gift cards, delivery, curbside pickup, curbside drop-off, drive-thru, takeout, shipping, online classes, virtual estimates, mobile services, and remote services.

The field of data containing information about COVID-19-related temporary closures consists of string data that encodes the anticipated end of the temporary closure or a 'FALSE' value if the business remains open. A new column was created with conditional filter expressions to provide Boolean data to filter the data or aggregate the count of closed businesses. For the data related to highlights of special COVID-19 related offerings, a new column was created with the length of that field for each business, with this quantity serving as a quantitative index of businesses offering these services for future statistical analysis (where higher numbers indicate more service offerings and vice versa).

The quantity of records in this dataset required pySpark as a more scalable tool for data manipulation. Because this addendum dataset contains only data for participating businesses, intuitively, checks for missing data were negative.

Yelp Open Dataset (business subset)

The business_id variable of this dataset allowed it to be joined via pySpark SQL with the matching column in the Yelp COVID-19 addendum dataset. In addition to providing various other information for businesses, one of the most important pieces of data that remained untouched was the state within which the business is located which is used in subsequent DataFrame joins. Notably relevant to this variable, however, is its subsequent use in aggregation during grouping the businesses by state to count the number of businesses.

The categories column was used to create new columns of Boolean values for various business categories and subcategories for subsequent use in filtering and aggregation of data to compare closure rates among various categories of business. This was accomplished with the INSTR substring function applied to the categories column. The City variable of this dataset was subsequently used in a final combined DataFrame to create a new column through conditional filter expressions to provide Boolean data for filtering on whether the business is located in the urban center of the metropolitan area or more toward the suburban or rural periphery. Here, the urban class was operationalized by location in the named metropolitan area city, and the "not urban" class by location in a city outside that named metropolitan area city but within the same state. During grouping by state, the mean value was chosen as the aggregated metric for

latitude and longitude so that geographic plotting tasks employed in the Analysis & Results section would have geo scatter marks centered on the metropolitan area within the state.

State/US Abbreviation Data Dictionary

During reading this into a Pandas DataFrame using the read_table function, the Geographic Names Information System (GNIS) codes were dropped before creation of the pySpark DataFrame. This DataFrame is used in joining the remaining state names and abbreviations to the combined Yelp datasets, and for subsequent joining with web-scraped data sets that are classified by state full name.

State-level Government COVID-19 Responses Wiki

Wikipedia HTML was requested with the requests package, the markup processed with BeautifulSoup, and the table of data read into a Pandas DataFrame. Minor cleanup of data columns containing dates was performed with removal of data source attributions with regular expressions. These columns of dates were converted to Python datetime objects, various arithmetic operations were applied between the columns of dates, and the day attribute was extracted from the calculated timedelta objects to create new columns: the time from state of emergency declaration to issuance of stay at home orders, each state's difference in declaration of state of emergency from the mean date of declaration calculated from all states, and each state's difference in issuance of stay at home orders from the mean date of issuance calculated from all states.

Small Business Administration Paycheck Protection Program Loan Data

The tabula-py package was used to read tables from a PDF on the web into a Pandas DataFrame. The table consisted of three separate columns, for which it was necessary to rename columns before concatenating the nested tables containing the separate columns into a comprehensive DataFrame. Currency-related characters were removed from quantitative columns for the number of loans disbursed per state and the total dollar amount of loans disbursed to businesses in each state. Each column in the Pandas DataFrame was cast to string type to avoid merge type errors during subsequent creation of a pySpark DataFrame. This data was joined with other source data on the state column, and in subsequent groupBy operations by state, the quantitative columns from this data were also grouped to avoid unwanted aggregation.

Web-scraped COVID-19 Tweets containing COVID-19 related hashtags (for sentiment analysis)

This public Kaggle dataset consists of 33 individual, daily CSV files representing the most recent sample of tweets for the month of April 2020. The dataset was cleaned by stripping the text of each tweet of stop words, punctuation, and special characters. A sentiment analysis was performed using the Python TextBlob library to transform these cleaned strings into TextBlob objects. Calling the sentiment method on these TextBlob objects yields a tuple with a polarity score and a subjectivity score. New columns for these scores were added by applying the method to the TextBlob objects, a float from -1 to 1 and a float from 0 to 1 (where 0 is very objective and 1 very subjective), for polarity and subjectivity, respectively. Regular expressions were applied to the location data to extract the state abbreviation by which the data was grouped and the mean of each sentiment score was aggregated.

To accomplish these tasks across the 33 files, Alteryx, a workflow designer, was used to filter for US specific tweets and the date of each file was written into the dataset. Ultimately all of the data files were appended into one large csv file that became the basis of the sentiment analysis of tweets as previously described.

Manipulation of the data in aggregate form

Joining of the various data sources followed the relationships laid out in the Entity Relationship Diagram that constitutes our Data Sources section. After all of the aforementioned data sources were manipulated and aggregated into a single DataFrame it was exported from pySpark to Pandas. Prior to this, some data columns contained string data, especially web-scraped data, that required using the cast method with DataType instances such as IntegerType and DoubleType so that the data could be ready for statistical analysis. Because several slices and subsets of the total data would be subject to statistical analysis, functions were defined to standardize the preparation of calculating correlation coefficients and p-values for a correlation matrix to be visualized as a heatmap. Finally, a column of ordinal variables was added to the DataFrame to represent each state's degree of either Democratic or Republican political affiliation.

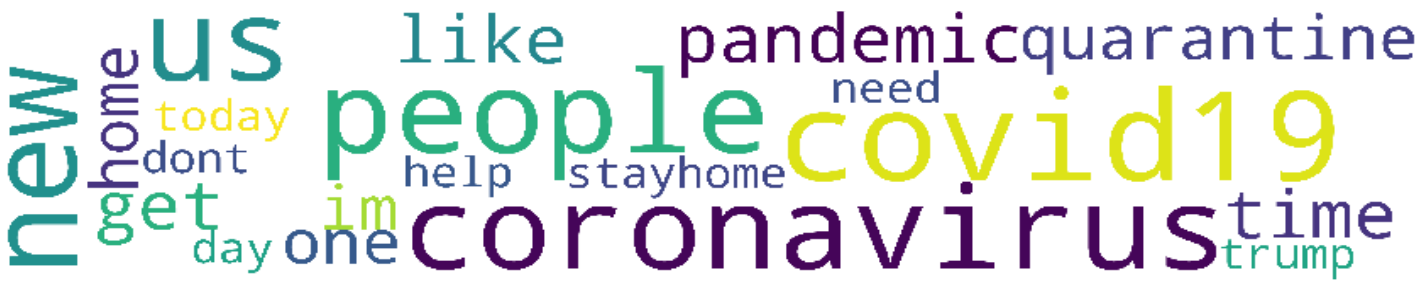
Though public health policy data had a sense of time, in addition to social media sentiment, the Yelp Open Dataset and addendum lack this data. Due to the large number of records and homogeneity for businesses not engaged in or not reporting highlights of extra COVID-19 services, questions related to the correlation between geographic and public health policy measures became difficult. The dependent variable in this exploration is dichotomous and bivariate correlation is therefore not possible, and point-biserial correlations were used instead. Having so much data from the Yelp Open Dataset and so little data from the COVID-19 Business Highlights Addendum also thwarts our ability to build a logistic classifier because of the enormous size and influence of the class where no highlighted services are offered, even after attempting to fit the classifier though stratified sampling. A Random Forest Classifier performed better with minority classes, with better accuracy and improved AUC - ROC, but precision and recall was still poor.

ANALYSIS & VISUALIZATION

There is much to be learned about how businesses are responding to the COVID-19 pandemic and the specific consumer demands for goods and services. In addition to keeping staff on the payroll, the Small Business Administration's Paycheck Protection Program (PPP) loans from Coronavirus Aid, Relief, and Economic Security (CARES) Act funds was designed to allow some businesses to remain open and even engage in innovative solutions for their customers. State participation in these types of stimulus and safety-net programs have always been variable, and typically responsive to the greater sociopolitical climate of the state, even in regards to socioeconomics of the private sector.

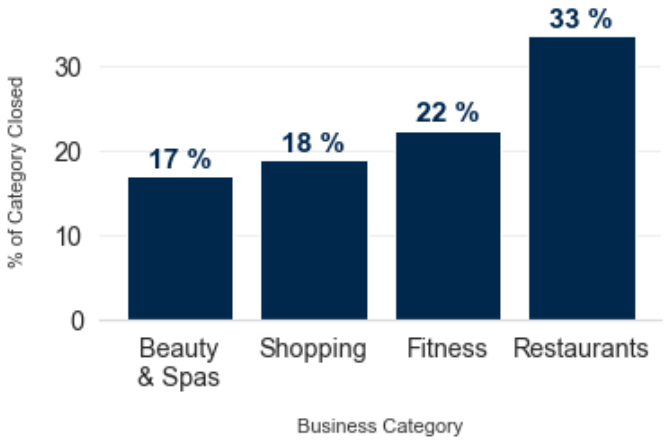
Social media's credibility for providing insight is demonstrated by *Figure 1*, with some clues into what will help us people get what we need and want during the COVID19 pandemic in a new time of staying home in quarantine.

Figure 1. Word cloud of most frequent words from social media sentiment analysis of Tweets from April 2020



While the data that is available surrounding initiatives such as offering gift cards, delivery, curbside pickup, curbside drop-off, drive-thru, takeout, shipping, online classes, virtual estimates, mobile services, and remote services is as of yet still fledgling, Yelp's COVID-19 addendum dataset serves to begin the documentation of these service highlights. This study provides a preview into what socioeconomic, political, and consumer pressures are present, and some foresight into the success of engaging in specific actions by assessing point-biserial correlations between the Yelp data and available data pertaining to public health measures, PPP loan funding, geography, the general political leaning of the state in which a business operates, and even social media consumer sentiment.

Figure 2. COVID-19 Business Closures



To begin, we get some perspective on the scope of business closures, despite examples of the beauty sector offering internet tutorials and virtual consultations, stores offering curbside pickup, gyms and trainers offering online training sessions, and restaurants switching to high-margin takeout and delivery, by looking at each major consumer sector in *Figure 2*. It is important to note that this data was mostly collected before stay-at-home orders went into place and the extent of public health policy was generally advisory, so it closely follows consumer behavior. In addition to not capturing reduced staffing and decreased hours of operation, not capturing mandatory measures makes it likely that these closure levels are an underestimation.

For restaurants - the largest category of closures - investigating correlation between geographic location and percentage of COVID-19 closures showed a significant ($p = 0.01$) strong negative correlation ($corr = -0.87$) between percentage of closures and longitude. With the idea that there is some relationship between closures and geography, we were motivated to consider geography, in addition to other sociopolitical factors, across several categories of businesses.

Postulating that more granular detail from the broad categories from *Figure 2* could possibly provide more insight into consumer behaviors, we compared the closure rates of several subcategories, beginning with restaurants (**Table 1**). Two subcategories were defined: indoor group dining experiences (buffets, fondue, hot pot, tapas) and home dining (fast food, drive-thru, and Chinese or pizza delivery or takeout). For the indoor dining subcategory, a highly significant ($p < 0.001$) strong positive correlation ($corr = 0.92$) was found between a state's political leaning and percentage of businesses closed (*Figure 3).* The Democratic leaning states have over 50% group dining closures, while Republican states have less than 50%. For the purposes of correlation analysis, state political leaning was operationalized on an integer ordinal scale with -2 being highly Republican and +2 being highly Democratic according to state wins in the 2016 presidential election.

Figure 3. COVID-19 Related Group Dining Business Closures

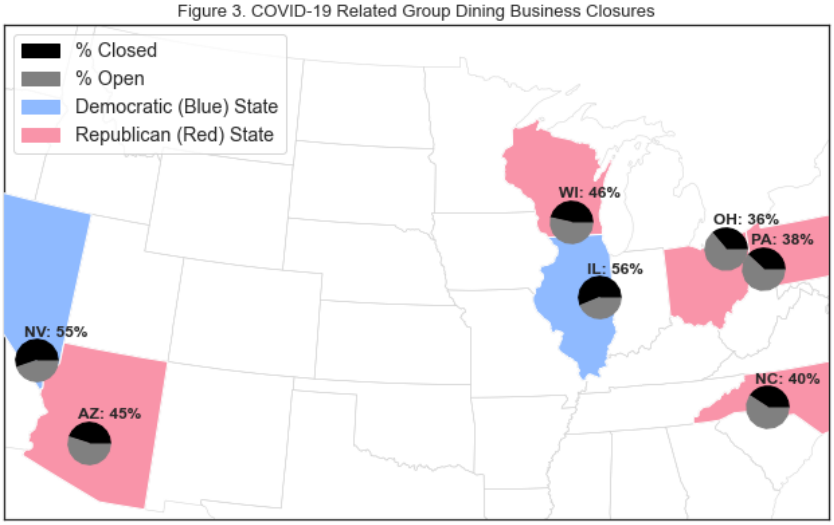
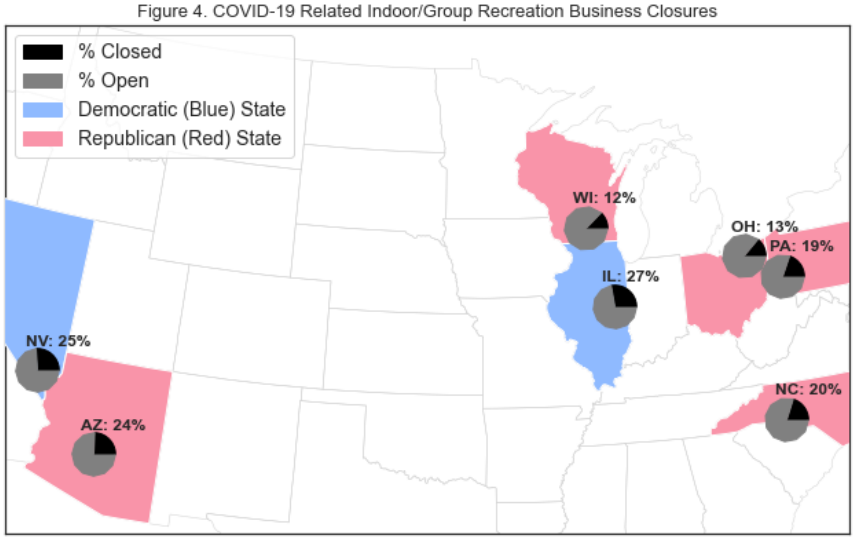


Table 1.	Sub-category	% Closed
	Indoor Group Dining	46.1
	Home Dining	22.9

For the next largest category - fitness - two subcategories were defined (**Table 2**): indoor/group recreation (axe throwing, bowling, escape games, fitness centers, go karts) and non-team outdoor recreation (amusement parks, bike rentals, fishing at lakes, golf, hiking at trails or parks, horseback riding, mini golf, mountain biking). Similar to indoor group dining, the indoor/group recreation subcategory had a significant (*p = 0.03*) strong positive correlation (*corr = 0.8*) between a state's political leaning and percentage of business closed. Here, the correlation was not as highly significant or strong as for indoor group dining, and other possible variables in play might be the difference in the spread between the subcategories' closed percentages or weather and climate differences based on geography. Nonetheless, it seems likely that political climate matters when attempting to characterize consumer behavior.

Table 2.	Sub-category	% Closed
	Non-Team Outdoor	12.6
	Indoor/Group	22.5

Figure 4. COVID-19 Related Indoor/Group Recreation Business Closures



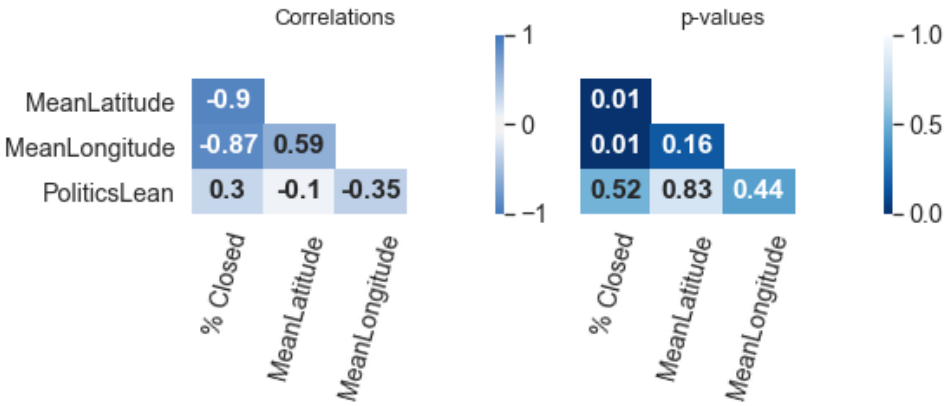
Next, we considered shopping; specifically subcategories related to essential food, household products, and personal care items. There has been no shortage of consumer sentiment and newsworthy behavior of the panic variety in seeking these essential services. Significant correlations found in other categories were not found here, perhaps an indication of the universality in consumer sentiment for these essential services, independent of politics or geography. **Table 3** does offer some insight into behaviors surrounding shopping for these necessities at three subcategories of stores: warehouse clubs, grocers, and drugstores. It could be possible that warehouse clubs remained open to a greater extent than grocers as people engaged in bulk-buying behaviors due to uncertainty early in the pandemic when this data was collected. An ever greater difference is seen with drug stores remaining open, perhaps servicing the demands for medicine and personal care items.

Bars have been a major news topic throughout the pandemic, and while liquor stores are usually regarded as essential services, it seems from media reports that perhaps many American also believe bars are essential. In **Table 4**, the percentage of liquor stores closed is less than that of nightlife (including bars, clubs, etc.), and significant negative correlations were identified between both latitude and longitude and closures (*Figure 5*).

Table 3.	Sub-category	% Closed
	Wholesale Stores	14.5
	Grocery Stores	17.1
	Drug Stores	7.2

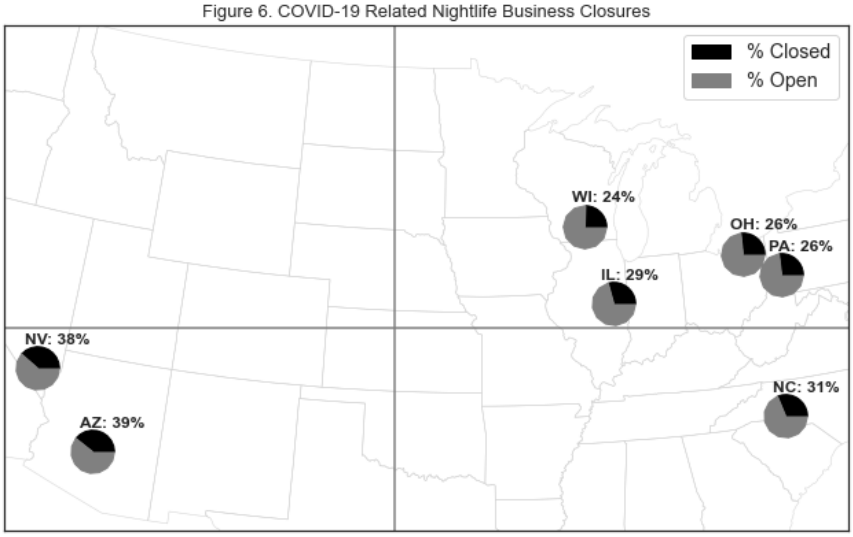
Table 4.	Sub-category	% Closed
	Nightlife	33.8
	Liquor Stores	27.4

Figure 5. Nightlife Category Correlations



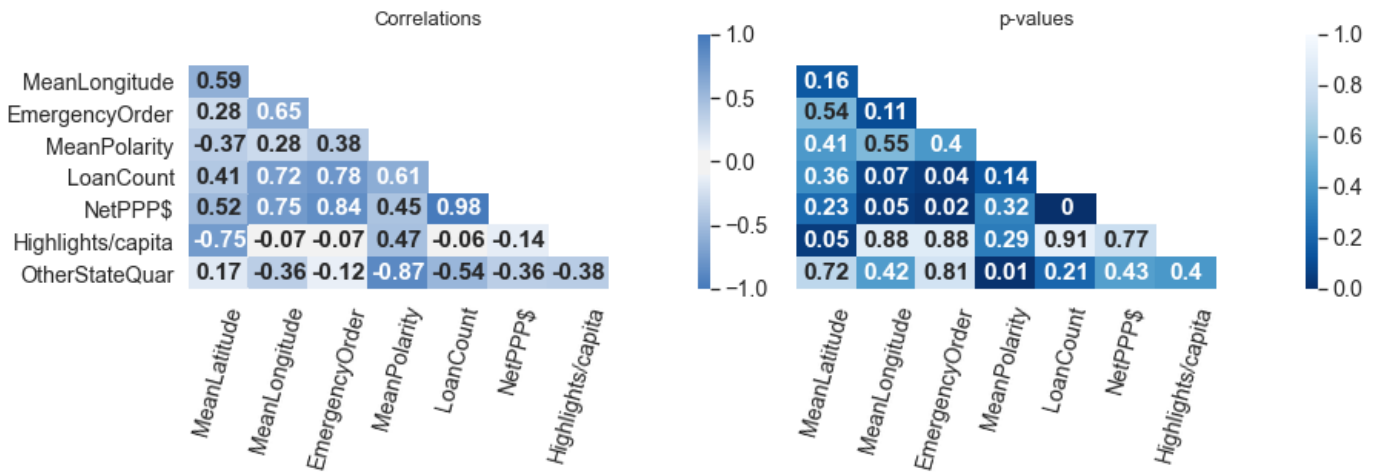
While geography is significant in regards to consumer behavior here, it is unclear if it is prevailing attitudinal differences that are not captured in our social media sentiment analysis, political policy decisions states have made in response to COVID-19, or another specific set of factors that would require more research. The trend toward staying out of bars in the western and southern states is seen in the bottom quadrants of *Figure 6*.

Figure 6. COVID-19 Related Nightlife Business Closures



Having nearly exhausted our granular explorations of consumer categories and subcategories where geography and geopolitics became a recurrent theme, we then shifted our attention to expanding our scope to businesses across all categories. We set out to explore what geographic and political pressures might be present as states begin major shifts in evolving public health policies and engaging in economic stimulus activities in an attempt to resuscitate spiraling economies, fractured supply chains, and escalating consumer uncertainty. In addition to uncovering another geographic variation related to consumer demand, where highlights of special COVID-19 services offered by businesses in a state have a significant (*p = 0.05*) strong negative correlation (*corr = -0.75*) with latitude, other significant relationships between indicators for economic stimulus, public health orders, and consumer sentiment were found (*Figure 7*).

Figure 7. All-Category Correlations



Perhaps signalling the importance of considering more interactions between public sentiment and public health policy as both continue to evolve through the pandemic, is the significant (*p = 0.01*) albeit weak negative correlation (*corr = -0.12*) between Tweet polarity and whether a state has implemented quarantines for out-of-state visitors. It also turns out that our premonition for geographic distribution of economic stimulus funds and state's varied participation and representation in the PPP loan program based on geography and public health decisions are not necessarily counterintuitive: with total PPP loan dollars in states being significantly (*p = 0.05 and 0.02*) strongly positively correlated (*corr = 0.75 and 0.84*) with longitude and the timeliness with which a state proclaimed a state of emergency, respectively.

With at least some of our premonitions regarding PPP loan program funds justified, we decided to take a deeper look at

PPP loans and business closure percentages in a greater level of geographic detail. Up until this point, the business data was considered in the scope of the state in which it operates because most of our analysis considered state-level socioeconomic variables. Already identifying the relationship between PPP loan funds and longitude, we instead operationalized two classes from the seven metropolitan areas represented in the Yelp Open dataset: urban-based businesses (business's city listed as metropolitan area name) and suburban/rural-based businesses (city listed as a name outside the urban core of the metropolitan area).

Overall, a higher percentage of businesses were closed in urban areas compared to non-urban, and this could be related to the overarching effects of the relationships between longitude and business closures like that resulting from our exploration of restaurants and nightlife (*Figure 5*). Nonetheless, it is difficult to ignore the significant (*p = 0.05 and 0.03*) strongly negative relationships (*corr = -0.76 and -0.8*) between suburban business closures and both the number of PPP loans and net dollars, respectively, as a signal for differential economic stimulus effects in regards to population density (*Figure 8*).

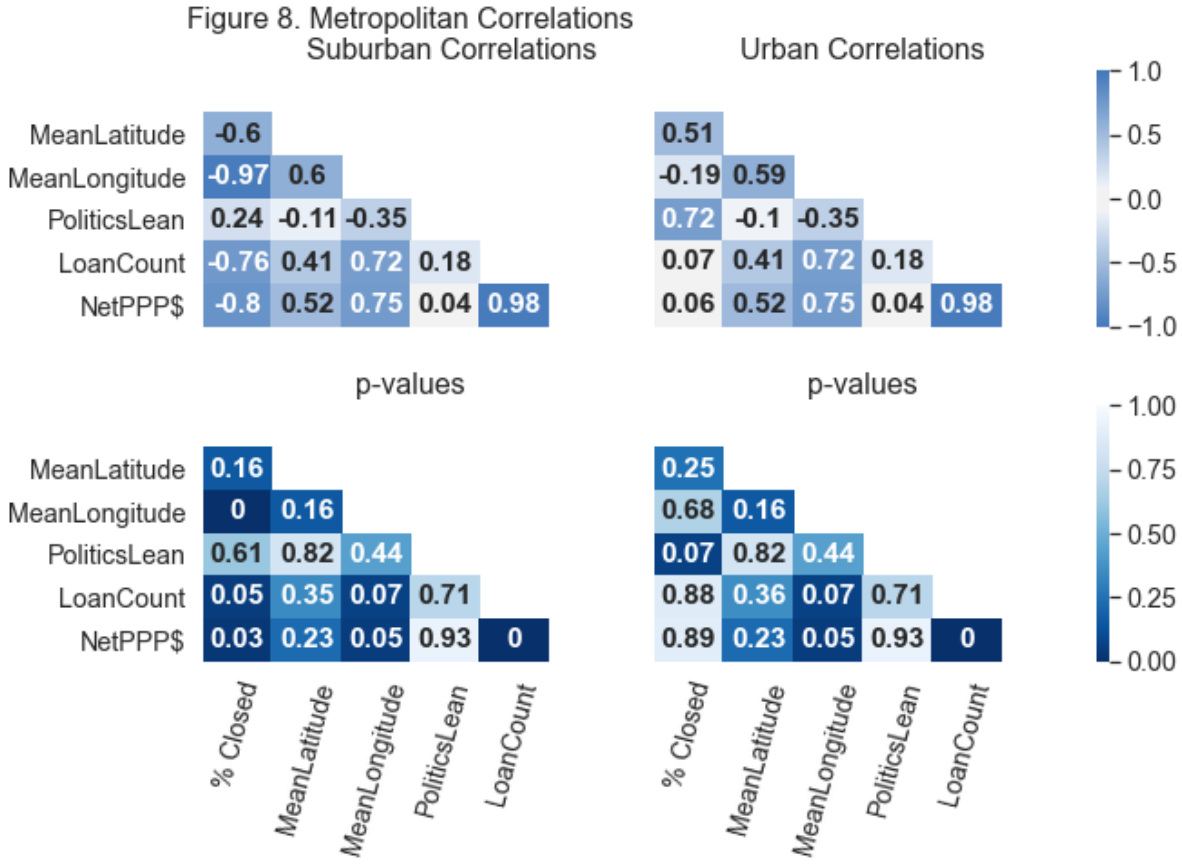


Figure 8. Metropolitan Correlations

In addition to not fully entrusting the relationship between differences in population density and PPP stimulus effects due to possible longitude and business closure relationship effects (*Figure 5*), there are also possible effects due to a relationship between longitude and PPP funds (*Figure 7: p = 0.05, corr = 0.75*). These issues of possible multicollinearity and the aforementioned suspected underestimation of closures made the attempt to successfully fit a logistic regression classifier for predicting business closure rates or offerings of COVID-19-related service highlights difficult due to drastically unequal balance in the dichotomous target outcomes. Engaging in this perhaps too-complicated-to-model task does not call for more data — the sample size of the Yelp dataset provides plenty statistical power — but higher quality data about closures and a more heterogeneous geopolitical sampling of more than seven metropolitan areas.

Nonetheless, this descriptive analysis can serve to inform any who wish to attempt such a predictive task in choosing candidate variables and with awareness of potential bias pitfalls.

CONCLUSION & NEXT STEPS

Ultimately, in our sample of seven metropolitan areas, more PPP dollars in an area was correlated with fewer closures overall. Additionally, states that declared a state of emergency sooner were likely to bring in more PPP dollars. Considering most of the Yelp data used in this study was collected before most stay-at-home orders, restaurants were unsurprisingly hit the hardest in terms of business closures. As exemplified in *Figure 3*, the Democratic leaning states in this study had a larger amount of closures (> 50%) surrounding group dining than Republican states (< 50%). This correlation, while not as strong, was also present regarding indoor/group recreation businesses (*Figure 4*). Concerning consumer sentiment, there was a significant weak negative relationship between public sentiment via Twitter data collected and whether a state enacted out-of-state quarantines that could benefit from further research. Some consistent behaviors of consumers discovered were a general increase in buying bulk, and a greater focus on necessary items such as drugs, personal care, and other necessities. In addition, consumers were more likely to avoid sit down restaurants generating a greater dependency on fast food, drive-thru, curbside pickups, and delivery. The closure of gyms signals an increase in outdoor recreational activity such as running and hiking, weather permitting, of course.

While the country continues to adjust and respond to the COVID-19 outbreak, on the horizon is the upcoming 2020 election, and in tandem to political sentiment beliefs surrounding the virus, continue to further divide the country. The implications of the relationships between COVID-19, future political climate, and the state of the country's small businesses could benefit from future-focused exploration. Owing to awareness from civil unrest, there seems to be increased interest in supporting minority businesses. Accordingly, Yelp has expanded options that will likely be available in future data sets for black- and women-owned businesses. Considering the relationships between these businesses and closure rates, sociopolitical features, and economic stimulus distribution and effect are areas of possible future research. Reaching out to industry experts for advice on refining the operationalization of variables such as indoor versus outdoor recreation business and soliciting review of study assumptions could lend more clarity and power to this type of study.

STATEMENT OF WORK

Tony Purkal, project lead, is experienced in GIS and leveraged this skill set for effectively communicating some data-driven insights. Tony performed web-scraping for data extraction as well as working on correlation analysis to determine insight into the general relationship between the variables, but fundamentally, their suitability for future predictive studies outside of the scope of this project. Tony completed the final draft of the project proposal.

Jenna Mekled, an experienced data analyst, completed the data extraction, cleaning, and Sentiment analysis on the Twitter Data to create visualizations. Jenna created the first draft of the project proposal based off of Tony's initial vision for the project and worked with the sentiment data used for analyzing correlations.

Jenna worked on the motivation section, including the opening visualization and Plotly interactive visualization, and conclusion of the project, as well as relevant sentiment sections within the Data Manipulation segment, including the word cloud of most frequently used words. Tony worked predominantly on the Data Manipulation and Analysis and Visualization section of the report creating the majority of crucial visuals that convey the findings of the project and interpreting results, as well as setting up the layout. Both Tony and Jenna performed formatting and testing the final project and package.

ADDITIONAL REFERENCES

(2015, June 07). Retrieved September 28, 2020, from https://planspace.org/20150607-textblob_sentiment

(Tutorial) Generate Word Clouds in Python. (n.d.). Retrieved September 28, 2020, from https://www.datacamp.com/community/tutorials/wordcloud-python

Chugh, A. (2018, November 04). Aakash Chugh. Retrieved September 28, 2020, from https://data-science-blog.com/en/blog/2018/11/04/sentiment-analysis-using-python/

Federal Information Processing Standard state code. (2020, June 14). Retrieved September 28, 2020, from https://en.wikipedia.org/wiki/Federal_Information_Processing_Standard_state_code

Red states and blue states. (2020, September 22). Retrieved September 28, 2020, from https://en.wikipedia.org/wiki/Red_states_and_blue_states

United States COVID-19 Cases and Deaths by State over Time. (n.d.). Retrieved September 28, 2020, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data