EPiC Series in Computing

Volume 75, 2021, Pages 22–31

CAINE 2020. The 33rd International Conference on Computer Applications in Industry and Engineering

Determining Anti-Curve-Flattening Behaviors for COVID-19 in the United States

Rehab El-Hajj[1], Sarah Davis[2], and Raymond Otieno[3]

[1] University of University, Edmonton, Alberta, Canada
relhajj@ualberta.ca
[2] University of University, Edmonton, Alberta, Canada
sdavis@ualberta.ca
[3] University of Alberta, Edmonton, Alberta, Canada
rotieno@ualberta.ca

Abstract

COVID-19 has arguably impacted every dimension of social living — be that employment, schooling, healthcare or recreational activities. In a matter of months, businesses have shut down and the workforce and schools have been redirected to online work in many regions of the world. One key element of the North American pandemic response has been the emphasis that the spread or prevention of the pandemic is largely dependent on the measures taken by residents of any region. As such, our research focuses on outlining the factors that determine if an individual is less likely to take this pandemic seriously (i.e. is taking fewer measures to prevent the spread of COVID-19). We have analyzed the results of a U.S. wide COVID-impact survey using random forest classification (RFC) to associate individual demographic factors to measures taken against the pandemic such as washing/sanitizing hands. Our results indicate that the top three influential factors are household size, the number of adults living in one household and the health of the respondent (poor to excellent). Using these insights, we used association rules to determine key combinations of features that may lead to an apathetic response to a global pandemic in U.S. citizens, such as lower income households.

Keywords

COVID-19, prevention, machine learning, data mining, random forest, association rules

1 Introduction

2019-nCoV, or COVID-19, is a novel coronavirus known to affect the respiratory system. It comes from the viral family *Coronaviridae*, which are positive-sense, single-stranded RNA viruses. Other well-known coronaviruses are SARS-CoV and MERS-CoV, which have both caused significant outbreaks in the past. COVID-19 is transmitted through aerosol droplets or direct contact with secretions from an infected individual [1, 2, 3]. COVID-19 has quickly

become a worldwide pandemic; there have been approximately 5.99 million cases of COVID-19 globally, resulting in 367K deaths [4]. More recently, the United States (U.S.) has become the pandemic's epicenter, reporting 1.76 million cases of COVID-19 patients since the first case in January 21, 2020 [5]. Of these, there have been 103K deaths (these numbers are reported as of 12:00 P.M. May 30^{th}, 2020). The severity of this virus has called for everyone to play a part in prevention of its spread; the general public has been advised to practice social distancing, many non-essential businesses were required to close, and countless research facilities, government institutions, and universities have focused their energy on researching the virology and epidemiology of COVID-19.

In the U.S., many state-level governments have implemented public policies to prevent the increased spread of COVID-19. While some states such as New York, California and Illinois have continued to impose statewide stay-at-home orders, other states such as Texas, Florida and Alabama have gradually lifted these restrictions and are beginning to reopen non-essential businesses and allowing some aspect of public gatherings — lifting of restrictions has largely been attributed to their relatively minute number of cases [6]. These measures require the public's cooperation to ensure everyone takes precautions to limit the spread of COVID-19 while mitigating the current effects this pandemic has had on the economy.

A key element for tackling the COVID-19 pandemic and preventing spread are the measures taken by individuals such as increased hygienic actions and practicing social distancing. Therefore, our paper aims to study and outline the factors that determine if an individual is more (or less) likely to take this pandemic seriously. This question holds great value, especially as many city-wide, state-wide and even country-wide governments are looking to prevent increased escalation of COVID-19 cases. Especially considering that even individuals who feel healthy must practice relevant safety measures such as social distancing and wearing face masks in order to protect the community as a whole (in case they are asymptomatic).

We used random forest decision tree analysis and association rules. Random forest is an ensemble model consisting of many decisions trees (visual representations of tree-like graphs which model decisions and their possible consequences) which can be used for both classification and regression [7]. We used random forest analysis to identify the most frequent or most *influential* factors for determining whether or not an individual will take extra measures against COVID-19. Association analysis is the process of discovering frequent sets of interesting relationships between data points. [8]. Association rules are of great use in our research in allowing us to outline associations between personal health decisions and individual factors. The combination of these machine learning tools has culminated in a number of interesting results with respect to what factors influence safety measures taken by groups of individuals.

1.1 Literature Review

Multiple public health policies have been implemented globally to try to limit the spread of COVID-19 and it is up to governments, businesses and the general public to follow these guidelines carefully. Often, "curve-flattening" policies include business closures, social distancing protocols, increased hygiene, and stay-at-home orders.Exploring potential evidence-based policies and their challenges from previous outbreaks is imperative to successfully gain insight on achieving the public's much needed cooperation.

Tuncer et al. conducted a study which evaluated the efficacy of certain control measures such as quarantining, isolation, education, social distancing, and safe burials that reduced transmission of the Ebola epidemic in Liberia [9]. In their study, they used a model selection analysis approach and obtained WHO reports to collect Ebola outbreak data. They were able

to develop a predictive model that best matched the Ebola infection and death data and were then able to use this model to determine that social distancing was the most impactful control measure in decreasing the spread of Ebola in Liberia.

Blair et al. conducted an in-person representative survey to determine public levels of distrust in Liberia to find out how this affected the public's compliance with control strategies during the recent Ebola outbreak [10]. They found no significant correlations between distrust in government and erroneous beliefs on the epidemic. In fact, they found that survey participants who were more informed about Ebola had less trust in the government than those less informed. Unsurprisingly, trust in government was negatively correlated with hardships experienced amidst the Ebola epidemic. They concluded that trust in government is an imperative determinant of the population's compliance with public health policies and they speculate that increasing this trust in government could be obtained by collaborating INGOs, which are more highly trusted. This study exemplifies how the public prioritizes their health and are willing to follow safety measures in order to decrease the spread of infectious diseases but gaining the public's trust of the government is crucial for all members of the nation to follow preventative protocols.

Tsai et al. conducted a study to determine what governments can do to increase trust when protecting the public's wellbeing [11]. The researcher's conducted a representative study on Liberians' attitudes towards their government during the Ebola outbreak partially using the data obtained from the study Blair et al. conducted in 2017. The researchers did a follow up survey three months later. They were able to conclude that government outreach had significant impact in public cooperation in following the guidelines, even with policies such as bans on social gatherings, which were quite controversial. Tsai et al. found that the approach of the government's outreach during the Ebola epidemic allowed Liberians to increase their knowledge of Ebola, as well as be more compliant and supportive of the preventative public policies implemented to reduce the spread of the deadly virus.

2 Materials & Methods

We have analyzed the results of a U.S. wide COVID-impact survey[1], to determine what individual factors determine if an individual will take the pandemic seriously [12]. To analyze these results, we first familiarized ourselves with the data-set (refer to section 2.1), we then prepped the data for in depth analysis (refer to section 2.2). Finally, we completed random forest and association rule analysis to outline what factors are most influential for determining who refrains from taking measures against COVID-19 (refer to section 2.3).

2.1 COVID-Impact Survey

The Data foundation, which hosted the COVID-impact survey, has delivered the same survey on two dates so far (April 30^{th}, 2020 and May 12^{th}, 2020) [12]. Each date, a different set of randomized individuals in various states in the U.S. are contacted. The April 30 survey has a total of 8790 distinct records and the May 12 survey has a total of 8974 distinct records, where each record represents a single individuals' survey results. There are 174 columns in the survey outlining participants' responses to many questions on their employment status, age, gender, race, ethnicity, region/place of residence, total household income, self-reported mental state in response to COVID-19, diagnosed medical conditions, whether they have/know someone who

[1] https://www.covid-impact.org/

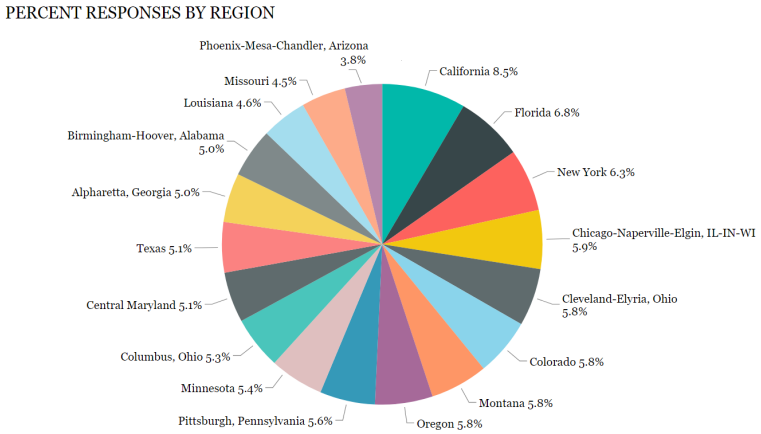

Figure 1: Percent of participant responses by region in the U.S. The pie chart above displays the distribution of survey participants' place of residence during the COVID-19 pandemic. There are a total of 18 distinct regions (8 are states, 10 are metropolitan areas).

Which of the following measures, if any, are you taking in response to the coronavirus?
Worn a face mask
Cancelled or postponed dentist or other appointments
Avoided some or all restaurants
Cancelled or postponed pleasure, social, or recreational activities
Stockpiled food or water
Avoided public or crowded places
cancelled or postponed dentist or other appointments
Avoided contact with high-risk people
Washed or sanitized hands
Kept six feet distance from those outside household
Wiped packages entering home

Table 1: The ten binary (yes/no) questions we used to represent taking the pandemic seriously.

has been confirmed with COVID-19 etc. For our analysis, we have chosen to focus on the survey administered on April 30^{th}, 2020.

Initially, we were interested in understanding the distributions of the survey participants based on their region Figure 1 outlines the distribution of regions for all survey participants. In total, 82.7% of the survey participants specified where they live, while 17.3% did not input their state or city of residence. Figure 1 displays the results of those who did input their place of residence. Of these participants, there were 18 distinct regions throughout the U.S. included in the data.

For our analysis, we selected a subset of questions to consider as our target variables. The original question had 19 sub-questions relating to preventative actions taken; however, we selected 10/19 to analyze as these questions were about general safety protocols and distancing procedures (i.e., not just to be taken by sick people). Each of these 10 sub-questions are binary, with a simple yes/no answer and no missing data. Table 1 displays all these questions. Refer to Figure 2 for the distribution of yes/no answers for all 10 questions/measures taken. Our focus

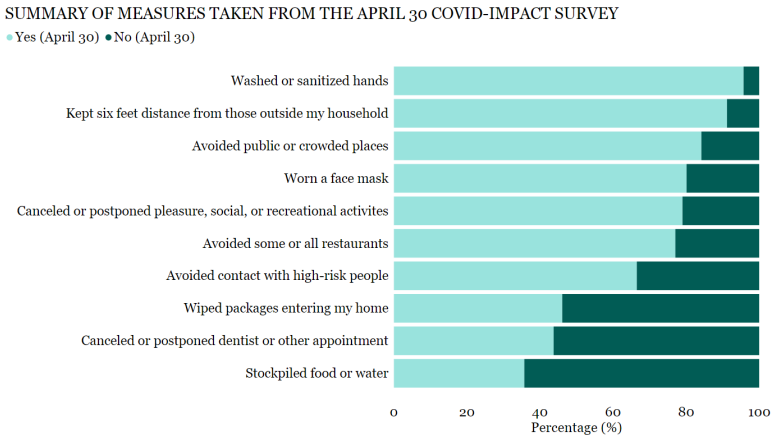

Figure 2: Summary of measures taken in response to the COVID-19 pandemic in 18 regions of the U.S. based on the results of a COVID-impact survey administered on April 30^{th}, 2020. The chart above shows the distribution of yes/no answers to whether each measure was taken by survey participants..

was not to demonstrate which factors affect those who *do* take measures against COVID-19 (i.e. answer 'Yes' to the questions displayed in Table 1), but more importantly, to unfold the factors associated with the participants who answer 'No' to the aforementioned questions, as understanding these populations is critical for ensuring we properly tackle this pandemic.

2.2 Data Preparation

To prime the data for analysis, any columns with over 20% missing values were dropped. Also, there were a few initial features/columns for region weights, these are not of use for our analysis and so these columns were dropped. Columns that contained no demographic information were also dropped, as these did not pertain to our research question. This left us with 55 features before dummy-encoding. For pre-processing, different types of data (numeric, Likert scale, binary or categorical) were dealt with separately.

Note that to deal with missing values, two methods were used. Either a NULL value was inserted to missing data cells (these cells were later imputed according to data type, as is explained below) or a new 'NaN' class was created, effectively dedicating a feature to missing values through dummy encoding. The method for pre-processing each data type is as follows:

- Numeric data fields where either the data was missing or the survey participant refused to answer the question were replaced with a NULL value.
- Ordinal data was processed so that each element on the scale was interpreted as an integer. All cases where there was no answer (either missing value or a survey participant refused to answer the question) were replaced with a NULL value.
- Binary data (questions with only two answer options) were converted to integers where the first answer choice coincided with a 0 (such as 'No') and the second answer choice coincided with a 1 (such as 'Yes'). Dummy encoding was used on any records with missing values, effectively creating three categories ('Yes', 'No', 'NaN').

- Categorical data features were dummy encoded, any missing values were replaced with the string 'NaN' to create a category for missing values.

To complete the data training and testing, the following algorithm was followed: First, under-sampling was done so the two target classes (1 and 0) for every target feature were present in equal amounts in both the training and test data. This was done to ensure our baseline accuracy was always exactly 50%. A 80-20 train-test split then applied to the data. The training data and test data were then imputed and normalized separately to avoid data information leakage. Numeric and Likert scale data were imputed with the feature median wherever a NULL value was found. No imputation was done on the binary and categorical data. All of the data pre-processing was done in python using *NumPy* and *Pandas* [13, 14].

2.3 Random Forest & Association Rules

Once the data was prepared, we used the *sklearn* implementation of random forest classification (RFC) to determine the three most influential factors for each of our targets. Before training the final RFC, we applied random-forest feature selection on the training data to remove all features with a below-average importance coefficient. A five-fold cross-validation approach was used to ensure our model was not over-fitting to the training data before reporting our classification accuracies for each of the measures in Table 1. The results are displayed in the Results section (section 3) in Table 2. Next, we used the important features that appeared for our most successful models (prediction accuracy of ¿ 0.6) to narrow our scope, then used Microsoft's Visual Studio Business Intelligence tool for association rule analysis on those features. The resulting association rules discovered are displayed in the Results section (section 3) in Table 3. For details on how to read and interpret association rules, refer to section 2.4.

2.4 Reading Association Rules

Recall that association rules are relationships that describe how a set of items is related with another set of items. There are two main building blocks in association rules, the *rule antecedent* and a *rule consequent*. The rule antecedent is the groups of items or factors that appear before the $\Rightarrow$ symbol while the rule consequent is the group of items that appear after the $\Rightarrow$ symbol. Therefore, to read the rule, we say that when all the items in the rule antecedent appear together, there is 'X' probability that the rule consequent will also appear in the data-set.
(EDUCATION = high school diploma or equivalent), (HOUSEHOLD INCOME = over $150,000) $\Rightarrow$ (worn a face mask = No)
probability = 0.56

For example, for the real association rule above, we can conclude that there is a probability of 0.56 that individuals with an income of $150,000 or higher and only a high school level education answer 'No' to wearing a face mask, compared to the average person with a 0.19 probability of answering 'No' to the same question.

3 Results

Looking at Table 2 (which shows the results of our feature selection using random forest), we see the same three factors appear multiple times. For all 10 measures taken in response to COVID-19, the top three influential factors are, in varying orders, household size (HHSIZE), health-state (HLTH, self-reported health as one of excellent, very good, good, fair or poor) and

household adults (HH18, the total number of people age ≥ 18 in a household). These results tell us that the health, family size and number of adults are all factors related to curve-flattening behaviors or not. We used these results to further narrow down our analysis and discovered a series of interesting relationships. Note that in the next step of association analysis, our goal was to focus on the associations between these factors that are related to *not* taking measures against COVID-19.

We have reported six association rules in Table 3 which all include at least one of the most influential factors determined via random forest. The probability is the chance that the association rule holds, the Prop. Measure is the proportion that an average individual will answer 'No' to the relevant measure. For example, while only a proportion of 0.199 survey participants answered 'No' to wearing masks, people with large households and an income between \$75,000 and \$100,000 will answer 'No' to wearing a mask with a probability of 0.485, which is about 2.5 times larger. Notice that all the probabilities in Table 3 are over 2 times larger than their relative Prop. Measure.

4 Discussion

For the top three influential factors seen in Table 2, we discovered six interesting associations, which are displayed in Table 3. For the first four rules, there is a relation with household income. The first rule shows an association between a household size of six or more persons and an income between \$75,000 and \$100,000 with not wearing a mask, while the second rule shows an association between a household size of five persons and an income ranging from \$40,000 to \$50,000 with not wearing a mask. As well, the third rule shows an association between a household size of six or more persons and an income under \$10,000 with not practicing social distancing. The USDA (U.S. Department of Agriculture) has estimated that it costs approximately \$12,980 to raise a single child annually. Multiplying this by 5 children results in a cost of approximately \$65,000. This does not include mortgages, bills, a college education or the parents' living expenses [15]. Therefore we can conclude that the incomes of these three mentioned rules are relatively low compared to the family size. This leads us to believe that there is a relationship between large families who may be struggling financially and wearing a mask and practicing social distancing.

Next we analyzed another one of the three most influential factors, health status. Recall that this is self-reported health as one of excellent, very good, good, fair or poor. The last three rows in Table 3 involve health status. The associations found here are quite surprising, as we discovered relationships between low income and both health extremes (excellent health and poor health). While one rule associated *poor* health and an income under \$10,000 with choosing to not cancel or postpone pleasure, social or recreational activities, the other rule associated *excellent* health and an income under \$10,000 with choosing to not cancel or postpone pleasure, social or recreational activities. The common factor between these rules was having a low income, where such a low income lead to an apathy towards and/or lack of ability to follow safety measures during the pandemic.

These association rules illustrated a common theme with regards to lower income and not following public health guidelines. However, upon further analysis of this relationship with income, we found that people with low income and large families — who were associated with not wearing a face mask — reported that they were washing and sanitizing their hands more frequently in response to COVID-19. This leads us to speculate that the reason for not following the safety measures specified in Table 3 is perhaps not due to apathy, but rather, the fact that purchasing these materials might only contribute to the household's already-present financial

Measure Taken in Response to COVID-19	P(yes)[1]	Accuracy[2]	Factor 1	Factor 2	Factor 3
Worn a face mask	0.801	0.613	HHSIZE[3]	HLTH[4]	HH18[5]
Cancelled or postponed dentist or other appointments	0.437	0.583	HLTH	HHSIZE	HH18
Avoided some or all restaurants	0.771	0.557	HHSIZE	HLTH	HH18
Cancelled or postponed pleasure, social, or recreational activities	0.790	0.615	HHSIZE	HLTH	HH18
Stockpiled food or water	0.357	0.583	HLTH	HHSIZE	HH18
Avoided public or crowded places	0.842	0.578	HLTH	HHSIZE	HH18
Avoided contact with high-risk people	0.665	0.546	HLTH	HHSIZE	HH18
Washed or sanitized hands	0.957	0.586	HHSIZE	HH18	HLTH
Kept 6 ft distance from those outside household	0.921	0.630	HHSIZE	HLTH	HH18
Wiped packages entering home	0.461	0.577	HLTH	HHSIZE	HH18

[1] P(yes) refers to the proportion of individuals who reported an answer of 'Yes' for each measure taken.
[2] Equal to the number of correct predictions by the RFC over the total number of predictions.
[3] HHSIZE refers to the household size, or number of people living in one house.
[4] HLTH is a self-reported health as one of excellent, very good, good, fair or poor.
[5] HH18 is the total number of people age $\geq$ 18 in a household.

Table 2: Feature Selection results for factors influencing measures taken in response to COVID-19. Results were obtained via random forest classification. Note that under-sampling was done to ensure the baseline accuracy was 50%.

strain. This is an intriguing (but not surprising) discovery and more research would need to be done to reach a firm conclusion on the matter.

The last rule involved both the health status and the number of persons above the age of 18 in a household. We find that households with 4-7 adults, who are all in relatively good health, are associated with not cancelling or postponing pleasure, social or recreational activities. It is not clear the average age of these adults from the survey data, but in the absence of this information, a potential explanation may lie in college-aged students (potentially living with other college-aged students) ignoring stay-at-home orders for recreational/social gatherings [16].

Association Rule	Prob.[1]	Prop. Measure[2]
(HOUSEHOLD SIZE[3]= six or more persons), (HOUSEHOLD INCOME [4]= $75,000 to under $100,000) ⇒ (worn a face mask = No)	0.485	0.199
(HOUSEHOLD SIZE = five persons), (HOUSEHOLD INCOME = $40,000 to under $50,000) ⇒ (worn mask = No)	0.407	0.199
(HOUSEHOLD SIZE = six or more persons), (HOUSEHOLD INCOME = under $10,000) ⇒ (cancelled or postponed pleasure, social, or recreational activities = No)	0.481	0.210
(HEALTH[5]= Excellent), (HOUSEHOLD INCOME = under $10,000) ⇒ (cancelled or postponed pleasure, social, or recreational activities = No)	0.520	0.210
(HEALTH = Poor), (HOUSEHOLD INCOME = under $10,000) ⇒ (cancelled or postponed pleasure, social, or recreational activities = No)	0.500	0.210
(HEALTH = Good), (HOUSEHOLD $\geq$ 18 [6]= 4.23-7.18) ⇒ (cancelled or postponed pleasure, social, or recreational activities = No)	0.441	0.210

[1] The probability that a participant that fits the antecedent answers 'No' to the consequent.
[2] The probability that an average survey participant answers 'No' to the measure in the rule consequent.
[3] HOUSEHOLD SIZE is the number of individuals living in a household.
[4] HOUSEHOLD INCOME is the total income made in a household.
[5] HEALTH is a self-reported health as one of excellent, very good, good, fair or poor.
[6] HOUSE OLD $\geq$ 18 is the total number of people age $\geq$ 18 in a household.

Table 3: Association rules for demographics and groups of factors associated with answering 'No' to taking increased safety measures in response to COVID-19.

4.1 Limitations & Future Work

In terms of limitations, we acknowledge that the accuracy values seen in Table 2 are overall lower than normally desirable. Fortunately, our final goal was not to predict if an individual person would likely break social distancing protocols, rather, we wanted an idea about the direction in which to take the association analysis. Some steps for future research may include tuning the RFC model to find parameters that lead to a more desirable result as this may lead to new feature importance discoveries. In terms of future work, there are numerous directions to take this project in the future such as parameter tuning, comparing differences between the April and May surveys or potentially combining the two surveys into a "master" data set with twice the survey responses.

5 Conclusion

It is evident that the COVID-19 pandemic has impacted many dimensions of daily life. As such, it is critical that we understand the complexities of personal health decisions made by individuals, as this can aid governments and health care systems in the process of informed

policy-making. We have analyzed a COVID-impact survey to highlight socioeconomic and demographic factors that are influential for determining whether someone will help flatten the curve by taking various safety measures such as washing hands more frequently, wearing a face mask, keeping a 6-foot distance from others etc. Our results connect household size, number of adults per household and health status with measures taken in response to COVID-19. We have also come across support that potentially vulnerable lower income households are less likely to follow measures that involve purchasing materials such as face masks. As Well, we found that over four healthy adults living in one household is associated with choosing not to cancel recreational activities during the pandemic. This research is critical for addressing the populations of individuals who may not be following health guidelines, as well as understanding how different demographics perceive the pandemic.

References

[1] Juan Wang and Guoqiang Du. Covid-19 may transmit through aerosol. *Irish Journal of Medical Science (1971 -)*, 03 2020.

[2] Zi-yu Ge, Lu-ming Yang, Jia-jia Xia, Xiao-hui Fu, and Yan-zhen Zhang. Possible aerosol transmission of covid-19 and special precautions in dentistry. *Journal of Zhejiang University-SCIENCE B*, 21, 03 2020.

[3] Alan D. Workman, D. Bradley Welling, Bob S. Carter, William T. Curry, Eric H. Holbrook, Stacey T. Gray, George A. Scangas, and Benjamin S. Bleier. Endonasal instrumentation and aerosolization risk in the era of covid-19: simulation, literature review, and proposed mitigation strategies. *International Forum of Allergy & Rhinology*, n/a(n/a), 2020.

[4] Covid-19 map.

[5] Covid-19 united states cases by county.

[6] 2020 Published: May 29. State data and policy actions to address coronavirus - notes and sources, May 2020.

[7] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *Forest*, 23, 11 2001.

[8] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Databases*, pages 487–499. ACM Press, 1994.

[9] Necibe Tuncer, Chindu Mohanakumar, Samuel Swanson, and Maia Martcheva. Efficacy of control measures in the control of ebola, liberia 2014–2015. *Journal of biological dynamics*, 12(1):913–937, 2018.

[10] Lily L Tsai, Benjamin S Morse, and Robert A Blair. Building credibility and cooperation in low-trust settings: persuasion and source accountability in liberia during the 2014–2015 ebola crisis. *Comparative Political Studies*, page 0010414019897698, 2020.

[11] Robert A Blair, Benjamin S Morse, and Lily L Tsai. Public health and public trust: Survey evidence from the ebola virus disease epidemic in liberia. *Social Science & Medicine*, 172:89–97, 2017.

[12] covid-impact.

[13] Ecosystem.

[14] pandas.

[15] Mark Lino, Rixt Luikenaar, Mary Anne Sherman, Thomas Desmond, James Augustine, Marie, Elise, Mike Erekson, Haley, Kate, and et al. The cost of raising a child, Feb 2020.

[16] Timothy Bella. 'if i get corona, i get corona': Miami spring breakers say covid-19 hasn't stopped them from partying, Mar 2020.