# Predicting Incidence of Increasing Mental Health Stress in Relationship to the COVID-19 Pandemic

Georgia Institute of Technology
CSE6242
December $3^{rd}$, 2021

Andrew Piscioneri {*apiscioneri3*}
Kevin Ayers {*kayers8*}
Eric Nagel {*enagel6*}
Jerrod Pelley {*jpelley3*}
Jeffrey Zhang {*jeffhz16*}
Vipul Koti {*vkoti7*}

**Note: All team members have contributed similar amount of effort.**

## I. Introduction

The impact of the COVID-19 pandemic on the United States and the world is still being measured, and understanding the lasting impacts of COVID-19 will be difficult. One of the largest impacts thus far is to members of the workforce. Each industry has been impacted in different ways, with employers and employees seeing the transition differently. Additionally, mental health has become an important focus of the healthcare community, with the COVID-19 pandemic highlighting the inequities and lack of mental health resources.

While data and reporting has been capturing sentiment, raw employment, and health data for months during the pandemic, this information is not generally available to the public in a consumable format, either requiring usage of government database APIs or requiring subscriptions. This presents a challenge and opportunity, in that those who are dis-proportionally affected by COVID-19 have the least access and visibility into data that can improve their livelihood. The following project report and available visualizations are focused on providing an easy to understand format on existing data that removes barriers to information and promotes democratization of data regarding the COVID-19 pandemic. Additionally, we highlight how remote work and the pandemic has impacted mental health and detail how mental health can be predicted using machine learning models, in hopes to promote preventative distribution of health resources.

## II. Problem Definition

We strove to provide an in-depth analysis on how remote work and the COVID-19 pandemic has affected mental health in the United States. Specifically, we looked at how COVID-19 infections, vaccination rates, and lockdowns have impacted remote work and mental health. Utilizing this information in conjunction with US Census Bureau (USCB) data, we were able to understand how the certain characteristics impact remote work and mental health, and how the two are related. Creating a simple and interactive visualization of this data, we attempted to predict mental stress at the individual and aggregate (state) level. This information is valuable to individuals who lack the means to access this data in addition to allowing entities to utilize this information and future iterations of this work to predict the distribution of mental health resources.

## III. Survey

The world changed after China reported suspicious cases of pneumonia to the local World Health Organization (WHO) facility on December 31, 2019 [10][11]. Roughly 3 weeks later, the first reported case of COVID-19 was documented on United States (US) soil in Snohomish county Washington on January 20, 2020 [7][11]. COVID-19 has since cascaded across the US and the rest of the world, with the WHO announcing a global pandemic on March 11, 2020 [11].

The US Department of Health and Human Services, among other organizations, described human to human transmission [8][10], which led government officials across the country to initiate mandated stay-at-home orders [9]. These stay-at-home orders effectively limited population movement across much of the US [9]. In this initial stage of the pandemic, remote workers were three times less likely to lose their job compared to those without the ability to work remotely [1]. Additionally, COVID-19 has been found to increase labor market inequality and that industries have seen different impacts in terms of remote work due to COVID-19 [2][4][5][17]. This inequality has brought a significant risk to workers, as COVID-19 related job displacement is typically unrelated to work performance [6]. These challenges have only increased existing trends in the landscape of 'where' work exists, by slowing metro-growth in already hard-hit areas, like the Midwest, where cities registered their lowest population gains in a decade [13][16]. Even further, COVID-19 has depressed the rate at which Americans are moving, meaning cities needing employees aren't able to find them [14]. These challenges make it more critical than ever to understand what employment data is telling us, so resources can be made available to meet the changing scope of work post-COVID.

Among the many activities that became restricted to the home, remote work was one that many people had to adapt to. Now that we are assimilating into a "new normal", remote work is here to stay; both a majority of hiring managers and employees have looked to increase remote work capabilities due to better than expected transitions [3][12][15]. These results have been overwhelming, in so far as 97% of employees both recommend to others the benefits of remote work, as well as saying they want to continue to work remotely [15].

We will go into more background about how the COVID-19 pandemic started and how this brought about the environment to promote remote work and population migration in this report, using widely available US Government data that has been captured during the pandemic [23-28], as well as proven methods for running data science related projects [18]. Such as, the use of visualization and modeling tools like Plotly, Dash, and XGBoost[29] to understand and predict how these trends will continue in the future.

## IV. Proposed Method

### A. *Intuition*

As touched on in the *Introduction*, data and reporting during the COVID-19 pandemic **is not generally available to the public in a consumable format, either requiring usage of government database APIs or requiring subscriptions**. Additionally, the COVID-19 pandemic has increased mental health disease and has brought to light the inequities and lack of access of healthcare.

The innovation that we aim to provide is a **simple, interactive Plotly/Dash hosted web page that combines Center for Disease Control (CDC) COVID-19 count, vaccination, and lockdown data with multiple US Census Bureau (USCB) datasets on remote work and mental health characteristics in order for people to understand and analyze the impact of COVID-19 on the individual at a state level**.

By analyzing the waxing and waning trends in COVID-19 data, and how that correlates to citizens forced into remote work and mental health disease, **predictions about future mental stress can be made at the individual or aggregate (state) level**. This can include, but is not limited to which locations might see more mental stress in the future, which states may see an aggregate increase in mental stress, and how likely a person is to be at high risk of mental stress. All of which have implications on the individual and aggregate level. This information can be used to assess the current state of mental health, which factors have the biggest impact on mental health, and how to allocate the constrained mental health resources accordingly.

The interactive visualizations will be made publically available for those interested in taking a deeper look into the data, while prediction results are made available in the Experiments and Evaluations section below.

### B. *Approaches*

*1) Data Collection and Storage:* Utilizing government data, we've established a master dataset for modeling (Appendix F). The master dataset aggregates data around remote work capability, sentiment, COVID-19 numbers and lockdown procedures from two USCB surveys [25,26,27] and two CDC surveys [23,24]. The following datasets, The USCB Current Population Survey (CPS), USCB Housing Pulse Survey (HPS), CDC COVID-19 Case Surveillance Public use Data, CDC COVID-19 Lockdown Data, and CDC COVID-19 Vaccination data (Included in HPS), can be found in Appendices D, E, A, and B, respectively.

These data-sets are joined together based on the month, year and FIPS (geographical code) to form a data-set of over 3M rows. Our modeling (described below) was predicted against this data-set to answer the problem statement of how remote work sentiment and remote work availability are impacting the actual practice of working remotely, under the guise of the COVID-19 pandemic. This is an established time series of data from early 2020 to 2021, with an additional geographical component at the state and county level using a FIPS code to allow for geographical visualization.

The Correlation map shown in Figure IV.I identifies inter-correlation between features. Of note, the industry of work sees strong negative correlation (to be expected), and the remote percentage of work in the respondent's location has a negative correlation to vaccination rates, but interestingly not to the amount of COVID-19 cases. This figure is available on the website as an interactive tool for further analysis and review.

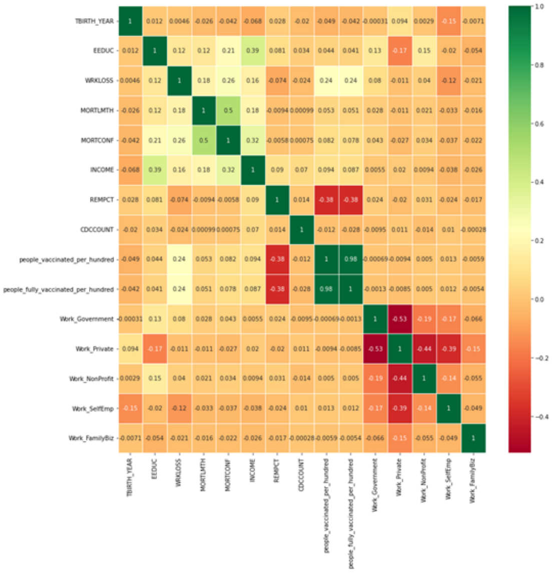


Figure IV.I: Correlation Heat Map of Features

As this is survey data, we needed to appropriately clean and update data. The majority of these efforts were surrounding the removal of non-answered survey questions. Specifically, respondents could either ignore a question (see it, but not answer it) or skip it completely (it was never seen). In both these cases, for all questions in our data-set, we removed those records such that they were *not* included in modeling.

Additionally, for vaccination data, since our full dataset begins with data collected in January 2020, we've replaced empty vaccination values for locales with 0. Vaccines were not available until January, so to accurately use these scalar features, they were modified to remove sparsity.

Features were also modified to create categorical variables as identified below:

Of note:

- **Mortgage Paid Last Month (MORTLMTH)**: Modified from a scalar variable to categorical, either (0) not paid or (1) paid.
- **Kind of Work (KINDWORK)**: Modified from a scalar variable to a categorical variable, pertaining to the industry of the employer.
- **States (STATES_C)**: The US State from which the data was collected was modified to create categorical variables but was removed due to low predicting power.
- **Month (YEARMONTH)**: The US State from which the data was collected was modified to create categorical variables but was removed due to low predicting power.

Finally, the data was split into two forms: aggregated data based on the mean values, grouped by Year/Month/FIPS, as well as individual data that was not grouped or aggregated. The original, raw data that was consumed was over five million records. The cleaned, non-sparse individual dataset comprises of 960,721 rows of data and the aggregated dataset is based off this data with 918 total rows based on  18 months of data for 50+ states and territories.

This data is being stored within the Google Cloud Platform using a publicly accessible storage bucket. Due to the size of the data files, and cooperative nature of the Google Cloud Platform with Python, we felt that this was the best resource for us to use for this project. All code artifacts are stored within a private Github repository that is shared between the team members.

*2) Algorithms:* Our algorithm approach is based on review and analysis. We have highlighted approaches such as XGBoost [29], which we believe is uniquely suited to our problem due to it's ensemble nature; it is computationally efficient and naturally resists multicollinearity as it is a decision tree based algorithm. However, our approach is to compare this to many machine learning models to understand how well it can predict.

Our development makes extensive use of existing Python libraries designed to promote ease-of-use. Please see Appendix G for a complete list of packages used.

For our Ensemble and NN learners (XGBoost, Random Forest, MLP), pre-processing of the data was completed within the steps described in the previous section *Data Collection and Storage*, as they are not impacted by multi-collinearity. For Regression analysis, principal component analysis (PCA) with two components was applied to the training and testing data-sets for dimensionality and multi-collinearity reduction purposes.

Our final models are also split into two sections: an Individual model designed to predict whether or not a person is at a high risk for mental stress, as well as an aggregated quantitative model designed to calculate the Mental Stress Index (MSI) sum (detailed in *Design of Experiments*) for use in understanding how, at a state level, mental stress is impacting the population. This output will be visualized on a map of the US to help resources focus on areas of high concern for mental health, as related to COVID-19.

*3) Interaction and Visualization:* For interaction, our final visualizations, comments and notes will be hosted on a public website that can be accessed for the public to interact with, as well as by embedding Dash applications for an interactive, data-driven dashboard that website visitors can review.

We plan on utilizing the following visualizations:

1) Maps within Plotly that define the COVID timeline, prediction results and data
2) Timeline Charts for trend analysis on month-over-month analysis
3) Density maps for patterns, heatmap-style analysis for correlation and other geographical patterns
4) A prediction form to predict a level of stress based on how you are feeling.

The final visualizations are available on the website, https://cse6242-delta6-project.com, but an example of the visualizations available can be seen in Appendix H.

*4) Website Design and Hosting:* Our interactive layer is a website that is stored and hosted via the Google Cloud Platform (GCP). An architecture diagram is available below in Appendix I. This website contains the hosted Dash apps for visualizations and is designed as a Single-Page Application (SPA) for ease-of-use.

To allow for public access, the website is hosted using Google Storage Buckets, which are connected to a Load Balancer to support web traffic. This load balancer has been set to a static IP that has been assigned by GCP, and the appropriate domain and domain name system (DNS) entries have all been aligned via Google services.

## V. Design of Experiments and Evaluation

### A. Description of Testbed

**This list addresses the questions that we designed our experiments to answer:**

1) Has remote work, brought on by the COVID-19 pandemic impacted the mental health of individuals in the US?
2) Can we with reasonable accuracy predict the incidence of mental health risks with a machine learning algorithm?
3) Can we predict the quantitative mental health incidence level throughout the United States?

Our hypothesis for this project is that the change to a remote capacity for employees would increase the stress level, not only in a direct way but also indirectly due to the COVID-19 pandemic causing remote work. The questions above can be appropriately answered from our predicted model. After initial analysis, we decided to take two approaches to test the accuracy of our hypothesis, as detailed below.

### B. Details of Experiments

*1) Model Selection and Analysis:* Our first approach to try and answer the questions and hypothesis above were to create a classification model based on the aggregate score of the three primary variables in our data-set, *Anxiety*, *Depression* and *Worry*. If the sum of these three values was greater than or equal to seven (indicating the individual believed they were at least a high risk in one of the specific mental health categories), the supervised data would consider this response a *high risk individual*. For a score of six or below, meaning the individual *did not* respond with at least a high risk in any of these categories, they would be considered low-risk. This output moving forward will be referred to as the *Mental Stress Index* (MSI).

Our second approach was to use our aggregated state data-set to predict the quantitative sum of these three values. Based on the initial analysis detailed in the *Approaches* section of the report, we found that the predictive importance of COVID-19 related data in general increased as the aggregation of the data increased. This is to say, what impacts an individual does not impact a community, or in our case, a state.

To measure success of these models, we looked at the predictive power, accuracy scores, and whether or not they reflect the ability to answer the listed questions.

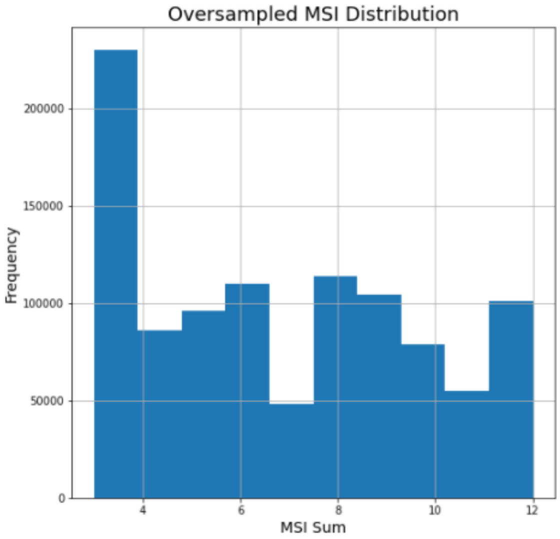


Figure V.II: Oversampled Data Distribution

*2) Individual MSI Classifier Model:* For the individual classifier, we used a variety of models to review the overall score and error. This data was predicted using our cleaned dataset, but based on the supervised results, we over-sampled the high risk individuals to provide a more normal distribution to use in modelling. The original data distribution can be found in Appendix J. Shown below is the over-sampled dataset. While still skewed, it presents a much cleaner distribution:

When reviewing the accuracy scores of the models, they perform well against the data when classifying individuals as high risk. We chose the XGBoost model because it has the highest accuracy with the lowest error metrics and also is computationally less intensive compared to Random Forest, which placed in close second.

| Individual MSI Results | | | |
|---|---|---|---|
| Model | MAE | RMSE | Acc. |
| **XGBoost** | **.20777** | **.45582** | **79.22%** |
| Random Forest | .20782 | **.45582** | **79.22%** |
| Naïve Bayes | .21237 | .46084 | 78.76% |
| MLP NN | .20967 | .45790 | 79.03% |
| Log. Regression | .20967 | .45790 | 79.03% |
| KNN | .20960 | .45790 | 79.04% |

Table V.I: Individual Model Predictive Accuracy

The top five most important features by F-score for the final XGBoost model are shown in Table V.II below. When reviewing our questions and hypothesis, the relative predictive strength of the COVID variables is low. In fact, only CDC Case Count made the important predictive features list, but its predictive strength is lowest of the five features selected.

| XGBoost Feature Importance | |
|---|---|
| Feature | F-Score |
| Mortgage Confidence | 180022.0 |
| Birth Year | 48015.5 |
| Work Loss | 21898.0 |
| Income Level | 14583.4 |
| CDC Case Count | 7543.4 |

Table V.II: F-Score of Important Features

Based on our hypothesis, the results are mixed. Our background literary sources suggest that COVID-19 caused loss of employment when it could not be shifted to remote, which is one of the top three features in terms of predictive importance in our model (Work Loss). However, the confidence in paying your mortgage only weakly correlated with work loss, which is by far the most important feature of the model. This would suggest that any disruption, COVID-19 related or not, would impact an individual's MSI risk level.

*3) State Aggregate MSI Sum Model:* In addition to our individual model, we also ran a set of models at an aggregated, state level. This data was was grouped by the year/month it was recorded, and then by state. This provided aggregated state level data to perform quantitative analysis on the MSI sum score. The models were tested as found in table V.III below.

| Aggregated MSI Results | | | |
|---|---|---|---|
| Model | MAE | RMSE | Acc. |
| XGBoost | .24457 | .08756 | 30.50% |
| MLP NN | .21419 | .07209 | 42.78% |
| **PCA Lin. Reg.** | **.18180** | **.05165** | **59.00%** |

Table V.I: Aggregated Model Predictive Accuracy

What's interesting is that comparing the Individual XGBoost vs. Linear Regression modeling in aggregate, we can see that vaccination information available is weighted much heavier in terms of prediction capability. In fact, while the individual metrics are still present, vaccination data is the most important prediction feature available for the model, as seen in Table V.III.

| Aggregate Linear Regression Information Gain | |
|---|---|
| Feature | MI Score |
| People Fully Vaccinated p. 100 | 0.4275 |
| People Vaccinated p. 100 | .4107 |
| Work Loss | .3963 |
| Mortgage Confidence | .2134 |
| Mortgage Paid Last Month | .1409 |

Table V.III: MI of Important Features (Aggregate)

Originally, without PCA, the $R^2$ = 0.69, however multi-collinearity was present. Upon introducing PCA analysis with two eigenvectors, the $R^2$ value dropped to 0.59, but multi-collinearity was reduced while still preserving a low amount of error.

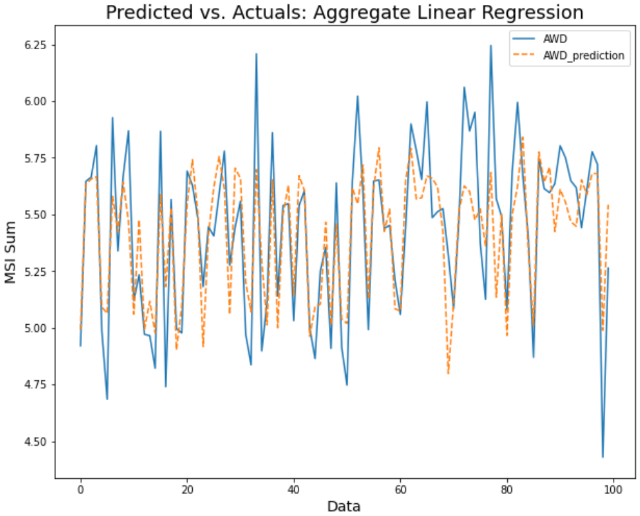


Figure V.III: Aggregated Predicted vs. Actuals

## VI. Discussion

Reviewing the final models and original questions, we feel confident that the original three questions can be reviewed. Firstly, to the question of whether or not remote work has impacted mental health of individuals, the answer is mixed. While at an individual level there are clear parallels to be drawn from COVID-19 causing those without the ability to work remotely to lose their jobs (as seen in the literary background information), it's certainly not the only reason for one's Mortgage Confidence level to be low. It is likely that while the pandemic impacted the financial situation for many, the evidence does not support a causal relationship between Remote Work and Mental Stress. Additionally, at an aggregate level, the percentage of remote work in a person's location had little explanatory power for the MSI index. Therefore, we would reject the our hypothesis; that is to say remote work does not conclusively impact mental stress.

For the second and third questions regarding prediction of mental health risks of individuals based on COVID-19 information, as well as quantitative aggregation of mental health stress levels throughout the United States, the answer is much more promising. Using any of the classifying models, the prediction accuracy of determining if an individual is at high-risk of mental stress was impressive at 79.22%. While the quantitative model only scored 59.0% accuracy, it too is likely something that can be built upon or used as an initial guide for federal agencies and organizations to see which states have a higher overall MSI aggregate value.

As we enter our second winter of the pandemic and into our third year, it could be highly advantageous to be able to distribute mental health resources where they are needed since the United States has a shortage of mental health professionals. Based on the results of the models, we believe they can be of assistance in this regard.

There are a number of limitations for our project. We attempted to predict the status of mental stress (anxiety, worry, and depression) using CDC COVID-19 data and US Census Bureau (USCB) data. Not only were we limited to the people who took the surveys and the accuracy of the government agencies who collected the data, but mental health diseases are complex in nature and causal relationships cannot be formed without extensive research. Mental health contains a genetic component, and those with history of mental health issues are more likely to have mental health issues in the future. Finally, over the past two years, we have had many other polarizing issues such as the political field and race riots in the United States. These among many other public issues could negatively affect ones mental health and would go unseen in our current data-sets, or could bias the data.

The scope of this project focused to predicting mental stress using the CDC and USCB datasets. With additional time, we could have increased the scope of our datasets to include any data that could be connected via county or state codes. Future work would look at adding additional data that could possibly add to the predictive accuracy of our current algorithm.

## VII. Conclusion

We aimed to highlight the impact of COVID-19 on the United States workforce. In conclusion. While there is evidence to suggest that COVID-19 has indirectly impacted mental health through remote work, there isn't enough statistical data to suggest a direct connection. However, it is possible to use this information to understand how resources can be used to prevent mental stress; not only from a geographical level in terms of regional resourcing, but also in terms of what services should be provided and to whom. Additionally, the predictive models highlighted by this research indicate there is a possibility of identifying mental stress of individuals and in aggregate. We encourage the use and additional research in this area to promote well-being and to reduce the chance, stigma and impacts of mental health inequality within the United States.

# REFERENCES

[1] Angelucci, M., Angrisani, M., Bennett, D. M., Kapteyn, A., Schaner, S. G. (2020). Remote Work and the Heterogeneous Impact of COVID-19 on Employment and Health (No. 27749). https://www.nber.org/papers/w27749

[2] Brynjolfsson, E., Horton, J. J., Ozimek, A., Rock, D., Sharma, G., Tuye, H.-Y., & Upwork, A. O. (2020). COVID-19 and Remote Work: An Early Look at US Data (No. 27344). https://www.nber.org/papers/w27344

[3] Ozimek, A. (2020). The Future of Remote Work. https://ssrn.com/abstract=3638597

[4] Althoff, L., Eckert, F., Ganapati, S., Walsh, C. (2020). The City Paradox: Skilled Services and Remote Work. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3744597

[5] Beland, L.-P., Brodeur, A., Wright, T. (2020). The Short-Term Economic Consequences of COVID-19: Exposure to Disease, Remote Work and Government Response. In IZA Discussion Paper (No. 13159; Issue No. 13159). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584922

[6] Tronco Hernandez, Y. A. (2020). Remote Workers During the COVID-19 Lockdown. What Are We Missing and Why Is Important. Journal of Occupational and Environmental Medicine, 62(11), e669–e672. https://doi.org/10.1097/JOM.0000000000002018

[7] Holshue, M. L., DeBolt, C., et al. (2020). First Case of 2019 Novel Coronavirus in the United States. New England Journal of Medicine, 382(10), 929–936. https://doi.org/10.1056/nejmoa2001191

[8] Ghinai, I., Woods, S., et al. (2020). Morbidity and Mortality Weekly Report Community Transmission of SARS-CoV-2 at Two Family Gatherings-Chicago, Illinois, February-March 2020. US Department of Health and Human Services/CDC, 69(15). https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e1.htm

[9] Moreland, A., Herlihy, C., et al.; (n.d.). Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement — United States, March 1–May 31, 2020. Retrieved October 10, 2021, from https://www.cdc.gov/mmwr/volumes/69/wr/mm6935a2.htm

[10] Carvalho, T., Krammer, F., & Iwasaki, A. (2021). The first 12 months of COVID-19: a timeline of immunological insights. In Nature Reviews Immunology (Vol. 21, Issue 4, pp. 245–256). Nature Research. https://doi.org/10.1038/s41577-021-00522-1

[11] Sullivan, K., & Laube, J. (2021, February 19). A Brief History of COVID, 1 Year In. Everyday Health. https://www.everydayhealth.com/coronavirus/a-brief-history-of-covid-one-year-in/

[12] Parker, K., Horowitz, J., Minkin, R., & Arditi, T. (2021). How the Coronavirus Outbreak Has-and Hasn't-Changed the Way Americans Work FOR MEDIA OR OTHER INQUIRIES. https://www.pewresearch.org/social-trends/2020/12/09/how-the-coronavirus-outbreak-has-and-hasnt-changed-the-way-americans-work/

[13] Frey, W. H. (2021). Pandemic population change across metro America: Accelerated migration, less immigration, fewer births and more deaths. https://www.brookings.edu/research/pandemic-population-change-across-metro-america-accelerated-migration-less-immigration-fewer-births-and-more-deaths/

[14] Patino, M. (2020). What We Actually Know About How Americans Are Moving During Covid. https://www.bloomberg.com/news/articles/2020-09-16/the-truth-about-american-migration-during-covid

[15] Buffer. (2021). The 2021 State of Remote Work. https://buffer.com/2021-state-of-remote-work

[16] Maddow-Zimet, I., Kost, K., & Finn, S. (2020, October). Pregnancies, Births, and Abortions in the United States, 1973-2016: National and State Trends by Age. Guttmacher Institute. https://www.guttmacher.org/report/pregnancies-births-abortions-in-united-states-1973-2016

[17] Dey, M., Frazis, H., Loewenstein, M. A., & Sun, H. (2020). Ability to work from home: evidence from two surveys and implications for the labor market in the COVID-19 pandemic. Monthly Labor Review US Bureau of Labor Statistics. https://doi.org/10.21916/mlr.2020.14

[18] Schermann, M. (2019). How to Run a Data Visualization Project. In A Reader on Data Visualization. Bookdown. https://mschermann.github.io/data_viz_reader/how-to-run-a-data-visualization-project.html

[19] Lee, B. (2021). COVID-19 Case Surveillance Public Use Data with Geography. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

[20] NIH: National Eye Institute. (2020, February 7). Nearsightedness (Myopia) Tables. NIH: National Eye Institute. https://www.nei.nih.gov/learn-about-eye-health/outreach-campaigns-and-resources/eye-health-data-and-statistics/nearsightedness-myopia-data-and-statistics/nearsightedness-myopia-tables

[21] FreshRemote.work. (n.d.). FreshRemote.work Salaries. Retrieved October 10, 2021, from https://salaries.freshremote.work/download/

[22] Kang, Y., Gao, S., Liang, Y., Li, M., Rao, J., & Kruse, J. (2020). Multiscale dynamic human mobility flow data-set in the U.S. during the COVID-19 epidemic. Scientific Data, 7(1). https://doi.org/10.1038/s41597-020-00734-5

[23] CDC Surveillance Review and Response Group. (n.d.). COVID-19 Case Surveillance Public Use Data with Geography. Centers for Disease Control and Prevention. Retrieved October 30, 2021, from https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

[24] Centers for Disease Control and Prevention. (2021, September 10). U.S. State and Territorial Stay-At-Home Orders: March 15, 2020 - August 15, 2021 by County by Day. Data.Cdc.Gov. https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Stay-At-Home-Orders-Marc/y2iy-8irm

[25] United States Census Bureau. (n.d.). Basic Monthly Current Population Survey (CPS). Retrieved October 30, 2021, from https://www.census.gov/data/data-sets/time-series/demo/cps/cps-basic.html

[26] United States Census Bureau. (n.d.). COVID-19 Data from the CPS. Retrieved October 30, 2021, from https://www.census.gov/data/data-sets/time-series/demo/cps/cps-supp_cps-repwgt/cps-covid.html

[27] United States Census Bureau. (n.d.). Household Pulse Survey Public Use File. United States Census Bureau. Retrieved November 2, 2021, from https://www.census.gov/programs-surveys/household-pulse-survey/data-sets.html

[28] No Reference. Place Holder for Sourcing Purposes.

[29] Jason Brownlee. (2020, August 5). How to Use XGBoost for Time Series Forcasting. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-for-time-series-forecasting/

[30] Marina Chatterjee. (2020, August 16). Introduction to Spectral Clustering. Great Learning. https://www.mygreatlearning.com/blog/introduction-to-spectral-clustering/

[31] Aishwarya Singh. (2018, September 27). A Multivariate Time Series Guide to Forecasting and Modeling (with Python Codes). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

## APPENDIX A

CDC Monthly COVID-19 Case Surveillance Public Use Data with Geography [23]:

- **state_fips**: State FIPS Code (01-56 for the 50 States and 6 US Territories)
- **case_month**: The month the data was received by the CDC
- **res_state**: State of residence
- **res_county**: County of residence
- **county_fips**: County FIPS code
- **case_count**: Total number of aggregated cases per County for the associated month.

## APPENDIX B

CDC COVID-19 Lockdown Data [24]

- **State_Tribe_territory**: US State, tribe, and territory names
- **County_Name**: US county names
- **FIPS_State**: US state FIPS codes
- **FIPS_County**: US count FIPS codes
- **date**: Daily date in dataset
- **Order_code**: Numeric order status grouped by display categories used on the Environmental Public Health Tracking Network
- **Stay_at_Home_Order_Recommendation**: Stay-At-Home status

## APPENDIX C

Empty Appendix

## APPENDIX D

USCB Current Population Survey [25,26]:

- **HRYEAR4**: Year
- **HRMONTH**: Month
- **GESTFIPS**: State FIPS code
- **REMPCT**: Percentage of people working from home

## APPENDIX E

USCB Household Pulse Survey [27]:

- **WEEK**: Week of year
- **TBIRTH_YEAR**: Year of birth
- **EEDUC**: Educational attainment (1) Less than high school (2) Some high school (3) High school graduate or equivalent (4) Some college, but degree not received or in progress (5) Associate's Degree (6) Bachelor's Degree (7) Graduate Degree
- **THHLD_NUMPER**: Total number of people in household
- **THHLD_NUMKID**: Total number of people under 18-years-old in household
- **THHLD_NUMADLT**: Total number of people above or equal to 18-years-old in household
- **WRKLOSS**: Recent household job loss (1) Yes (2) No (-99) Question seen but category not selected (-88) Missing
- **KINDWORK**: Sector of employment (1) Government (2) Private company (3) Non-profit organization including tax exempt and charitable organizations (4) Self-employed (5) Working in a family business (-99) Question seen but category not selected (-88) Missing
- **ANXIOUS**: Frequency of anxiety over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **WORRY**: Frequency of worry over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **DOWN**: Frequency of feeling depressed over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **MORTLMTH**: Pay last month's mortgage or rent on time (1) Yes (2) No (3) Payment was deferred (-99) Question seen but category not selected (-88) Missing
- **MORTCONF**: Confidence in ability to pay mortgage or rent next month (1) Not at all confident (2) Slightly confident (3) Moderately confident (4) Highly confident (5) Payment is/will be deferred (-99) Question seen but category not selected (-88) Missing
- **INCOME**: Total household income (before taxes) (1) Less than $25,000 (2) $25,000 - $34,999 (3)

$35,000 - $49,999 (4) $50,000 - $74,999 (5) $75,000 - $99,999 (6) $100,000 - $149,999 (7) $150,000 - $199,999 (8) $200,000 and above (-99) Question seen but category not selected (-88) Missing

- **people_vaccinated**: Number of people vaccinated against COVID-19 with either 1 or 2 shots
- **people_vaccinated_per_hundred**: Number of people vaccinated against COVID-19 with either 1 or 2 shots per 100 population count
- **people_fully_vaccinated**: Number of people fully vaccinated against COVID-19
- **people_fully_vaccinated_per_hundred**: Number of people fully vaccinated against COVID-19 per 100 population count

## Appendix F

The master dataset represents individuals who participated in our surveys with population level COVID-19 and lockdown information appended. The time-frame is monthly.

Master Dataset:

- **YMFIPS**: Year, Month, FIPS Code (YYYY-MM-FIPS)
- **YEAR**: Year
- **MONTH**: Month of year
- **FIPS**: State FIPS Code (01-56 for the 50 States and 6 US Territories)
- **STATE**: State Name
- **STATE_CODE**: State abbreviation
- **TBIRTH_YEAR**: Year of birth
- **EEDUC**: Educational attainment (1) Less than high school (2) Some high school (3) High school graduate or equivalent (4) Some college, but degree not received or in progress (5) Associate's Degree (6) Bachelor's Degree (7) Graduate Degree
- **THHLD_NUMPER**: Total number of people in household
- **THHLD_NUMKID**: Total number of people under 18-years-old in household
- **THHLD_NUMADLT**: Total number of people above or equal to 18-years-old in household
- **WRKLOSS**: Recent household job loss (1) Yes (2) No (-99) Question seen but category not selected (-88) Missing
- **KINDWORK**: Sector of employment (1) Government (2) Private company (3) Non-profit organization including tax exempt and charitable organizations (4) Self-employed (5) Working in a family business (-99) Question seen but category not selected (-88) Missing
- **ANXIOUS**: Frequency of anxiety over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **WORRY**: Frequency of worry over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **DOWN**: Frequency of feeling depressed over previous 2 weeks (1) Not at all (2) Several days (3) More than half the days (4) Nearly every day (-99) Question seen but category not selected (-88) Missing
- **MORTLMTH**: Pay last month's mortgage or rent on time (1) Yes (2) No (3) Payment was deferred (-99) Question seen but category not selected (-88) Missing
- **MORTCONF**: Confidence in ability to pay mortgage or rent next month (1) Not at all confident (2) Slightly confident (3) Moderately confident (4) Highly confident (5) Payment is/will be deferred (-99) Question seen but category not selected (-88) Missing
- **INCOME**: Total household income (before taxes) (1) Less than $25,000 (2) $25,000 - $34,999 (3) $35,000 - $49,999 (4) $50,000 - $74,999 (5) $75,000 - $99,999 (6) $100,000 - $149,999 (7) $150,000 - $199,999 (8) $200,000 and above (-99) Question seen but category not selected (-88) Missing
- **REMPCT**: Percentage of people working from home
- **CDCCOUNT**: CDC Count of COVID-19 infections
- **people_vaccinated**: Number of people vaccinated against COVID-19 with either 1 or 2 shots
- **people_vaccinated_per_hundred**: Number of people vaccinated against COVID-19 with either 1 or 2 shots per 100 population count
- **people_fully_vaccinated**: Number of people fully vaccinated against COVID-19
- **people_fully_vaccinated_per_hundred**: Number of people fully vaccinated against COVID-19 per 100 population count
- **lockdown**: Was the person in lockdown (1) yes (2) no

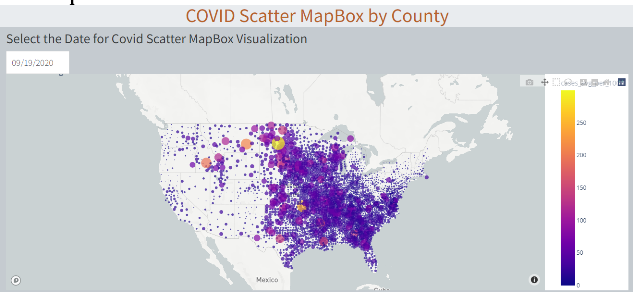
## Appendix G

Python Libraries:

- *sklearn.ensemble*: Random Forest Modeling
- *sklearn.naive_bayes*: Gaussian NB Modeling
- *sklearn.neural_network*: MLP Neural Networking
- *sklearn.linear_model*: Linear and Logistic Regression
- *sklearn.decomposition*: Principal Component Analysis
- *xgboost*: XGBoost ensemble learning models

## Appendix H

Example of the website visualizations:

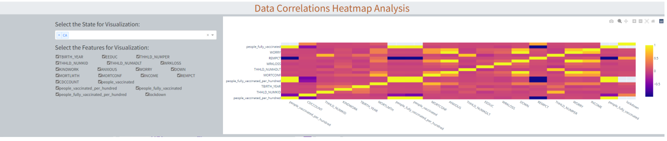


COVID Scatter MapBox by County

Select the Date for Covid Scatter MapBox Visualization

09/19/2020

Map Scatter Visualization



Data Correlations Heatmap Analysis

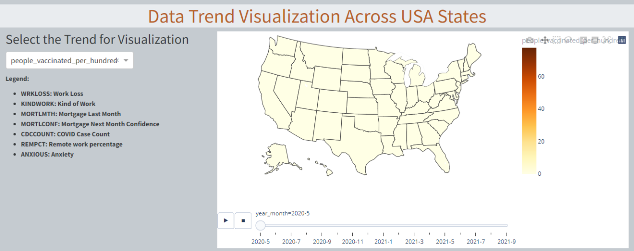
Select the State for Visualization:

× CA

Select the Features for Visualization:

☑TBIRTH_YEAR ☑EEDUC ☑THHLD_NUMPER
☑THHLD_NUMKID ☑THHLD_NUMADLT ☑WRKLOSS
☑KINDWORK ☑ANXIOUS ☑WORRY ☑DOWN
☑MORTLMTH ☑MORTCONF ☑INCOME ☑REMPCT
☑CDCCOUNT ☑people_vaccinated
☑people_vaccinated_per_hundred ☑people_fully_vaccinated
☑people_fully_vaccinated_per_hundred ☑lockdown

Heat-map Plot Visualization



Data Trend Visualization Across USA States

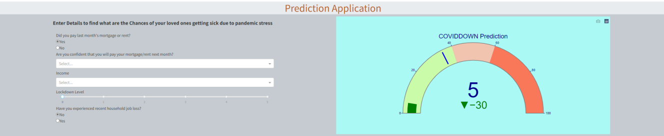
Select the Trend for Visualization

people_vaccinated_per_hundred

**Legend:**

- **WRKLOSS: Work Loss**
- **KINDWORK: Kind of Work**
- **MORTLMTH: Mortgage Last Month**
- **MORTLCONF: Mortgage Next Month Confidence**
- **CDCCOUNT: COVID Case Count**
- **REMPCT: Remote work percentage**
- **ANXIOUS: Anxiety**

Map Visualization



Prediction Application

**Enter Details to find what are the Chances of your loved ones getting sick due to pandemic stress**

Did you pay last month's mortgage or rent?

● Yes
○ No

Are you confident that you will pay your mortgage/rent next month?

Select...

Income

Select...

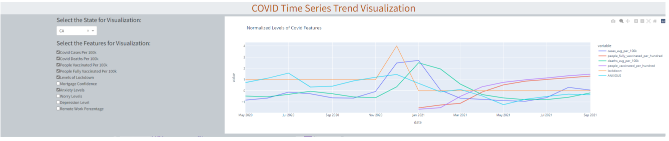
Lockdown Level

0 1 2 3 4 5

Have you experienced recent household job loss?

● No
○ Yes

Prediction Gauge Visualization



COVID Time Series Trend Visualization

Select the State for Visualization:

CA

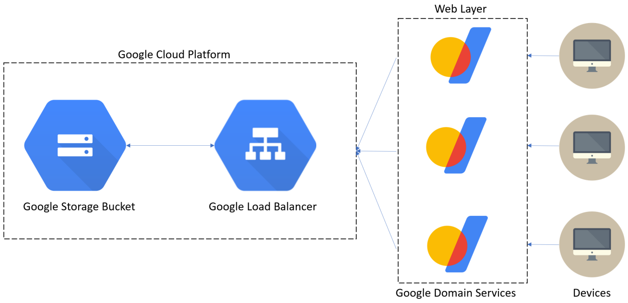
Select the Features for Visualization:

☑Covid Cases Per 100k
☑Covid Deaths Per 100k
☑People Vaccinated Per 100k
☑People Fully Vaccinated Per 100k
☑Levels of Lockdown
☐Mortgage Confidence
☑Anxiety Levels
☐Worry Levels
☐Depression Level
☐Remote Work Percentage

Time, Trend Visualization

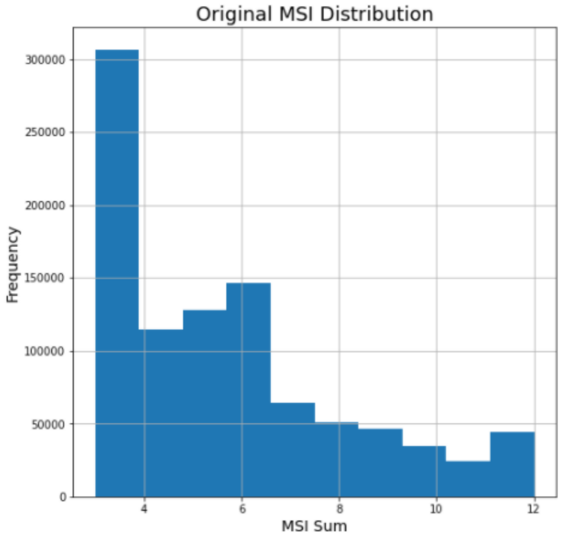
## Appendix I

Example of the website visualizations:



Architecture Diagram

## Appendix J



Original Data Distribution