

# Mental Health Across US States: A COVID-19 Impact Analysis

---

**Group members:** Fan Fan (ff2249), Sridevi Turaga (st4282), Sonali Mhatre (ssm10093), Sneha Trichy Shekar (skt9904)

## Abstract:

The COVID-19 pandemic has significantly impacted mental health across the United States. In this paper, we aim to compare and analyze the factors that influence people's mental health during COVID-19 period in different states. We will utilize datasets from Kaiser Family Foundation (KFF) primarily and join it with demographics, economic conditions, and early COVID-19 positivity rates within each state for our analysis. We expect to find significant evidence of increasing rates of anxiety and depression in the pandemic time in the United States. Based on research, we can know which group is prone to anxiety or depression and find whether the other factors will affect the anxiety or depression rate.

## Introduction:

The COVID-19 pandemic has brought about significant difficulties to the United States for the public health crisis, particularly with regard to mental health. Anxiety and depression rate can reflect how COVID-19 influences people's mental health. In light of the anticipated increase in anxiety and depression rates from 11% in 2019 to over 41% by 2021, our research aims to provide guidance for focused interventions and customized support networks that cater to the distinct mental health requirements of various communities.(Vankar, Preeti.Statistics,2023) This trend suggests the seriousness of mental health issues during a pandemic and emphasizes the pressing need to learn more about the people and factors that contribute to them. The purpose of this study is to provide a thorough analysis of the various elements affecting mental health during a pandemic in each state by fusing precise KFF data with economic and COVID-19 case rate data. We pay particular attention to the ways in which these variables interact and impact the state of mental health in various communities. To try and determine which groups of people are more likely to experience anxiety or depression, as well as how wealth and the report of positivity rate will affect the anxiety and depression rate in each state, we employ time series, correlation, network, SVM, Random Forest, and Bayesian inferences.

## Research Question:

How has the COVID-19 pandemic affected rates of anxiety and depression in the United States, with a particular focus on potential variations by demographics, economic factors and the possible influence of early COVID-19 cases?

## Literature Review:

According to continuous surveys monitoring Americans' psychological distress levels since the pandemic began, the COVID-19 pandemic has sparked a complex and dynamic pattern of psychological suffering among them. Together with earlier evaluations, the September 2022 survey shows a stable but unpredictable trend in psychological discomfort among adult Americans. Although 21% of respondents in the most recent survey were classified as suffering high distress, a greater proportion (41%) reported feeling high distress at least once in the previous 2.5 years, highlighting the inconsistent pattern of distress attacks. The results of intersectional analyses show relationships between distress and age, income, and disability status. Higher levels of distress were reported by those with disabilities or restricting health problems, and higher levels of distress were shown by lower-income households, which aggravated already-existing socioeconomic inequities.( Keeter, Giancarlo Pasquini and Scott,2022)

Disparities in distress are also visible among demographic categories, with women, persons with lower incomes, and adults who identify as Black or Hispanic being disproportionately impacted. This is consistent with public health alerts about vulnerable populations experiencing increased distress during the pandemic (In CDC survey, 37% of U.S. high school students report regular mental health struggles during COVID-19 pandemic, 2022). Despite anxieties about the virus, the study highlights other complex stressors that contribute to discomfort, such as societal and economic difficulties. Two-thirds of people reported having anxiety or difficulty sleeping, while despair and loneliness were also significant distress markers ( Giancarlo, Scott,2022)

The majority of respondents reported feeling optimistic about the future, while a significant minority reported feeling infrequently or rarely hopeful, highlighting the intricate relationship between resilience and suffering. Although the study offers insightful information, it may have overlooked longer-term implications for mental health due to its focus on self-reported measures and immediate discomfort experiences. Therefore, more longitudinal research is necessary. In conclusion, considering the changing nature of psychological distress during the COVID-19 pandemic, focused interventions are essential to addressing a range of mental health needs, especially among vulnerable groups ( Pasquini, Giancarlo.2022)

## **Data:**

Data links(drive): [Data](#)

We primarily used the KFF datasets to analyze factors affecting anxiety rates in the context of individual states. In addition to the dataset which calculates the overall anxiety and depression rates for each state, we also used the anxiety and depression rate dataset with age and gender minute statistics. There are also annual income and positivity rates for each state.

## **Data prepossessing:**

For datasets, we check whether they have missing or NaN values and then drop them off and fill them with zero at first. After that, because some data shows as NSD which means no statistical number in the dataset, we replace all NSD with zero to make sure our cell only has numeric without words. Then changed the datetime format to the timeline. Filter the columns and state that we do not need to make sure the dataset is clean. We also standardized part of our dataset in the late part for the analysis.

## **Methods:**

### Time Series Analysis/Model ( Stanislav Sobolevsky Applied Data Science Session1)

For the project we used ARIMA, PACF, ACF plots to understand the time-related patterns. Time series is mainly used for pattern recognition such as signal detection, financial trends and predictive modeling like stock prices. ARIMA (AutoRegressive Integrated Moving Average) combines autoregressive (AR) and moving average (MA) models and integrates the concept of differencing to make the time series data stationary. An ARIMA model is typically defined by three parameters: (p, d, q), where 'p' is the number of autoregressive terms, 'd' is the number of non-seasonal differences needed for stationarity, and 'q' is the number of lagged forecast errors in the prediction equation.

$$(1 - \sum_{i=1}^p \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^q \theta_i L^i)\varepsilon_t$$

Seasonal ARIMA-SARIMA extends the ARIMA model by adding seasonal terms. This model is particularly useful for time series with clear seasonal patterns. Its defined by the parameters, (p, d, q)(P, D, Q)s, where 'P', 'D', and 'Q' are the seasonal autoregressive, differencing, and moving average terms

respectively, and 's' is the length of the seasonal cycle.

PACF is Partial Autocorrelation Function and is used to determine the extent of the association between a time series and its lags. It measures the direct effect of past data points on the current data point by eliminating the effects of earlier data points. The PACF is useful in identifying the appropriate lag length 'p' in an AR model.

ACF, Autocorrelation Function is used to measure the linear relationship between lagged values of the time series. Unlike the PACF, which finds the correlation of the residuals with the next lag value, thus holding constant the rest, the ACF considers all the intermediate components. ACF plots are used to determine the 'q' parameter in the MA model.

These tools are integral components of the time series forecasting toolbox and are used to understand data behaviors, detect patterns, and build predictive models. Understanding the underlying data generating processes and assumptions for each model is crucial for producing reliable forecasts.

### Correlation Analysis

We used the Correlation Analysis to compare variables such as rates and demographics features. Correlation Analysis helps us in assessing the strength and direction of the linear relationship between two continuous variables. It is often measured by the correlation coefficient, typically represented by Pearson's r. The coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. In machine learning, correlation analysis helps us in feature selection by identifying pairs of features that are highly correlated, which may imply redundancy. The formula for Pearson's r is as follows, where  $x_i$ ,  $y_i$  are the individual sample points indexed with i, while  $\bar{x}$  and  $\bar{y}$  are the mean values of the corresponding datasets.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

### Network Analysis

For this project we created graphs for various States using correlation and variables like gender, income, population and anxiety and depression rates. Network Analysis technique helps us to analyze complex networks by using graphs composed of nodes (representing entities) and edges (representing relationships between the entities). Some metrics in network analysis include degree centrality, betweenness centrality, and clustering coefficient, without a specific universally applicable formula. These metrics help in understanding the importance of nodes, the flow or control of information, and the tendency for clustering within the network. Machine learning applications related to Network Analysis include link prediction and community detection.

### SVM

SVM is a supervised machine learning algorithm that is used for classification or regression challenges. It works by finding the hyperplane that best divides a dataset into classes. The support vectors are the data points closest to the hyperplane, and they influence the position and orientation of the hyperplane. The optimization objective of SVM is to maximize the margin between the data points and the hyperplane. For nonlinear problems, SVM uses the kernel trick to transform the feature space to make the data linearly separable. For our project we tried to predict Anxiety values for each State. The decision function for a simple linear SVM is given by:  $f(x) = w * x + b$ , where  $w$  is the weight vector,  $x$  is the feature vector, and  $b$  is the bias.

### Random Forest

Random Forest is an ensemble learning technique that builds multiple decision trees and merges them to get a more accurate and stable prediction. Each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. The general prediction function of a Random Forest for classification or regression can be represented as:

$$y = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

where n is the number of trees,  $f_i(x)$  is the prediction of i'th tree and x is the input feature vector.

### Bayesian Inference

Bayesian Inference is a method of statistical inference where Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayes' theorem is expressed as:

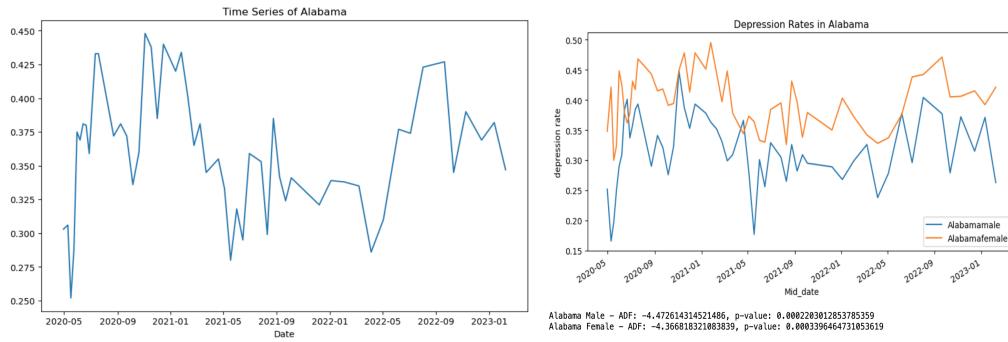
$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

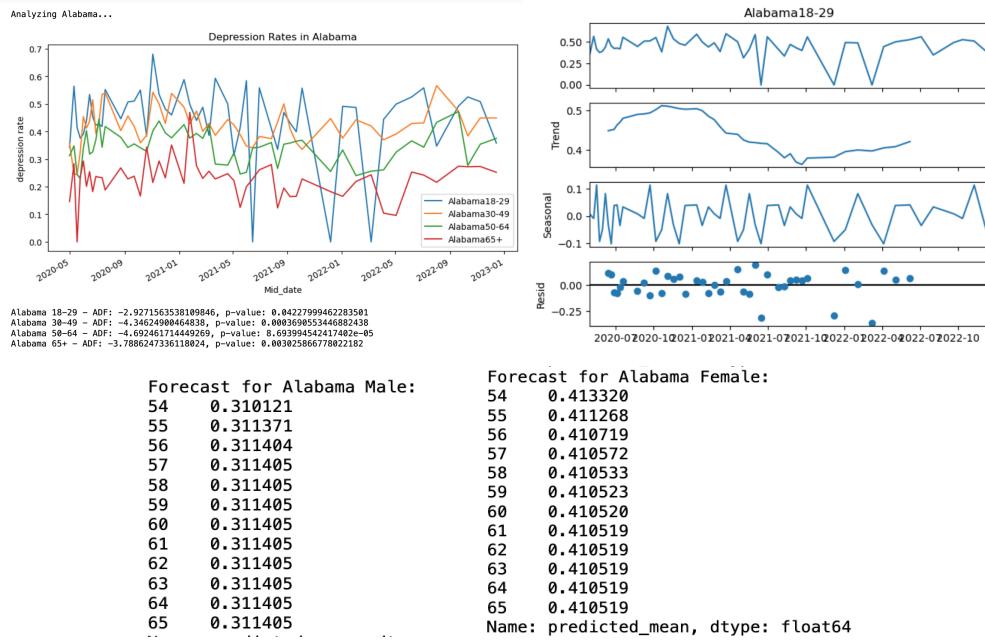
where  $P(H|E)$  is the posterior probability of hypothesis H given the evidence E,  $P(E|H)$  is the likelihood of observing evidence E given that hypothesis H is true,  $P(H)$  is the prior probability of hypothesis H, and  $P(E)$  is the marginal likelihood of evidence E. Bayesian Inference is heavily utilized in machine learning to deal with uncertainty and to incorporate prior knowledge into the model.

## **Results:**

### Time Series Analysis

We performed time series analysis on three datasets from KFF. We used ACF, PACF, ARIMA, adfuller for each state for both sexes and different age groups. Our example here is the state of Alabama. There are also other state graphs in the code. In the state as a whole, anxiety rates fluctuate widely, with women having higher rates than men. For the age groups, there are significant variations in depression rates between age groups. The younger age group (18-29 years old) in particular exhibited greater and more fluctuating rates of depression, which might be traced to other life questions, financial difficulties, or social pressures. We can see that p-value is less than 0.05. This indicates that the sequence does not contain any structure that is self-correlated over time, enabling more time-series analysis to be done, like forecasting depression/anxiety rates in the future. There is one prediction graph between females and male. We can find that the prediction of depression rate for females is higher 10 percent than male. This may reflect that females are more likely to get depression compared with male. The prediction rate of Alabama is 0.358609. This result is a bit different from the random forest one, but similar to the SVM model. We reviewed all the graphs in the code, and found that most of the states showed the same trend: that a brief rise was followed by a decline and then another minor rise. For age group, the slope for the second minor rise is different that 65+age group's slope is steep than other groups. Maybe because of less work pressure.





## Correlation Analysis

For the correlation analysis, we use the states anxiety/depression dataset and the states income dataset. We want to find the correlation between these two variables. We expect the result will show that each state's income has negatively correlated with the anxiety/depression rate. When the income decreases, the anxiety/depression rate will increase. But the result shows a difference with our thinking. I think there might also be other factors that we ignored and need to add them for future research. We can find some states have a positive relationship and some have a negative relationship. Positive correlation means the income increases, the anxiety/depression rate also increases. Negative correlation means the income increases, the anxiety/depression rate decreases. We can find some strong correlation in some states such as Massachusetts, Michigan, Alaska and Maine. Some of the states' results are not same as our opinion and some states do not show strong correlation between income and anxiety/depression rate, so this means that there are other reasons that influence the anxiety/depression rate that we do not relate for this time.

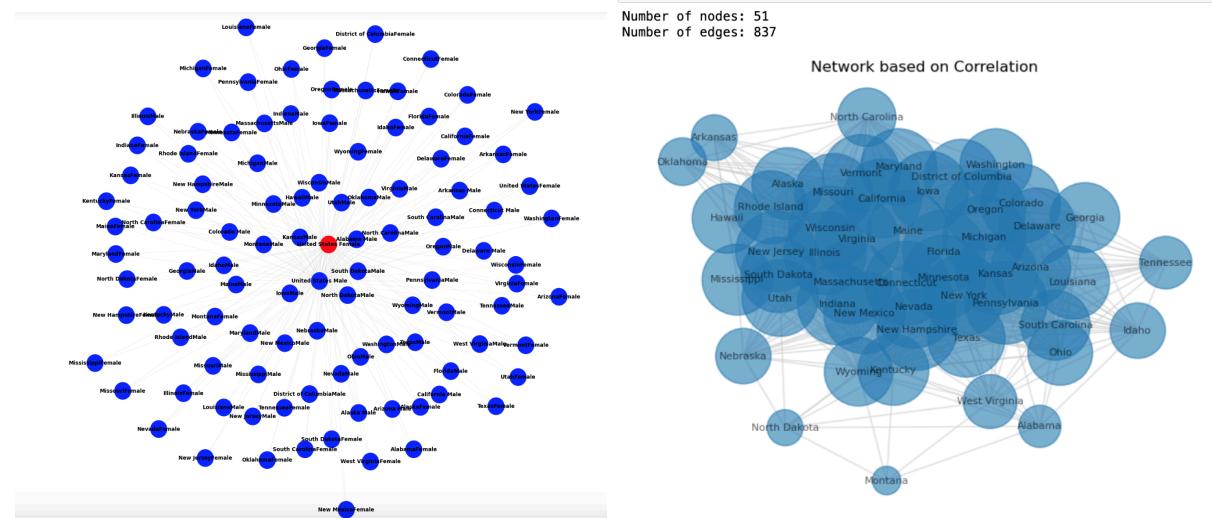
0	Alabama	0.216348
1	Alaska	-0.989968
2	Arizona	0.374389
3	Arkansas	0.848879
4	California	0.584043
5	Colorado	0.987531
6	Connecticut	0.134363
7	Delaware	-0.127167
8	District of Columbia	-0.398939
9	Florida	0.168869
10	Georgia	-0.571599
11	Hawaii	0.415717
12	Idaho	-0.155377
13	Illinois	0.337409
14	Indiana	0.225207
15	Iowa	0.361810
16	Kansas	0.383188
17	Kentucky	0.551067
18	Louisiana	-0.514861
19	Maine	-0.996258
20	Maryland	-0.180887
21	Massachusetts	0.998778
22	Michigan	0.999732
23	Minnesota	0.016193
24	Mississippi	0.201181
25	Missouri	-0.249311
26	Montana	-0.420250
27	Nebraska	-0.601568
28	Nevada	-0.609498

## Network Analysis

In this project, through the network analysis for prediction we understand that the impact of nodes for United States Male and United States Female depends on the degree of centrality, betweenness centrality,

and closeness centrality where the results for Average Shortest Path Length between the nodes is 1.9623. The visualization explains the vulnerability of the State according to the gender, age, income and positivity of the population during the COVID-19. In conclusion, the nodes closer to the center are likely to be more vulnerable than the ones away from the center.

Average Shortest Path Length: 1.9623



### SVM model

We want to use the SVM model to predict the future anxiety/depression rate in each state. The result of mean square error(mse) can measure the average of the squares of the differences between the model's predicted and actual values. The result of mse totally shows positively that the model is fit. The model's predicted accuracy increases as the MSE value decreases. For California and Maryland, the model brings best results that the mse extremely close to zero. The significant variations in MSE could be caused by changes in the quality of data collection or the volatility of anxiety rate data across states. It might also be the result of the model not being able to account for important variables that affect anxiety levels in particular states. We may need to change other variables to make the result become more accurate. The prediction for Alabama is 0.35473 which is similar to the result of time series one.

```

Mean Squared Error for Alabama: 9.536207775469392e-05
Mean Squared Error for Alaska: 0.00055782981301939
Mean Squared Error for Arizona: 3.291135734071951e-05
Mean Squared Error for Arkansas: 0.0030361066674361344
Mean Squared Error for California: 0.0006567530201600491
Mean Squared Error for Colorado: 0.0002122611284241397
Mean Squared Error for Connecticut: 8.809633733456482e-07
Mean Squared Error for Delaware: 5.657942828562883e-07
Mean Squared Error for District of Columbia: 0.0008369347491535836
Mean Squared Error for Florida: 4.7229964989227595e-06
Mean Squared Error for Georgia: 2.9786520852569814e-06
Mean Squared Error for Hawaii: 0.002807140658664202
Mean Squared Error for Idaho: 0.0003719146083410288
Mean Squared Error for Illinois: 0.0009073835026161877
Mean Squared Error for Indiana: 6.229224857648565e-05
Mean Squared Error for Iowa: 0.000197599742283774
Mean Squared Error for Kansas: 3.4176333102492784e-06
Mean Squared Error for Kentucky: 4.944771468144073e-06
Mean Squared Error for Louisiana: 8.378672283779635e-05
Mean Squared Error for Maine: 0.0001608891966759005
Mean Squared Error for Maryland: 0.0007386236582409973
Mean Squared Error for Massachusetts: 9.553347664666083e-05

```

### Random Forest

In this project, the random forest model provides predictions for COVID-19 impact on states based on states. The Mean Squared Error and the R square value is used to quantify the Random Forest Model. The Mean Squared Error (MSE) value of 5.2678e-05 indicates the average squared value difference between actual and predicted COVID-19 impact value. The R-squared value of 0.9608 indicates the variance in COVID-19 impact, showing that higher the value, the better the fit. This means that the model is a good fit without overfitting.

```
{'Alabama': 0.3738399999999995,
 'Alaska': 0.3160400000000004,
 'Arizona': 0.3469900000000035,
 'Arkansas': 0.3949499999999995,
 'California': 0.3305899999999955,
 'Colorado': 0.33139,
 'Connecticut': 0.3276099999999973,
 'Delaware': 0.2588099999999997,
 'District of Columbia': 0.3308899999999935,
 'Florida': 0.322810000000003,
 'Georgia': 0.2957399999999995,
 'Hawaii': 0.3068600000000013,
 'Idaho': 0.3322900000000014,
 'Illinois': 0.3229500000000003,
 'Indiana': 0.3811900000000001,
 'Iowa': 0.2701999999999983,
 'Kansas': 0.2974999999999965,
 'Kentucky': 0.4033400000000003,
 'Louisiana': 0.4031699999999997,
 'Maine': 0.3632599999999996,
 'Maryland': 0.31286,
 'Massachusetts': 0.2770999999999985,
 'Michigan': 0.33151,
 'Minnesota': 0.2775300000000005,
 'Mississippi': 0.4385399999999997,
 'Missouri': 0.3065800000000035,
 'Montana': 0.3609099999999995,
 'Mississippi': 0.4385399999999997,
 'Missouri': 0.3065800000000035,
 'Montana': 0.3609099999999995,
 'Nebraska': 0.3044699999999995,
 'Nevada': 0.3830900000000002,
 'New Hampshire': 0.2880800000000017,
 'New Jersey': 0.3598099999999995,
 'New Mexico': 0.3945400000000045,
 'New York': 0.3194600000000013,
 'North Carolina': 0.3231000000000004,
 'North Dakota': 0.3265700000000002,
 'Ohio': 0.2822099999999996,
 'Oklahoma': 0.4238099999999963,
 'Oregon': 0.3449400000000036,
 'Pennsylvania': 0.3563899999999996,
 'Rhode Island': 0.3732900000000007,
 'South Carolina': 0.3134499999999973,
 'South Dakota': 0.2792399999999993,
 'Tennessee': 0.3453399999999965,
 'Texas': 0.3701199999999984,
 'Utah': 0.3825700000000024,
 'Vermont': 0.25885,
 'Virginia': 0.3200900000000002,
 'Washington': 0.2850299999999998,
 'West Virginia': 0.3720399999999999,
 'Wisconsin': 0.2758500000000004,
 'Wyoming': 0.3408799999999946}
```

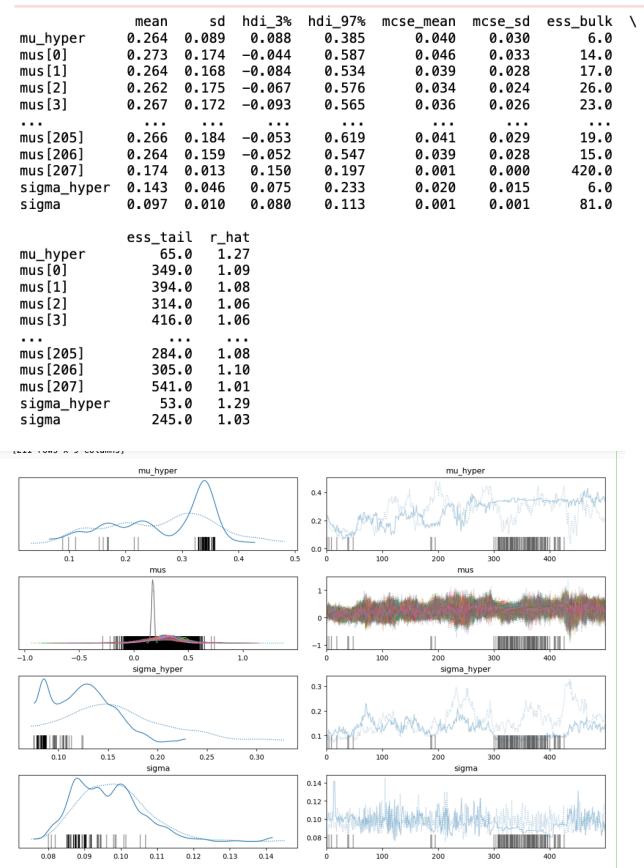
**MSE: 5.267799999999995e-05**

**Rsquare: 0.9607857687764544**

Root Mean Squared Error : RMSE =  $\sqrt{(\text{MSE})} \approx 0.0073$ .

### Bayesian Inference

We also try to make a model by Bayesian Inference. We want to create a model to predict the anxiety/depression rate of different age groups in each state. But the result does not show well, we do not use this model to do the prediction. With a standard deviation of 0.089 and a mean value of roughly 0.264 for mu\_hyper (the hyperparameter  $\mu$ ), the model's estimate of this hyperparameter is comparatively concentrated and has minimal uncertainty. The r\_hat should ideally be around 1.0, which implies that the sampling has converged. While mu\_hyper and sigma\_hyper have higher r\_hat values (1.27 and 1.29), indicating that these parameters may not have reached a good convergence and that more iterations or an update of the model structure may be necessary, the majority of the parameters have r\_hat values close to or equal to 1.0 based on the output. These results show that we need to improve our model to achieve prediction.



## Conclusion:

Based on our research, it appears that gender, age, and socioeconomic factors significantly influence anxiety rates across different states. In particular, anxiety is more common in women and younger people, which suggests that specific mental health interventions are needed. The varied relationship between income and anxiety levels highlights the intricate relationship between mental health and economic circumstances even more. Going forward, it will be critical that we focus on the health of young women, making use of trends seen in monthly data to create timely and efficient interventions. Furthermore, broadening the scope of our analysis to incorporate a wider range of parameters may improve the prediction and comprehension of anxiety trends among various demographic groups. This all-encompassing strategy helps larger initiatives to enhance mental health outcomes in diverse populations in addition to helping to create customized health policies.

## Job Description:

**Writing** - Fan Fan, Sridevi Turaga, Sonali Mhatre, Sneha Trichy, Anne Mercado (responsible for overseeing the written submission and slides)

**Research** - Fan Fan, Sonali Mhatre, Sridevi Turaga, Sneha Trichy Shekar, Anne Mercado (responsible for background study and collating existing research material)

**Data Engineering** - Fan Fan, Sridevi Turaga, Sneha Trichy Shekar, Anne Mercado (responsible for data collection cleaning and feature engineering)

**Modeling** - Fan Fan, Sridevi Turaga, Sneha Trichy Shekar, Anne Mercado (responsible for data modeling)

**Data Interpretation + Visualizations** - Sonali Mhatre (responsible for turning data work into cohesive

thoughts for the final submission)

**References:**

Keeter, Giancarlo Pasquini and Scott. "Young Adults Are Especially Likely to Have Experienced High Psychological Distress since March 2020." *Pew Research Center*, Pew Research Center, 12 Dec. 2022, [www.pewresearch.org/short-reads/2022/12/12/at-least-four-in-ten-u-s-adults-have-faced-high-levels-of-psychological-distress-during-covid-19-pandemic/](https://www.pewresearch.org/short-reads/2022/12/12/at-least-four-in-ten-u-s-adults-have-faced-high-levels-of-psychological-distress-during-covid-19-pandemic/).

Pasquini, Giancarlo. "At Least Four-in-Ten U.S. Adults Have Faced High Levels of Psychological Distress during COVID-19 Pandemic." *Pew Research Center*, Pew Research Center, 12 Dec. 2022, [www.pewresearch.org/short-reads/2022/12/12/at-least-four-in-ten-u-s-adults-have-faced-high-levels-of-psychological-distress-during-covid-19-pandemic/](https://www.pewresearch.org/short-reads/2022/12/12/at-least-four-in-ten-u-s-adults-have-faced-high-levels-of-psychological-distress-during-covid-19-pandemic/)

Sobolevsky, Stanislav. Time Series Analysis slides

Sobolevsky, Stanislav. Network Analysis slides

Sobolevsky, Stanislav. Bayesian Inference slides

Vankar, Preeti. "Adults Reporting Anxiety or Depression U.S. 2019 vs. 2021." *Statista*, 29 Nov. 2023, [www.statista.com/statistics/1221102/anxiety-depression-symptoms-before-since-covid-pandemic-us/](https://www.statista.com/statistics/1221102/anxiety-depression-symptoms-before-since-covid-pandemic-us/)