

Final Project Report

Group 4

Members: Manoj Kumar Selvaraj, Amritesh Koul, Megha
Karatela, Alan Chen, Ishika Bhargava, Geethalakshmi
Venkateswara Raju, Saisha Shori

Professor Naser Islam

MIS 6380.002

November 27, 2021

Executive Summary

The COVID-19 pandemic has had an unprecedented impact on communities throughout the world. It has undoubtedly had a psychological impact on the people and many have suffered from mental health disorders such as anxiety and depression to name a few.

With this project we aim to study the impact of the consequences of the pandemic on the mental health of the population. Lockdown measures implemented post the virus outbreak have hindered the regular lifestyle of people, eventually deteriorating their mental health. We have utilized two datasets for the analysis of impact of the pandemic with regards to mental health, one of them being the mental health data from CDC and the other one being the Covid dataset from Our World in Data. We intend to explore the following hypotheses:

- Will the spike in Covid cases lead to a spike in mental health disorder amongst the population of the United States?
- Will certain age groups have a higher severity of depression than other age groups?
- Will gender play a factor in the state of mental health during Covid?
- Will location have an impact with regards to mental health disorder and Covid?

By conducting analysis and forming visualizations with regards to the state of mental health of people during the Covid era, we were able to derive crucial insights for each hypothesis. We could observe a correlation between the population and the state of mental health due to Covid. The greater the number of Covid cases, more were the cases of depression. The age groups that had shown high effects of depression were

the youth, ranging from 18-29 and 30-39 years. Furthermore, the female gender consistently showed a higher depression value and percent of depressed cases than males. Lastly, the impact caused by Covid was minimal with regards to location and the depression severity. Overall, we explored the state of mental health with regards to the entire population during the pandemic year; then, we delved deeper and explored factors such as age, gender, and location in conjunction to this trend and their impact.

INDEX

I.	Data Description.....	4
II.	Data Cleaning.....	5
III.	General Introduction.....	9
IV.	Insights & Findings.....	10
	A. Covid Cases & Mental Health Disorder.....	10
	B. Age & Depression Severity.....	12
	C. Gender & Mental Health.....	14
	D. Location & Mental Health.....	16
V.	Conclusion.....	23

Data Description

Project Datasets:

- Mental Health Data

The U.S. Census Bureau, in collaboration with five federal agencies, launched the Household Pulse Survey to produce data on the social and economic impacts of Covid-19 on American households. The Household Pulse Survey was designed to gauge the impact of the pandemic on employment status, consumer spending, food security, housing, education disruptions, and dimensions of physical and mental wellness. It has cases of people with symptoms of depression and anxiety disorder. In our data set, we have a timeline with various categorial subgroups like Age, Gender, Location. Apart from that, we have a Depression Value variable representing the numerical equivalent of the severity of depression.

- Covid Data

The Covid dataset from Our World In Data, represents the global covid cases and covid deaths across a timeline for all countries. In addition, this dataset contained situational factors such as hospital beds, ICU patients, and vaccination availability.

Data Cleaning

Dataset 1 was the Mental Health Data from *Center for Disease Control & Prevention*.

Dataset 2 was the Covid Data from *Our World in Data*. The core aim with regards to the cleaning process was to merge and format the datasets in such a manner that the time specifications and granularity are synchronized while the unique attributes of each dataset is preserved. Before proceeding forward with the cleaning process, we conducted a brainstorming session. During this gathering, we analyzed the datasets using excel where we tried to understand the characteristics and perceivable features of the data. Then, we mapped the important attributes for each hypothesis. Once the necessary dimensions were finalized, we began cleaning each dataset separately and then merged the respective components. The Pandas library for the Python language was primarily utilized to carry out the cleaning procedures.

The Mental Health data had various demographic subgroups. However, we decided to primarily focus on factors such as age, gender, and location. The different subgroups were not separate columns; instead they were row values connected with the Group column, which identified the type of Subgroup. Thus, to filter the desired factors, the Group column had to be parsed. By parsing the Group column with the desired factors, the Subgroup column was automatically filtered. This resulted in the dataset showcasing all the different variable values for the corresponding subgroup type noted in the Group column. Then, certain data entries within the Subgroup column had different formats. Therefore, few regex operations had to be performed so all the values

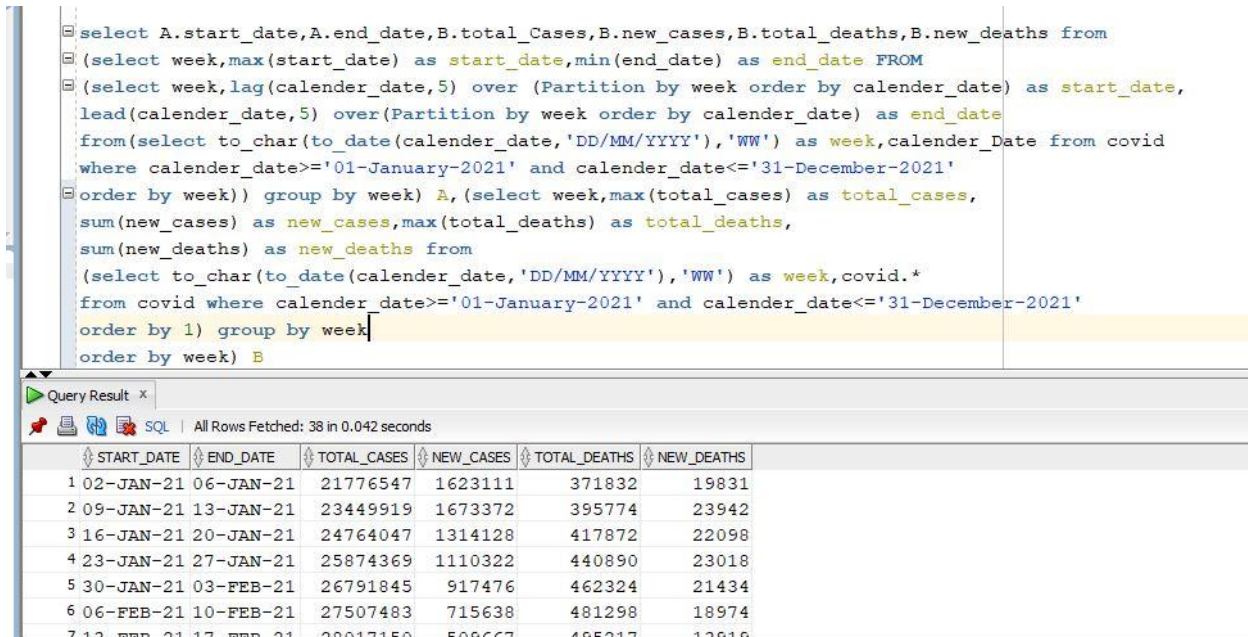
are of a standardized format. Lastly, some of the extra dimensions were removed as they did not play a role in the analysis.

The Covid dataset presented pandemic data from across the world along with various dimensions ranging from positivity rates to hospital beds availability index. There was a plethora of information in relation to progression of covid over time accompanied by various situational factors. Firstly, the scope of the dataset was reduced to only the United States. Then, many of the redundant and unnecessary columns were removed as they did not factor into the analysis. The records with missing and null values needed to be fixed. Luckily, the amount of such records was minimal. Hence, we dropped them from the dataset. Finally, the granularity of the dataset was converted into weeks such that it matches the granularity of the Mental Health dataset.

Once both the datasets were cleaned, they were merged to form the comprehensive dataset. The common dimension shared by both the datasets was the time component. As the time dimensions are now synchronized, both the datasets were merged on this column with the base being the Mental Health dataset. Lastly, this newly generated dataset now contained all the mental health and covid dimensions necessary for performing analytics and forming visualizations.

Data Cleaning Process, Sample Screenshots:

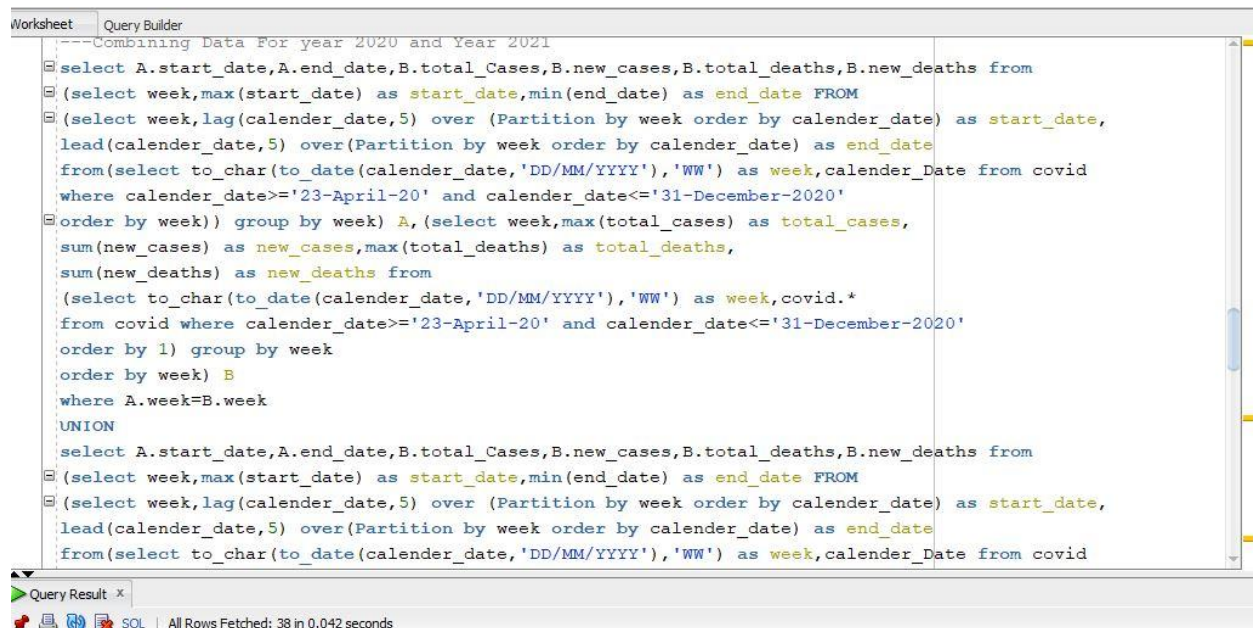
- SQL Files



```

select A.start_date,A.end_date,B.total_Cases,B.new_cases,B.total_deaths,B.new_deaths from
(select week,max(start_date) as start_date,min(end_date) as end_date FROM
(select week,lag(calender_date,5) over (Partition by week order by calender_date) as start_date,
lead(calender_date,5) over(Partition by week order by calender_date) as end_date
from(select to_char(to_date(calender_date,'DD/MM/YYYY'),'WW') as week,calender_Date from covid
where calender_date>='01-January-2021' and calender_date<='31-December-2021'
order by week)) group by week) A,(select week,max(total_cases) as total_cases,
sum(new_cases) as new_cases,max(total_deaths) as total_deaths,
sum(new_deaths) as new_deaths from
(select to_char(to_date(calender_date,'DD/MM/YYYY'),'WW') as week,covid.*
from covid where calender_date>='01-January-2021' and calender_date<='31-December-2021'
order by 1) group by week
order by week) B
  
```

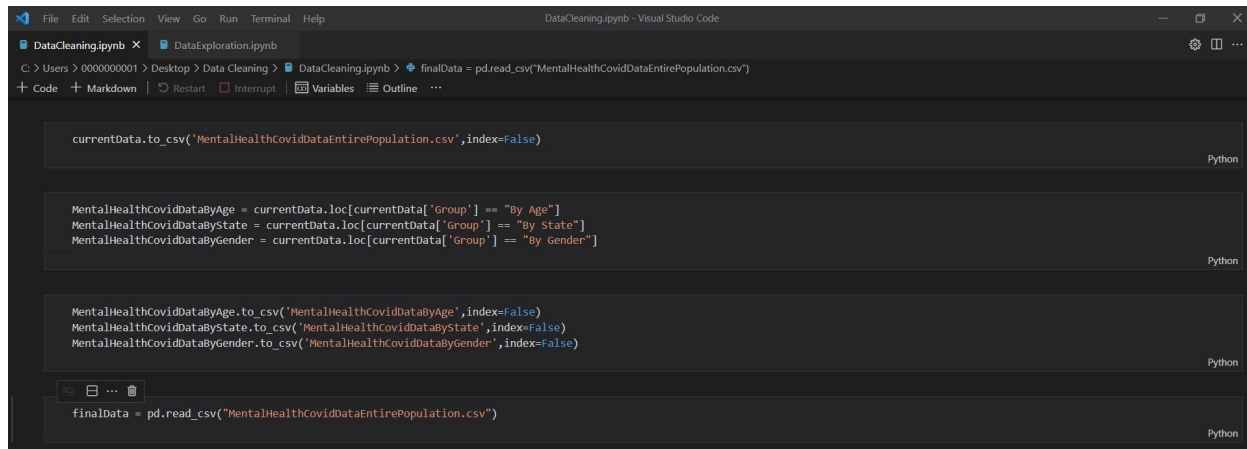
	START_DATE	END_DATE	TOTAL_CASES	NEW_CASES	TOTAL_DEATHS	NEW_DEATHS
1	02-JAN-21	06-JAN-21	21776547	1623111	371832	19831
2	09-JAN-21	13-JAN-21	23449919	1673372	395774	23942
3	16-JAN-21	20-JAN-21	24764047	1314128	417872	22098
4	23-JAN-21	27-JAN-21	25874369	1110322	440890	23018
5	30-JAN-21	03-FEB-21	26791845	917476	462324	21434
6	06-FEB-21	10-FEB-21	27507483	715638	481298	18974
7	13-FEB-21	17-FEB-21	28017150	500667	495217	13010



```

---Combining Data For year 2020 and Year 2021
select A.start_date,A.end_date,B.total_Cases,B.new_cases,B.total_deaths,B.new_deaths from
(select week,max(start_date) as start_date,min(end_date) as end_date FROM
(select week,lag(calender_date,5) over (Partition by week order by calender_date) as start_date,
lead(calender_date,5) over(Partition by week order by calender_date) as end_date
from(select to_char(to_date(calender_date,'DD/MM/YYYY'),'WW') as week,calender_Date from covid
where calender_date>='23-April-20' and calender_date<='31-December-2020'
order by week)) group by week) A,(select week,max(total_cases) as total_cases,
sum(new_cases) as new_cases,max(total_deaths) as total_deaths,
sum(new_deaths) as new_deaths from
(select to_char(to_date(calender_date,'DD/MM/YYYY'),'WW') as week,covid.*
from covid where calender_date>='23-April-20' and calender_date<='31-December-2020'
order by 1) group by week
order by week) B
where A.week=B.week
UNION
select A.start_date,A.end_date,B.total_Cases,B.new_cases,B.total_deaths,B.new_deaths from
(select week,max(start_date) as start_date,min(end_date) as end_date FROM
(select week,lag(calender_date,5) over (Partition by week order by calender_date) as start_date,
lead(calender_date,5) over(Partition by week order by calender_date) as end_date
from(select to_char(to_date(calender_date,'DD/MM/YYYY'),'WW') as week,calender_Date from covid
  
```


• Python Files



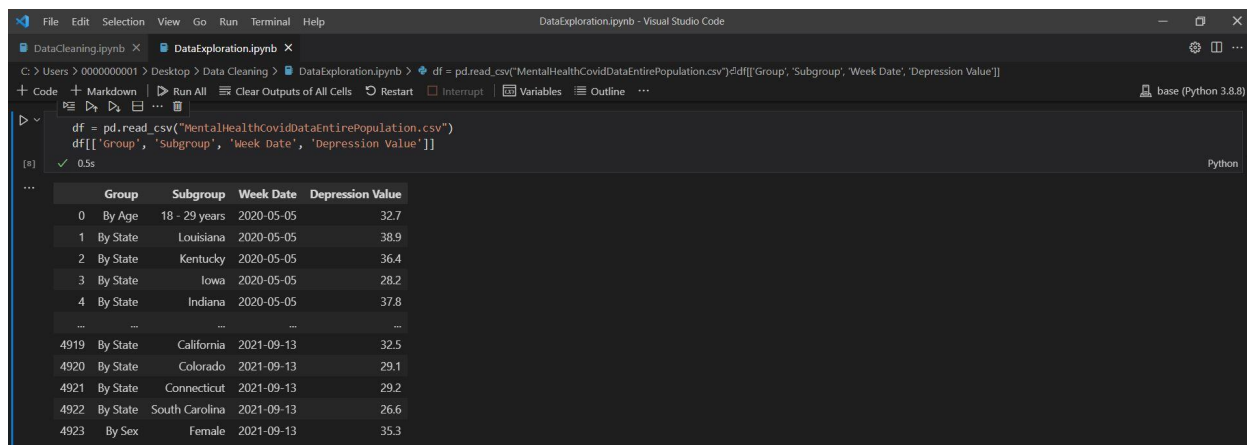
The screenshot shows a Jupyter Notebook with four cells. The first cell saves the current data to a CSV file. The second cell filters the data by group into three separate DataFrames. The third cell saves each of these DataFrames to individual CSV files. The fourth cell reads the original data back into a DataFrame named 'finalData'.

```
currentData.to_csv('MentalHealthCovidDataEntirePopulation.csv', index=False)

MentalHealthCovidDataByAge = currentData.loc[currentData['Group'] == "By Age"]
MentalHealthCovidDataByState = currentData.loc[currentData['Group'] == "By State"]
MentalHealthCovidDataByGender = currentData.loc[currentData['Group'] == "By Gender"]

MentalHealthCovidDataByAge.to_csv('MentalHealthCovidDataByAge.csv', index=False)
MentalHealthCovidDataByState.to_csv('MentalHealthCovidDataByState.csv', index=False)
MentalHealthCovidDataByGender.to_csv('MentalHealthCovidDataByGender.csv', index=False)

finalData = pd.read_csv("MentalHealthCovidDataEntirePopulation.csv")
```



The screenshot shows a Jupyter Notebook with two cells. The first cell reads the data from a CSV file into a DataFrame named 'df'. The second cell displays a preview of the data as a table.

```
df = pd.read_csv("MentalHealthCovidDataEntirePopulation.csv")
df[['Group', 'Subgroup', 'Week Date', 'Depression Value']]
```

	Group	Subgroup	Week Date	Depression Value
0	By Age	18 - 29 years	2020-05-05	32.7
1	By State	Louisiana	2020-05-05	38.9
2	By State	Kentucky	2020-05-05	36.4
3	By State	Iowa	2020-05-05	28.2
4	By State	Indiana	2020-05-05	37.8
...
4919	By State	California	2021-09-13	32.5
4920	By State	Colorado	2021-09-13	29.1
4921	By State	Connecticut	2021-09-13	29.2
4922	By State	South Carolina	2021-09-13	26.6
4923	By Sex	Female	2021-09-13	35.3

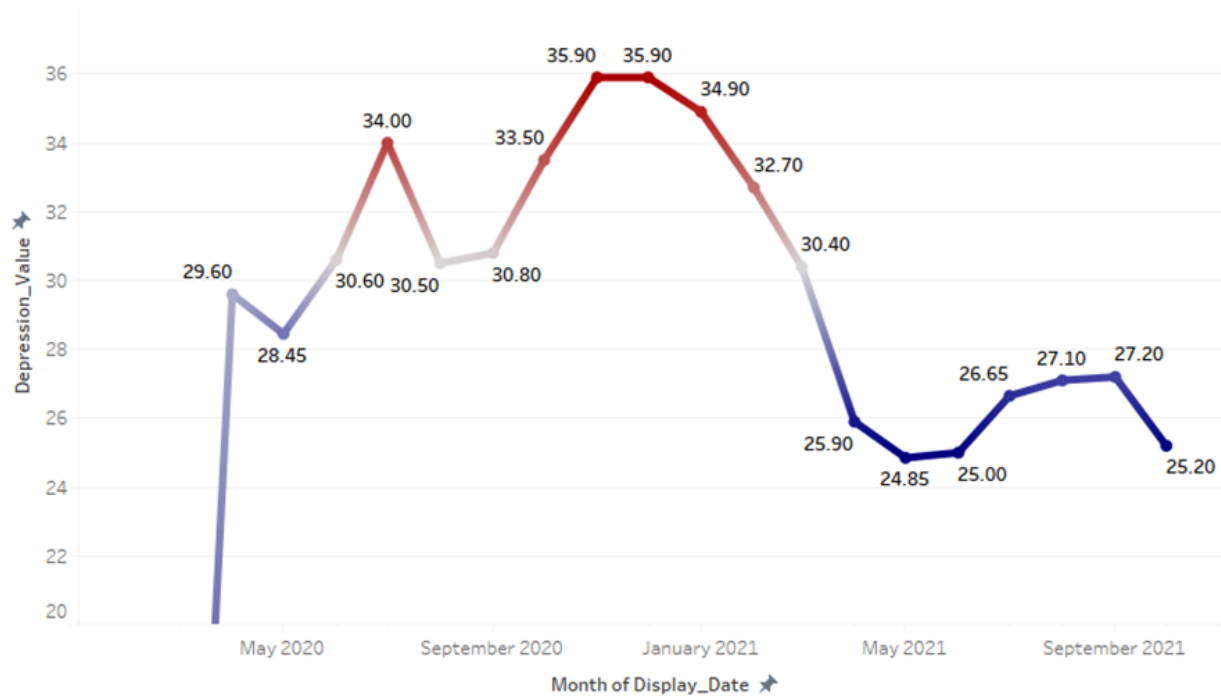
General Introduction

During the Covid pandemic, individuals' mental health has been in a state of rut as a result of the lockdown and isolation. As a result, we looked into the trends and links between mental health and covid by examining factors such as population, age, gender, and location. We investigated our hypotheses as to how mental health and Covid trends correlated. Based on the formed assumptions, we aimed to discover if the increase in Covid instances leads to an increase in mental health. We looked at which age groups have the highest levels of depression. We also examined how men and women's mental health fared throughout the pandemic year. Lastly, we also wanted to explore how mental health affected people living in different states and the impact caused by Covid with regards to this relation. In the upcoming section, we have documented our findings derived from the performed analytics and generated visualizations.

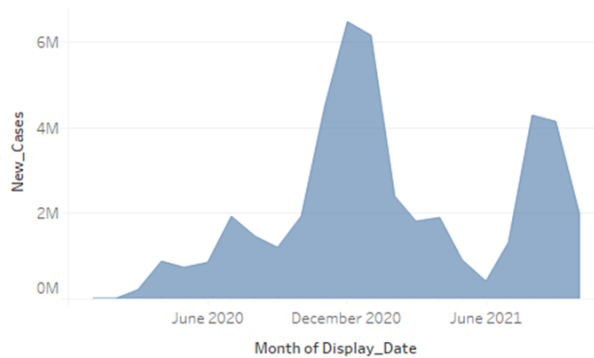
Insights & Findings

- Covid Cases & Mental Health Disorder

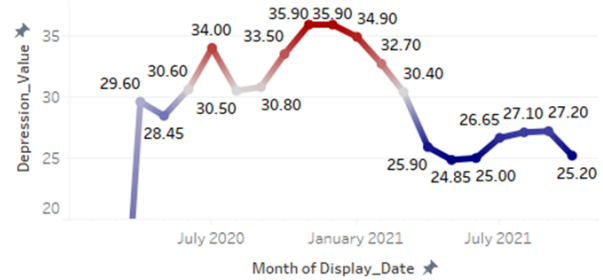
Depression Value - October 2021



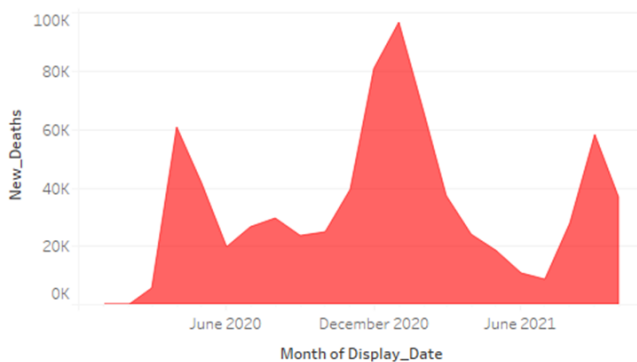
Covid Cases - October 2021



Depression Value - October 2021



Covid Deaths - October 2021

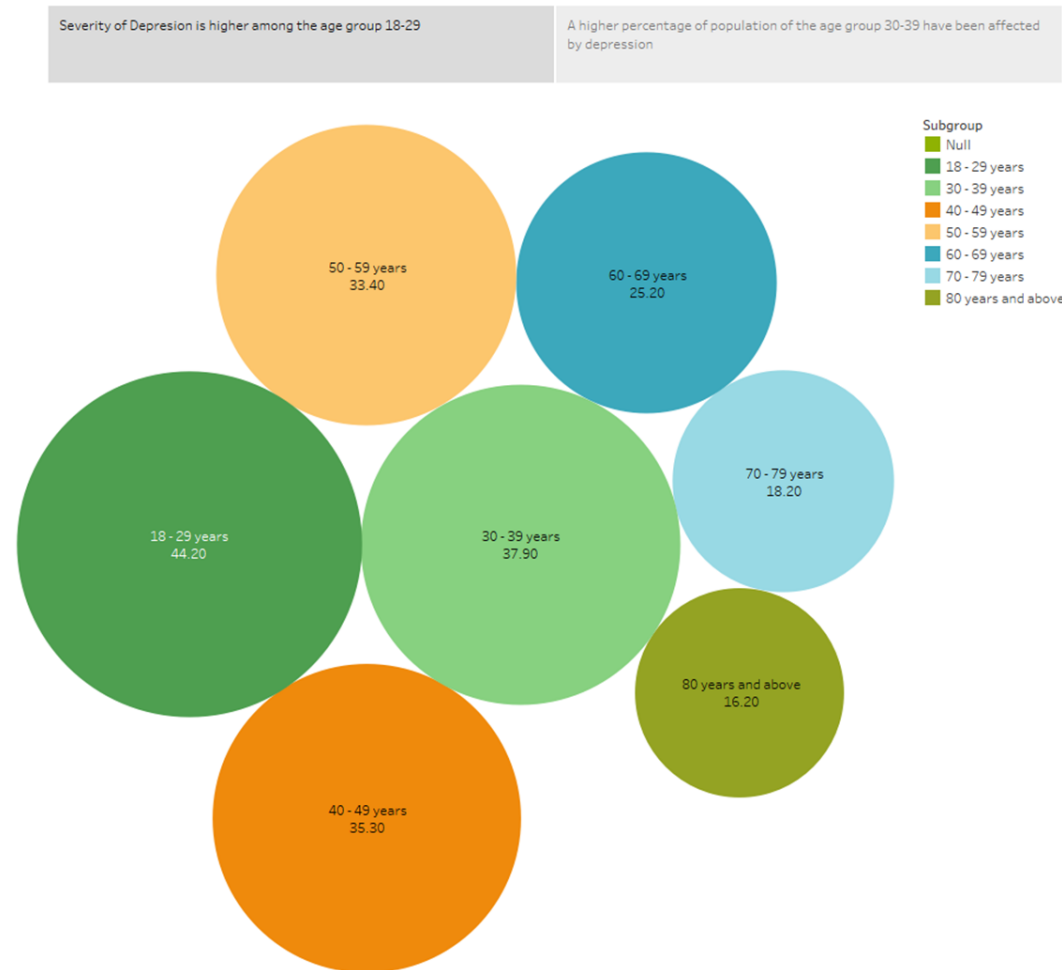


Month - Year
October 2021
☒ Show history

We compared covid cases and depression severity and observed that as covid cases increased from May 2020 to July 2020, depression cases also increased. When covid cases took a dip from July 2020 to September 2020, depression cases also showed a noticeable decrease. Our significant observation was that when the covid cases and deaths were at their peak in December 2020, depression cases also peaked at their highest at 35.90. Hence, we can assume that there is a strong correlation between the rise in covid cases and depression severity. The above visualizations provide evidence that the number of covid cases were at their peak from December 2020 to January 2021, which is also the time when we can see a peak in depression value.

• Age & Depression Severity

Depression Instances Across Age

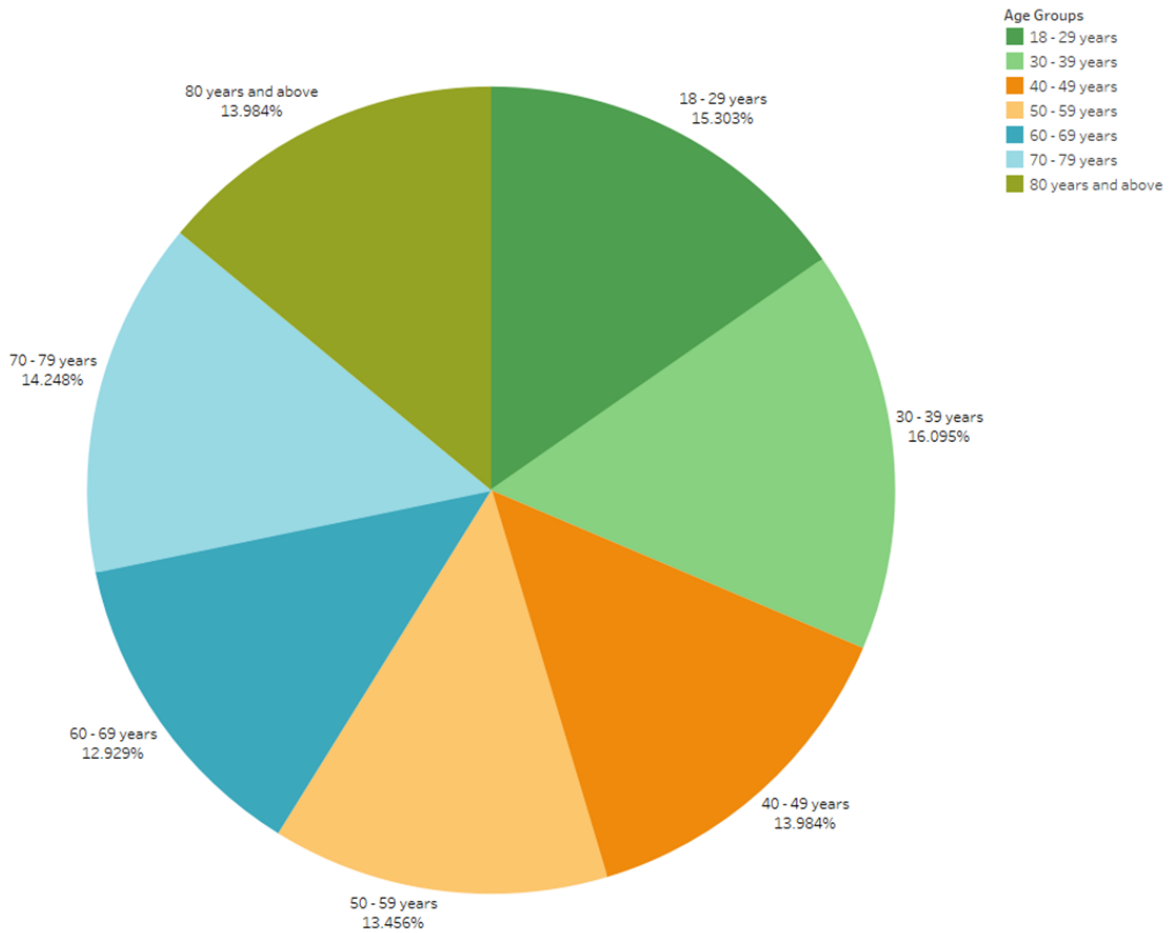


The packed bubbles visualization indicates the increase in median depression values for different age groups due to consequences of the Covid pandemic. Age is a crucial parameter of how likely a person is to suffer from depression. The visualization clearly indicates that the maximum severity of depression can be seen in the 18 - 29 age group estimated at 44.20 and the least could be seen in age group 80+ estimated at 16.20.

Depression Instances Across Age

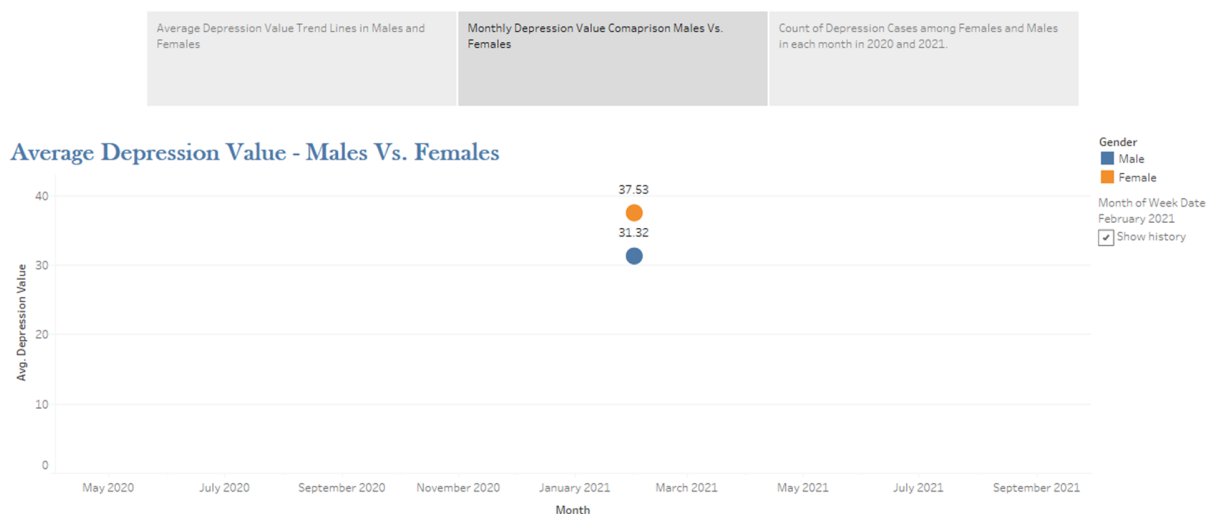
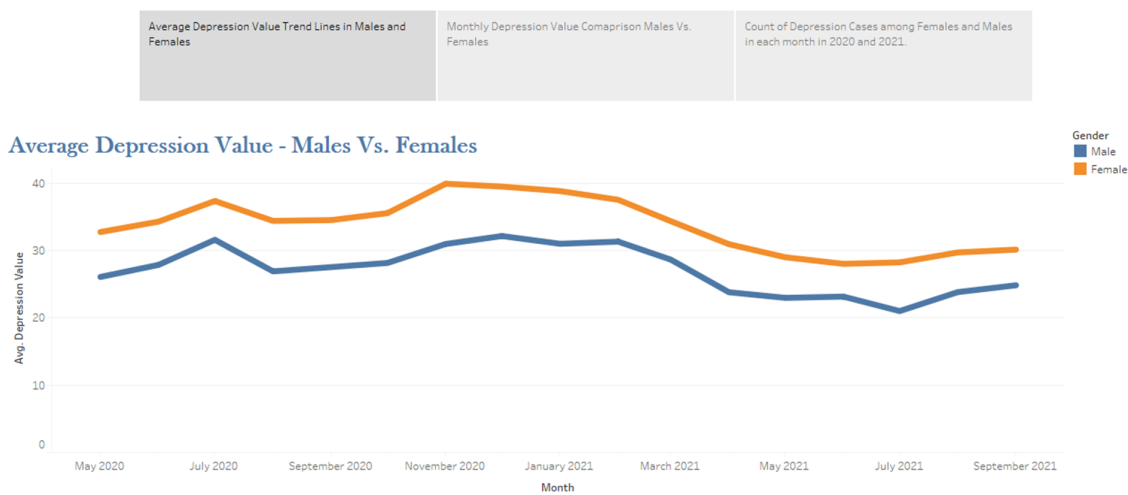
Severity of Depression is higher among the age group 18-29

A higher percentage of population of the age group 30-39 have been affected by depression

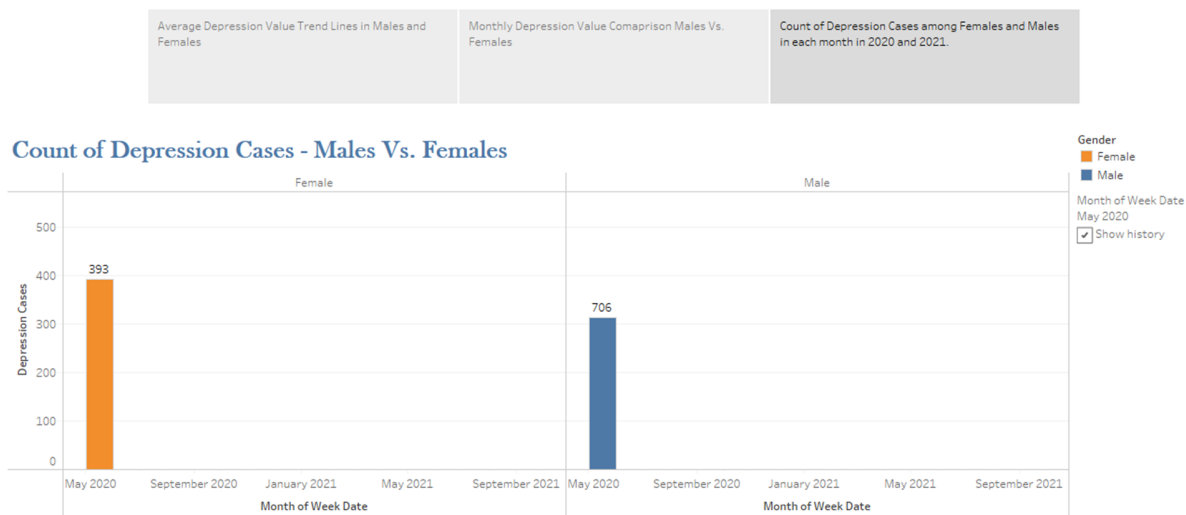


The pie chart visualization gives the percentage of people affected by depression during Covid based on the age group. The chart clearly indicates that younger age groups were more affected by the pandemic. For example, the people who are in the age group between 18 - 29.

● Gender & Mental Health



The dual line graph represents the average depression values based on gender. The orange line represents the females and the blue line represents the males. This graph gives details about the gender, month, average depression value. The depression values of females are higher than those of males. Depression reached its peak in November 2020 for females and December 2020 for males.



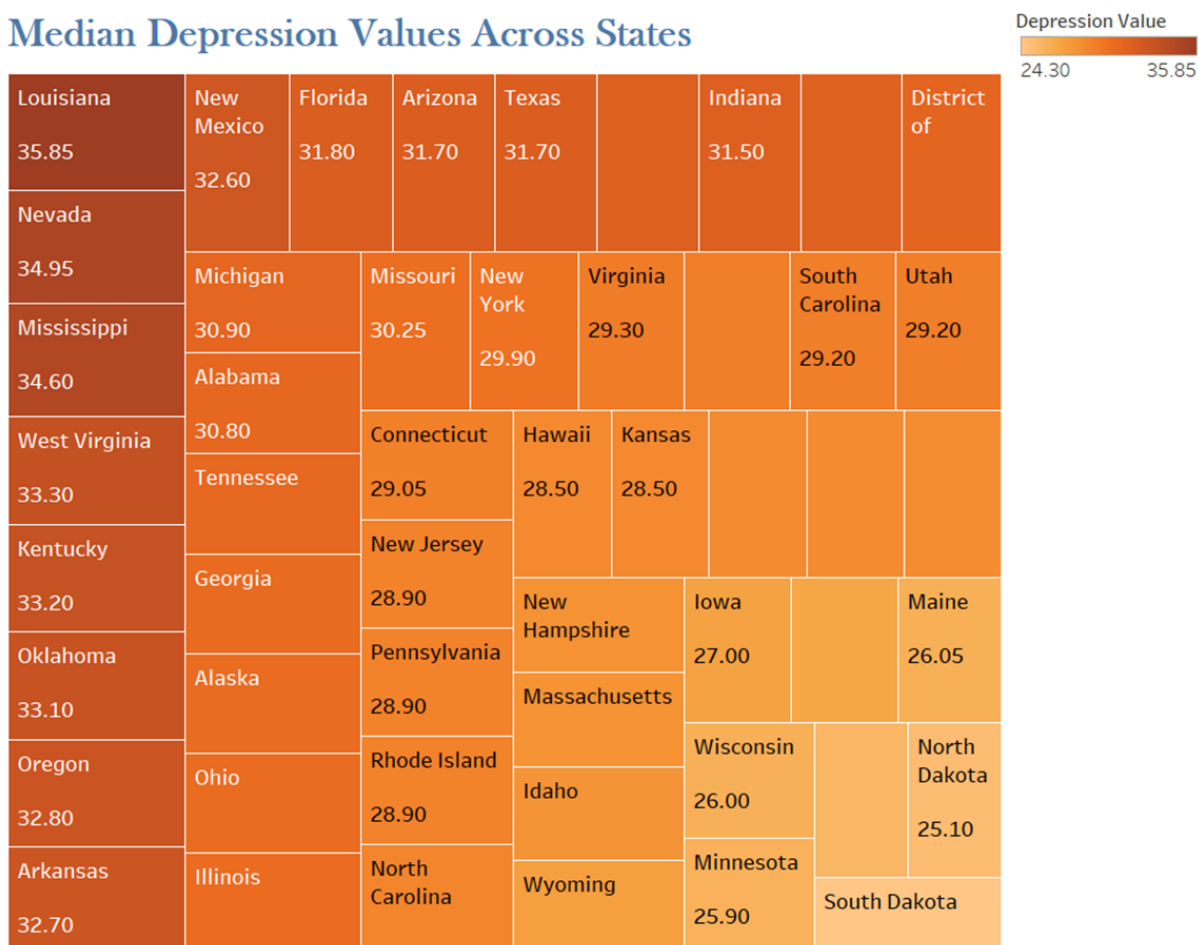
The bar graph represents the count of depression cases based on gender. This graph gives details about gender, month, and depression cases. The number of depression cases were high in females when compared to males. The highest count reached by females was 514 cases and the highest count reached by males was 418 cases.

• Location & Mental Health

Southern & Western States seem to showcase higher severity of depression. The severity of depression seems to be less amongst Midwestern & Northeastern States.

The median depression range hovers around 24 to 36, indicating predominantly Phase 2.

Median Depression Values Across States



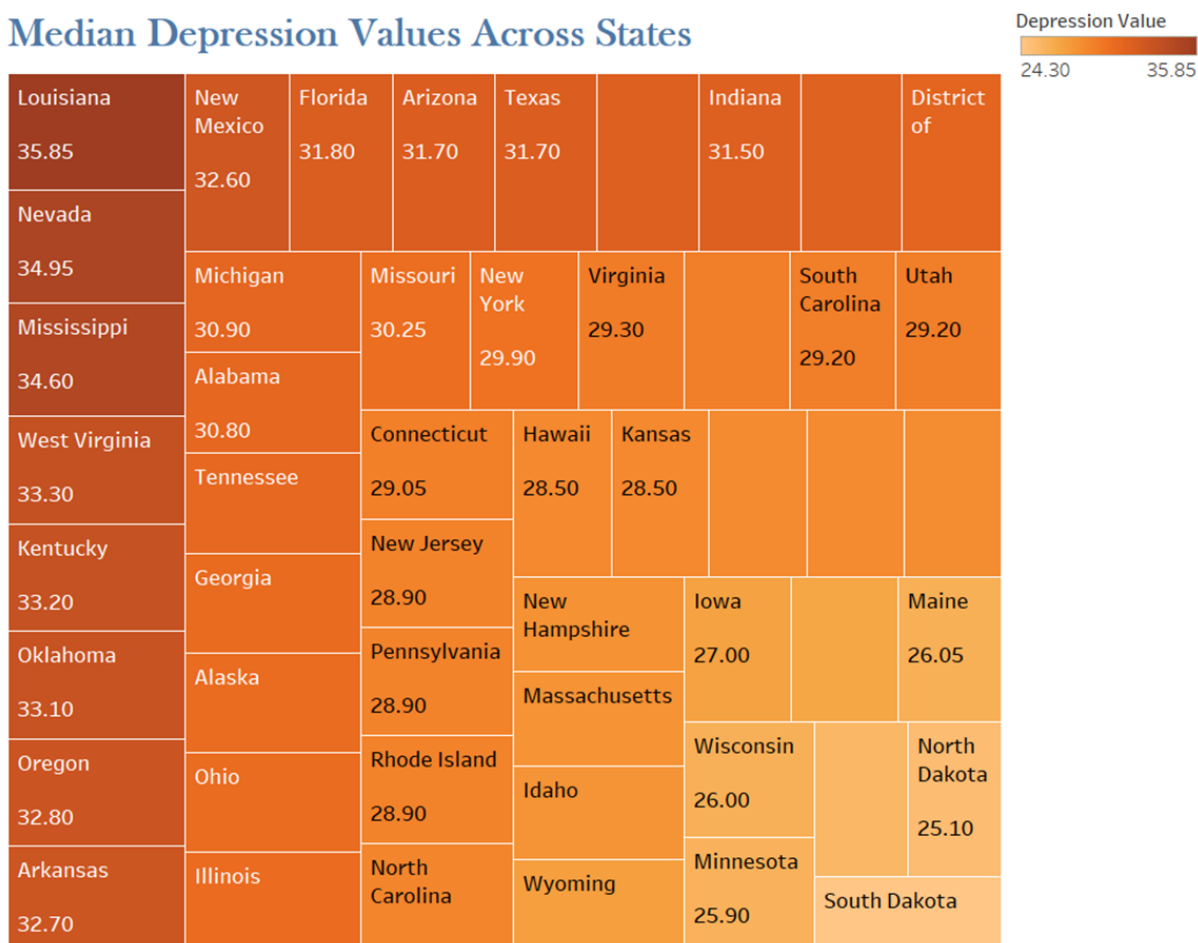
The tree map shows relational information about the median depression values across various states in the US. Southern and Western states have higher depression severity. Midwestern and Northeastern states have lower depression severity. Louisiana has the highest median depression value at 35.85.

Southern & Western
States seem to show..

The median depression range hovers around 24 to 36, indicating predominantly Phase 2.

The Southern & Western
States region cluster a..

Median Depression Values Across States



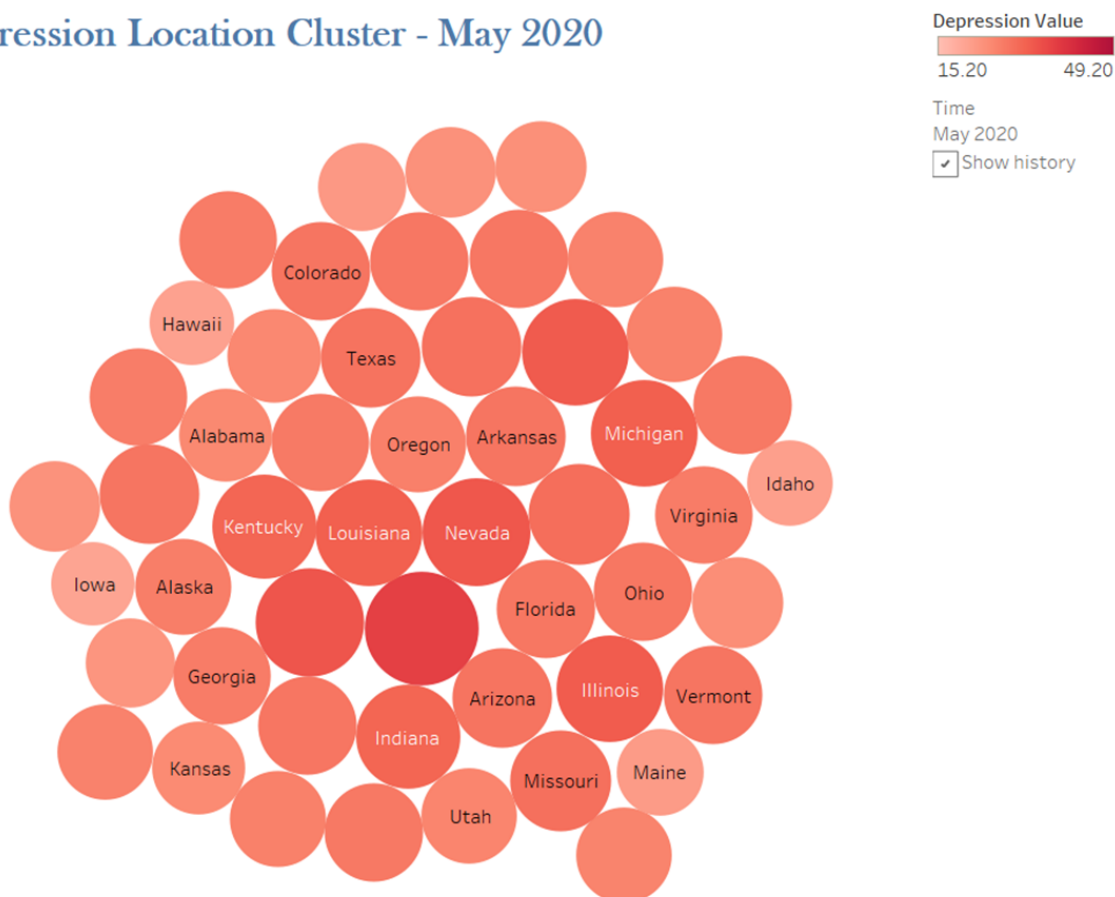
The tree map shows that the median depression range hovers around 24 to 36, indicating predominantly Phase 2, which is related to Anxiety or Depressive Disorder.

The median
depression range ho..

The Southern & Western States region cluster are the most dominant in depression severity and are represented in the center.

The Southern Belt
States had the highest ..

Depression Location Cluster - May 2020



The packed bubbles animation is comparing depression values of each state for every month from May 2020 to September 2021. The depression severity is emphasized via darker colored bubbles and is of central position. Southern and Western regions are the most dominant in depression severity.

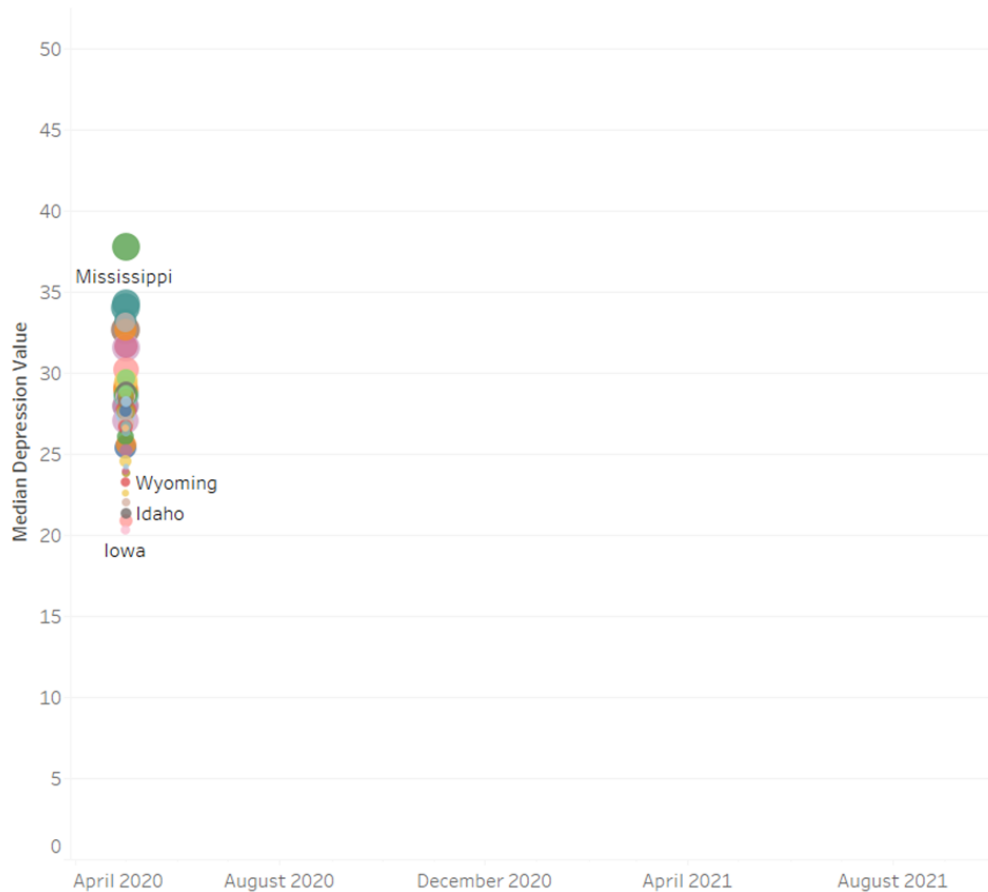
The Southern & Western States regi..

The Southern Belt States had the highest depression severity and the Midwestern states had the lowest depression severity through the course of the pandemic year.

There is a weak correlation with regar..

Depression Location Spread - May 2020

Time
May 2020
☒ Show history



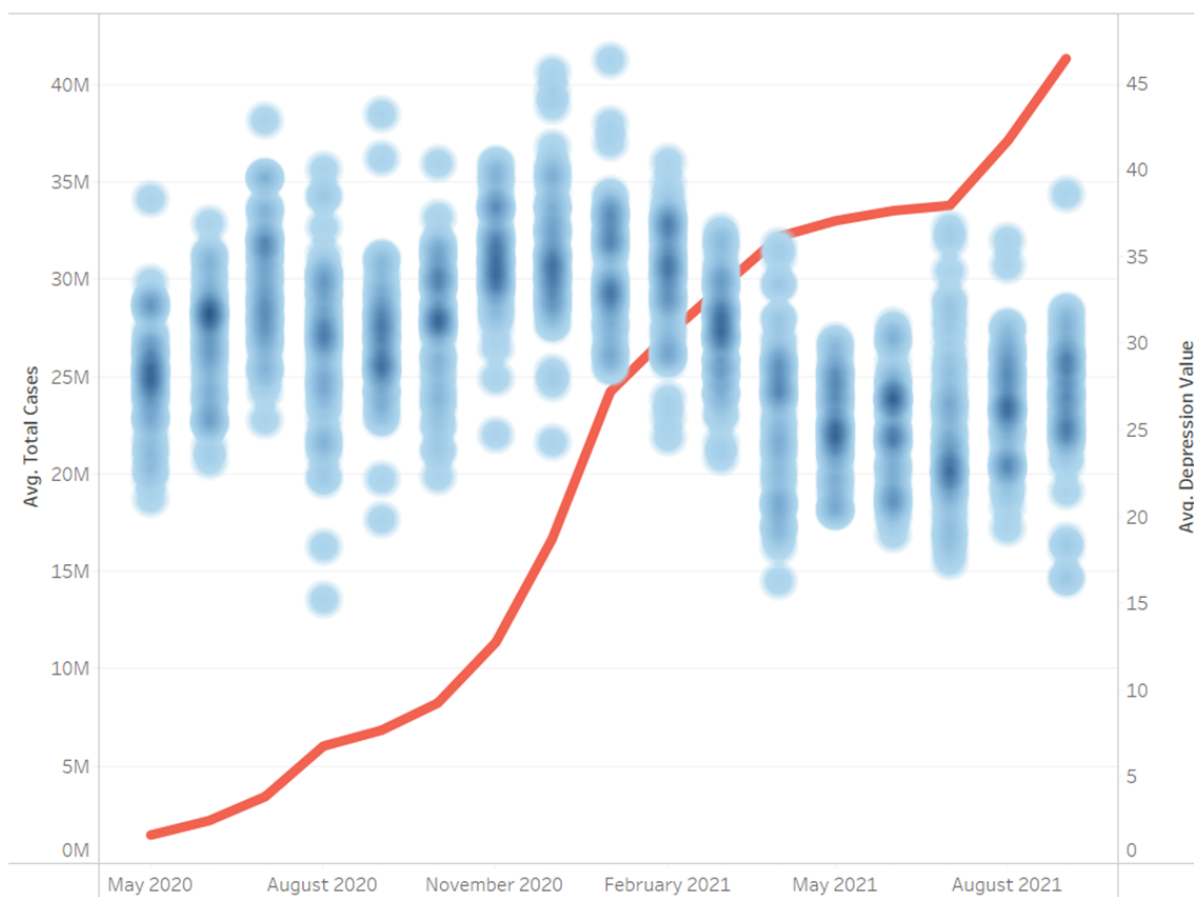
The bubble chart animation represents the median depression values for all states at each time point from May 2020 to September 2021. This visualization format helps us identify outliers and trend repetitions. The Southern belt states had the highest depression severity and the Midwestern states had the lowest depression severity during the pandemic.

The Southern Belt
States had the highe..

There is a weak correlation with regards to location factors and depression when
compared against the covid trend line.

Western &
Northeastern States h..

Location & Depression Against Covid Cases



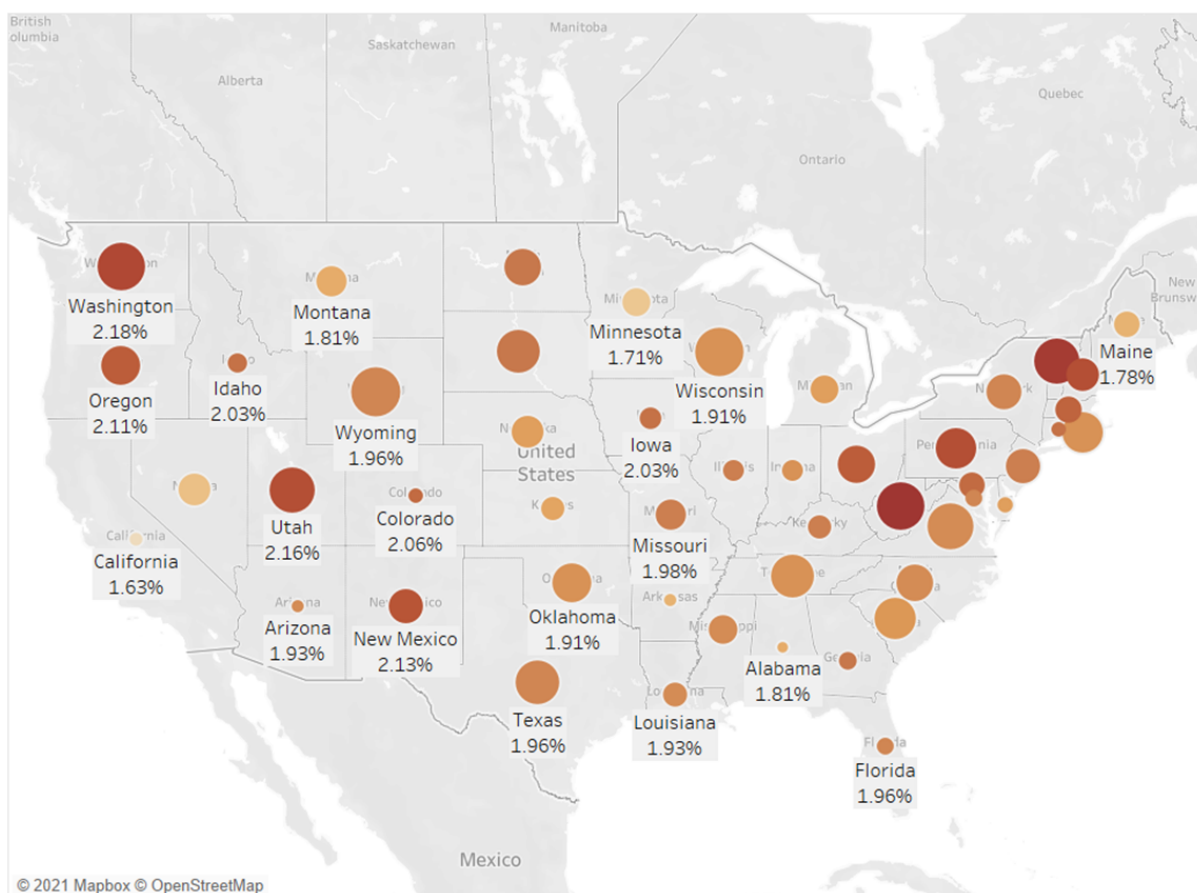
The density chart represents the depression density of states over time with regards to covid cases trend line. Overall, there is a weak correlation with regards to location factors and depression when compared against covid trend line.

There is a weak correlation with reg..

Western & Northeastern States have the highest concentration of reporting instances.

The population magnitude does not sh..

Depression Instances Across States

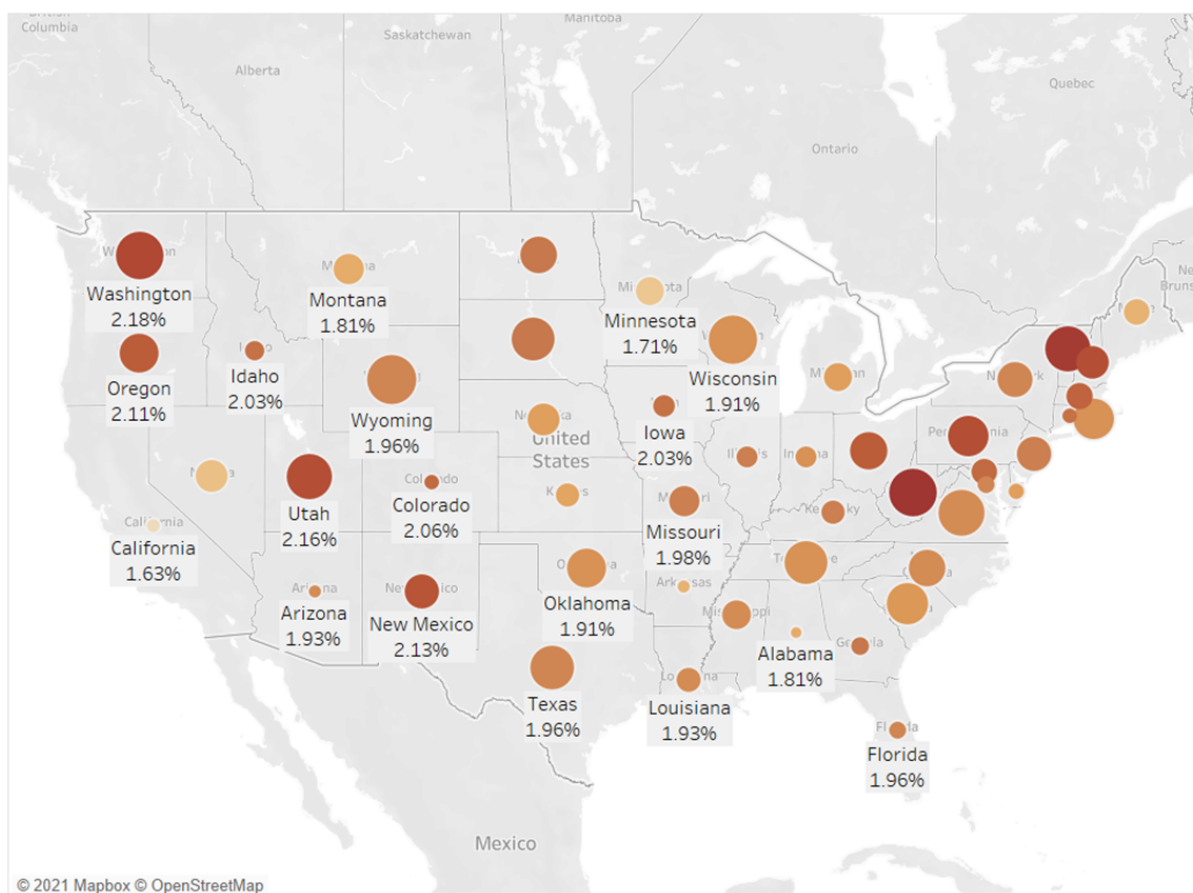


The geographic map spatially represents the spread of the reporting instances as to how many percent of cases stemmed from each state. From this map, we can deduce that Western and Northeastern states have the highest concentration of reporting instances.

Western & Northeastern States have the highest concentration of reporting instances.

The population magnitude does not show correlation with regards to the reporting magnitude.

Depression Instances Across States



The geographic map gives insights as to how population magnitude does not show correlation with regards to the reporting magnitude. Small states like Washington have higher reporting instances than a state like Texas.

Conclusion

In conclusion, the mental health of individuals was impacted by COVID-19 throughout 2020 and 2021. Overall, the Mental Health data by CDC and the Covid data by Our World in Data were analyzed and the insights were presented through data visualizations to show trends and correlations.

The factors for analysis of mental health included population, age, gender, and location. There was a correlation presented between the population and the mental health due to COVID-19. The greater the number of COVID-19 cases and the greater the cases of depression. The age ranges that had shown greater effects of depression were the youth population ranging from 18-29 and 30-39 years. The female gender consistently showed higher depression values and percent of depressed cases than males throughout the pandemic. Lastly, the impact of Covid was minimal with regards to location and the depression severity. Overall, we explored the state of mental health with regards to the entire population during the pandemic year; then, we delved deeper and explored factors such as age, gender, and location in conjunction to this trend and their impact.