Assignment 4

Analysis of COVID-19 Cases in Florida

&

Vaccination rate in rural and urban areas in USA

Team K

Neeraj Sinha​

Charlotte Wade​

Eddie Hernandez​

Keerthana Darwaja​

Sarvan Vavilthota

**Business Objective​**

Since the onset of COVID-19 at the beginning of 2020, many people worldwide have contracted the virus. Although a vast number of people were able to recover, COVID-19 has proven to be extremely deadly. Several states within the U.S. have been identified as having a critically substantial number of recorded cases and deaths. The state of Florida is one of those critical states. Being able to understand all parameters and risk factors will attribute to overcoming and surviving this deadly pandemic.

*Some of the questions our group asked included*:

Did age play a factor in the number of deaths?

How did travel affect the high rate of COVID-19 cases and would that have attributed to the death rate?

Does race or gender play a factor?

Why rural areas are more infected and less vaccinated compared to the urban areas?

**Preliminary Hypotheses and Intuition**

Group K formulated several hypotheses along with rationales for why we believed our hypotheses could be proven.

***Hypothesis 1***

Black American twice as likely to catch coronavirus as compared to White Americans and therefore death rates are higher in Black Americans as compared to White Americans.

Rationale:

Black people are more likely to have preexisting conditions that make them more vulnerable to COVID-19 infection, less likely to have health insurance, and more likely to work in jobs that do not accommodate remote work.

***Hypothesis 2***

COVID-19 cases are higher within the 25 – 34 Age Group due to travel.

Rationale:

Young people, referred to as millennials, typically have more free time as well as fewer major obligations. Therefore, the thought process is that those within that age group will have more freedom to travel.

***Hypothesis 3***

In the age groups 55-64 and above, both genders are more likely to be hospitalized and die compared to age groups 25-34 and below.

Rationale:

Since the beginning of the Covid pandemic, it was known to have had a higher impact within the older age groups and preexisting conditions.

***Hypothesis 4***

Vaccination rate in urban areas are higher when compared to rural areas.

Sub Hypothesis:

Rural areas with a higher population will have more covid cases.

Rationale:

Rural areas often have a higher proportion of residents who lack health insurance and have limited access to health care facilities and are less educated

**Data Source**

<https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-case-line-data/about>

https://covidtracking.com/race/about#download-the-data

<https://www.nationalpopularvote.com/rural-states-are-almost-entirely-ignored-under-current-state-state-system>

<https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker>

<https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>

**Data Description**

The first dataset has 16 columns and 63,939 rows. We have taken this dataset from the Florida Department of Health’s website. It is a case line data of all counties of Florida from March 2020 to June 2020.

The second dataset is the CRDT dataset which has all the tracks of covid cases for various races and ethnicities from April 12, 2020, to March 7, 2021, has 5320 observations and 54 variables in total.

The third dataset for vaccination rate in urban and rural areas in the USA thatthe collected from 3 websites one is from the national popular vote to get the rural and urban area population and from the 2nd website we got Data from vaccine-tracker to get the vaccination percentage of population in USA and the last one we got from USA facts to collect the data of covid cases and deaths totally we have 52 rows of states in USA and 11 columns

**Data Preprocessing Steps**

# **Florida COVID-19 Case-Line Dataset**

The data prepossessing steps included the following steps:

1. Importing the dataset.
2. Determining the variable types to understand how each variable could be used to process the data.
3. Determine which columns contained missing data as well as the total number missing data for each row by using the ColSums processing step in R.
   1. The Age Variable contained 20 rows of missing data
   2. Origin Variable had 59,067
   3. ED Visit Variable had 2347 rows
   4. Hospitalized Variable 473 rows
   5. Contact had 5941 rows of missing date
   6. The Died column had 61,152 missing rows of data; however, it was concluded that if the row contained NA, then it was safe to assume that the patient did not die.
4. Replaced missing values
   1. Missing values in the Died column were replaced with ‘No’ by aggregating and creating a table.
   2. For the Age column, the mean of Age was calculated using na.rm to ignore the missing values - for which the mean was then used to replace the 20 missing rows of data.
5. Create plots to determine outliers and correlations.
   1. A histogram was created to show the Age plot after the missing values were replaced.
   2. A table to aggregate the numbers for Age Group was created along with a Bar Plot.
   3. A Bar Plot and Pie Chart was created for the Died column.

# **CRDT: All Race & Ethnicity Covid Data:**

1. Importing the dataset:

We used read.csv() to import the dataset. With the help of view() and dim(), we can see there are 5320 observations and 54 variables.

2. Checking the missing values:

We need to check the missing values in our target variables based on the hypothesis I.e., “Cases\_Black” and “Cases\_White”. We used is.na() to find if there are NA values. We observed that there were 3 NA values in both of our target variables which we also confirmed with the help of sum(is.na()).

3. Replacing the missing values:

We replaced the blank values with the mean of the column. We then did the final check for NA by using the view()and we observed all those blanked values have been replaced with the mean of the column.

4. Identifying outliers:

We used boxplot() to identify if there are any outliers and we observed that there are so many outliers.

5.Selecting the required variables:

We used tidyverse library to create a new dataset which has all the required columns needed for hypothesis testing.

**Vaccination RATE in urban & Rural areas**

The data prepossessing included the following steps:

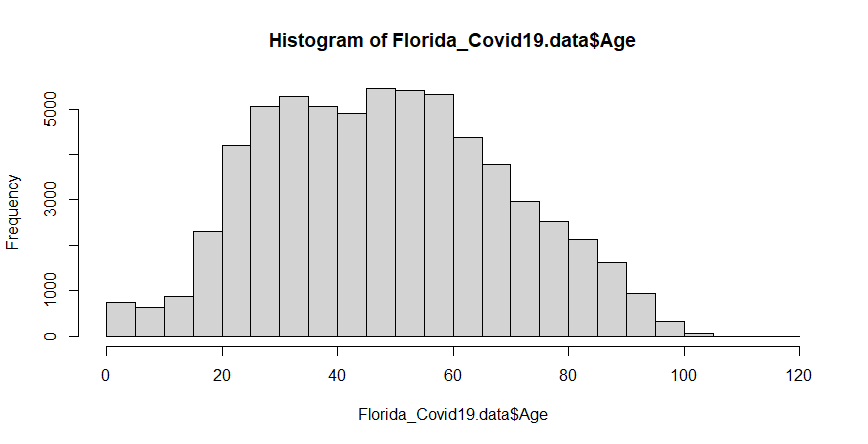
1. Importing the dataset in csv file format using read . csv()
2. Viewing the data to make sure it is in right format
3. Replacing the missing values like (spaces , /,.) with an underscore
4. Removing % and other symbols in columns and converting it to numeric column
5. Computing the average vaccination rate for rural and urban areas using group by() .
6. Using bar plot to view the vaccination rate for rural and urban areas

**Specific Approach and Analysis**

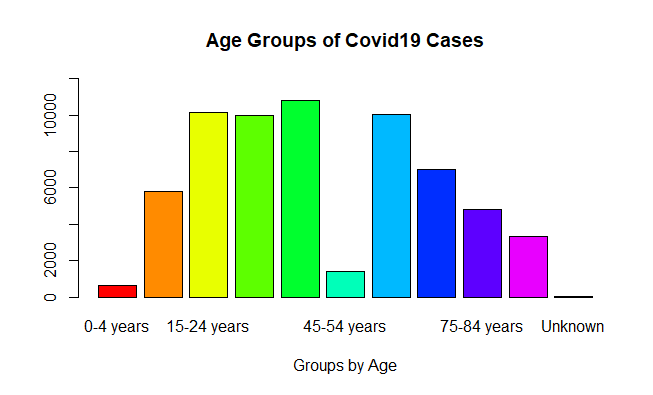
***Hypothesis 1***

***Hypothesis 2***

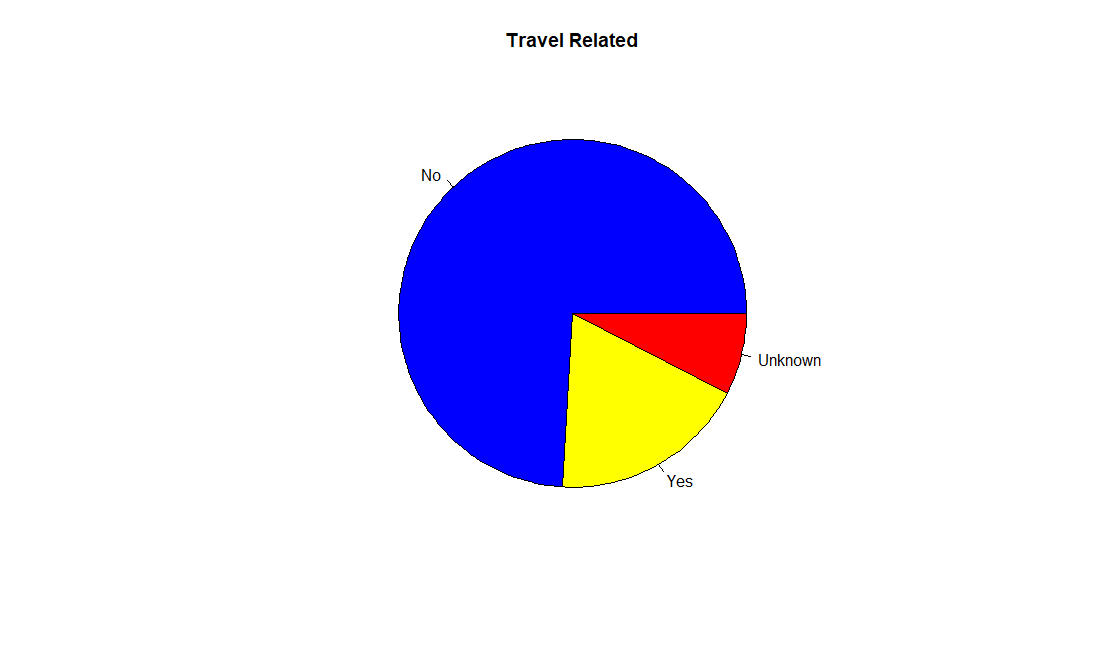
To determine if travel was a determining factor that results in a higher number of COVID-19 cases with people from the 25 – 34 Age Group, data must be analyzed relating the number of COVID-19 cases to the various age groups. Although the number of missing values within the Age column was very negligible, missing values were replaced to allow data to be attributed to the appropriate Age Group. Figure 2.a shows a Histogram of the number of COVID-19 cases according to Age after the missing numbers were replaced.

 Figure 2.a

Those numbers must then be correlated to whether the contraction of the virus was due to travel. A Bar Plot was created to show the number of COVID-19 cases according to Age Group (Figure 2.b) Though the numbers appear to very low, additional analysis must be done to replace the unknown category for Age Group.

 Figure 2.b

When performing ColSum, the Travel-Related column did not indicate there were missing values. However, there are some rows that show a nominal output as ‘Unknown’. A Pie Chart was created to show that output. (See Figure 2.c.) Additional processing will be performed to change the ‘Unknown’ output.

 Figure 2.c

More preprocessing, including filtering functions using dplyr, will be done to compare Travel-Related numbers to Age Group.

***Hypothesis 3***

The initial approach to determine if older age groups have more deaths is to find all deaths by age group. After that, we can draw preliminary conclusions if the data can clearly show older age groups are affected the most. To further prove or disprove, we’ll run functions in r that will generate accurate and useful information.

We found a very helpful function, tabyl, that can show us if they died or not. Additional packages like ggplot2 and dplyr will be installed to run certain functions. With that, we can provide better comparisons in deaths in all age groups. In Florida, their governor was really against mask mandates and vaccinations, so we believe this had an affect leading to more cases and more deaths in the state. We hope to find our conclusion to be true.

***Hypothesis 4***

The initial approach is to show that the urban areas are more vaccinated compared to rural areas for that we have done an average of both the rural areas of vaccination rate in states and urban area vaccination rate in usa states to prove our hypothesis that rural areas are less vaccinated and more infected compared to urban area of usa for that we used groupby() and a library(tidyverse) and we replaced all the missing values with underscore and a bar plot was created to show the visualization of rural and urban area vaccination percentage and we got this bar plot from the dataset which I choose

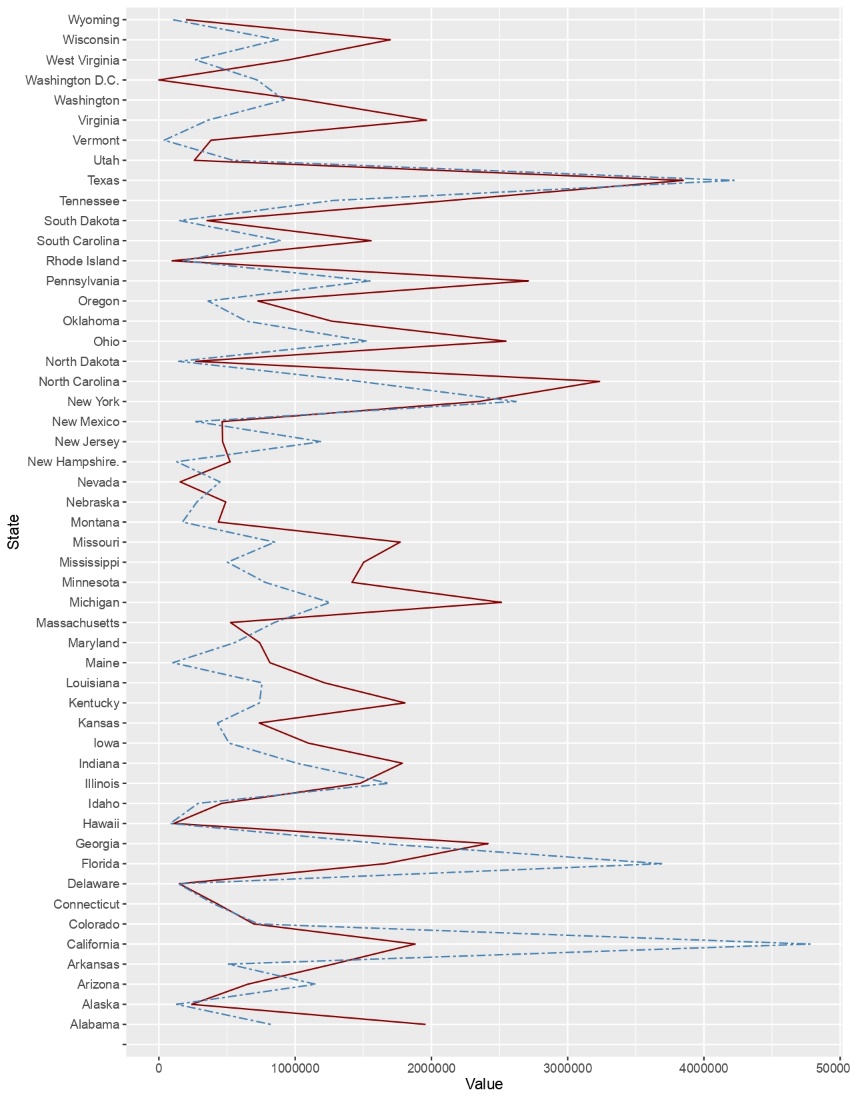
BAR PLOT FOR VACCINATION RATE IN RURAL & URBAN AREAS

Chart, bar chart

Description automatically generated

***Sub Hypothesis 4***

This approach is to show how Rural areas are most affected by Covid virus. All the values given were accurate and there were very fewer missing values. To show that Rural areas with higher population have more Covid cases, we read the data from the dataset and drew a line graph by using ggplot. We used two columns, Rural population, and Total cases to get the line chart. The image below proves this.



Specific Outcomes and Insights

*Hypothesis 1*

*Hypothesis 2*

When viewing the initial completed charts, the highest number of COVID-19 related cases fall within the ‘15 - 24’ and ‘35 - 44’ year old Age Groups. The Pie Chart indicates that the largest number of COVID-19 related cases are not Travel Related.

*Hypothesis 3*

The initial approach to finding deaths by age groups was to sort data to see what age group is affected the most. At first glance, it seems like deaths were not too high across all age groups. Figure 3.a below shows that the total deaths were relatively small compared to cases that did not lead to death. This comes from having so many more cases than actual deaths, therefore it diminishes the fact that older people are dying, while younger people are not.

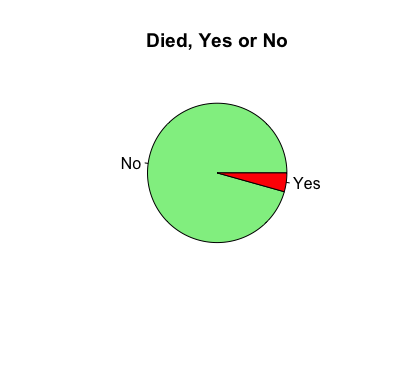
Figure 3.a

Figure 3.b below shows deaths in older age groups by far exceed the lower age groups. Additionally, the data shows there is more covid cases that did not lead to death for the younger age groups.

Figure 3.b

***Hypothesis 4***

After getting the results from the bar plot results is like rural areas are more vaccinated we got according to our dataset compared to urban area

*Sub Hypothesis*

*If we look at the Line graph, we can observe that Rural areas with a larger population are much effected by Covid than less populated areas.*

**Summary and Conclusions**

*Hypothesis 1*

*Hypothesis 2*

The preliminary analysis of the output charts indicates that more people within the 25 – 34-year old Age Group have not contracted COVID-19 due to travel primarily because the Travel-Related output shows that most of the cases are not Travel Related.

*Hypothesis 3*

Our preliminary conclusion is that Hypothesis 3 is correct because our data output in r demonstrates age groups 55 and above have more deaths even though the younger age groups had many more cases. Covid is more deadly for the older generation compared to younger ones.

Even though cases were not as high compared to younger age groups, it could result from older age groups taking more caution while being in public areas. Being against mask mandates, vaccines, and safety measures in the state, could have also had an impact on death rates among older age groups.

*Hypothesis 4*

Usually, Urban areas are more populated, and they have much vaccinated percentage, whereas in rural areas vaccinated percentage is low and if rural areas are heavily populated there is a chance of getting infected easily it can be because of no awareness about vaccinations in such places.

Key Learning

References:

<https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-case-line-data/about>

<https://www.cidrap.umn.edu/news-perspective/2020/08/us-blacks-3-times-more-likely-whites-get-covid-19>

<https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-case-line-data/about>

https://covidtracking.com/race/about#download-the-data

<https://www.nationalpopularvote.com/rural-states-are-almost-entirely-ignored-under-current-state-state-system>

<https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker>

<https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>