Abinesh Senthil Kumar (#50320934)

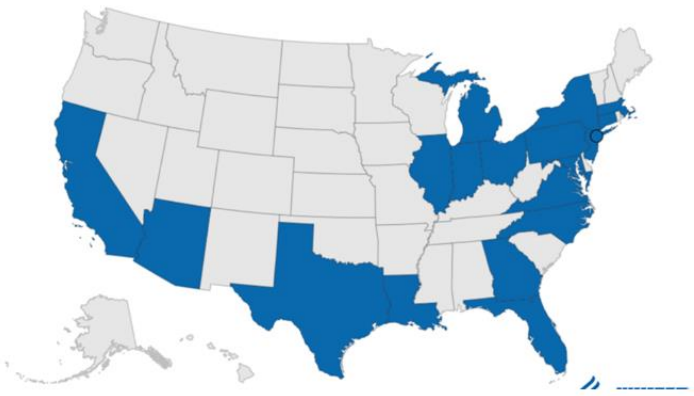
Proof of Work – Individual problem (Fall 2020)

This document serves as proof of my work in analyzing survey data to understand physician mental health working as a frontline worker during COVID pandemic.

Introduction

The study aims to evaluate the risk factors that healthcare workers face while working as a frontline worker to treat COVID-19 patients. A survey is designed and collected on behalf of American Medical Association (AMA) and it has been sent out to top 18 states that has the highest infection count in USA. The survey captures information on responders that help us assess demographics and different mental health conditions including but not limited to Post Traumatic Stress Disorder (PTSD), Resilience, work changes, Burnout etc.

States captured in the survey:



- New York
- California
- New Jersey
- Illinois
- Texas
- Massachusetts
- Florida
- Pennsylvania
- Michigan
- Georgia
- Maryland
- Virginia
- North Carolina
- Arizona
- Louisiana
- Connecticut
- Ohio
- Indiana

Instructions to use the code:

Run the "data_wrangling.R" code fully first before running "plots_code.R" code.

- data_wrangling.R – contains all the data wrangling operations performed including manual functions.
- Plots_code.R – contains all the codes for plots and tables produced.

Dataset,

- The initial dataset has 1467 samples and 253 variables.
- The analysis was performed only on people who are not retired and people who said "yes" to the consent.
- After the filtering, the new dataset has 1223 samples and all the analysis were carried out on this dataset.

Packages used

- **Readr** – read excel files
- **Plyr** – Data wrangling
- **Dplyr -** Data wrangling
- **Tidyverse -** Data wrangling
- **reshape2 -** Data wrangling
- **writexl** – Write excel files
- **hrbrthemes** – theme for ggplot
- **Hmisc -** Data wrangling
- **Lubridate** – Dealing with date and time values
- **BBmisc** – data normalization techniques
- **Tidyr -** Data wrangling
- **zipcodeR** – contains entire zip code data in USA
- **ggplot2** – Creating graphs and plotting
- **scales** – aesthetics for ggplot, determine labels and breaks
- **gridExtra** – combining multip le plots
- **grid -** combining multiple plots
- **table1** – create html tables
- **usmap** – plotting values in USA map
- **fmsb** – radar chart

Manual functions:

Density plot for groups:

The function "indivgroupdensityplot" takes seven inputs and gives out a density plot with groups and annotations. Legend is an optional input which can be moved as per requirement. The columns "group" and "pts" are extracted from the dataset, aggregation is done on the extracted values before saving to "meanbptsd". A data frame named "annot" is created and assigned x,y axis to "meanbptsd" followed by rounding of values to second decimal. To visualize, ggplot is used with each category taking different colors, density plot is produced by geom_density. Text and vertical dashed lines are added using syntax geom_vline, geom_text.

Input arguments:

- **Dataset** – dataset containing the required columns
- **Points** – continuous variable to be plotted in the density plot
- **Groups** – categorical variable to be plotted
- **Annotx, Annoty** – corresponding x,y axis value annotations for mean/median
- **Titled** – title of the plot to be given
- **Legend_pos** (optional argument) – optional argument given for the position of the legend.

Revert one hot encoded value:

The function "revert_onehot_with_id" takes only one input and stacks column one below another. The column of the data set is changed to "ID". A new function "sum_of_missing" is created to count the NA values. Using condition, the missing values are checked row wise, compared with the column of the original dataset and finally the result is saved to dataframe named "dset1missing_index". The column name is changed to "ID". Exception handling is used to overcome warnings/ errors. Syntax "melt" is used to stack each column one below another. Rbind command is used to bind rows together and is returned along the number of original missing values.

Input arguments:

- **Dset** – dataset containing the one hot encoded columns plus the id variable in the end.

Data wrangling

Directly worked with COVID patients (Q2.1)

The levels of directly worked with COVID variable are changed to "directly worked" and "not directly worked" from "yes" and "No"

Age (Q5.1)

Age is calculated from year of birth question and data cleaning is done from misspelt numbers in date of birth variable

Medical speciality (Q5.7)

Physician medical specialty is captured by 40 different questions individually, it is reversed into single column using manual function written before.

Physician Race (Q5.4)

Physician race is captured by 6 different questions individually, it is reversed into single column using manual function written before.

Average working hours before and after COVID (Q2.6) (Q2.7.1)

Average working hours per week before and during COVID variables are captured by two different variables. They are cleaned individually for redundant values.

A new function is written that splits each value by "- " and takes the average value of the two splitted values. This is done because few people have given the number of hours in a range, i.e 40-50 hrs so we split and take the average of 40 and 50 to determine them as 45.

This process is done for both the variables average work hours before and during COVID.

Type of hospital (Q5.11.1)

Type of hospital is captured by 15 questions individually, it is reversed into single column using manual function written before.

Work setting within the hospital (Q5.11.2)

Work setting within the hospital is captured by 8 questions individually, it is reversed into single column using manual function written before

Organizational support (Q4.1)

Organizational support has 8 individual questions in it with answers as strings, it is changed to,

- “Strongly disagree" = 0
- "Moderately disagree" = 1
- "Slightly disagree" = 2
- "Neither agree nor disagree" = 3
- "Slightly agree" = 4
- "Moderately agree" = 5
- "Strongly agree" = 6

Final score is calculated by summing all the above 8 questions.

Identifying ways that you have been coping with the stress in your life (Q4.4)

Identifying ways to cope with covid Pandemic is assessed by set of 28 questions. The responses ar changed to,

- "I haven't been doing this at all" = 1
- "I've been doing this a little bit" = 2
- "I've been doing this in medium amount" = 3
- "I've been doing this a lot" = 4

Individual coping attributes are found by summing specific questions.

- Self-distraction – 1st question + 19th question
- Active coping – 2nd question + 7th question
- Denial – 3rd question + 8th question
- Substance use – 4th question + 11th question
- Emotional support – 5th question + 15th question
- Instrumental support – 10th question + 23rd question
- Disengagement – 6th question + 16th question
- Venting – 9th question + 21st question
- Reframing – 12th question + 17th question
- Planning – 14th question + 25th question
- Humor – 18th question + 28th question

- Acceptance – 20th question + 24th question
- Religion – 22nd question + 27th question
- Self-blame – 13th question + 26th question

PTSD – Post Traumatic stress disorder

For all the PTSD (PCL-5) variables (20 questions) the answers were change to,

- Not at all – 0
- A little bit – 1
- Moderately – 2
- Quite a bit – 3
- Extremely – 4

The final score for PTSD variables is created by summing up all the 20 variables.

PTSD criterion

Different criterions/clusters for PTSD are calculated individually,

- Criterion B – 1st question + 2nd question + 3rd question + 4th question + 5th question
- Criterion C – 6th question + 7th question
- Criterion D – 8th question + 9th question + 10th question + 11th question + 12th question + 13th question + 14th question
- Criterion E – 15th question + 16th question + 17th question + 18th question + 19th question + 20th question

PTSD final classification:

First the initial DSM-5 criteria is checked as follows,

- Criterion B – at least one question should be having a value greater than 2
- Criterion C – at least one question should be having a value greater than 2
- Criterion D – at least two questions should be having a value greater than 2
- Criterion E – at least two questions should be having a value greater than 2

After checking for initial DSM conditions then the classifications are given by,

Probable PTSD

- Satisfying any 3 or all 4 DSM-5 criteria & PTSD checklist score >=33

Subclinical

- Satisfying 2 DSM-5 criteria & PTSD checklist score >=33
- Satisfying 2 or more DSM-5 criteria & PTSD score >=12 & <33

Pre-Subclinical

- Satisfying 1 DSM-5 criteria & PTSD checklist score >=12 & <33
- Satisfying 1 DSM-5 criteria & PTSD checklist score >33
- Satisfying no criteria but score >=12 & <=21

No PTSD

- Any score <12 irrespective of DSM-5 criteria

Depression:

For all the depression (PHQ-9) variables (9 questions) the answers were changed to,

- Not at all – 0
- Several days – 1
- More than half the days – 2
- Nearly every day – 3

The final depression scores are calculated by summing up all the 9 variables.

The final scores are then again categorized into different levels such as,

- 0 to 5 = 'None-Minimal depression',
- 5 to 10 = 'Mild depression',
- 10 to 15 = 'Moderate depression',
- 15 to 20 = 'Moderately Severe depression',
- 20 to 27 = 'Severe depression'

Resilience:

For all the resilience (CDRISC10) variables (10 questions) the responses were changed to,

- Not true at all – 0
- Barely true – 1
- Sometimes true – 2
- Often true – 3
- True nearly all the time – 4

The final resilience scores were calculated by summing up all the 10 scores.

Along with the final scores, the individual hardiness values are calculated by summing up specific questions.

- flexibility – 1st question + 5th question.
- sense of self efficacy – 2nd question + 4th question + 9th question.
- regulate emotion – 10th question.
- optimism – 3rd question + 6th question + 8th question.
- cognitive focus – 7th question.

Zip code and counties to infection and death rate

- Redundant information is corrected in the zip code column, few of them have entered invalid zip codes.
- A package named **ZipcodeR** has all the available zip codes in the US and their corresponding state and county information.
- We then match our zip code to the entire zip code database to find the counties and state information.
- Infection and death rate for USA counties is obtained from 'https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv' – NYTimes open data. It has date, county, state, infection and death rates.
- Group by operation is done to get the current date's infection and death rate.
- Our zip code with county and state name is then matched with the NYTimes data to get the final output.

Plots

All plots for the analysis are created using ggplot in R

- To compare categorical with categorical variable, **side-by-side stacked bar plots** are used
- To compare categorical with numerical variable, **Density plots** and **violin plots** showing mean/median are used
- To compare three or more continuous variables, **radar chart** is used by normalizing and comparing the mean/median of them
- To summarize large categories, an HTML table is used. It is created by function **table1** from **table1** package.
- To combine multiple plots together, a function called **grid.arrange** is used from **gridExtra** package

Radar charts

- A radar chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point. – Wikipedia
- In order to display multiple dimensions, we'll have to normalize all the required variables to a common scale.

Normalization formula

$$x''' = (b - a)\frac{x - \min x}{\max x - \min x} + a$$

Where,

- a = lower limit in required range [a,b].
- b = upper limit in required range [a,b].
- x = number to be normalized.
- min x = minimum number from the non-normalized set of values.
- max x = maximum number from the non-normalized set of values.
- X^m = normalized value.
- Then the average (mean) scores for each of the respective dimensions are calculated and plotted in the radar chart.