Covid-19 Project

Theodore Koby-Hercsky

5/15/2021

html_document: https://rpubs.com/theoKoby/770127

Working directory set to DSC520 file

setwd("~/Documents/Bellevue University Classes/DSC520/Final Project")

Covid-19 Vaccination and Death Rate Introduction

The topic I will be researching and addressing is the vaccination rate for the Pfizer, Moderna, and Johnson and Johnson vaccines to determine which companies' vaccine is being taken and or accessible in each state. While also researching into death rates of Covid-19 by using variable such as sex, age, race, education, and the state the individual was living in. These statistics will help shine some light on the current pandemic and see the demographics on the current death rates and what vaccines are being taken and the rate of vaccination in each state. While also determining who is currently dying from the virus. As I feel as if this is a topic that has affected everyone in the world to some extent over the past year and can help educate individuals on which vaccine has been distributed in their state and the United States the most and what the death rate is in regard to these variables. As the current problem we are dealing with is who is currently at the highest risk of death and should you be getting a vaccine and if so which one should you get? In our findings we will see that this topic can be seen as a data science problem as we have mass amounts of data that can be sifted and analyzed through the use of R Studios to help our current problem.

Research questions

1. Out of the three Pfizer, Moderna, and Johnson and Johnson vaccine for the week of 05/10/2021 which company has the highest number of vaccinations administered?

a. In regard to the number of vaccinations administered which state had the highest rate for each company?

b. In regard to the number of vaccinations administered which state had the lowest rate for each company?

c. View(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Pfizer)

d. View(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Moderna)

e. View(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Janssen)

2. In the time period of 01/01/2020 to 01/30/2021 what percentage of deaths reported were solely due to Covid-19?

a. When it comes to age and sex which combination had the highest death rate for each sex and age group?

b. As most think the elderly population is the highest at risk for death from the virus but what percentage of deaths in the elderly population was from Covid-19 and what percentage for other illnesses besides Covid-19?

c. View(AH_Provisional_COVID_19_Deaths_by_Educational_Attainment_Race_Sex_and_Age)

3. When it comes to race which ethnicity did, we find to have the highest death rate from Covid-19 from the time period 01/01/2020 to 01/30/2021?

a. Did the education level have an effect on the death rate of different ethnicities?

b. View(AH_Provisional_COVID_19_Deaths_by_Race_and_Educational_Attainment)

4. In the time period of 01/01/2020 to 04/30/2021 did we see an increase in Covid-19 deaths as the age of the individuals rise? If so, did sex have an effect on the number of deaths?

a. View(Provisional_COVID_19_Deaths_Counts_by_Age_in_Years)

5. In regard to Covid-19 deaths how many are solely related to Covid-19 from the period of 01/01/2020 to 05/08/2021? On another note, what other illnesses had a large mortality rate in this time period and where these illnesses combined with Covid-19 deaths?

a. View(Provisional_COVID_19_Death_Counts_by_Sex_Age_and_State)

Provide a concise explanation of how you plan to address this problem statement.

I will be addressing this problem statement on which vaccines currently are being received the most and what variables are having the highest rate of death due to Covid-19. I will filter my new data files and create tables and charts to help show the outcome of variables such as age, sex, race, and state living in to determine the highest mortality rate. I will also be researching into the three vaccinations that are currently being used known as Pfizer, Moderna, and Johnson and Johnson vaccine that we can determine the sum of all vaccines received for each company and which states are currently having the highest rate of vaccination.

Approach

The estimated known cases of the COVID-19 virus in the United States have reached an estimate of 32.9 million cases in accordance Google statistics. As the estimated death toll in the united states from the COVID-19 virus hitting 584,000 deaths. Which means just from these statistics alone not taking into account any other variable we see that you would have a 1.78% of dying from the virus. This does not seem so bad but let's do some research and see if sex, race, education, and or age has any direct correlation with the chances of an individual dying from the virus. Also, you might be wondering should I get a shot and if yes which one should I get? Well, we are going to look at some data on the Pfizer, Moderna, and Johnson and Johnson vaccine to determine which vaccine has been most widely available in each state and if the country and or state is favoring a certain company shot over another.

Discuss how your proposed approach will address (fully or partially) this problem.

My approach will help show who is at the highest risk of death and who should get the vaccine if they haven't already done so. While also determining if variables such as age, sex, and or race have any effect on the death rate and if so, who is the most vulnerable. We will be able to see if individuals that a lower education level have are more impacted than others that have college degrees. Another way my approach will address my problem will be seen in the data on vaccines administered as we can determine which states have the highest rate of vaccination and which vaccine is used the most. This will show our viewers which vaccines are readily available in their state and which vaccine most are getting.

Covid-19 Data Sets

```
## I am importing readr from the library so I can use the read_csv function
to create my student survey data frame.
library(readxl)
library(readr)

## Vaccine administered for Pfizer by state by week
COVID_19_Vaccine_Distribution_Pfizer <- read_excel("Updated Data Final
Project/COVID-19_Vaccine_Distribution_Pfizer.xlsx")
View(COVID_19_Vaccine_Distribution_Pfizer)
head(COVID_19_Vaccine_Distribution_Pfizer)

## # A tibble: 6 x 4
##   Company Jurisdiction `Week of Allocations` `1st Dose Allocations`
##   <chr>   <chr>        <dttm>                                  <dbl>
## 1 Pfizer  Oklahoma     2021-04-05 00:00:00                     56400
## 2 Pfizer  Texas        2021-04-05 00:00:00                    392100
## 3 Pfizer  Iowa         2021-04-05 00:00:00                     45800
## 4 Pfizer  Kansas       2021-04-05 00:00:00                     41800
## 5 Pfizer  Missouri     2021-04-05 00:00:00                     89600
## 6 Pfizer  Nebraska     2021-04-05 00:00:00                     27600

## Provisional_COVID-19_Death_Counts_by_Sex__Age__and_State data frame Covid-
19 data.
Provisional_COVID_19_Death_Counts_by_Sex_Age_and_State <- read_csv("CDC Covid
Data/Provisional_COVID-19_Death_Counts_by_Sex__Age__and_State.csv")
View(Provisional_COVID_19_Death_Counts_by_Sex_Age_and_State)
head(Provisional_COVID_19_Death_Counts_by_Sex_Age_and_State)

## # A tibble: 6 x 16
##   `Data As Of` `Start Date` `End Date` Group Year  Month State Sex   `Age
Group`
##   <chr>        <chr>        <chr>      <chr> <lgl> <lgl> <chr> <chr> <chr>
## 1 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … All
Ages
## 2 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … Under
1 ye…
## 3 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … 0-17
years
## 4 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … 1-4
years
## 5 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … 5-14
years
## 6 05/12/2021   01/01/2020   05/08/2021 By T… NA    NA    Unit… All … 15-24
years
## # … with 7 more variables: COVID-19 Deaths <dbl>, Total Deaths <dbl>,
## #   Pneumonia Deaths <dbl>, Pneumonia and COVID-19 Deaths <dbl>,
## #   Influenza Deaths <dbl>, Pneumonia, Influenza, or COVID-19 Deaths
<dbl>,
## #   Footnote <chr>
```

```
## Covid-19 Deaths by race, sex, and age
AH_Provisional_COVID_19_Deaths_by_Educational_Attainment_Race_Sex_and_Age <-
read_csv("CDC Covid Data/AH_Provisional_COVID-
19_Deaths_by_Educational_Attainment__Race__Sex__and_Age.csv")
View(AH_Provisional_COVID_19_Deaths_by_Educational_Attainment_Race_Sex_and_Ag
e)
head(AH_Provisional_COVID_19_Deaths_by_Educational_Attainment_Race_Sex_and_Ag
e)

## # A tibble: 6 x 9
##   `Data as of` `Start Date` `End Date` `Education Level`  `Race or Hispan…
Sex
##   <chr>        <chr>        <chr>      <chr>              <chr>
<chr>
## 1 02/03/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 2 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 3 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 4 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Fema…
## 5 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Male
## 6 02/02/2021   01/01/2020   01/30/2021 Associate degree … Hispanic
Male
## # … with 3 more variables: Age Group <chr>, COVID-19 Deaths <dbl>,
## #   Total Deaths <dbl>

## Covid-19 Deaths by education and race
AH_Provisional_COVID_19_Deaths_by_Race_and_Educational_Attainment <-
read_csv("CDC Covid Data/AH_Provisional_COVID-
19_Deaths_by_Race_and_Educational_Attainment.csv")
View(AH_Provisional_COVID_19_Deaths_by_Race_and_Educational_Attainment)
head(AH_Provisional_COVID_19_Deaths_by_Race_and_Educational_Attainment)

## # A tibble: 6 x 7
##   `Data as of` `Start Date` `End Date` `Education Level` `Race or Hispanic
Orig…
##   <chr>        <chr>        <chr>      <chr>             <chr>
## 1 02/01/2021   01/01/2020   01/30/2021 8th grade or less Hispanic
## 2 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
American …
## 3 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
Asian
## 4 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
Black
## 5 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic More
than…
## 6 02/01/2021   01/01/2020   01/30/2021 8th grade or less Non-Hispanic
```

```
Native Ha…
## # … with 2 more variables: COVID-19 Deaths <dbl>, Total Deaths <dbl>

## Covid-19 Deaths by age in years
Provisional_COVID_19_Deaths_Counts_by_Age_in_Years <- read_csv("CDC Covid
Data/Provisional_COVID-19_Deaths_Counts_by_Age_in_Years.csv")
View(Provisional_COVID_19_Deaths_Counts_by_Age_in_Years)
head(Provisional_COVID_19_Deaths_Counts_by_Age_in_Years)

## # A tibble: 6 x 7
##   `Data as of` `Start Date` `End Date` Sex   `Age Years` `Total deaths`
##   <chr>        <chr>        <chr>      <chr> <chr>                <dbl>
## 1 05/10/2021   01/01/2020   04/30/2021 Male  <1 year              13525
## 2 05/10/2021   01/01/2020   04/30/2021 Male  01 Years               949
## 3 05/10/2021   01/01/2020   04/30/2021 Male  02 Years               635
## 4 05/10/2021   01/01/2020   04/30/2021 Male  03 Years               515
## 5 05/10/2021   01/01/2020   04/30/2021 Male  04 Years               406
## 6 05/10/2021   01/01/2020   04/30/2021 Male  05 Years               381
## # … with 1 more variable: COVID-19 Deaths <dbl>

## Vaccine administered for Moderna by state for the week of 05/10/2021
COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Moderna <-
read_csv("CDC Covid Data/COVID-
19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Moderna.csv")
View(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Moderna)
head(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Moderna)

## # A tibble: 6 x 4
##   Jurisdiction  `Week of Allocation… `1st Dose Allocation… `2nd Dose
Allocation…
##   <chr>         <chr>                                <dbl>
<dbl>
## 1 Connecticut   05/10/2021                           41300
41300
## 2 Maine         05/10/2021                           15800
15800
## 3 Massachusetts 05/10/2021                           79500
79500
## 4 New Hampshire 05/10/2021                           15900
15900
## 5 Rhode Island  05/10/2021                           12400
12400
## 6 Vermont       05/10/2021                            7500
7500

## Vaccine administered for Johnson and Johnson by state for the week of
05/10/2021
COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Janssen <-
read_csv("CDC Covid Data/COVID-
19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Janssen.csv")
```

```
View(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Janssen)
head(COVID_19_Vaccine_Distribution_Allocations_by_Jurisdiction_Janssen)

## # A tibble: 6 x 3
##   Jurisdiction  `Week of Allocations` `1st Dose Allocations`
##   <chr>         <chr>                                 <dbl>
## 1 Connecticut   05/10/2021                             6400
## 2 Maine         05/10/2021                             2500
## 3 Massachusetts 05/10/2021                            12300
## 4 New Hampshire 05/10/2021                             2500
## 5 Rhode Island  05/10/2021                             2000
## 6 Vermont       05/10/2021                             1200
```

Original source where the data was obtained is cited and, if possible, hyperlinked.

The original source of all my data that is hyper linked below comes directly from the CDC.

Provisional_COVID-19_Death_Counts_by_Sex__Age__and_State

COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Pfizer

COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Moderna

COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Janssen

AH_Provisional_COVID-19_Deaths_by_Educational_Attainment__Race__Sex__and_Age

AH_Provisional_COVID-19_Deaths_by_Race_and_Educational_Attainment

Provisional_COVID-19_Deaths_Counts_by_Age_in_Years

Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

The original purpose of this data was to keep track of the current vaccination and illness rate of the COVID-19 virus along with the death rate.

The dataset titled Provisional_COVID-19_Death_Counts_by_Sex__Age__and_State started being collected on May 1, 2020 with 15 variables with year, month, and foot note variables returning a NA in their fields, but this dataset holds mass amounts or data that is very useful

The dataset titled COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Pfizer started being collected on December 14, 2020 with 4 variables with the last variable second dose providing the same info for 1st dose.

The dataset titled COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Moderna started being collected on December 19, 2020 with 4 variables with the last variable second dose providing the same info for 1st dose.

The dataset titled COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Janssen started being collected on February 26, 2021 with 4 variables with the last variable second dose providing the same info for 1st dose.

The dataset titled AH_Provisional_COVID-19_Deaths_by_Educational_Attainment__Race__Sex__and_Age started being collected on February 3, 2021 with 9 variables with all variables seeming very useful at the moment.

The dataset titled AH_Provisional_COVID-19_Deaths_by_Race_and_Educational_Attainment started being collected on February 1, 2021 with 7 variables with all variables seeming very useful as we can see the education level and race deaths from the COVID-19 virus and other deaths not related to the virus.

The dataset titled Provisional_COVID-19_Deaths_Counts_by_Age_in_Years started being collected on August 7, 2020 with 7 variables with all variables seeming very useful as we can see the sex and age of the deaths from COVID-19.

Identify the packages that are needed for your project.

Required Packages

rmarkdown is a package I will need as it will allow me to make an Rmarkdown report.

tidyverse is a package that includes other packages that you're likely to use in everyday data analyses. These packages consist of ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcats.

The ggplot2 package will be needed as it creates graphics, based on The Grammar of Graphics from the data we provide and tell ggplot2 how to map variables to aesthetics, what graphical primitives to use.

The dplyr package will be needed as it provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges.

The tidyr package will also be needed as this helps us tidy our data in a consistent form by having every variable go in a column, and every column being a variable.

The readr package is needed as this provides a fast and friendly way to read rectangular data like csv, tsv, and fwf files.

The ppcor package is needed as this will be used to create a partial correlation from the cor() function we use to correlate different variables such as age and death rate.

The effects package is needed as this will be used to create displays for linear, generalized linear, and other models that will be needed to show the outputs of our predictions. By using Graphical and tabular effect displays to interact with various statistical models with linear predictors.

The readxl package will be used to help pull data into R from gdata, xlsx, xlsReadWrite Excel files.

The statmod package consist of algorithms and functions to aid statistical modeling which will come in handy when we are trying to run a comparison on the growth curve or check on mixed linear models.

We can also use the reshape2 package to help melt and form the data into what we need. While also using the scale package to bring in some extra plotting features that can be used when creating plot for our correlation models.

The Knitr package will allow us to use the kable function that will help create nice looking tables that are adapted to the type of output document to show to viewers.

What types of plots and tables will help you to illustrate the findings to your research questions?

We can use a scatterplot matrix to compare variables such as age and race to the death rate of individuals that have died from Covid-19.

Create a table that will show the correlations between different variables so we can make it easier for the viewers to see and use in plots

Create a heatmap in regard to the correlation between our different variable and the rate of death from the Virus. This will show us visually how the virus is correlating to variables such as the populations education level, their age, and even their sex.

We can create a Histogram from a two-sample T-test that compares males and females and the death rate of the COVID-19 virus which will determine if the distribution on either variable can be seen as normal.

Create a scatterplot that compares two variables such as age and the death rate of the Virus.

I can create a table that combines all three companies' vaccines together to make it easier to access and use in a plot.

Use the line plot to show separate line for each state as our x axis will be equal to week of allocation, while Y axis will be equal to dose allocation

What do you not know how to do right now that you need to learn to answer your research questions?

Questions for future steps

My main issue I cannot seem to fix is pulling my data in my Rmarkdown file when I knit it. The data sets show up on a seperate pop up so I had to use the Head function to show the top first couple lines of my data sets. Is there a way to see my whole data set or is this not possible with the amount of data each data set contains?

At the moment I do not know or remember how to create a ggplot out of multiple data frames at once. If this is not doable, I will try and combine multiple data frames together so I can plot them together.

Another issue I am running into is updating my data sets to take out unneeded words such as "80 years" as it can just be 80 for the age. How can I remove this in R or do I need to do this in Excel?

In regard to data sets if I have cells inside a variable that are either blank or NA how can I get rid of those or do I need to keep them?