FinalPT2

Marc Riley

5/24/2021

Probelm

Now that Covid-19 has been prevalant in the United States for over a year I would like to determine which groups are most at risk. Hopefully this project will help people understand how concerned they need to be with contracting this disease. This can be accomplished by looking at data sets provid by the CDC that shows deaths by age group, race, and location.

Potential questions

1. What age group needs to be most cautios?
2. What gender is more suseptable?
3. Are you at higher risk of getting covid in certain states?

Data sets

- https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg hcku/data
- https://covid.cdc.gov/covid-data-tracker/#vaccination-demographics-trends
- https://covid.cdc.gov/covid-data-tracker/#variant-proportions

Variables

1. Age
2. Sex
3. Location
4. Deaths
5. Time

Packages needed

ggplot/tidyvers, psych, readxl, purr, dplyr

Plots/Tables

Historgrams wiht trend lines, Discard, Cbind, Scatterplots

Step 2

How to import/Clean data

All three files have been imported as csv file and changed to an excel file. There are many variables that need to be split or removed because they are unnecessary for my analysis.

What does the final data set look like?

-Adjustments are still being made to the data set as of now the data set shows age broken down by key years, states, covid deaths, flu deaths, gender, and time frame.

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.0.5

Covid_df <- read_excel("CovidData.xlsx")
summary(Covid_df)

##    Data As Of           Start Date
##  Min.   :2021-05-19   Min.   :2020-01-01 00:00:00
##  1st Qu.:2021-05-19   1st Qu.:2020-03-24 06:00:00
##  Median :2021-05-19   Median :2020-08-16 12:00:00
##  Mean   :2021-05-19   Mean   :2020-08-13 04:48:00
##  3rd Qu.:2021-05-19   3rd Qu.:2021-01-01 00:00:00
##  Max.   :2021-05-19   Max.   :2021-05-01 00:00:00
##
##     End Date                      Group              Year          Month
##  Min.   :2020-01-31 00:00:00   Length:55080       Mode:logical
Mode:logical
##  1st Qu.:2020-06-22 12:00:00   Class :character   TRUE:52326
TRUE:46818
##  Median :2020-11-15 00:00:00   Mode  :character   NA's:2754      NA's:8262
##  Mean   :2020-10-26 08:24:00
##  3rd Qu.:2021-03-07 18:00:00
##  Max.   :2021-05-15 00:00:00
##
##     State               Sex             Age Group         COVID-19 Deaths
##  Length:55080       Length:55080       Length:55080       Min.   :     0.0
##  Class :character   Class :character   Class :character   1st Qu.:     0.0
##  Mode  :character   Mode  :character   Mode  :character   Median :     0.0
##                                                           Mean   :   362.6
##                                                           3rd Qu.:    60.0
##                                                           Max.   :574045.0
##                                                           NA's   :13303
##   Total Deaths     Pneumonia Deaths   Pneumonia and COVID-19 Deaths
##  Min.   :      0   Min.   :     0.0   Min.   :     0.0
##  1st Qu.:     41   1st Qu.:     0.0   1st Qu.:     0.0
##  Median :    151   Median :    16.0   Median :     0.0
##  Mean   :   2640   Mean   :   348.1   Mean   :   176.3
```

```
##  3rd Qu.:    686   3rd Qu.:    81.0   3rd Qu.:    27.0
##  Max.   :4554041   Max.   :501850.0   Max.   :281084.0
##  NA's   :8149      NA's   :16803      NA's   :12865
##  Influenza Deaths   Pneumonia, Influenza, or COVID-19 Deaths   Footnote
##  Min.   :   0.000   Min.   :     0.0                          Length:55080
##  1st Qu.:   0.000   1st Qu.:     0.0                          Class
:character
##  Median :   0.000   Median :    24.0                          Mode
:character
##  Mean   :   5.215   Mean   :   550.4
##  3rd Qu.:   0.000   3rd Qu.:   123.0
##  Max.   :9129.000   Max.   :802666.0
##  NA's   :10139      NA's   :16462

str(Covid_df)

## tibble [55,080 x 16] (S3: tbl_df/tbl/data.frame)
##  $ Data As Of                               : POSIXct[1:55080], format:
"2021-05-19" "2021-05-19" ...
##  $ Start Date                               : POSIXct[1:55080], format:
"2020-01-01" "2020-01-01" ...
##  $ End Date                                 : POSIXct[1:55080], format:
"2021-05-15" "2021-05-15" ...
##  $ Group                                    : chr [1:55080] "By Total" "By
Total" "By Total" "By Total" ...
##  $ Year                                     : logi [1:55080] NA NA NA NA NA
NA ...
##  $ Month                                    : logi [1:55080] NA NA NA NA NA
NA ...
##  $ State                                    : chr [1:55080] "United States"
"United States" "United States" "United States" ...
##  $ Sex                                      : chr [1:55080] "All Sexes"
"All Sexes" "All Sexes" "All Sexes" ...
##  $ Age Group                                : chr [1:55080] "All Ages"
"Under 1 year" "0-17 years" "1-4 years" ...
##  $ COVID-19 Deaths                          : num [1:55080] 574045 74 295
36 103 ...
##  $ Total Deaths                             : num [1:55080] 4554041 25108
44028 4522 7306 ...
##  $ Pneumonia Deaths                         : num [1:55080] 501850 273 733
146 211 ...
##  $ Pneumonia and COVID-19 Deaths            : num [1:55080] 281084 12 58 6
23 ...
##  $ Influenza Deaths                         : num [1:55080] 9129 21 182 64
76 ...
##  $ Pneumonia, Influenza, or COVID-19 Deaths: num [1:55080] 802666 356 1152
240 367 ...
##  $ Footnote                                 : chr [1:55080] NA NA NA NA ...

head(Covid_df)
```

```
## # A tibble: 6 x 16
##   `Data As Of`        `Start Date`        `End Date`          Group  Year
Month
##   <dttm>              <dttm>              <dttm>              <chr>  <lgl>
<lgl>
## 1 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## 2 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## 3 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## 4 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## 5 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## 6 2021-05-19 00:00:00 2020-01-01 00:00:00 2021-05-15 00:00:00 By To~ NA
NA
## # ... with 10 more variables: State <chr>, Sex <chr>, Age Group <chr>,
## #   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>
```

Questions for future steps.

1. should other factors such as vaccines be looked at?
2. Should gender be included even though I do not expect it to be a factor?
3. Should states be removed to make the overall process easier?

What information is not self-evident?

Currently no operations have been run to see the correlation between age and deaths, but by looking at the raw data we are able to tell there is a correlation between those factors.

How do you plan to slice and dice the data?

If needed I can alter the age variable to narrow down exactly at what age sees the most significant increase in deaths.

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

The plan is to still use Histograms, Scatterplots and run correlation tests.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain

Once I am more comfortable with machine learning I would like to incorporate it to see any trends that might pop up as more data becomes avalibale.