Hyper Link for data sets

* COVID-19\_Case\_Surveillance\_Public\_Use\_Data\_with\_Geography
  + <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4>
* **Provisional\_COVID-19\_Death\_Counts\_by\_Sex\_\_Age\_\_and\_State**
  + <https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku>
* **COVID-19\_Vaccine\_Distribution\_Allocations\_by\_Jurisdiction\_-\_Pfizer**
  + <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg>
* COVID-19\_Vaccine\_Distribution\_Allocations\_by\_Jurisdiction\_-\_Moderna
  + <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws>
* COVID-19\_Vaccine\_Distribution\_Allocations\_by\_Jurisdiction\_-\_Janssen
  + <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh>
* AH\_Provisional\_COVID-19\_Deaths\_by\_Educational\_Attainment\_\_Race\_\_Sex\_\_and\_Age
  + <https://data.cdc.gov/NCHS/AH-Provisional-COVID-19-Deaths-by-Educational-Atta/3ts8-hsrw>
* AH\_Provisional\_COVID-19\_Deaths\_by\_Race\_and\_Educational\_Attainment
  + <https://data.cdc.gov/NCHS/AH-Provisional-COVID-19-Deaths-by-Race-and-Educati/i6ej-9eac>
* Provisional\_COVID-19\_Deaths\_Counts\_by\_Age\_in\_Years
  + <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-Counts-by-Age-in-Years/3apk-4u4f>
* sex\_of\_people\_fully\_vaccinated
  + https://covid.cdc.gov/covid-data-tracker/#vaccination-demographic
* age\_groups\_of\_people\_fully\_vaccinated
  + https://covid.cdc.gov/covid-data-tracker/#vaccination-demographic
* race\_ethnicity\_of\_people\_fully\_vaccinated
  + https://covid.cdc.gov/covid-data-tracker/#vaccination-demographic
* u.s.\_covid19\_vaccine\_delivered\_by\_vaccine\_type
  + https://covid.cdc.gov/covid-data-tracker/#vaccinations
* u.s.\_covid19\_vaccine\_administration\_by\_vaccine\_type
  + https://covid.cdc.gov/covid-data-tracker/#vaccinations
* number\_of\_people\_fully\_vaccinated\_in\_the\_u.s.in\_the\_u.s.\_by\_covid19\_vaccine\_series\_type
  + https://covid.cdc.gov/covid-data-tracker/#vaccinations
* trends\_in\_number\_of\_covid19\_vaccinations\_in\_the\_us
  + <https://covid.cdc.gov/covid-data-tracker/#vaccination-trends>
* deaths\_by\_sex\_\_all\_age\_groups
  + https://covid.cdc.gov/covid-data-tracker/#demographics
* cases\_by\_sex\_\_all\_age\_groups
  + https://covid.cdc.gov/covid-data-tracker/#demographics
* deaths\_by\_age\_group
  + https://covid.cdc.gov/covid-data-tracker/#demographics
* cases\_by\_age\_group
  + https://covid.cdc.gov/covid-data-tracker/#demographics
* deaths\_by\_race\_ethnicity\_\_all\_age\_groups
  + https://covid.cdc.gov/covid-data-tracker/#demographics
* cases\_by\_race\_ethnicity\_\_all\_age\_groups
  + <https://covid.cdc.gov/covid-data-tracker/#demographics>

## 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

### 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?

Class 1

Start Date: June 7, 2021

End Date: August 14, 2021

Class title:  Data Exploration and Analysis

Credit Units: 3

Class number: DSC 530

Course type: Graduate Class

Class Description:  This course introduces complex techniques needed for profiling and exploring data. Students use programming and statistics-based inference to ask and answer insightful questions of data.

Estimated Tuition Cost: 1,785.00

Estimated Book Cost: $34.99

Text

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

**Discussion 10.1: What is the z statistic?**

The Z-Score is known as a standard score that can be used to calculate if an individual score on an exam was in the top 10% of the classes scores. You can calculate this by using the Z-Score calculation formula known as **Z= (value – mean)/ (Standard Deviation).** As we have learned that a z score of 1 tells you that the observation is at a distance of one standard deviation towards the right from the center. While a Z Score of -1 tells you that the observation is one standard deviation towards the left of the center.

While in R Studios we can calculate the Z Score if a p-vaule is given by using the function “qnorm()”. As this function takes an argument of the p value and gives the z score. As seen below we see the qnorm function being used by inputting the p value and receiving the Z score in return:

A picture containing text

Description automatically generated

As seen above this value being below 1 means that the point that separates the lower 75% observations and upper 25% observations is within one standard deviation of the average, towards the right. While the link below will show you more info on z scores and other ways we can calculate it.

Reference: <https://www.programmingr.com/statistics/z-score-in-r/#:~:text=Z%3D%20(value%20%E2%80%93%20mean)%2F%20(Standard%20Deviation)&text=In%20other%20words%2C%20it%20should,in%20terms%20of%20standard%20deviations>.

The Z-Score is known as a standard score that can be used to calculate if an individual score on an exam was in the top 10% of the classes scores. You can calculate this by using the Z-Score calculation formula known as Z= (value – mean)/ (Standard Deviation). As we have learned that a z score of 1 tells you that the observation is at a distance of one standard deviation towards the right from the center. While a Z Score of -1 tells you that the observation is one standard deviation towards the left of the center.

While in R Studios we can calculate the Z Score if a p-vaule is given by using the function “qnorm()”. As this function takes an argument of the p value and gives the z score. As seen below we see the qnorm function being used by inputting the p value and receiving the Z score in return:

As seen above this value being below 1 means that the point that separates the lower 75% observations and upper 25% observations is within one standard deviation of the average, towards the right. While the link below will show you more info on z scores and other ways we can calculate it.

Reference: https://www.programmingr.com/statistics/z-score-in-r/#:~:text=Z%3D%20(value%20%E2%80%93%20mean)%2F%20(Standard%20Deviation)&text=In%20other%20words%2C%20it%20should,in%20terms%20of%20standard%20deviations.

400

As seen in our data frame that focused on race, sex, and age of individuals that have died from the Covid-19 virus I found a correlation between age and Covid-19 deaths. As I used the cor function that allowed me to calculate the correlation between the two as seen below that showed us that we have a positive correlation at 0.221864 which means as the age goes up, we see that the number of deaths also increase

cor(COVID\_19\_Deaths\_by\_Educational\_Race\_Sex\_and\_Age$`Age Group`, COVID\_19\_Deaths\_by\_Educational\_Race\_Sex\_and\_Age$`COVID-19 Deaths`)

As seen below from our scatter plot below we see that as age increase the number of deaths also increase which backs up our outcome from our corelation test. Another interesting factor we see is that in our eldest age group which is the 65 plus in this data frame we see that females out do the males in the death rate as the females almost reach 80,000 deaths due to Covid-19.

ggplot(data = COVID\_19\_Deaths\_by\_Educational\_Race\_Sex\_and\_Age, mapping = aes(x =`Age Group`, y =`COVID-19 Deaths`)) + geom\_point(aes(color = Sex))

These statistics above show us that our target audience for concern is the elder population such as individuals in their late 40s and up as our 65 plus individuals are at the highest risk for death if they were to catch this virus. While the female population in the 65 plus category saw the highest rate of death which would make them the most vulnerable.

Limitations that I have seen in my analysis would stem directly from the amount of data that has been collected on the data sets. As for the data frames on the vaccines I wish I had more information on age groups and races for the individuals in each state as this would help determine what age category, gender, and race were receiving the vaccine. This would allow me to do further research and see if the number of vaccines administered for each race, age group, and or gender influenced the number of deaths that have been see in our data frames on Covid-19 death rates. While another limitation I have seen is that I am not sure if all the data on Covid-19 deaths is complete such as each race did not seem to have an accurate number of deaths that have been reported to the CDC. As this can affect the accuracy of our data plots and statistics as our data stated whites had the highest death rate while Asian race showed very little to no deaths. While I can say either I our someone else can improve and or build on this report by either running further analysis and or pull new data from the CDC to get updated information.