Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?

**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.

* **Draft 5-10 Research questions that focus on the problem statement/topic.**

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?
   1. In regard to the number of vaccinations administered which state had the highest rate for each company?
   2. In regard to the number of vaccinations administered which state had the lowest rate for each company?
   3. View(COVID\_19\_Vaccine\_Distribution\_Allocations\_by\_Jurisdiction\_Pfizer)
   4. View(COVID\_19\_Vaccine\_Distribution\_Allocations\_by\_Jurisdiction\_Moderna)
   5. 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?
   1. When it comes to age and sex which combination had the highest death rate for each sex and age group?
   2. As most think the elderly population is most at risk for death from the virus but what percentage of deaths in the elderly population was from other illnesses besides Covid-19?
   3. 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?
   1. Did the education level have an effect on the death rate of different ethnicities?
   2. 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?
   1. 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?
   1. 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, 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.

Approach

<https://www.google.com/search?q=covid-19+deaths+in+us&oq=&aqs=chrome.0.69i59i450l8.108387505j0j15&sourceid=chrome&ie=UTF-8>

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, 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.

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. is used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables. Thanks Jiteshwar Anjale for putting this in a better perspective as the multinomial logistic regression Example 1. People’s occupational choices might be influenced by their parents’ occupations and their own education level. We can study the relationship of one’s occupation choice with education level and father’s occupation. The occupational choices will be the outcome variable which consists of categories of occupations.

this makes total sense in regards to the multicollinearity is the more common problem that happens when in multiple regression modelling two or more of the explanatory variables are approximately linearly related.

is to examine the correlation between each pair of explanatory variables. If two of the variables are highly correlated, then this may the possible source of multicollinearity.