Final Project

Jackson Lyons

2023-11-18

Introduction

The goal of this paper is to look at the relationship between Covid Vaccine data and hospitalization rates vs Influenza vaccine and hospitalization data. The overall goal of this paper is to look at how closely related covid and flu vaccines are as well as their hospitalization rates. Our central problem for the paper is to look at how we can allocate medial supplies by predicting covid and flu hospitalization rates using vaccination rates for both illnesses.

Research Questions

Some research questions that will be addressed through this paper includes what kind of relationship, if any, exists between covid vaccination rates and flu vaccination rates on a state by state basis? What kind of relationship exists between flu vaccination rates and hospitalizations on a year by year basis or in other words, how effective is the flu vaccine annually? What kind of relationship exists between covid vaccination rates and hospitalizations on a year by year basis or in other words, how effective is the covid vaccine annually? How do different states compare between covid and flu vaccination? How related are all the different variables and are there relationships between the two that could be useful to predict illnesses? Finally, can we use covid and flu data to predict which states are more likely to have high rates of hospitalization from flu and covid?

Approach

To answer these research questions, there will be multiple steps that needs to be done. The first it to gather, clean, and merge the different data for covid vaccines, flu vaccines, coivd hospitalizations and flu hospitalizations together. Next, we will look at the relationships between the different variables to see what methods would work best to conduct further analysis. Finally, we will look at creating a model to use covid and flu data to predict which states will have the most hospitalizations based on vaccine data to help distribute supplies as needed.

Using this approach, we will be able to see how much of an impact vaccination rates have on serious illness prevention. Once we know how much this effect is, we can use current vaccination rates to predict how many hospitalizations will occur and thus allocate medical supplies based on this prediction.

We will need the ggplot2 package for this project. At this time, there are no other packages that will be required but this may change going forward based on the project needs.

Data

The data from this project all comes from the CDC 2021-2022 flu season as well current covid vaccination rates and hospitalization rates. The data is all by state so we can analyze the impact on a state as well as a national level. The data for this project comes from, https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-people-fully-percent-tot, https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html, and https://www.cdc.gov/flu/fluvaxview/dashboard/vaccination-adult-coverage.html. The flu data contains information about each state on a weekly basis. The Covid data is also by state, but is cumulative. We will need to conver the flu data so that it is compatible with our covid data for analysis.

Plots and Table Needs

The plots that we need will be scatterplots between all the different variables to see what kind of relationships exist, if any. When we analyze our models we create, we will need to analyze the residuals as well. A table will be needed to compare the state by state flu and covid data.

Questions for Further Steps

For further steps, we need to figure out if there is any relationship between the variables. If there are no apparent relationships, our models are likely to not have any predictive power. Another question might be, what other data could we include with our model to improve our preditive power.

```
flu_vaccine <- read.csv("Flu_Vacine_Data.csv")
covid_vaccine <- read.csv("Covid19_Vaccination_Data_May_11_2023.csv")
flu_hospital <- read.csv("Flu_Surveillance_Data.csv")
```

After reading in the data, we can see that our data frames are not formatted in the same way. The flu vaccine data is broken down by state and further divided by race. In order to use this data with the covid vaccine data, we are going to need to look at the cumulative vaccine data for each state.

First, we are going to clean up the flu vaccine data. We don't need the data to be broken down by race and age group so our first step is going to be to only keep the overall data for all adults.

```
new_flu_vaccine <- flu_vaccine[(flu_vaccine$Race %in% "Overall"),]
new_flu_vaccine <- new_flu_vaccine[(flu_vaccine$Age_Group == "All Adults (18+)"),]
```

Next, we don't need any columns except for the state name as well as the estimate for how many flu vaccines had been distributed. So, we will drop all the unnecessary columns. We will also rename the column names to be consistent across all dataframes.

```
new_flu_vaccine <- new_flu_vaccine[,c("Geography", "Point_Estimate")]
colnames(new_flu_vaccine)[1] = "State"
colnames(new_flu_vaccine)[2] = "Flu_Vaccines"
head(new_flu_vaccine)
```

```
##               State Flu_Vaccines
## 723         Florida         22.0
## 727        Kentucky         25.1
## 905           Texas         37.2
## 906         Wyoming         37.5
## 907         Florida         38.2
## 908 City of Houston         38.6
```

Now we will clean up the data in the flu hospitalization data. Since the hospitalization data gives us activity levels as strings and not as numeric values, this will be our first step. We will split the string using the space as our delimiter and we will keep only the number.

```
new_flu_hospital <- mutate(flu_hospital, ACTIVITY.LEVEL=sapply(strsplit(flu_hospital$ACTIVITY.LEVEL, sp
```

When looking through the data frame, there are values of "Surveillance" and "News" under the activity levels so we will drop these rows.

```
new_flu_hospital <- new_flu_hospital[!(new_flu_hospital$ACTIVITY.LEVEL %in% c("Surveillance","News")),]
```

Now, we can convert the activity levels into numeric values to help us with further analysis. We are also going to drop any rows with NA values.

```
new_flu_hospital$ACTIVITY.LEVEL = as.numeric(as.character(new_flu_hospital$ACTIVITY.LEVEL))
new_flu_hospital <- na.omit(new_flu_hospital)
```

Finally, we do not need all the extra columns besides the state name, activity level so we will drop the rest of the columns for clarity. This gives us the values that we will need going forward. We will also rename the columns to be the same as our other data frames.

```
new_flu_hospital <- new_flu_hospital[,c("STATENAME", "ACTIVITY.LEVEL")]
colnames(new_flu_hospital)[1] = "State"
colnames(new_flu_hospital)[2] = "Activity_Level"
head(new_flu_hospital)
```

```
##        State Activity_Level
## 1    Alabama              1
## 2     Alaska              1
## 3    Arizona              1
## 4   Arkansas              2
## 5 California              1
## 6   Colorado              2
```

Finally, we do not have to clean up the covid vaccine data too much. For this data frame we are just going to get rid of the columns that are not relevant to our analysis.

```
new_covid_vaccine <- covid_vaccine[,c("Jurisdiction..State.Territory..or.Federal.Entity",
                                      "Percent.of.total.pop.with.a.completed.primary.series",
                                      "Percent.of.total.pop.with.at.least.one.dose")]
```

These are very long column names so we will shorten them to more readable names.

```
colnames(new_covid_vaccine)[1] = "State"
colnames(new_covid_vaccine)[2] = "Primary_Series"
colnames(new_covid_vaccine)[3] = "One_Dose"
head(new_covid_vaccine)
```

```
##            State Primary_Series One_Dose
## 1  United States           69.5     81.4
## 2         Alaska           65.3     73.2
## 3        Alabama           53.3     65.1
## 4       Arkansas             57     70.1
## 5 American Samoa           89.7       95
## 6        Arizona           66.2     78.4
```

```
new_covid_vaccine$Primary_Series <- as.numeric(new_covid_vaccine$Primary_Series)
```

```
## Warning: NAs introduced by coercion
```

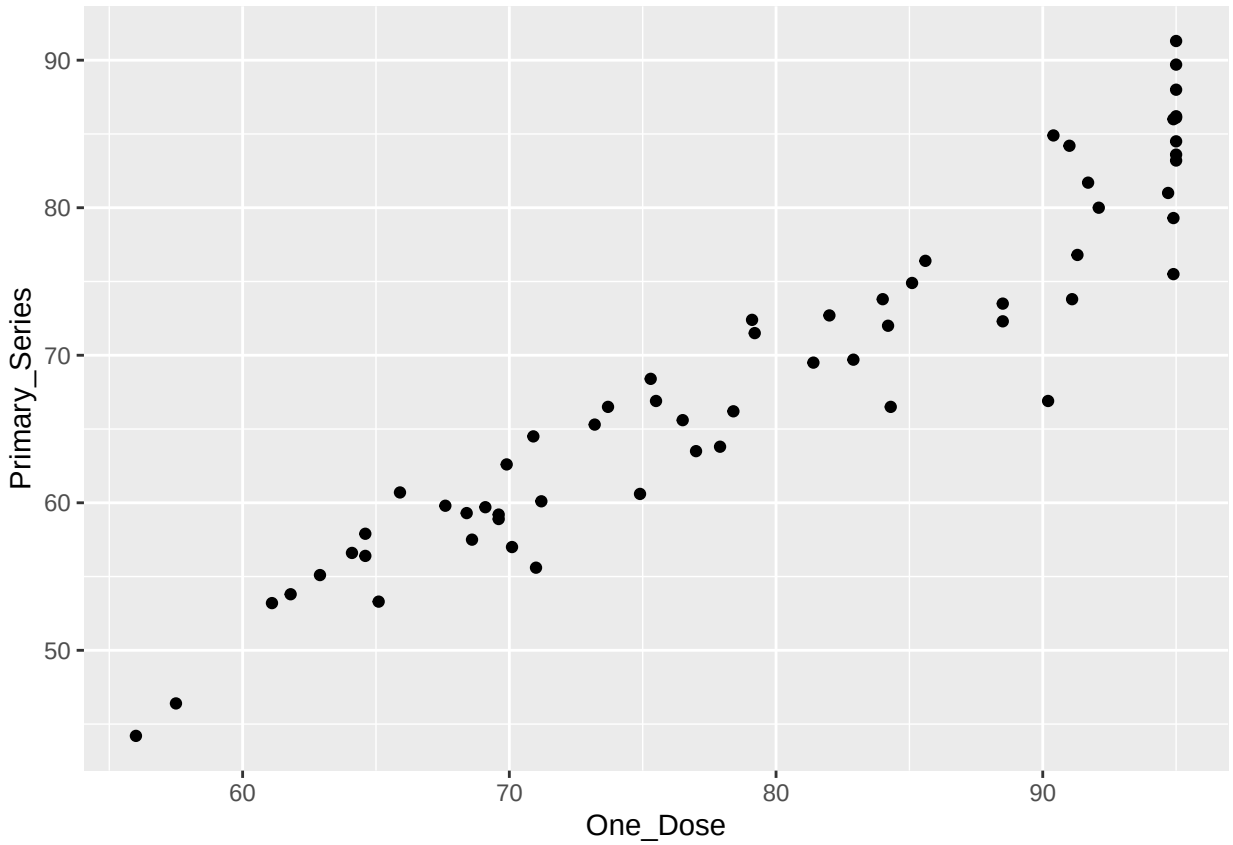
```
new_covid_vaccine$One_Dose <- as.numeric(new_covid_vaccine$One_Dose)
```

```
## Warning: NAs introduced by coercion
```

Now, our data frames for each data set are clean with consistent and descriptive column names. For the time being, we are going to keep the data frames separate so we can do some more modifications to our data. We still need to get the average activity level for each state for the flu hospitalization data as well as doing some analysis within each of the data frames.

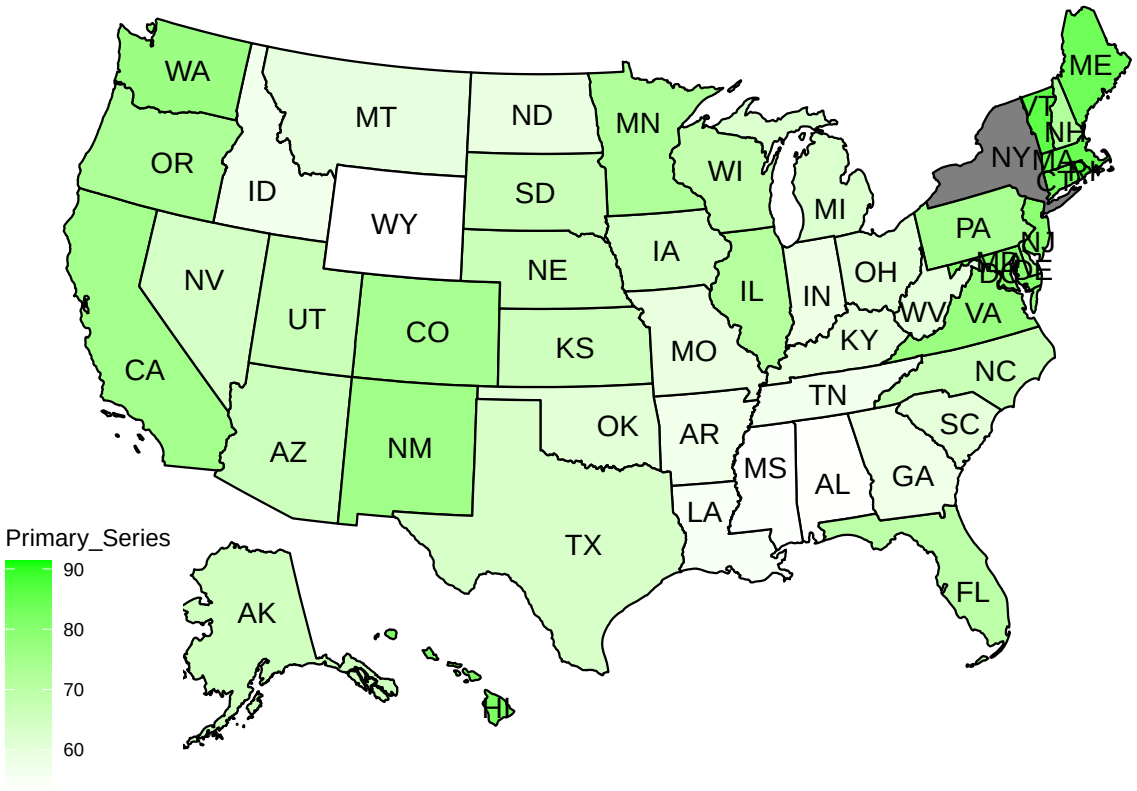
```
ggplot(new_covid_vaccine) + geom_point(aes(One_Dose, Primary_Series))
```

```
## Warning: Removed 3 rows containing missing values (`geom_point()`).
```

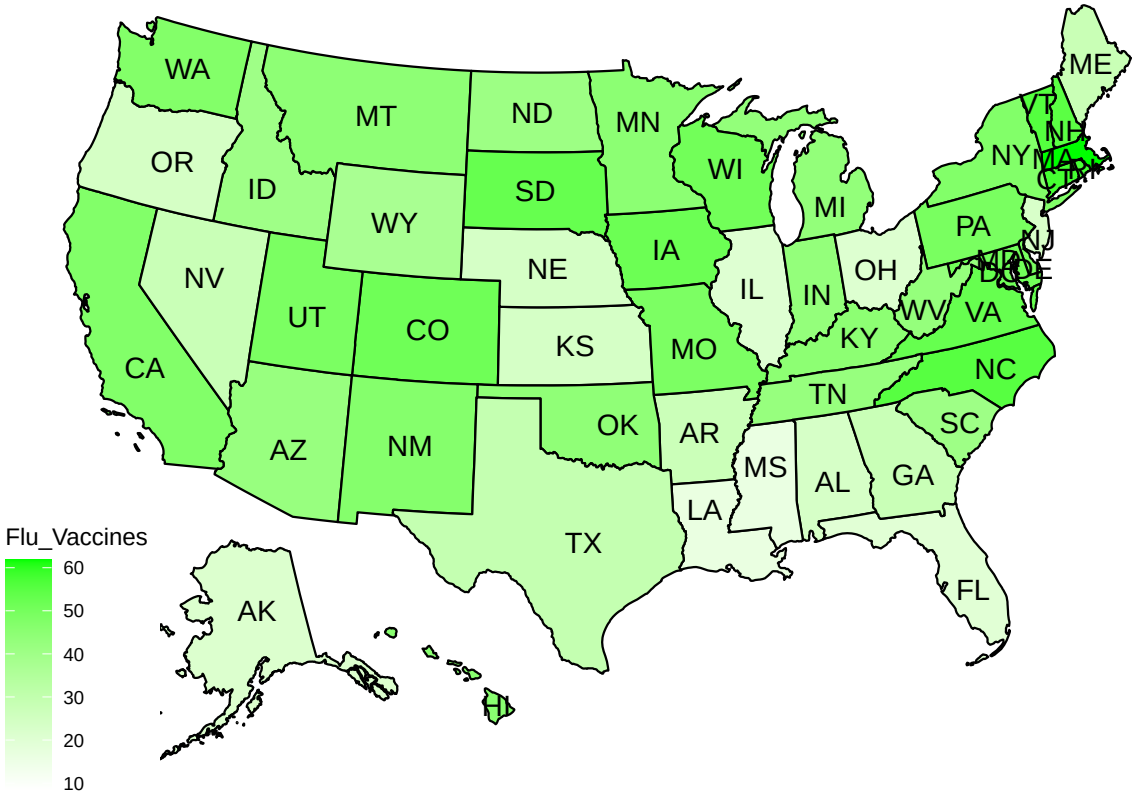


We can see from the plot of one dose vs primary series a clear linear relationship between the two variables. This is to be expected because most people who get one dose of the covid vaccine is likely to get the second dose.

```
plot_usmap(data = new_covid_vaccine, values = "Primary_Series", labels = TRUE) +
    scale_fill_gradient(low = "white", high = "green")
```



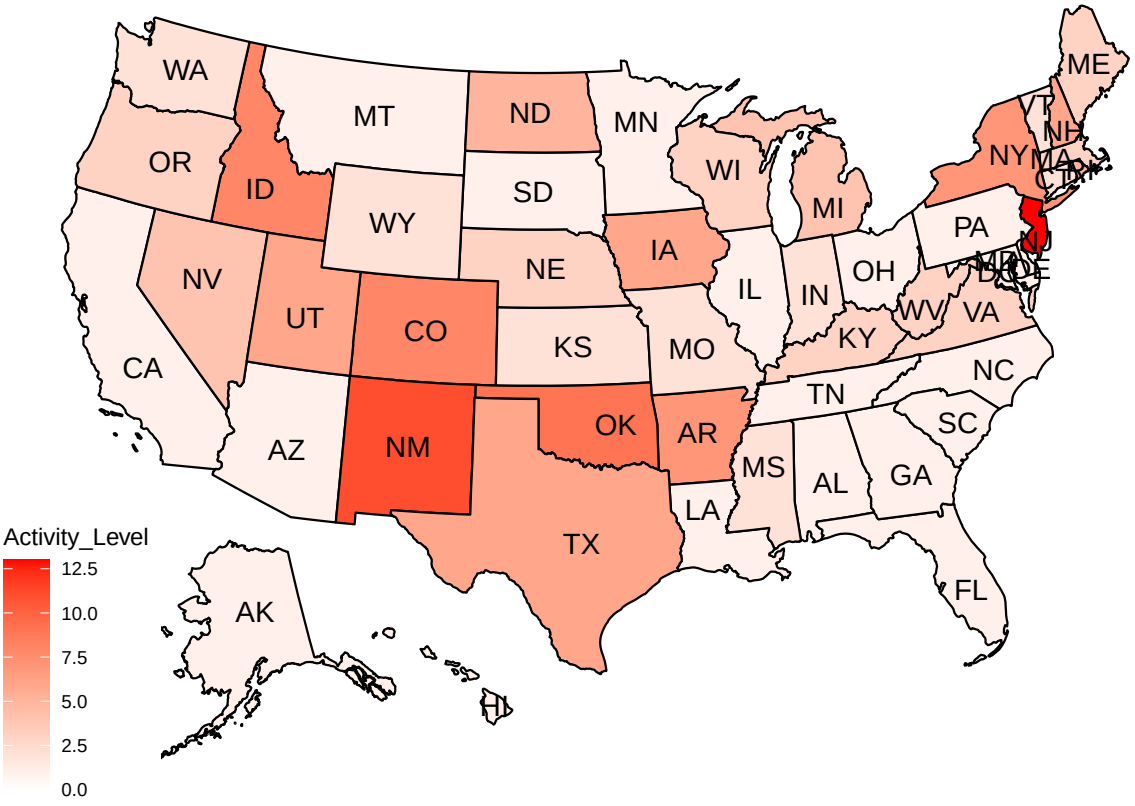

```
plot_usmap(data = new_flu_vaccine, values = "Flu_Vaccines", labels = TRUE) +
    scale_fill_gradient(low = "white", high = "green")
```



Looking at the previous two plots, we can see the difference between states for covid and flu vaccination rates. In

both plots, the darker the green the higher the vaccination rate. This gives us a nice visual to look at and compare between states. We can see that there tends to be some more uniformity among states in regards to flu vaccination. We see many states have the same, darker green shade compared to the covid vaccines. However, we can see from the scales of the two that the flu vaccination rate is much lower than that of the covid vaccination rate.

```
plot_usmap(data = new_flu_hospital, values = "Activity_Level", labels = TRUE) +
    scale_fill_gradient(low = "white", high = "red")
```



The plot above shows the hospitalizations by each state using the max activity level for each state. The darker the red the value, the higher the max activity level was.

```
merged <- merge(new_covid_vaccine, new_flu_vaccine, by = "fips", all = TRUE)
merged <- merged[,c("fips", "State.x", "Primary_Series", "One_Dose", "Flu_Vaccines")]
colnames(merged)[2] = "State"
```

```
merged <- merge(merged, new_flu_hospital, by = "fips", all = TRUE)
merged <- merged[,c("fips", "State.x", "Primary_Series", "One_Dose", "Flu_Vaccines", "Activity_Level")]
colnames(merged)[2] = "State"
```

Now that we have merged all of our data, we can discuss what would be the steps going forward. There is still quite of bit of work to be done before we could implement a model for our data. Our data would need to be cleaned a little bit more to remove NA values as well as the duplicate flu vaccine values for the different weeks. The next step would be to look at how each of the variables relates with the others. This would be accomplished through looking at scatter plots as well as correlation coefficients. We could also look at the values on a state by state level to see how our data varies by state. Finally, we could implement the model for our data. We would more than likely utilize a linear regression model to model our hospital load by using the flu and covid vaccines as predictors.

Conclusion

To conclude, there are a couple things I would have done differently in this analysis. The first would be with the data. While all the data came from trustworthy sources, it was not the easiest to use or the most intuitive. There were variables that were confusing with what they were and required more research to understand. There was also too much data for what I was trying to achieve in this research. The flu vaccine data included vaccine values for each week of the flu season. This could be useful