Final Paper

Lalitha Vadrevu

June, 2021

INTRODUCTION

It has been over 468 days since the declaration of the corona virus pandemic on 11 March 2020, and an unimaginable amount of data has been found related to the novel corona virus that appeared in December of 2019. As it has enveloped every aspect of our lives for nearly a year and a half, naturally we have some questions about this disease.

Throughout our analysis of a data sheet we downloaded from the CDC's website on 29 May 2021, we answered two questions we had been curious about. The first question is which comorbidities were most commonly associated with COVID19 deaths. The second question we wished to explore is how the age groups and geographic region of the population affect the number of COVID-19 deaths.

One might ask why anyone would care to answer these questions. Our first question is extremely important, as discovering which types of diseases are most commonly associated with dying from COVID-19 could provide some insight into fighting the disease. In the future, knowing what kind of disease something is could be used to combat it as well. For example, we know COVID-19 is a respiratory illness, so logically it should be combated with techniques that are used to fight other respiratory diseases. However, we also know it affects more than just our lungs, so knowing what diseases people have along with COVID-19 when they die from the disease could be used to prevent further deaths later on. Our second question is equally important, as we can see which age group is most affected by the virus. Additionally, our second question involving geographic region can allow us to look further into each state's action towards the pandemic, and understand the number of COVID-19 deaths associated with each state.

DATA

For this project, we downloaded a data set from the Center for Disease Control official website that was a record of the pre-existing health conditions that contribute to COVID-19 deaths sorted by states and age groups. The data was collected from 1 January 2020 to 29 May 2021.

The variables we decided to use include: State (includes the United States as the grand total), Condition.Group(pre-existing health condition), Condition(specific health condition with in each Condition.Group), Age.Group, and COVID.19.Deaths(number of COVID deaths).The State variable describes which State the data was collected in. The Condition.Group variable combines multiple pre-existing health conditions to summarize the data. The Condition breaks down Condition.Group and shows each of the conditions within each individual disease that contributed to COVID-19 deaths.The Age.Group divides up people by age to show how many people died in each group. Finally, the COVID.19.Deaths is the count of deaths associated with each condition.

The sample contains the number of COVID-19 deaths with pre-existing health conditions and also the COVID-19 deaths in general. Each observation in our table is grouped by the State, Condition.Group, the specific Condition, and the Age.Group,and generates the COVID-19 deaths associated with this specific grouping. The number of observations in our data set were 248,400.

Group	State	Condition.Group	Condition	Age.Group	COVID.19.Deaths
By Total	United States	Respiratory diseases	Influenza and pneumonia	0-24	1569
By Total	United States	Respiratory diseases	Influenza and pneumonia	25-34	5804
By Total	United States	Respiratory diseases	Influenza and pneumonia	35-44	15080
By Total	United States	Respiratory diseases	Influenza and pneumonia	45-54	37414
By Total	United States	Respiratory diseases	Influenza and pneumonia	55-64	82668
By Total	United States	Respiratory diseases	Influenza and pneumonia	65-74	129005

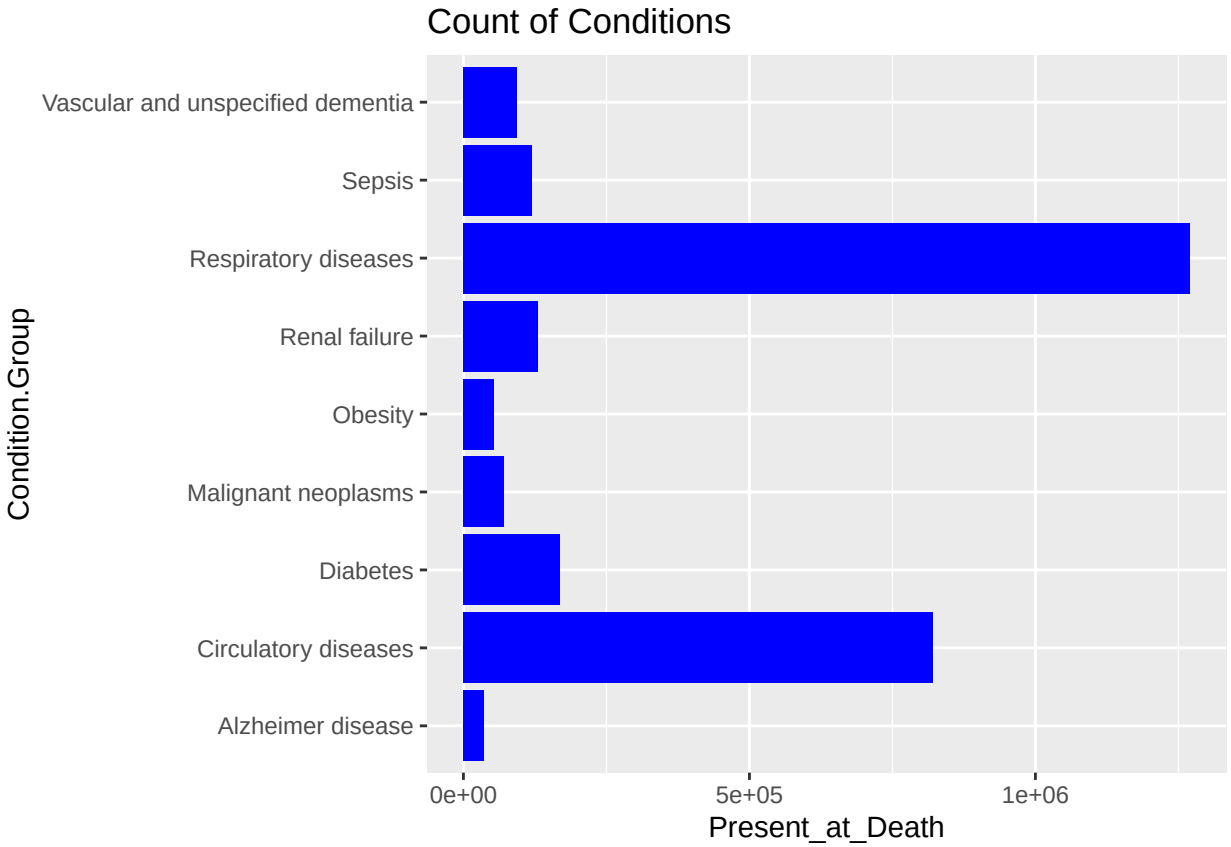
This table shows us the number of deaths present for each pre-existing health conditions.

Condition.Group	Present_at_Death
Alzheimer disease	35377
Circulatory diseases	821161
Diabetes	168830
Malignant neoplasms	71669
Obesity	53321
Renal failure	129674
Respiratory diseases	1269923
Sepsis	119000
Vascular and unspecified dementia	93117

RESULTS

Question 1: Which groups of comorbidities are most commonly associated with COVID-19 Deaths?

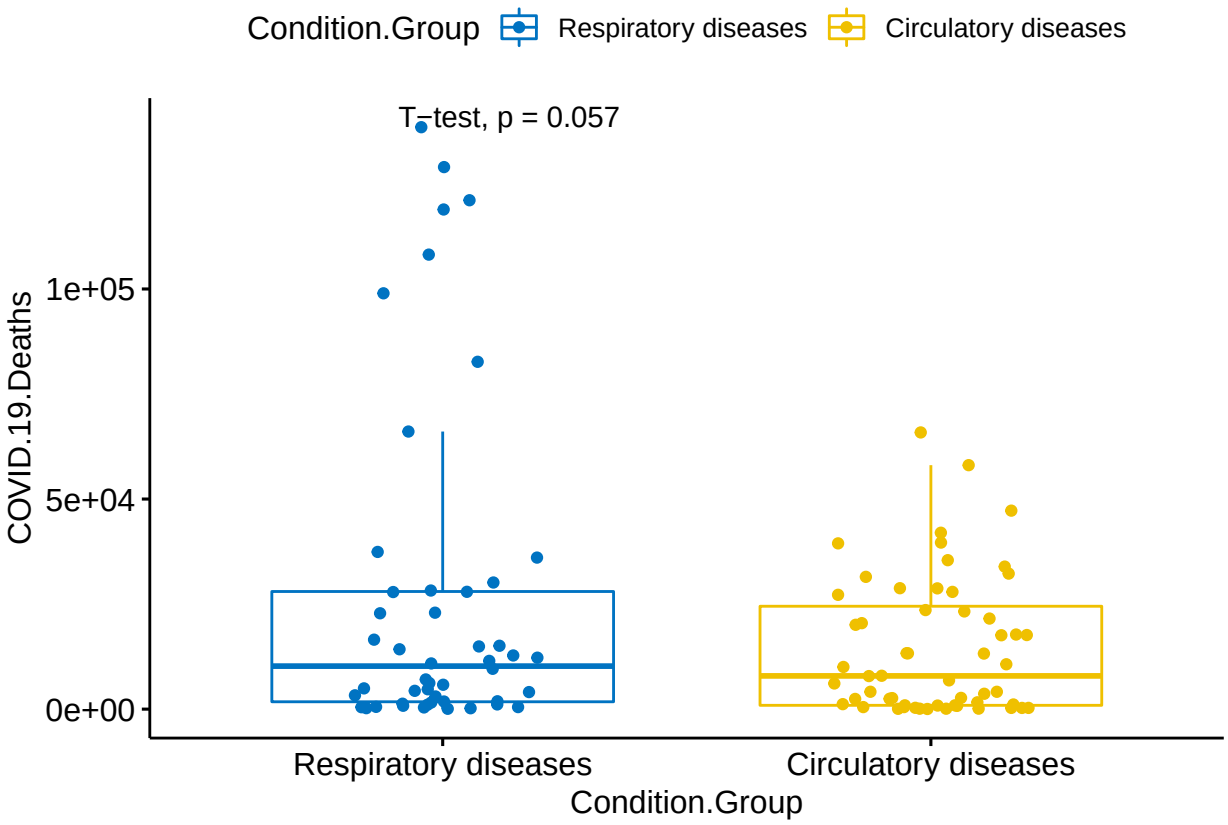
To discover which disease groups were most commonly associated with COVID-19 deaths, I filtered the data set to exclude duplicate counts. This meant filtering out the total COVID-19 deaths, the total deaths associated with each disease, total deaths by State, and the total deaths in the age column, as each of these numbers doubled the actual count. After doing this, I was left with the reported count of each condition within their associated condition groups in the United States as a whole. The result of this filtered data was then put into both a box plot and a bar graph to visualize it. From this visualization, it was very clear that respiratory and circulatory diseases were most commonly associated with COVID-19 deaths. Considering COVID-19 is a primarily respiratory disease, it is not surprising to find that there were actually more instances of respiratory diseases as underlying conditions than there were COVID-19 deaths with over 600,000 reported instances, showing that many people actually had multiple respiratory illnesses concurrently as there were under 600,000 COVID-19 deaths at the time the data was taken. What was more surprising was finding that circulatory diseases also stood out from the other condition groups, with over 400,000 reported cases of a COVID-19 deaths associated with a circulatory disease. To understand more about why respiratory and circulatory diseases were the primary condition groups associated with COVID-19 deaths, I decided to break down each of these groups to find out which conditions in particular, if any, were most often associated with COVID-19 deaths. We wanted to explore the conditions within pre-existing health condition groups; therefore, we decided to see which conditions within the respiratory disease affected the COVID-19 deaths.



T-Test:

To compare the two most common diseases in the bar chart above, we ran a T-test to visualize the difference in the impact of these diseases. As the test shows, respiratory disease has a significantly larger presence in those who died by COVID, with some of the data points reaching up to more than 60000. On the other hand, circulatory diseases were comparatively lower in presence with the highest data point being less than 40000 deaths.

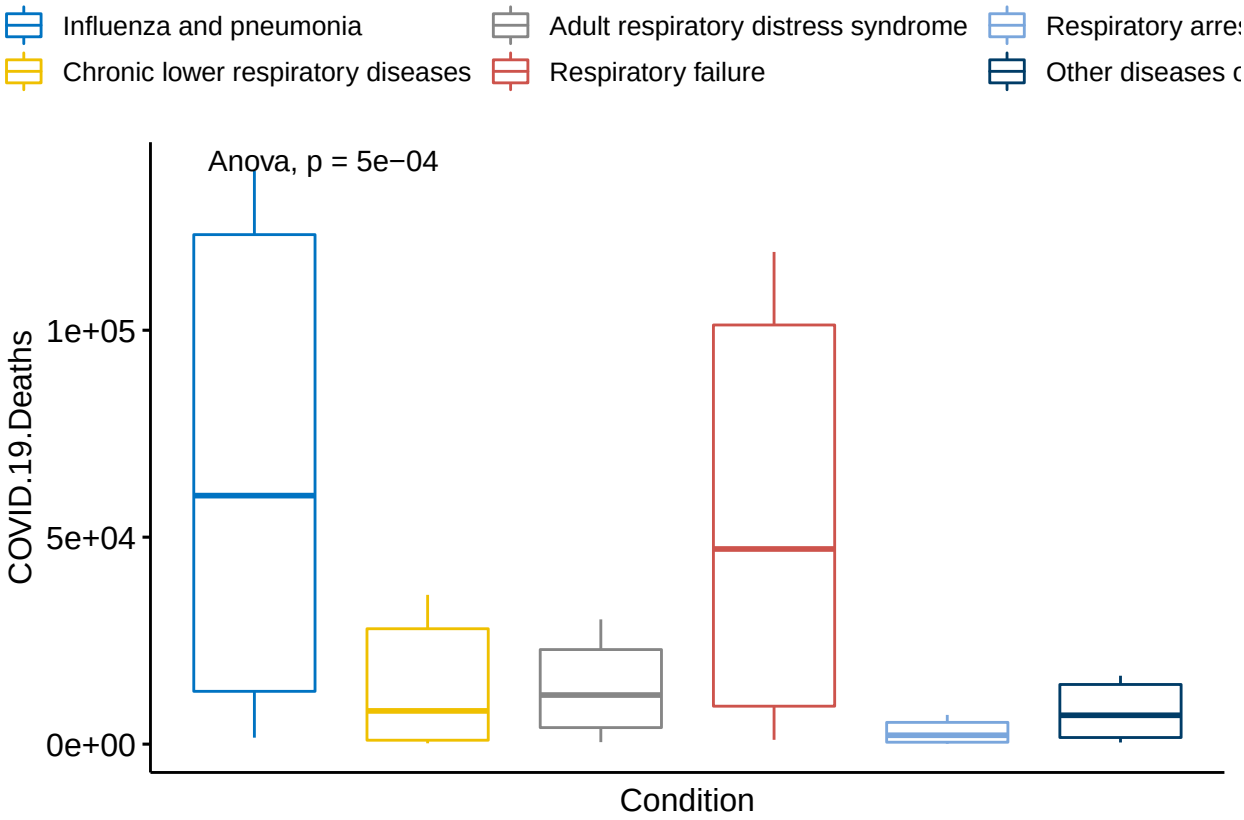
.y.	group1	group2	p	p.adj	p.format	p.signif	method
COVID.19.Deaths	Respiratory diseases	Circulatory diseases	0.3804822	0.38	0.38	ns	Wilcoxon



ANOVA Test for Respiratory Disease:

We ran an Anova test on the subcategories of respiratory disease in order to see which conditions within this disease affected the COVID-19 deaths the most. As the Anova test shows, the largest presence of death by corona virus was of "Influenza and pneumonia" while the second largest subcategory was "Respiratory failure". The p-value after running the Anova test is statistically significant, so we can assume that there is a positive association between the pre-existing condition of "Influenza and pneumonia" and COVID-19 deaths.

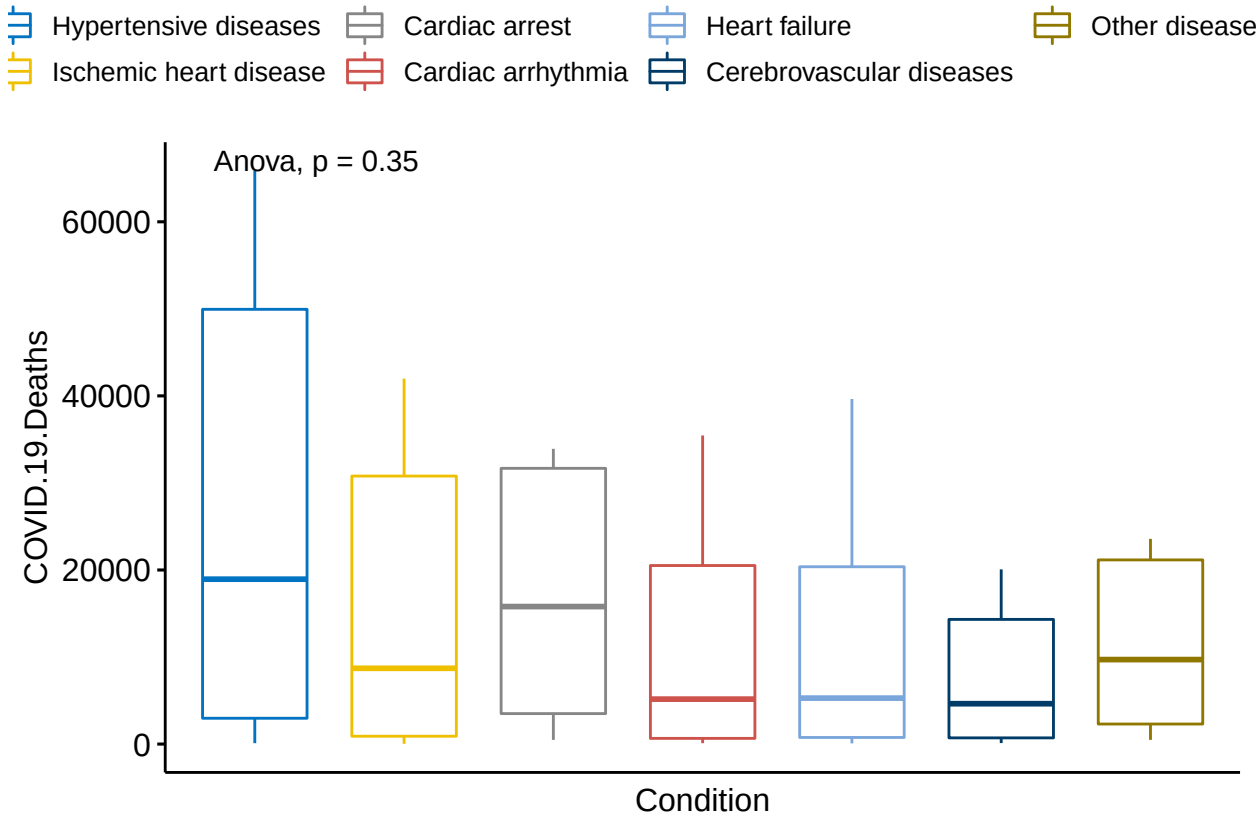
.y.	p	p.adj	p.format	p.signif	method
COVID.19.Deaths	0.0004998	5e-04	5e-04	***	Anova



ANOVA Test for Circulatory Disease:

We ran an Anova test on the subcategories of circulatory disease in order to see which conditions within this disease affected the COVID-19 deaths the most. We can see that hypertensive diseases are by far the most commonly present in those patients who die by COVID-19. We can thus infer that Pneumonia and Hypertensive diseases, while may not be a causative factor for COVID-19, but they definitely aggravate and increase the chances of fatality by corona virus.

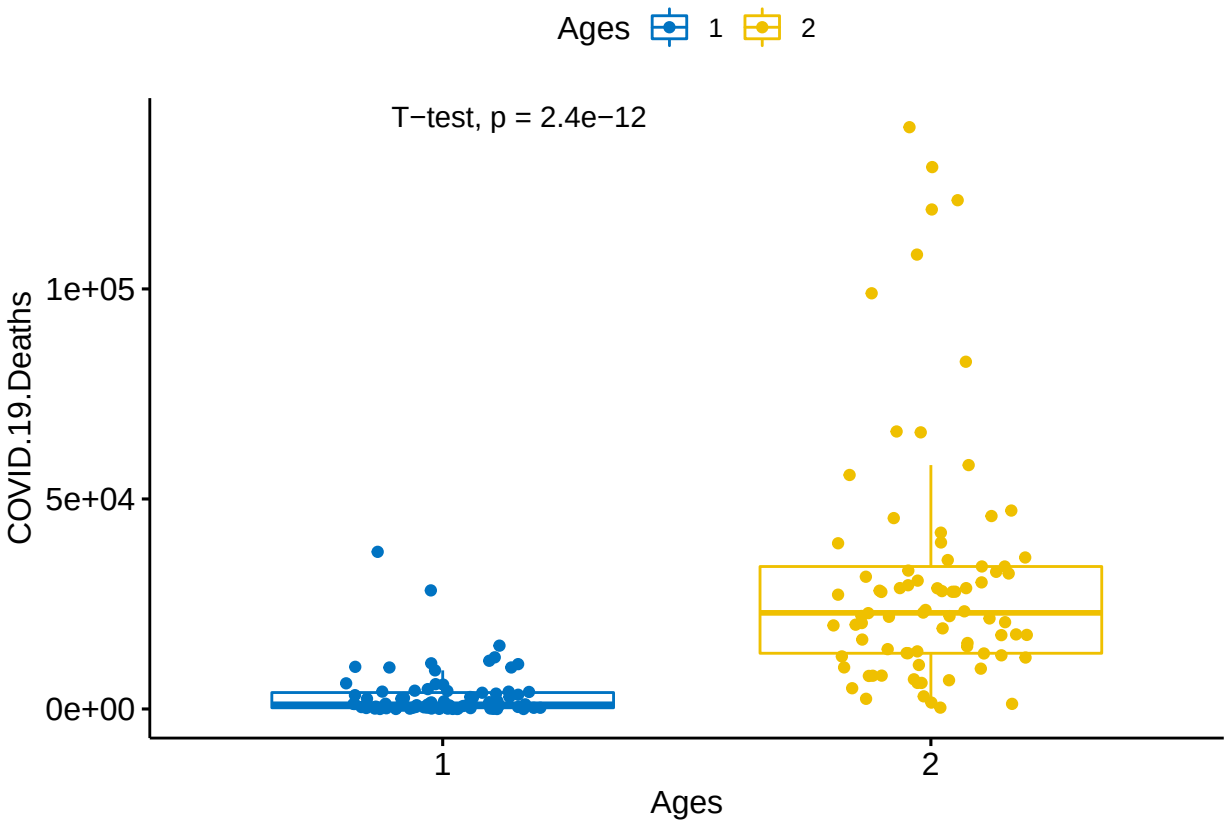
.y.	p	p.adj	p.format	p.signif	method
COVID.19.Deaths	0.3516849	0.35	0.35	ns	Anova

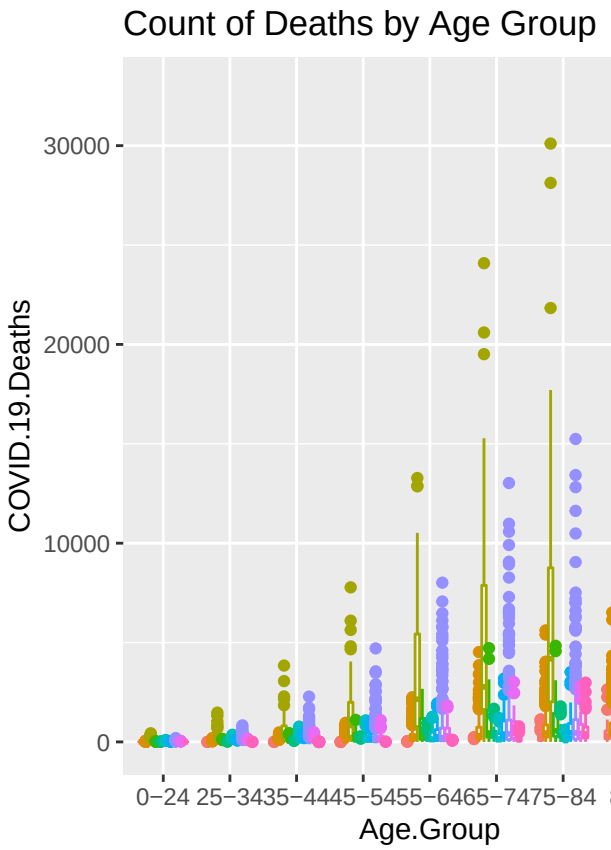


Question 2: How does the age group and geographical region of the population affect the number of COVID-19 deaths?

To analyze the relation between age and COVID-19 deaths, we divided all the age groups into two main categories. The first category (1) contains those people between the ages of 1-45 while the second category (2) contains people beyond 45-85+. There is a clear and visible difference between the death counts among both groups. The first category of younger people has a significantly lower number of deaths compared to the second category. The upper bound of the first category is less than 20000 while the upper bound of the category containing older patients is beyond 60000. Additionally, the p-value is statistically significant; therefore, we can assume that the older population is more affected by COVID-19 compared to the younger population.

.y.	group1	group2	p	p.adj	p.format	p.signif	method
COVID.19.Deaths	1	2	0	0	<2e-16	****	Wilcoxon





This is a visualization to explore both of the questions of our data analysis.

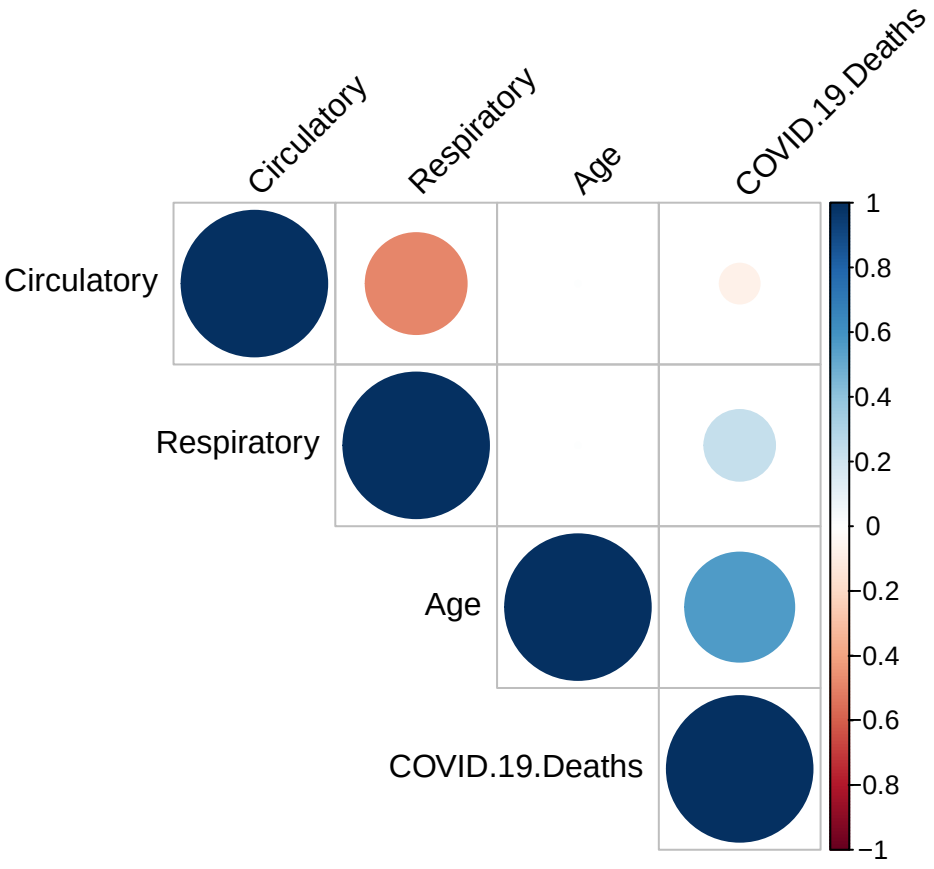
Multiple Regression:

Multiple regression is basically an extension of linear regression into relationship between more than two variables. In a linear relation we have one predictor(x) and one response(y) variable, but in multiple regression we have more than one predictor variable and one response variable.

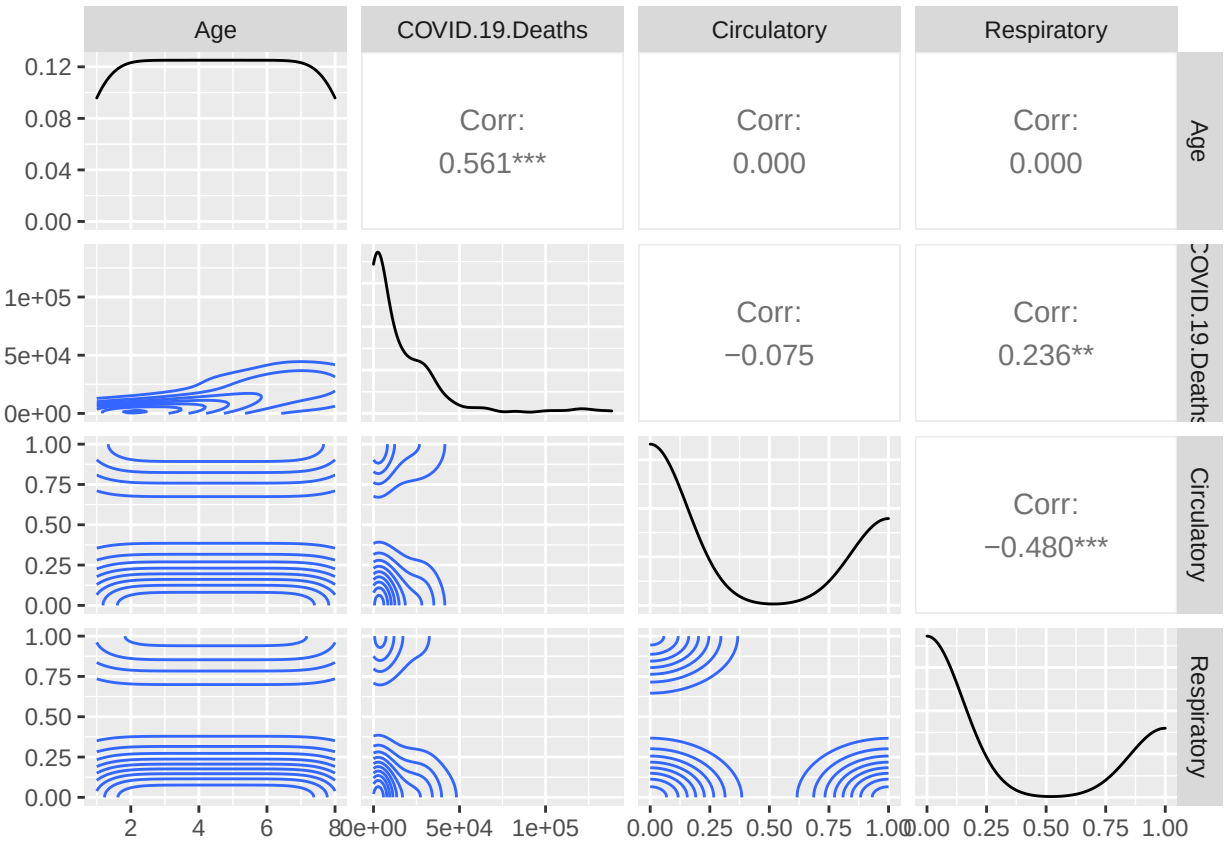
Since our data set had a limited number of numerical variables, we decided to create dummy variables for the disease, in order compare how having a respiratory/circulatory disease affected one's chances of surviving COVID-19 versus their chance without the pre-existing health condition.

VISUALIZATIONS/TABLES generated from Multiple Regression procedures:

	Age	COVID.19.Deaths	Circulatory	Respiratory
Age	1.00	0.56	0.00	0.00
COVID.19.Deaths	0.56	1.00	-0.07	0.24
Circulatory	0.00	-0.07	1.00	-0.48
Respiratory	0.00	0.24	-0.48	1.00



Age	COVID.19.Deaths	Circulatory	Respiratory
1	1569	0	1
2	5804	0	1
3	15080	0	1
4	37414	0	1
5	82668	0	1
6	129005	0	1



Optimization by Backward Elimination:

Backward elimination can be performed manually while considering what variables are eligible for removal.The predictors in our model were Age, Respiratory disease, and Circulatory disease. Then we decided to refit the model and eliminate the remaining least significant predictor provided its p-value is greater than 0.05. After continuing the back-end elimination procedure, we can drop all of the non-significant predictors. At the end of this procedure only the significant variables remain (i.e., COVID.19.Deaths ~ Age + Respiratory).

```
##
## Call:
## lm(formula = COVID.19.Deaths ~ Age + Circulatory + Respiratory,
##     data = plot)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41257  -9298  -2604   5131  96429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16129.9     4193.8  -3.846 0.000174 ***
## Age           6247.0      706.4   8.844 1.87e-15 ***
## Circulatory   2681.6     3868.9   0.693 0.489263
## Respiratory  14474.6     4026.9   3.595 0.000435 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20470 on 156 degrees of freedom
## Multiple R-squared:  0.3723, Adjusted R-squared:  0.3602
```

```
## F-statistic: 30.84 on 3 and 156 DF,  p-value: 1.036e-15
```

After continuing the back-end elimination procedure, we can drop all of the non-significant predictors. At the end of this procedure only the significant variables remain (i.e., COVID.19.Deaths ~ Age + Respiratory).

Optimization by AIC:

We used the AIC to estimate the relative amount of information lost by the given model. So the less amount a model loses, the higher the quality. Basically the AIC calculates the trade-off between the quality of the fit of the model and the simplicity of the model.AIC is minimized by choosing only two of the three predictors originally used (i.e., COVID.19.Deaths ~ Age + Respiratory).

```
## Start:  AIC=3180.53
## COVID.19.Deaths ~ Age + Circulatory + Respiratory
##
##               Df  Sum of Sq        RSS    AIC
## - Circulatory  1 2.0135e+08 6.5583e+10 3179.0
## <none>                      6.5381e+10 3180.5
## - Respiratory  1 5.4151e+09 7.0796e+10 3191.3
## - Age          1 3.2781e+10 9.8162e+10 3243.6
##
## Step:  AIC=3179.02
## COVID.19.Deaths ~ Age + Respiratory
##
##               Df  Sum of Sq        RSS    AIC
## <none>                      6.5583e+10 3179.0
## - Respiratory  1 5.7959e+09 7.1378e+10 3190.6
## - Age          1 3.2781e+10 9.8364e+10 3241.9
```

```
##
## Call:
## lm(formula = COVID.19.Deaths ~ Age + Respiratory, data = plot)
##
## Coefficients:
## (Intercept)          Age  Respiratory
##      -14789         6247        13134
```

Optimization by Adjusted R^2:

The maximum adjusted-R^2 criterion also suggests a model with only 2 predictors (i.e., COVID.19.Deaths ~ Age + Respiratory).

```
## Subset selection object
## Call: regsubsets.formula(COVID.19.Deaths ~ ., data = plot)
## 3 Variables  (and intercept)
##             Forced in Forced out
## Age             FALSE      FALSE
## Circulatory     FALSE      FALSE
## Respiratory     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          Age Circulatory Respiratory
## 1  ( 1 ) "*" " "         " "
## 2  ( 1 ) "*" " "         "*"
## 3  ( 1 ) "*" "*"         "*"
```

```
## [1] 2
```

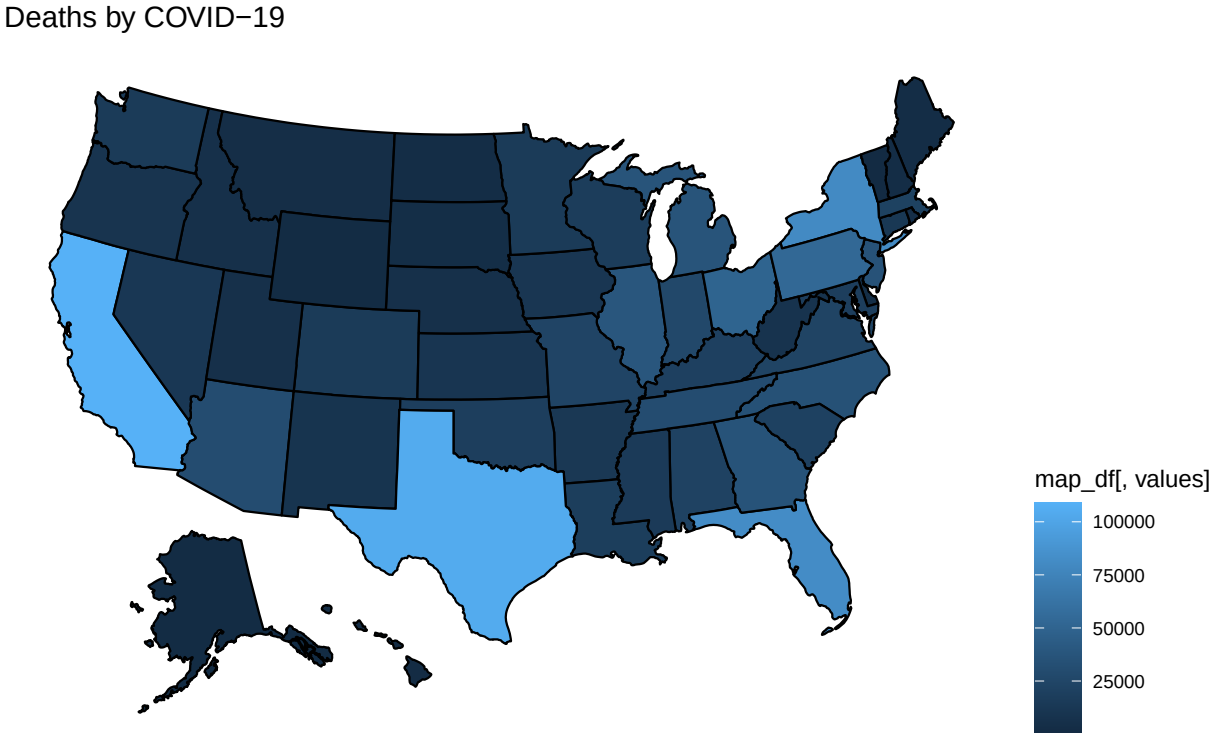
After using the three procedures(Backward Elimination, AIC, Adjusted R^2) of Multiple Regression, we can see consistent results of a regression model with only two predictor variables. The two predictor variables that all three procedures/criterion result in are "Age" and "Respiratory".

Geographical Region vs. COVID-19-Deaths:

According to this map, the more populous a state is, the higher the amount of COVID-19 deaths there are. The states with the highest populations such as California, Texas, New York show a stark contrast in the light blue color compared with darker states with lower populations. We also found that while Florida is the third most populous state, the number of deaths are not as high as other highly populated states. Generally, the Eastern regions of the United States have far more COVID-19 deaths than the Western regions, however this is not surprising as the Eastern half of the United States is more populated. The northwest areas have the lowest amount of deaths which can be associated with the low populations in colder climate regions and healthier lifestyles. Western states that are more sparsely populated also experienced lower death tolls because there are not many people there to begin with.

State	COVID.19.Deaths
United States	1146242
Alabama	21513
Alaska	1491
Arizona	30309
Arkansas	12661
California	109130

state	COVID_Deaths
AK	1491
AL	21513
AR	12661
AZ	30309
CA	109130
CO	15382



CONCLUSION

We all know that COVID-19 has been the worst pandemic in many decades with a death toll of around 3.9 million deaths worldwide. It is crucial that we analyze the reasons that have increased the mortality rate so we can prevent and treat such patients in a timely manner. Therefore, we decided to explore our data set which contains different factors such as Age Group, Geographical Region, and Condition Group(pre-existing health condition) of the population.

The two main questions we wanted to explore were: Which co-morbidity condition groups are most commonly associated with COVID-19 Deaths? & How does Age Group and Geographical Region affect the number of COVID-19 Deaths?

For our first question, we used a bar chart to see which Condition Group resulted in the highest number of COVID-19 deaths. Once we saw that "Respiratory disease" and "Circulatory disease" resulted in the most number of COVID-19 deaths, we wanted to further explore the sub-conditions within each disease. In order to do so, we ran an ANOVA test on the subcategories of the specific disease. The ANOVA test on the conditions within Respiratory disease resulted in a statistically significant p-value, which affirms our prediction that having a respiratory disease as a pre-existing health condition results increases the chances of dying with COVID-19.

For the second question, we used a t-test and also explore more of multiple regression procedures. Using the procedures Backward Elimination, AIC, and Adjusted R^2, we were able to derive a regression model with two proper predictors("Age" and "Respiratory disease"). The multiple regression procedures produce a regression model which affirms our prediction that age group affects the number of COVID-19 deaths. Additionally, we included a map representing the number of COVID-19 deaths in each state. Before performing any tests on our data, we expected that the pre-existing health condition of having a Respiratory Disease and that the age group of the population would have an affect on the number of COVID-19 deaths. We did wish to work with a data set with more continuous variables, as we could have performed better multiple regression procedures and classification on our data set; however, we are happy with the results from the tests, and we think that the results from our data analysis can help people understand how health conditions and one's age can increase/decrease their survival rate with COVID-19.