Investigating Early Stages of the COVID-19 Outbreak

Due: Wednesday April 29th

Team 2: Delaney Demark, Jenny Huang, Abby Mapes, Harshavardhan Srijay

Cleaning Data

Section 1: Introduction

Several recent reports suggest that older age groups show more severe symptoms in the face of COVID-19, causing the virus to be more deadly for older age groups. We want to test whether this is true by comparing the death rate of older individuals to the death rate of younger individuals. Our null hypothesis (H_0) is that the death rate for older individuals is the same as the death rate for younger individuals, while our alternative hypothesis (H_1) is that the death rate for older individuals is higher than that of younger individuals.

We plan on working with the data from the early stages of the COVID-19 outbreak from 1/20/2020 to 2/15/2020. This data set, from Kaggle, was first extracted from information provided by Johns Hopkins University. Johns Hopkins University collected this data from the World Health Organization, the Center for Disease Control and Prevention, the European Centre for Disease Prevention and Control, the National Health Commission of the People's Republic of China, among other state and national government health departments. Each observation in the data set is a case in which an individual tested positive for COVID-19. The variables include the ID number of the individual, the number that the case is in the country, the date the case was reported, the location of the case, the gender of the individual, the age of the individual, the age group of the individual, the date of the onset of symptoms, the date of the hospital visit, the start and end dates of exposure to the virus, if the individual visited Wuhan, if the individual was from Wuhan, if the patient died, and if the patient recovered.

Link to Data: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019- dataset#COVID19_line_list_data.csv

Section 2: Data Analysis Plan

Through our analysis, we will use death as our outcome variable by analyzing the proportion of patients who have died, indicated by a value of "yes" for death. To do so, we will use the following predictor variables: age_group, gender, visited_wuhan, from_wuhan, country.

Using these variables, we will attempt to determine not only if age affects one's survival outcome due to COVID-19, but also if any of these other characteristics are associated with one's survival outcome. To do this, we will employ a host of statistical techniques such as confidence intervals, hypothesis testing, and logistic regression to better understand the impact of these variables on survival outcome due to COVID-19.

To start, we will preform some preliminary exploratory data analysis to learn more about our data.

First, we will determine the death rate, proportion of those dead, of all patients in our dataset. As we can see, there is a small percentage, about 6%, of patients who died from COVID-19 in our data set.

```
# A tibble: 2 x 3
  death     n prop_dead
  <chr> <int>     <dbl>
```

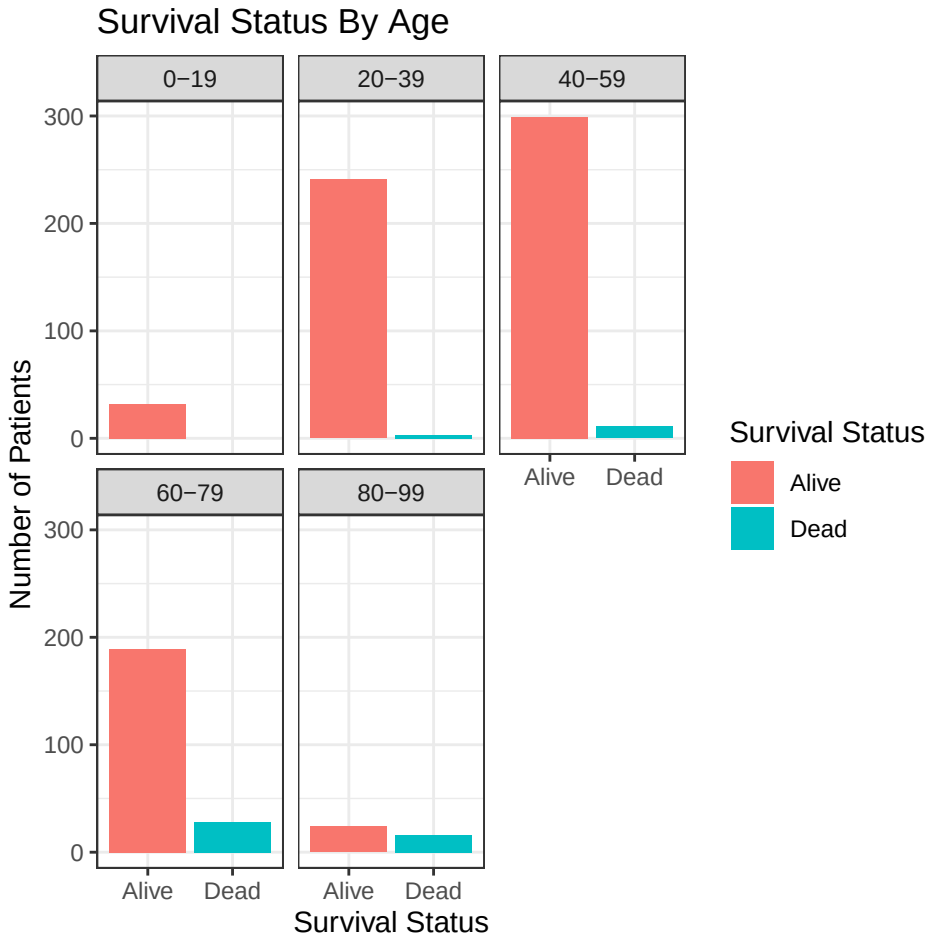
```
1 Alive  1022    0.942
2 Dead     63    0.0581
```

Now, we will look at the survival rate by age group.

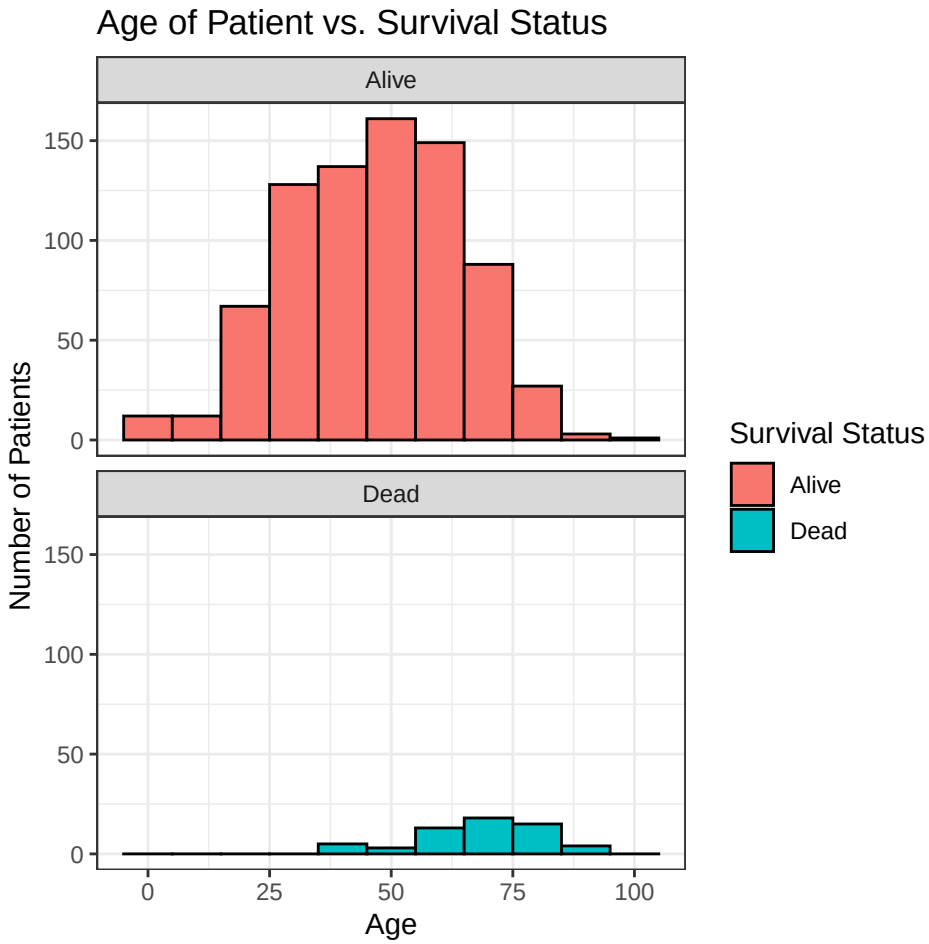
```
# A tibble: 11 x 4
# Groups:   age_group [6]
   age_group death     n prop_dead
   <chr>     <chr> <int>     <dbl>
 1 0-19      Alive    32    1
 2 20-39     Alive   241    0.988
 3 20-39     Dead      3    0.0123
 4 40-59     Alive   299    0.965
 5 40-59     Dead     11    0.0355
 6 60-79     Alive   189    0.871
 7 60-79     Dead     28    0.129
 8 80-99     Alive    24    0.6
 9 80-99     Dead     16    0.4
10 NA        Alive   237    0.979
11 NA        Dead      5    0.0207
```

The death rate is lowest for the 20-39 age group while highest for the 80-99 age group. Older age groups, then, tend to have a higher death rate.

Now, we will visualize survival status faceted by age_group.



We will also visualize the ages of patients for each survival status.



Below is a summary table of the mean age of patients who have died from COVID-19 and those that haven't, out of all patients in our data set that reported their ages. As we can see, the average age of patients that have died is greater than the average age of patients that are alive. These ages give us a reference point to determine what age will be considered "old" and what ages will be considered "young" for our exploratory data analysis.

```
# A tibble: 2 x 2
  death mean_age
  <chr>    <dbl>
1 Dead      68.6
2 Alive     48.1
```

To understand some of our other explanatory variables, we will calculate some statistics to get a sense of our data in terms of gender, country, and patients who have been to Wuhan recently. As we can see below, around 35% of patients are female, 48% of patients are male and 17% of patients are not classified. Noting that more male patients are included in our data set will be important and helpful when performing our exploratory data analysis.

```
# A tibble: 3 x 3
  gender     n  prop
  <chr>  <int> <dbl>
1 female   382 0.352
2 male     520 0.479
3 <NA>     183 0.169
```

Additionally, the patients in our data set come from 38 different countries.

```
# A tibble: 1 x 1
  total_countries
            <int>
1              38
```

However, as we can see below, 8 countries in our dataset have at least 1 reported death, with China having the most number of deaths by a significant margin, as of the date of our dataset. It will be helpful to know that only 8 of the 38 total countries in our data set have reported deaths when we consider the explanatory variable 'country' in our exploratory data analysis.

```
# A tibble: 8 x 3
# Groups:   country [8]
  country     people_dead prop_dead
  <chr>             <int>     <dbl>
1 China                39    0.198
2 France                2    0.0513
3 Hong Kong             2    0.0213
4 Iran                  4    0.222
5 Japan                 5    0.0263
6 Phillipines           1    0.333
7 South Korea           9    0.0789
8 Taiwan                1    0.0294
```

Additionally, from the summary table below, we see the mean age of the patients who have died for each country where there are reported deaths. In China, the mean age is about 71. In Taiwan, the mean age is 65. In Hong Kong, the mean age is about 54. However, since there were less than ten recorded deaths in countries other than China, the use of country as a predictor may yield misleading results, so we plan to group the categories into "China" and "Other Countries."

```
# A tibble: 7 x 2
  country     mean_death_age
  <chr>                <dbl>
1 Japan                 82.5
2 China                 71.1
3 France                70
4 Taiwan                65
5 South Korea           57.7
6 Hong Kong             54.5
7 Phillipines           44
```

From the table below, we see the majority of patients included in our data set are not from Wuhan, nor have they reported that they have previously visited Wuhan. Even though the majority of patients in our data set are not from or have been to Wuhan, it will be interesting to see if time in Wuhan is associated with one's survival outcome.

```
# A tibble: 3 x 2
  from_wuhan    prop
  <fct>        <dbl>
1 0          0.853
2 1          0.144
3 <NA>       0.00369
```

```
# A tibble: 2 x 2
  visiting_wuhan  prop
  <fct>          <dbl>
1 0              0.823
```

```
2 1              0.177
```

We plan to use the following statistical methods to answer our research questions about whether variables like age_group, gender, country, visited_wuhan, from_wuhan play a role in variable death rate:

a) Use estimation via bootstrap to create a confidence interval for the mean age of affected individuals:

For each age_group, we will take bootstrap samples, calculate the mean death rate, and obtain a 95% confidence interval for the mean death rate of that particular age group. If the confidence intervals for the mean death rates do not overlap, we can conclude that the mean death rate for the older age group is significantly different from the mean death rate of the younger age group.

If there are any confounding variables, we will estimate the mean age of affected individuals faceted by the confounding variables.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0365      0.0662
```

We are 95% confident that the true mean death rate of the population age 20-39 is between the range of (0.03668033 and 0.06598361).

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0455       0.159
```

We are 95% confident that the true mean death rate of the population age 80-99 is between the range of (0.04545455 and 0.1590909).

Since the two confidence intervals overlap, we cannot claim that there is significant difference between the mean death rate of individuals in the 80-99 age group with those in 20-39 age group. However, both the lower and upper end of the confidence interval for the older age group (0.04545455 and 0.1590909) is higher than the interval for the younger age group (0.03668033 and 0.06598361).

b) To get more information than simply seeing whether the confidence intervals overlap, we will also use simulation-based hypothesis testing to give the exact probability estimate (p-value) for the alternate hypothesis.

p is the mean death rate of a given age group:

$$H_0 : p_{young} = p_{old}$$
$$H_1 : p_{young} < p_{old}$$

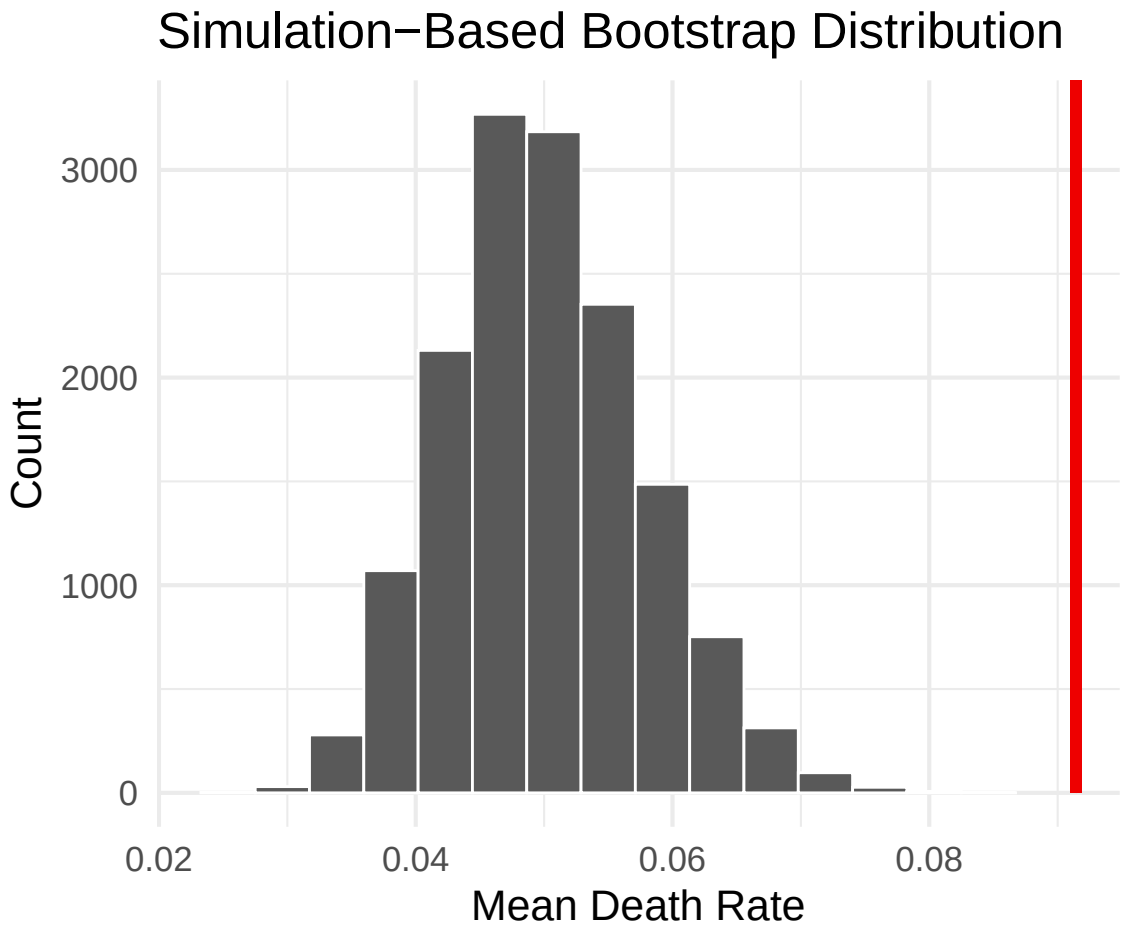
Here, by using the bootstrap distribution, we will see where the elderly mean death rate lies on the null distribution for the young mean death rate.

First, we estimate the mean death rate among the older age group through using bootstrap estimation.

```
[1] 0.09142742
```

Next, we create a null distribution for the mean death rate of the younger age group.

By graphing the sample mean for the older age group on the null distribution of the younger age group, we see that the probability of observing a sample mean the same or greater than that of the estimated mean for the older age group is very low.



We confirm this by doing a hypothesis test and getting the p_value, which comes out to be 6.666667e-05. Therefore, the probability of seeing a mean death rate of 0.09094273 (the sample mean death rate of the older age group) for the younger age group is approximately zero.

Thus, with an α level of 0.05, we have sufficient evidence to reject the null hypothesis that the mean death rate for the age group "80-99" is equal to the mean death rate of age group "20-39."

We will also conduct hypothesis tests for other variables such as gender or from_wuhan.

c) Calculate the conditional probability of P(Death | Age) to determine if there are any confounding variables: gender, visited_wuhan, from_wuhan, country. We will also keep in mind that the conclusions we draw about death rate may be confounded with variables that are not included in our dataset, such as smoking/drinking status, chronic disease status, BMI, etc.

d) Finally, use a logistic regression model to obtain predicted probabilities of an outcome of "death" given the explanatory variables in our model.

The covid_class dataset is created below, selecting for the variables of interest (death, age, gender, visiting_wuhan, from_wuhan) and removing all "NA" values.

We picked 100 random observations (10% of the data) to set aside as our testing data.

We created training and testing data sets covid_train and covid_test.

Also, a vector of the class labels for the training dataset, train_type, and a vector of the true class labels for the test dataset, true_type, were created.

The R function glm() requires for logistic regression that the response variable takes on values of 0 or 1. Create a new variable in the training dataset named bin_type that is 0 if the patient did not die, and 1 if the

patient died.

Below is a logistic regression model with relevant predictor variables.

```
# A tibble: 5 x 2
  term            estimate
  <chr>              <dbl>
1 (Intercept)      -9.04
2 gendermale        1.41
3 age               0.0787
4 visiting_wuhan1  -0.824
5 from_wuhan1       2.37
```

Holding all other variables constant, for each unit increase in age, we would expect the log-odds of a covid-19 case resulting in deaths to increase by approximately 0.07870473.

We plan to create a classifier to classify survival status based on the above variables. Then, we will determine how well these variables predict death by looking at the accuracy of the logistic model.

Section 3: Data

The dimensions of our dataset are 1,085 rows by 16 columns.

```
Observations: 1,085
Variables: 17
$ id                    <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
$ reporting_date        <date> 2020-01-20, 2020-01-20, 2020-01-21, 2020-01-...
$ location              <chr> "Shenzhen, Guangdong", "Shanghai", "Zhejiang"...
$ country               <chr> "China", "China", "China", "China", "China", ...
$ gender                <chr> "male", "female", "male", "female", "male", "...
$ age                   <dbl> 66, 56, 46, 60, 58, 44, 34, 37, 39, 56, 18, 3...
$ symptom_onset         <date> 20-01-03, 2020-01-15, 20-01-04, NA, NA, 2020...
$ if_onset_approximated <fct> 0, 0, 0, NA, NA, 0, 0, 0, 0, 0, 0, 0, NA, 0, ...
$ hosp_visit_date       <date> 20-01-11, 2020-01-15, 2020-01-17, 2020-01-19...
$ exposure_start        <date> 2019-12-29, NA, NA, NA, NA, NA, NA, 20-01-10...
$ exposure_end          <date> 20-01-04, 20-01-12, 20-01-03, NA, NA, NA, NA...
$ visiting_wuhan        <fct> 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, ...
$ from_wuhan            <fct> 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, ...
$ death                 <chr> "No", "No", "No", "No", "No", "No", "No", "No...
$ recovered             <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ source                <chr> "Shenzhen Municipal Health Commission", "Offi...
$ age_group             <chr> "60-79", "40-59", "40-59", "60-79", "40-59", ...
```

Section 4: Methods and Results

I. Confidence Intervals for Each Age Group

First, we will focus on one's age, and its impact on death rate from COVID-19. For each age group, we will take bootstrap samples, calculate the mean death rate, and obtain a 95% confidence interval for the true mean death rate of that particular age group to see if the mean death rates overlap for each age group. To do so, we must assume that our sample is representative of the population. Additionally, as seen below, in all age groups, there are more than 5 observations from our original sample, so, our original sample for each

confidence interval is greater than 5 for all cases, so it is not too small. Since our original sample is not too small and is representative of the population, we can create a bootstrap confidence interval.

```
# A tibble: 5 x 2
  age_group     n
  <chr>     <int>
1 0-19         32
2 20-39       244
3 40-59       310
4 60-79       217
5 80-99        40
```

First, we will create a 95% confidence interval for population mean death rate of individuals between 0 and 19 who are infected with COVID-19. As seen below, we are 95% confident that the actual population mean death rate of infected individuals between 0 and 19 is between 0.06 and 0.09.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0558      0.0871
```

Now, we will create a 95% confidence interval for population mean death rate of individuals between 20 and 39 who are infected with COVID-19. As seen below, we are 95% confident that the actual population mean death rate of infected individuals between 20 and 39 is between 0.04 and 0.07.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0365      0.0660
```

Now, we will create a 95% confidence interval for population mean death rate of individuals between 40 and 59 who are infected with COVID-19. As seen below, we are 95% confident that the actual population mean death rate of infected individuals between 40 and 59 is between 0.03 and 0.07.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0334      0.0739
```

Now, we will create a 95% confidence interval for population mean death rate of individuals between 60 and 79 who are infected with COVID-19. As seen below, we are 95% confident that the actual population mean death rate of infected individuals between 60 and 79 is between 0.03 and 0.08.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0293      0.0772
```

Finally, we will create a 95% confidence interval for population mean death rate of individuals between 80 and 99 who are infected with COVID-19. As seen below, we are 95% confident that the actual population mean death rate of infected individuals between 80 and 99 is between 0.05 and 0.16.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0455       0.161
```

The confidence intervals all seem to overlap, so we cannot claim that there is significant difference between the mean death rate of individuals in the different age groups. However, the upper end of the confidence

interval for the oldest age group, 80 to 99 year old patients, is at least .07 higher than the upper end of the confidence intervals for any other age group. We will discuss this further in our discussion.

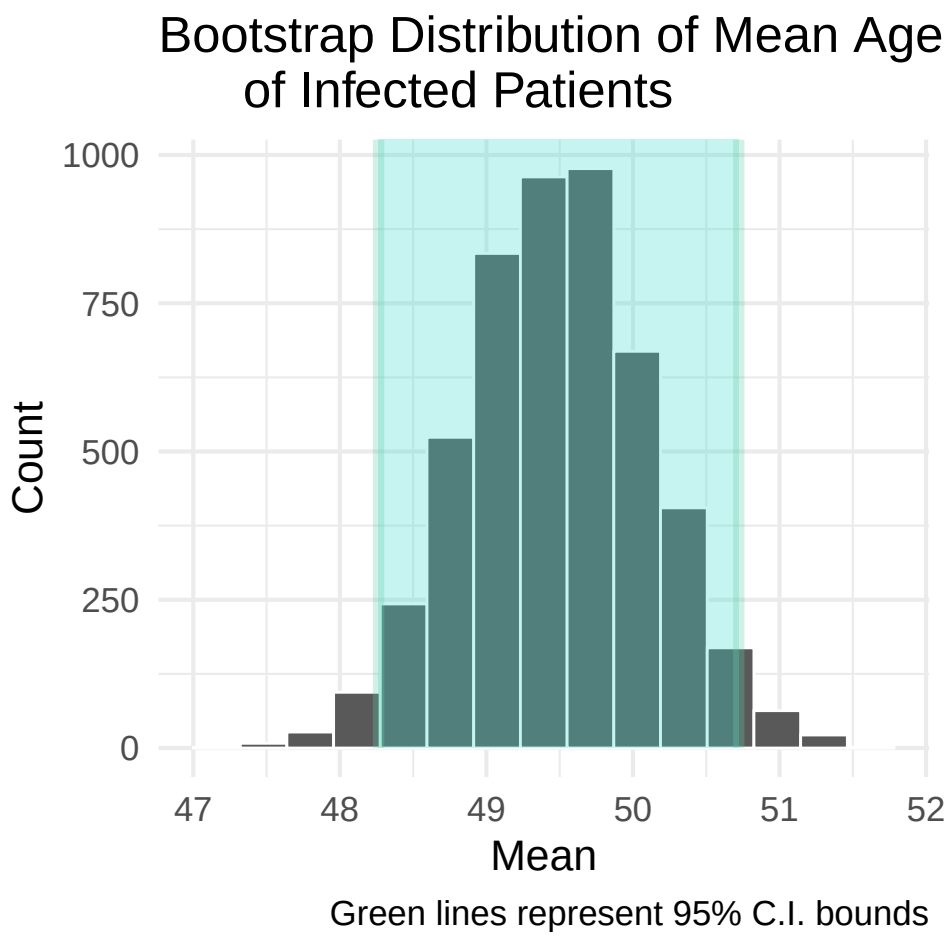
II. Mean Age of Affected Individuals

Now, rather than comparing confidence intervals for each age group, we will attempt to compare confidence intervals for the mean age of each disease outcome. This will show us how age impacts disease outcomes. We will create a 95% confidence interval for the population mean age of individuals infected with COVID-19, the population mean age of those infected with COVID-19 who survive, and the population mean age of those infected with COVID-19 who die. To do so, we must assume that our sample is representative of the population. Additionally, in all three cases, there are more than 5 observations in our original sample, since our dataset includes 1085 total infected patients, 1022 alive patients, and 63 dead patients, so, our original sample is greater than 5 for all cases, so it is not too small. Since our original sample is not too small and is representative of the population, we can create a bootstrap confidence interval.

```
# A tibble: 2 x 2
  death     n
  <chr> <int>
1 No     1022
2 Yes      63
```

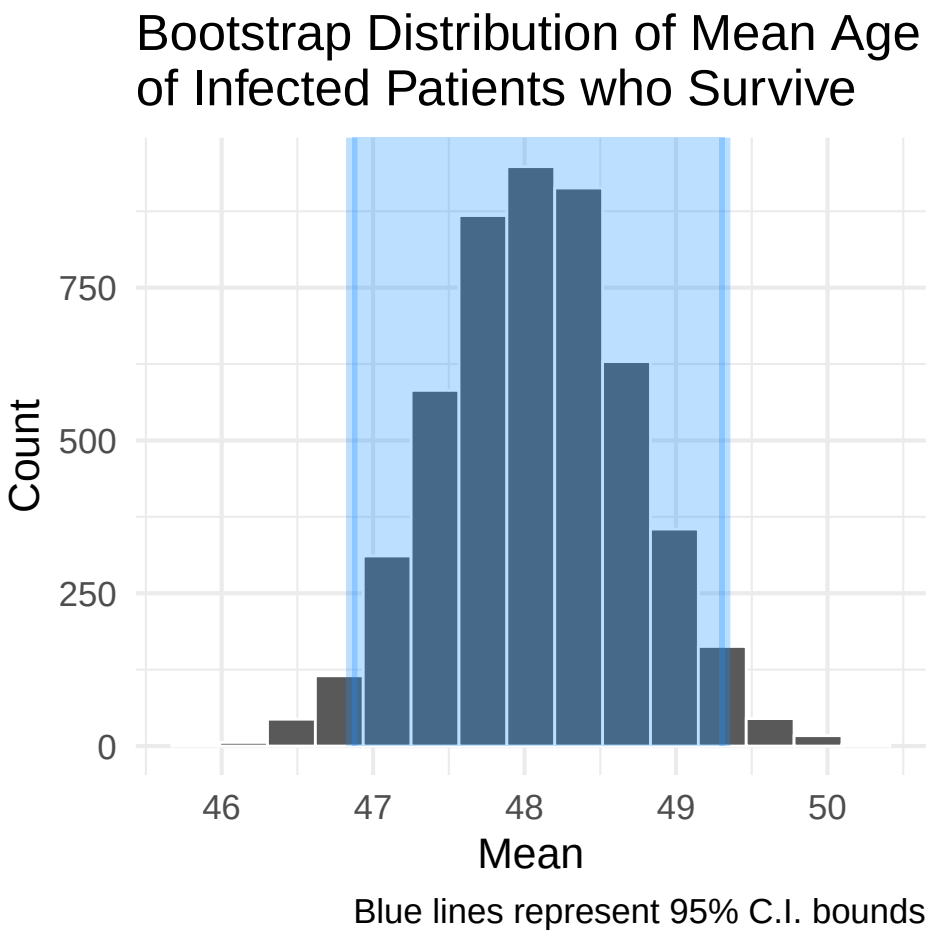
First, we will create a 95% confidence interval for the population mean age of individuals infected with COVID-19. As seen below, we are 95% confident that the actual population mean age of infected individuals is between 48.30 and 50.73.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        48.3        50.7
```



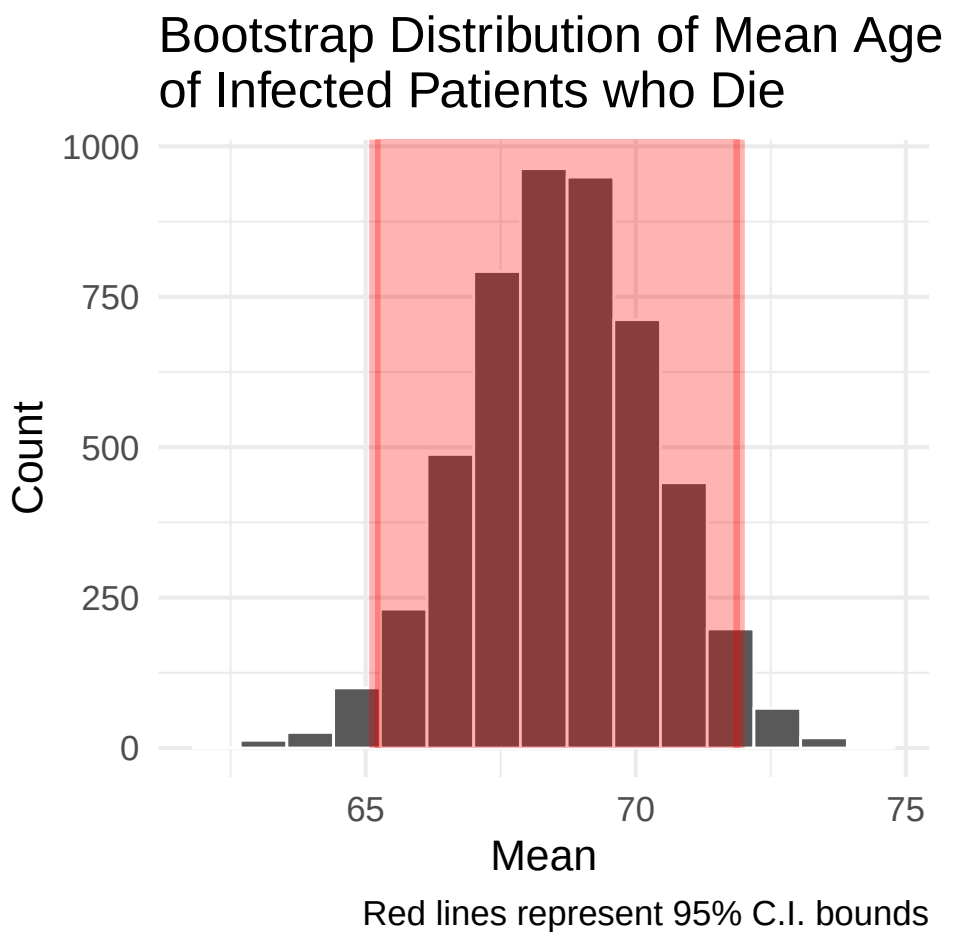
Now, we will create a 95% confidence interval for the population mean age of individuals infected with COVID-19 who survive. As seen below, we are 95% confident that the actual population mean age of infected individuals who survive COVID-19 is between 46.82 and 49.30.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        46.9        49.3
```



Now, we will create a 95% confidence interval for the population mean age of individuals infected with COVID-19 who die. As seen below, we are 95% confident that the actual population mean age of infected individuals who die from COVID-19 is between 65.13 and 71.98.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        65.2        71.9
```



From the three 95% confidence intervals calculated above, we observe that the 95% confidence interval for the true mean age of infected individuals who survive is the lowest, overlapping a year with the 95% confidence interval for the true mean age of all infected individuals. The 95% confidence interval for the true mean age of infected individuals who die is about 15 years higher than the other two intervals.

III. Hypothesis Testing

To further our understanding of whether death rate can be attributed to our variables of interest, such as age, we will now employ a more rigorous method of testing for the effect of these variables on death rate, beyond that of confidence intervals, using simulation-based hypothesis tests, under the $\alpha = 0.05$ level.

First, we conduct a one-sided hypothesis test for the age variable. We chose to delineate "old" observations from "young" observations through the 60 years of age mark, as the mean age of deaths due to COVID-19 in our sample was approximately 68, and the mean age of survival was approximately 48. Ideally, we would have chosen 58 then, to be our cut-off age, but as our age group was defined to be 40-59, we chose 60 to be our cut-off age.

Unlike our initial attempt in our data analysis section, here we will create a null distribution from the difference in death proportions between our two groups, rather than a distribution of mean death rates for only the "young" group. Though computationally very similar, we felt this approach was more intuitive to read from a plot.

In this hypothesis test, our null hypothesis was that the proportion of deaths in both the younger and older groups would be equal.

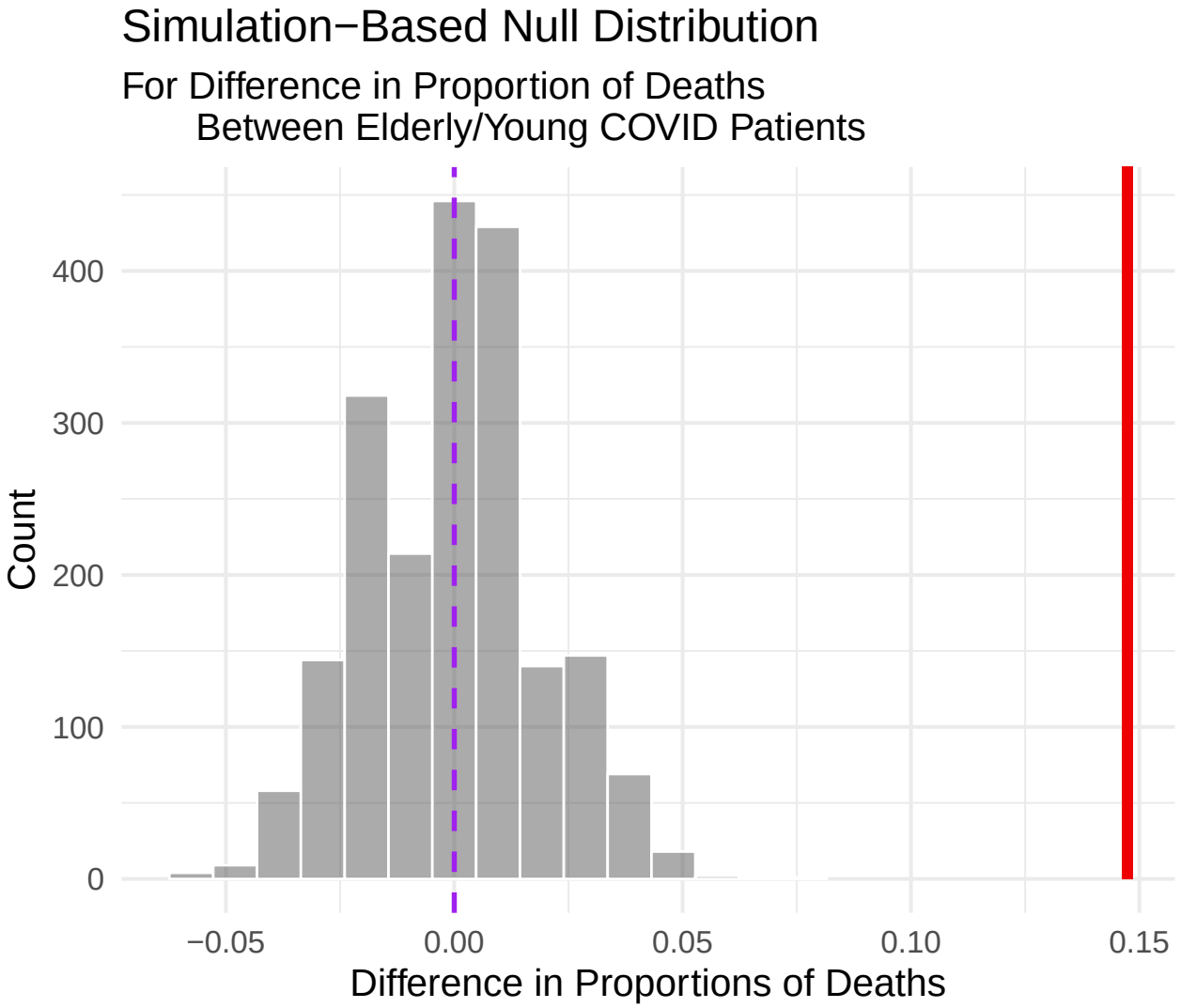
$$H_0 : p_{old} = p_{young}$$

Our alternative hypothesis was that the true proportion of deaths (p_{old}) in the older population is greater than the true proportion of deaths (p_{young}) in the younger age group.

$$H_A : p_{old} > p_{young}$$

```
# A tibble: 1 x 1
   stat
  <dbl>
1 0.147
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```



From this simulation-based hypothesis test, we obtained a p-value of 0, meaning we can reject our null hypothesis, as this is less than our α level of 0.05. This p-value means that assuming there is no difference

in the death rates of these two groups, the probability that the difference in death rate between older and younger people is 0.147 or more, which is what we observed in our sample data, is essentially 0. So, we can say that the death rate due to COVID-19 of people aged 60 or older is significantly higher than that of people younger than the age of 60. This makes intuitive sense from the plot depicted above as well, as the red line indicating our sample difference in death rate is significantly higher than the null distribution.

To test the effect of gender, we will calculate the difference in the proportion of deaths between men and women in our sample. Our null hypothesis is that there is no difference in the proportions of deaths between genders:

$$H_0 : p_{male} = p_{female}$$

Our alternative hypothesis is that the proportion of deaths between the two genders is not equal:

$$H_A : p_{male} \neq p_{female}$$

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.006
```

Here, our p-value was calculated to be 0.006. This is lower than our aforementioned α level of 0.05, meaning that we have enough evidence to suggest that the difference in proportion of deaths between men and women is significantly different.

Similarly, we conducted a simulation-based hypothesis test for whether a person is from Wuhan, using the following hypotheses:

$$H_0 : p_{fromWuhan} = p_{notfromWuhan}$$
$$H_A : p_{fromWuhan} \neq p_{notfromWuhan}$$

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

Here, our p-value was 0, meaning that we have enough evidence to reject our null hypothesis. This means that there is a significant difference in the proportion of deaths between people from Wuhan and people not from Wuhan.

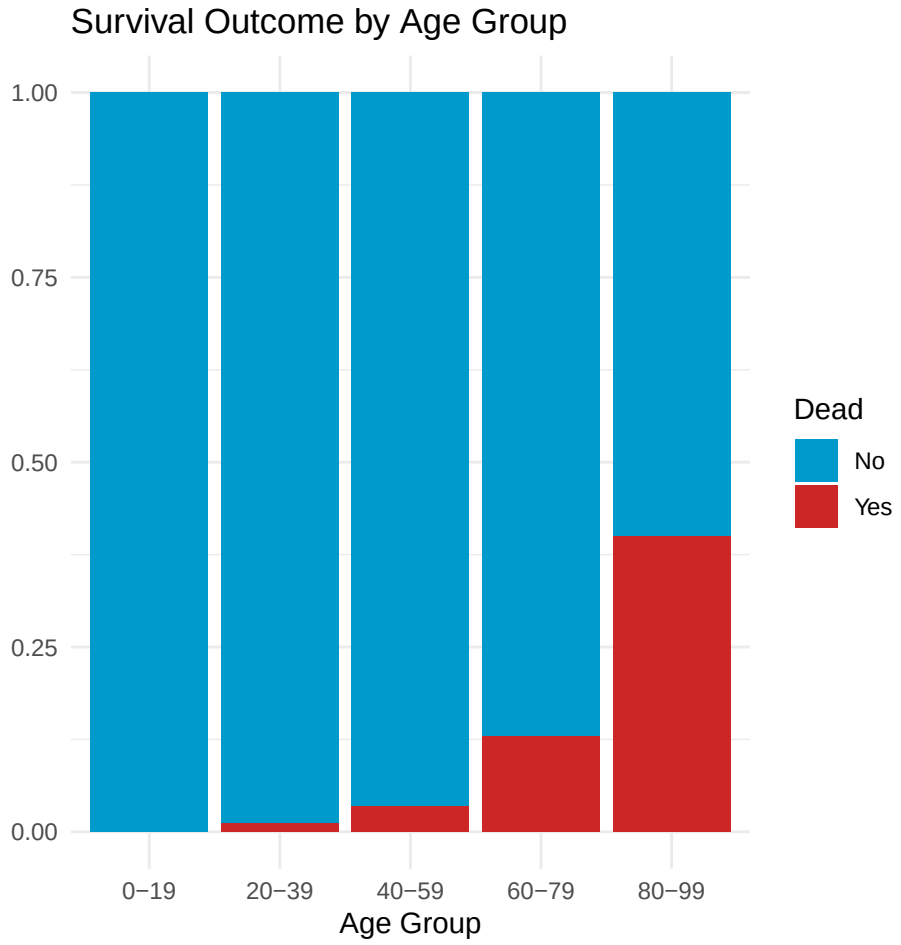
IV. Conditional Probabilities

Given our results from hypothesis testing, to find causal relationships between our aforementioned variables of interest and death rate, we must first calculate the conditional probability of P(Death | Age) for each age group and determine if there are any confounding variables: gender, visited_wuhan, from_wuhan, country.

Below, we calculate and visualize P(Dead|Age Group). From the visualization, we see that P(Dead|Age Group) increases as the age groups increase. For this section of our methods, we will filter out any patients with unrecorded ages in our data set, as we will not be able to take these into consideration with calculating P(Death | Age).

```
# A tibble: 9 x 3
# Groups:   age_group [5]
  age_group death prop_dead
  <chr>     <chr>     <dbl>
```

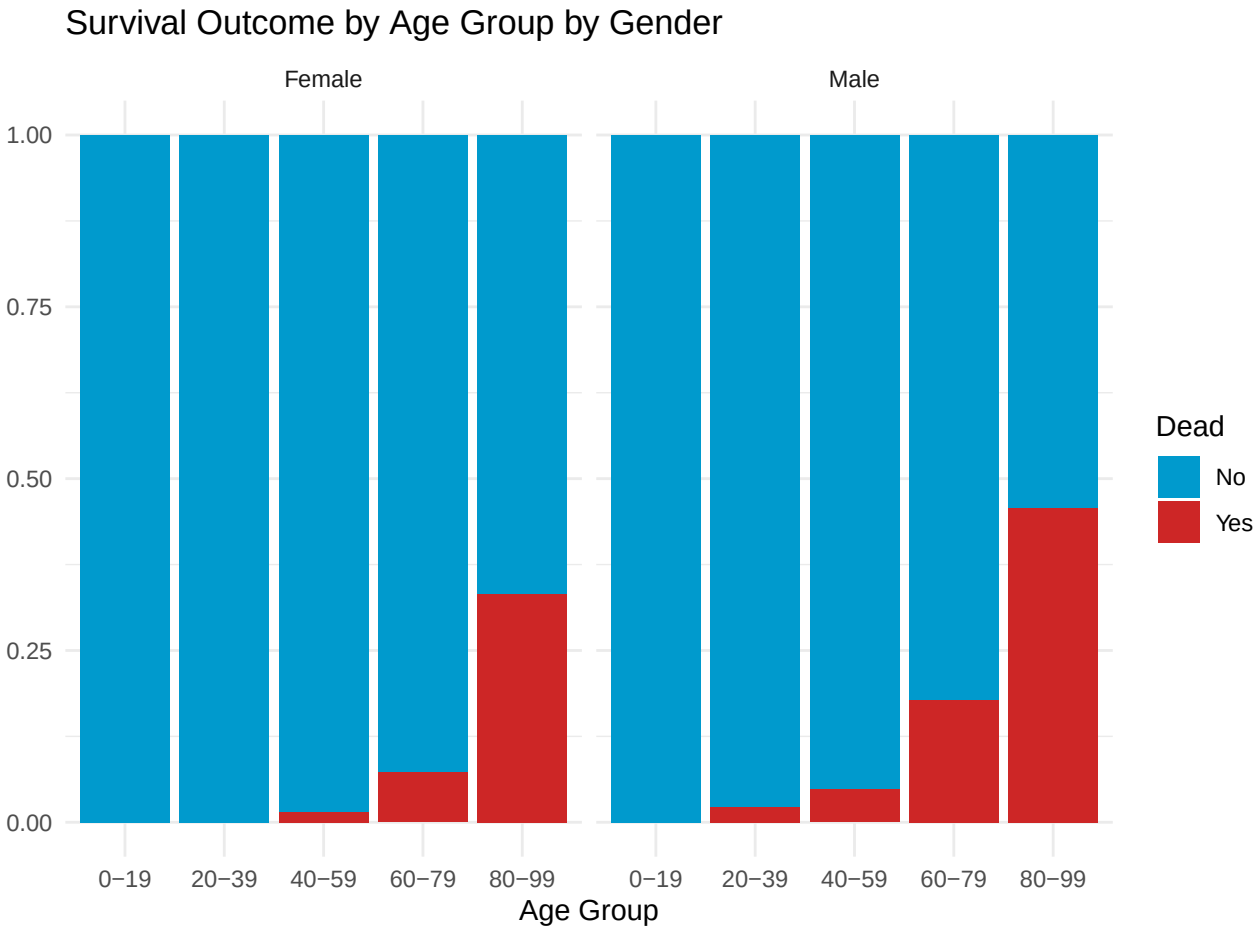
```
1 0-19      No       1
2 20-39     No       0.988
3 20-39     Yes      0.0123
4 40-59     No       0.965
5 40-59     Yes      0.0355
6 60-79     No       0.871
7 60-79     Yes      0.129
8 80-99     No       0.6
9 80-99     Yes      0.4
```



Now, we will divide the observations in our data by gender, allowing us to consider gender as a possible explanation of our data. As seen in the calculation and visualization of P(Dead|Age Group) for each gender, just as we saw in the visualization above, P(Dead|Age Group) increases as the age groups increase for both men and women. Thus, gender doesn't seem to act as a confounding variable that is correlated with both the explanatory and response variables.

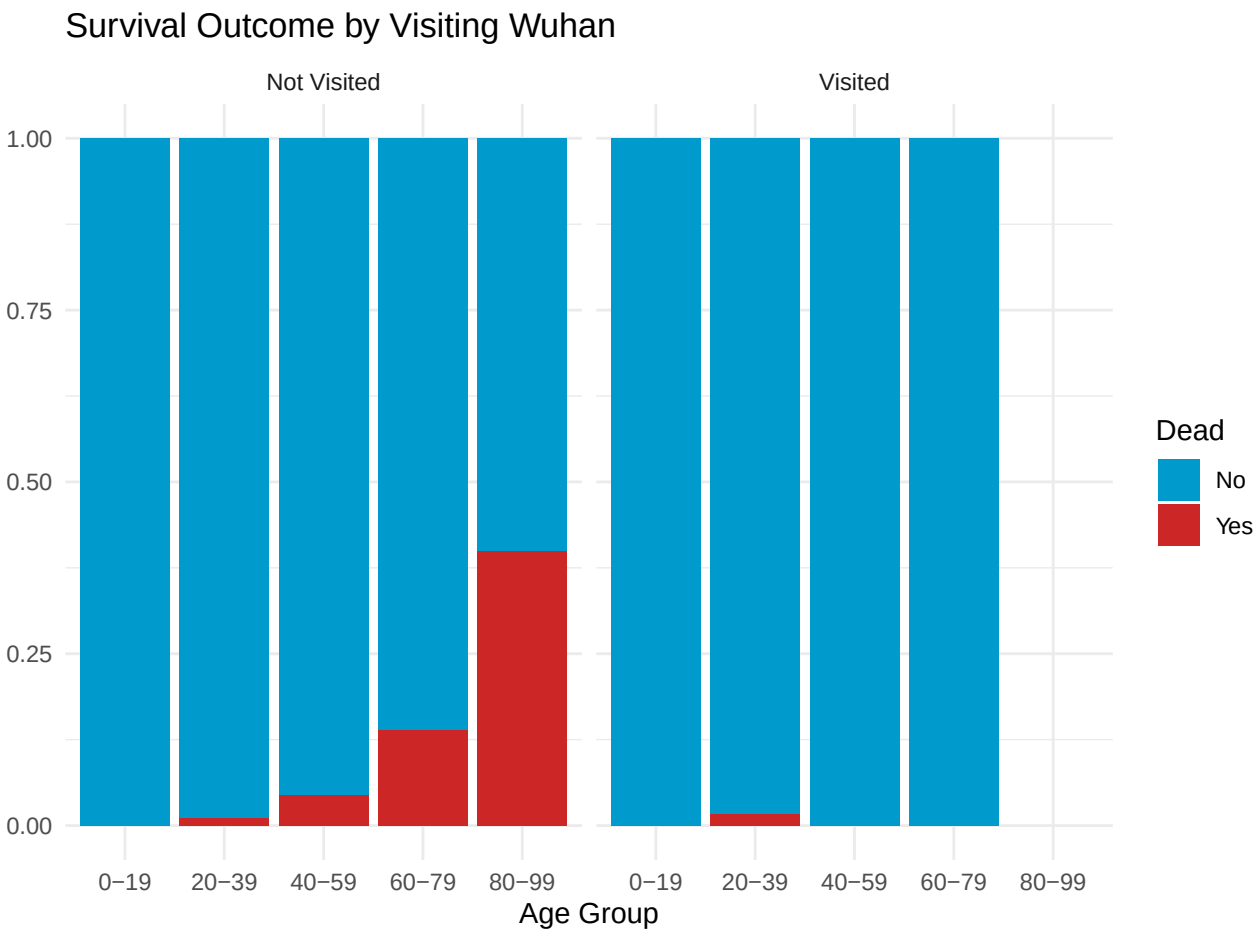
```
# A tibble: 17 x 4
# Groups:   gender, age_group [10]
   age_group death gender prop_dead
   <chr>     <chr> <chr>      <dbl>
 1 0-19      No    female     1
 2 20-39     No    female     1
 3 40-59     No    female     0.984
 4 40-59     Yes   female     0.0161
 5 60-79     No    female     0.926
```

```
 6 60-79     Yes   female    0.0737
 7 80-99     No    female    0.667
 8 80-99     Yes   female    0.333
 9 0-19      No    male      1
10 20-39     No    male      0.978
11 20-39     Yes   male      0.0224
12 40-59     No    male      0.951
13 40-59     Yes   male      0.0489
14 60-79     No    male      0.822
15 60-79     Yes   male      0.178
16 80-99     No    male      0.542
17 80-99     Yes   male      0.458
```



Now, we will divide the patients in our dataset by whether or not the patient has visited Wuhan recently, allowing us to consider visiting Wuhan as a possible explanation of our data. As seen in the calculation and visualization of P(Dead|Age Group), P(Dead|Age Group) increases as the age groups increase for patients that have not visited Wuhan. For patients that haven't visited Wuhan, P(Dead|Age Group) is 0. However, as we saw in our exploratory data analysis (Section 2), only 18% of patients in our data set had indicated that they had visited Wuhan recently and the other 82% indicated that they had not. Of the 18% of patients who indicated that they had visited Wuhan recently, only 1 of them died. So, we don't have many data points to calculate P(Dead|Age Group) for patients that have visited Wuhan recently, so we can't determine whether visiting wuhan acts as a confounding variable that is correlated with both the explanatory and response variables.

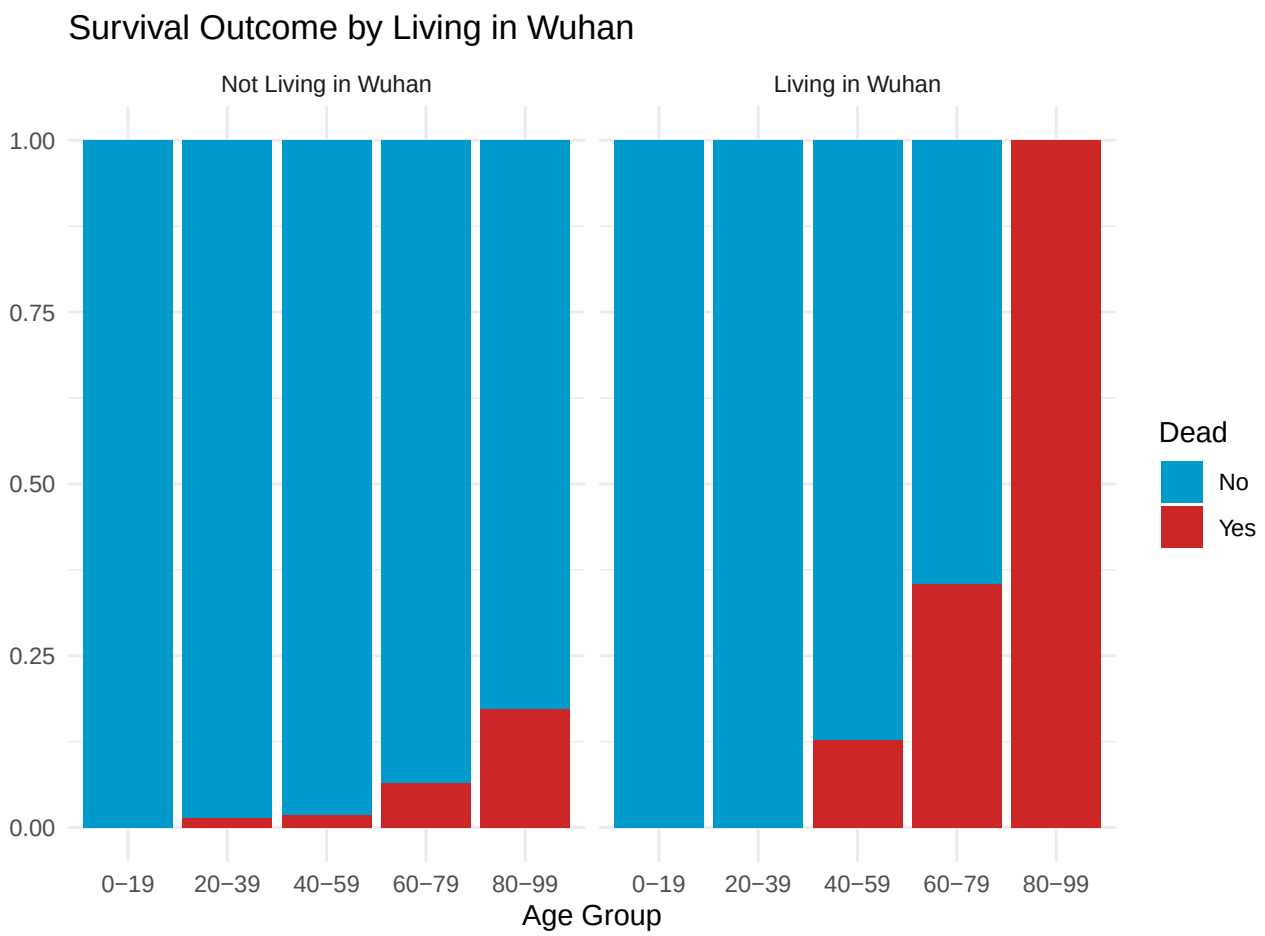
```
# A tibble: 14 x 4
# Groups:   visiting_wuhan, age_group [9]
   age_group death visiting_wuhan prop_dead
   <chr>     <chr> <fct>               <dbl>
 1 0-19      No    0                   1
 2 20-39     No    0                   0.989
 3 20-39     Yes   0                   0.0108
 4 40-59     No    0                   0.955
 5 40-59     Yes   0                   0.0451
 6 60-79     No    0                   0.86
 7 60-79     Yes   0                   0.14
 8 80-99     No    0                   0.6
 9 80-99     Yes   0                   0.4
10 0-19      No    1                   1
11 20-39     No    1                   0.983
12 20-39     Yes   1                   0.0172
13 40-59     No    1                   1
14 60-79     No    1                   1
```



Now, we will divide the patients in our dataset by whether or not the patient lives in Wuhan, allowing us to consider visiting Wuhan as a possible explanation of our data. As seen in the calculation and visualization of P(Dead|Age Group) below, P(Dead|Age Group) increases as the age groups increase for patients that live in Wuhan and patients that don't live in Wuhan. Interestingly, for patients that live in Wuhan, P(Dead|Age Group) is higher for all age groups over 40 years old than P(Dead|Age Group) for patients that don't live in

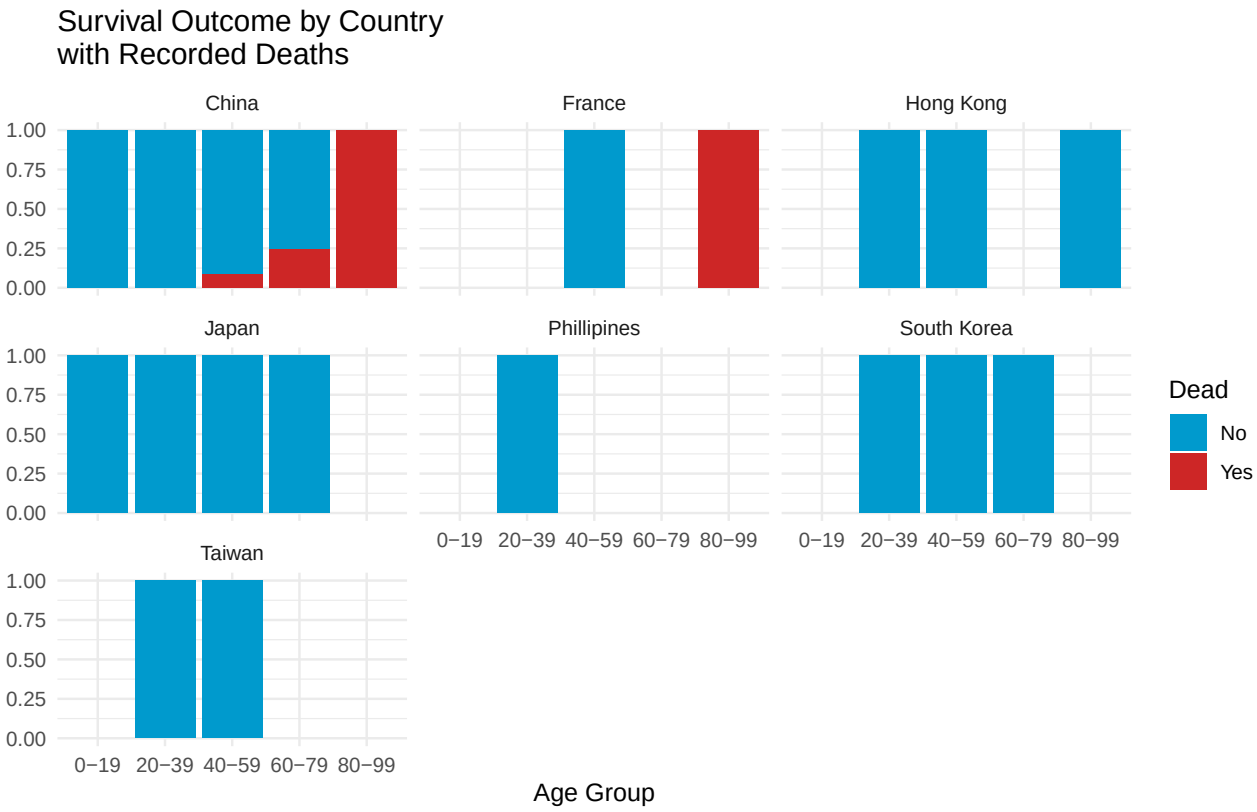
Wuhan. For 80-99 year old patients in our dataset, 100% of them who live in Wuhan have died, while only 17% of them who don't live in Wuhan have died. Keep in mind, the majority of patients in our data set are not from Wuhan and only 14% indicated that they lived in Wuhan. However, all 11 of the 80-99 year old patients from Wuhan died while only 5 of the 24 80-99 year old patients who were not from Wuhan died.

```
# A tibble: 16 x 5
# Groups:   from_wuhan, age_group [10]
   age_group death from_wuhan     n prop_dead
   <chr>     <chr> <fct>      <int>     <dbl>
 1 0-19      No    0             26    1
 2 20-39     No    0            198    0.985
 3 20-39     Yes   0              3    0.0149
 4 40-59     No    0            258    0.981
 5 40-59     Yes   0              5    0.0190
 6 60-79     No    0            157    0.935
 7 60-79     Yes   0             11    0.0655
 8 80-99     No    0             24    0.828
 9 80-99     Yes   0              5    0.172
10 0-19      No    1              5    1
11 20-39     No    1             41    1
12 40-59     No    1             41    0.872
13 40-59     Yes   1              6    0.128
14 60-79     No    1             31    0.646
15 60-79     Yes   1             17    0.354
16 80-99     Yes   1             11    1
```



Now, we will divide the patients in our data by country, allowing us to consider country of origin as a possible explanation of our data. We will only consider countries with recorded deaths in our data set. Note: we filtered any patients with unrecorded ages, so these deaths are not included. As seen in the calculation and visualization of P(Dead|Age Group) for each country below, P(Dead|Age) for each country group never decreases as age increases. However, we do not have enough data for each country to properly visualize P(Dead|Age) for each country and consider country of origin as a possible explanation of our data.

```
# A tibble: 22 x 4
# Groups:   country, age_group [20]
   age_group death country   prop_dead
   <chr>     <chr> <chr>         <dbl>
 1 0-19      No    China         1
 2 20-39     No    China         1
 3 40-59     No    China         0.909
 4 40-59     Yes   China         0.0909
 5 60-79     No    China         0.75
 6 60-79     Yes   China         0.25
 7 80-99     Yes   China         1
 8 40-59     No    France        1
 9 80-99     Yes   France        1
10 20-39     No    Hong Kong     1
# ... with 12 more rows
```



Ultimately, a patient's gender, whether they visited Wuhan recently, and what country they are from doesn't seem to affect the association between a patient's age group and their survival status from COVID-19.

V. Logistic Modeling to Classify Death

Given our understanding now of the confounding effects within our variables, we will now use a logistic regression model to obtain predicted probabilities for an outcome "death == yes" given the explanatory variables in our model. We will determine whether the variables of interest are good predictors of "death" through looking at the accuracy of our logistic regression model.

A new dataset called covid_class is created, selecting the variables (death, age, gender, visiting_wuhan, from_wuhan) and removing all "NA" values.

We picked 100 random observations to set aside as our testing data.

We then created training and testing data sets covid_train and covid_test.

Also, a vector of the class labels for the training dataset, train_type, and a vector of the true class labels for the test dataset, true_type, were created.

The R function glm() requires for logistic regression that the response variable takes on values of 0 or 1. We will create a new variable in the training dataset named bin_type that is 0 if the patient did not die, and 1 if the patient died.

Below is our logistic regression model.

```
# A tibble: 5 x 2
  term            estimate
  <chr>              <dbl>
1 (Intercept)      -9.04
2 gendermale        1.41
```

```
3 age               0.0787
4 visiting_wuhan1  -0.824
5 from_wuhan1       2.37
```

A patient that is female, age 0, and has neither visited nor is from Wuhan is predicted, on average, to have the log-odds of the death status being "Yes" (dead) to be -9.04349028; in this case, the y-intercept is simply a placeholder without practical meaning, as it is impractical for the age of patient to be 0.

Holding all other variables constant, for each unit increase in age, we would expect the log-odds of a covid case resulting in death to increase by approximately 0.07870473, which is consistent with the trend of older individuals having higher death rates.

Holding all other variables constant, if a patient is male, we would expect the log-odds of a covid case resulting in death to increase by approximately 1.40861782, which is consistent with the trend of males having higher death rates.

Holding all other variables constant, if a patient has visited Wuhan, we would expect the log-odds of a covid case resulting in death to decrease by approximately 0.82439051.

Holding all other variables constant, if a patient is from Wuhan, we would expect the log-odds of a covid case resulting in death to increase by approximately 2.37039736. This is consistent with the finding in our hypothesis test, that there is a significant difference in proportion of deaths between people from Wuhan and people not from Wuhan.

Here, we created a classifier with the cut-off value of 0.5. We will display the first 15 classifications:

```
# A tibble: 15 x 2
   pred_probs classified_types
        <dbl> <chr>
 1      0.407 No
 2      0.052 No
 3      0.008 No
 4      0.014 No
 5      0.152 No
 6      0.008 No
 7      0.075 No
 8      0.019 No
 9      0.013 No
10      0.052 No
11      0.075 No
12      0.15  No
13      0.002 No
14      0.076 No
15      0.009 No
```

```
[1] 0.9
```

As we can see above, the prediction accuracy is 90% for our logistic regression model. Given this fairly high prediction accuracy, we find the predictor variables age, gender, visiting_wuhan, and from_wuhan to be good predictors of death.

Section 5: Discussion

I. Confidence Intervals for Each Age Group

Using bootstrapping, we were able to calculate 95% confidence intervals for the mean death rate of each age group in our dataset. For the 0-19 age group, our 95% confidence interval was (.06, .09). For the 20-39 age

group, our 95% confidence interval was (.04, .07). For the 40-59 age group, our 95% confidence interval was (.03, .07). For the 60-79 age group, our 95% confidence interval was (.03, .08). Finally, for the 80-99 age group, our 95% confidence interval was (.05, .16). These intervals all overlap, so we cannot claim that there is a significant difference in mean death rate between patients in different age groups.

We do note that the confidence interval for 80-99 year old patients includes higher death rates than the intervals for younger age groups. This does not mean that the true mean death rate for 80-99 year olds is higher than other age groups. However, we are 95% confident that the true mean death rate for 80-99 year olds is contained in an interval that includes higher death rates than the 95% confidence intervals for younger age groups.

Additionally, we see that the width of the 95% confidence intervals for the 0-19, 20-39, 40-59, and 60-79 age groups range from around .03 to .05, while the width of the 95% confidence interval for the 80-99 age group is around .11. So, our 95% confidence interval for the oldest age group is wider than the confidence interval for all other age groups. Since the confidence level is fixed for each age group, this indicates that the standard error is higher for the 80-99 year old age group. The standard error depends on the sample size and variation. In our dataset, there are 40 patients that are 80-99 years old. This sample size is smaller than some of the other age groups in our dataset; however it is larger than the 0-19 year old age group that includes 32 patients. The width of the 95% confidence interval for the 0-19 age group was around .03, which is much smaller than .11, indicating that there is more variation in death rate within the 80-99 year old age group than variation in death rate within the 0-19 year old age group.

II. Mean Age of Affected Individuals

To help us further determine if there exists a relationship between age and survival, we created 95% confidence intervals for the mean age of all patients, the patients that survived, and the patients that died. The 95% confidence interval for the surviving patients was slightly lower than the confidence interval for all patients, whereas the confidence interval for the mean age of patients who died was approximately 15 years higher than the others.

III. Hypothesis Testing

To answer our research question of whether age, gender, and other variables affect survival outcome, the hypothesis tests we created to compare the differences in mean death rates is effective because it gives us the probability that our observed difference in means is due to random chance alone, giving us a better sense of the variables most likely associated with death. It is slightly more rigorous than a confidence interval-based test, but it is still not perfect.

For example, though we found all of our results to be statistically significant, there is still a chance that we simply committed a type I error, where we mistakenly rejected a true null hypothesis. Though an α level of 0.05 limits this possibility, there is still a 5% chance such a mistake occurs.

Another notable disability of our hypothesis tests is that they fail to find whether our variables of interest are linked in a causal relationship to death rate. Though we did find statistically significant differences in death rates between some groups, the hypothesis test fails to identify whether such a difference is due to the variable in question itself, or some other "confounding" variable that we failed to recognize. So, further analysis through logistic regression must be used to fully answer our question, as the results obtained from these tests, though meaningful, are incomplete.

IV. Conditional Probabilities

In the early stages of our data analysis, we found that the death rates for each age group increased as the ages increased, so we explored the impact of a variety of variables on the death rate of each age group.

For each of the potential explanatory variables, we analyzed the proportion of patients that died, grouping them by the categories of the variable. First, we investigated whether gender was a possible confounding variable to explain the increase in death rates among the age groups. Upon examining the proportions of dead patients per age group faceted by gender, it seemed as though gender was not a confounding variable. Although we wanted to determine if patients visiting Wuhan had an impact on the death rate, only 18% of patients indicated that they had visited Wuhan, so the data for this relationship was insufficient to reach any conclusions. Next, we looked into the impact of the patients living in Wuhan on the survival rate within each age group; we found that the death rate among patients living in Wuhan was higher than those not living in Wuhan for age groups above 40. Again, the limitations of the data set prevented us from concluding any significant impact the country of the case could have on the death rate. Therefore, we found that living in Wuhan had the greatest impact on the death rates within each age group, especially for the oldest age group.

V. Logistic Modeling to Classify Death

Through performing a logistic regression, we were able to create a classifier for survival status (death = yes or no). Through our logistic regression model, we saw that, holding all other variables constant, for each unit increase in age, we would expect the log-odds of a covid case resulting in death to increase by approximately 0.07870473. This pattern was consistent with our hypothesis that cases in which the patient is older have higher likelihoods of resulting in death. We also saw that, if a patient is male, we would expect the log-odds of a covid case resulting in death to increase by approximately 1.40861782, which is consistent with the trend of males having higher death rates. From the accuracy of the classifier (90%) we created from our model, we determined that our variables of interest are good predictors of death.

To answer our research questions, implementing a logistic regression was appropriate, since our response variable (death) is categorical and our predictor variables include both categorical and continuous variables. The conditions to perform logistic regression are: 1. the observations are independent of one another and 2. the predictors are not highly correlated. Given the nature of disease spread, we cannot be certain that our observations are independent from one another– one case could have been linked to another. Furthermore, there may be some correlation between predictor variables "from_wuhan" and "visited_wuhan," as people who visit wuhan are likely to be from regions around Wuhan. These above-mentioned factors may act to lower the reliability of our logistic model.

Finally, there were several missing values present among each of the variables of interest. Given that we filtered out these "NA" values before creating the model, we have to consider that the incompleteness of the data makes our model our less accurate and representative of the true situation.

VI. Reliability and Validity of Data

Throughout our data exploration, we were often unable to conclude whether
there were correlations among different variables and the death rates because often there was insufficient data to do so. Even though our dataset included over 1,000 observations, many of these observations did not include values for all 17 variables. Many patients in our dataset had unidentified gender and age which made it hard to determine whether these variables impacted patient survival status since we couldn't consider these observations. Additionally, in creating confidence intervals and performing hypothesis tests, we had to assume that our dataset was representative of the population. However, COVID-19 is a pandemic that infects hundreds of people every day. Thus, any dataset including COVID-19 patients becomes quickly outdated as the virus spreads and infects more individuals. Also, the impacts of a virus varies based on individuals' situations. Countries have different medical resources, climates, and populations. So, it is hard to determine if our dataset is representative of an overall population that varies within itself. Ultimately, due to the difficulties in obtaining reliable, complete and valid data about a real, current global pandemic, we proceeded with caution when performing our statistical analysis, keeping these limitations in mind.

VII. Next Steps

If we were to start over on this project, we would pick a data set with more observations outside of China. For example, when considering country of origin as a possible determiner of death rate, we did not have enough data for each country to properly visualize P(Dead|Age) for each country. More inclusive and larger data sets about COVID-19 are easier accessible now, but they weren't when we started this project. Furthermore, since our research question focused on whether age was a determiner of death rate, age could easily have been confounded with other variables such as heart disease or diabetes status or other pre-existing health conditions. Next time we explore the COVID-19 death rates, we would find a data set containing more information about underlying health issues and a wider set of patients from more regions. Finally, another method we could have used to determine the significance of each explanatory variable on death rate would be inference for regression. The next steps we would do to continue this project would be to update the data frame with current COVID-19 data. Our data frame is small, contains very few data points outside of China, and only focuses on the early days of the pandemic (from Jan. - Feb.) The vast amount of data that has since been added and collected about the pandemic would provide more reliable conclusions to our research questions surrounding whether death outcome is correlated with factors like age and gender. Moreover, given that there are very few data points outside of Asian countries, we were unable to draw meaningful conclusions to our research questions outside of these countries. Thus, our exploration cannot be extended accurately to the global population. It would be interesting to find whether our conclusions would hold in the U.S. and Europe, now that COVID-19 has spread to other countries around the globe. Ideally, we would want to create a way to update our data so that our analyses can be done and adjusted on real time data as the pandemic progresses.