Investigating the COVID-19 Outbreak

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 (H0) is that the death rate for older individuals is the same as the death rate for younger individuals, while our alternative hypothesis (H1) 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.

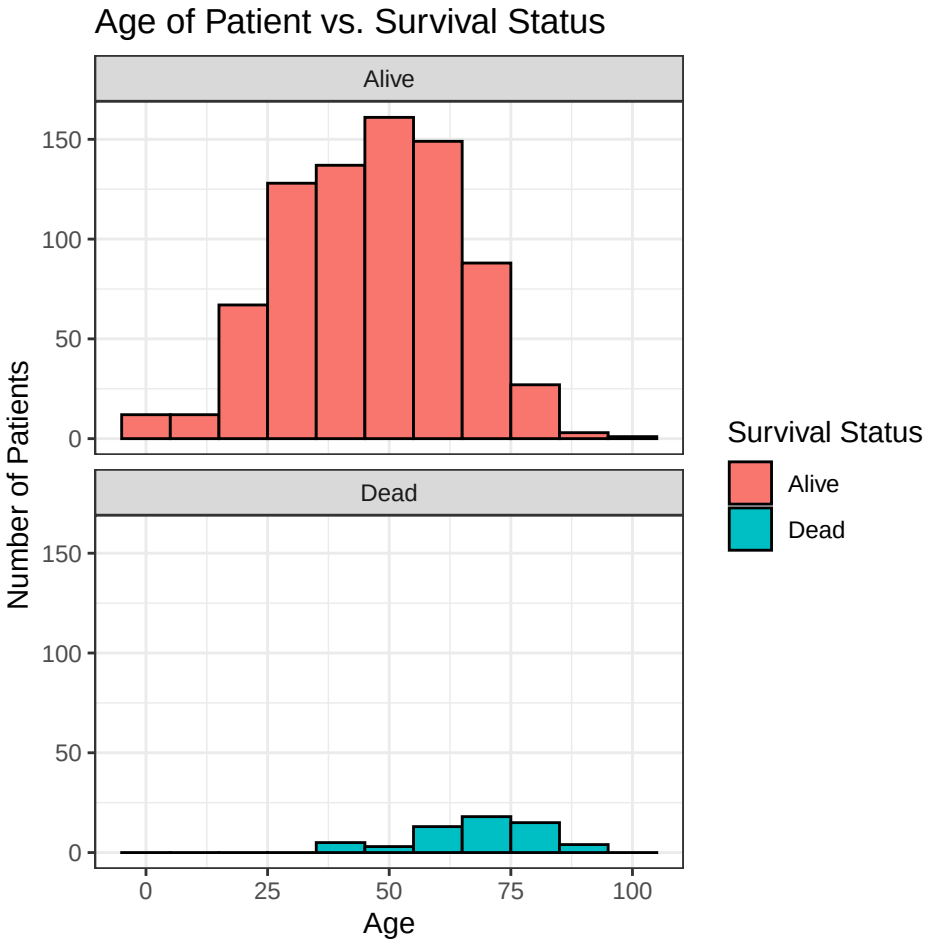
Using these variables, we will attempt to determine not only if age affects one's survival outcome due to COVID-19, but also if any of these other characteristics are associated with one's survival outcome.

To 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 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.

```
# 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
```

In our exploratory data analysis, we plan to use the following statistical methods to answer some of these questions:

a) Calculate the conditional probability of P(Death | Age) and determine if there are any confounding variables: gender, visited_wuhan, from_wuhan, country.

b) Take a bootstrap sample to estimate a confidence interval for the mean age of affected individuals. If there are any confounding variables, we will estimate the mean age of affected individuals faceted by the confounding variables.

c) To answer the question of whether age_group is associated with death rate, we will be using simulation via bootstrap.

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.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1      0.0363      0.0658
```

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).

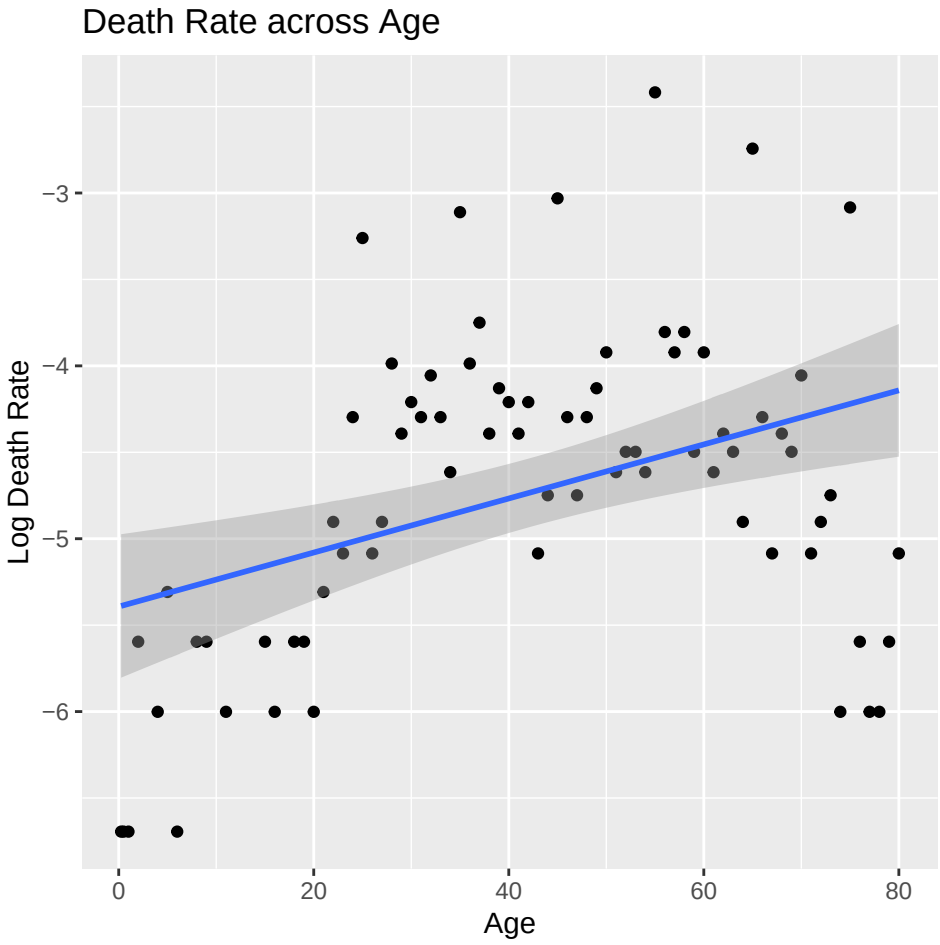
Since the two confidence intervals overlap, we cannot claim that there is significant difference between the mean death rate of individuals in the 80-99 age group with those in 20-39 age group. However, both the lower and upper end of the confidence interval for the older age group (0.04545455 and 0.1590909)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

In addition to age, we will repeat this process for the other explanatory variables: gender, country, visited_wuhan, and from_wuhan.

d) Another way we will measure whether the explanatory variable age, gender, visited_wuhan, from_wuhan, and country are significant determiners of death rate is by creating logistic regression models.

```
# A tibble: 2 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)  -5.39
2 age           0.0156
```

```
[1] 0.1500937
```



From this visualization, we can see that death rate peaks around age 60.

After creating logistic regression models for each variable, we will determine whether each variable is significant in influencing death rates by looking at P-value for each coefficient.

Then, we can use AIC/BIC model selection criteria to select for a model with the highest adjusted R^2 using backwards elimination.

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...
```

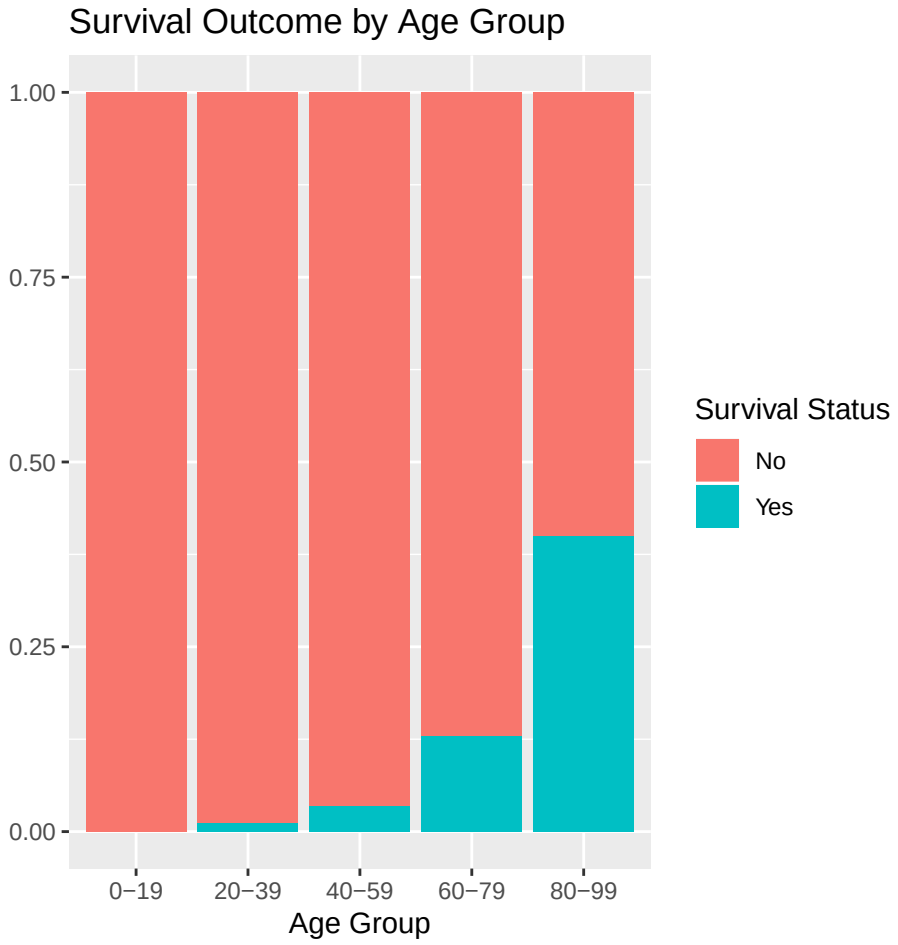
```
$ visiting_wuhan        <fct> 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, ...
$ from_wuhan            <fct> 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, ...
$ death                 <chr> "No", "No", "No", "No", "No", "No", "No", "No...
$ recovered             <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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. Conditional Probabilities

First, we will 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 part, 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).

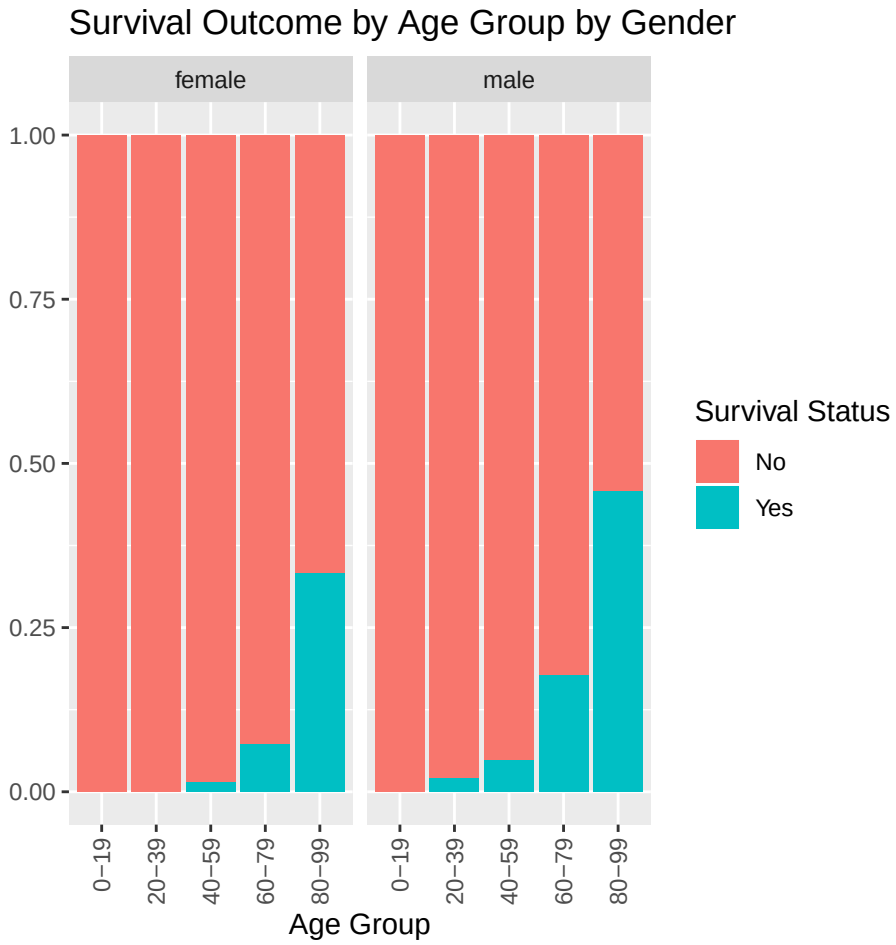
```
# A tibble: 9 x 3
# Groups:   age_group [5]
  age_group death prop_dead
  <chr>     <chr>     <dbl>
1 0-19      No        1
2 20-39     No        0.988
3 20-39     Yes       0.0123
4 40-59     No        0.965
5 40-59     Yes       0.0355
6 60-79     No        0.871
7 60-79     Yes       0.129
8 80-99     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
```

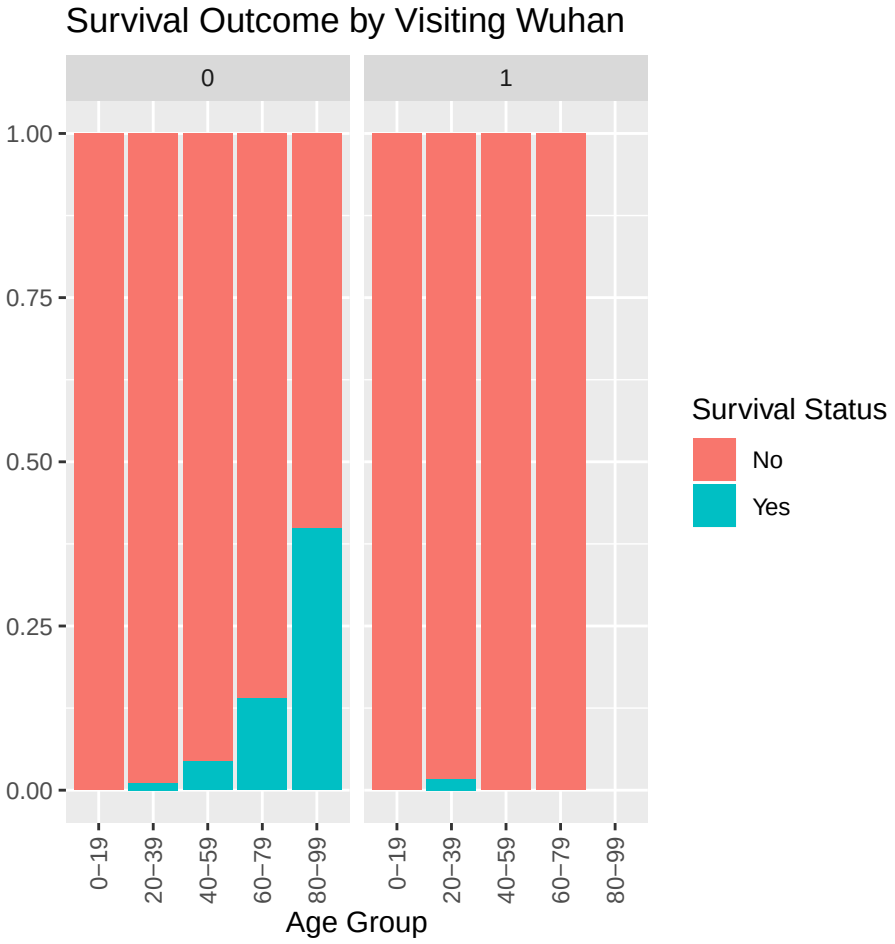
```
15 60-79     Yes   male      0.178
16 80-99     No    male      0.542
17 80-99     Yes   male      0.458
```



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
```

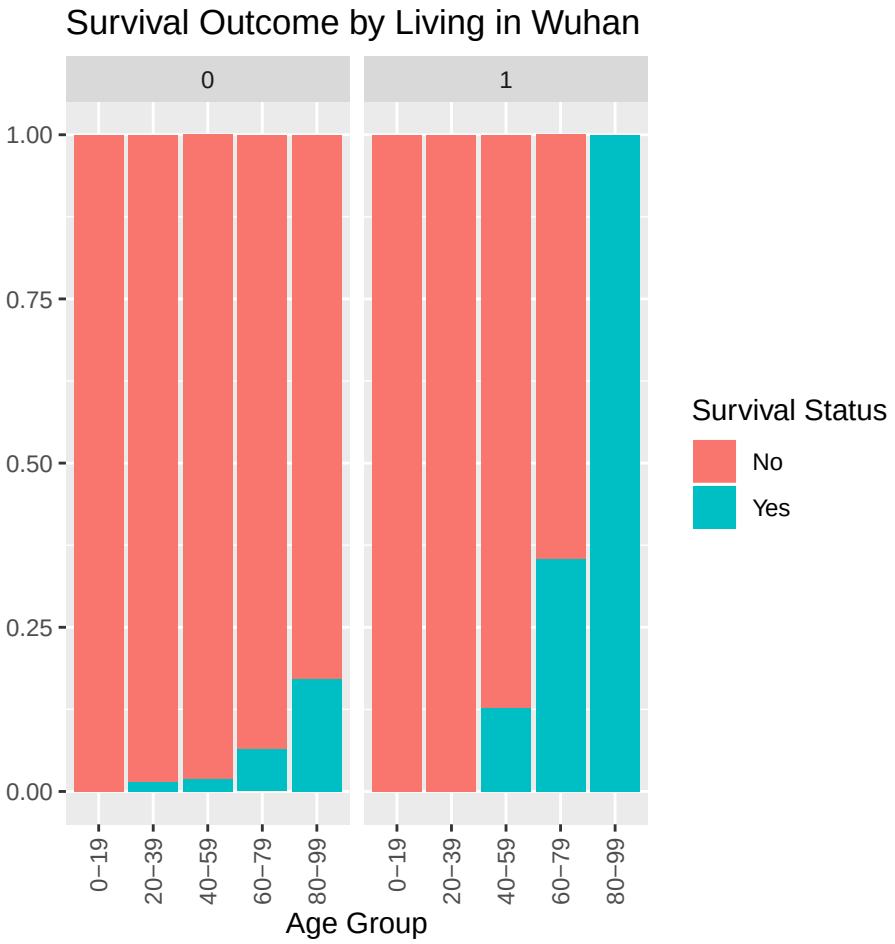
```
 7 60-79     Yes   0                0.14
 8 80-99     No    0                0.6
 9 80-99     Yes   0                0.4
10 0-19      No    1                1
11 20-39     No    1                0.983
12 20-39     Yes   1                0.0172
13 40-59     No    1                1
14 60-79     No    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 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
```

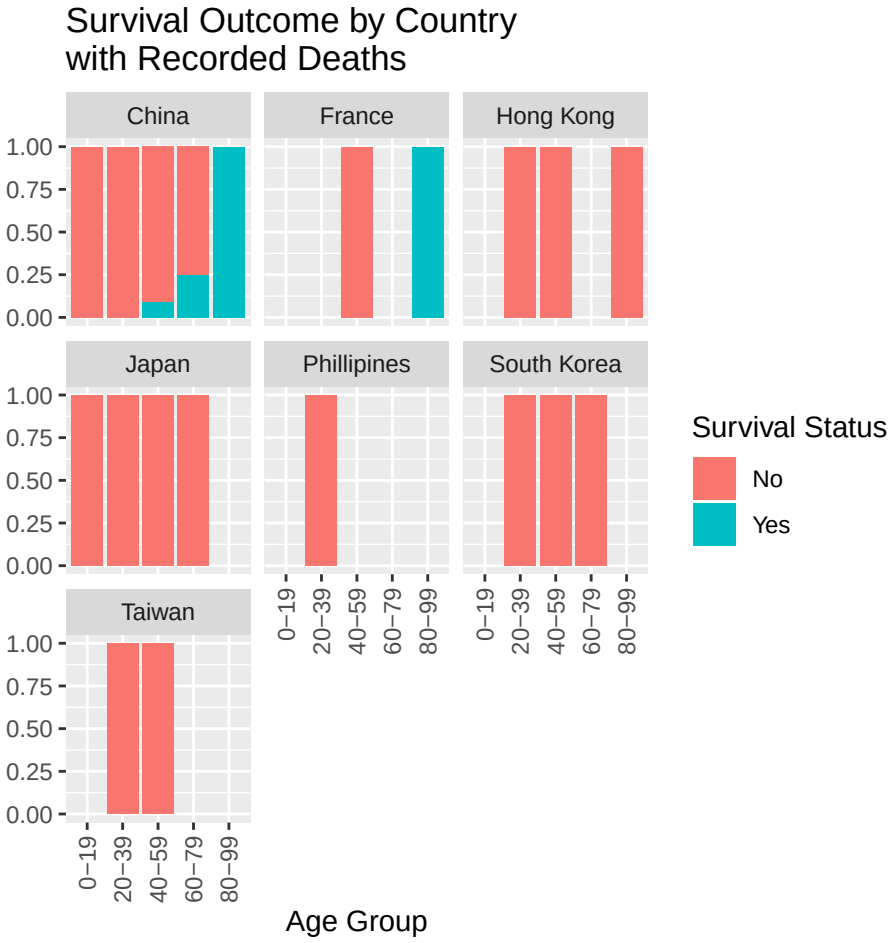
```
 3 20-39     Yes   0             3    0.0149
 4 40-59     No    0           258    0.981
 5 40-59     Yes   0             5    0.0190
 6 60-79     No    0           157    0.935
 7 60-79     Yes   0            11    0.0655
 8 80-99     No    0            24    0.828
 9 80-99     Yes   0             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, 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 increeases. However, we do not have enough data for each country to properly visualze 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
```



Ulitmately, a patient's gender, whether they visited Wuhan recently, whehter they live in Wuhan, 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.

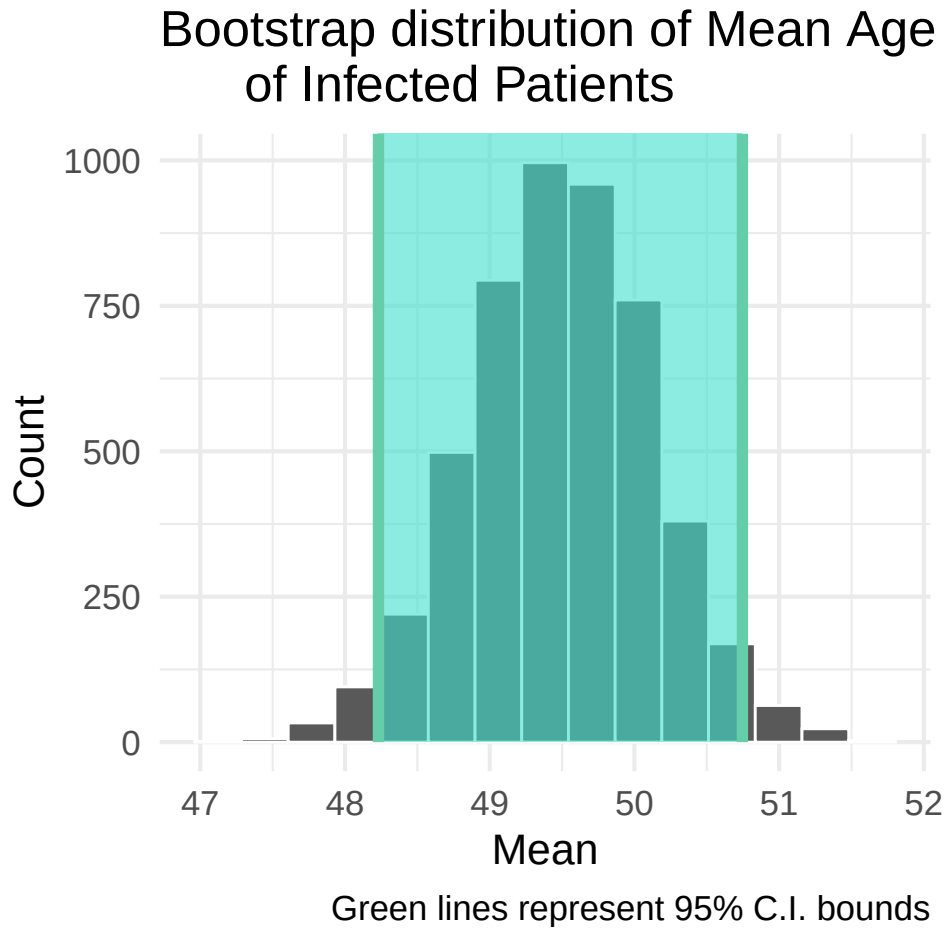
II. Mean Age of Affected Individuals

Now, 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 data 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.

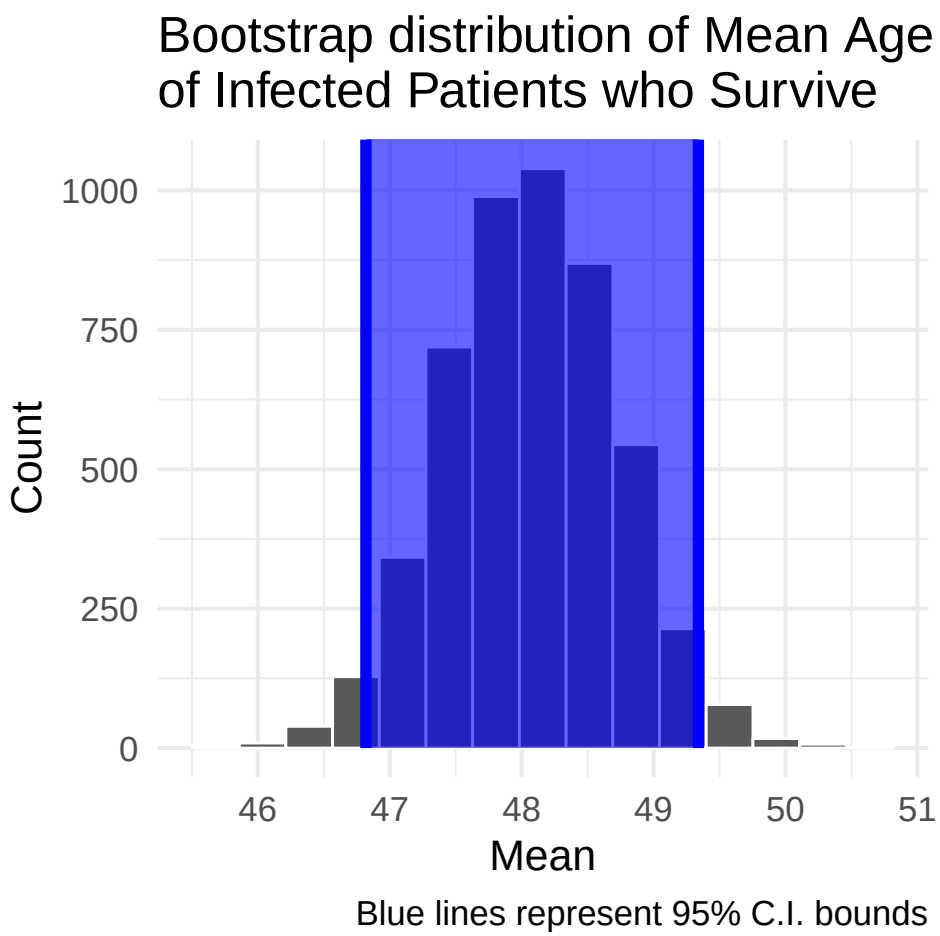
First, we will create a 95% confidence interval for the population mean age of individuals infected with COVID-19. As seen below, we are 95% confident that the actual population mean age of infected individuals is between 48.30 and 50.73.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        48.2        50.7
```



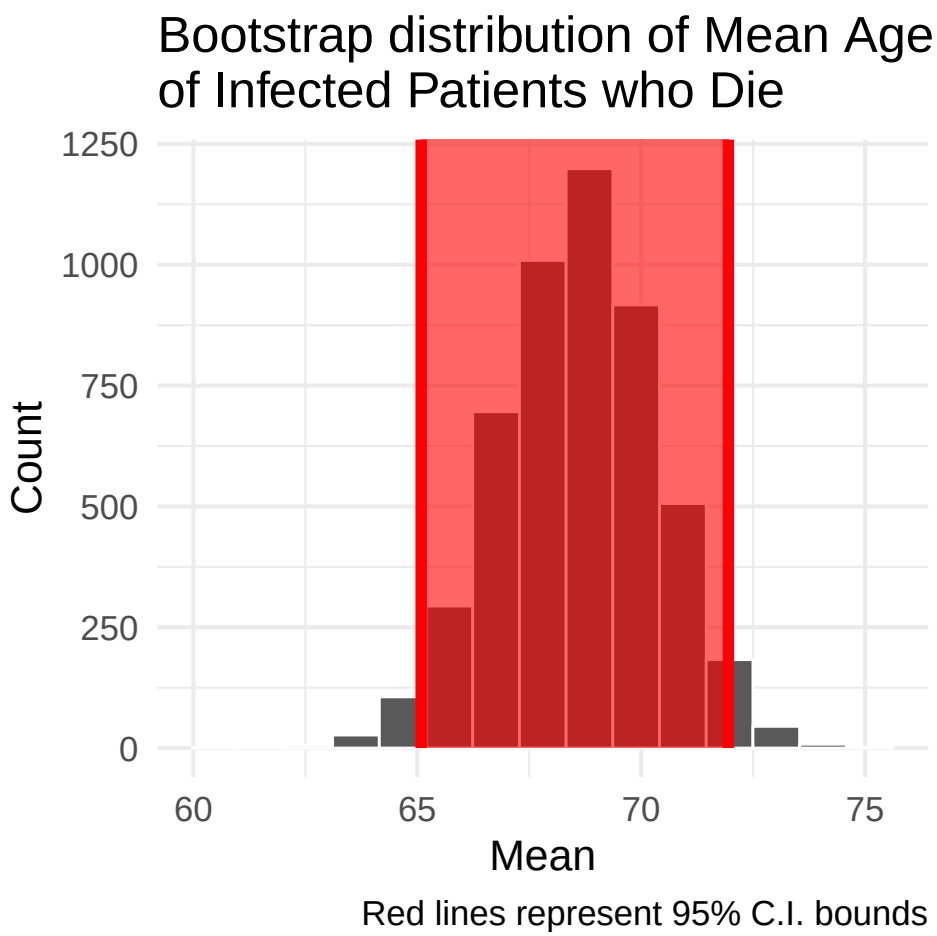
Now, we will create a 95% confidence interval for the population mean age of individuals infected with COVID-19 who survive. As seen below, we are 95% confident that the actual population mean age of infected individuals who survive COVID-19 is between 46.82 and 49.30.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        46.8        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        65.1        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. Association between Age Group and Death Rate

IV. Modeling Death Rate

Section 5: Discussion

- II: mean age CI interval for those infected is closer to interval to those that survive isn't the case that just older people are infected, those that are infected usually survive but of those infected the ones who are older are the ones that die (?)