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 dataset, 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 dataset 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.

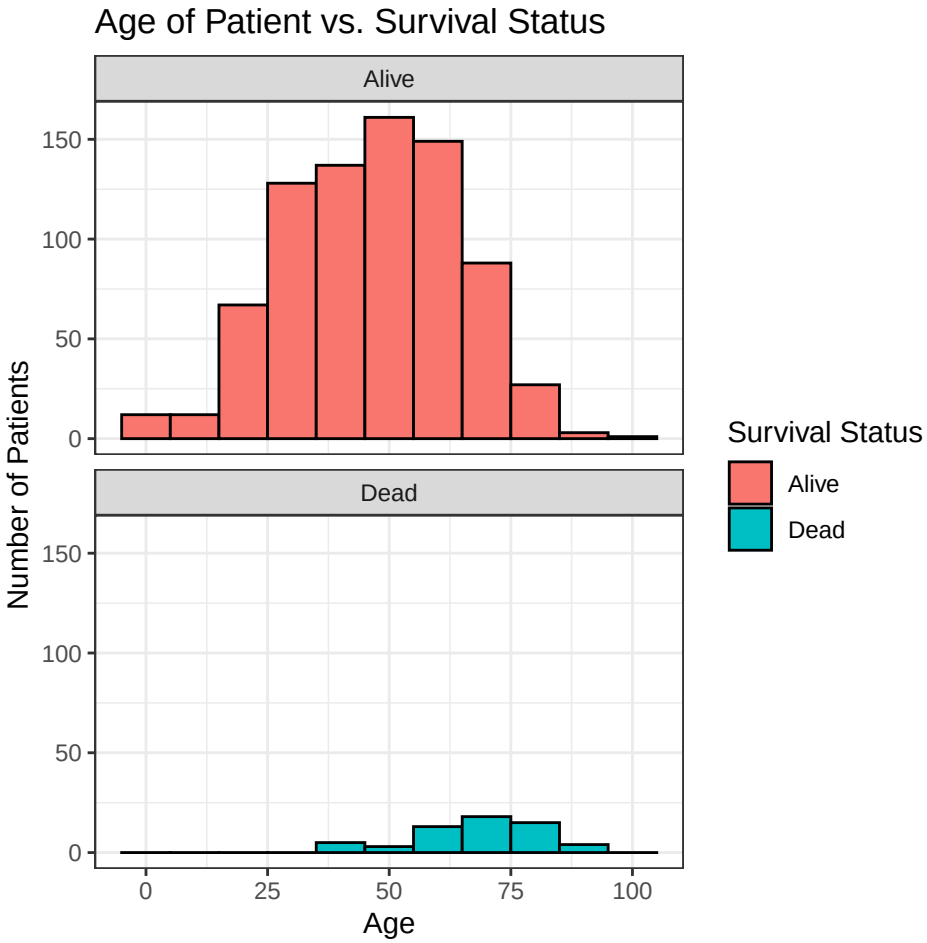
Using these variables, we will attempt to determine not only if age affects one's survival outcome due to COVID-19, but also if any of these other characteristics are associated with one's survival outcome.

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      0.0365      0.0664
```

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       0.161
```

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).

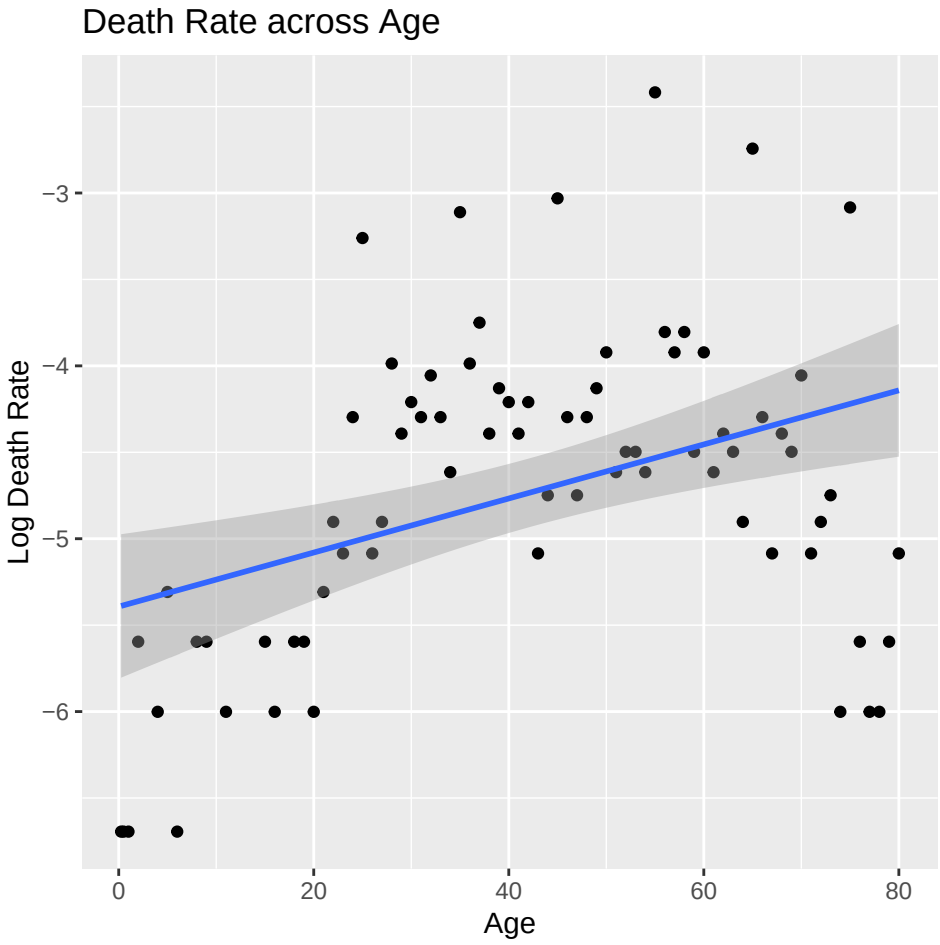
Since the two confidence intervals overlap, we cannot claim that there is significant difference between the mean death rate of individuals in the 80-99 age group with those in 20-39 age group. However, both the lower and upper end of the confidence interval for the older age group (0.04545455 and 0.1590909)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