Preliminary Project Submission

Elijah Chou

# My Variables of Interest

|  |  |  |  |
| --- | --- | --- | --- |
| *Variable Name (in R)* | *Variable Type* | *Response/Explanatory* | *Variable Description* |
| death | Dichotomous | Response | Did the patient die from the disease? Yes, No |
| lag.onset.hosp | Numerical (discrete) | Response | The delay in days between the onset of symptoms and seeking hospital care |
| gender | Dichotomous | Explanatory | The gender of the patient Male, Female |
| Continental.Region | Categorical | Explanatory | The continental region in which the patient sought care. Categories are:  - Australia - Asia - Europe - North America |
| age | Numerical (discrete) | Explanatory | The age of the patient |

# Description of all recoding procedures:

To make the graphs easier to read, I changed the categories of the death variables from 0 and 1 to “Did Not Die” and “Died”. This would make more sense on the legends attached to the bar plots.

For the lag.onset.hosp variable, I had to recode two values to NA: -1 and 999. I believe that the -1-day delay was a misinput, so I decided to set it to NA. Because the codebook of the data set says that the 999 values are NA values, I also had to go ahead and set those to NA as well.

There were two “male” categories and two “female” categories for the gender variable. I combined the similar categories into one so that there were only one “male” and one “female” category.

I recoded the Country variable into a Continental Region. I thought it would make more sense to study continental regions as an explanatory variable rather than countries themselves so that I could generalize the relationships to broader regions. It wouldn’t make much sense to see the differences in death counts between India and Nepal, for example, which both only have one case each. As reference, I used the M49 standard presented by the Statistics Division of the United Nations to divide the countries into their respective continental regions, specifically for Asia (United Nations Statistic Division).

For the age variable, the only thing I had to do was to recode the 999 values into their respective NA values.

# Research Questions

1. RQ 1: What is the relationship between the patient’s gender and the delay in days between the onset of symptoms and seeking hospital care?
   1. There could be societal pressures of men/women to not seek medical attention that could delay their reporting of symptoms to healthcare providers. Depending on culture, men or women could be discouraged from seeking help if their symptoms do not seem severe.
2. RQ2: What is the relationship between the continental region in which the patient sought care the delay in days between the onset of symptoms and seeking hospital care?
   1. Different regions of the world have different healthcare systems. As a result, different regions are likely to show differences in the delay of reporting the disease for various reasons. These reasons may include distrust in healthcare, cost of seeking medical attention, etc.
3. RQ 3: What is the relationship between the patient’s age the delay in days between the onset of symptoms and seeking hospital care?
   1. When you are a young adult, you are less likely to see a doctor when you get a fever or a cough because you are more likely to be confident enough in your own immune system to fight off the disease. Children and elderly people, on the other hand, are more likely to seek medical attention immediately because they are much more vulnerable to diseases due to a weaker/weakened immune system.
4. RQ 4: What is the relationship between the patient’s gender and whether the patient died of the disease?
   1. According to recent data, there seems to be more men than women who die of the disease. The same report suggests that many death cases also report “existing co-morbities” that “tend to more burdensome among men globally” (Blog | BMJ Global Health, 2020).

# Plots

Figure 1: Proportion of Reported Female/Male COVID-19 Patients Who Died From the Disease

A screenshot of a cell phone

Description automatically generated

Figure 2: Proportion of Deaths in Reported COVID-19 Cases in Different Continents

A screenshot of a cell phone

Description automatically generated

Figure 3: Distributions of Age of COVID-19 Cases Categorized by Death or No Death

A screenshot of a social media post

Description automatically generated

Figure 4: Distribution of the Delay in Days Between Onset of Symptoms & Seeking Hospital Care Based on COVID-19 Patients’ Gender

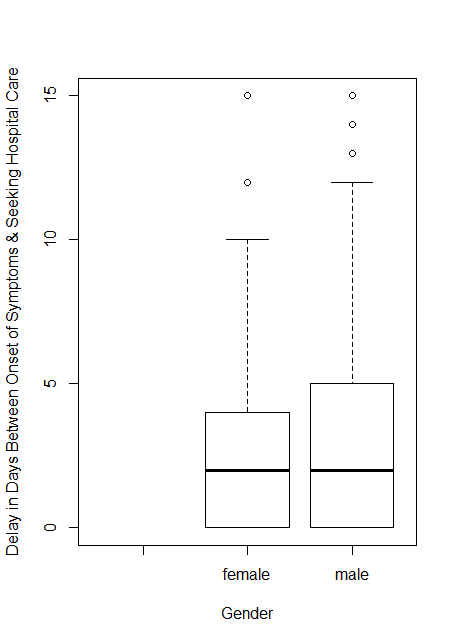


Figure 5: Distribution of the Delay in Days Between Onset of Symptoms & Seeking Hospital Care Based on Continent in which the COVID-19 Patient Sought Hospital Care

A screenshot of a cell phone

Description automatically generated

Figure 6: Scatterplot of Age vs. Delay in Days Between Onset of Symptoms & Seeking Hospital Care, with No Apparent Correlation

A screenshot of a cell phone

Description automatically generated

# Works Cited

2020. “Sex, gender and COVID-19: Disaggregated data and health disparities”. *Blog | BMJ Global Health*. <https://blogs.bmj.com/bmjgh/2020/03/24/sex-gender-and-covid-19-disaggregated-data-and-health-disparities/>

“Standard country or area codes for statistical use (M49)”. *United Nations Statistics Division*. <https://unstats.un.org/unsd/methodology/m49/#qa>

# Appendix: R Code

###Setting up Environment###

#Set up working directory

setwd("C:/Users/ecool/Desktop/R Code")

#Import data

corona <- read.csv("C:/Users/ecool/Desktop/R Code/Data Sets/COV Dataset.csv")

####Inspecting Dataset####

summary(corona)

####Recoding variables####

##Recoding the "death" variable

summary(corona$death) #check death variable, need to change categories from numbers to labels

corona$death2<-factor(NA, levels=c("Died", "Did Not Die")) #create new categorical variable

corona$death2[corona$death==0]<-"Did Not Die" #reassigns those who did not die to respective category

corona$death2[corona$death==1]<-"Died" #reassigns those who did die to respective category

summary(corona$death2) #verify changes to variable

##Recoding the "lag.onset.hosp" variable

summary(corona$lag.onset.hosp) #check lag.onset.hosp variable, noticed that there are two problems: a -1 value and many 999 values

corona$lag.onset.hosp2<-corona$lag.onset.hosp #create duplicate "lag.onset.hosp" variable

corona$lag.onset.hosp2[corona$lag.onset.hosp2<0]<-NA #recode all entries of this variable where delay is less than 0 to NA

corona$lag.onset.hosp2[corona$lag.onset.hosp2==999]<-NA #recode all "coded NAs" of 999 back to NA

summary(corona$lag.onset.hosp2) #verify changes to variable

##Recoding the "gender" variable

summary(corona$gender) #check gender variable, need to combine the two "male" groups into one, do same for "female" categories, and remove 999 values

corona$gender2<-corona$gender #create duplicate "gender" variable

corona$gender2[corona$gender2=="MALE"]<-"male" #recode erroneous male values

corona$gender2[corona$gender2=="Female"]<-"female" #recode erroneous female values

corona$gender2[corona$gender2==999]<-NA #recode all "coded NAs" of 999 back to NA

corona$gender2<-factor(corona$gender2) #drop erroneous categories

summary(corona$gender2) #verify changes to variable

##Creating the "Continental.Region" variable from "Country" variable

summary(corona$Country) #check country variable to determine which countries there are

corona$Country[corona$Country=="china"]<-"China" #recode erroneous china values

corona$Continental.Region<-factor(NA, levels=c("Australia", "Asia", "Europe", "North America")) #creates new categorical variable with the four continental regions

corona$Continental.Region[corona$Country=="Australia"]<-"Australia" #reassign values of the new Australia Continental Region

corona$Continental.Region[corona$Country=="Cambodia" | corona$Country=="China" |

corona$Country=="Hong Kong" | corona$Country=="India" |

corona$Country=="Japan" | corona$Country=="Malaysia" |

corona$Country=="Nepal" | corona$Country=="Russia" |

corona$Country=="Singapore" | corona$Country=="South Korea" |

corona$Country=="Sri Lanka" | corona$Country=="Taiwan" |

corona$Country=="Thailand" | corona$Country=="UAE" |

corona$Country=="Vietnam"]<-"Asia" #reassign values of the new Asia Continental Region

corona$Continental.Region[corona$Country=="France" | corona$Country=="Germany" |

corona$Country=="Italy" | corona$Country=="UK"]<-"Europe" #reassign values of the new Europe Continental Region

corona$Continental.Region[corona$Country=="Canada" | corona$Country=="USA"]<-"North America" #reassign values of the new North America Continental Region

summary(corona$Continental.Region) #verify successful creation of new variable

##Recoding the "age" variable

summary(corona$age) #check age variable, need to recode 999 to NA

corona$age2<-corona$age #create duplicate "age" variable

corona$age2[corona$age2==999]<-NA #recode all "coded NAs" of 999 back to NA

summary(corona$age2) #verify changes to variable

####Creating Graphs & Figures####

#Figure 1: death vs. gender

table1 <- table(corona$death2, corona$gender2) #create two-way contigency table

proptable1 <- prop.table(table1, margin=2) #create column proportions table

barplot(proptable1, beside=TRUE, legend.text = T, xlab = "Gender", ylab = "Proportion") #create bar plot using proportions with side-by-side bars

#Figure 2: death vs. Continental.Region

table2 <- table(corona$death2, corona$Continental.Region) #create two-way contigency table

proptable2 <- prop.table(table2, margin=2) #create column proportions table

barplot(proptable2, beside=T, legend.text = T, xlab = "Continental Region", ylab = "Proportion") #create bar plot using proportions with side-by-side bars

#Figure 3: death vs. age

#creates side-by-side boxplots of the distributions of age based on whether they died or not

boxplot(corona$age2 ~ corona$death2, xlab = "Did Patient Die of Disease?", ylab = "Age")

#Figure 4: lag.onset.hosp vs. gender

#creates side-by-side boxplots using lag.onset.hosp and gender variables

boxplot(corona$lag.onset.hosp2 ~ corona$gender2, xlab = "Gender", ylab = "Delay in Days Between Onset of Symptoms & Seeking Hospital Care")

#Figure 5: lag.onset.hosp vs. Continental.Region

#creates side-by-side boxplots using lag.onset.hosp and Continental.Region variables

boxplot(corona$lag.onset.hosp2 ~ corona$Continental.Region, xlab = "Continental Region", ylab = "Delay in Days Between Onset of Symptoms & Seeking Hospital Care")

#Figure 6: lag.onset.hosp vs. age

#creates scatterplot using lag.onset.hosp and age variables

plot(corona$age2, corona$lag.onset.hosp2, xlab = "Age", ylab = "Delay in Days Between Onset of Symptoms & Seeking Hospital Care")