CPSC 375 Project 1 Report

GROUP MEMBERS:
NATHAN MARCOS, RYAN TEOH

Part 1: Data Wrangling

Covid Confirmed Cases and Deaths Data:

- Read in data:

**> deaths <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv")**

**> confirmed <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv")**

- Tidy the tables to a column-centric format by using pivot_longer, rename Country/Region to Country, aggregate State/Province by grouping Country and Day and summarizing the sum of Deaths/Confirmed, order the dates and arrange the names so the data is easier to read:

```
> deaths <- deaths %>% pivot_longer(-(1:4), names_to="Day",
values_to="Deaths")
> deaths <- deaths %>% rename(`Country`= `Country/Region`)
> deaths <- deaths %>% group_by(Country, Day) %>%
summarize(Deaths=sum(Deaths))
> deaths <- deaths[order(as.Date(deaths$Day, format="%m/%d/%Y")),]
> deaths <- deaths %>% arrange(Country)

> confirmed <- confirmed %>% pivot_longer(-(1:4), names_to="Day",
values_to="Confirmed")
> confirmed <- confirmed %>% rename(`Country`= `Country/Region`)
> confirmed <-  confirmed %>% group_by(Country, Day) %>%
summarize(Confirmed=sum(Confirmed))
> confirmed <- confirmed[order(as.Date(confirmed$Day, format="%m/%d/%Y")),]
> confirmed <- confirmed %>% arrange(Country)
```

- Join the Deaths and Confirmed datasets:

> covid <- deaths %>% full_join(confirmed)

Hospital Beds Data:

- Read in data:

> beds <- read_csv("C:/Users/username/Downloads/hospitalbeds.csv")

- Filter for only the beds counted in the most recent year of each country:

> beds <- beds %>% group_by(Country) %>% filter(Year == max(Year))

- Remove the years from the dataset as they are not needed:

> beds <- beds %>% select(-Year)

Demographics Data:

- Read in data:

> demographics <- read_csv("C:/Users/username/Downloads/demographics.csv")

- Select Country Name, Series Code, YR2015, as they have the information we need, tidy it using pivot_wider, and rename Country Name to Country:

> demographics <- demographics %>% select(`Country Name`, `Series Code`, YR2015)
> demographics <- demographics %>% pivot_wider(names_from = `Series Code`, values_from = YR2015)
> demographics <- demographics %>% rename(`Country`= `Country Name`)

- Rename names by mutating so that they are easier to read, then remove the data with the unwanted names:

> demographics <- demographics %>% mutate("Population (Total)"=SP.POP.TOTL.FE.IN+SP.POP.TOTL.MA.IN, "Population (Urban)"=SP.URB.TOTL, "Population (80+)"=SP.POP.80UP.FE+SP.POP.80UP.MA, "Population (65+)"=SP.POP.65UP.FE.IN+SP.POP.65UP.MA.IN, "Population (15-64)"=SP.POP.1564.MA.IN+SP.POP.1564.FE.IN, "Population (0-14)"=SP.POP.0014.MA.IN+SP.POP.1564.FE.IN)

> demographics <- demographics %>% select(-(2:16))

Joining Datasets:

- Rename as much countries as possible to match each other across all datasets, then join them (we mutated South Korea, Iran, United Kingdom, United States, Congo, Venezuela, and Bolivia):

> covid <- covid %>% mutate(Country = replace(Country, Country == "Korea, South", "South Korea")) %>% mutate(Country = replace(Country, Country == "US", "United States")) %>% mutate(Country = replace(Country, Country == "Congo (Brazzaville)",

"Congo")) %>% mutate(Country = replace(Country, Country == "Congo (Kinshasa)", "Congo"))

> beds <- beds %>% mutate(Country = replace(Country, Country == "Republic of Korea", "South Korea")) %>% mutate(Country = replace(Country, Country == "Iran (Islamic Republic of)", "Iran")) %>% mutate(Country = replace(Country, Country == "United Kingdom of Great Britain and Northern Ireland", "United Kingdom")) %>% mutate(Country = replace(Country, Country == "Bolivia (Plurinational State of)", "Bolivia")) %>% mutate(Country = replace(Country, Country == "United States of America", "United States")) %>% mutate(Country = replace(Country, Country == "Democratic Republic of the Congo", "Congo")) %>% mutate(Country = replace(Country, Country == "Venezuela (Bolivarian Republic of)", "Venezuela"))

> demographics <- demographics %>% mutate(Country = replace(Country, Country == "Korea, Dem. People's Rep.", "South Korea")) %>% mutate(Country = replace(Country, Country == "Korea, Rep.", "South Korea")) %>% mutate(Country = replace(Country, Country == "Iran, Islamic Rep.", "Iran")) %>% mutate(Country = replace(Country, Country == "Congo, Dem. Rep.", "Congo")) %>% mutate(Country = replace(Country, Country == "Congo, Rep.", "Congo")) %>% mutate(Country = replace(Country, Country == "Venezuela, RB", "Venezuela"))

> mydata <- covid %>% full_join(beds) %>% full_join(demographics)

Part 2: Modeling

The first step to linear modeling the number of deaths in a country was to create a predictive model to compare our combination of predictor models to. This predictive model was based on Confirmed cases because it is the most relevant predictor. Based on that model, we tested a different combination of predictor variables which will be described below.

The combinations we will be using are:

1. Deaths~Confirmed+Hospital Beds
    - Does the amount of hospital beds help us predict Covid related deaths?

2. Deaths~Confirmed+Urban Population
    - Urban areas usually have more compact areas and could make spreading the virus easier, so would this help predict deaths?

3. Deaths~Ages65UP+Population Total
    - According to the news, elderly may be affected by Covid more than the younger population. Would a large population with a higher amount of elderly help predict if there are more or less deaths?

4. Deaths~Confirmed+(Urban Population/Population Total)

- As mentioned in combination 2, urban areas seem to be more compact, so does a population with a higher urban ratio help predict that there will be more deaths?

5. Deaths~Ages0-14
    - According to the news, younger people may be less affected by Covid in a way where they might not feel any side effects and have a higher chance of living. Does the number of children 0-14 in a population help predict deaths?

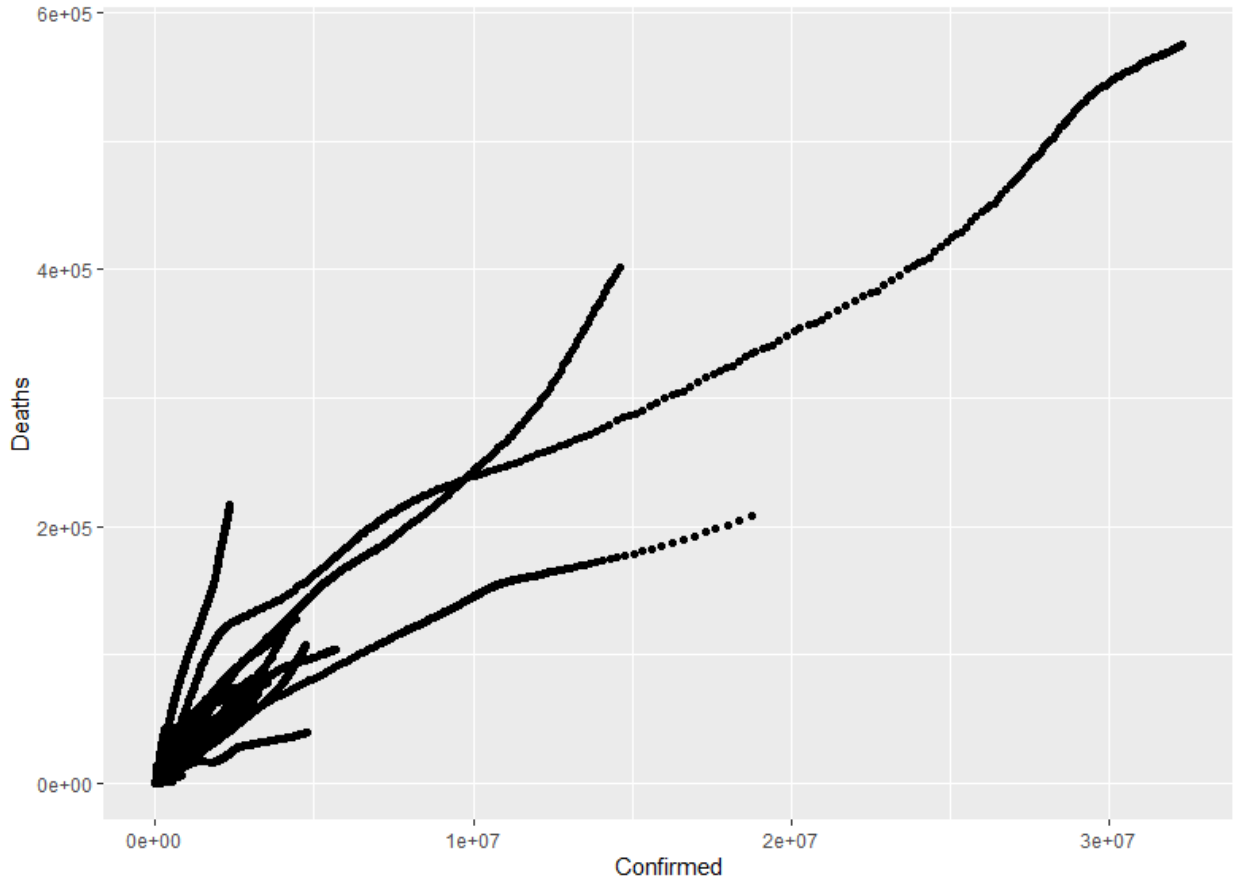
Combination 0:

Dependent Variable: Deaths

Independent Variable: Confirmed Cases

- Plot the data:

> ggplot(data=mydata)+ geom_point(mapping = aes(x=Confirmed, y=Deaths), na.rm = TRUE)



> modConfirmed <- lm(data=mydata, formula=Deaths~Confirmed)

> summary(modConfirmed)

Call:
lm(formula = Deaths ~ Confirmed, data = mydata)

Residuals:
Min 1Q Median 3Q Max
-155366 -1224 -1195 -1140 170025

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.195e+03 3.094e+01 38.63 <2e-16 *
Confirmed 1.932e-02 2.241e-05 862.09 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 9244 on 91870 degrees of freedom
(93 observations deleted due to missingness)
Multiple R-squared: 0.89, Adjusted R-squared: 0.89
F-statistic: 7.432e+05 on 1 and 91870 DF, p-value: < 2.2e-16

Combination 1:

Dependent Variable: Deaths

Independent Variables: Confirmed Cases and Beds

> modBeds <- lm(data=mydata, formula = Deaths~Confirmed + `Hospital beds (per 10 000 population)`)

> summary(modBeds)

Call:

lm(formula = Deaths ~ Confirmed + `Hospital beds (per 10 000 population)`,

data = mydata)

Residuals:

Min 1Q Median 3Q Max

-155090 -1426 -1387 -1313 169888

Coefficients:

	Estimate	Std. Error
(Intercept)	**1.394e+03**	**5.471e+01**
Confirmed	**1.929e-02**	**2.420e-05**
`Hospital beds (per 10 000 population)`	**-3.113e-01**	**1.553e+00**

	t value	Pr(>|t|)
(Intercept)	**25.48**	**<2e-16 *****
Confirmed	**797.27**	**<2e-16 *****
`Hospital beds (per 10 000 population)`	**-0.20**	**0.841**

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 9898 on 79805 degrees of freedom

(12157 observations deleted due to missingness)

Multiple R-squared: 0.8885, Adjusted R-squared: 0.8885

F-statistic: 3.179e+05 on 2 and 79805 DF, p-value: < 2.2e-16

Combination 2:

Dependent Variable: Deaths

Independent Variables: Confirmed Cases and Urban Population

```
> modUrban <- lm(data=mydata, formula = Deaths~Confirmed + `Population (Urban)`)
> summary(modUrban)

Call:
lm(formula = Deaths ~ Confirmed + `Population (Urban)`, data = mydata)
```

Residuals:

Min	1Q	Median	3Q	Max
-156444	-1382	-1153	-1112	169433

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.118e+03	3.577e+01	31.24	<2e-16	***
Confirmed	1.910e-02	2.529e-05	755.44	<2e-16	***
`Population (Urban)`	1.222e-05	5.108e-07	23.92	<2e-16	***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 9749 on 81661 degrees of freedom

(10301 observations deleted due to missingness)

Multiple R-squared: 0.8894, Adjusted R-squared: 0.8894

F-statistic: 3.283e+05 on 2 and 81661 DF, p-value: < 2.2e-16

Combination 3:

Independent Variable: Deaths

Predictor Variables: Confirmed Cases, Ages 65+, and Population Total

```
> modElderlyTotal <- lm(data=mydata, formula = Deaths~Confirmed + `Population (65+)` +`Population (Total)`)
> summary(modElderlyTotal)
```

Call:

lm(formula = Deaths ~ Confirmed + `Population (65+)` + `Population (Total)`,

data = mydata)

Residuals:

Min	1Q	Median	3Q	Max
-133541	-1581	-1362	-756	171159

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.440e+03	3.585e+01	40.16	<2e-16	***
Confirmed	1.914e-02	2.499e-05	765.65	<2e-16	***
`Population (65+)`	5.795e-04	8.277e-06	70.02	<2e-16	***
`Population (Total)`	-4.675e-05	6.702e-07	-69.76	<2e-16	***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 9647 on 78876 degrees of freedom

(13085 observations deleted due to missingness)

Multiple R-squared: 0.8952, Adjusted R-squared: 0.8952

F-statistic: 2.246e+05 on 3 and 78876 DF, p-value: < 2.2e-16

<u>Combination 4:</u>

Dependent Variable: Deaths

Independent Variables: Confirmed Cases and Urban Population to Total Population Ratio

```
> urbanratiodata <- mydata %>% mutate(UrbanTotalRatio = `Population (Urban)`/`Population (Total)`)

> modUrbanTotal <- lm(data=urbanratiodata, formula = Deaths~Confirmed + UrbanTotalRatio)

> summary(modUrbanTotal)


Call:

lm(formula = Deaths ~ Confirmed + UrbanTotalRatio, data = urbanratiodata)


Residuals:

    Min      1Q  Median      3Q     Max

-151350   -2304   -1073     533  168668


Coefficients:

                 Estimate Std. Error t value Pr(>|t|)

(Intercept)     -2.550e+03  9.647e+01  -26.43   <2e-16 ***

Confirmed        1.919e-02  2.419e-05  793.29   <2e-16 ***

UrbanTotalRatio  6.830e+03  1.559e+02   43.82   <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 9832 on 78877 degrees of freedom

  (13085 observations deleted due to missingness)

Multiple R-squared:  0.8912,      Adjusted R-squared:  0.8912

F-statistic: 3.229e+05 on 2 and 78877 DF,  p-value: < 2.2e-16
```

Combination 5:

Dependent Variable: Deaths

Independent Variable: Confirmed + Ages 0-14

```
> modYoung <- lm(data=mydata, formula = Deaths~Confirmed + `Population (0-14)`)
> summary(modYoung)

Call:
lm(formula = Deaths ~ Confirmed + `Population (0-14)`, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-152736   -1509   -1466   -1265  169964

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.494e+03  3.695e+01   40.44   <2e-16
Confirmed           1.940e-02  2.545e-05  762.10   <2e-16
`Population (0-14)` -7.199e-06  5.435e-07  -13.24   <2e-16

(Intercept)         ***
Confirmed           ***
`Population (0-14)` ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9940 on 78877 degrees of freedom
  (13085 observations deleted due to missingness)
```

Multiple R-squared: 0.8888, Adjusted R-squared: 0.8887

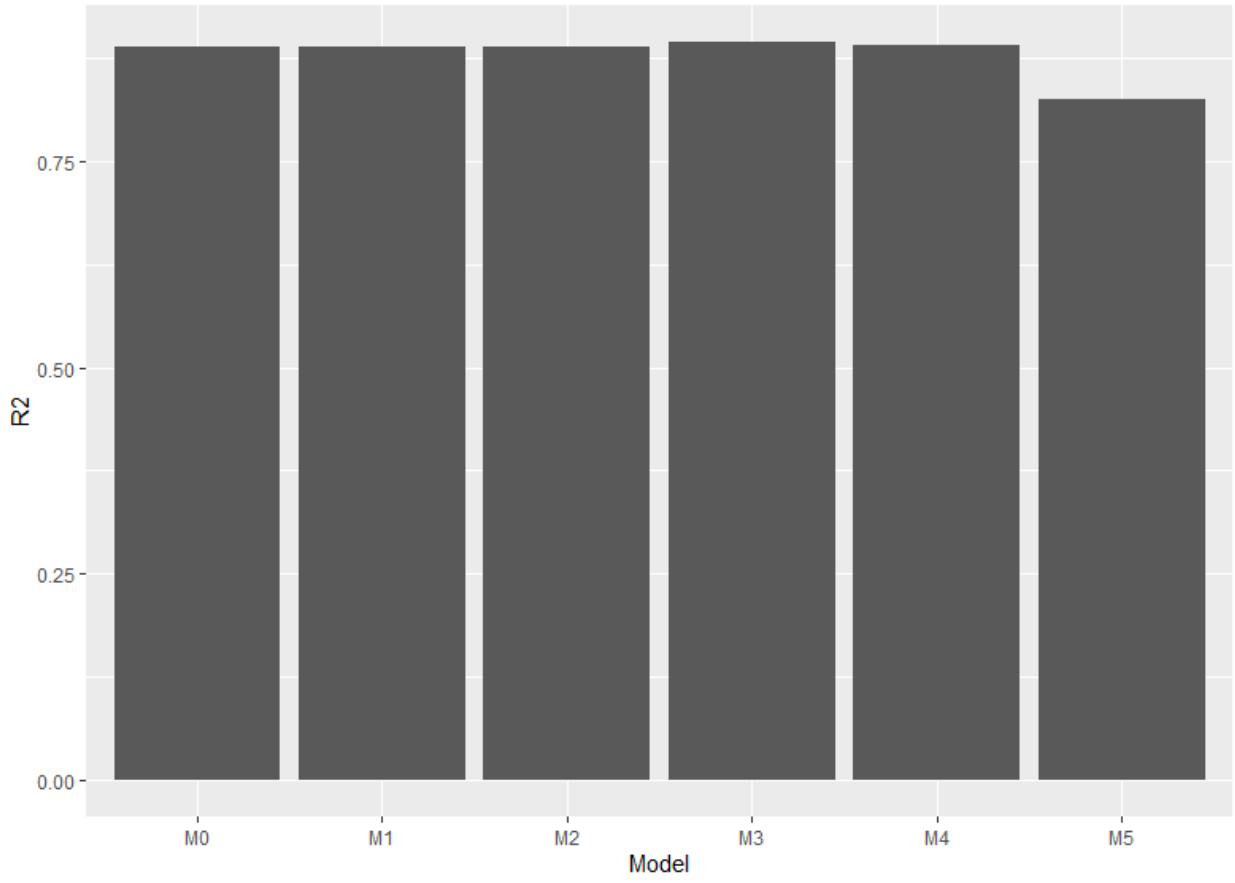
F-statistic: 3.151e+05 on 2 and 78877 DF, p-value: < 2.2e-16

Results and Analysis:

Results:

```
> myresults <- tibble(model=c("Deaths~Confirmed", "Deaths~Confirmed+Beds",
"Deaths~Confirmed+Urban", "Deaths~Confirmed+65UP+Total",
"Deaths~Confirmed+Urban/Total", "Deaths~Confirmed+0-14"), R2 =
c(0.89,  0.8885, 0.8894, 0.8952,  0.8912, 0.8888))

> myresults

# A tibble: 6 x 2

  model                    R2

  <chr>                 <dbl>

1 Deaths~Confirmed          0.89

2 Deaths~Confirmed+Beds     0.888

3 Deaths~Confirmed+Urban    0.889

4 Deaths~Confirmed+65UP+Total  0.895

5 Deaths~Confirmed+Urban/Total 0.891

6 Deaths~Confirmed+0-14     0.889


> ggplot(data=myresults) + geom_col(mapping = aes(x=model, y=R2)) +
labs(x="Model")
```

 # M0-M5 follow the the myresults tibble names respectively (we renamed it so the names fit on the chart)



Conclusion:

Based off the results of our modeling about Covid 19 related deaths, we can determine which were the significant factors and which were not:

- Compared to the Confirmed cases predictor (R2 = 0.89), the first combination (Confirmed+Hospital Beds [R2 = 0.888]) had a lower R2 value, and therefore using Hospital Beds when modeling resulted in a worse model.

- The second combination (Confirmed+Urban Population [R2 = 0.889]) was very close in R2 value, but because we are omitting the rest of the population it seems that it makes this model worse. However, because the R2 value is so close, we can assume that urban areas most likely have the most confirmed cases compared to non-urban areas, a much higher proportion at that.

- The third combination (Confirmed+Ages65UP+Population Total [R2=0.895]) proved to have a higher R2 value and is therefore a better model. It seems that the higher the population, and the more elderly there are, the easier it is to predict the number of deaths. This may have to do with what we stated earlier that the elderly are more susceptible to side effects and a higher population increases the chances of the virus spreading.

- The fourth combination (Confirmed+Urban Population/Population Total [R2 = 0.891]) also had higher R2 value which makes it a better model. By transforming the Urban Population and Population into a ratio, we can see the proportion of a country's urban population to its total population helps us predict the number of deaths.

- The final combination (Confirmed+Ages0-14 [R=0.889]) was very close in R2 value, but because we are omitting the rest of the population it seems that it makes this model worse. However, because the R2 value is so close, we can assume that the number of young people in a country helps us strongly with predicting deaths.