Luke Duggan
CPSC 375
Dr. P

Project 1 Wrangling and Conclusion

Data Wrangling

1. **Read in the data using read.csv and give variable names**

```
confdata <- 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")

deathdata <- 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")

beddata <- read.csv('C:/Users/luked/Downloads/WHS6_102.csv')

demodata <- read.csv("C:/Users/luked/Downloads/demographics.csv")
```

2. **Fix country names that might not match other data sets**

```
beddata <- beddata %>% mutate(Country = replace(Country, Country == "United Kingdom of Great Britain and Northern Ireland", "United Kingdom"))

beddata <- beddata %>% mutate(Country = replace(Country, Country == "Iran (Islamic Republic of)", "Iran"))

beddata <- beddata %>% mutate(Country = replace(Country, Country == "Republic of Korea", "South Korea"))

demodata <- demodata %>% mutate(Country = replace(Country, Country == "Korea, Rep.", "South Korea"))

demodata <- demodata %>% mutate(Country = replace(Country, Country == "Iran, Islamic Rep.", "Iran"))
```

3. **Rename hospital beds column and only contain most recent bed year data**

```
beddata <- beddata %>% rename(beds = Hospital.beds..per.10.000.population.)

updbeddata <- beddata %>% group_by(Country) %>% filter(Year==max(Year))
```

4. **Rename Country.Region column for future join**

```
confdata <- confdata %>% rename(Country = Country.Region)

deathdata <- deathdata %>% rename(Country = Country.Region)

demodata <- demodata %>% rename(Country = ï..Country.Name)
```

Luke Duggan
CPSC 375
Dr. P

5. **Select only country name and day columns for confirmed and death data**

```
confdata1 <- confdata %>% select(2,5:440)

deathdata1 <- deathdata %>% select(2,5:440)
```

6. **Group together confirmed data and death data by country and sum together the groups (this is for the countries that had a state)**

```
confdata2 <- confdata1 %>% group_by(Country) %>% summarise(across(2:436,sum))

deathdata2 <- deathdata1 %>% group_by(Country) %>% summarise(across(2:436,sum))
```

7. **Join together confirmed and death data tables and fix country name values wont match other tables**

```
coviddata <- deathdata3 %>% inner_join(confdata3)

coviddata <- coviddata %>% mutate(Country = replace(Country, Country == "Korea, South", "South Korea"))
```

8. **Join death and confirmed table with bed data table and remove bed year column**

```
covidBed <- coviddata %>% inner_join(updbeddata)

covidBed <- covidBed %>% select(1:4, 6)
```

9. **Remove unnecessary series codes that don’t relate to pop total, or age brackets**

```
popdata <- demodata %>% filter(!Series.Code == 'SP.DYN.LE00.IN', !Series.Code == 'SP.URB.TOTL', !Series.Code =='SP.DYN.AMRT.FE', !Series.Code == 'SP.DYN.AMRT.MA', !Series.Code == 'SP.POP.TOTL.FE.IN', !Series.Code == 'SP.POP.TOTL.MA.IN')
```

10. **Put each set of series codes for male and female into their own variable to be summed together**

```
pop14 <- popdata %>% filter(Series.Code=='SP.POP.0014.MA.IN'| Series.Code=='SP.POP.0014.FE.IN')

pop1564 <- popdata %>% filter(Series.Code=='SP.POP.1564.MA.IN'| Series.Code=='SP.POP.1564.FE.IN')

pop65up <- popdata %>% filter(Series.Code=='SP.POP.65UP.MA.IN'| Series.Code=='SP.POP.65UP.FE.IN')

pop80up <- popdata %>% filter(Series.Code=='SP.POP.80UP.MA'| Series.Code=='SP.POP.80UP.FE')

poptotal <- popdata %>% filter(Series.Code=='SP.POP.TOTL')
```

Luke Duggan
CPSC 375
Dr. P

11. **Convert totals to integer for each age group so male and female can be summed together**

```
pop14$YR2015 <- as.integer(pop14$YR2015)

pop1564$YR2015 <- as.integer(pop1564$YR2015)

pop65up$YR2015 <- as.integer(pop65up$YR2015)

pop80up$YR2015 <- as.integer(pop80up$YR2015)

poptotal$YR2015 <- as.integer(poptotal$YR2015)
```

12. **Sum together the male and female**

```
pop14 <- pop14 %>% group_by(Country) %>% summarise(Total=sum(YR2015, na.rm = TRUE))

pop1564 <- pop1564 %>% group_by(Country) %>% summarise(Total=sum(YR2015, na.rm = TRUE))

pop65up <- pop65up %>% group_by(Country) %>% summarise(Total=sum(YR2015, na.rm = TRUE))

pop80up <- pop80up %>% group_by(Country) %>% summarise(Total=sum(YR2015, na.rm = TRUE))
```

13. **Rename columns before joining**

```
pop14 <- pop14 %>% rename(POP.0014 =Total)

pop1564 <- pop1564 %>% rename(POP.1564=Total)

pop65up <- pop65up %>% rename(POP.65.UP=Total)

pop80up <- pop80up %>% rename(POP.80.UP=Total)

poptotal <- poptotal %>% rename(POP.TOTL=YR2015)
```

14. **Join together series code data with covid conf/death/bed data table**

```
covidBed <- covidBed %>% inner_join(poptotal)

covidBed <- covidBed %>% inner_join(pop80up)

covidBed <- covidBed %>% inner_join(pop65up)

covidBed <- covidBed %>% inner_join(pop1564)

covidBed <- covidBed %>% inner_join(pop14)

covidBed <- covidBed %>% select(1:5, 9:13)
```

Data is now wrangled and in complete format.

Luke Duggan
CPSC 375
Dr. P

Creating Linear Models

Using different combinations of predictor variables

mod <- lm(deaths ~ confirmed+beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014, data=covidBed)

Adjusted R2 = 0.8308

mod1 <- lm(deaths ~ confirmed+beds+POP.80.UP+POP.65.UP+POP.1564, data=covidBed)

Adjusted R2 = 0.8208

mod2 <- lm(deaths ~ confirmed+beds+POP.TOTL+POP.80.UP+POP.65.UP, data=covidBed)

Adjusted R2 = 0.822

mod3 <- lm(deaths ~ beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014, data=covidBed)

R2 = 0.3001

mod4 <- lm(deaths ~ confirmed+beds+POP.TOTL, data=covidBed)

R2 = 0.7987

Luke Duggan
CPSC 375
Dr. P

Creating Bar Graph

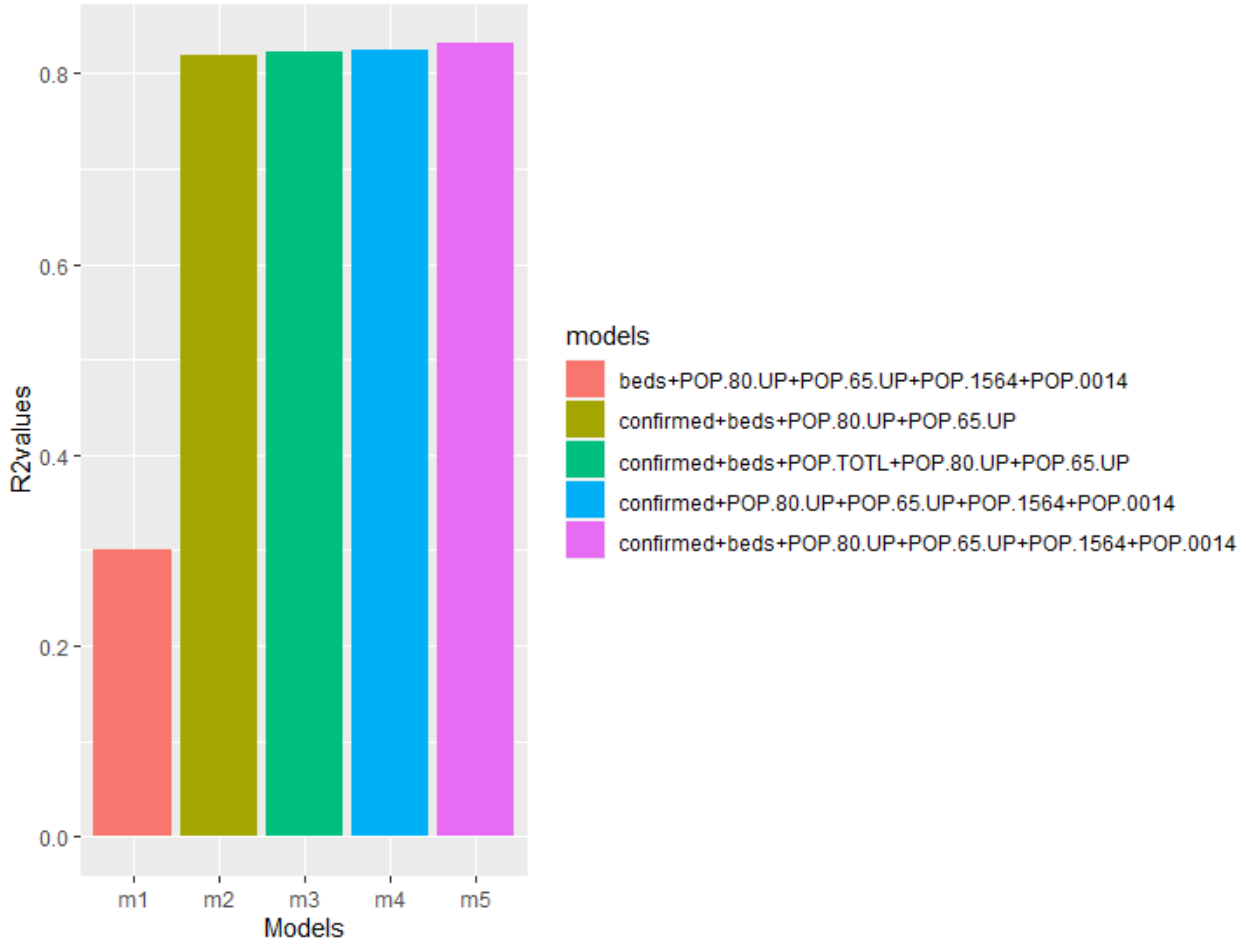
Used solely for R2 comparison between models

```
R2values <- c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared,
summary(mod3)$adj.r.squared, summary(mod4)$adj.r.squared, summary(mod5)$adj.r.squared)

models <- c('beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014',

        'confirmed+beds+POP.80.UP+POP.65.UP',

        'confirmed+beds+POP.TOTL+POP.80.UP+POP.65.UP',

        'confirmed+POP.80.UP+POP.65.UP+POP.1564+POP.0014',

        'confirmed+beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014')

models <- factor(models, levels=c('beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014',

                'confirmed+beds+POP.80.UP+POP.65.UP',

                'confirmed+beds+POP.TOTL+POP.80.UP+POP.65.UP',

                'confirmed+POP.80.UP+POP.65.UP+POP.1564+POP.0014',

                'confirmed+beds+POP.80.UP+POP.65.UP+POP.1564+POP.0014'))

Models <- c('m1','m2','m3','m4','m5')

R2s <- data.frame(models, R2values)

# The ggplot code needs to be ran manually for the bars to show

ggplot(data=R2s, aes(x=Models, y=R2values, fill=models))+geom_bar(stat='identity')
```

Luke Duggan
CPSC 375
Dr. P

Conclusion

The bar graph displays the R2 values for each of the 5 linear models I created using the data. There is also a legend to the right displaying which predictor variables were used for each of the models.



As said in the project instructions, confirmed was a very important variable to include since deaths is highly correlated with confirmed cases. I decided to see what would happen if I didn't include confirmed, and as you can see the first bar has a significantly lower R2 value than the other bars.

I then played around with the other predictor variables since I thought only including older ages would provide a better model since covid has a higher mortality rate for those who are older.

I proceeded to develop the linear model with the best R2 by simply using all the predictor variables except POP.TOTL (I felt this one wasn't necessary if you are including all the other populations) and was able to create the best model using **confirmed, beds, pop.80.up, pop.65.up, pop.1564, and pop.0014**. These variables, as far as I know, were the best for predicting the deaths in a country although the residual standard error sometimes permitted 20-30% differences in the number of deaths predicted by the linear model vs the actual deaths for some of the data rows.