Vincent Diep
vincentdiep@csu.fullerton.edu
CPSC 375 Project 1

1. Linear model of deaths vs. confirmed.
   (Number of deaths vs number of confirmed cases)

```
> model_confirmed <- lm(data=covid19_clean_data, formula=deaths~confirmed)
> summary(model_confirmed)

Call:
lm(formula = deaths ~ confirmed, data = covid19_clean_data)

Residuals:
        Min     1Q  Median   3Q      Max
-180427   -1116   -1029      -975  166472

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.029e+03  3.564e+01   28.88   <2e-16 ***
confirmed   2.091e-02  3.869e-05  540.42   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9403 on 73774 degrees of freedom
Multiple R-squared:  0.7983,    Adjusted R-squared:  0.7983
F-statistic: 2.92e+05 on 1 and 73774 DF,  p-value: < 2.2e-16
```

2. Linear model of deaths vs. Proportion of Urban Population.
   (Prop. of Urban Pop. = Urban Population / Total Population)

```
> covid19_clean_data <- covid19_clean_data %>% mutate(URBAN.POP.PROP =
URBAN.POP / POP.TOTL)
> model_urb_pop_prop <- lm(data=covid19_clean_data,
formula=deaths~URBAN.POP.PROP)
> summary(model_urb_pop_prop)

Call:
lm(formula = deaths ~ URBAN.POP.PROP, data = covid19_clean_data)

Residuals:
   Min  1Q Median    3Q      Max 
-11566  -6812  -3781   -702 391675 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2881.6         213.5  -13.50   <2e-16 ***
URBAN.POP.PROP  14448.1       339.1   42.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20680 on 73774 degrees of freedom
Multiple R-squared:  0.02402,    Adjusted R-squared:  0.02401 
F-statistic:  1816 on 1 and 73774 DF,  p-value: < 2.2e-16
```

3. Linear model of deaths vs. beds + proportion of urban population.
   (beds = number of hospital beds per 10,000 population)

```
> model_bed_urb <- lm(data=covid19_clean_data, formula=deaths~`Hospital beds (per
10 000 population)`+ URBAN.POP.PROP)
> summary(model_bed_urb)

Call:
lm(formula = deaths ~ `Hospital beds (per 10 000 population)` + 
    URBAN.POP.PROP, data = covid19_clean_data)

Residuals:
   Min  1Q Median    3Q      Max 
-12602  -6585  -3770   -402 391010 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             -2580.816     214.900  -12.01   <2e-16 ***
`Hospital beds (per 10 000 population)`   -40.107         3.486  -11.50   <2e-16 ***
URBAN.POP.PROP                          15851.097     360.058   44.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20670 on 73773 degrees of freedom
Multiple R-squared:  0.02577,    Adjusted R-squared:  0.02574 
F-statistic: 975.6 on 2 and 73773 DF,  p-value: < 2.2e-16
```

4. Linear model of deaths vs. beds + confirmed.

```
> model_bed_conf <- lm(data=covid19_clean_data, formula=deaths~(confirmed /
`Hospital beds (per 10 000 population)`))
> summary(model_bed_conf)
```

```
Call:
lm(formula = deaths ~ (confirmed/`Hospital beds (per 10 000 population)`),
        data = covid19_clean_data)

Residuals:
        Min      1Q  Median   3Q      Max
-150531        -906    -791   -744  169342

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                              7.910e+02  3.513e+01   22.52   <2e-16 ***
confirmed                                1.872e-02  5.385e-05  347.58   <2e-16 ***
confirmed:`Hospital beds (per 10 000 population)` 1.088e-04  1.902e-06   57.19
<2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9201 on 73773 degrees of freedom
Multiple R-squared:  0.8069,    Adjusted R-squared:  0.8069
F-statistic: 1.541e+05 on 2 and 73773 DF,  p-value: < 2.2e-16
```

5. Linear model of deaths vs. total population.

```
> model_totlpop <- lm(data=covid19_clean_data, formula=deaths~POP.TOTL)
> summary(model_totlpop)
```

```
Call:
lm(formula = deaths ~ POP.TOTL, data = covid19_clean_data)

Residuals:
   Min  1Q Median    3Q      Max
-56815  -4382  -4024  -3135 389308
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.000e+03  7.674e+01   52.12   <2e-16 ***
POP.TOTL    3.853e-05  4.854e-07   79.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20100 on 73774 degrees of freedom
Multiple R-squared:  0.07869,    Adjusted R-squared:  0.07867
F-statistic:  6301 on 1 and 73774 DF,  p-value: < 2.2e-16
```

6. Linear model of deaths vs. proportion of confirmed from total population
   (Proportion = confirmed / Total Population)

```
Call:
lm(formula = deaths ~ (confirmed/POP.TOTL), data = covid19_clean_data)

Residuals:
   Min     1Q Median     3Q    Max
-88239   -509   -216   -205 154825

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         2.093e+02  2.943e+01    7.113 1.15e-12 ***
confirmed           2.749e-02  4.667e-05  589.035  < 2e-16 ***
confirmed:POP.TOTL -1.031e-11  5.378e-14 -191.692  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7682 on 73773 degrees of freedom
Multiple R-squared:  0.8654,    Adjusted R-squared:  0.8654
F-statistic: 2.371e+05 on 2 and 73773 DF,  p-value: < 2.2e-16
```

7. Linear model of deaths vs. proportion of beds to confirmed cases
   (Proportion = beds / confirmed cases)

```
> model_bed_conf <- lm(data=covid19_clean_data, formula=deaths~(`Hospital beds
(per 10 000 population)` / confirmed))
> summary(model_bed_conf)
```

```
Call:
lm(formula = deaths ~ (`Hospital beds (per 10 000 population)`/confirmed),
          data = covid19_clean_data)

Residuals:
   Min  1Q Median    3Q      Max
-96606  -3911  -2153         529 212461

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         5.656e+03  8.429e+01   67.10   <2e-16 ***
`Hospital beds (per 10 000 population)`             -1.450e+02  2.381e+00  -60.91   <2e-16
***
`Hospital beds (per 10 000 population)`:confirmed  6.103e-04  2.181e-06  279.87
<2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14580 on 73773 degrees of freedom
Multiple R-squared:  0.5151,    Adjusted R-squared:  0.515
F-statistic: 3.918e+04 on 2 and 73773 DF,  p-value: < 2.2e-16
```

For the process of data wrangling, both the covid19 deaths data and confirmed cases data were downloaded and loaded into RStudio. The data were subset to remove the Lat and Long columns since they were not needed. The column ‘Country/Region’ was then renamed to ‘Country’ in both these data sets. These datasets were then pivoted longer and the counts were summarised to create a day column and a confirmed/deaths count column. The two datasets were then joined together by Country and day.

The hospital bed data from the World Health Organization was then downloaded and read into RStudio. In order to get the most recent year of each country, the data was grouped by Country then summarised to get the most recent year. Inner join was used to match the bed counts with the most recent years. The country names in the bed data were then renamed in order to match the confirmed+deaths data country names. The two tables were then joined and the most recent year column from the bed data was removed.

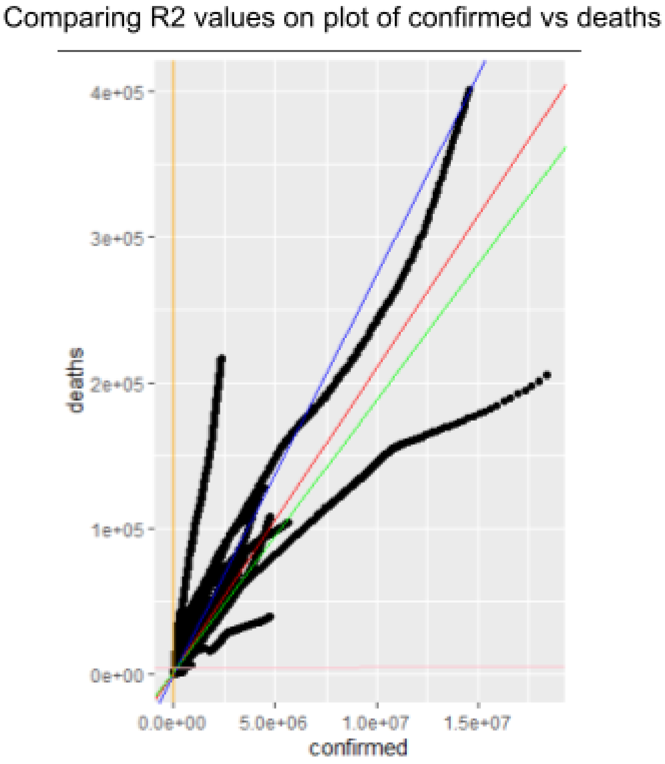
The demographics table was then downloaded from Titanium and had the “Country Code” and “Series Name” columns removed. The “Country Name” column was renamed to “Country”. The data was then pivoted wider to create variable columns for the age group populations, life expectancies, urban population, and total population of each country. The population columns were renamed to remove the “SP.” prefix and columns that were separated into female/male counts were added together. The female/male counts for population were then removed. Several country names were then renamed in order to match the confirmed+deaths+bed data. The demographic data was then joined with the confirmed+deaths+bed data to create covid19_data.

In order to find which countries were not included, rows with missing values were isolated and summarised by death sums. A few of the countries were then renamed in order for the country names to match across the demographic, confirmed/deaths, and bed data. After adding more countries to the final data, rows with missing data were removed to create covid19_clean_data.

For the linear modeling, the different combinations of predictor variables that were tried were : deaths vs. confirmed, deaths vs. Proportion of urban population, deaths vs. deaths vs. beds + proportion of urban population, deaths vs. beds + confirmed, deaths vs. proportion of confirmed from total population, and deaths vs. proportion of beds to confirmed cases.

The confirmed cases were chosen since they are the most obvious variable to be relevant to deaths. Proportion of urban population was then chosen to test whether or not countries with higher urban densities were more at risk than others. Since the urban population proportion had a very little R2, beds were added to check for beds + urban population proportion to see if there would be a significant change. The proportion of confirmed cases to beds was then tested to see if increasing the number of beds would significantly affect the relationship to deaths. The variable total population was then tested to see if a country's population density could be a core issue. The proportion of confirmed cases to the total population was then chosen to see if a country's infected proportion could be a better predictor than the confirmed cases alone. The variable proportion of beds to confirmed cases was then chosen to test the significance of availability of hospital beds to confirmed cases.

```
ggplot(data=covid19_clean_data)+ geom_point(mapping = aes(x=confirmed, y=deaths))
+ geom_abline(intercept = model_confirmed$coefficients[1],
slope=model_confirmed$coefficients[2], color="red")+ geom_abline(intercept =
model_conf_tpop$coefficients[1], slope=model_conf_tpop$coefficients[2], color="blue")
+ geom_abline(intercept = model_bed_conf$coefficients[1],
slope=model_bed_conf$coefficients[2], color="green")+ geom_abline(intercept =
model_totlpop$coefficients[1], slope=model_totlpop$coefficients[2], color="cyan")+
geom_abline(intercept = model_urb_pop_prop$coefficients[1],
slope=model_urb_pop_prop$coefficients[2], color="yellow")+ geom_abline(intercept =
model_totlpop$coefficients[1], slope=model_totlpop$coefficients[2], color="pink")+
geom_abline(intercept = model_bed_urb$coefficients[1],
slope=model_bed_urb$coefficients[2], color="orange")
```



The most significant standalone factor in relation to Covid19 deaths is the number of confirmed cases. The least significant variables that were tested were tested without incorporating the variable confirmed cases were beds, total population and proportion of urban population. However, by finding the proportion of confirmed cases in respect to the total population, a much greater R2 can be found as opposed to only constructing a linear model with confirmed cases vs. deaths. Deaths vs. proportion of confirmed from total population yielded a R2 of 0.8654 compared to the R2 of 0.7983 from deaths vs. confirmed cases.