Eugene Lee
Jarrett Chien
Joshua Roberts

Project 1 Report

For data wrangling, we needed to put the data into a tidy format. For example, tidying hospital bed data by getting the most recent hospital bed numbers per 10,000 for each country in the dataset. See full data wrangling techniques in the attached R file.

```
tidy_hospital_beds <- hospital_beds %>% group_by(Country)%>% mutate(maxYear = max(Year)) %>% filter(Year==maxYear) %>% select(-Year)
```

The variables we chose were mainly age related, as we wanted to see which age groups are vulnerable to COVID-19. We also were interested in seeing correlation with available hospital beds and expectancy. We performed data transformation on population above 65 divided by total population to obtain proportion of population above 65. We repeated this step for those under 15.

```
Call:
lm(formula = num_deaths ~ num_cases, data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
-7648.3    40.7    46.8    46.8  9026.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.684e+01  5.175e+00  -9.051   <2e-16 ***
num_cases    8.351e-02  3.732e-04 223.795   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 583.6 on 13138 degrees of freedom
Multiple R-squared:  0.7922,    Adjusted R-squared:  0.7922
F-statistic: 5.008e+04 on 1 and 13138 DF,  p-value: < 2.2e-16
```

Eugene Lee
Jarrett Chien
Joshua Roberts

The next model describes the number of deaths in terms of the number of cases and the number of citizens aged 80 and up in each country. Adding the population 80 and up for the country did not add much to the R2 value, so we concluded that the predictor was not very good.

```
lm(formula = num_deaths ~ num_cases + POP.80UP, data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
-7647.8     2.6    35.6    62.7  9028.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.1308409 13.4938289   3.122   0.0018 **
num_cases    0.0832723  0.0003739 222.687   <2e-16 ***
POP.80UP    -0.0257370  0.0036061  -7.137    1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 582.5 on 13137 degrees of freedom
Multiple R-squared:  0.793,     Adjusted R-squared:  0.793
F-statistic: 2.516e+04 on 2 and 13137 DF,  p-value: < 2.2e-16
```

This model describes the number of deaths in terms of the number of cases and number of citizens aged 65 and up in each country. Lowering the age increased the R-squared value to 0.8239, which indicates a higher correlation between being over 65 and dying from COVID-19.

```
Residuals:
    Min      1Q  Median      3Q     Max
-8273.0   -18.0     6.1    31.3  7950.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.721e+01  1.265e+01   4.522 6.18e-06 ***
num_cases    9.047e-02  3.773e-04 239.782  < 2e-16 ***
POP.80UP    -1.592e-02  3.420e-03  -4.655 3.28e-06 ***
POP.65UP    -1.950e-05  4.074e-07 -47.877  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 539 on 13046 degrees of freedom
  (90 observations deleted due to missingness)
Multiple R-squared:  0.824,     Adjusted R-squared:  0.8239
F-statistic: 2.036e+04 on 3 and 13046 DF,  p-value: < 2.2e-16
```

Eugene Lee
Jarrett Chien
Joshua Roberts

This model describes the number of deaths in terms of the number of cases and number of citizens older than 65 and younger than 15 in each country. Adding the under 15 population increased the R-squared value to 0.8309, which indicates a slightly higher correlation than being over 65 and dying from COVID-19.

```
Call:
lm(formula = num_deaths ~ num_cases + POP.80UP + POP.65UP + POP.0014,
    data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
-8354.0   -28.9     9.4    45.6  7710.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.688e+01  1.240e+01   4.587 4.54e-06 ***
num_cases    9.269e-02  3.819e-04 242.690  < 2e-16 ***
POP.80UP    -2.034e-02  3.357e-03  -6.061 1.39e-09 ***
POP.65UP    -3.436e-05  7.552e-07 -45.498  < 2e-16 ***
POP.0014     5.437e-06  2.346e-07  23.176  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 528.2 on 13045 degrees of freedom
  (90 observations deleted due to missingness)
Multiple R-squared:  0.8309,    Adjusted R-squared:  0.8309
F-statistic: 1.603e+04 on 4 and 13045 DF,  p-value: < 2.2e-16
```

The next model describes the number of deaths by the proportion of the population older than 65. With an R-squared of 0.8096, there is a high correlation of an increasing proportion of population older than 65 and COVID-19 deaths.

```
Call:
lm(formula = num_deaths ~ POP.65_prop + num_cases, data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
-8054.0    12.9    15.3    19.3  8440.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.450e+01  5.076e+00  -2.856  0.00429 **
POP.65_prop -4.967e-03  1.435e-04 -34.607  < 2e-16 ***
num_cases    8.671e-02  3.701e-04 234.295  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560.5 on 13047 degrees of freedom
  (90 observations deleted due to missingness)
Multiple R-squared:  0.8097,    Adjusted R-squared:  0.8096
F-statistic: 2.775e+04 on 2 and 13047 DF,  p-value: < 2.2e-16
```

Eugene Lee
Jarrett Chien
Joshua Roberts

This model describes the number of deaths by the number of people 65 and older, or 14 and younger, and also the proportions of these age groups. With an R-squared of 0.832, having a large proportion AND quantity of population older than 65 or younger than 15 correlates highly with COVID-19 deaths

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.250e+02  2.947e+01  -7.636 2.40e-14 ***
num_cases      9.256e-02  3.804e-04 243.309  < 2e-16 ***
POP.0014_prop  3.666e+02  7.148e+01   5.128 2.97e-07 ***
POP.65_prop    1.282e+03  1.261e+02  10.166  < 2e-16 ***
POP.0014       6.071e-06  2.481e-07  24.471  < 2e-16 ***
POP.65UP      -3.625e-05  7.953e-07 -45.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 526.5 on 13044 degrees of freedom
  (90 observations deleted due to missingness)
Multiple R-squared:  0.832,     Adjusted R-squared:  0.832
F-statistic: 1.292e+04 on 5 and 13044 DF,  p-value: < 2.2e-16
```

This model describes the number of deaths in terms of the number of available hospital beds. With an R-squared of 0.002949, there is very little correlation between having more available beds and more COVID-19 deaths.

```
Call:
lm(formula = num_deaths ~ beds, data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
 -489.7  -181.8  -131.6  -100.4 23937.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.3389    17.1604   4.565 5.04e-06 ***
beds          2.8769     0.4557   6.313 2.82e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1278 on 13138 degrees of freedom
Multiple R-squared:  0.003025,  Adjusted R-squared:  0.002949
F-statistic: 39.86 on 1 and 13138 DF,  p-value: 2.819e-10
```

Eugene Lee
Jarrett Chien
Joshua Roberts

This model describes the number of deaths in terms of the life expectancy. With an R-squared of 0.01856, there is very little correlation between having a longer life expectancy and COVID-19 deaths.

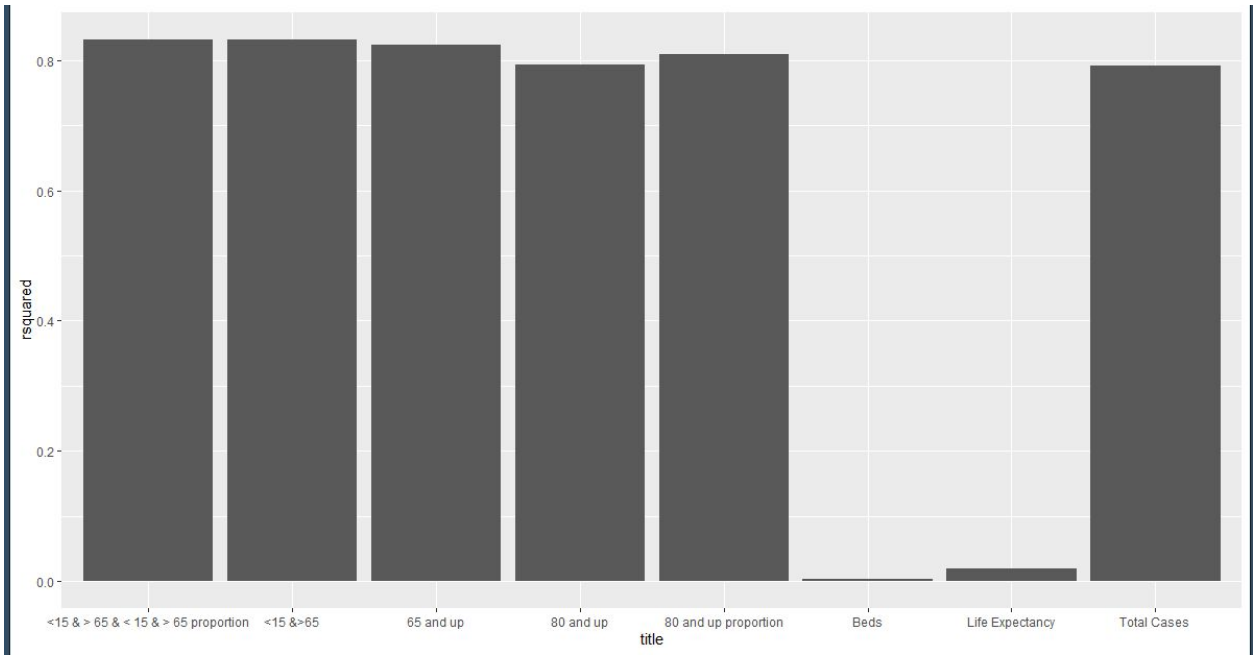
```
Call:
lm(formula = num_deaths ~ lifeExpectancy, data = tidy_covid_19)

Residuals:
    Min      1Q  Median      3Q     Max
 -413.8  -263.3  -173.7     5.3 23717.2

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.193e+03  1.494e+02  -14.68   <2e-16 ***
lifeExpectancy  8.093e-01  5.123e-02   15.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1268 on 13138 degrees of freedom
Multiple R-squared:  0.01864,   Adjusted R-squared:  0.01856
F-statistic: 249.5 on 1 and 13138 DF,  p-value: < 2.2e-16
```

The following bar chart depicts the evaluation metrics (adjusted R-squared values) of our models.



Eugene Lee
Jarrett Chien
Joshua Roberts

Conclusion

In conclusion, our models suggest that there is a high correlation of being younger than 15 or older than 65 and dying from COVID-19. In addition, there is also a high correlation for populations with larger proportions of younger than 15 or older than 65 and higher COVID-19 deaths. On the other hand, there was almost no correlation between more hospital beds and COVID-19 deaths. There was very little correlation with life expectancy of a country's population and COVID-19 deaths, as well.