Data Preparation & Information Modeling Report

Human Freedom vs Suicide Rates

This project will use two datasets obtained from kaggle.com to explore relationships between a country's human freedom – measured by their Human Freedom Index (HFI) score – and their suicide rates, population, and gross domestic product (gdp) per capita (Sutter, 2020; Yates, 2018). I joined the two datasets by country and year to determine if there is any link between freedom, number of suicides scaled by population, a country's gdp per capita, and its total population. More specifically my analysis examined the links between population size and freedom, female freedom scores and female suicide rates, and important factors that contribute to human freedom. I posited that high female freedom countries will have lower female suicide rates than low female freedom countries and that small population countries will have more freedom than large population countries.

Data Preparation

Before beginning exploratory data analysis (EDA), it's important to assess the initial data format and clean up variable types and names. By default when R reads in a csv file, it assigns categorical variables a character format when often a factor format is more appropriate (R Core Team, 2021). Applying this approach, the categorical variables *country, region, sex,* and *age* were reassigned to be factors. Each *age* value initially had the suffix 'years' - like '5-14 years' for the first age bucket - but it was chopped off as 'years' doesn't add any value to the label. In comparing the HFI and suicides datasets, I noticed that HFI records start back in 2008, while data on suicides began in 1987. However, my project is primarily focused on the HFI freedom dataset, so the suicides data was filtered to remove records before 2008 prior to exploration and joining. Suicide counts initially were broken down by age and sex categories (in addition to year and country). Some analyses compare suicide rates to freedom scores and do not involve sex or age, and this required aggregating suicides and population into overall counts and dividing them to get a scaled rate of suicides per one hundred thousand persons. This scaling and formatting is common in the medical and sociological fields and gives a rate proportional to a country's population in a usable format - short values like 5.6 rather than long values like 0.0000056. For a later analysis on female freedom, a few variables that measure female freedom – inheritance rights, movement, female to female marriages, and divorce – are averaged together to resolve some missing values in each individual variable and provide a better correlation plot (Vásquez & McMahon, 2020).

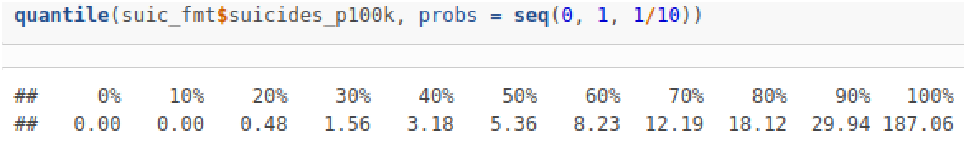
Joining Datasets. While joining two datasets on the same two variables – country and year – may seem simple, it did involve a fair amount of work to identify and fix country name mismatches, like Cabo Verde vs Cape Verde. I checked country names with the Duckduckgo search engine and Wikipedia entries ("Wikipedia", 2021). I was able to fix six country names so they could join correctly, but over fifty more were present in only one dataset or were present in both but the years were not aligned. In total, eighty-four countries were joined together to create a clean dataset consisting of seventeen variables across almost seven thousand rows.

Suicides Data Exploration

I am fairly comfortable with Python and R, but I prefer the pipe workflow in R with its *tidyverse* suite of packages – readr for reading csv files, dplyr for data manipulation, tidyr for column work, stringr for string functions (v1.3.0; Wickham et al., 2019). I've used R for past analysis projects in this masters program, but this time I wanted to learn a new approach to EDA with R Markdown (Xie et al., 2018). Doing EDA this way took longer, but the upside is the exploration process is more detailed and accessible for other data analysts. R Markdown combines code with markup comments and code results all in one output so that analysts can show their thought process alongside the results from each step.

Suicides data. I started with the typical *summary*, *head*, and *tail* functions and saw that there were no missing values in the data and nothing out of the ordinary looking at the top or bottom of the dataset. The value in checking the top and bottom is verifying that all records were loaded and no additional metadata – column names, number of rows, date created, etc – snuck in from the top or bottom of the file. I first sorted the data by descending suicide counts to see what years and countries had the highest suicide counts. It turns out that the US and Russian men in the age bracket 35-54 were at the top, but after moving to the scaled variable *suicides_p100k* they were no longer present. This is due to the large populations that US and Russia have. Without scaling, their populations would incorrectly weight their suicide counts.

Scaled suicide rates. R's *summary* gives some detail on the distribution of suicides per one hundred thousand persons, but it's *quantile* and *boxplot* functions provide more resolution on the shape of suicide rates in the data. From Figure 1 below I can see that a significant portion of the records have a count of zero and those that don't remain small until the last quantile between ninety-one and one hundred percent.

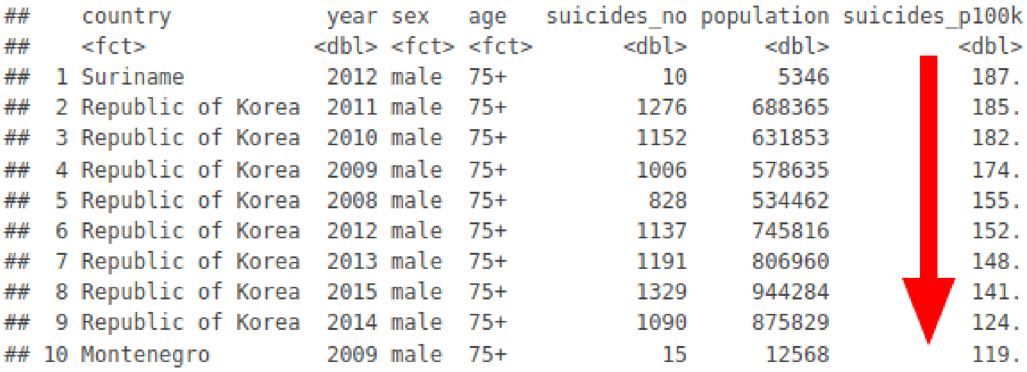


```
quantile(suic_fmt$suicides_p100k, probs = seq(0, 1, 1/10))
```

```
##     0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100% 
##   0.00   0.00   0.48   1.56   3.18   5.36   8.23  12.19  18.12  29.94 187.06
```

Figure 1: Quantile breakdown for Scaled Suicide Rates

This indicates a right skewed distribution with the majority of values between zero and twenty with a few large positive values – coming mainly from elderly males from the Republic of Korea (as shown in Table 1 below).



```
##    country            year sex   age   suicides_no population suicides_p100k
##    <fct>             <dbl> <fct> <fct>       <dbl>      <dbl>          <dbl>
##  1 Suriname           2012 male  75+            10       5346           187.
##  2 Republic of Korea  2011 male  75+          1276     688365           185.
##  3 Republic of Korea  2010 male  75+          1152     631853           182.
##  4 Republic of Korea  2009 male  75+          1006     578635           174.
##  5 Republic of Korea  2008 male  75+           828     534462           155.
##  6 Republic of Korea  2012 male  75+          1137     745816           152.
##  7 Republic of Korea  2013 male  75+          1191     806960           148.
##  8 Republic of Korea  2015 male  75+          1329     944284           141.
##  9 Republic of Korea  2014 male  75+          1090     875829           124.
## 10 Montenegro         2009 male  75+            15      12568           119.
```

Table 1: Descending Scaled Suicide Rates, Korea at top

Zero suicides? Greater than ten percent of records have stated zero suicides for a given year and country. R counted the number of records each country has with zero and non-zero suicide counts and those with the highest percentage of zero suicide counts – Antigua and Barbuda, Grenada, Barbados, Maldives – were not included in the analysis due to their not being present in the HFI dataset.

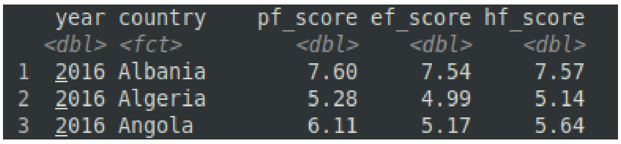
Age and Sex. Though age and sex are not featured in this analysis, besides female suicide rates and freedom correlation, I decided to explore their contributions to the suicides data as part of the EDA process. Suicides seem to be more common in men than women and in elderly vs young folk, consistent with what Table 1 above is showing. For a more appealing visual of this trend and all other project resources, please see my github page.

Distributions. Each of the numeric variables – *suicides_no, population, suicides_p100k,* and *gdp_per_capita* – had a heavy right skew, enough to where each histogram looked a lot like a letter L. While log transformations are a bit harder to interpret, I felt it was necessary to use them here to get a better view of the variable distributions.

Time-series. Lastly, I wanted to get some feel for suicide rate trends over time and created an interactive plotly graphic. It is quite busy since there are ninety-four countries being displayed at one time. Fortunately, the interactive nature allowed me to selectively show only a few countries at a time with a few clicks of the mouse. While this dataset didn't have any missing data in the form of NAs, I could see from this plot that several countries suicide records were not included in the dataset. These tended to be in the smaller countries that were ultimately removed after joining to the HFI dataset, but it's good to keep in mind.

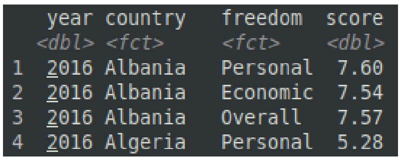
HFI Data Exploration

The HFI dataset is much larger and contains over one hundred variables, so it was important to me to limit the scope and focus in on the ones that would help answer the questions I laid out. I narrowed down the variables to year, country, region, five female freedom variables, personal freedom score, economic freedom score, and overall human freedom score. Like before with the suicides dataset, I started with *summary*, *head*, and *tail* functions to check if all data was loaded in and was in a good format.



```
   year country    pf_score ef_score hf_score
  <dbl> <fct>         <dbl>    <dbl>    <dbl>
1  2016 Albania        7.60     7.54     7.57
2  2016 Algeria        5.28     4.99     5.14
3  2016 Angola         6.11     5.17     5.64
```

Figure 2: Before: Wide with 3 Score Variables



```
   year country freedom  score
  <dbl> <fct>   <fct>    <dbl>
1  2016 Albania Personal  7.60
2  2016 Albania Economic  7.54
3  2016 Albania Overall   7.57
4  2016 Algeria Personal  5.28
```

Figure 3: After: Long format

Freedom scores. Since the overall freedom score, will be the outcome variable in our modeling step, I looked at its distributions first compared to its personal and economic subscores with a density plot using R's ggplot2 package (Wickham, 2016). To construct a superposed plot like this, R needs data to be in a tidy long format. The *pivot_longer* function can pivot from wide to long formats as shown in Figures 2 and 3 above. Figure 4 below is a

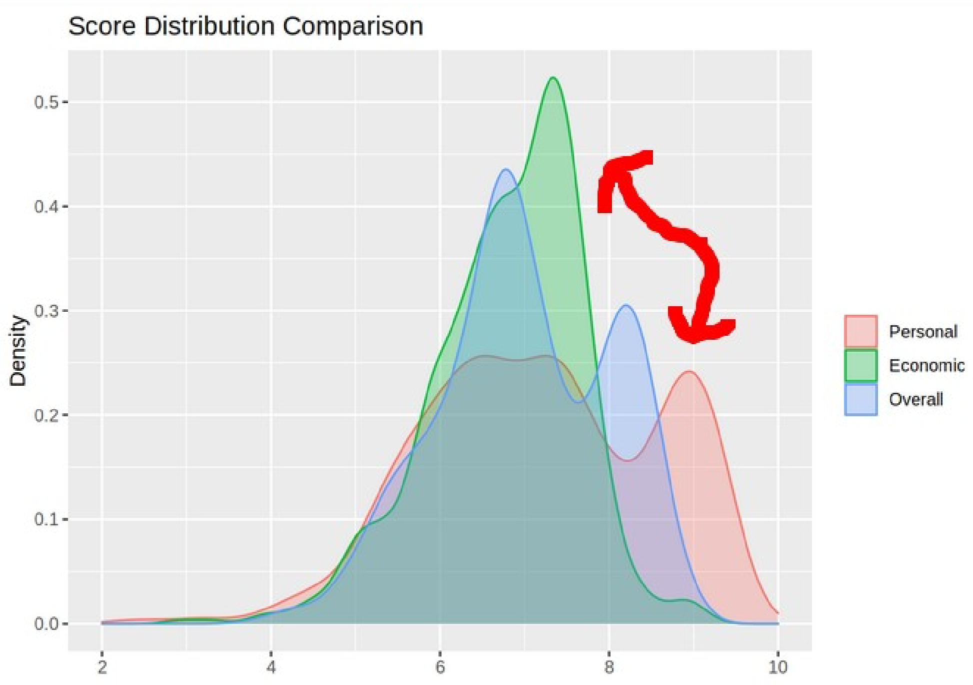


Figure 4: Score Distribution by Type

plot showing the distribution of scores for each of the three types of scores and highlights that economic scores tend to be in a lower tighter range than either personal or overall freedom scores. As the arrow annotations indicate, countries were more likely to have higher personal freedom scores than economic or overall scores.

Female freedom. Turning our attention to the five female freedom variables, there are quite a few missing values ranging from around five to sixty percent. This led me to average each row's five scores together to partially resolve the missing values problem and lead to a simpler plot representation. Figure 5 shows that the most common average female freedom score was ten or very near ten, making up more than half of the records.

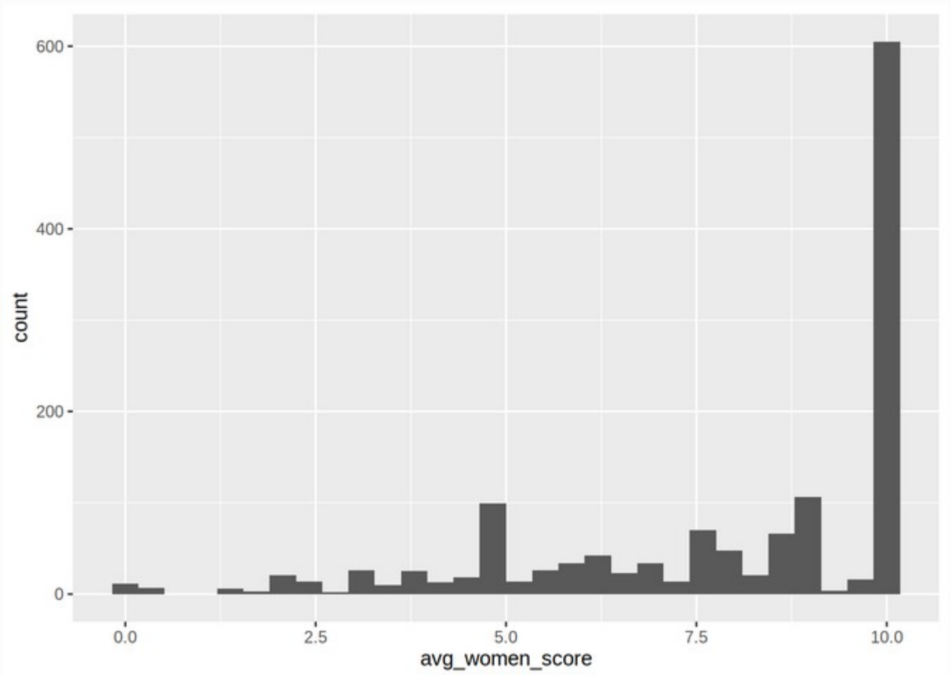


Figure 5: Average Women Score Distribution

Information Modeling

The earlier data preparation section details joining the datasets together, so now that the two datasets have fused into one I can start find correlations and apply other techniques to answer questions. To identify any relationships between the three freedom scores, scaled suicide rates, and gdp per capita, I used the R package *psych* which provides a nice pairs plot function *pairs.panels* (Revelle, 2021). It creates scatter plots with loess smoothed lines, histograms and density plots for each variable along the diagonal, and correlation coefficients with their level of significance ("Correlation Plot in R Correlogram [WITH EXAMPLES]", n.d.). Figure 6 below shows the results of this plot, and it seems that gdp per capita has the strongest correlation with overall freedom score (0.55), but scaled suicides curiously is correlated with overall freedom and personal freedom, but not with economic freedom or gdp per capita. Perhaps happiness is more strongly related to suicide rates than personal, economic, or overall human freedom? This sounds like a good idea for a future project.

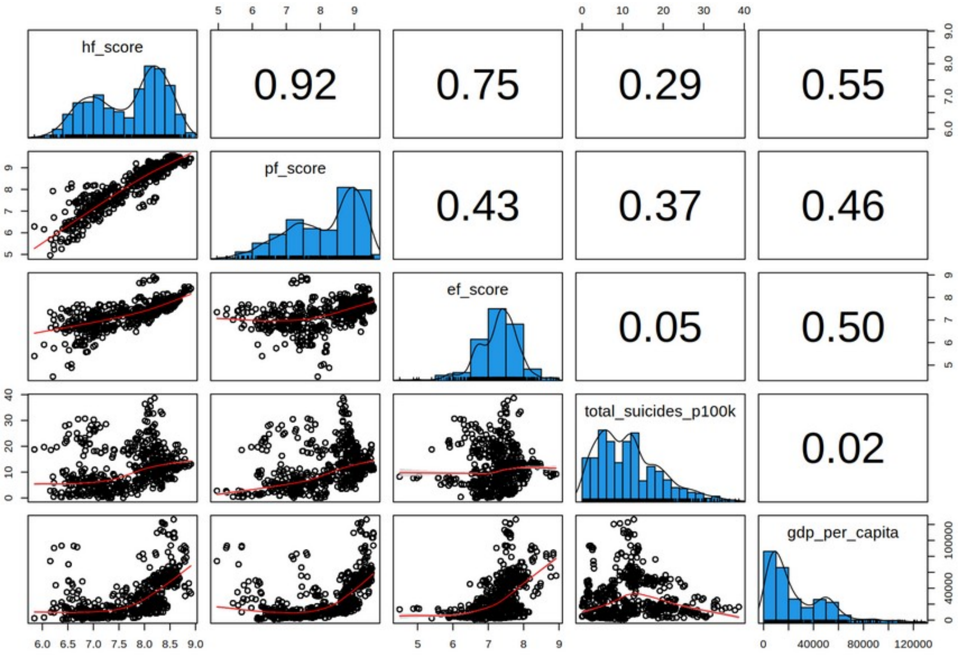


Figure 6: Psych package Correlation Plot

Female freedom and suicides. One hypothesis I had stated that countries with high female freedom scores will have lower female suicide rates than countries with low female freedom scores. For this test I focused on data from only the most recent year (2015), sorted it on average female freedom score – an average of five female-related freedom scores – and finally split the data into low and high freedom groups. With these two groups I can perform

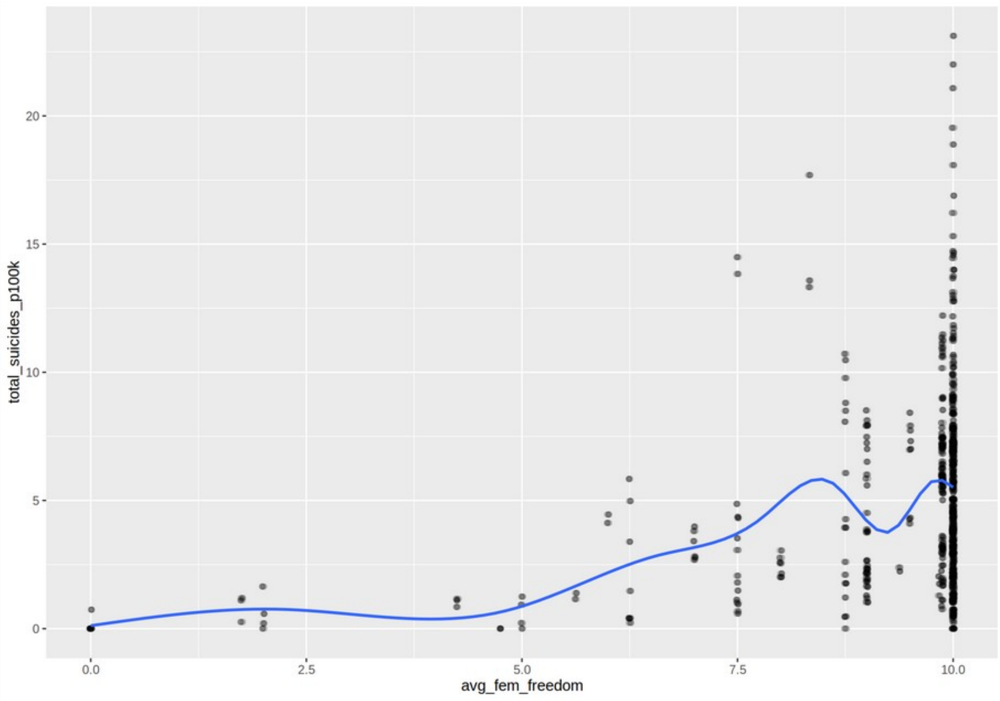
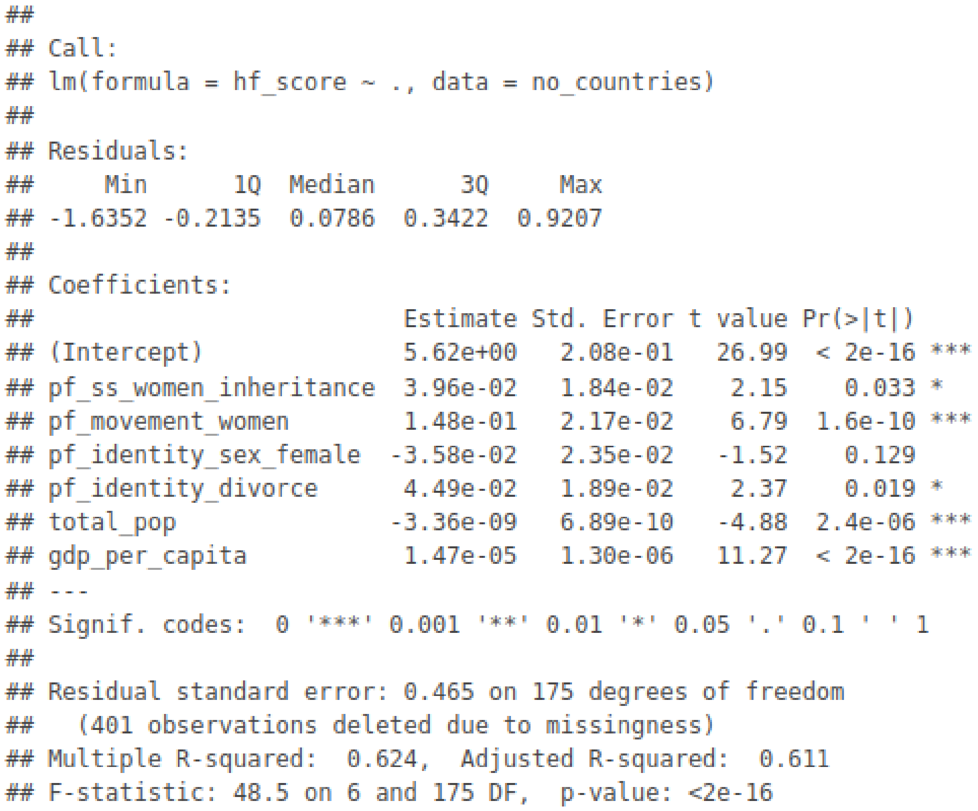


Figure 7: Female Scaled Suicide Rates vs Average Female Freedom

a two sample t-test to determine if there is a significant difference between them. The t-test results had a p-value of 0.01, indicating that there it is very likely that there is a difference in suicide rates between the two groups. “That’s excellent news” one might say, but unfortunately as Figure 5 and 7's plots show, almost sixty-five percent of countries had a female freedom score of ten! Since more than half of the countries had a perfect score, which should be put in the high freedom category and which in the low freedom one? Ultimately, there doesn't seem to be any relationship between female freedom score and female suicide rates as the scatter plot highlights that suicide rates range from low to high for countries with the same top freedom score.

Population size related to human freedom. The second hypothesis I put forward was that countries with smaller populations will have more freedom. I performed a similar process to the prior female freedom analysis where I sorted and split the data into a small and large population group. While the t-test results were only significant at ninety percent confidence level, at least the groups were constructed correctly with each having a different population size that didn't overlap the other. The freedom scores for both groups were not significant at a ninety-five percent confidence level (p-value of 0.06). I failed to reject the null hypothesis and concluded that it's possible that small population countries have more freedom than large population, but is only significant at the lowest significance level.

Human freedom factors. To cap off this analysis I briefly looked at factors that could have a significant impact on human freedom. I included four female freedom variables, total population, and gdp per capita to see if their linear combination could predict a country's human freedom score. I fit a linear regression model and Figure 8 shows that a few variables were significant – women inheritance rights, women's freedom of movement, divorce rights, total population, and gdp per capita. Only the female to female relationship variable was not significant. Not surprisingly, countries with higher gdp per capita also had higher human freedom scores. I explored the negative relationship between population size and freedom score earlier, but this model says that population size is significant in explaining the variance in human freedom. I planned to examine more variables in this model but time ran out.



```
##
## Call:
## lm(formula = hf_score ~ ., data = no_countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6352 -0.2135  0.0786  0.3422  0.9207
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.62e+00   2.08e-01   26.99  < 2e-16 ***
## pf_ss_women_inheritance  3.96e-02   1.84e-02    2.15    0.033 *
## pf_movement_women        1.48e-01   2.17e-02    6.79  1.6e-10 ***
## pf_identity_sex_female  -3.58e-02   2.35e-02   -1.52    0.129
## pf_identity_divorce      4.49e-02   1.89e-02    2.37    0.019 *
## total_pop               -3.36e-09   6.89e-10   -4.88  2.4e-06 ***
## gdp_per_capita           1.47e-05   1.30e-06   11.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.465 on 175 degrees of freedom
##   (401 observations deleted due to missingness)
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.611
## F-statistic: 48.5 on 6 and 175 DF,  p-value: <2e-16
```

Figure 8: R's Linear Regression Model Summary

References

Auguie, B. (2015). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.0.0. http://CRAN.R-project.org/package=gridExtra

Correlation Plot in R Correlogram [WITH EXAMPLES]. Retrieved 13 November 2021, from https://r-coder.com/correlation-plot-r/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.9, https://CRAN.R-project.org/package=psych

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A. and Crowley, J. (2021). GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally

Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

Sutter, G. (2020). The Human Freedom Index. Retrieved 28 October 2021, from https://www.kaggle.com/gsutters/the-human-freedom-index

Vásquez, I., & McMahon, F. (2020). *the Human Freedom Index 2020*. CATO and Fraser Institutes. Retrieved from https://www.cato.org/sites/cato.org/files/2021-03/human-freedom-index-2020.pdf

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wikipedia. (2021). Retrieved 13 November 2021, from https://www.wikipedia.org/

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown*. CRC Press.

Yates, R. (2018). Suicide Rates Overview 1985 to 2016. Retrieved 28 October 2021, from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016