Project Recap & Lessons Learned

Human Freedom and Suicide Rates

This project focused on two datasets obtained from kaggle.com – Human Freedom Index (HFI) and Suicide Rates Overview – to explore relationships between a country's human freedom, suicide rates, population, and gross domestic product (gdp) per capita (Sutter, 2020; Yates, 2018). The authors of the HFI, Vásquez and McMahon (2020), describe it as "a broad measure of human freedom" for many countries around the globe, where freedom is the absence of coercive constraint, also known as negative liberty. I would have liked to have seen positive liberty factors included as well, but I recognize these can be difficult to measure and can vary from person to person. For example, a 16 year old female in the United States has no laws prohibiting her from attending college or seeking medical treatment for a bad back (negative liberty), but her financial or health circumstances may keep her from choosing a particular action and fulfilling her potential (positive liberty).

Data. Both datasets contained the variables *year* and *country* and were joined accordingly. The suicide data was initially broken down by age group and sex and in some cases was aggregated to count total suicides by country and year. Suicide rates were scaled by a country's population at that time to facilitate comparison across countries and years.

Initial Requirements. With this data I can use the variables *country, year, sex, suicides per 100,000 persons, gdp per capita, population,* and 5 female-related freedom variables to explore relationships and test my hypotheses. Specifically, my analysis will examine the correlations between human freedom, personal freedom, economic freedom, total suicides per 100,000 persons, and gdp per capita, the impact a country's population size may have on its freedom, and the relation between a country's average female freedom score and its female suicide rates. I also look at factors that may contribute a country's freedom utilizing a linear regression model.

Hypotheses

I posited that high female freedom countries (top half) will have lower female suicide rates than low female freedom countries. Secondly, I state that small population countries (bottom half) will have more freedom than large population countries.

Data Preparation

Since data was obtained from kaggle, it was fairly clean and only needed a short cleaning and preparation cycle. By default R's *read_csv* function reads in categorical variables from a csv as character values and not factors – R's categorical format (R Core Team, 2021). The variables *country, region, sex,* and *age* were thus converted to factors after loading the data sets. In later programs I learned how to specify variable types within the *read_csv* function so they are correct upon data loading. For simplicity, the *age* variable was trimmed from values like '5-14 years' to just '5-14' as the audience will understand that a person's age is measured in years. While the suicide rates data already contained scaled suicide rates per 100,000 persons, often I had to sum the raw suicide and population counts and recalculate the scaled values in order to remove the age and sex categories and get overall rates.

Timeframe. Both data sets contained the same *country* and *year* variables enabling easy joining, but the available years and countries in each were different. HFI has freedom scores dating back to 2008, while suicide rates extends further to 1987. I chose to filter both to start in 2008 to simplify later analyses. 2016 records were also removed as there were significantly fewer records during that year relative to all other years. Figure 1 below shows this record difference.

```
## # A tibble: 9 x 2
## # Groups:   year [9]
##    year     n
##   <dbl> <int>
## 1  2008   912
## 2  2009   960
## 3  2010   948
## 4  2011   924
## 5  2012   876
## 6  2013   864
## 7  2014   840
## 8  2015   672
## 9  2016   150
```

Figure 1: Record Counts by Year

Joining Issues. Unlike my classmates I didn't have to deal with geolocation translations, but I still had to resolve over twenty countries that didn't immediately match up between the two data sets. I was able to fix six county names because the countries were the same, only their names were abbreviated in different ways, like "Russian Federation" vs "Russia" or "Krygyz Republic" vs "Krygyzstan". At the end the joined data set contained eighty-four countries with seventeen variables and almost seven thousand rows.

Tidy Data. For my final analysis on the temporal component of freedom, I wanted to see trends in the overall freedom score over time next to its two subscores – personal and economic freedom. To include all three on the same plot encoded with different colors required transforming my wide data set to a long data set – from three individual columns to one key and one value column. Figures 2 and 3 below illustrate this transformation.

```
##    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 Scores

```
     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
5  2016 Algeria Economic  4.99
6  2016 Algeria Overall   5.14
```

Figure 3: After: Long Format

Female Freedom. One problem I noticed during the exploratory data process is that my five variable average of female freedom was very left skewed with more than half of the records having a perfect score of ten! This made my hypothesis about high freedom countries and their suicide rates practically useless as I cannot accurately select the top half of countries ranked by female freedom.

Correlation Pairs Plot

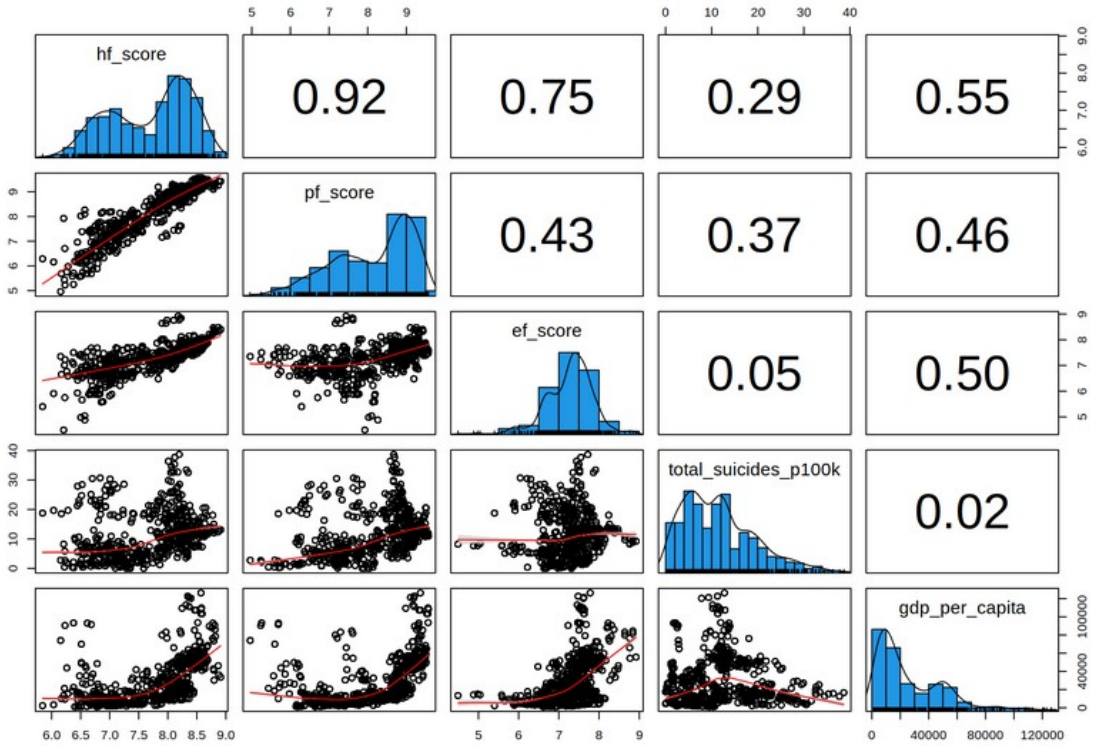

Figure 4: Psych package Correlation Pairs Plot

I employed R's *psych* package and its *pairs.panels* function to create the correlation pairs plot with histograms, density, and scatter plots shown in Figure 4. With this graphic I can see that the personal and economic freedom subscores are correlated with their overall freedom score (as expected), though personal freedom seems to be more strongly related. Gdp per capita was most strongly correlated with the overall freedom score followed by the economic score. Total suicide rates were weakly related to personal and overall freedom and curiously not

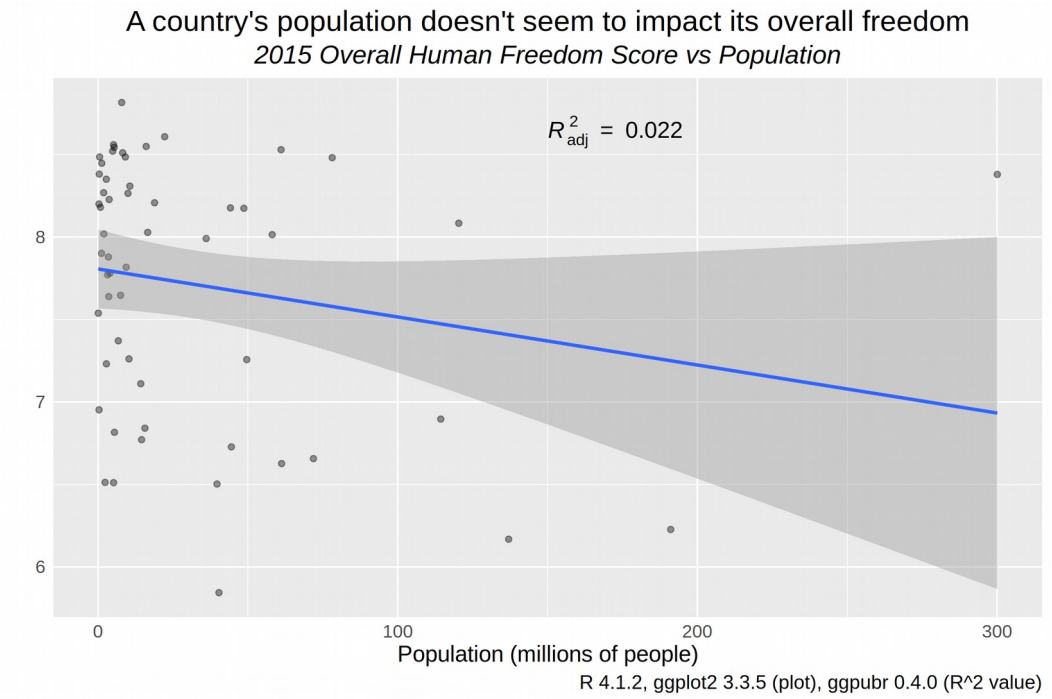

Figure 5: Scatter plot exploring relationship between Freedom and Population

related to economic freedom or gdp per capita. Kaggle had another data set on a world happiness index and perhaps happiness exhibits a stronger relationship to these variables. I am leaving this for a future analysis project.

Freedom vs Population

While my hypothesis that smaller countries had more freedom was based on speculation, I used a scatter plot and Welch's t-test to validate my claim. Figure 5 above illustrates the lack of relationship between the population of a country and its overall freedom. Very few points fall along the smoothed linear regression line in blue. The adjusted R^2 value – a measure of outcome variance explained by an input variable – is near zero. Figure 6 plots a different perspective of this relationship and its freedom zigzags shows either no relationship or a very weak negative relationship.

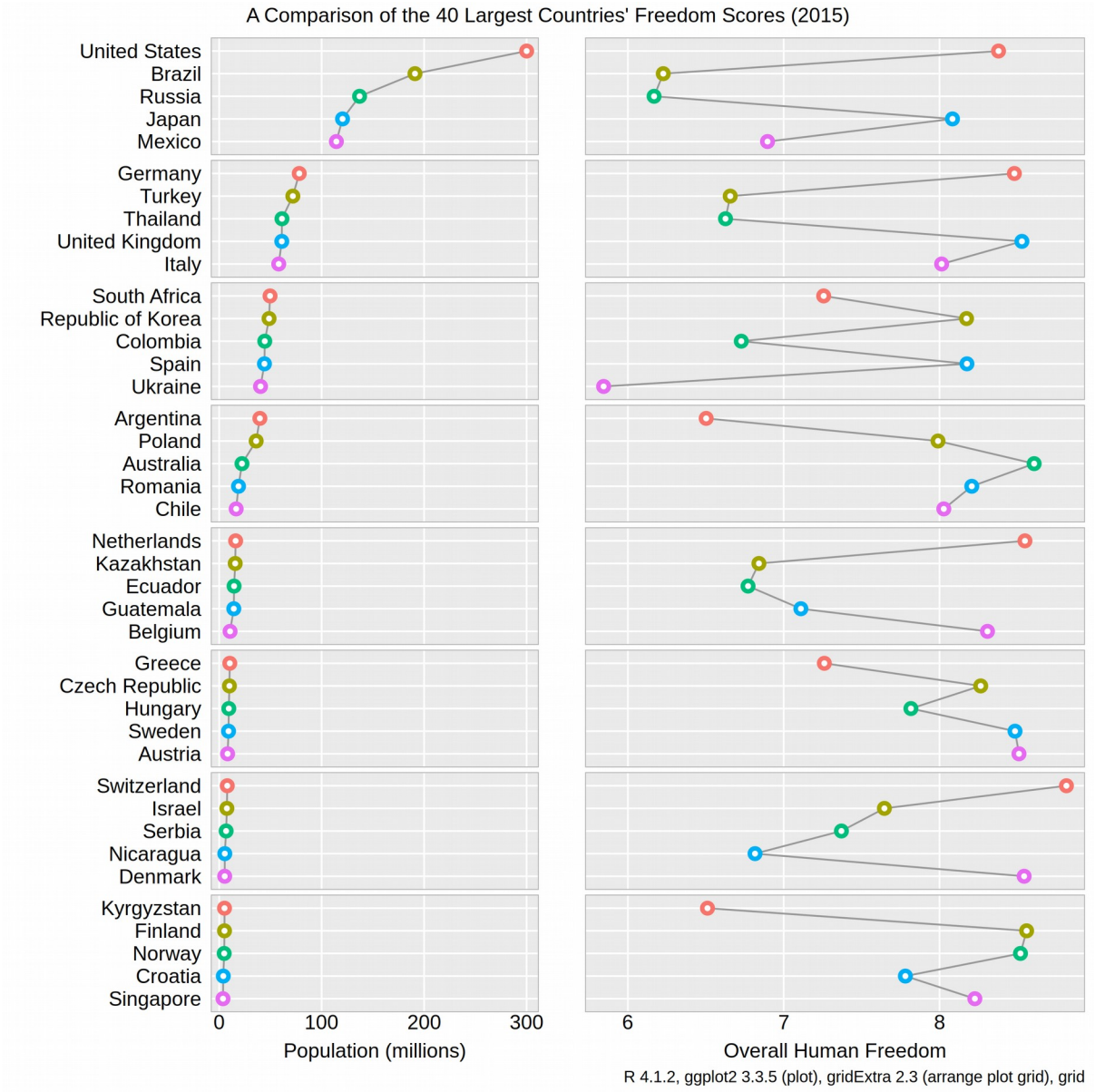

Figure 6: Row-labeled dot plot of Freedom vs Population

What is a bit crazy is that Welch's t-test comparing freedom in small and large population countries resulted in a p-value of 0.06, almost significant at a 95% confidence level. Ultimately, I fail to reject the null hypothesis that there is no difference in freedom

between small and large countries and conclude that there is not sufficient evidence to find a statistically significant difference in freedom.

Female Freedom and Suicide Rates

As mentioned earlier, I had issues determining if there was a relationship between female freedom and female suicide rates on account of the abundance of perfect ten freedom scores. I used similar techniques from the prior population analysis to determine if a relationship existed between female freedom and female suicide rates. The x-axis is an average of five women-related freedom indicators – female genital mutilation, inheritance for widows and daughters, movement, female to female relationships, and divorce. Due to the discrete nature of the average female freedom score, I chose a series of boxplots instead of a scatter plot to visualize the relation between female freedom and suicide rates and resolve an overplotting problem (Figure 7). The boxplots and outlier dots highlight the wide range of suicide rates that a country could have even with a high female freedom score.

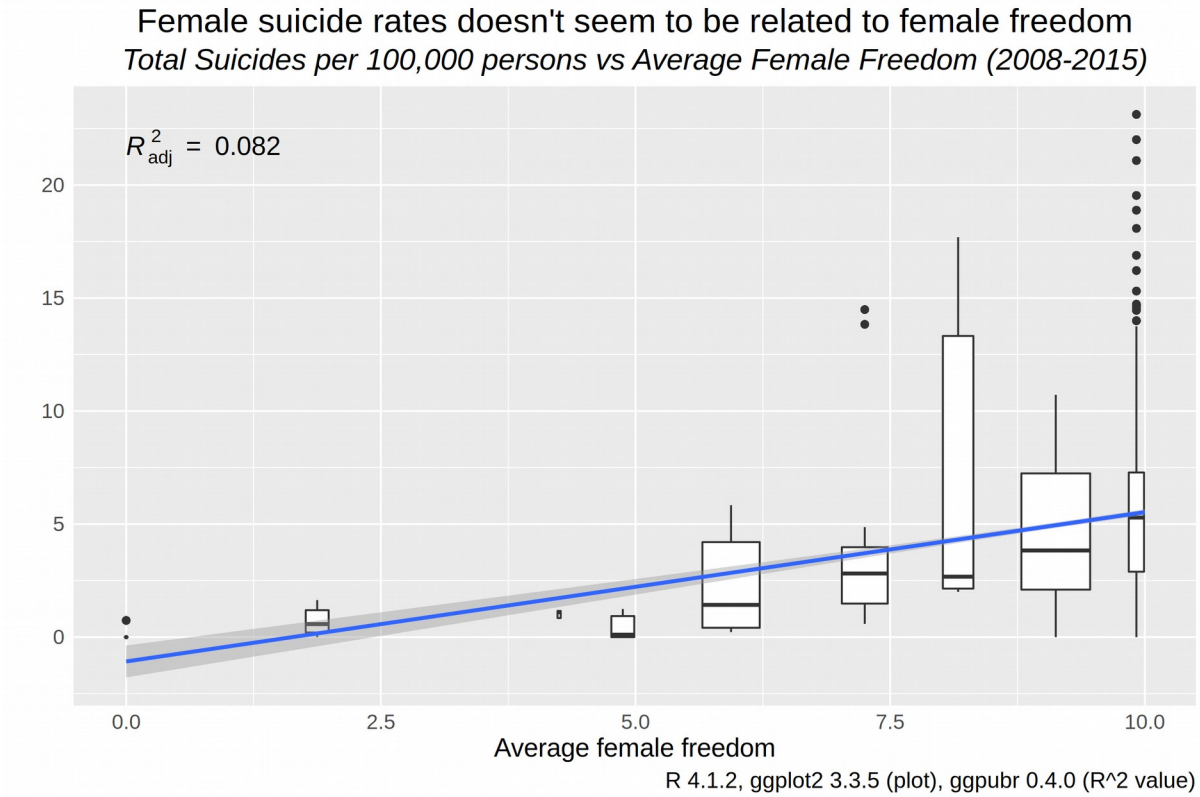

Figure 7: Boxplots showing relationship between Female Freedom and Female Suicide Rates

Though there are problems with a two sample t-test for determining the significance of relationship between female freedom and female suicide rates, the t-test did result in a p-value of 0.01. I would reject the null hypothesis and conclude that it is likely there exists a relationship between female freedom and suicide rates. While I split the group down the middle and saw freedom scores of 10 in both groups, maybe a better approach would be to include all the 10s in the 'high' freedom category and the rest in the 'low category despite their sizes being different. This seems like a simple solution to the problem of having so many countries with the same top score.

Overall Freedom and Suicide Rates

Returning to the previous 40 largest countries dataset, I can see if any overall trends appear between freedom and suicide rates. Figure 8 uses another row-labeled dot plot to

show each country's freedom and suicide rates side by side and utilizes perceptual groups of 5 countries to ease processing (Carr & Pickle, 2010). If there is any trend present it is a very weak positive one where countries with higher freedom have higher suicide rates. The lines connecting each perceptual group serve to accent any sharp changes in values from one country to the next. This helps the audience identify South Korea, Kazakhstan, Russia, and Ukraine that have abnormal suicide rates compared to their perceptual peers. I would have liked to make this into a linked micromap where the location of each country can be shown alongside its freedom and suicide rates, but sadly I was not able to find a way to do this. This would have allowed the audience to diagnose freedom from a regional or geographical perspective. Unfortunately, it seems R's *micromapST* package does not yet support world maps for its micromap legend.

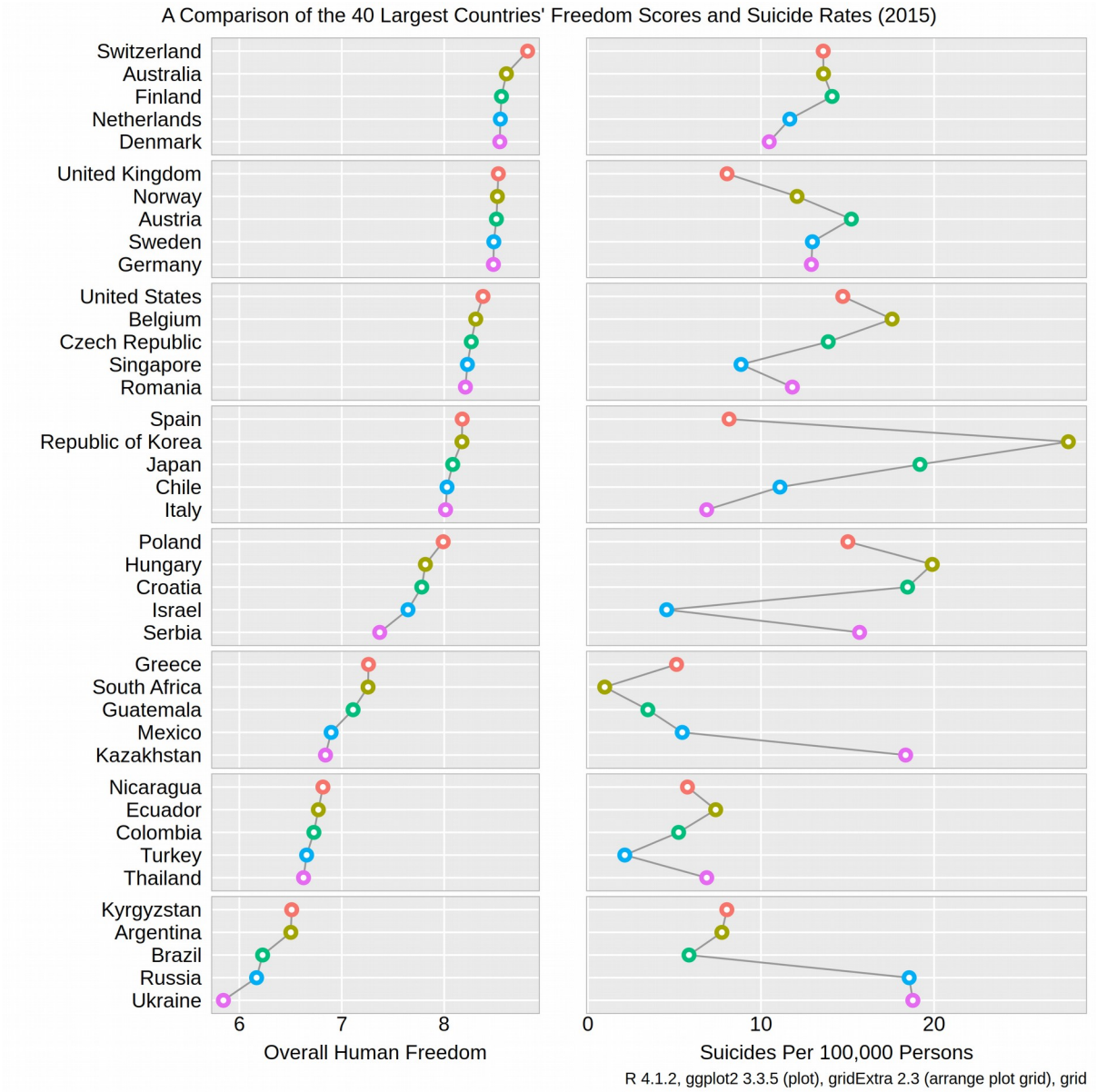

Figure 8: Row-labeled dot plot of Overall Freedom vs Suicide Rates

GDP per Capita vs Freedom

Figure 4's correlation plot stated the correlation coefficient *R* was 0.55 for gdp per capita and human freedom and Figure 9 below illustrates the strongest trend in this analysis.

With some exceptions, such as Argentina and Romania, the positive relationship between these two variables is quite evident. Countries with a vast amount of wealth near the top of the plot seem to also provide lots of freedom to their citizens, or perhaps because their citizens are so free they have become more wealthy. This plot only describes a relationship and I can only speculate as to the direction or causality of it.

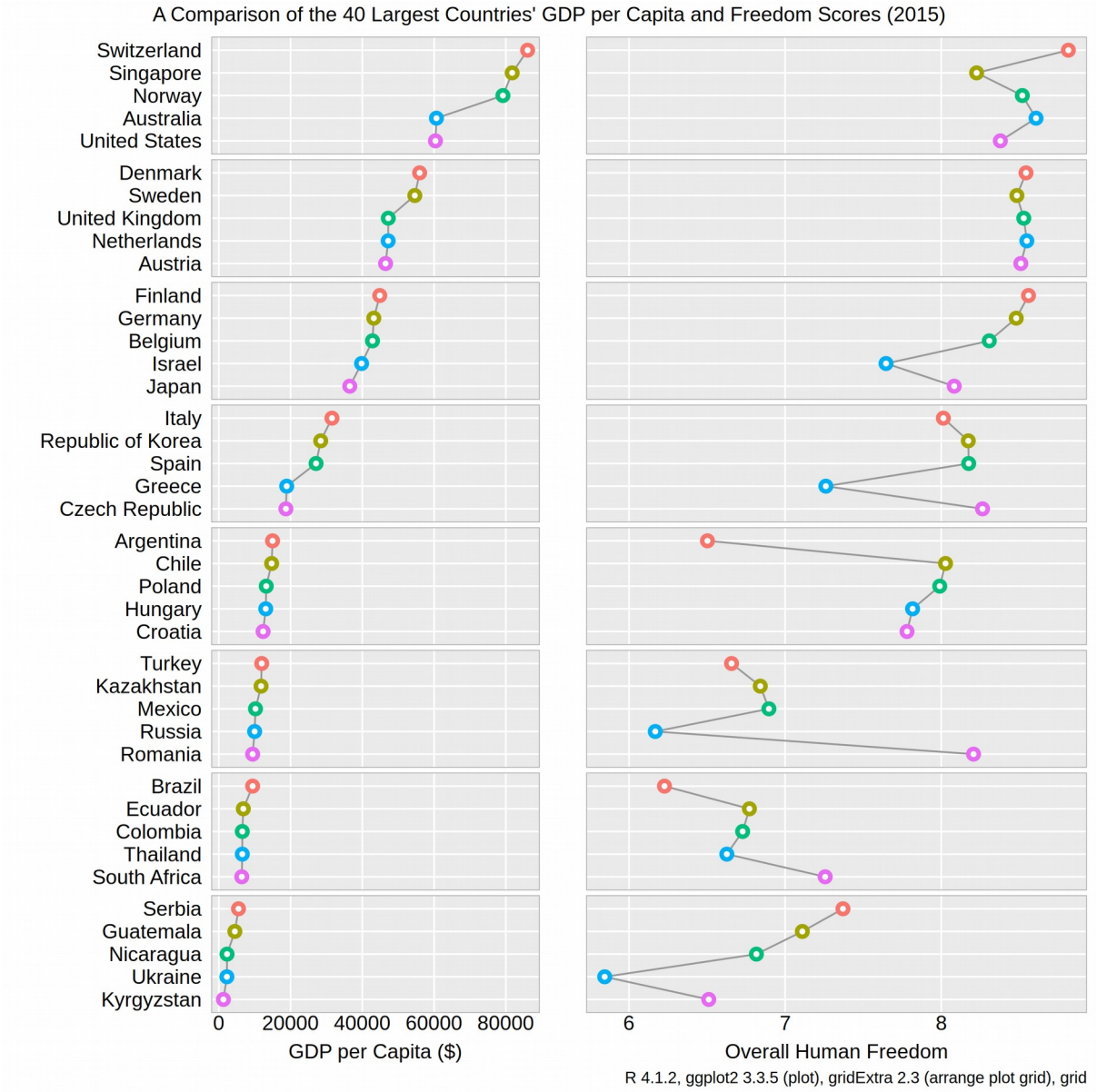

Figure 9: Row-labeled dot plot of Freedom vs GDP per Capita

Human Freedom Factors

This analysis focused on only a handful of variables that might contribute to a country's human freedom. Figure 10 below shows how I used R Markdown to display the code, model output, and comments all in one document. First the data set was aggregated to the top level by removing age and sex breakdowns for suicide rates, then the set was narrowed to a select few variables. Last the data was fed into R's *lm* function to perform a linear regression explaining the *hf_score* – overall human freedom score – with all other variables present. Four factors were very significant – women inheritance, free movement of women, total population, and gdp per capita. The adjusted R^2 value of 0.611 means this

model is fair at predicting human freedom with this variable subset. Often a linear regression model is used as a baseline model for comparison to more complex models, such as random forest, lasso regression, or neural networks. I ran out of time in this project, but I would have liked to look at the residual plots and distributions of each variable to check if any regression model assumptions were violated.

```
no_countries <- joined_no2016 %>%
    group_by(year, country) %>%
    mutate(total_suicides = sum(suicides_no),
           total_pop = sum(population),
           total_suicides_p100k = total_suicides / total_pop * 100000) %>%
    ungroup() %>%
    select(pf_ss_women_inheritance, pf_movement_women, pf_identity_sex_female,
           pf_identity_divorce, hf_score, total_pop, gdp_per_capita) %>%
    unique()

lm1 <- lm(hf_score ~ ., data = no_countries)
summary(lm1)
```

```
##
## 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 10: Linear Regression Model to Identify Freedom Factors

Temporal Analysis

So far every visualization has focused on relations between freedom and other variables but has not investigated the time component at play. Starting in 2008, 8 years of freedom and its two subscores can be aggregated by medians and means to show how they have changed over time. With the median line being greater than the mean line in all cases, there seems to be a few countries with very low freedom scores that are adversely affecting the mean. This indicates a left skew to the data. Economic freedom has been rising slightly

over the time period, while personal freedom has fallen a bit and overall freedom has remained relatively flat.

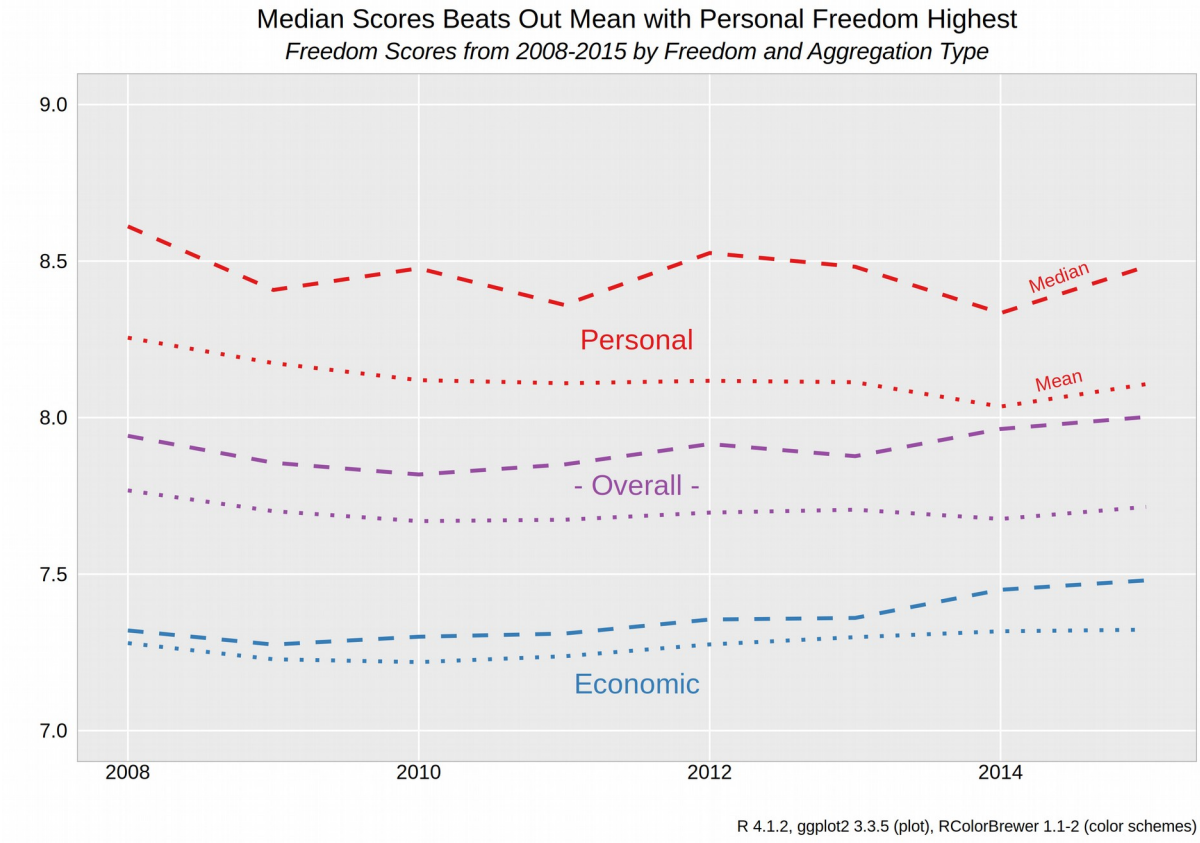

Figure 11: Time-series plot of 3 types of Freedoms Aggregated by Median and Mean

Conclusion

Many avenues were explored in this analysis and while some didn't find much evidence of relationships, it was still worthwhile, educational, and enjoyable. A data analysis is not a failure even if no significant insights are found as the absence of a relationship between variables can still be valuable information. While I speculated that countries with more freedoms given to their citizens would result in lower suicide rates, there certainly are numerous factors at play that are contributing to suicide rates.

For both hypotheses I looked at the visual evidence through scatter plots and conducted Welch's t-tests. These are basic tests to determine if two populations are significantly different with regards to a single variable. I used a linear regression model to identify important factors that contribute a country's freedom.

While I was fairly comfortable doing data analyses in R coming into this course, I am glad I tried using R Markdown as well (Xie, et. al, 2018). I see why so many analysts like using it to share their work with others and grow the data science community. Having code and results side by side with markdown comments to explain the analyst's though process makes it more accessible to a range of audiences from novice to expert data scientists.

Upon reflection of each deliverable, I believe I need to read and plan ahead better. My second report on data preparation and information modeling should focus more on exploration – cluster analysis, decision trees, and other exploratory techniques – and less on answering the specific questions I had asked at the beginning. This made my second report and my third report on information visualization overlap quite a bit. Next time I will

thoroughly read the project request and plan my analyses according to their associated part. It seems like I continually get ahead of myself and want to skip straight to the fun part – creating a variety of visualizations. This is a dangerous habit that needs to reigned in as exploring and understanding a dataset is crucial to conducting a proper analysis and communicating accurate results.

My github page for all project files:
https://github.com/doug-cady/gmu_daen/tree/master/AIT664_InfoReprProcVisuals/project

References

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

Carr, D., & Pickle, L. (2010). *Visualizing data patterns with micromaps* (p. 24). London: Chapman & Hall.

Chang, W. (2021). 3.10 Making a Cleveland Dot Plot | R Graphics Cookbook, 2nd edition. Retrieved 21 November 2021, from https://r-graphics.org/recipe-bar-graph-dot-plot

Chang, W. (2021). 5.5 Dealing with Overplotting| R Graphics Cookbook, 2nd edition. Retrieved 21 November 2021, from https://r-graphics.org/recipe-scatter-overplot

Clark, Z. (2014). How to add whitespace to an RMarkdown document?. Retrieved 21 November 2021, from https://stackoverflow.com/questions/24425786/how-to-add-whitespace-to-an-rmarkdown-document

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

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/.

Saving grid.arrange() plot to file. (2015). Retrieved 21 November 2021, from https://stackoverflow.com/a/28136155

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.

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