Final Report

2019 Economic Freedom Index

Intro to Data Science, Spring 2021

Index

Background

Our data set comes from The Heritage Foundation, a group that has attempted to categorize the scale of economic freedom based on 12 different factors for a country. The process of creating these scales can be summarized by some more overarching categories: Rule of Law, Government Size, Regulatory Efficiency, and Open Markets. These categories are broken down into more sub-categories, 12 quantitative and qualitative factors in total (3 for each main category), that comprise the overall rating for each country in its respective section. Each of these are graded on a scale from 0 to 100, and the country's overall score comes from the average of the subcategories. All of the economic freedom components are weighed equally, there are no types of economic freedom classified as more important than a different one and all are equal in determining their world ranking.

Each observation represents one specific country. in the world This data set has 186 countries in total, covering all parts of the world. Our variables cover the advancement in economic freedom, prosperity, and opportunity and promote these ideas in their homes, schools, and communities. The Index covers 12 freedoms, from property rights to financial freedom, in these countries. Some categorical variables include the ones that name and identify the countries/observations, as well as regions. Quantitative variables included our freedom scores, as well as other miscellaneous variables like unemployment rate, population, and GDP.

We had a multitude of questions to ask of our data including how countries are ranked and what variables have the most influence on the 2019 Score variable. We also wanted to make a model that could predict a country's rank. We also wanted to see how countries' scores compared to their respective regions' scores as a whole.

To answer these questions, we had to do a good amount of preprocessing. There were a handful of countries that did not have enough raw data to even be given a rank, so these countries had to be removed from the data frame. In addition, some countries had some scattered NA values, so these had to be changed to zero. Lastly, we wanted to condense our data frame to only include our categorical variables and the 12 economic freedom variables. We also created some variables that measured the mean for these 12 economic freedom variables, in order to make better comparisons.

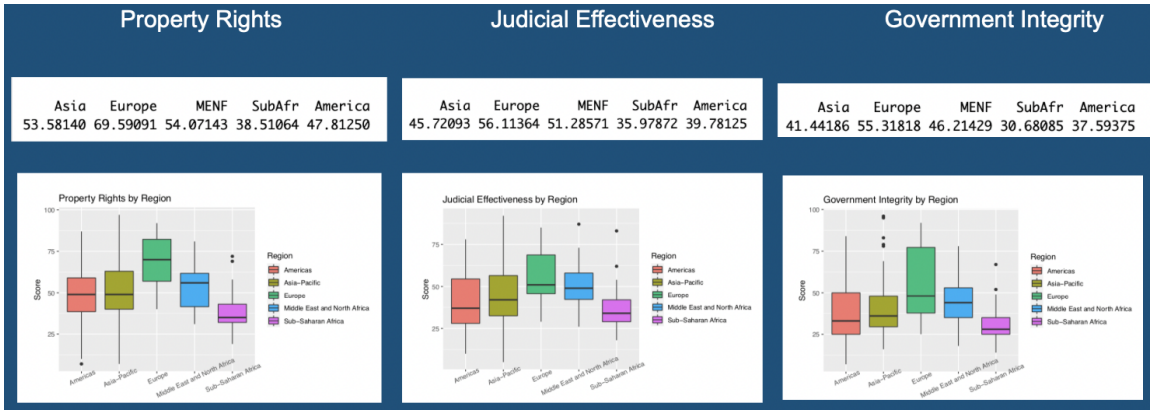
We used this new dataset to predict how countries' scores were calculated. We used a decision tree, clustering, and a linear regression model to do so. We split our data into a training and test set, and based our performance on bootstrap measures and R^2 values. We saw from very high R^2 values, that the linear regression model was the best.

Data Sources

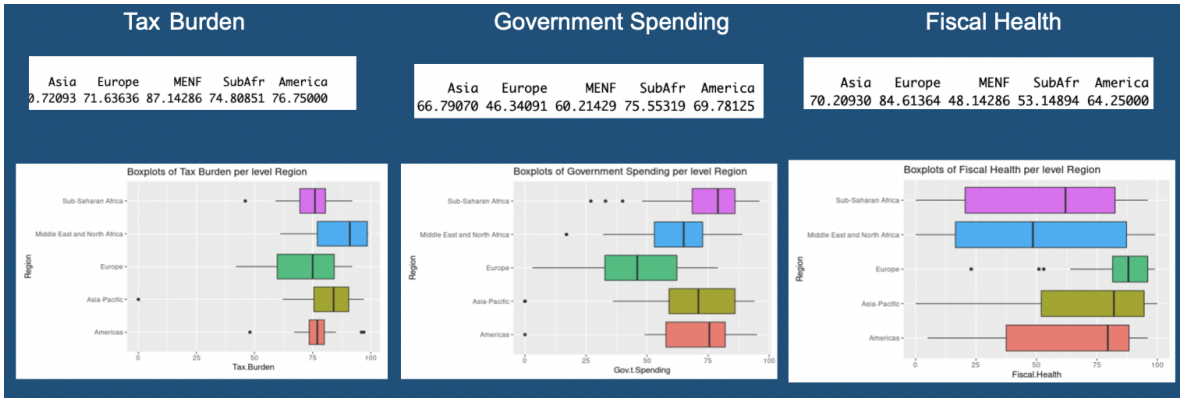
- https://www.kaggle.com/lewisduncan93/the-economic-freedom-index

---

Exploratory Data Analysis



Our first category is rule of law, where we see all three variables having pretty similar distributions, and a few outliers in government integrity. Each rule of law variable can be defined by the following, property rights: the theoretical and legal ownership of resources and how they can be used, government integrity: how reliable the government is on fulfilling its commitments and keeping its word as an agent to the public, judicial effectiveness: how fair are court decisions and are they implemented in a fair way that equally protects all citizens. By understanding what each variable means, this allows us to further develop and process the relationships. We can see that Europe had the highest score in all rule of law category variables and that Sub-Saharan Africa consistently has the lowest score in the same variables.



Next we have government size, where there is certainly more variation between the 3 variables. Tax burden is very scattered regionally, while government spending and fiscal health have opposite regional relations. These three variables shown in the visualizations above can be defined as- Government spending: spending by the public sector on goods and services such as education, health care and defence, Tax burden: a measure of the tax burden imposed by government, it includes direct taxes, in terms of the top marginal tax rates on individual and corporate incomes, and overall taxes, including all forms of direct and indirect taxation at all levels of government, as a percentage of GDP, and Fiscal health: is the ability of local governments to plan, manage, and pay for critical public services and investments; this ability grows more important every day as the world's cities confront rapid growth and local governments shoulder increasingly complex responsibilities. The medians are very similar between the different regions in the tax burden boxplot, with the average score for all regions being around 78. The government spending variable has more variance between the regions. There are 6 outliers and the regions have larger spreads. Europe has the lowest median score for government spending while Sub-Saharan Africa has the highest median score. Lastly, in fiscal health, the spreads are the largest in this variable for this category. Middle East and North Africa have a spread that is nearly 100 (actually 99), in addition to Europe having

all three outliers from this variable.

For regulatory efficiency, we see similar distributions for each variable. A common theme we have seen thus far is Europe dominating most of the variables, followed by Asia-Pacific and the Americas. It is pretty safe to say that European countries typically have better economic freedom. The three variables in the regulatory efficiency category can be defined as the following- business freedom: this is an overall indicator of the efficiency of government regulation of business, the business freedom score for each country is a number between 0 and 100, with 100 equaling the freest business environment, labor freedom: this is a quantitative measure that considers various aspects of the legal and regulatory framework of a country's labor market, including regulations concerning minimum wages, laws inhibiting layoffs, severance requirements, and measurable regulatory restraints on hiring and hours worked and monetary freedom: combines a measure of price stability with an assessment of price controls. Both inflation and price controls distort market activity, the score for the monetary freedom component is based on two factors: The weighted average inflation rate for the most recent three years and. By looking at the visualizations we generated above we can conclude that the average score was higher for tax burden than in the other two variables in this category. The spread was also less significant in the monetary freedom variable in addition to having 10 outliers. For business freedom, there were 6 outliers and the spread varied depending on the region with Asia-Pacific having the largest spread. Lastly, when analyzing the labor freedom variable, we can conclude that there are a total of three outliers, the average center score for all regions will be around 60.

In financial freedom, the Asia-Pacific region has the lowest median as well as the largest range. Europe

has the highest median of financial freedom, which means the government does not intervene much with the financial services of the market itself. Europe and the Middle East/North Africa also have the most outliers. The Americas and Sub-Saharan Africa have similar spreads, ranging from a score of 10 in financial freedom to 79. Sub-Saharan Africa has the lowest median for trade freedom. This means that this region has the most restricting policies and tariffs on imports and exports. Europe has the highest median score in trade freedom, which means that this region has the most relaxed practices around imports and exports. In addition, compared to the medians' relations to each other in the Financial Freedom box plot, the medians in this box plot are much closer to one another. This could mean two things. One, trade freedom is not a big indicator of a high freedom index score or two, trade freedom is rather similar among all of the regions therefore it is not very significant due to similar values. The Asia-Pacific region has the lowest median score again in the box plot analysis. This makes sense when looking at our market openness plot because Asia-Pacific has the second-lowest Market Openness score. The Americas has the most outliers (ranging from 0 to 14), with a median of 70. Europe has the highest median score in Investment freedom with a value of around 77. We can infer that Europe has the most relaxed policies around investing in and out of the country and in different sectors. Even though Asia-Pacific has the lowest median score, they do have the largest spread- ranging from 0 to 83. One interesting observation to make is that there are at least three countries that have a score of zero as their investment freedom score

---

Model Evaluation and Validation

Linear Regression

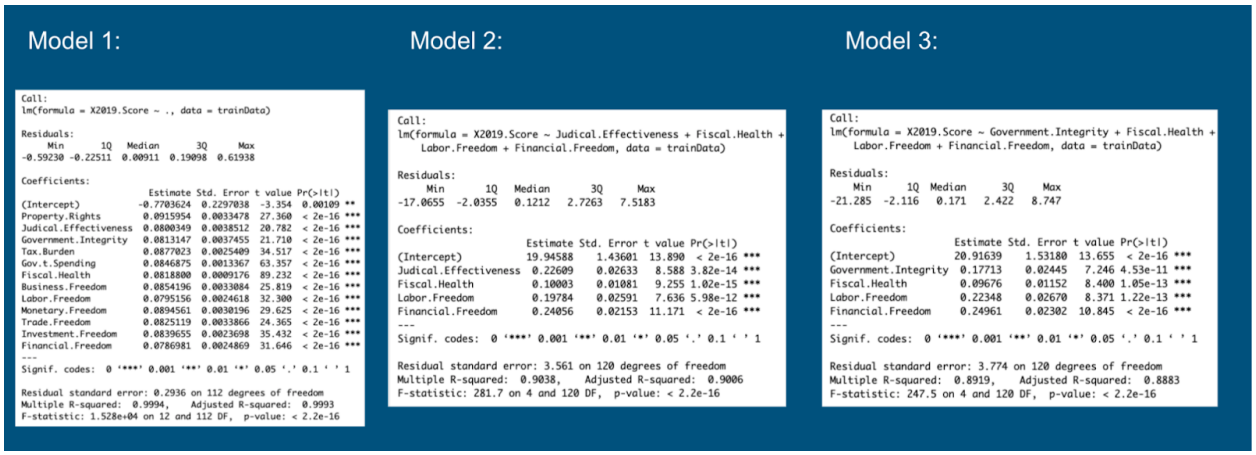
The first thing we did was narrow our data down to only the variables that we wanted to look at. These variables are each of the ones that define economic freedom.

```
```{r}
freedomModelDat <- freedom[,7:19]
#freedomModelDat
```
```

Next we created a 70/30 split of our data to separate it into training, and testing data.

```
```{r}
set.seed(12)
trainSel = sample(1:nrow(freedomModelDat), nrow(freedomModelDat) * .7)
trainData = freedomModelDat[trainSel, ]
testData = freedomModelDat[-trainSel, ]
```
```

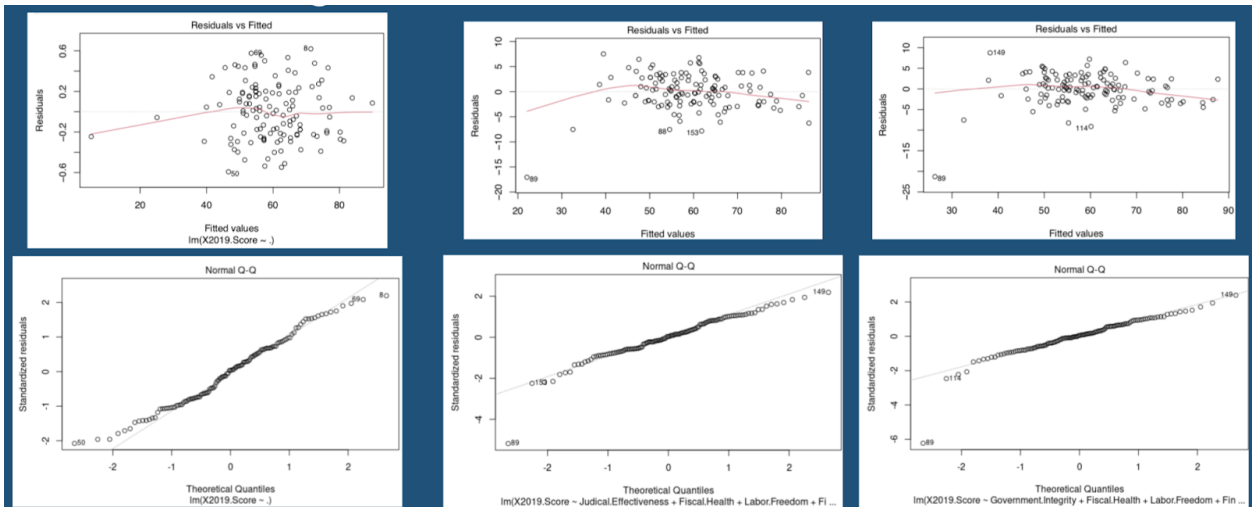
After the data was split we created three different linear regression models, each time trying to improve upon the last one through a different selection of variables.

```r
out = lm(X2019.Score ~ ., data = trainData)
summary(out)
plot(out, which= c(1:2))

#Model with variables chosen based off first model and picking the variables for each category (Rule of Law, Regulatory Efficiency,
Government Size, and Market Openness) that have the lowest, closest to zero.
out2 = lm(X2019.Score ~ Judical.Effectiveness + Fiscal.Health + Labor.Freedom + Financial.Freedom, data = trainData)
summary(out2)
plot(out2, which= c(1:2))

#Model with variables chosen based off our intuition of which models were most important and impactful.
out3 = lm(X2019.Score ~ Government.Integrity + Fiscal.Health + Labor.Freedom + Financial.Freedom, data = trainData)
summary(out3)
plot(out3, which= c(1:2))

```

These are the results we received from each linear regression model:



And here is the plots (residuals and Q-Q plots) we created from each of the models:



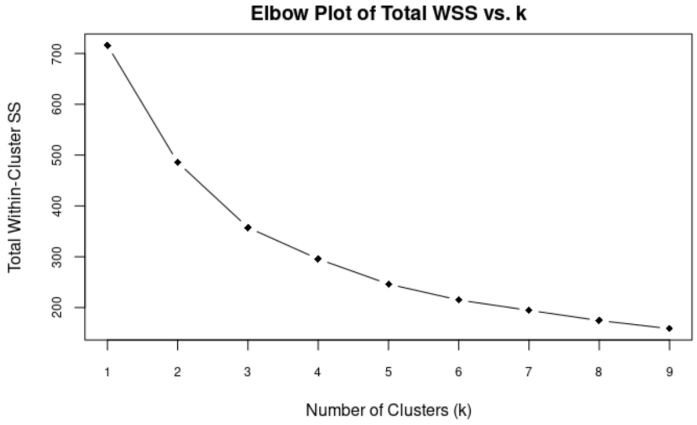
Model Interpretations:
Based on the original model, we were able to narrow down which variables seemed to have more impact, and we ended up selecting one variable from each category that was closest to zero for its estimate and created a new model with those four variables. As you can see the information in this model is a lot more meaningful and allows us to get a better understanding of how those four variables play a role in

estimating X2019.Score. The variables we ended up choosing were Judicial Effectiveness, Fiscal Health, Labor Freedom, and Financial Freedom

Plot Interpretations:
The first two plots are from the first model we created with all of the variables. We can see that its Q-Q plot does seem to be relatively linear with respect to the outliers on both ends. However, its residual plot does have quite a large spread and is not very condensed which leads to a bad best line of fit. On the other hand, the Q-Q plot for the model with only four variables is really good. A majority of the data seems to be pretty linear again with some exceptions for outliers but overall it seems to indicate a better model. Its residual plot also is better than the last. Although it still has some variability it is a lot more condensed compared to the original model. The last model we created seemed to be the best, and this was the one that we chose the variables that we concluded to be the most impactful. And while it did not change too much from the second model it seemed to be a better model based on the residuals and Q-Q graphs. It's residual graph has a much better line of fit and is closer to being straight. And the Q-Q plot looks to be even more on line and have fewer outliers compared to the previous models.

Clustering

The first thing we did for our clustering model was scale the variables that we wanted to use (which were the variables we used for our third linear regression model). Then we created an elbow plot to determine which value to use for k.




```r
```{r}
set.seed(12)
freedom.sc = scale(freedom[, c("Government.Integrity", "Fiscal.Health", "Labor.Freedom", "Financial.Freedom")])

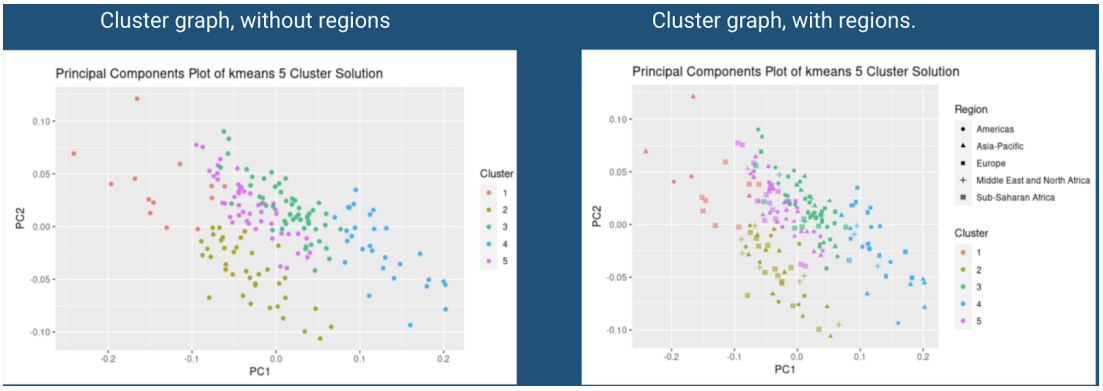

km1 = kmeans(freedom.sc, centers = 1, nstart = 50)
km2 = kmeans(freedom.sc, centers = 2, nstart = 50)
km3 = kmeans(freedom.sc, centers = 3, nstart = 50)
km4 = kmeans(freedom.sc, centers = 4, nstart = 50)
km5 = kmeans(freedom.sc, centers = 5, nstart = 50)
km6 = kmeans(freedom.sc, centers = 6, nstart = 50)
km7 = kmeans(freedom.sc, centers = 7, nstart = 50)
km8 = kmeans(freedom.sc, centers = 8, nstart = 50)
km9 = kmeans(freedom.sc, centers = 9, nstart = 50)


WSS = NULL
for(k in 1:9){
  WSS[k] = kmeans(freedom.sc, centers = k, nstart = 50)$tot.withinss
}


plot(1:9, WSS, type = "b", pch = 18, xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster SS", main = "Elbow Plot of Total WSS vs. k",
     xaxt = "n", cex.axis = .75)

axis(1, 1:9, 1:9, cex.axis = .75)
```

After deciding to use five for our k value we then created the cluster model, both with and without each point showing which region it is in.


```

To then assess the accuracy of our clustering model we then looked at the bootstrap means for each cluster.

```{r}
library(fpc)
set.seed(12)
table(km5$cluster, freedom$Region)

cboot.kmeans = clusterboot(freedom.sc, clustermethod = kmeansCBI, k = 5, count = FALSE)
cboot.kmeans
print("------------------")
cboot.kmeans$bootmean
```

```
  Americas Asia-Pacific Europe Middle East and North Africa Sub-Saharan Africa
1        2           3      0                           0                 8
2        8           9      1                           7                13
3       12           7     23                           3                 8
4        3           6     16                           3                 1
5        7          18      4                           1                17
* Cluster stability assessment *
Cluster method:  kmeans
Full clustering results are given as parameter result
of the clusterboot object, which also provides further statistics
of the resampling results.
Number of resampling runs:  100

Number of clusters found in data:  5

 Clusterwise Jaccard bootstrap (omitting multiple points) mean:
[1] 0.7618570 0.9176265 0.7951892 0.7242621 0.5749265
dissolved:
[1] 14  2 13 11 52
recovered:
[1] 57 91 69 39 26
[1] "------------------"
[1] 0.7618570 0.9176265 0.7951892 0.7242621 0.5749265
```

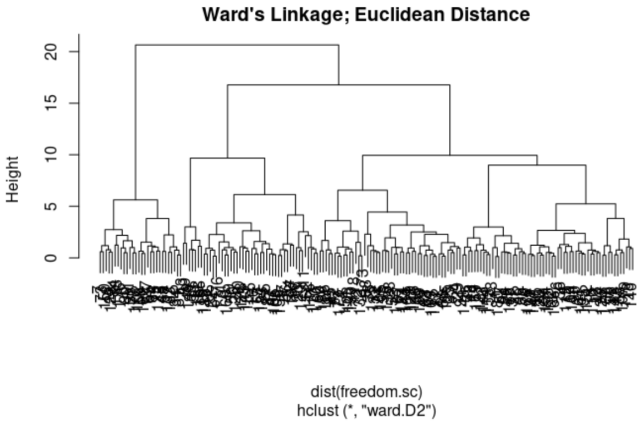
The results from the bootstrap means indicate that we have highly stable clusters. Although they do not classify each cluster for an individual region the prediction of which countries are in each cluster does seem to be accurate. And we can see from the plots that our model does not predict what we expected to see accurately, but this could be because of the variables that we chose, or the fact that not every country will align in its scoring with other countries in its region.

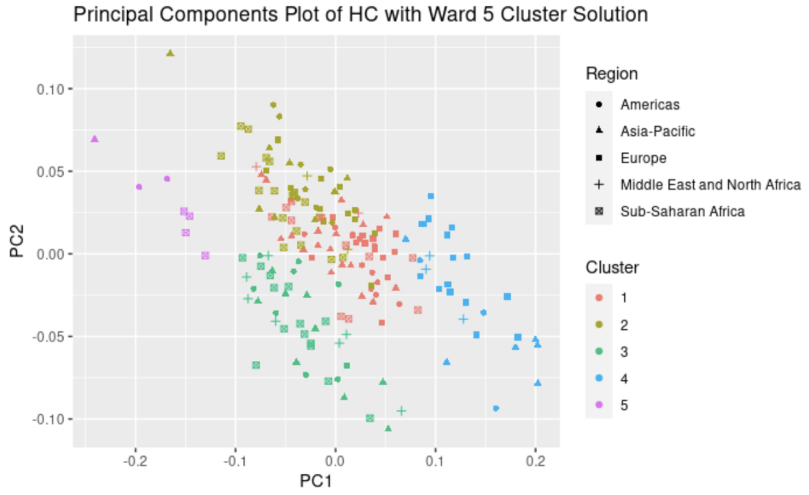
Linkage Model

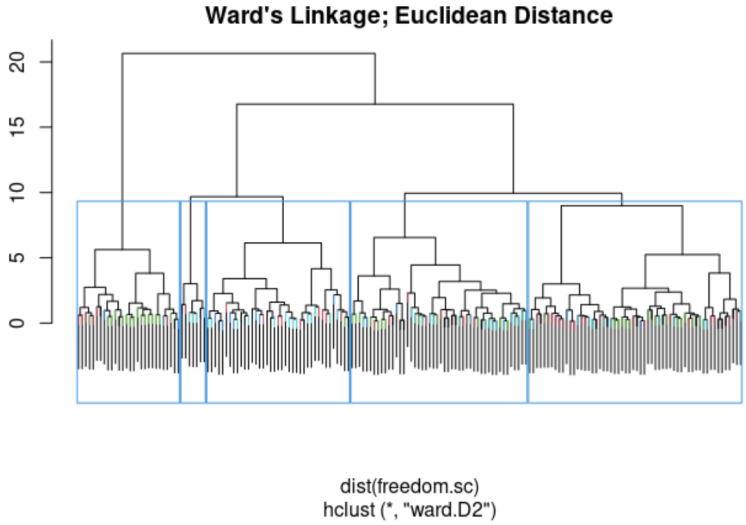
We again used the same variables that we used in the previous model as well as our third linear regression model. We then chose to use the Wards Linkage Method to illustrate our model.

```{r}
library(palmerpenguins)
library(sparcl)
hc.ward = hclust(dist(freedom.sc), method = "ward.D2")
plot(hc.ward, main="Ward's Linkage; Euclidean Distance")


tab = table(cutree(hc.ward, 5), freedom$Region)
tab

cboot.hclust = clusterboot(freedom.sc, clustermethod = hclustCBI, method = "ward.D2", k = 5, count = FALSE)
cboot.hclust$bootmean

princ = prcomp(freedom.sc)
nComp = 2
project = predict(princ, newdata = freedom.sc)[,1:nComp]
project.plus = cbind(as.data.frame(project), Cluster = as.factor(cutree(hc.ward, 5)), Region = freedom$Region)
ggplot(project.plus, aes(x = PC1, y = PC2)) +
  geom_point(aes(shape = Region, color = Cluster)) +
  ggtitle("Principal Components Plot of HC with Ward 5 Cluster Solution")

ColorDendrogram(hc.ward, y = factor(freedom$Region), main = "Ward's Linkage; Euclidean Distance")
rect.hclust(hc.ward, 5, border = 4)
```

The block of code above depicts all of our work done for the linkage model. In addition to creating the plot of the linkage model we also created another cluster graph to depict the distribution of points so we could compare it to our previous one. We again decided to use five clusters in our linkage model so we see the differences between the two different models. Here are the results of our linkage model:







We can see that again this did not accurately represent each cluster into an individual region. However, according to our bootstrap means most of the clusters were still almost all stable. In addition we can also notice that this model does not perform as well as our previous clustering model does, however, it still does have relatively high accuracy.

```
  Americas Asia-Pacific Europe Middle East and North Africa Sub-Saharan Africa
1       11           19     15                            2                 11
2        8            8     12                            2                 18
3        8            9      1                            7                 14
4        3            6     16                            3                  0
5        2            1      0                            0                  4
[1] 0.4891465 0.5247137 0.8399420 0.9068513 0.5996814
```

Conclusions and Results

Using the linear regression model, we found high R squared values that informed us of a strong relationship between the variables. The values we found for our three models were 0.9944, 0.9038, and 0.8191. The model with the highest R squared value was when we used all of the variables, the second model consisted of the four variables with the closest to zero estimate, and the last was made up of handpicked variables. While the first model had the highest R squared value, it also had the most outliers. The model with the least amount of outliers was the third. We were able to determine that the regression model fit our observations well. With the clustering models, we found that there is a high stability throughout all of the clusters. The fifth cluster had the lowest stability, but still had a bootstrap mean of above 0.5. When observing the clustering graph with regions, one can see that all of the clusters had representation of all the regions within them. This answers the research question of "Does each country in a certain region behave the same as the region as a whole," with a no. Overall, the linear regression model was the more successful of the two, with high R squared values, f-statistics, and a very low p-value. However, the clustering model visually helped us more in determining some of the information needed to answer the research questions. A limitation that our group faced in analyzing this dataset was that we replaced NA values with 0s, which could have led to skewed results.

Challenges

Throughout the process of cleaning and generating the dataset and model we experienced some challenges. One of the logical challenges we faced in our dataset revolved around time itself. Our data measures and is composed of economic data; due to common change of government, which leads to change in economic policy and well being, our observations were easily impacted due to the time variable. Another challenge we faced was eliminating effects of all outside variables. Weather is a variable that could have impacted countries' GDP and expenditure but we were not able to measure for the impact. Another challenge we faced was the effect of wars and disagreements between countries. War and conflict could easily slow down trade, financial freedom and many other variables we analyzed.

What We Would Do to Improve Our Model

Our linear regression model was very accurate in predicting the 2019 score variable. Using all 12 economic freedom variables proved to be very effective in doing so.Ways to improve our model could include using some of the other miscellaneous variables like unemployment rate, population, etc. We could cycle through these variables, using them in our model accompanying our economic freedom variables, seeing if we can predict the 2019 score any better. Though our model was overall successful, it didn't accurately predict each cluster to one individual region. We could improve on this issue by finding a different way to create a model that accurately predicts what region a country would fall into. Overall,

we want to improve by taking our model and analysis another step forward and fully answering our main research question and completing our goal.

Works Cited

The Heritage Foundation. 2019. *The Economic Freedom Index* [Data set].

https://www.kaggle.com/lewisduncan93/the-economic-freedom-index

---

Data Preprocessing

Our ideal data frame consists of only the countries with a valid rank and with full data input for each variable. Some examples of countries we removed due to lack of sufficient data include Somalia and Yemen. Both of these countries are examples of data we removed during preprocessing. In addition, during data preprocessing, we simplified the variables we used and chose to focus on 12 variables that represented the 12 economic freedoms. These variables were the most significant due to our team being able to clearly define what each meant as well as the economic freedom index and our sources focusing on these variables. For our team to be able to simplify our data to what we wanted, we used data selection techniques. The data selection process we used selected entire columns of variables we were interested in in addition to removing rows of countries that we did not want to be included in the final ranking due to missing data. The last data preprocessing step we took was changing any remaining 'NA' values to 0.

Code for Functions

```r
rmse <- function(y, yhat){
 return(sqrt(mean((y - yhat)^2)))
}

rmse(y=trainData$X2019.Score, yhat=preds.train)

rsquared <- function(y, yhat){
  return(1-sum((y-yhat)^2)/sum((y-mean(y))^2))
}

rsquared(y=trainData$X2019.Score, yhat=preds.train)
```

Data Files

economic_freedom_index2019_data.csv

Glossary

F-statistic	the value that is the output when a function is run with the goal to find the mean for two chosen populations
GDP	Gross Domestic Product
MENF	Middle East and North Africa region- this term is seen being used in our outputs and visualizations
NA	in the coding language we use, 'NA' means that there was no value available for the desired
P-value	the value that represents the probability that a statistical measure of a set probability distribution will be greater than or equal to the observed results
R squared value	a statistical measure that represents the distance of the data point to the line of best fit.