Freedom and Prosperity Research Project

Jonathan Kang, Muhan Zhang, Henry Yin, Matthew Feinberg

December 8th, 2022

Introduction

Do countries need freedom to achieve prosperity? Are there any relationships between a country's wealth and the level of freedom a country enables for its citizens? As we are seeing massive spikes of gentrification negatively impacting the economy worldwide, we are interested in investigating the scores provided by the Atlantic Council and the underlying relationships behind the categories and factors.[1] With our team consisting of members from different majors, we aim to analyze the socio-economic situation of different countries. This is done by investigating the relationship between freedom and prosperity, then developing regression models to extract the most significant variables in determining a country's level of freedom and prosperity.
The purpose of this is to understand the most significant factors of freedom in affecting a country's level of prosperity and what may be the most impactful factors in measuring a country's economic situation and national prosperity.

Our project is based around analysis of the Freedom and Prosperity dataset from the Atlantic Council Freedom and Prosperity Center.
In this project, we will bgein by exploring freedom and prosperity scores and their categorical scores. Next, we will investigate and generate multiple types of learning models to generate one with the best prediction accuracy possible. The correlations and statistical learning models will help provide an insight into the most significant variables we should extract and examine from the raw data from World Bank.

Exploratory Analysis - Descriptive Statistics

Dataset Description

The source of our dataset comes from the website: "https://www.atlanticcouncil.org/in-depth-research-reports/report/do-countries-need-freedom-to-achieve-prosperity/ (https://www.atlanticcouncil.org/in-depth-research-reports/report/do-countries-need-freedom-to-achieve-prosperity/)". [1]

This dataset has detailed data on 174 countries of the world, split into 6 geographical regions. The regions, along with their abbreviations, are listed below. We will be using these abbreviations throughout our project reports.

- America: **AME** (32 countries)
- Asia Pacific: **AP** (28 countries)
- Europe and Central Asia: **ECA** (18 countries)
- Middle Eastern Asia: **MENA** (18 countries)
- Sub Saharan Africa: **SSA** (47 countries)
- Western Europe: **WE/EU** (31 countries)

All data points in our dataset were recorded multiple times over a 15 year period - during 2006, 2011, 2016, and 2021. This initial summary will cover the data points from 2021.

The data covers freedom and prosperity, with detailed categories for each to give further insight into specifics of each country's freedom and prosperity conditions. The freedom data is split into three categories:

- **Economic Freedom** (4 additional subcategories)
- **Political Freedom** (3 additional subcategories)
- **Legal Freedom** (10 additional subcategories)

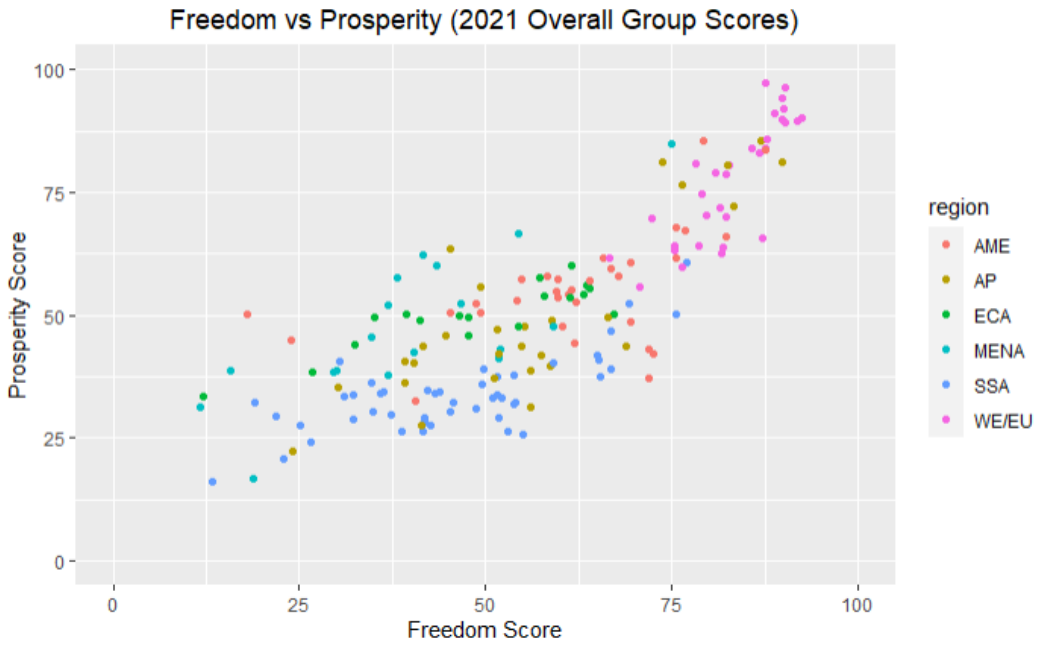
The prosperity data is split into five categories:

- **Income**
- **Environment**
- **Minority Rights**
- **Health**
- **Happiness**

Values for the overall categories were calculated by taking the average of all individual subcategory values.

Initial Analysis

We start by analyzing the overall scores for freedom and prosperity. Here is a graph displaying reedom and prosperity scores for each country, colored by region. This coloring will stay consistent throughout the initial descriptive data analysis.



Both freedom and prosperity scores are scaled to values between 0 and 100 inclusive. Here are some basic statistics for both scores (rounded to two decimal places):

Freedom Score

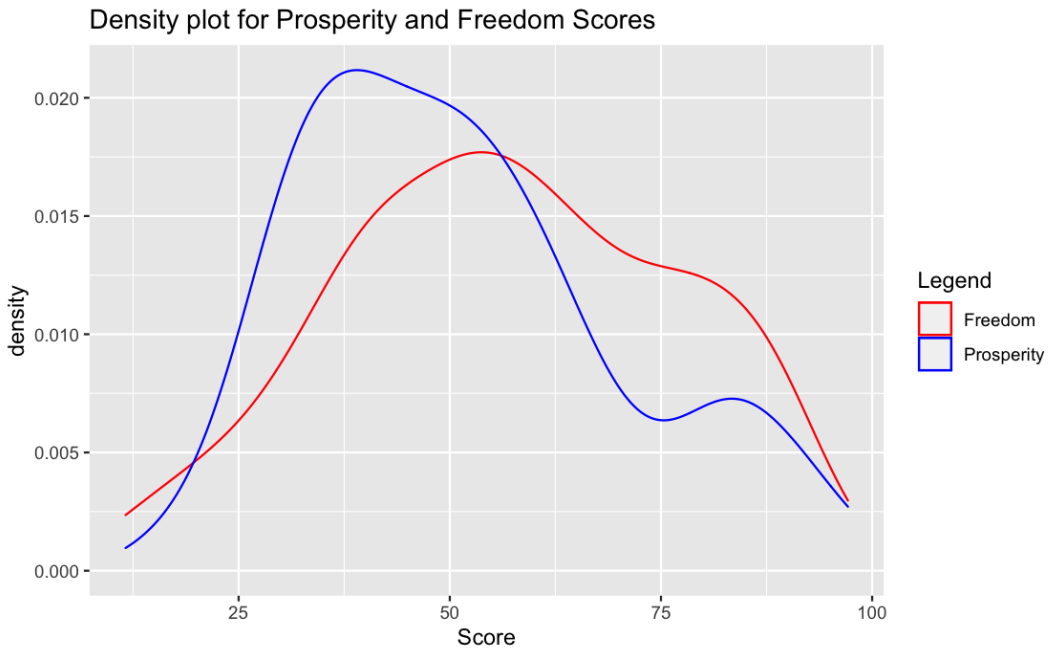
- Mean: 56.69
- Standard Deviaton: 19.94
- Minimum: 11.61
- Maximum: 92.45

Prosperity Score

- Mean: 51.24

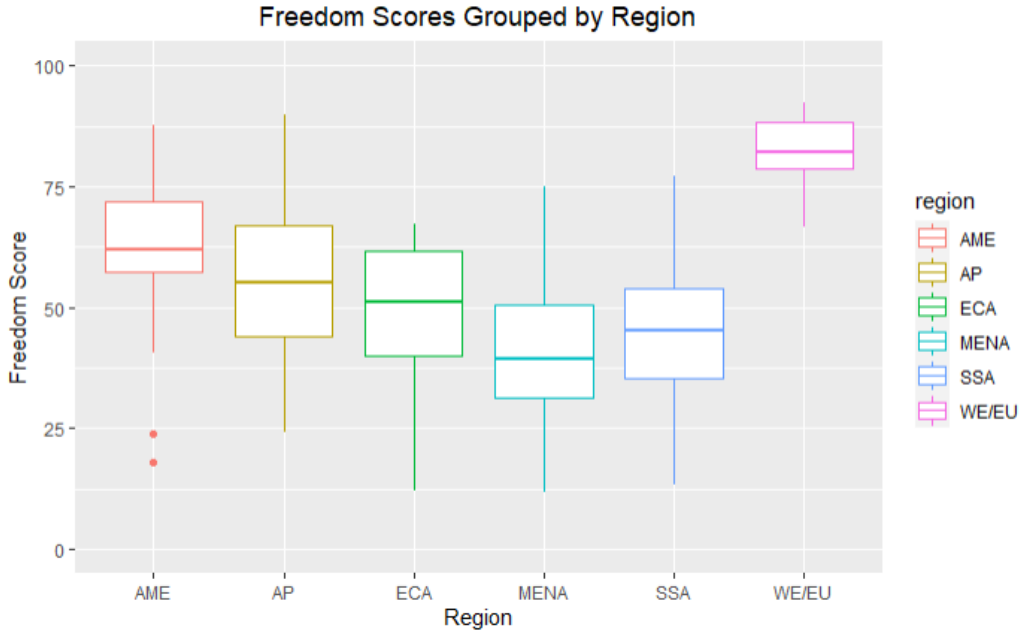
- Standard Deviation: 18.77
- Minimum: 15.96
- Maximum: 97.13

The following plot shows the distribution of the points:



Freedom Score Analysis

We can analyze the freedom scores for each specific region. We can use a box plot to visually see the freedom scores grouped by region.

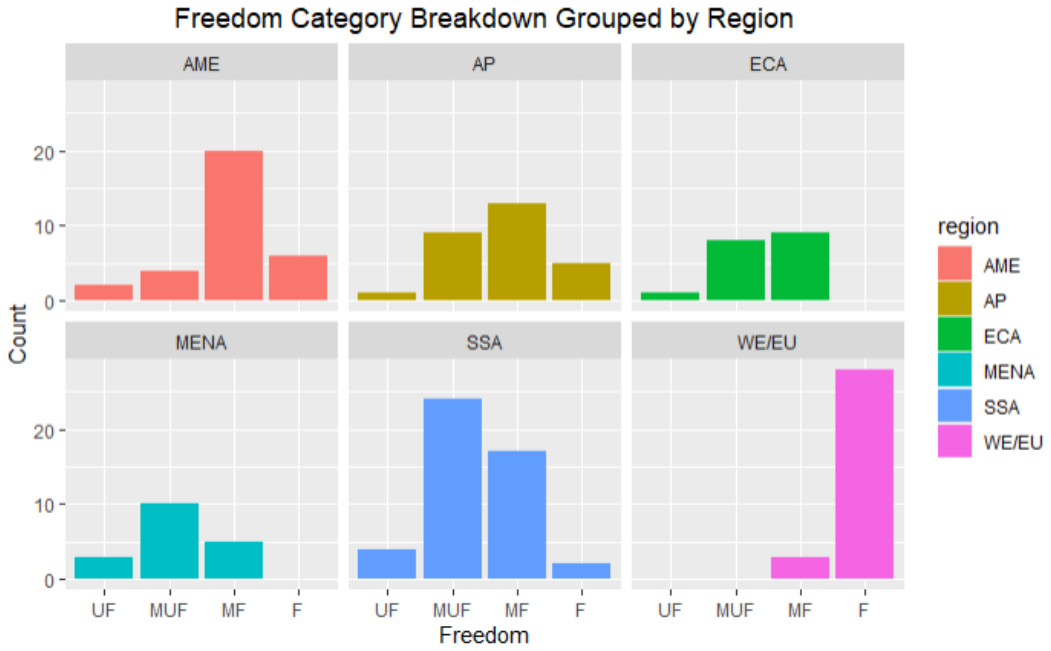


Our dataset has an additional categorical variable for freedom scores. The categorical variable is assigned as follows:

- **Unfree (UF)**: freedom score from 0 to 25 (11 countries)
- **Mostly Unfree (MUF)**: freedom score from 25 to 50 (55 countries)
- **Mostly Free (MF)**: freedom score from 50 to 75 (67 countries)

- **Free (F)**: freedom score from 75 to 100 (41 countries)

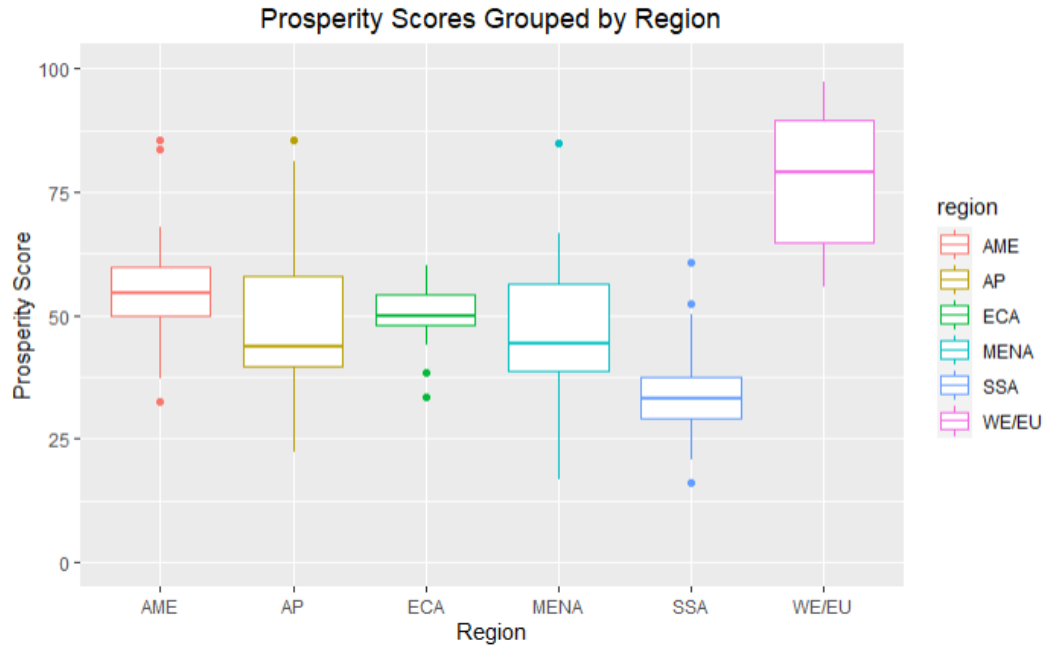
We can use this categorical variable to generate an additional visualization for country freedom. These bar plots are shown below.



From these two plots, we can see that the average freedom score is much higher for Western European countries compared to every other region. There are also 28 Western European countries that are Free out of a total of 41 Free countries.

Prosperity Score Analysis

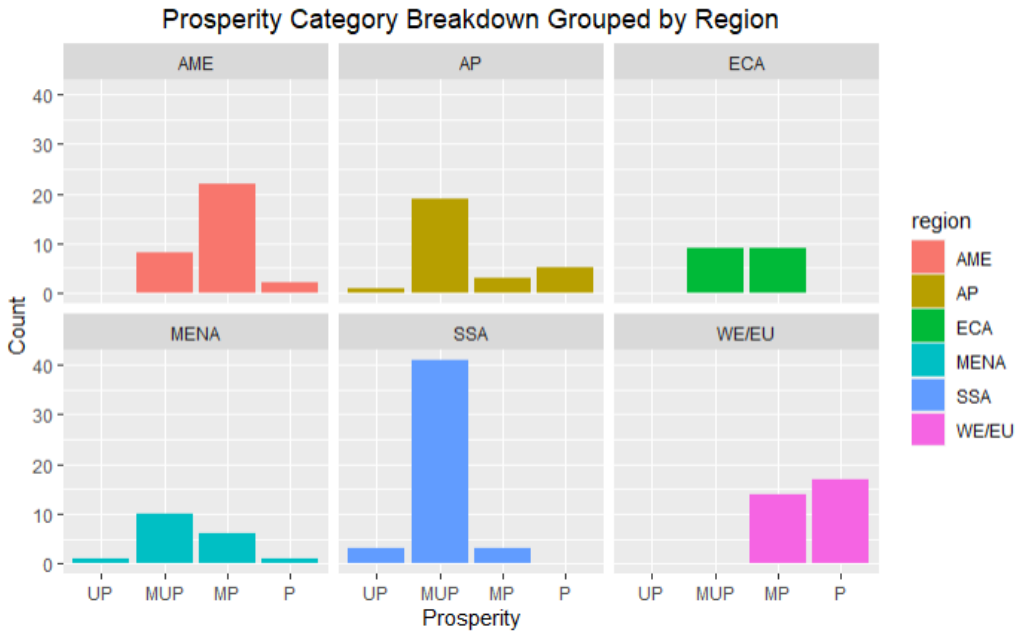
Similarly to what we have just done for the freedom scores, we can analyze the prosperity scores by region. We can use a box plot to visually see the prosperity scores grouped by region.



Similarly to the freedom scores, our dataset has an additional categorical variable for prosperity scores. The categorical variable is assigned as follows:

- **Unprosperous (UP)**: prosperity score from 0 to 25 (5 countries)
- **Mostly Unprosperous (MUP)**: prosperity score from 25 to 50 (87 countries)
- **Mostly Prosperous (MP)**: prosperity score from 50 to 75 (57 countries)
- **Prosperous (P)**: prosperity score from 75 to 100 (25 countries)

We can use this categorical variable to generate an additional visualization for country prosperity. These bar plots are shown below.

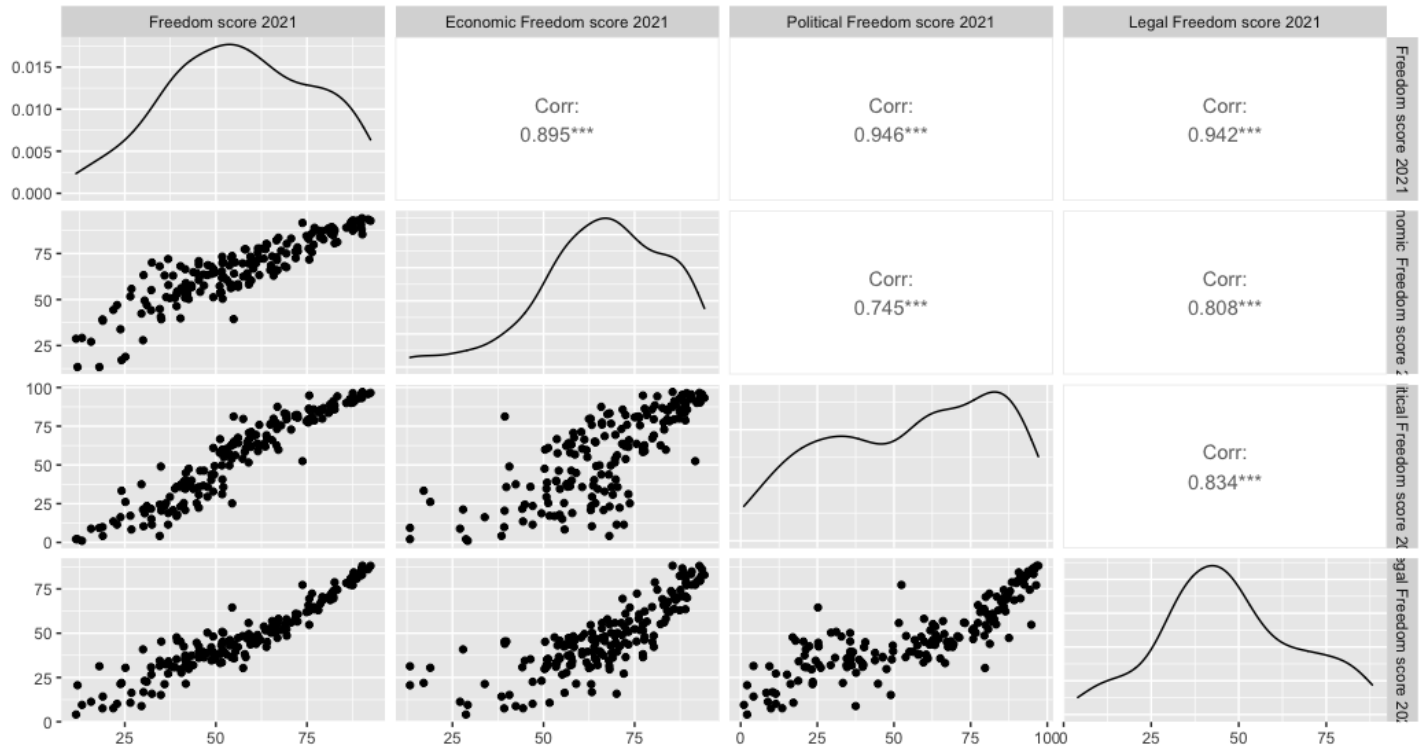


Once again, Western European countries have a higher average prosperity score than the other regions. Western European countries also account for 17 of the 25 Prosperous countries.

Exploratory Analysis - Correlations

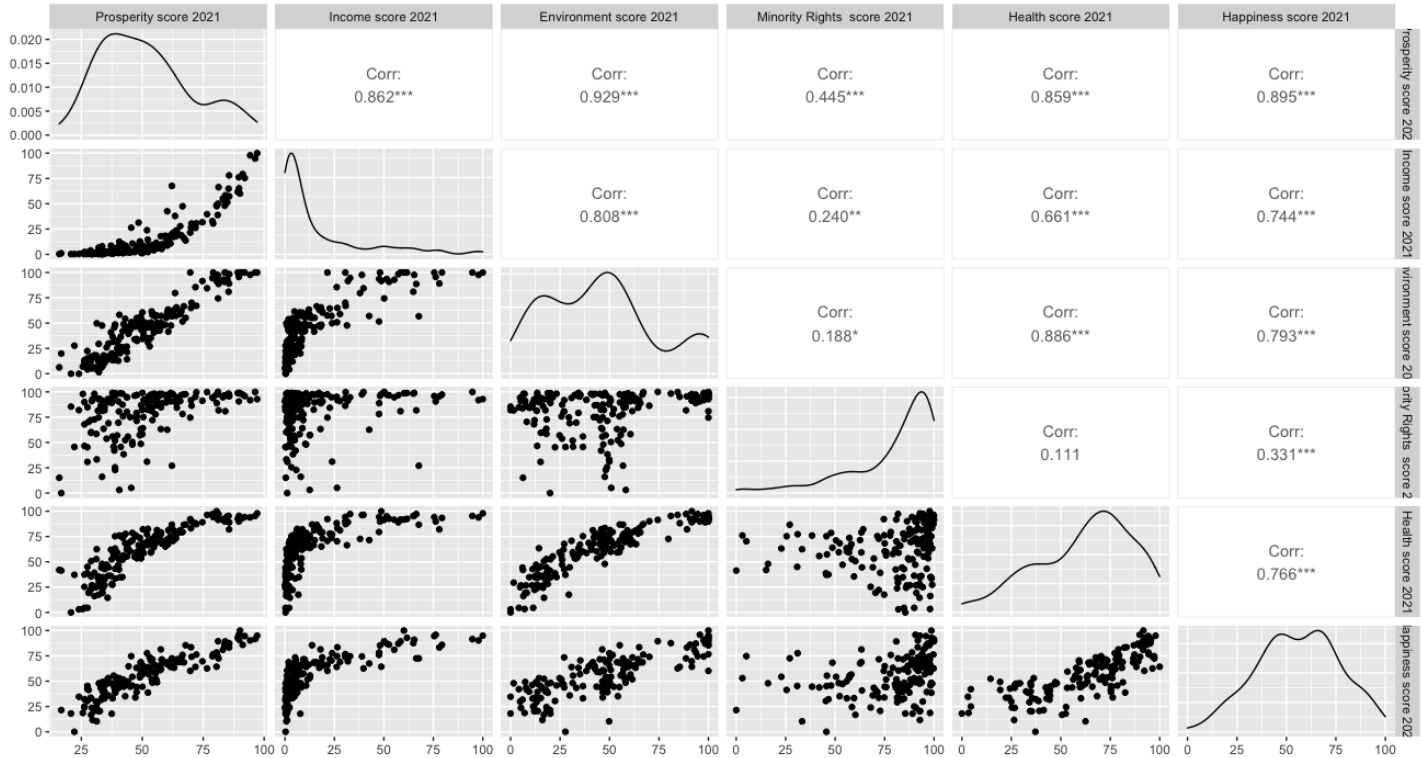
We will investigate the correlations between the main categories of Freedom and Prosperity against their subcategories. Next, we will dive into the specific factors of Freedom and obtain the factors that have are highly correlated with the subcategories of Freedom.

Freedom and its Main Categories



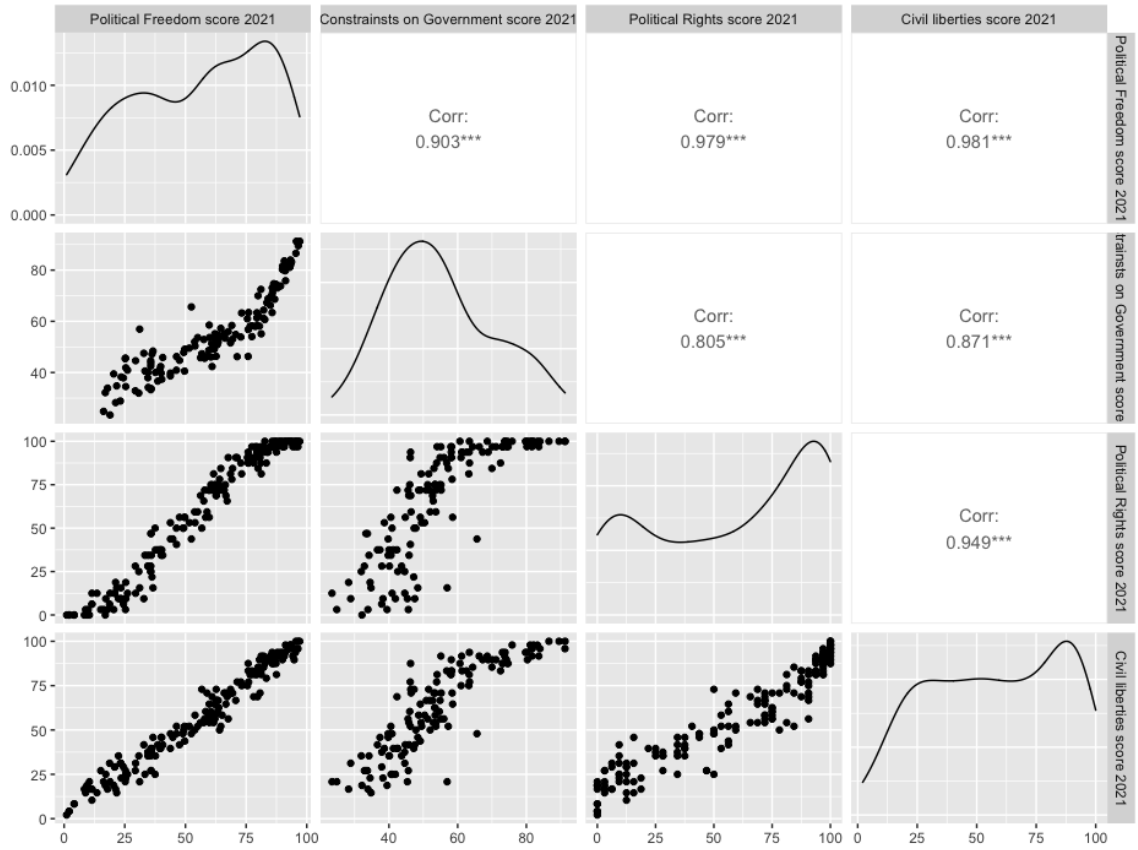
Freedom Correlations

Prosperity and its Main Cateogries



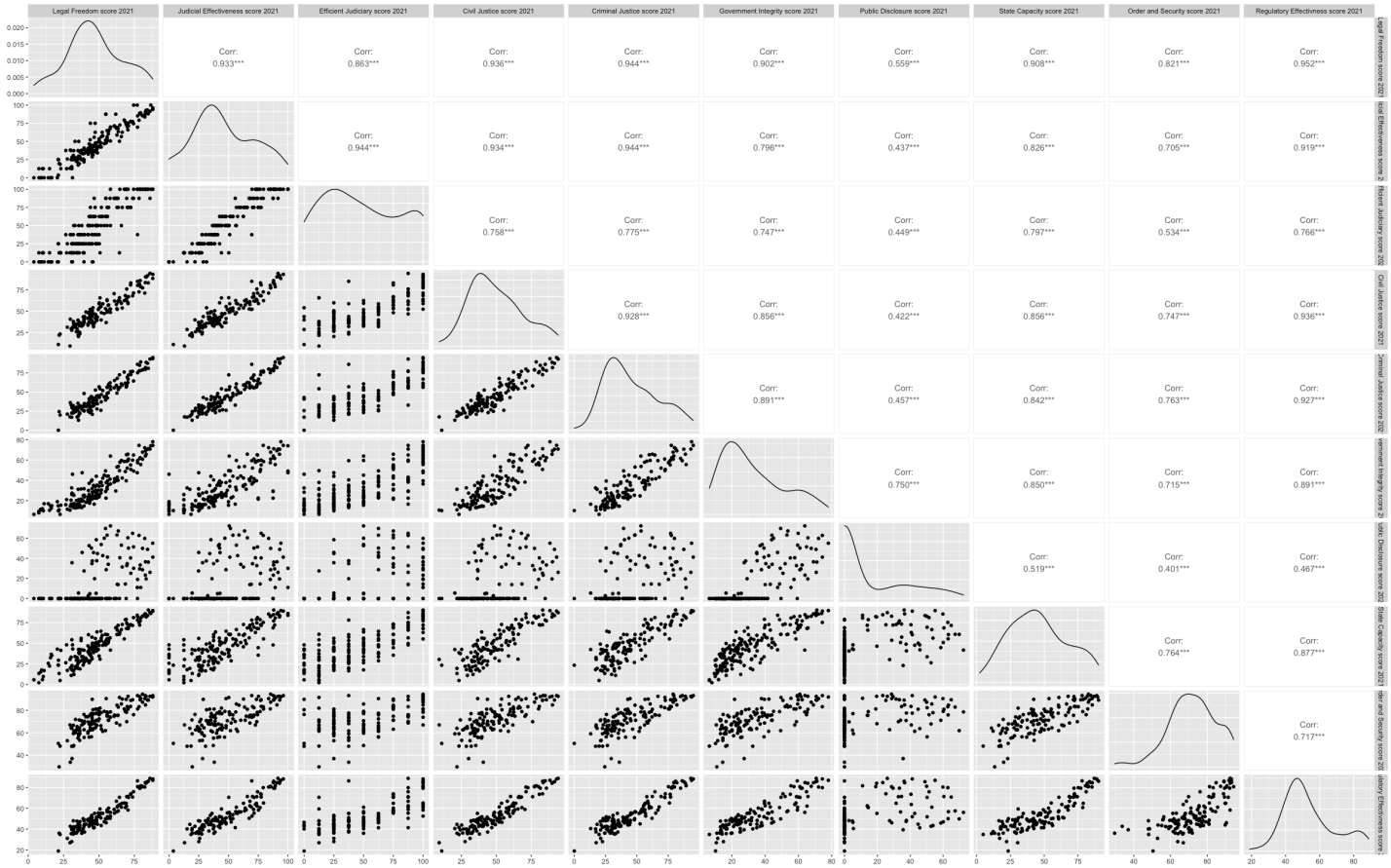
Prosperity Correlations

Freedom Subcategory (1)



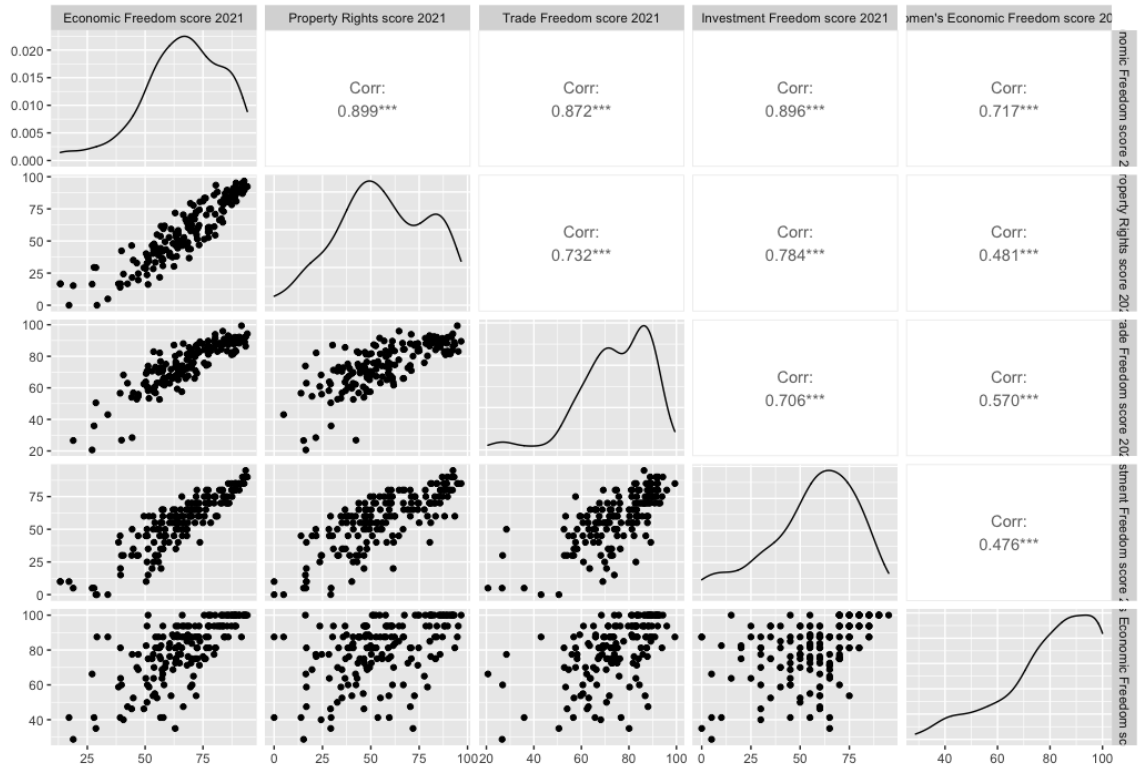
Political Freedom

Freedom Subcategory (2)



Legal Freedom

Freedom Subcategory (3)



Economic Freedom

Statistical Learning Methods

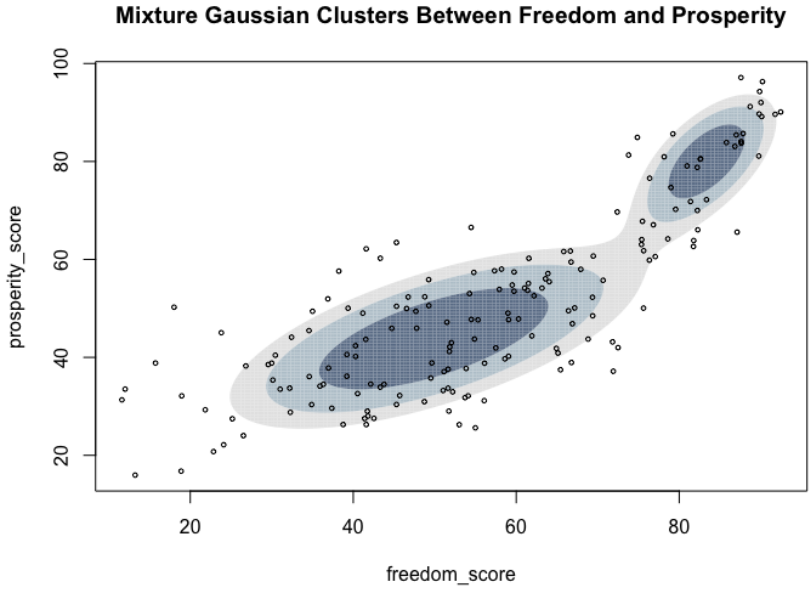
For the following slides, we will be looking at various statistical models that will help with this data analysis project.

Linear Regression - General Scores

Linear Regression

From this plot we can infer that there seems to be an upward curve resembling an exponential relationship. As there seems to be two large clusters, we will attempt to cluster the groups using Gaussian Mixture Model method.

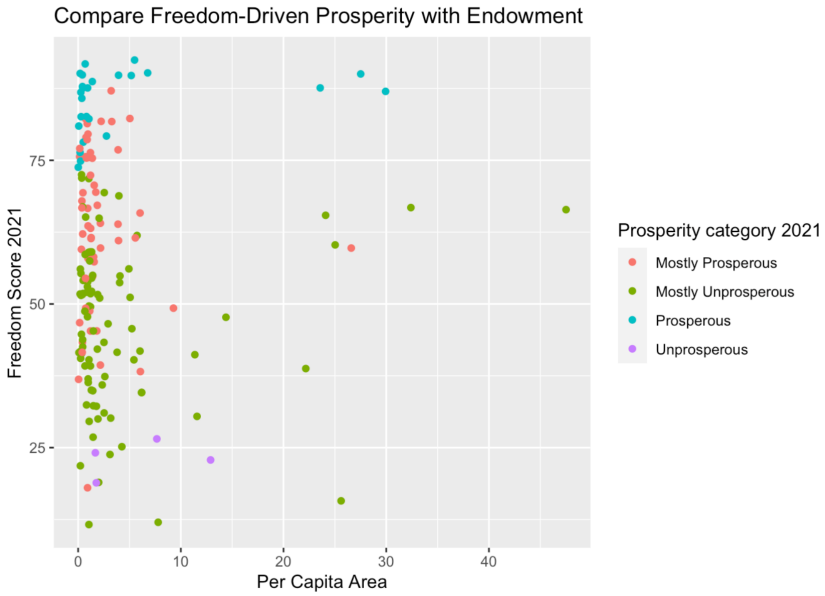
Gaussian Mixture Model Clustering

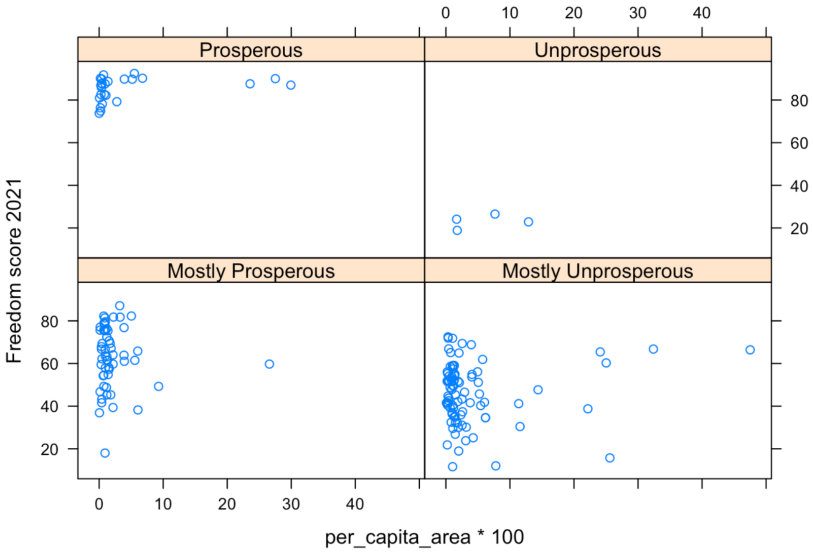


Gaussian Mixture Model

We can see from the clusters above that there are two clustered groups between the free and prosporous countries against the unfree and unprosperous countries.

Regression Model With Interaction Term





```
## Start:  AIC=834.25
## `Prosperity score 2021` ~ `Freedom score 2021` * I(per_capita_area *
##     100)
##
##                                                Df Sum of Sq   RSS    AIC
## - `Freedom score 2021`:I(per_capita_area * 100)  1    7.8334 20986 832.31
## <none>                                                      20979 834.25
##
## Step:  AIC=832.31
## `Prosperity score 2021` ~ `Freedom score 2021` + I(per_capita_area *
##     100)
##
##                             Df Sum of Sq   RSS     AIC
## - I(per_capita_area * 100)  1         7 20994  830.37
## <none>                                  20987  832.31
## - `Freedom score 2021`      1     38300 59287 1008.94
##
## Step:  AIC=830.37
## `Prosperity score 2021` ~ `Freedom score 2021`
##
##                         Df Sum of Sq   RSS     AIC
## <none>                               20994  830.37
## - `Freedom score 2021`  1     38294 59288 1006.94
```

```
##
## Call:
## lm(formula = `Prosperity score 2021` ~ `Freedom score 2021`,
##     data = ready_data)
##
## Coefficients:
##          (Intercept)  `Freedom score 2021`
##               8.1319                0.7606
```

The plots show endowment has a little to do with prosperity. The statistics also echo this conclusion as only freedom score was kept in the final model using an AIC based variable selection process.

Machine Learning Models for porsperity score prediction

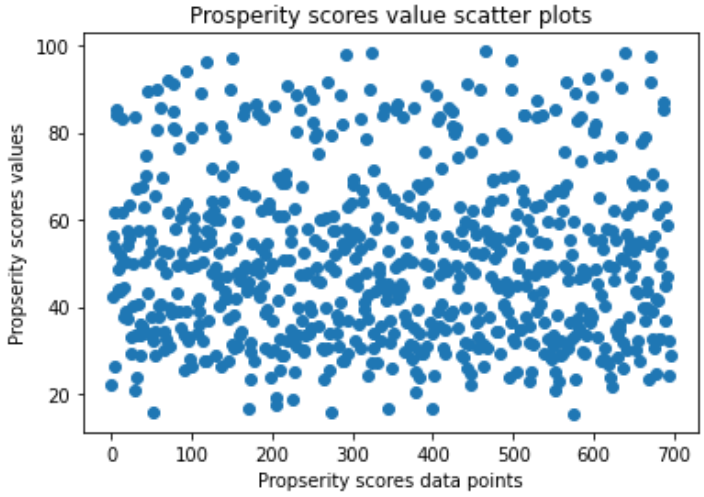
We use several supervised learning methods to build prediction models for prosperity scores based on all the freedom subcategories, and compare their performances.

Data Processing and prediction settings

Our data source is from the Atlantic Council. For every country beginning from 2006 and in every 5 years, the Atlantic Council give it a score for every freedom subcategories and prosperity. Using these scores, we form the training and testing data set. Since the algorithmic models we intend to use are designed for predicting discrete values, we have to classify the prosperity scores into reasonable partitions.In achieving this, the following basic statistical inferences are made:

mean	median	maximum	minimum	standard deviation
48.38	50.57	98.63	15.47	19.05

The following is the visualization of distribution of prosperity scores amongst all data.



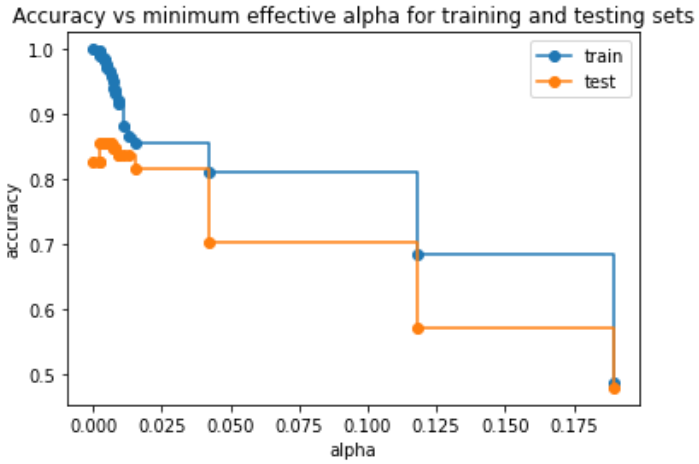
We observe that most countries have prosperity score in between 75 and 25 and a few are below 75. It is important to note that by Atlantic Council's definition, 75 is the threshold for developing countries and developed countries and 50 is the threshold for under developing countries and developing countries. We partition our data accordingly by this standard. We set the label 1, 2, and 3, where 1 stands for under developing countries, 2 stands for developing countries, and 3 stands for developed countries. Combining all the data from 2021, 2016, 2011, and 2006, we find that there are 384 under developing labels, 213, developing labels, and 99 developed labels.

labels	meaning	threshold	size
1	under developing countries	score <= 50	384
2	developing countries	score <= 75 and score > 50	213
3	developed countries	score > 75	99

Next, we delete all rows where there exists at least one null values. In separating training data and testing data, we use the default 0.8 value, whereby 80% of the data are for training and 20% of the data are for testing. After this process, we find that we have 395 data set for training and 98 for testing. It is also worth while to state that we have 17 features.

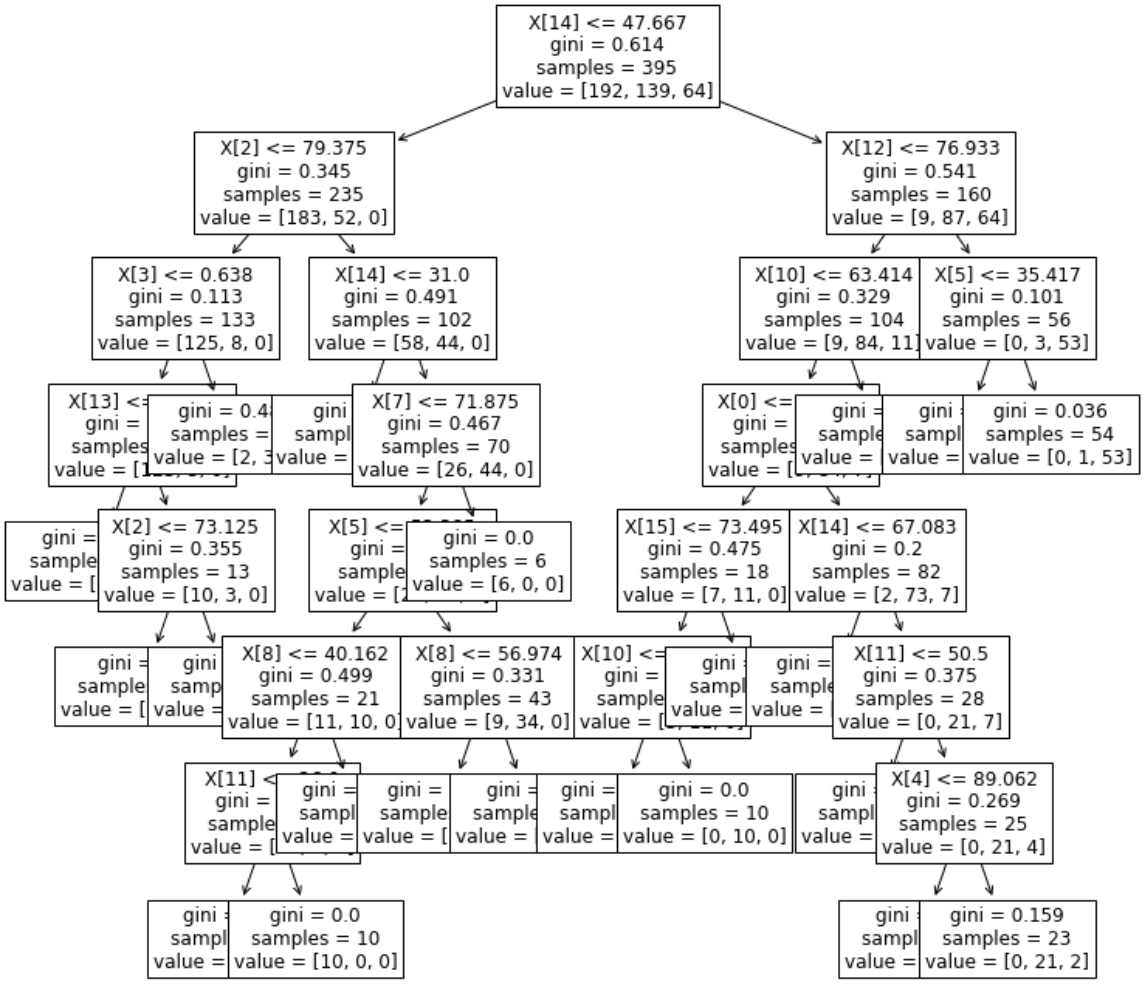
Decision Tree

Our initial decision tree model yields a 100% accuracy on training data and 82.65% accuracy on testing data. In avoiding over fitting, we then conduct post pruning. Our method is to continuously delete the node with the lowest impurity (least able to classify data) until the test data accuracy is maximized. The lowest impurity is manifested by "effective alpha". The following graph shows the accuracy with every node we prune (the x axis is the lowest effective alpha of ever tree).



While training accuracy decreases monotonically with nodes being removed, testing accuracy increases to a pinnacle, and then decreases monotonically. This is performing as expected since as the nodes are being removed, the problem of over fitting is being addressed, and as it starts to decrease, the model becomes to under fitting. In light of this observation, we pick the pinnacle of testing accuracy to be our final decision tree model. We reach a 85.71% accuracy for testing data and 95.94% accuracy for training data.

In addition, the simplicity of decision tree model also allows us to open the “black box” of the model itself. The following is the visualization of our post pruning model.



Decision tree visualization

Such visualization allows us to obtain the nodes with highest impurity. It implies that the nodes are amongst the strongest in separating labels, and hence, they are the most determining factors in implicating prosperity. From descending importance, these variables are: “State capacity”, “Absence of Corruption”, “Government integrity”, and “Efficient judiciary”. In light of this finding, we state that an effective governmental apparatus in maintaining law and order correlates with a country’s prosperity.

Other attempts

Although we reached a decent accuracy based on random forest, it would be beneficial if we test the performance of other major models. We also attempted random forest, gradient boosting, and neural network. For random forest, we obtain the optimal results after 300 iterations, and for gradient boosting, it is 100 iterations. For neural network, we setup 1 hidden layer and 50 training epochs. The following is the performance of all our models.

Model	training data accuracy	testing data accuracy	training data F1 score	testing data F1 score
Decision tree	0.960	0.857	0.960	0.859
Random Forest	1.000	0.867	1.000	0.864
Gradient Boosting	1.000	0.878	1.000	0.876
Neural Network	0.486	0.480	0.654	0.648

Gradient Boosting and Decision Tree performed slightly better than Decision Tree, while Neural Network’s performance is catastrophic, this is possibly due to lack of training data, and preponderance of features. Simple model such as decision tree almost performed as well as Random Forest and Gradient Tree.

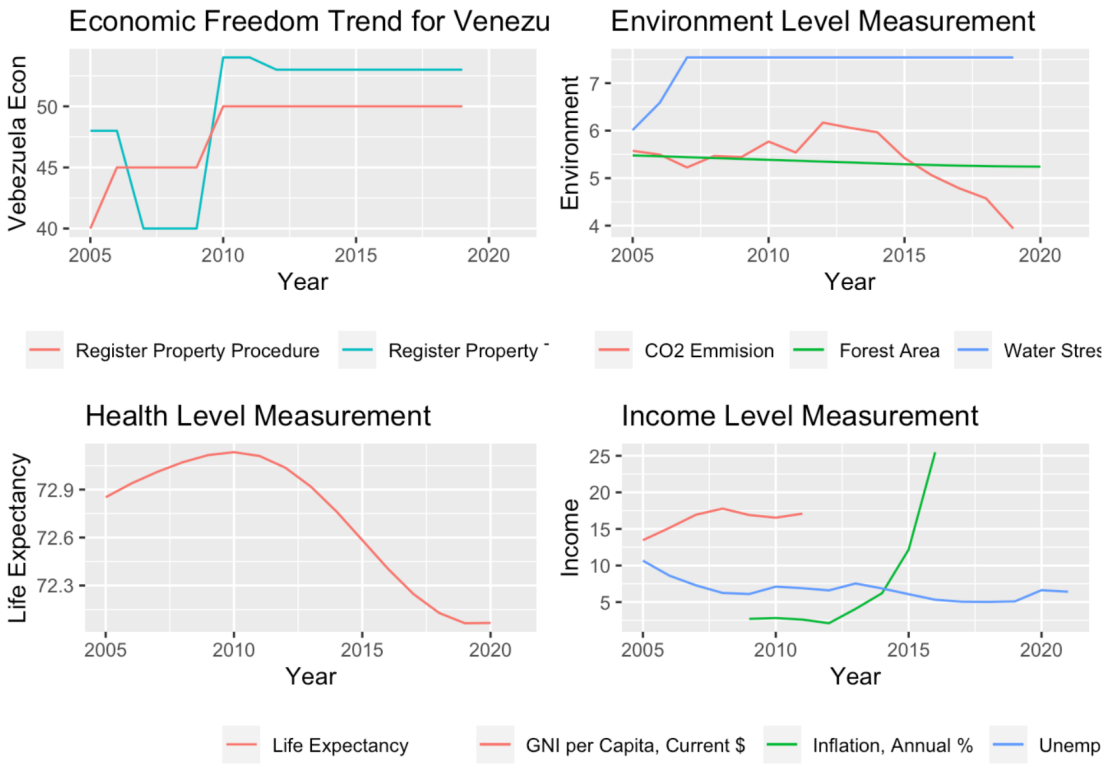
Time Series Data Analysis with Raw Data

Understanding what variables are important according the correlations of the normalized scores, we will check the current situation of some countries and investigate if the raw data aligns with our findings from the scored dataset. Countries are chosen based on three criterion, biggest change in freedom/prosperity score, richness in raw data, and representation of the world. Based on these, four countries were chosen: Venezuela, Hungary, European Union, and USA.

The raw data were fetched from the World Bank using the package WDI. Several indicators were selected to replicate each of the category the original dataset used.

Venezuela, RB

According to the scores analysis, Venezuela has a massive decrease of freedom score by -17 (-42%) and prosperity by -13.5 (-23.2%) between 2006 and 2021.



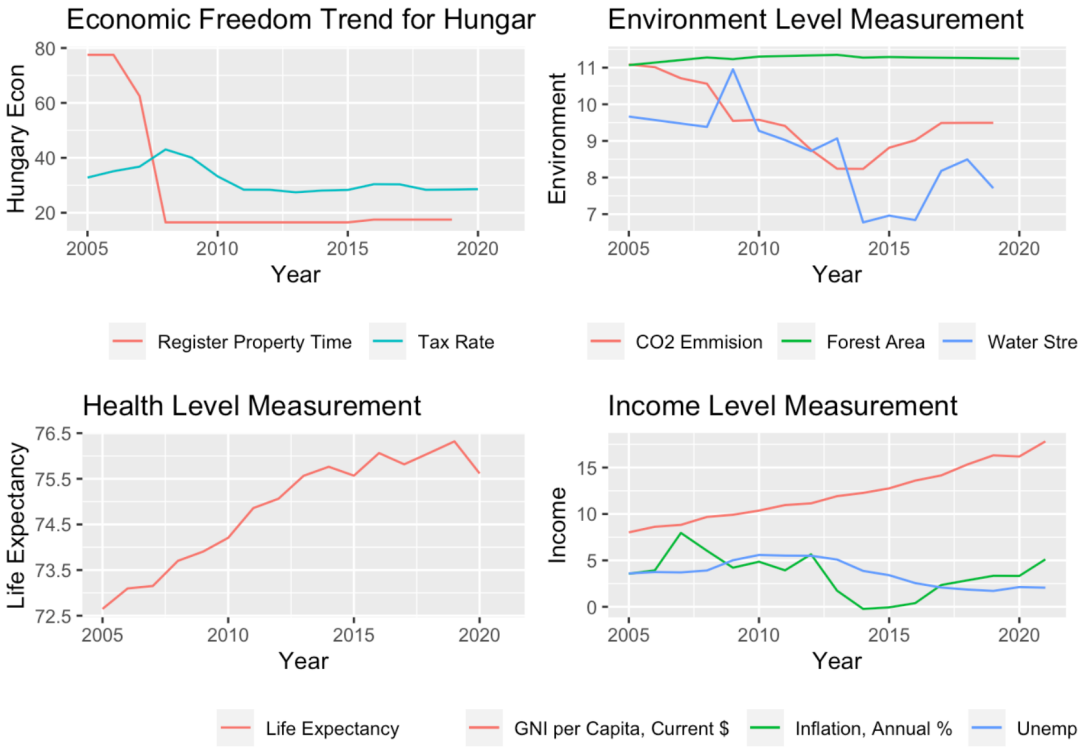
Venezuela Time Series

- The Property Registration procedures used to measure a nation's legal freedom is steadily increasing.
- Based on the environment plot, there is a significant downward trend for the level of CO2 emission which is a positive.
- The average life expectancy in Venezuela is decreasing since 2010.
- There is a lack of information regarding GNI, but the rapid increase in inflation and decrease in unemployment signifies that hyperinflation is most likely present, which is negative for a nation's development.

In summary, the trends in the income level raw measurements of Venezuela aligns with the freedom index trends and the net change of the health level measurements align with the prosperity trends. Property registry procedures and environmental factors did not possess any significant trends that contradicts with the score index.

Hungary

According to the scores analysis, Hungary has a massive decrease of freedom score by -9.6 (-12.6%) but an slight increase in prosperity by 4 (6.5%) between 2006 and 2021.



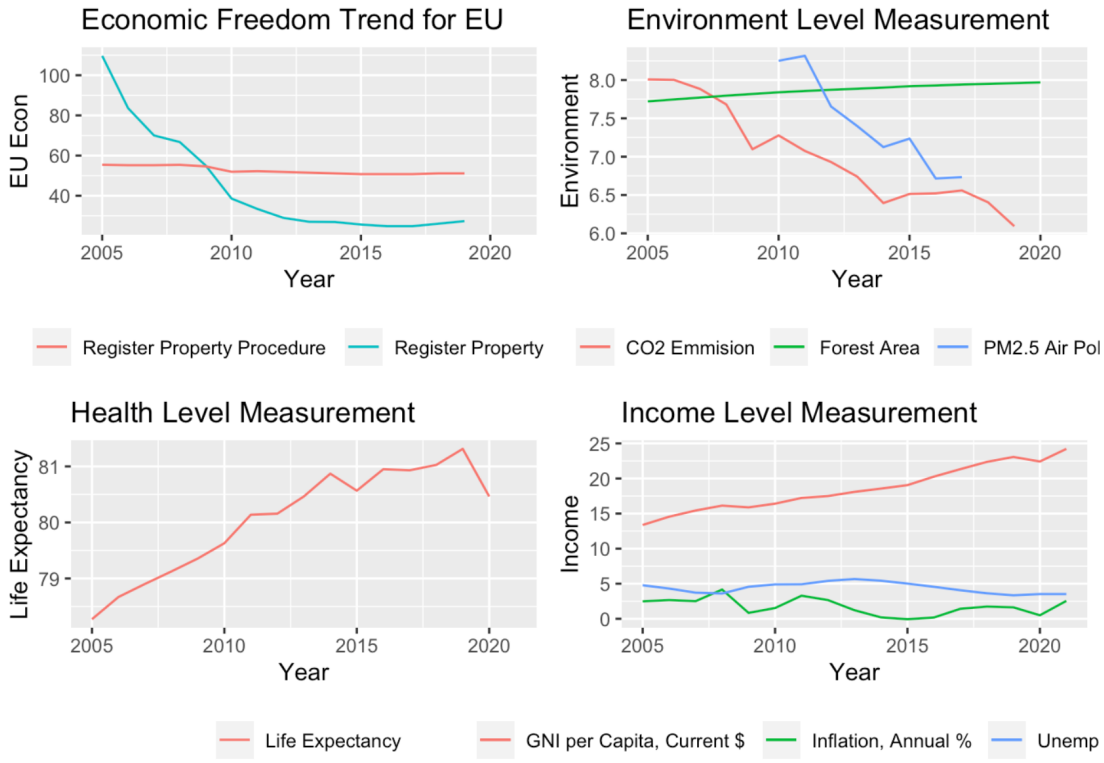
Hungary Time Series

- Significant decrease in registry property time, which is a positive for a country's level of legal freedom.
- Overall decreasing trends in CO2 emissions and water pollution, which is a positive for Hungary's prosperity.
- Significant increase in overall life expectancy.
- GNI has a significant increase, while unemployment and inflation fluctuates around the national equilibrium. This is an ideal sign of positive economic growth.

Unlike Venezuela, Hungary's raw data seems to show an overall positive trend in the nation's freedom - economically, politically, and legally. Thus for this case, the data from the scores for freedom does not align with the significant factors trends for Hungary. However, the prosperity index trend aligns with Hungary's environment and health level measurements.

European Union

To work with a larger dataset, we have decided to proceed with looking at a regional level, with the first example being the EU. There is an overall increase of 1.8 (3.5%) units of score in freedom and 1.1 (3.9%) unit of prosperity.



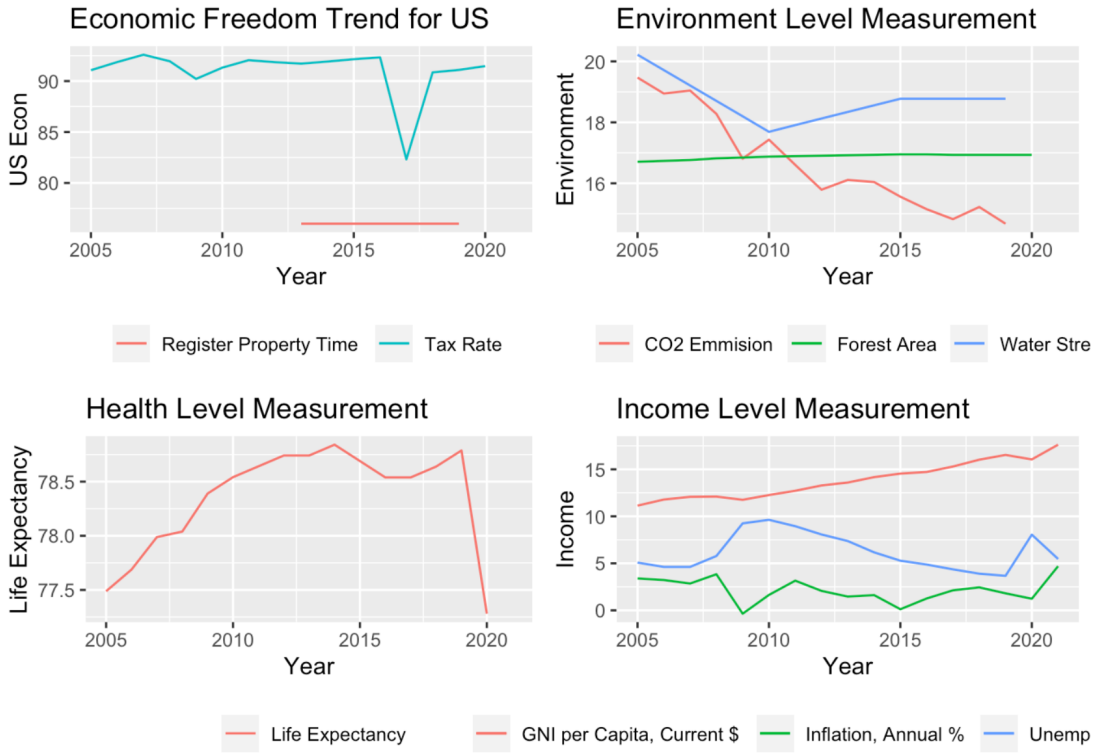
EU Time Series

- The rapid decline of property registry represents a positive increase in the region's freedom.
- The overall decrease in air pollution and CO2 emissions along with the increase in forest areas signifies a positive impact on the region's prosperity.
- Overall increasing trend in life expectancy aligns with the positive prosperity change.
- The steady increasing trend in income levels and fluctuations is an ideal representation of positive economic growth.

To summarize, every factor shown in the four plots for EU aligns with the overall trend for both freedom and prosperity.

USA

The final case study is performed on US. The data from AC shows a slight change of freedom score by -2.3 (-2.8%) units, while 0.32 (0.39%) increase for prosperity. Though there is little change in prosperity, the United States is still worthy for our study due to the abundance of data available.



USA Time Series

- CO2 emissions are rapidly decreasing.
- Though the life expectanc was increasing, there has been a recent massive dip.
- Though the nation's GNI is increasing, the unemployment rate and inflation rate is also slightly high.

To summarize, there are several factors in the US that seems to be both positively and negatively affecting the country's level of freedom and prosperity. It is interesting to see how the massive changes in different factors within the US does not seem to have much of an impact regarding the overall status of the country.

In conclusion, it is clear to see that environment, health, and income have been improving over time as a country get freer. In general, the trends of the raw data echoes with the trends evaluated by the Atlantic Council's scores.

Conclusion

To summarize, we have performed exploratory analysis on the dataset provided by the Atlantic Council, generated machine learning models to identify the significant variables in determining a country's freedom and prosperity, and have used these information to examine and analyze the raw data provided by World Bank.

The overall scores for freedom and prosperity follows a normal distribution, with two evident groups discovered by Gaussian clustering. From the decision tree classification, we have discovered that an effective governmental apparatus in maintaining law and order correlates with a country's prosperity.

The time series analysis performed for multiple cases generally aligns with our findings and resulting models, but realized that the lack of data provided by the World Bank hinders us from getting a more hollistic analysis.

Limitations and Future Work

There are several limitations regarding the this research project and possibilities for other researchers to expand upon:
- Investigate scores from past years

- Predict the futuristic scores of nations
- Instead of using the general correlation, other methods could be used to determine significant variables (e.g. PCA method) - Attempt on building a generalized linear model that can give an estimate to a country's freedom score given the raw data

Works Cited:

[1] Negrea, Dan, and Matthew Kroenic. "Do Countries Need Freedom to Achieve Prosperity?" Atlantic Council, July 7, 2022. https://www.atlanticcouncil.org/in-depth-research-reports/report/do-countries-need-freedom-to-achieve-prosperity/ (https://www.atlanticcouncil.org/in-depth-research-reports/report/do-countries-need-freedom-to-achieve-prosperity/).