Project Proposal

Group 4

2023-06-29

Question 1

Research Question

One of our research question is what are the key factors that most strongly affect the political corruption index, and can we use these factors to predict the corruption index? Hypothesis 1: Higher levels of electoral democracy (v2x_polyarchy), participatory democracy (v2x_partipdem), and liberal democracy (v2x_libdem) will be negatively correlated with the political corruption index. Hypothesis 2: Lower levels of civil society organization repression (v2csreprss) and higher scores on the core civil society index (v2xcs_ccsi) will be negatively correlated with the political corruption index. Hypothesis 3: Higher levels of clean elections (v2xel_frefair) and greater civil society participation (v2x_cspart) will be negatively correlated with the political corruption index. Hypothesis 4: Freedom House scores related to civil liberties (e_fh_cl), political rights (e_fh_pr), and the rule of law (e_fh_rol) will be negatively correlated with the political corruption index. To investigate the factors that have the greatest impact on the political corruption index, a multiple linear regression analysis will be conducted using a stepwise approach. The aim is to identify the 3-5 variables that exhibit the strongest association with the political corruption index. The dataset will be divided into a training set and a validation set to assess the predictive capability of the developed model. By evaluating the models performance on the validation set, it will be determined whether the selected factors can effectively predict the corruption index. This approach will enable a focused exploration of the key factors influencing political corruption and provide insights into the potential for developing a reliable predictive model for corruption levels.

By following these steps, we aim to identify the most influential factors contributing to political corruption and develop a model that can predict the corruption index based on these factors. The analysis will provide insights into the relationship between various political, economic, and sociological variables and political corruption, contributing to a better understanding of the factors influencing corruption levels globally.

Data for Question 1

We will look at all predictor variables included in the data set, except for the variables used to identify the data points - country_name, histname, year, and our dependent variable v2x_corr. In other words, we would feed all these predictors to help build our multiple linear regression model (and find out which top 3 or 5 predictors are most closely associated/correlated with higher corruption level. We would split the data into training and validation sets to test whether our model is able to predict/estimate corruption level in nations in the validation set. On another note, we could also do PCA to figure out which nations are most similar to one another in terms of all variales except country_name, histname, year, and our dependent variable v2x_corr. Then, we could cross-check with the variables in our multiple linear regression model to see if the variables in both approaches overlap.

Question 2

Research Question

Our second research question idea explores the relationship between prosocial/ social behaviors and civil society index. This question is of interest because it pulls together politics and measures of wellbeing to explore what effects civil society, which is the area outside of business and government, such as family and community. We want to explore how various measures of community engagement and support, both positive and negative, predict measure civil society scores. This will allow us to practice both clustering and regression modeling to answer this question. We will also be able to practice cleaning and merging data. While we cannot say for certain what relationships we expect to see, finding relationships between the democracy and happiness index datasets is an exciting possibility. Specific relationships we could see include a potentially negative between civil society organization oppression, freedom to make life choices, and wars/coups, a potentially positive relationship between social support, participation rate in civil society, generosity, education, and maybe government corruption. We expect places with high civil society indexes to have higher amounts of things that bring people together and lower rates of things that bring people apart or make meeting with other people difficult.

Alternately, if the idea above doesn't seem feasible, we could explore predicting civil society participation instead of the index. We could also explore whether these predictors of civil society also predict happiness, which is said to be one of the reasons for participation in civil society. Civil society participation could even be a more interesting question than predicting the index.

Data for Question 2

Since our topic is about how civil society index be affected by prosocial behaviors, we will use the variables listed above, which include: civil society repression effort scores, social support scores, freedom to make life choices, presence of war, presence of coups, participation rate in civil society, generosity, education, and government corruption. The data sets we will be using are vdem (democracy) and wellness. Combining these variables allow us to look at how civil society and connections between people exist from different lenses. Combining data that is political and sociological in nature allows us to get a fuller perspectice and understanding of the relationship between different behaviors that connect people, prosocial behaviors, and their relationship to civil society in different countries.

In the first part of the data processing, we will determine which of the factors will be important/influentional/significant in predicting civil society index (or participation…). After choosing the siginificant factors, both supervised and unsupervised methods could be conducted by either linear models and clustering, or if we focus on prediction/grouping, doing k-means clustering and hierarchical clustering. This would allow us to explore the difference in k-means and hierarchical clustering to see how the prediction/results change.