Springboard Data Science

CAPSTONE PROJECT #1

by Jason Green-Lowe

# Executive Summary

Will ambassadors to the UN General Assembly sell their votes in exchange for foreign aid? Based on publicly available data from the United Nations, the World Bank, and the US State Department, the answer is no: **US foreign aid has no practical effect on voting patterns in the UN General Assembly.** A country that has received billions of dollars in US foreign aid is no more likely to vote alongside the United States than a country that has never received aid. Instead, voting patterns appear to be closely tied to the nature of each country's domestic economy, as measured by statistics such as GDP per capita, exports, the GINI inequality index, and foreign debt. On average, richer countries with less internal inequality are much more likely to vote the same way that the United States did on any given UN resolution.

Countries also show voting patterns in the General Assembly based on geography and/or shared culture. Using Principal Component Analysis (PCA), I reduced the information from thousands of different UN votes into two arbitrary dimensions that appear to roughly represent each country's political alignment. Most of the countries on this plot are tightly clustered together, with one cluster for Scandinavia and Western Europe, one cluster for Eastern Europe, one cluster for central Asia, and one cluster for Arab Muslim states.

## Data Sources and Data Cleaning

**UN General Assembly**

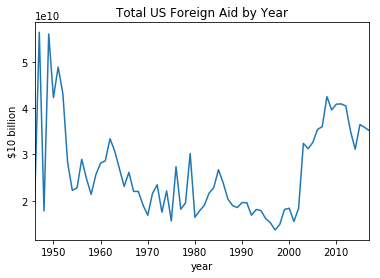
* The UN General Assembly provides CSV tables of all countries' votes on all resolutions that have come up for a roll-call vote.
* The primary difficulty in working with this data was converting the votes into values that would be mathematically useful. The UN codes 'yes' votes as 1, 'abstain' votes as 2, 'no' votes as 3, 'absent' votes as 8, and votes for which a country was not yet a member as 9. I reclassified absent votes as abstentions (because most absences from the UN are deliberate), and dropped records where a country was not yet a member (because a country that did not have the opportunity to cast a vote does not provide us with useful information).
* To best estimate each country's voting alignment with the USA, I coded instances where both countries cast the same vote as "0", instances where one country abstained and the other country voted as "1", and instances where both countries voted but in opposite directions as "4". This implicitly models a yes vote as "+1", an abstension as "0", and a no vote as "-1", and then provides a "voting difference" variable that is the square of the "distance" between the two countries' votes.

**World Bank**

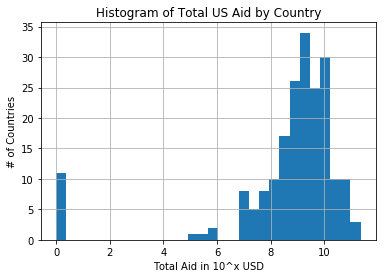
* The World Bank provides CSV tables of several key demographic variables for all registered countries.
* The primary difficulty in using the World Bank data was that they do not use the same country codes or country names as the United Nations, so I had to compile a list of all of the country codes and names used in each dataset in order to reconcile them. Using text-based searches for fragments of each country's name, I was able to confirm that each country code was representing a unique country, i.e., the same three-letter sequence never corresponded to more than one nation-state across the two databases.
* There were several instances where a country code was found in only one of the two databases, but this was always because the two institutions disagreed over who should be considered a country. For example, the World Bank keeps statistics on Kosovo and Bermuda, but Kosovo's membership in the UN has been blocked by Russia, and Bermuda is not a UN member because it is not considered sovereign; it is instead an autonomous possession of the United Kingdom. Conversely, the World Bank did not keep separate statistics on East Germany or North Yemen, even though these countries cast separate votes in the United Nations during some years.

## Data Exploration and Descriptive Statistics

**Tendencies in US Foreign Aid**

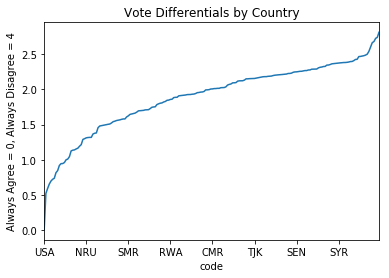


Grouping the dataset by year and plotting the sum of each group shows that there were two main peaks in total US aid -- one in the 1950s, during the Marshall Plan, and one in the 2000s and early 2010s. The aid is measured in constant dollars, so the increased aid volume in recent years cannot be attributed to inflation alone. One possibility is that the September 11th attacks on the World Trade Center triggered an increase in aid to US partners in the fight against terrorism.



The total amount of aid that the USA has given to each country varies quite widely, but the logarithm of total aid falls within a normal distribution. Most countries have received a lifetime total of US foreign aid ranging between 108 and 1010 constant US dollars. Eleven countries have never received any foreign aid, and three countries received more than 10 billion in aid: Egypt, Israel, and Afghanistan. Most of the countries in the detached section in the center of the histogram (near 105 USD) are geographically tiny, e.g., Tuvalu, or Brunei., which could explain why so little aid was delivered.

**Tendencies in UN Voting Patterns**



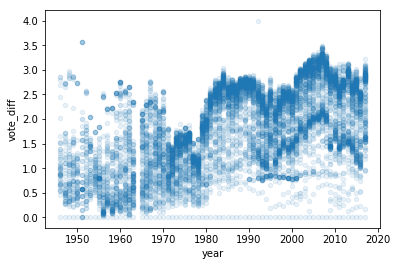
Is it common for countries' voting patterns to align with the votes of the USA? The chart to the right shows each country's average voting alignment. Resolutions where both countries vote the same way (e.g. both vote 'yes') are scored as a difference of 0.0. Resolutions where one country votes yes or no and the other country abstains are scored as a difference of 1.0. Resolutions where the countries cast opposite votes (e.g. one votes 'yes' and one votes 'no') are scored as a difference of 4.0.

Most countries voted against the United States somewhat more frequently than they voted with the United States. The range is actually somewhat narrow: on a scale of 0 to 4, all countries fell between 0.5 and 2.8.

The countries most likely to vote with the United States were clustered in Western Europe (Britain, France, Belgium, Netherlands, Italy) and in the Pacific islands (Taiwan, Palau, Micronesia, Marshall Islands), and also included Israel, Canada, Australia, and New Zealand. Unexpectedly, the #2 most US-aligned state was Zanzibar. The countries least likely to vote with the United States were often in the general vicinity of Zanzibar (Djibouti, Eritrea, Namibia, Zimbabwe) and also included Soviet-aligned opponents such as Vietnam, Syria, Cuba, and North Korea.

|  |  |
| --- | --- |
| **Strongest Agreement** ISR 0.519  EAZ 0.588  TWN 0.651  WGERMANY 0.695  PLW 0.720  GBR 0.735  FSM 0.814  CAN 0.844  FRA 0.917  LUX 0.943  BEL 0.944  NLD 0.959  ITA 0.998  AUS 1.008 | **Strongest Disagreement**  SYR 2.376  CUB 2.378  LBY 2.381  DJI 2.502  ERI 2.542  ZWE 2.602  NAM 2.666  VNM 2.681  BRN 2.727  YEM 2.741  PRK 2.814 |

Another general trend in the data is that countries have voted against the United States more often over time. This appears to be primarily the result of the creation of new countries via de-colonization and the breakup of the Soviet Union. As soon as they were created, these countries began voting differently from the United States on UN resolutions.

The graph on the right is a scatterplot where each dot represents one country's average voting alignment with the USA (averaged across all resolutions that came to a vote that year), set at alpha = 0.10, so that the fully-colored-in dark blue areas represent at least 10 countries per circle. The two waves of dark blue, which begin in about 1972 and 1992, might represent the votes from de-colonized countries (sharply disagreeing with the USA) and votes from newly created former Soviet republics (moderate disagreement with the USA), respectively.

## Data Analysis and Inferential Statistics

**Predicting UN Votes using Multivariate Linear Regression**

Linear regression estimates the value of a dependent variable by summing the products of one or more independent variables and their respective coefficients. In this case, each country's voting alignment with the United States (the dependent variable) was estimated based on independent variables such as foreign aid received, total debt, GDP per capita, and energy usage per capita. The linear regression tools provided by *statsmodels* will automatically select the coefficients that minimize the sum of the squared errors between each country's predicted voting alignment and each country's actual voting alignment.

One insight that was immediately apparent from the linear regression was that domestic, demographic variables were much better predictors of a country's voting alignment than the country's history of receiving foreign aid. Although foreign aid was able to very weakly predict voting alignment when foreign aid was the only variable used in the regression, foreign aid lost all predictive power once demographic variables were included. As shown in the table below, any entanglement between foreign aid and voting alignment was so weak that it would be expected to occur by chance alone over 51% of the time. Meanwhile, the demographic variables were so well-correlated with voting alignment that each correlation had less than an 0.1% chance of arising through chance alone.

OLS Regression Results

==============================================================================

Dep. Variable: vote\_diff R-squared: 0.044

Model: OLS Adj. R-squared: 0.044

Method: Least Squares F-statistic: 264.6

Date: Mon, 11 Nov 2019 Prob (F-statistic): 1.73e-277

Time: 19:06:09 Log-Likelihood: -56968.

No. Observations: 28780 AIC: 1.139e+05

Df Residuals: 28774 BIC: 1.140e+05

Df Model: 5

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 1.5688 0.065 24.033 0.000 1.441 1.697

**total\_aid** 3.526e-13 5.45e-13 0.647 **0.518** -7.16e-13 1.42e-12

GDPpercap -1.63e-05 6.14e-07 -26.527 0.000 -1.75e-05 -1.51e-05

debt -0.0034 0.000 -9.862 0.000 -0.004 -0.003

gini 0.0178 0.001 13.649 0.000 0.015 0.020

exports 0.0075 0.001 14.031 0.000 0.006 0.009

==============================================================================

Omnibus: 122395.966 Durbin-Watson: 1.537

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3522.188

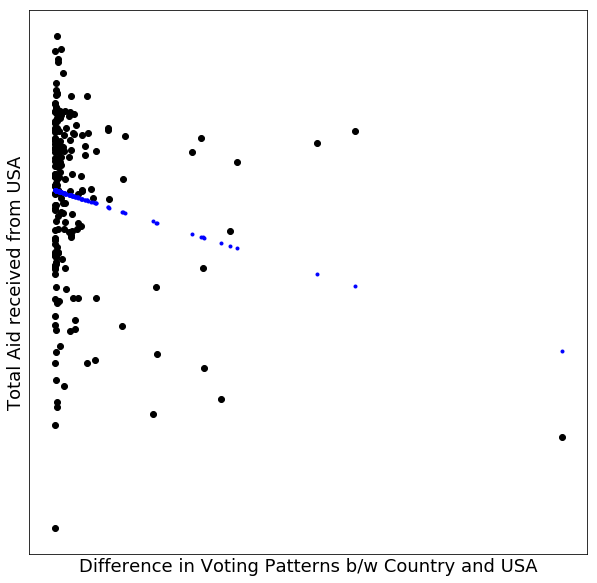
Skew: -0.031 Prob(JB): 0.00

Kurtosis: 1.287 Cond. No. 1.48e+11

==============================================================================

**The Model is Reasonably Robust**

An important issue to check in any regression is the effect of outliers. If a handful of data points with extreme values are unduly influencing the summary statistics, then the summary may not be theoretically compelling even if it is mathematically accurate. In this case, the surprising result was that foreign aid did not seem to influence voting alignment, so I needed to check whether there was a general influence on voting alignment that was being cancelled out by a handful of outliers with opposite values, i.e., a few countries who received enormous amounts of aid yet voted against the USA.

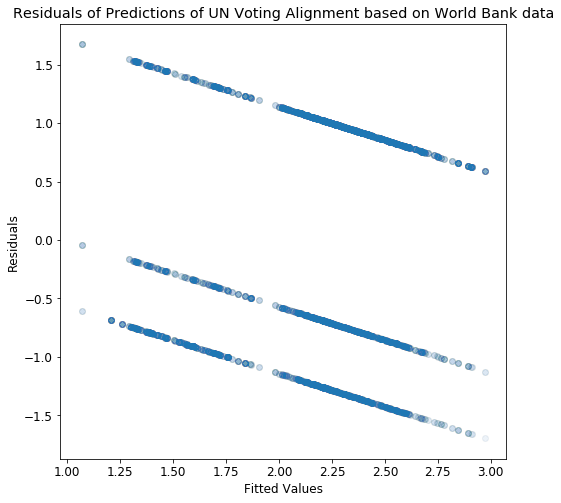


Fortunately, this does not seem to have been the case. Although the countries that received large amounts of aid often had unusually intense voting alignments, these alignments were evenly distributed -- some of these countries frequently voted with the USA, and some of these countries frequently voted against the USA. This tendency is illustrated in the chart on the left, where the large black dots show actual voting alignments for each country, and the small blue dots show the voting alignments expected for those countries based on their aid totals. Based on the chart, none of the outliers appear to have particularly high leverage.

To further confirm that the lack of correlation is not being driven by outliers, I sorted the database based on the absolute value of the residuals. The USA itself was the biggest outlier, as the USA by definition always votes with itself, and thus had a voting alignment of 0.0 despite receiving zero foreign aid, which the model was not able to successfully predict. The next largest residual was Zanzibar, with a residual of 1.34. The residuals continued to decrease; the tenth-largest residual was Austria, with a residual of 0.92. All ten of these residuals were removed from the model, and the next ten residuals continued to gradually decrease; the twentieth-largest residual was Afghanistan, at 0.75. However, even after removing these residuals from the model, the r-squared value of foreign aid was only 0.06, up from 0.04 in the model with all available data.

Another hazard in multivariable regressions is collinearity: if the so-called independent variables are not actually independent from one another, then minor hiccups in independent variables can cause large and chaotic shifts in which of those variables are thought by the model to 'explain' changes in the dependent variable. For example, suppose we are studying automobile accidents, and we are trying to understand why cars do not work after serious crashes. We could build a model that inputs the damage to the engine, damage to the axles, damage to the transmission, etc., and tries to predict the car's maximum speed based on this damage. However, because damage to each feature of the car is highly correlated with damage to other features of the car, the model could reasonably 'conclude' that any one feature was primarily responsible for the car's lack of speed, even though this has no basis in theory and is unlikely to be borne out by subsequent samples of data.

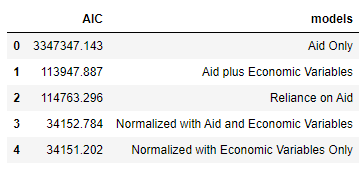
To measure the extent to which independent variables are dangerously intercorrelated, the *statsmodel* library provides a function that returns the variance inflation factor ("VIF") of a regression model. The VIF suggests the extent to which the variance reported by the model must be expected to increase in subsequent samples as a result of multi-collinearity. Values of 5 or greater generally indicate a serious concern. However, the primary model (with both foreign aid and domestic economic variables) showed VIFs between 1.02 and 1.22. This suggests that while there is understandably some correlation among different economic indicators in each country, there is no particular reason to doubt that the different economic variables are each capturing a meaningful effect on the countries' voting alignments.



One odd and somewhat concerning feature of the data is that the residuals from the multi-variable model are tightly connected to the fitted values of the model, with all data points being constrained to three very narrow diagonal lines, as shown in the chart on the right. One possible reason for this effect could be the way that the final dependent variable is inherently discrete, i.e., it can only take on the values of -4, -1, 0, 1, or 4. A continuous prediction would therefore be expected to show a certain amount of 'jumpiness' as the actual value leaps from one bracket to the next. However, this does not adequately explain why the data are so tightly contained within these three lines, nor does it explain why there are three lines and not five lines. Something is most likely being incorrectly modelled or interpreted here, but it is not clear what is causing the problem. One possible technique for further exploring this anomaly is to re-frame the analysis using logistic regression, coding exact matches (e.g. two yes votes, or two abstensions) as a '1' and all other pairs of votes as a '0'.

**Foreign Aid Adds No Explanatory Power to the Model**

To further test the theory that foreign aid is adding no explanatory power to the model, I measured several different indicators and tried re-calculating the value of foreign aid using different formulas. All of these techniques confirmed the initial result: foreign aid does not add any appreciable explanatory power to the 'domestic' model based on each country's exports, GINI inequality, GDP per capita, and foreign debt.

The standard measure of explanatory power is 'r-squared', which measures the fraction of the variability in the dependent variable that can be accounted for by variation in all of the independent variables taken together. In the model with foreign aid alone, r-squared is 0.003, indicating that less than 1% of the variability can be explained by foreign aid. In the model with demographic variables alone, the r-squared was 0.044, indicating that about 4% of the variability can be explained by demographic variables. In the model with both demographic variables and foreign aid, the r-squared was again exactly 0.044. In other words, foreign aid had negligible explanatory power, and added no detectable explanatory power to the demographic model.

Another measurement of explanatory power, the Akaike Information Criterion (AIC), measures the relative amount of information that is 'lost' by reducing a population to a relatively simple model. Lower values of AIC indicate that less data is being lost. The AIC includes a penalty term that increases the AIC when a model contains extraneous variables so as to properly disincentivize over-fitting. Here, the best available model that included foreign aid had an AIC of 34,152. The best available model that excluded foreign aid had an AIC of 34,151. This indicates that removing foreign aid from the model decreases the amount of information that is being lost, i.e., including foreign aid in the model does not perform any useful explanatory work.

A third confirmation that foreign aid is not a useful explanatory variable in this dataset is that the coefficient of regression for foreign aid is so small as to be of no practical significance. The coefficient is -3.5 \* 10-13, meaning that providing a country with $350 billion in foreign aid will, on average, decrease that country's measure of voting disagreement by 0.01 on the 4-point scale. A movement of 0.01 is too small to be reliably detected by the available data; the standard deviation for this variable is 1.72. Even pairs of countries that are very close together in voting alignment still typically differ by more than 0.05. On the other hand, no country ever received $350 billion in foreign aid -- the maximum total aid (in constant 2010 US dollars) was only $227 billion. Therefore, moving from the very minimum amount of foreign aid ($0) to the very maximum amount of foreign aid ($227 billion) would, on average, produce a change in voting alignment (<0.01 on a 4-point scale) that is far too small to convey any useful information. This is further evidence that foreign aid is not associated with voting alignment in the UN General Assembly.

I was concerned that these results could in some way be driven by the extreme difference in scale between the explanatory variable and the dependent variable. With billions of dollars in aid and only 4 points available on the voting alignment scale, was the effect of foreign aid somehow being artificially diminished. As it turns out, this was not the case. Normalizing all variables to use a range between 0 and 1 had no effect on the p-values or r-squared values. Normalization was helpful in lowering the AIC of all models, both with and without foreign aid, but normalizing the variables did not tend to disproportionately lower the AIC of models that included foreign aid.

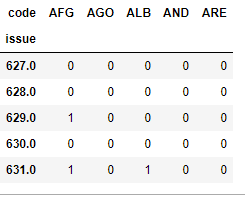
My other concern was that the results were being thrown off by the fact that some countries have existed for a longer period of time, giving them more of a chance to accumulate a higher total amount of foreign aid. If the countries that have existed since 1948 have different political leanings than countries that were only created during, e.g., the de-colonization of the 1970s, then total aid summed from 1948 through 2017 might not be an appropriate metric. To correct for this issue, I created a function that calculated the amount of aid a country has "recently" received, with aid received in the more distant past weighted less heavily than aid received in the current year. This allows for more straightforward comparisons between older and newer countries, and also embodies the realistic assumption that governments may not feel grateful or beholden based on aid that was received many years ago. I calculated each country's recent aid by going back over a window of 5 years, with a discount rate of 25% per year, so that last year's aid would count at 75% of the current value, the aid from two years ago would count at (0.75)2 of the current value, the aid from three years ago would count at (0.75)3 of the current value, and so on. This did not impact the significance of the variables -- foreign aid still had no explanatory power. Similarly, I tried ignoring all aid except the aid that was received in the same year that a vote was cast, but this also did not cause foreign aid to have any explanatory significance.

## Unsupervised Machine Learning

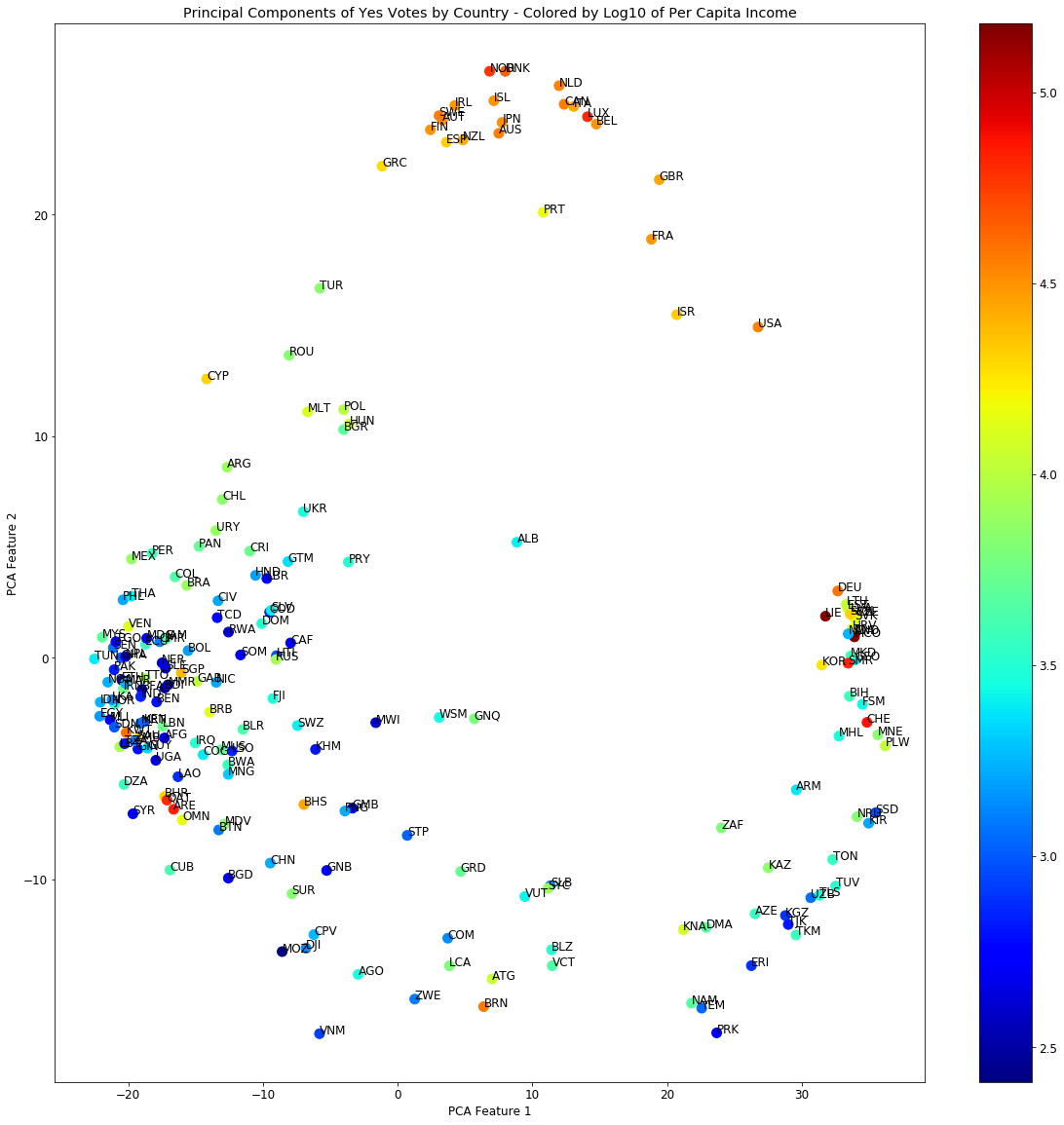
**Principal Component Analysis**

In addition to testing hypotheses that I generated myself, I also wanted to see what hypotheses would arise organically out of the data. Without using any exogenous variables, what can we tell about the way countries do and do not tend to vote with each other? Leaving aside any special emphasis on the United States, what blocs of countries are surprisingly aligned or unaligned?

The UN assigns 'category codes' to its resolutions that mark them as dealing with broad topic areas such as human rights, nuclear proliferation, and so on, but these category codes do not indicate whether a given resolutions is "for" or "against" a given topic. For example, both a resolution that restricted the spread of nuclear materials and a resolution that called for wider sharing of nuclear secrets would both be coded as dealing with nuclear proliferation. As a result, there is no easy way to build a model of countries' political preferences based on the specific content of the resolutions that they have voted for. Instead, it makes more sense to treat all resolutions as featureless observations of equal weight, and then to try to see which pairs or groups of countries are most likely to vote together on an arbitrary resolution. This allows us to visualize which countries appear to share each other's political opinions even if we cannot tell exactly what these political opinions are.

To perform this analysis, I reduced the UN voting data to a matrix of 0's and 1's, treating "yes" votes as 1's and all other votes (including no votes and abstentions) as 0's. The choice of which type of vote to code as "1" was based on both readability and silhouette scores -- it is possible to do a very similar analysis while coding "no" votes as "1", but the resulting graph is harder to read and has a lower maximum silhouette score for its optimized clusters (0.58 instead of 0.64). On the right is a sample of the matrix produced by treating "yes" votes as 1's. Each row of the matrix represents a country, and each column of the matrix represents a resolution that came up for a vote. Each entry in the matrix thus captures whether a particular country voted "yes" on a particular resolution.

This matrix can then be subjected to principal component analysis (PCA), a technique that identifies a handful of trends or tendencies in multi-dimensional data which account for most of the variance in that data, and then uses those tendencies to plot the data on a grid with relatively few dimensions. The initial matrix (before PCA) has over a thousand dimensions, because each roll call vote is theoretically a new opportunity for a country to express its political opinions in a unique way, but the reduced matrix (after PCA) has only two dimensions, because all of these political opinions have been condensed into these two dimensions. The choice of two dimensions is for convenience; it is easiest to illustrate the resulting trends on a two-dimensional grid. Countries that tend to vote 'yes' on the same resolutions will appear closer to each other on the resulting graph, shown below.



The colors in the graph above show the logarithm of each country's per capita income, so red represents 10^5 USD per capita, green represents 10^4 USD per capita, and blue represents 10^3 USD per capita. The income is averaged across the country's voting history, which is why even a rich country Norway is still reported at only about $20,000 per capita per year, because we're looking at an average of Norway's income from 1948 to 2017.

There are clear geographic and economic trends visible from the PCA graph. The Soviet Union's former satellites in Eastern Europe, for example, are all in the center-right portion of the graph, with Romania, Poland, Hungary, Bulgaria, and Ukraine all quite close together. These countries also tend to vote similarly to Cyprus and Malta and Turkey, which is interesting because they're neighbors geographically, but don't necessarily have a shared political heritage.

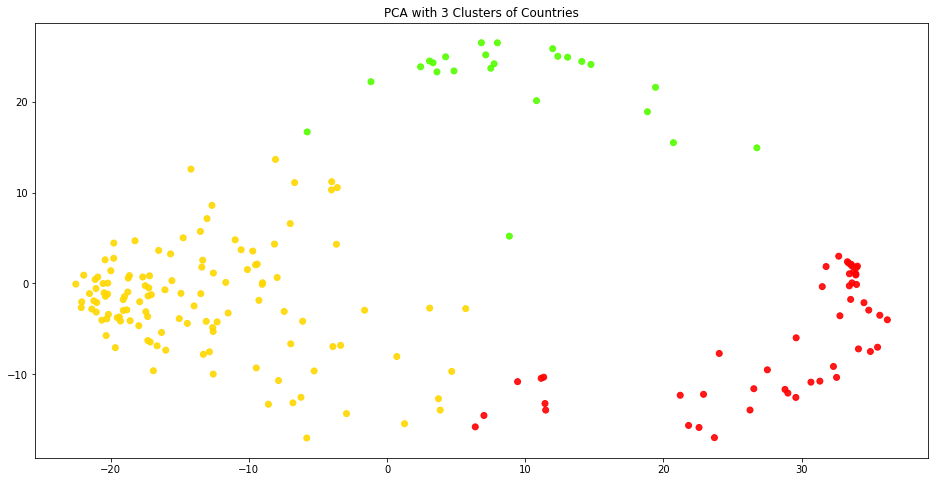
The richest countries, colored in orange and red, seem to fall into two groups: the USA, Israel, France, and Great Britain (democratic war hawks?) are off a bit to the right-of-center, somewhat distinct from the (less militaristic? more socialist?) countries at the very top of the graph like Netherlands, Canada, Belgium, Iceland, Norway, Japan, and Finland.

Finally, it looks like the very poorest countries, colored dark blue, are mostly split into two groups, one in the center-left, and one in the bottom-right. The group in the bottom-right has several central Asian countries, and the group in the center-left has several Muslim Arab countries, but it's not clear what if anything is causing that split.

**Clustering**

To formally test the significance of the groupings that appear in the PCA graph, I performed a K-Nearest Neighbors (KNN) analysis to sort the data points in the PCA graph into clusters. Because this is unsupervised learning, we do not know the true number of clusters in advance, so I performed the KNN analysis with all possible numbers of clusters between 2 and 10, and then calculated the "silhouette score" for each analysis. The silhouette score measures the extent to which points that were assigned to a given cluster tend to be closer to other points in their own cluster than to points from outside their cluster. KNN analyses that yield higher silhouette scores are more likely to have the 'correct' number of clusters, and silhouette scores that average above 0.5 suggest that a reasonable structure of clusters has been found.

With this data, the optimal number of clusters was 3, which yielded a silhouette score of 0.64. The silhouette score decreased monotonically in both directions as the number of clusters grew further from 3, suggesting that there is no need to try numbers of clusters outside the range [2, 10]. The three clusters identified through the KNN method are shown in the graph below, with each color representing a different cluster..



These clusters appear visually reasonable, and correspond fairly well to the political and geographical clusters that were subjectively identified before using KNN. One way to further test the stability of these clusters would be to randomly re-sample subsets of the total database of votes, and use those subsets to create new PCA graphs. If the new PCA graphs continue to fit neatly within the original clusters, then that suggests the clusters are representing a stable feature of the data.

## Implications and Directions for Further Research

* UN members may feel relatively free to vote in ways that further their domestic political interests, because UN General Assembly votes are typically non-binding. This suggests that countries may be 'paying' for US foreign aid by providing support to the United States in other, less publicly visible forums, e.g., patent enforcement, or trade deals, or covert military cooperation. The General Assembly offers what may be the highest possible ratio of visibility to impact. That could explain why UN voting patterns are based more on domestic political concerns than on foreign aid.
* Further research into UN voting patterns may help develop an understanding of how other countries conceive of their political identities; it would be interesting to try to measure the extent to which the clusters in the PCA correspond to specific cultural features such as religion, language, support for human rights, and/or ethnicity.
* An alternate theory of UN voting patterns is that smaller or poorer countries vote in favor of whichever patron has provided *more* aid. On this theory, US foreign aid alone might not predict voting patterns because the US might be engaged in a bidding war against, e.g., the Soviet Union or China (depending on the decade) for other countries' political allegiances. Follow-up research could collect more data on the foreign aid provided by other, non-US countries, and then compare each country's alignment with the US to its alignment with other large countries and then try to determine whether the difference in alignment (between, e.g., the USA and the USSR) correlates with a difference in aid (received from, e.g., the USA vs. received from the USSR).