Final_Project_Workup_Pingatore

Ross Pingatore

12/9/2020

Introduction: Political Stability Across the Globe

- Why does political stability vary across the globe? Are some nations innately built upon stability producing institutions while others are doomed, or can instability be triggered even within the most stable regimes? My research project will investigate the factors the produce political stability, or fail to, within the various countries across the globe.
- There is a vast deal of research and theories that investigate the mechanisms that produce political stability within a country. Many scholars will argue that the quality of a nations institutions are responsible for the stability, or lack thereof, that a country enjoys. The difficulty falls in attempting the quantitatively prove this theory. What kind of institutions are we categorizing, how many categories will we have, who will categorize them? The list of problems that would arise from such a task would be insurmountable. With that being said, my research aims to scratch the surface of this goal. The research conducted, utilizes a large N data set of countries and incorporates more rudimentary predictors such as GDP and Literacy rate in order to better understand variation in political stability. Let us be clear however that the predictors in this research are not mechanisms that produce political stability themselves. Metrics such as a stable GDP are preceded by the economic institutions that provide the conditions hospitable for a stable GDP. At most, our research will provide insight into the quality of a states institutions through our predictor metrics, which will give us some indication into the level of political stability of a country.

The Data

- We will utilizes The World Bank's data set on political stability measured by the absence of violence and terrorism. The time-series data contains 213 countries and provides estimates on political stability from 1996 to 2019. The estimate of political stability ranges from -2.5 (weak stability) to 2.5 (strong stability).
- The data is combine with additional data from The World Bank the includes predictors: population, fuel exports, military expenditure, ease of conducting business, inflation rate, literacy rate, and access to electricity. The compiled data frame is saved for additional research within this repository as compile_stability.csv.

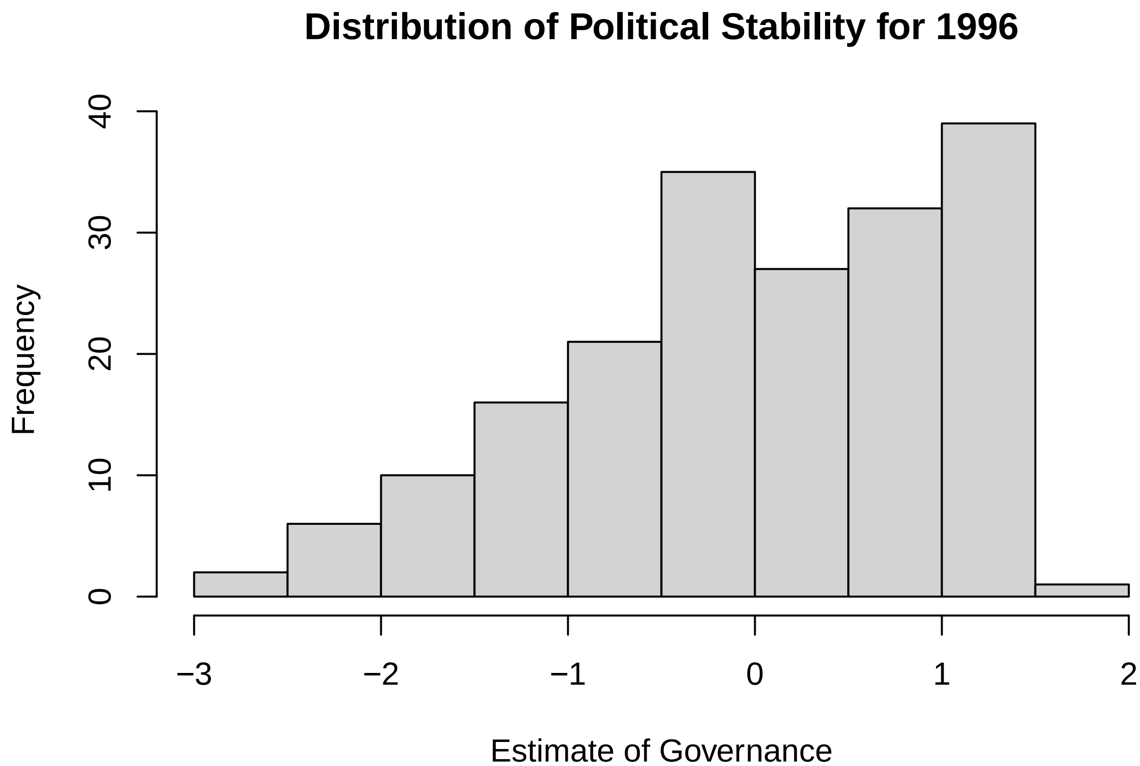
Preprocessing

- The data required a great deal of cleaning and pre-processing. Functions such as pivot-longer and dplyr's join functions were used to get the data in a usable format. The goal was to have the country, the country code, the year, and the estimate as variables in our data frame. Our initial cleaned data frame contains 3969 observations with 4 variables.

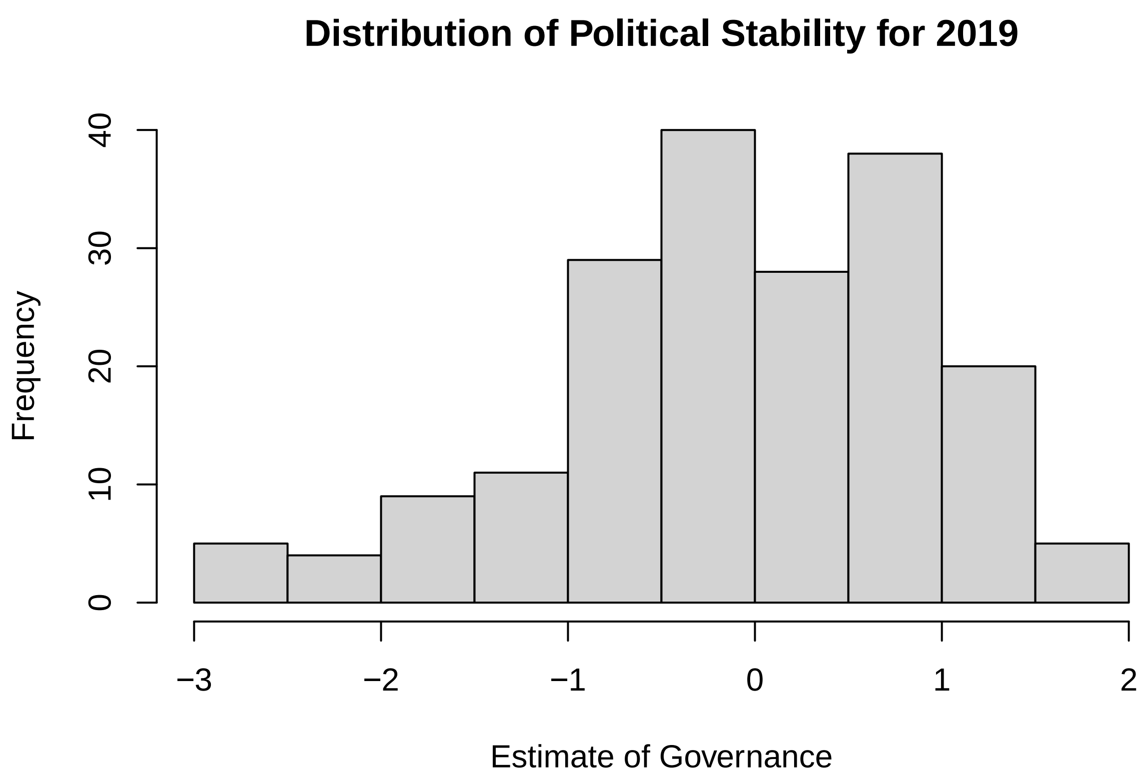
```
head(clean_stability)
```

```
## # A tibble: 6 x 4
##   country code   year estimate
##   <chr>   <chr> <dbl> <chr>
## 1 Andorra ADO    1996 1.1701573133468628
## 2 Andorra ADO    1998 1.1836445331573486
## 3 Andorra ADO    2000 1.1670020818710327
## 4 Andorra ADO    2002 1.282038688659668
## 5 Andorra ADO    2003 1.4649856090545654
## 6 Andorra ADO    2004 1.4014873504638672
```

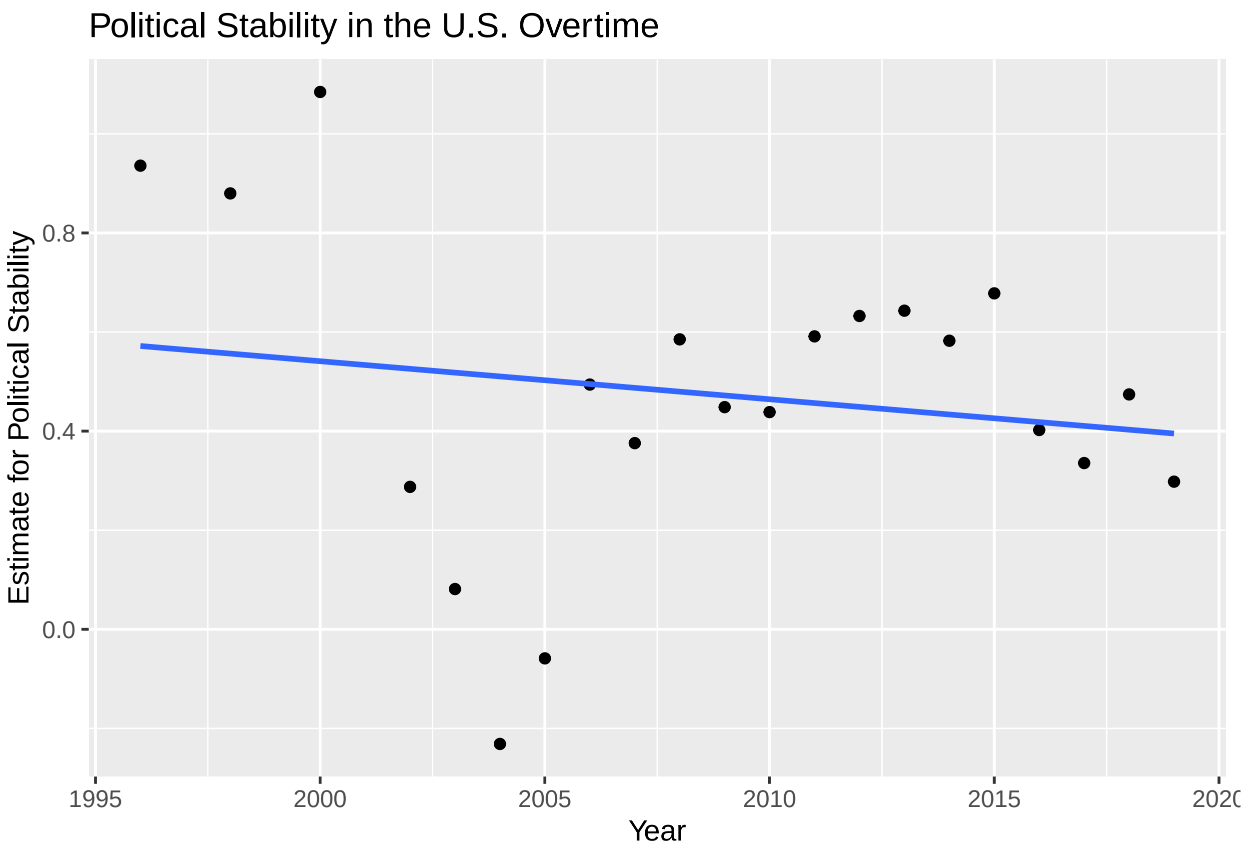
Preliminary Investigation



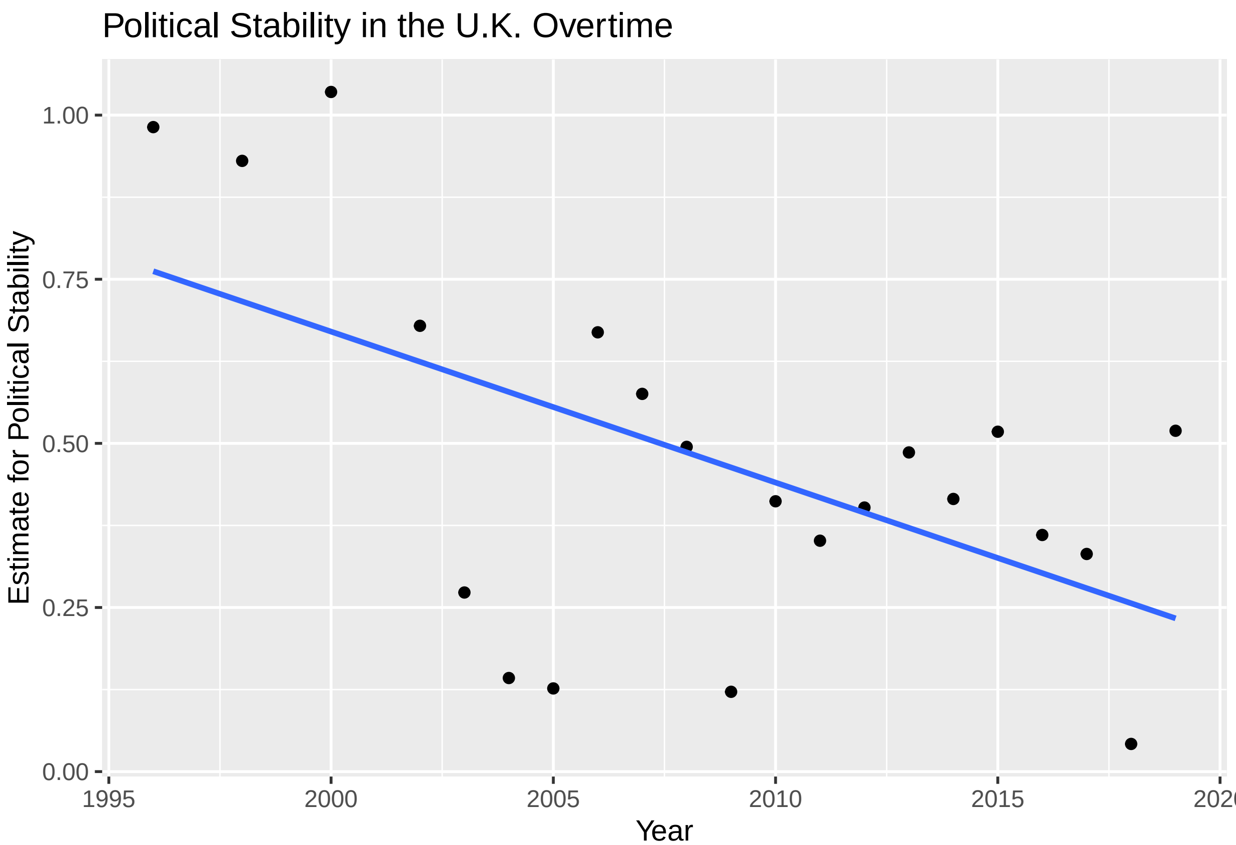
Preliminary Investigation



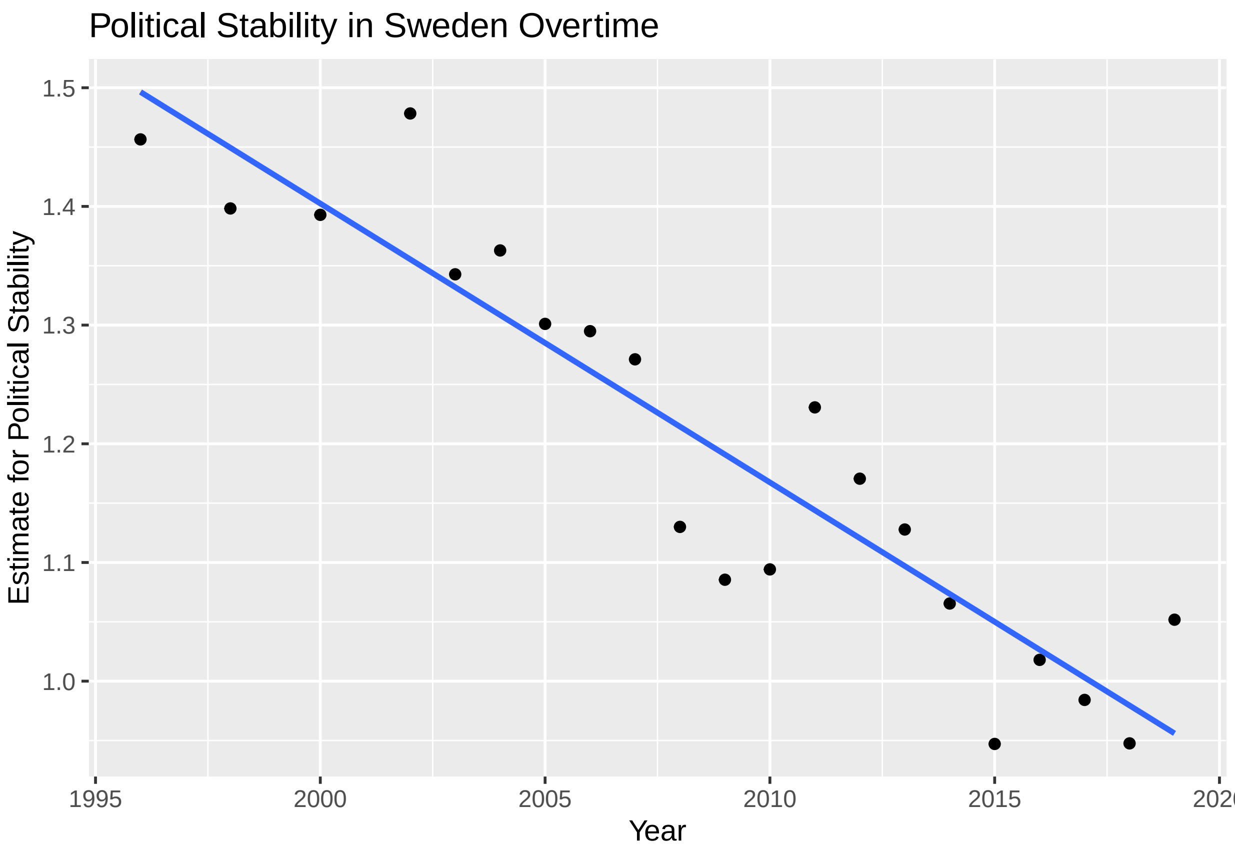
Preliminary Investigation



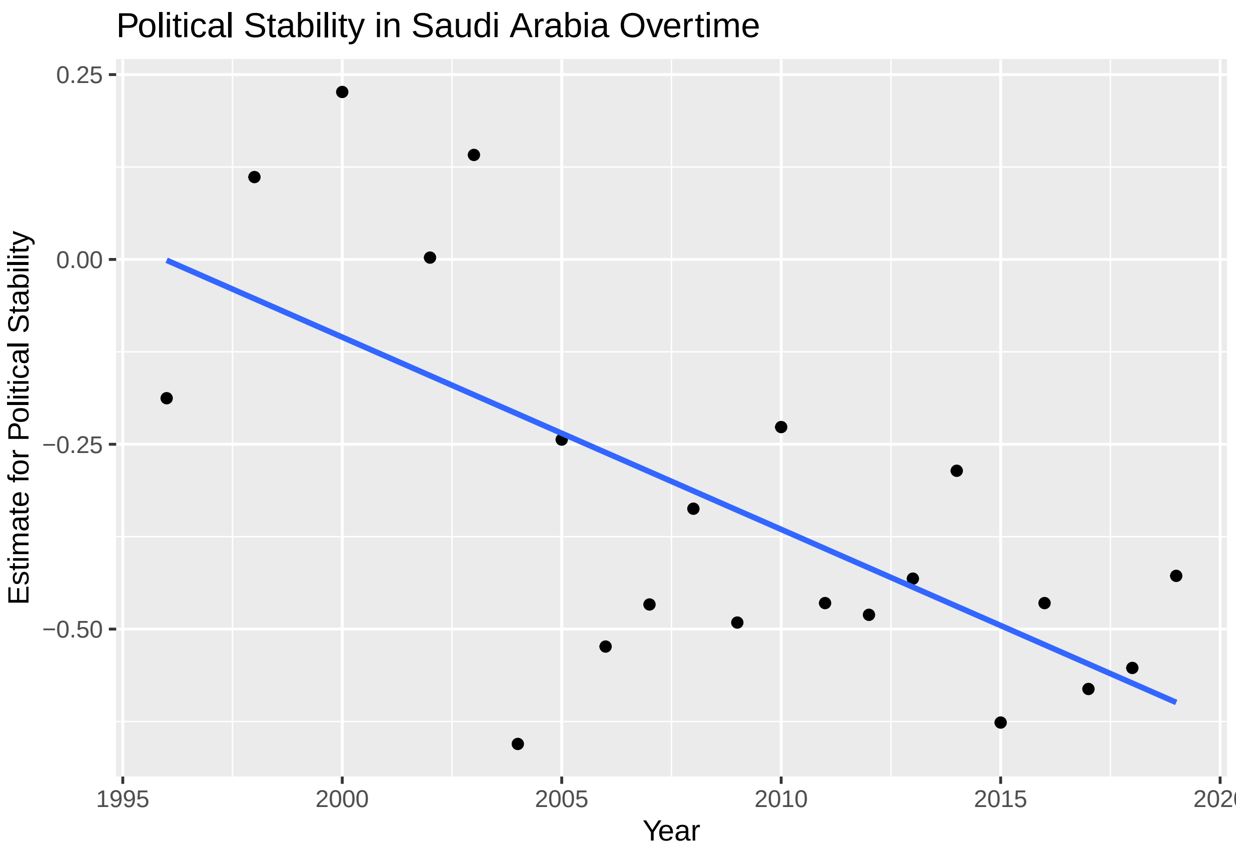
Preliminary Investigation



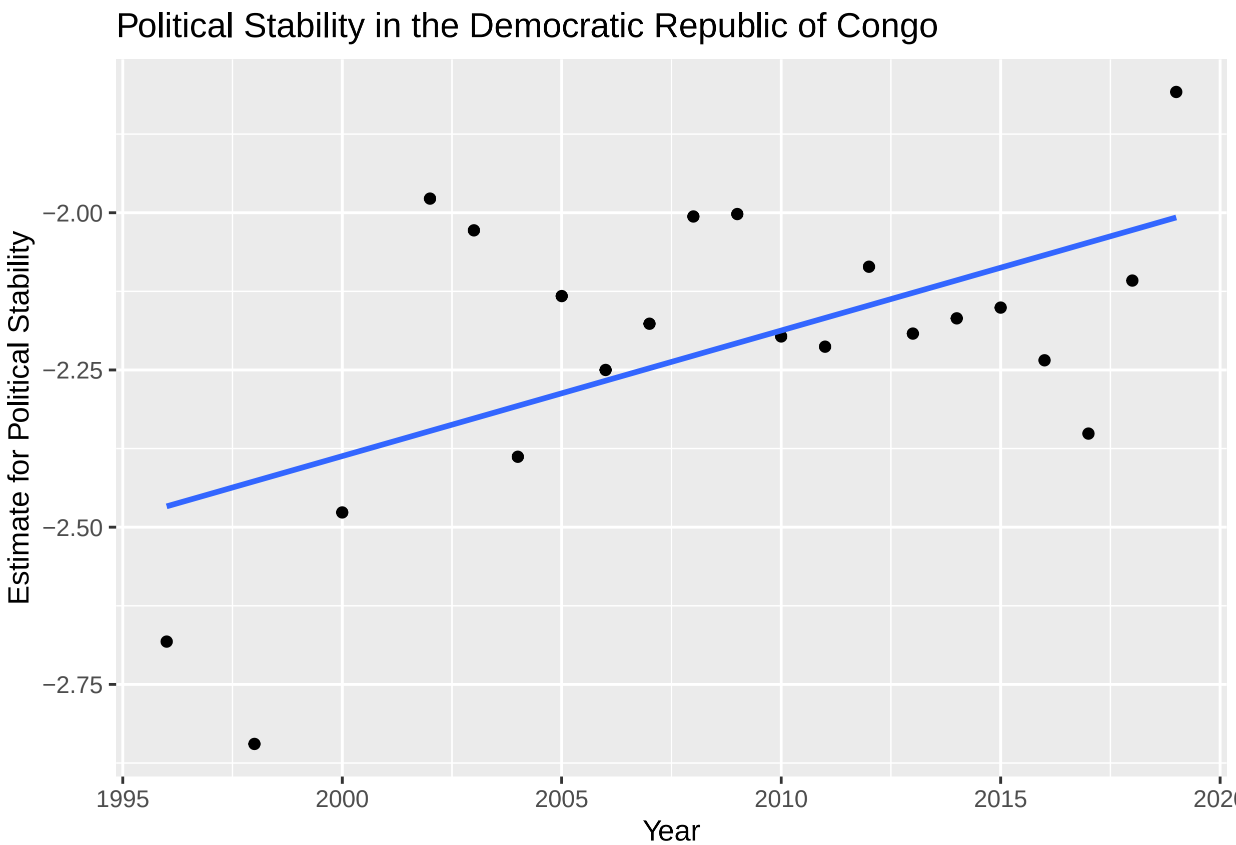
Preliminary Investigation



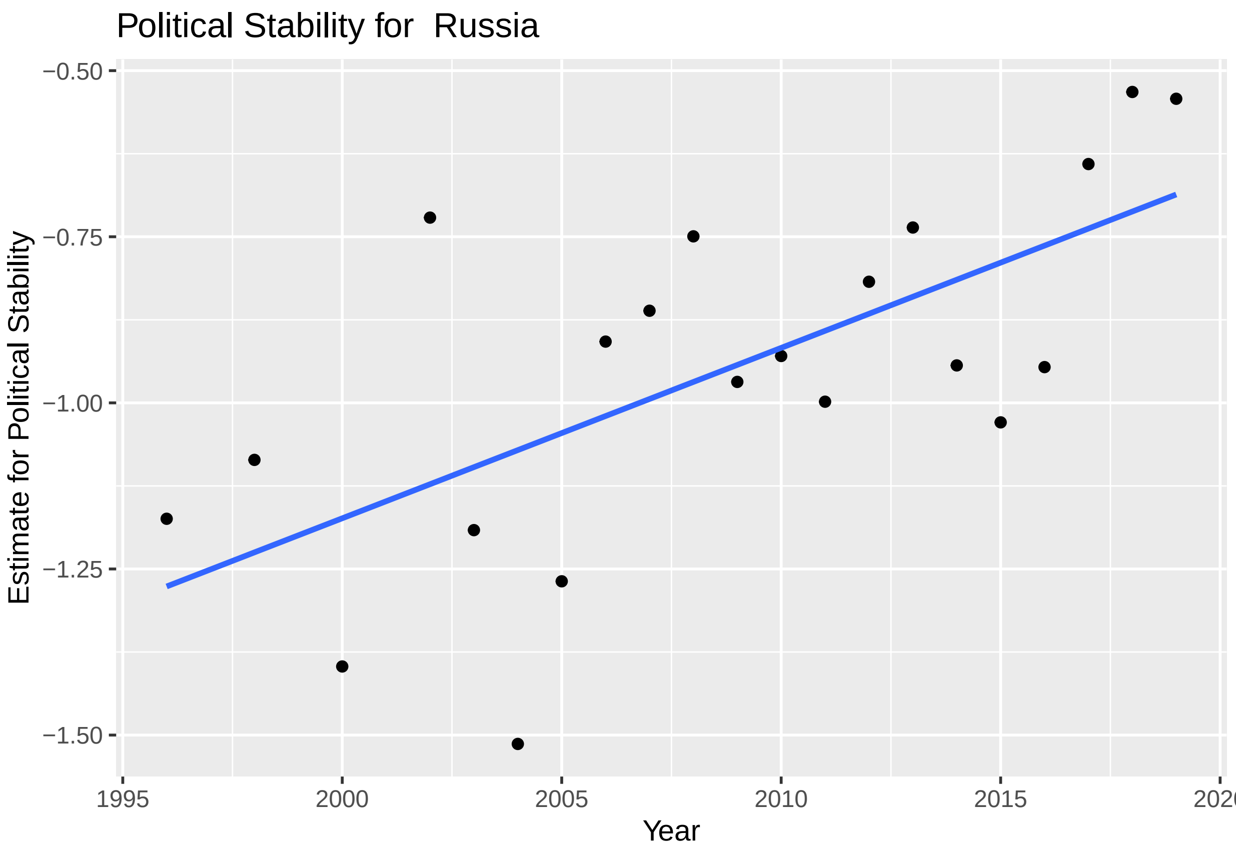
Preliminary Investigation



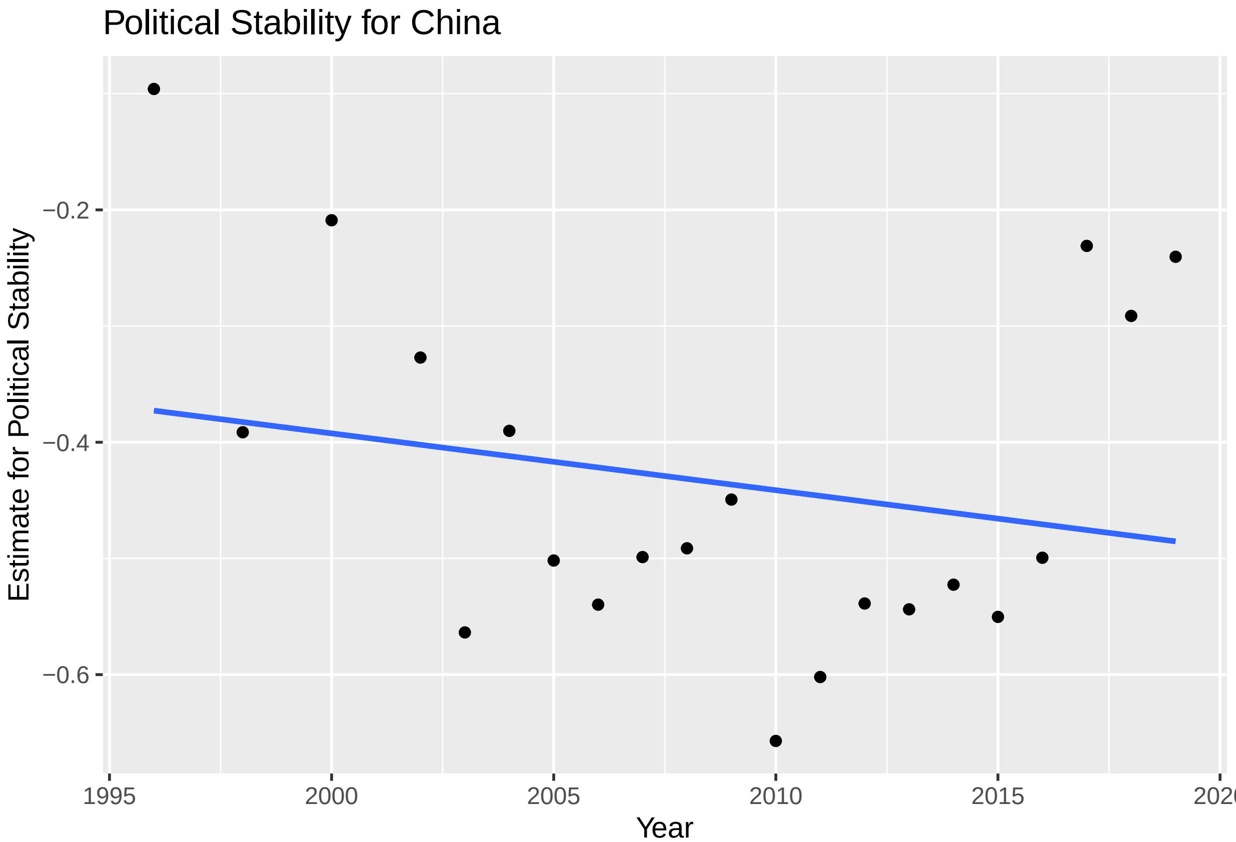
Preliminary Investigation



Preliminary Investigation



Preliminary Investigation



Adding More Data

```
##  [1] "country"              "code"                  "year"
##  [4] "estimate"             "gdp"                   "population"
##  [7] "fuel_ex"              "military_expenditure" "inflation"
## [10] "lit_rate"             "electric_access"
```

```
##    country               code                year         estimate
##  Length:495         Length:495         Min.   :1996   Min.   :-2.8100
##  Class :character   Class :character   1st Qu.:2007   1st Qu.:-0.7873
##  Mode  :character   Mode  :character   Median :2011   Median :-0.2753
##                                        Mean   :2011   Mean   :-0.3078
##                                        3rd Qu.:2015   3rd Qu.: 0.2518
##                                        Max.   :2018   Max.   : 1.6153
##       gdp              population          fuel_ex       military_expenditure
##  Min.   :6.975e+08   Min.   :8.372e+04   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:1.748e+10   1st Qu.:6.094e+06   1st Qu.: 1.374   1st Qu.: 1.048
##  Median :6.091e+10   Median :1.527e+07   Median : 6.673   Median : 1.511
##  Mean   :3.096e+11   Mean   :5.612e+07   Mean   :21.303   Mean   : 2.034
##  3rd Qu.:2.717e+11   3rd Qu.:4.729e+07   3rd Qu.:29.116   3rd Qu.: 2.672
##  Max.   :1.389e+13   Max.   :1.393e+09   Max.   :99.986   Max.   :12.035
##    inflation          lit_rate     electric_access
##  Min.   :-4.863   Min.   :12.85   Min.   :  3.696
##  1st Qu.: 2.283   1st Qu.:77.20   1st Qu.: 76.887
##  Median : 4.199   Median :92.06   Median : 98.035
##  Mean   : 5.649   Mean   :83.92   Mean   : 82.415
##  3rd Qu.: 7.313   3rd Qu.:95.86   3rd Qu.:100.000
##  Max.   :63.293   Max.   :99.97   Max.   :100.000
```

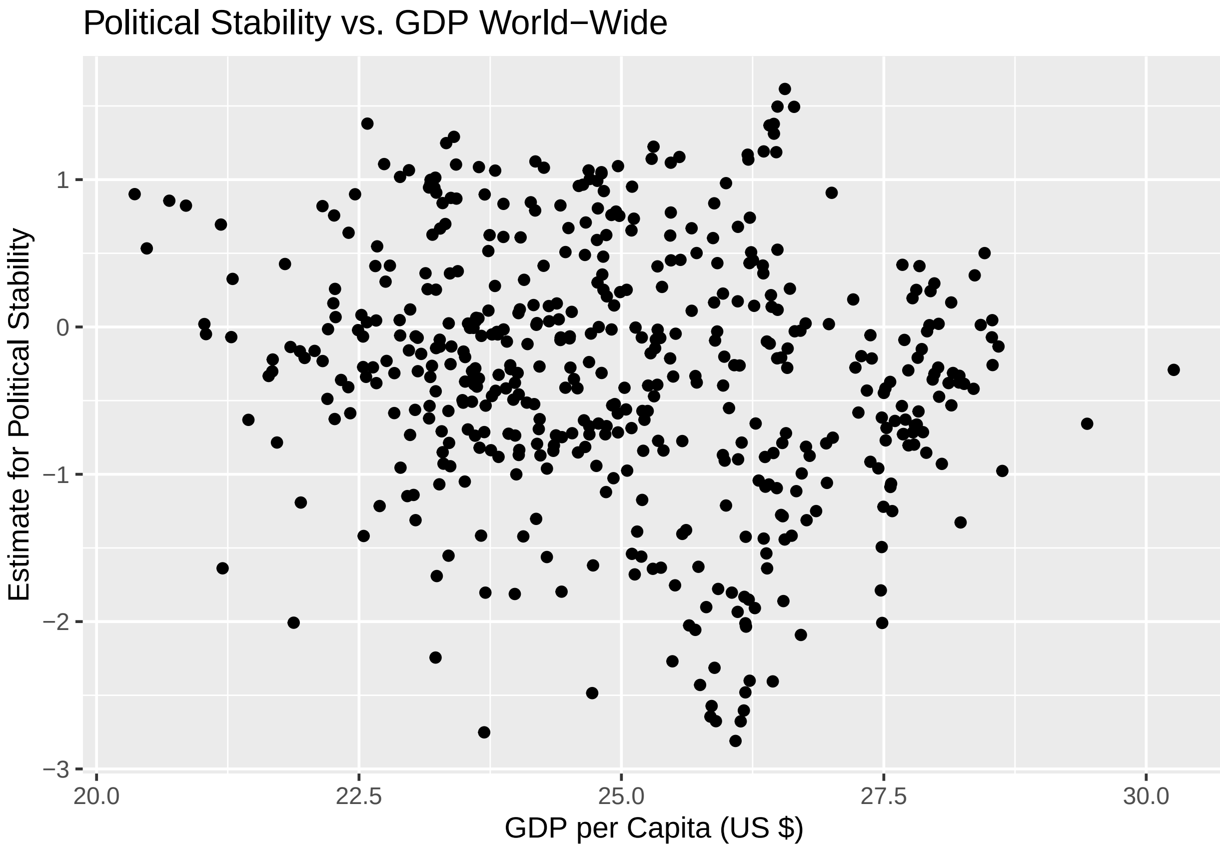
The New Compiled Data

- Upon merging all data frames into our compiled predictor_stability data frame, we were left with 495 observations and 13 variables. This is after removing NA values that did not align to all variables within a given observation.

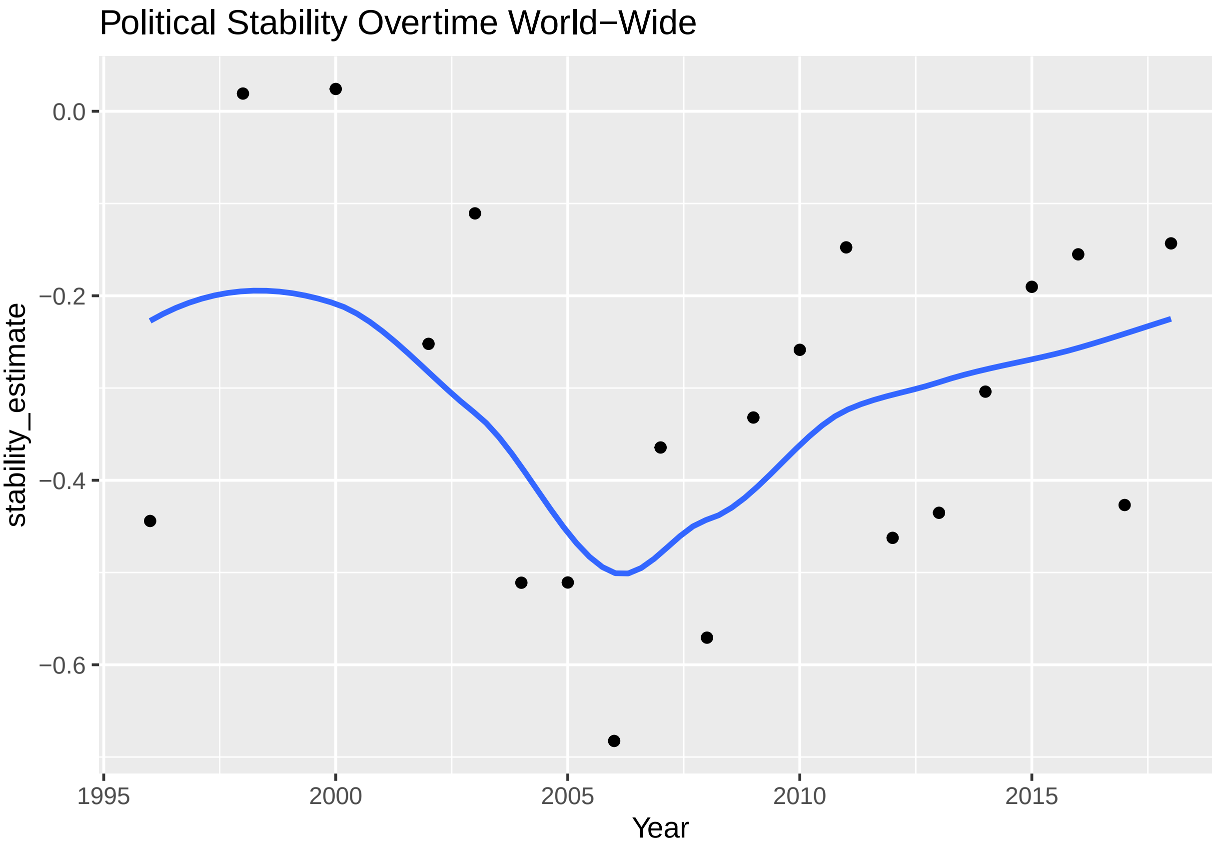
```
## # A tibble: 6 x 11
##   country code   year estimate     gdp population fuel_ex military_expend~
##   <chr>   <chr> <dbl>    <dbl>   <dbl>      <dbl>   <dbl>            <dbl>
## 1 Afghan~ AFG    2018  -2.75   1.95e10   37172386   10.5             1.01
## 2 Angola  AGO    2014  -0.333  1.46e11   26941779   96.2             4.70
## 3 Albania ALB    2012  -0.144  1.23e10    2900401   26.6             1.49
## 4 Albania ALB    2018   0.378  1.51e10    2866376    1.66            1.17
## 5 Argent~ ARG    2018   0.0192 5.20e11   44494502    4.29            0.745
## 6 Armenia ARM    2011  -0.0639 1.01e10    2876538    8.43            3.85
## # ... with 3 more variables: inflation <dbl>, lit_rate <dbl>,
## #   electric_access <dbl>
```

```
## [1] 495  11
```

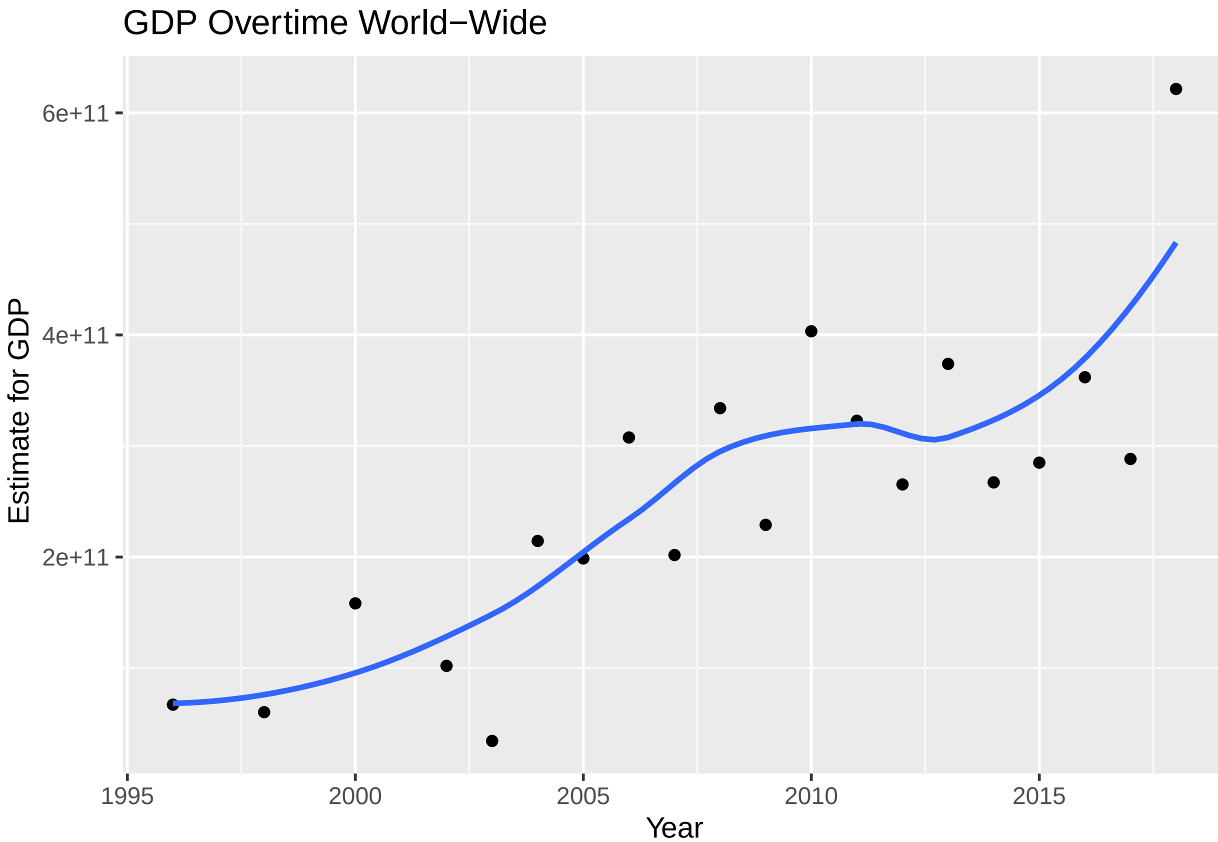
Further Investigation



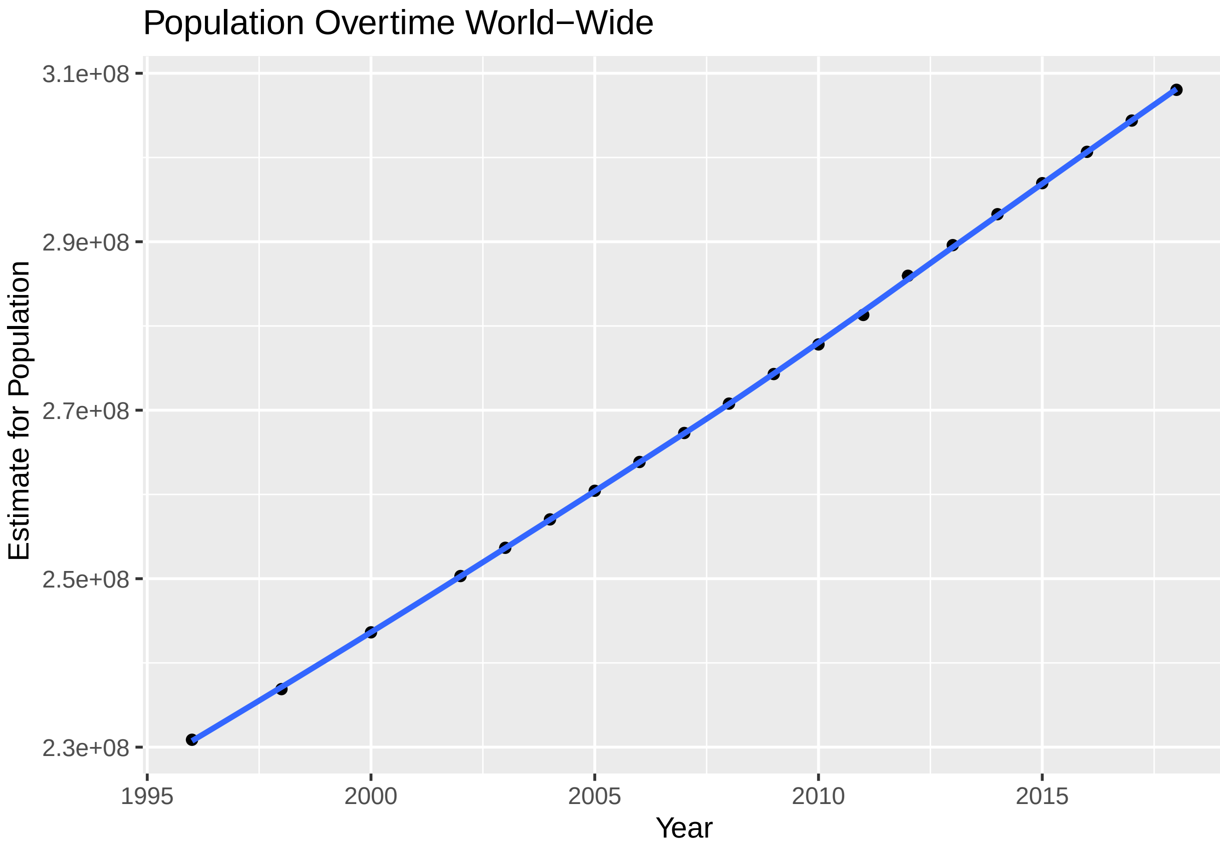
Further Investigation



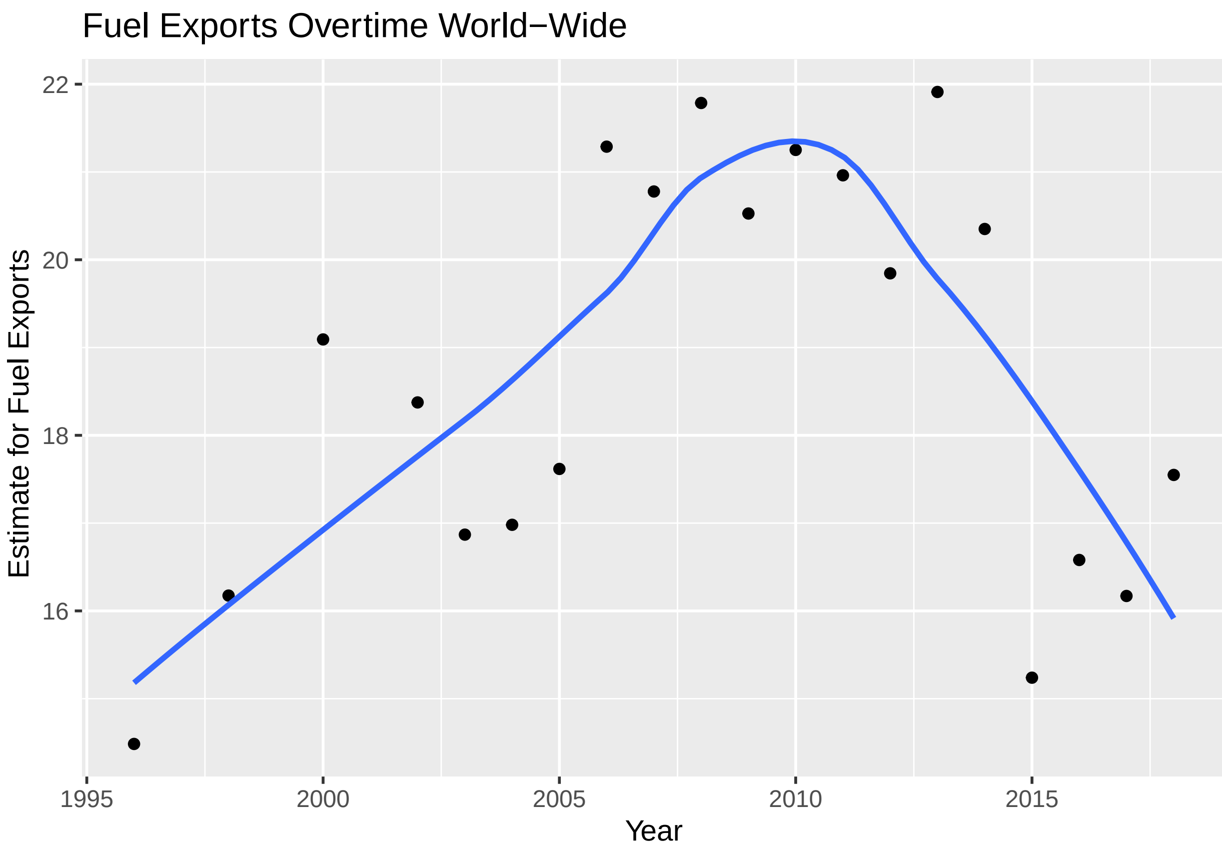
Further Investigation



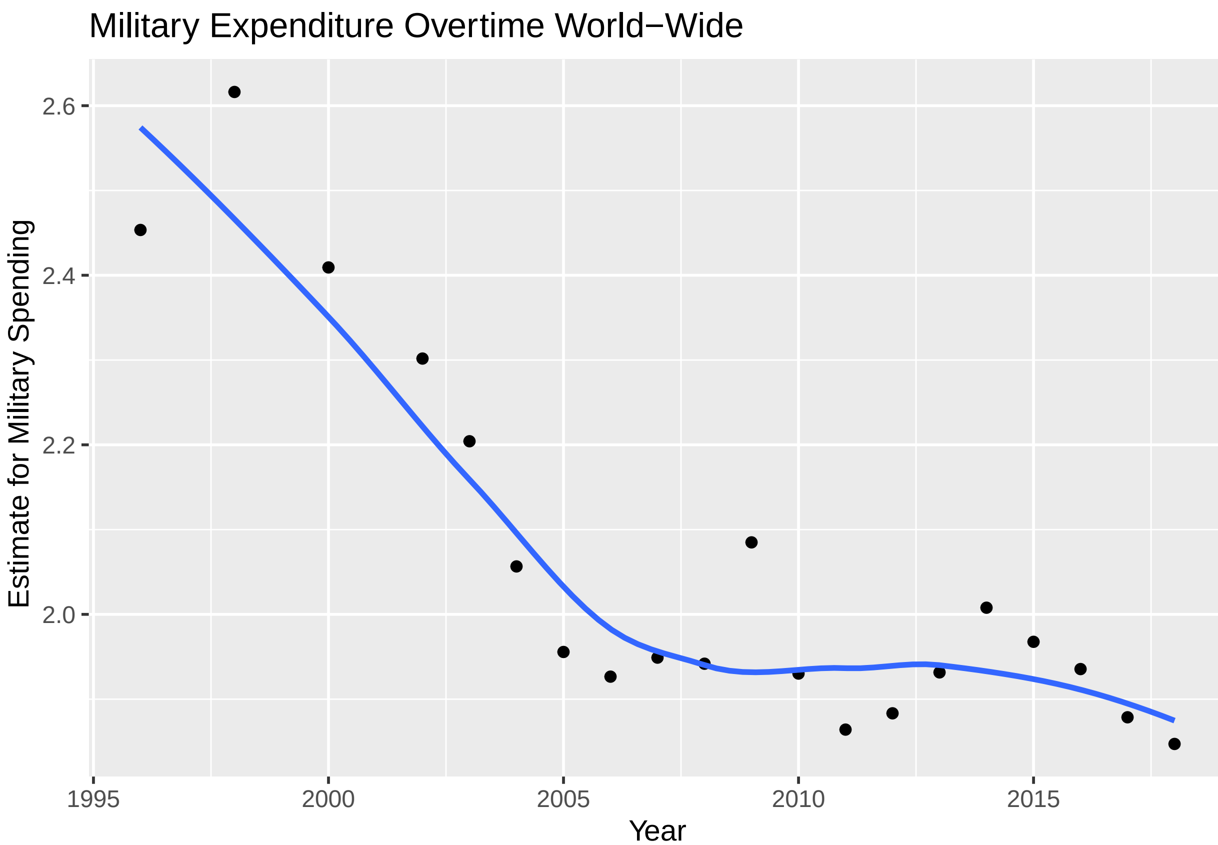
Further Investigation



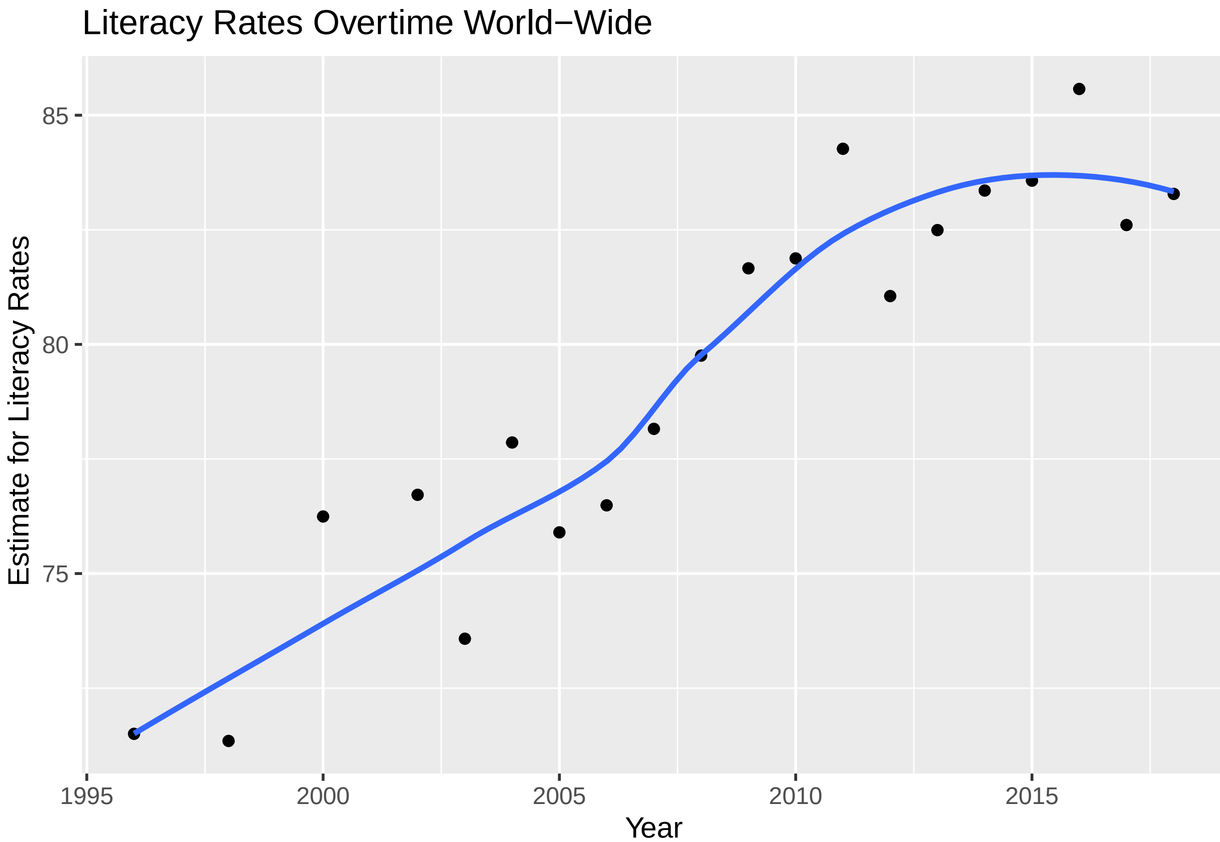
Further Investigation



Further Investigation



Further Investigation



Analysis

Variable Selection and Model Building

```
##
## Call:
## lm(formula = estimate ~ gdp + population + fuel_ex + military_expenditure +
##     inflation + lit_rate + electric_access, data = predictor_stability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98265 -0.48470  0.00826  0.51707  1.87188
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.078e+00  1.837e-01  -5.866 8.25e-09 ***
## gdp                   6.298e-14  5.949e-14   1.059   0.2902
## population           -1.440e-09  3.228e-10  -4.460 1.02e-05 ***
## fuel_ex              -3.654e-03  1.363e-03  -2.680   0.0076 **
## military_expenditure -1.496e-02  2.320e-02  -0.645   0.5195
## inflation            -3.326e-02  6.065e-03  -5.484 6.70e-08 ***
## lit_rate              1.713e-02  3.022e-03   5.668 2.47e-08 ***
## electric_access      -3.760e-03  2.001e-03  -1.879   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7665 on 487 degrees of freedom
## Multiple R-squared:  0.2196, Adjusted R-squared:  0.2083
## F-statistic: 19.57 on 7 and 487 DF,  p-value: < 2.2e-16
```

Variable Selection and Model Building with Lasso

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                                 s0
## (Intercept)          -1.103828e+00
## (Intercept)           .
## gdp                   .
## population           -1.863675e-09
## fuel_ex              -1.203555e-03
## military_expenditure  .
## inflation            -1.846982e-02
## lit_rate              1.124294e-02
## electric_access       .
```

```
##
## Call:
## lm(formula = estimate ~ gdp + military_expenditure + lit_rate +
##     electric_access, data = predictor_stability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05590 -0.56239  0.02311  0.53882  1.80241
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          -1.544e+00  1.811e-01  -8.524  < 2e-16 ***
## gdp                  -1.013e-13  4.577e-14  -2.212   0.0274 *
## military_expenditure -3.673e-02  2.210e-02  -1.662   0.0971 .
## lit_rate              1.966e-02  3.153e-03   6.235 9.78e-10 ***
## electric_access      -3.733e-03  2.083e-03  -1.792   0.0737 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8146 on 490 degrees of freedom
## Multiple R-squared:  0.1132, Adjusted R-squared:  0.1059
## F-statistic: 15.63 on 4 and 490 DF,  p-value: 4.802e-12
```

Variable Selection and Model Building with Step wise Variable Selection

```
##                                    Best Subsets Regression
## ----------------------------------------------------------------------------------------
## Model Index    Predictors
## ----------------------------------------------------------------------------------------
##      1         lit_rate
##      2         inflation lit_rate
##      3         population inflation lit_rate
##      4         population fuel_ex inflation lit_rate
##      5         population fuel_ex inflation lit_rate electric_access
##      6         gdp population fuel_ex inflation lit_rate electric_access
##      7         gdp population fuel_ex military_expenditure inflation lit_rate electric_access
## ----------------------------------------------------------------------------------------
##
##                                                      Subsets Regression Summary
## -------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)         AIC          SBIC           SBC          MSEP
## -------------------------------------------------------------------------------------------------
##   1        0.0908      0.0889      0.0813    76.3565    1215.0370    -190.2548    1227.6507    334.69
##   2        0.1438      0.1403      0.1311    45.2819    1187.3053    -217.8958    1204.1235    315.82
##   3        0.1875      0.1826      0.1685    19.9809    1163.3423    -241.5970    1184.3651    300.30
##   4        0.2112      0.2048      0.1885     7.1925    1150.6886    -254.0036    1175.9159    292.13
##   5        0.2170      0.2090      0.1911     5.5984    1149.0608    -255.5315    1178.4927    290.59
##   6        0.2189      0.2093      0.1823     6.4156    1149.8610    -254.6708    1183.4974    290.49
##   7        0.2196      0.2083      0.1778     8.0000    1151.4388    -253.0481    1189.2798    290.84
## -------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

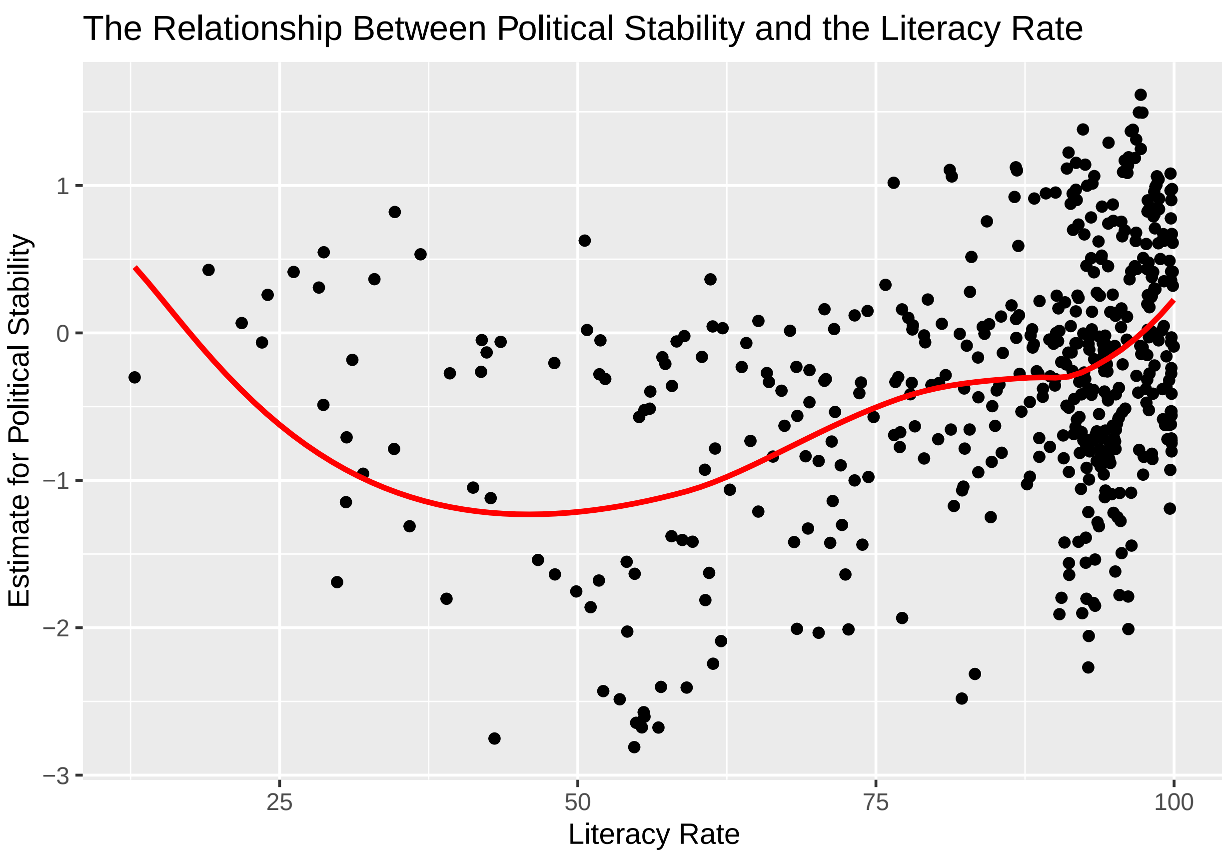
Optimal Linear Model

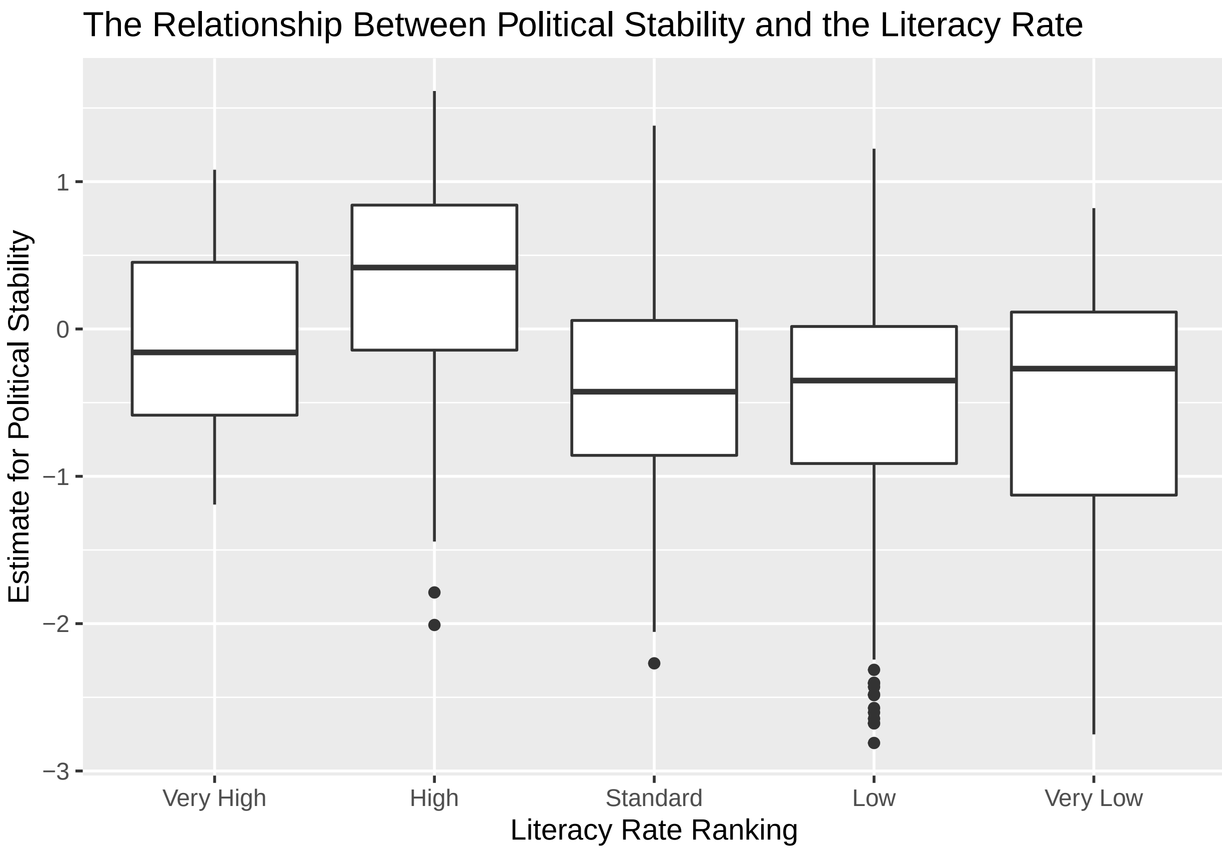
- After fitting a model with all possible predictors as well as running lasso variable selection and step wise variable selection, we find that our optimal model was found by our step wise variable selection method, based on adjusted R2. Our full linear model with all predictors offers an adjusted R2 of 0.2083. Our optimal lasso model offers an adjusted R2 of 0.1059. Lastly, our optimal step wise model with 5 predictors offers an adjusted R2 of 0.2090. This model surpasses the other adjusted R2, while offering fewer predictor variables than the initial full model. We will consider this our optimal model, containing predictors: population, fuel_ex, inflation, lit_rate, and electric_access.

- When we investigate our optimal model, we find that our only positive coefficient is the literacy rate. This is to say, for every unit increase in the literacy rate of a country, the political stability of that country can expect to increase by an average of 0.01774 units. All other coefficients are negative indicating that for every unit increase in the population, fuel exportation, inflation, and electric access; the political stability of that country can expect to decrease by the following coefficients in the table below on average. Additionally, we find that all of our coefficients are significant past the 0.05 alpha cut off, other than the predictor electric_access which produced a p-value of 0.0585. Again, we can see that our optimal model currently explains 20.9% of the variation in political stability across the globe.

```
##
## Call:
## lm(formula = estimate ~ population + fuel_ex + inflation + lit_rate +
##     electric_access, data = predictor_stability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00954 -0.47720  0.01616  0.51391  1.90392
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.141e+00  1.756e-01  -6.498 2.01e-10 ***
## population      -1.204e-09  2.347e-10  -5.130 4.19e-07 ***
## fuel_ex         -4.156e-03  1.219e-03  -3.410 0.000704 ***
## inflation       -3.320e-02  6.012e-03  -5.522 5.44e-08 ***
## lit_rate         1.774e-02  2.980e-03   5.953 5.03e-09 ***
## electric_access -3.784e-03  1.995e-03  -1.897 0.058474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7662 on 489 degrees of freedom
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.209
## F-statistic:  27.1 on 5 and 489 DF,  p-value: < 2.2e-16
```

- When we investigate the relationship between the literacy rate and the estimate for political stability we find an interesting relationship. The trend seems to follow something that resembles a parabolic function. As literacy rates remain low, the country maintains relatively high levels of political stability. As literacy rates increase, to where less than 50% of the country is literate, political stability in the country drops to an estimate below -1. Then as literacy rates increase past 50%, stability gradually increases on a similar relative path.



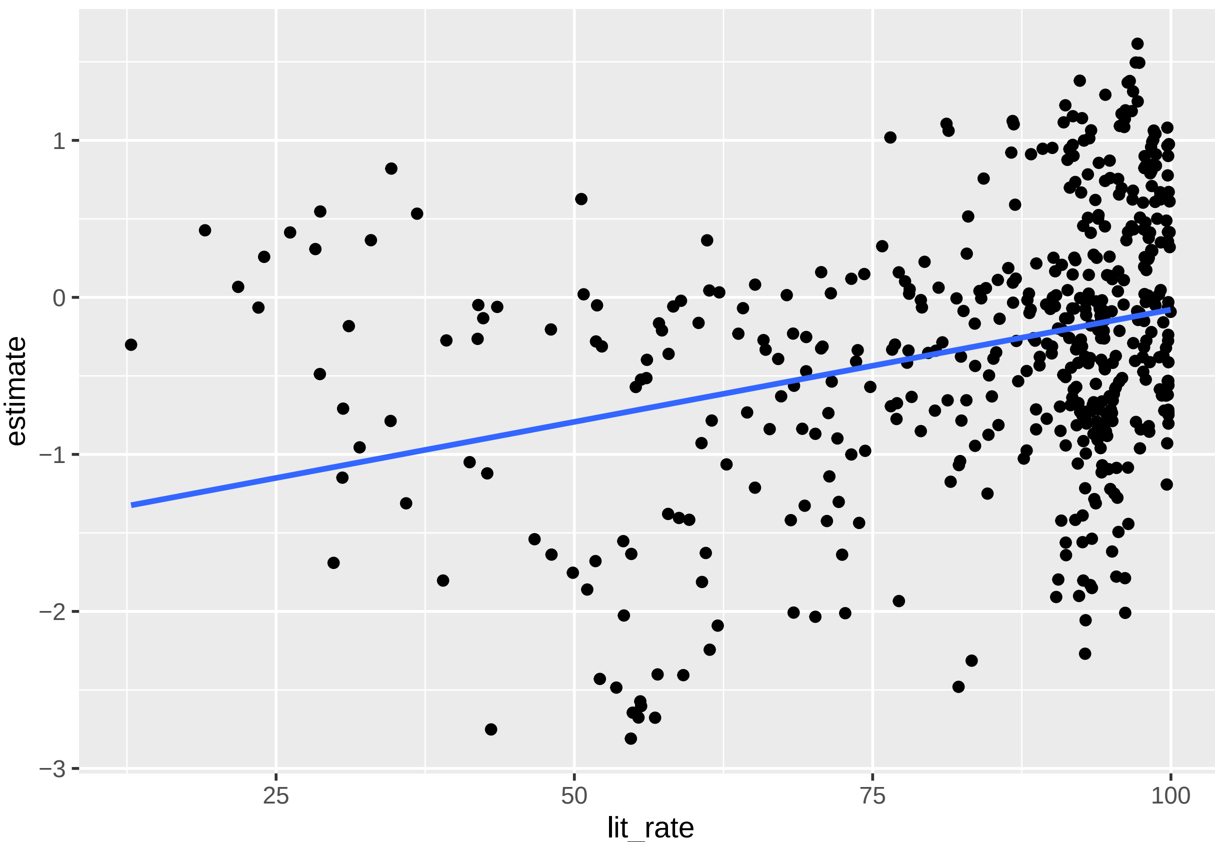

The Relationship Between Political Stability and the Literacy Rate
Estimate for Political Stability
1
0
−1
−2
−3
Very High
High
Standard
Low
Very Low
Literacy Rate Ranking

Lack of Fit Test

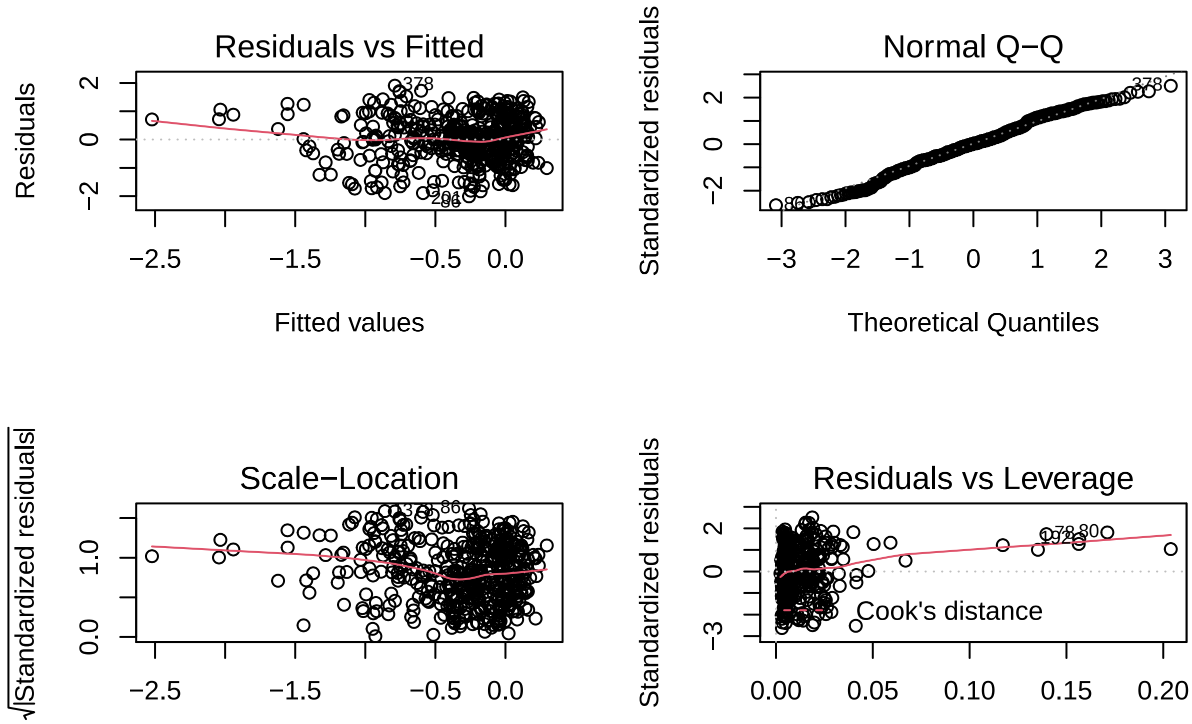
- Test assumption of linearity between political stability and literacy.
- With a small p-value we may have evidence against linearity. This need to be further investigated.

```
## Analysis of Variance Table
##
## Model 1: estimate ~ lit_rate
## Model 2: estimate ~ as.factor(lit_rate)
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1    493 333.35
## 2      2   0.01 491    333.34 148.11 0.006729 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linearity investigation



Plotting Residuals of Optimal Model



KNN

- Our first round when K = 3 gives a classification accuracy rate of the following.

```
##            knn.results
## Y.testing   Low Standard Very Low
##   High       70       11       12
##   Low        80       22       18
##   Standard   83       19       20
##   Very High  23        7        1
##   Very Low  102        9       12
```

```
## [1] 0.2269939
```

KNN

- A loops is then used to test our classification accuracy rate for all values of K from 1 to 20. We find that our optimal K with the highest classification accuracy rate is when K = 1 which gives a classification accuracy rate of the following.

```
## [1] 4
```

```
## [1] 0.2453988
```

```
write_csv(x = predictor_stability, file = "compiled_stability")
```

Findings

- Our optimal model explains 20.9% of the variation in political stability worldwide. Our final model included: population, fuel exports, inflation, literacy rates, and electric access as predictors. This produces the model: estimate ~ population + fuel_ex + inflation + lit_rate + electric_access. Our model results in the linear regression line: estimate = -1.141 - 1.204e-09 * population -4.156e0-03 * fuel_ex - 3.320e-02 * inflation + 1.774e-02 * lit_rate - 3.784e-03 * electric_access.

- Our predictor variable of literacy rate is arguably the most interesting for several reasons. We found that literacy was our only positive coefficient, the relationship between literacy and stability follows a "U" pattern, and it proved to be one of most significant predictor variables based on p-value. All of our other predictor variables were significant under the 0.05 alpha level, other than our electric access variable. Furthermore, we find that our KNN modeling was effectively able to classify approximately 30% of countries as having the correct political stability ranking.

- Apart from our modeling, we did uncover some interesting trends within our data. To note a few of the trends: political stability in the US has decreased overtime, we did not detect a relationship between GDP per capita and political stability, GDP per capita has increase dramatically overtime as has the population globally, fuel exports globally have been decreasing since 2010, military expenditure has also decreased dramatically since 1995, and lastly literacy rates have been on a steady upward climb since 1995 but began to plateau around 2012.

- As noted early, the reader should be cautioned against concluding that higher literacy rates lead to higher political stability for several reasons. First, that is not what the findings show. The relationship between literacy and political stability is shown to be more complex than a linear relationship. Secondly, as discussed previously, literacy does not produce political stability. The institutions in a country make an environment more or less hospitable to a literate populace which intern may make a country more or less hospitable to political stability.

Limitations

- One significant limitation is the random sample of countries that ends up in my compiled data frame after I remove NA values. The number of countries is cut in half as shown below. This could introduce a significant level of bias within my analysis as I would prefer to have kept all countries that were within the original stability data set. The data frame containing NA's was obviously not feasible for analysis and thus a sacrifice to the data set was made.

```
##   [1] "Afghanistan"            "Angola"                 "Albania"
##   [4] "Argentina"              "Armenia"                "Azerbaijan"
##   [7] "Burundi"                "Benin"                  "Burkina Faso"
##  [10] "Bangladesh"             "Bulgaria"               "Bahrain"
##  [13] "Bosnia and Herzegovina" "Belarus"                "Bolivia"
##  [16] "Brazil"                 "Brunei Darussalam"      "Botswana"
##  [19] "Chile"                  "China"                  "Cameroon"
##  [22] "Congo, Rep."            "Colombia"               "Costa Rica"
##  [25] "Cyprus"                 "Czech Republic"         "Dominican Republic"
##  [28] "Ecuador"                "Egypt, Arab Rep."       "Spain"
##  [31] "Estonia"                "Ethiopia"               "Fiji"
##  [34] "Georgia"                "Ghana"                  "Guinea"
##  [37] "Gambia, The"            "Greece"                 "Guatemala"
##  [40] "Guyana"                 "Honduras"               "Croatia"
##  [43] "Hungary"                "Indonesia"              "India"
##  [46] "Iran, Islamic Rep."     "Iraq"                   "Italy"
##  [49] "Jamaica"                "Jordan"                 "Kazakhstan"
##  [52] "Kenya"                  "Kyrgyz Republic"        "Cambodia"
##  [55] "Korea, Rep."            "Kuwait"                 "Lao PDR"
##  [58] "Lebanon"                "Sri Lanka"              "Lesotho"
##  [61] "Lithuania"              "Latvia"                 "Morocco"
##  [64] "Moldova"                "Madagascar"             "Mexico"
##  [67] "North Macedonia"        "Mali"                   "Malta"
##  [70] "Mongolia"               "Mozambique"             "Mauritius"
##  [73] "Malawi"                 "Malaysia"               "Namibia"
##  [76] "Niger"                  "Nigeria"                "Nicaragua"
##  [79] "Nepal"                  "Oman"                   "Pakistan"
##  [82] "Panama"                 "Peru"                   "Philippines"
##  [85] "Papua New Guinea"       "Poland"                 "Portugal"
##  [88] "Paraguay"               "Qatar"                  "Russian Federation"
##  [91] "Rwanda"                 "Saudi Arabia"           "Sudan"
##  [94] "Senegal"                "Singapore"              "El Salvador"
##  [97] "Slovenia"               "Seychelles"             "Syrian Arab Republic"
## [100] "Togo"                   "Thailand"               "Trinidad and Tobago"
## [103] "Tunisia"                "Turkey"                 "Tanzania"
## [106] "Uganda"                 "Ukraine"                "Uruguay"
## [109] "Venezuela, RB"          "Vietnam"                "Yemen, Rep."
## [112] "South Africa"           "Zambia"                 "Zimbabwe"
```

- An additional limitation is the number of predictors included. As we saw, our variable selection methods performed well however our optimal model only explains 20.9% of the the variation in political stability. For further research, additional predictor variables should be added to the compiled data set and the variable selection methods should be rerun in hopes of improving the adjusted R2 of the optimal model. Ideally, metrics such as institutional performance, colonial history, and immigration policies would greatly aid the analysis and improve the performance of the model. The difficulty comes from finding reliable forms of these metrics the would merged well the existing data frame. Ideally, a more robust data set would lead our variable selection methods on a more optimal path.

- A final limitation comes from the methods themselves. Our optimal model remained linear, yet additional variations could have provided a better fit. For example we discovered that the relationship between stability and literacy rates seems to follow a parabolic or "U" shape. Additionally, there a more robust machine learning models other than KNN that could have been employed in order to improve the overall classification accuracy of our final output. In further research, these avenues should certainly be investigated.

Conclusion

- To conclude, our optimal model explained 20.9% of variation in political stability globally. While this was disappointing, it highlights the difficulty of the research and the areas of specific improvement in future research. Our optimal model selected 5 predictors which included: population, fuel exports, inflation, literacy rates, and electric access. All were significant at the 0.05 cutoff other than electric access. Our KNN classification model accurately classified 30% of our observations correctly. Overall, the data compiled is valuable for future research and our methods highlight avenues of strength and weakness for further investigations.