Statistical Modelling & Computation Report

Spring 2023 Devyani Mardia, Michael Ryvin, Ryan Ali, Joseph Sheppard 954:567.02

1. Introduction

This project aims to implement Statistical Modelling concepts on our varieties of democracy data (source vdem website/GitHub). Vdem is a unique way of measuring democracies of different types. Our data includes various countries, years, continents, multiple indicators, and democratic indices. The 5 High-level democracy indices of interest are the Electoral democracy index, Liberal democracy index, Participatory democracy index, deliberative democracy index, and Egalitarian democracy index. Many indices are formed using these five indices, but as these are key in understanding the type of democracy within each country, it was essential to understand them in depth. It consists of N/A values throughout the data due to missing data. This directly leads us to investigate and analyze the type of missing values and apply statistical methods to impute these values to the maximum possible extent. As mentioned before, various indicators are present in this data, which led us to think about how to analyze this data and use tools like survival analysis to understand the time to event for a particular event/ or indicator in our data. We were mainly interested in analyzing how fast the leader in different continents(for all the countries that achieved in that particular continent) change with the number of years.

2. Data set

The data set we used for this project is the V-Dem dataset, which stands for the "Varieties of Democracy" dataset. The V-Dem dataset is a comprehensive database that provides information on various aspects of democracy and governance worldwide. The dataset is based on a unique measurement framework that considers the formal and informal elements of democratic governments and covers multiple indicators related to topics such as civil liberties, the rule of law, electoral processes, etc.

The V-Dem dataset consists of 4602 columns and 27555 rows or observations with 202 unique country names, data ranging from dates as early as the late 1700s to the most recent years. Each row of the dataset is structured as a country name, the year, and the various measurements of democracy of that particular country during that time. Out of the 4602 columns that measure multiple aspects of democracy, the V-Dem distinguishes between five high-level principles of democracy: electoral, liberal, participatory, deliberative, and egalitarian, and collects data to measure these principles.

3. Methodologies: Missing Data and Survival Analysis

The first goal of our project is to use missing value imputation techniques to fill in missing data in the V-Dem dataset. Similar to other large datasets, the V-Dem dataset suffers from the problem of missing data. For the five high-level democracy indices of interest, only about 69% have complete sets of row data, and about 4.6% have no values at all. The missing values can limit the accuracy and completeness of subsequent analyses. If we can use missing value imputation techniques, we will create a more complete and accurate dataset for subsequent studies.

First, we must identify the types of missing values to achieve our goal dataset. This includes conducting hypothesis testing to determine whether the missing value in the dataset is missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Depending on the type of missing data in the V-Dem dataset, we must select appropriate imputation techniques to fill in those missing values.

The second goal of our project is to conduct a survival analysis to determine how long it takes a country to change its leader after achieving independence. Our analysis will involve identifying the factors that may contribute to or hinder the achievement of this milestone, such as political institutions, cultural norms, and historical events.

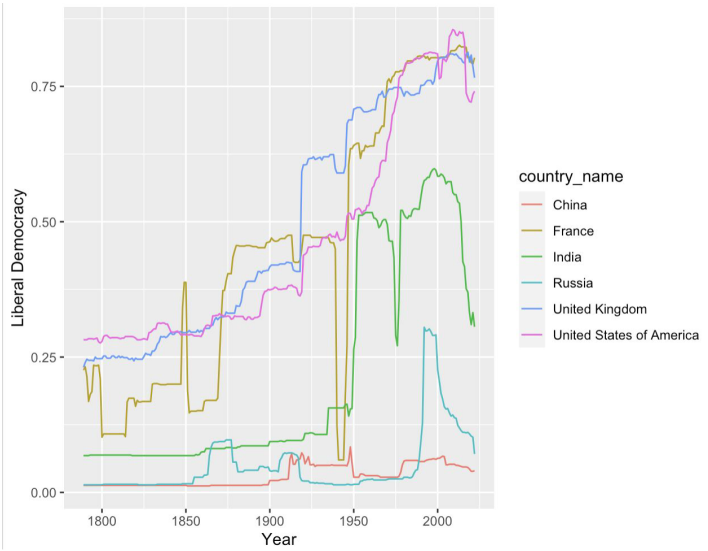
We used survival analysis, a statistical technique often used to analyze time-to-event data, such as the time from treatment until death. This analysis will identify the independent variables that may affect the time to the event. By conducting a survival analysis, we will be able to identify the factors that may contribute to or hinder the occurrence of the event of leadership change and gain a better understanding of the complex factors that shape political leadership transitions in newly independent countries.

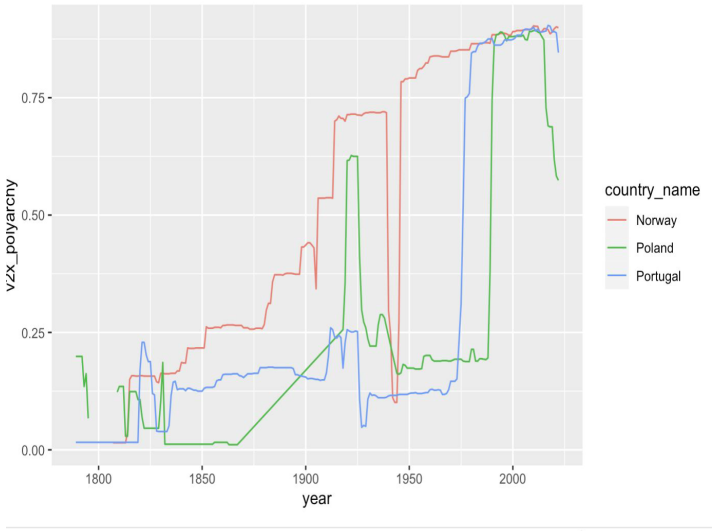
4. Pre-application visualizations/ Data Analysis

Conducting an exploratory data analysis (EDA) to better understand the data and identify potential issues or patterns can be helpful. Below are a few plots and visualization we created to understand the V-Dem dataset better.

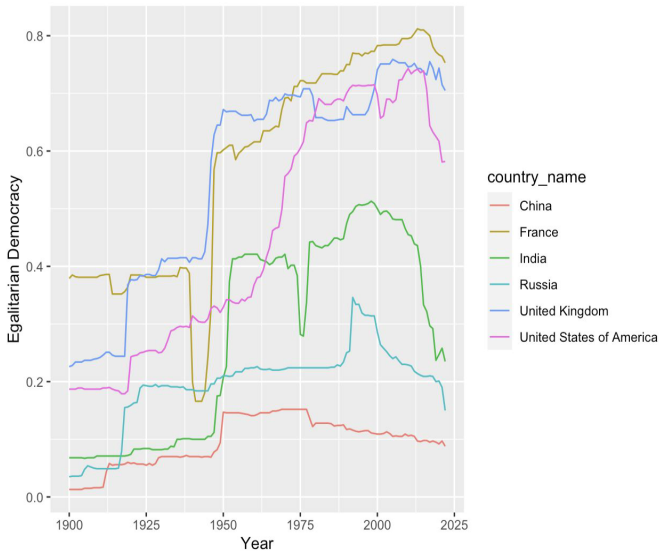
- Electoral Democracy Indices:
    - The figure on the left shows how the electoral democracy index of major countries has changed over time from approximately 1800 to 2022. Despite fluctuations in a few periods, the electoral democracy index of major countries has generally increased over the years.

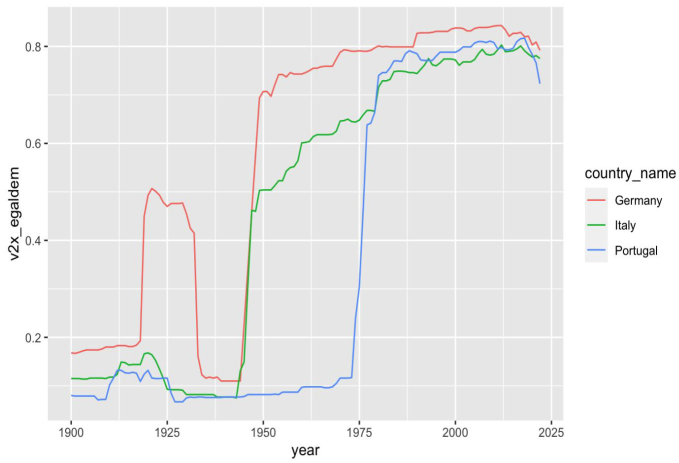
- The figure on the right shows the countries that have experienced the highest fluctuations of the electoral democracy index from 1800 to 2022. Interestingly, these countries have undergone several huge fluctuations before settling down on a score.



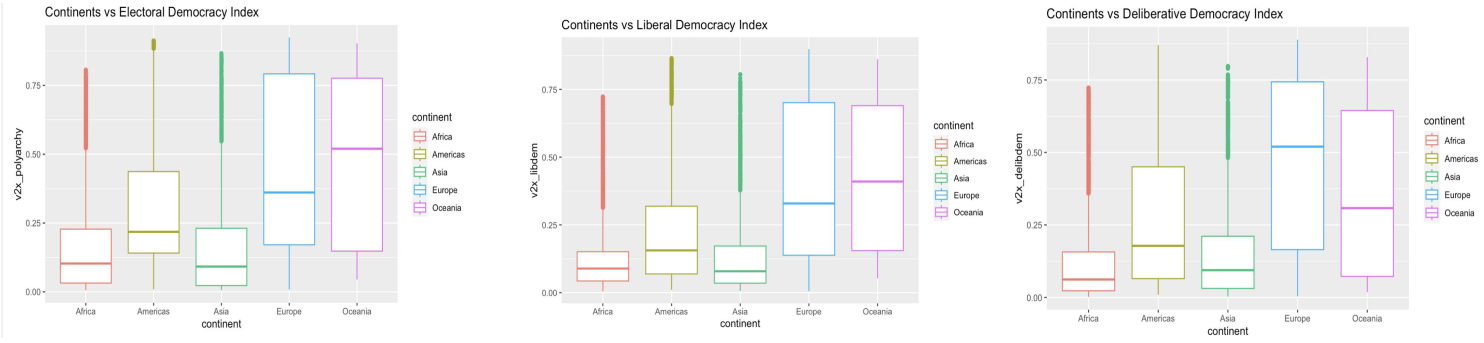


- Egalitarian Democracy Indices:
  - The figure on the left shows how the egalitarian democracy index of major countries has changed over time from 1800 to 2022. Despite fluctuations in certain periods, the egalitarian democracy index of major countries has generally increased over the years.
  - The figure on the right shows the countries that have experienced the most significant fluctuation in the egalitarian democracy index during the period of 1800 to 2022. Interestingly, these countries have only gone over one significant fluctuation before settling down on a score, unlike the electoral democracy index.





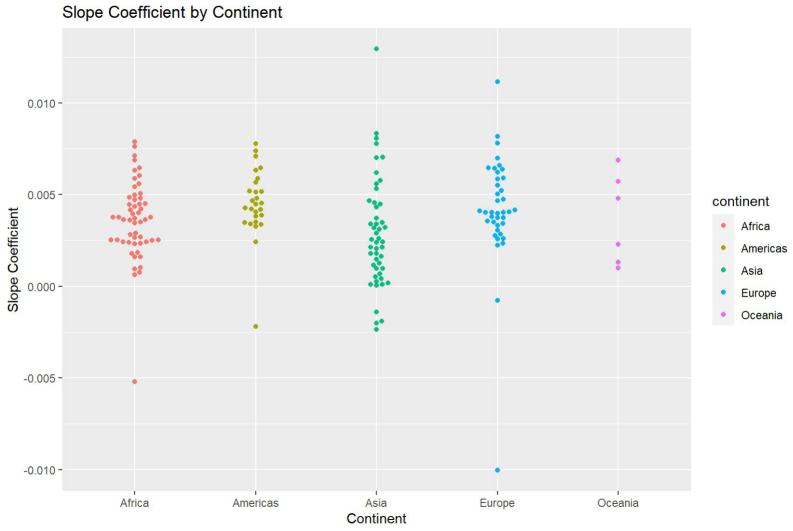
- Continent vs. Democracy Indices:
    - The five plots below show the distribution of the five high-level democracy indices among different continents.
    - Based on the plots we analyzed, we noticed a trend where European and Oceanian countries tend to have a higher democracy index than countries from America, Africa, and Asia. Additionally, we observed that the variation in the democracy index within the European and Oceanian countries is more significant than that of the other regions.
    - Also, there are many more outlier countries in Africa and Asia than on other continents. These extreme values indicate that countries in Africa and Asia either have a very high democracy index or a very low one.



- Increase in Democracy Score by Continent since 1900:
    - In the following beeswarm plot, we see how the democracy score of the different continents has changed. A coefficient value above 0 indicates that a particular

country has increased in democracy over the last century. A coefficient value below 0 indicates that a country has decreased in democracy over the last century.

- We see that the vast majority of countries have increased in democracy over the last century with only a few outliers. We also observe that countries in Asia and Africa have more variability than countries in Europe or the Americas.



5. Applications: Apply Missing Data Techniques and Survival Analysis.

Missing Data Analysis

To achieve our goal of using missing value imputation techniques to fill in missing data in the V-Dem dataset, we first want to visualize missing data before imputation. Visualizing the missing data patterns in the dataset before performing imputation can provide valuable insights into the nature and extent of the missing data. It helps inform the imputation technique we want to use for effective imputations and lesser errors.

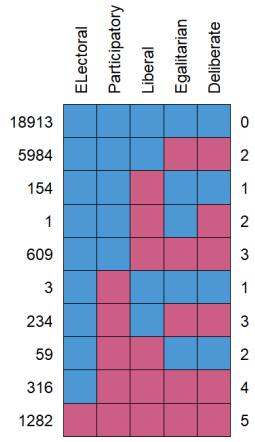
Missing Data Visualization:

The following two figures are the output of the function md.pattern() from the mice package, which provides a visualization of the missing data in the dataset. These figures focus specifically on the five high-level democracy indices and help to identify any patterns or trends in the missing data.

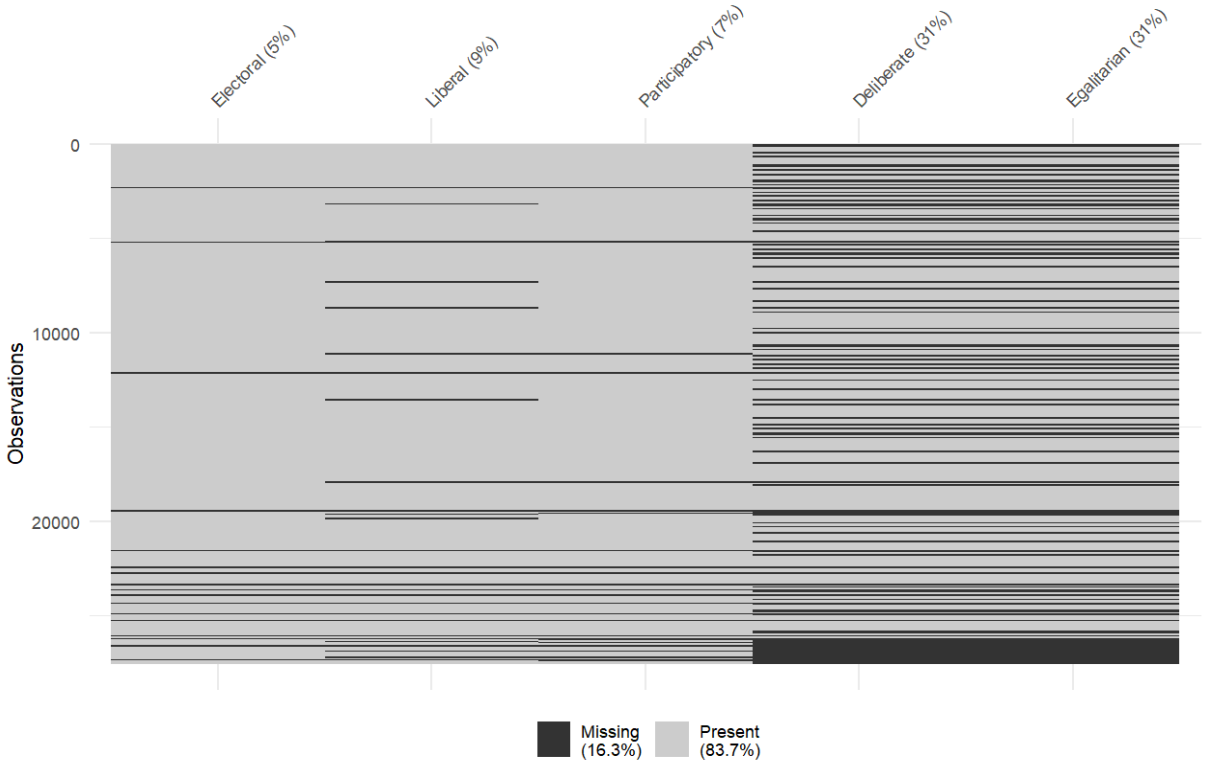
The output reveals that the V-Dem dataset contains 22,448 missing values. Of the 27,555 rows in the dataset, approximately 69%, or 18,913 rows, have a complete data set, while about 4.7%, or 1,282 rows, have no values at all. Among the five indices, the Egalitarian and Deliberative democracy indices exhibit the highest portion of missing values, with 8,425 and

8426 missing values each. These similarities in missing values prompted further investigation into the missing value pattern in these indices.

	ELectoral	Participatory	Liberal	Egalitarian	Deliberate	
18913	1	1	1	1	1	0
5984	1	1	1	0	0	2
154	1	1	0	1	1	1
1	1	1	0	1	0	2
609	1	1	0	0	0	3
3	1	0	1	1	1	1
234	1	0	1	0	0	3
59	1	0	0	1	1	2
316	1	0	0	0	0	4
1282	0	0	0	0	0	5
	1282	1894	2421	8425	8426	22448



The plot below, generated using the vis_miss() function from the visdat package, provides a visual representation of the missing data patterns in the dataset. Notably, the plot reveals a distinct pattern in the missing values of the Deliberative and Egalitarian democracy indices, where they have all missing values together, only with one exception. Upon further investigation, we discovered this pattern because these two indices have only data values from 1900 to 2022, whereas the other three indices span from 1800 to 2022.



Given that the missing values in the dataset are related to the time period, it appears likely that the missing data are not missing completely at random. However, in order to confirm this suspicion, we plan to conduct a hypothesis test. The following output is generated from the mcar_test() function from the naniar package. The function uses Little's (1988) test statistic to assess if data is missing completely at random (MCAR). The null hypothesis in this test is that the data are MCAR, and the test statistic is a chi-squared value. From the output, we see that the hypothesis test has a p-value of 0, which means we reject the null hypothesis and conclude that the data are not missing completely at random.

	statistic	df	p.value	missing.patterns
	<dbl>	<dbl>	<dbl>	<int>
1	4257.	22	0	10

Performing the Missing Data Imputation:

To account for the fact that the V-dem dataset is not missing completely at random, we plan to use multiple imputation methods to fill in the missing values. We will evaluate the performance of each imputation method by comparing the imputed data to the observed data

using appropriate metrics, such as the Root Mean Squared Error and Mean Absolute Error. By selecting the imputation method with the lowest error, we can create a more accurate and complete dataset for subsequent analyses.

To impute the missing values in the V-dem dataset, we have selected the Missing Value Imputation by Chained Equations (MICE) algorithm, which is available in the mice package. We will use three different imputation methods with MICE: Predictive Mean Matching, Random Forest, and Bayesian Linear Regression. For each process, we will generate five sets of imputed values using ten iterations.

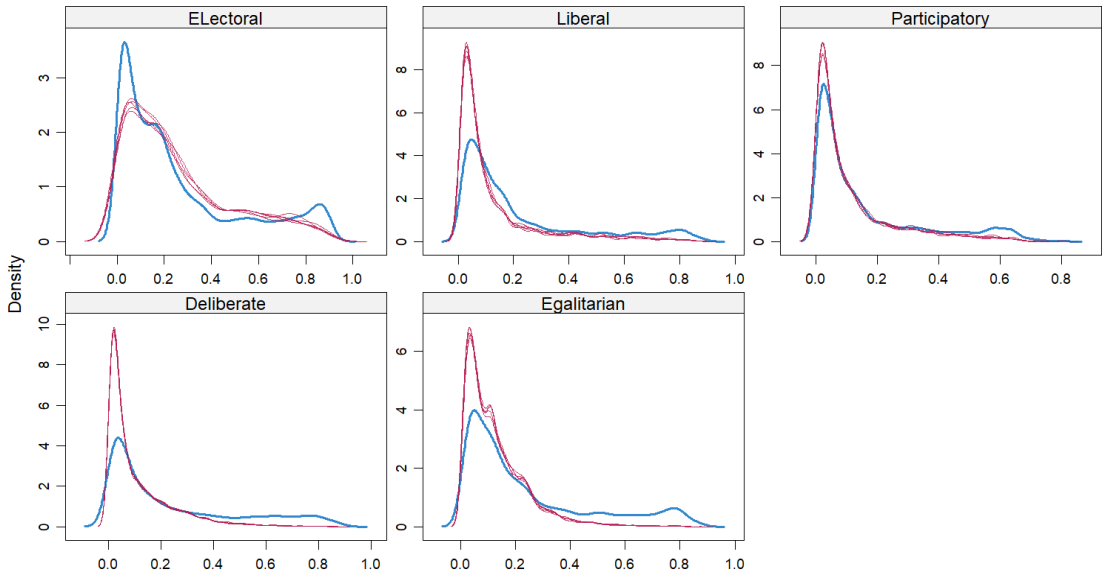
To calculate the error metrics, we took all the complete rows, filled in NA values with MAR assumption, and then used the three methods to impute those missing values. The metrics were then calculated by taking a difference between original and imputed values and applying the necessary operators(squared, mean, etc.) to get the required metric.

The following table shows the performance of each method based on Root Mean Squared Error and Mean Absolute Error. This table shows that the random forest method performs the best out of the three methods, followed by predictive mean matching and Bayesian linear regression in the last place.

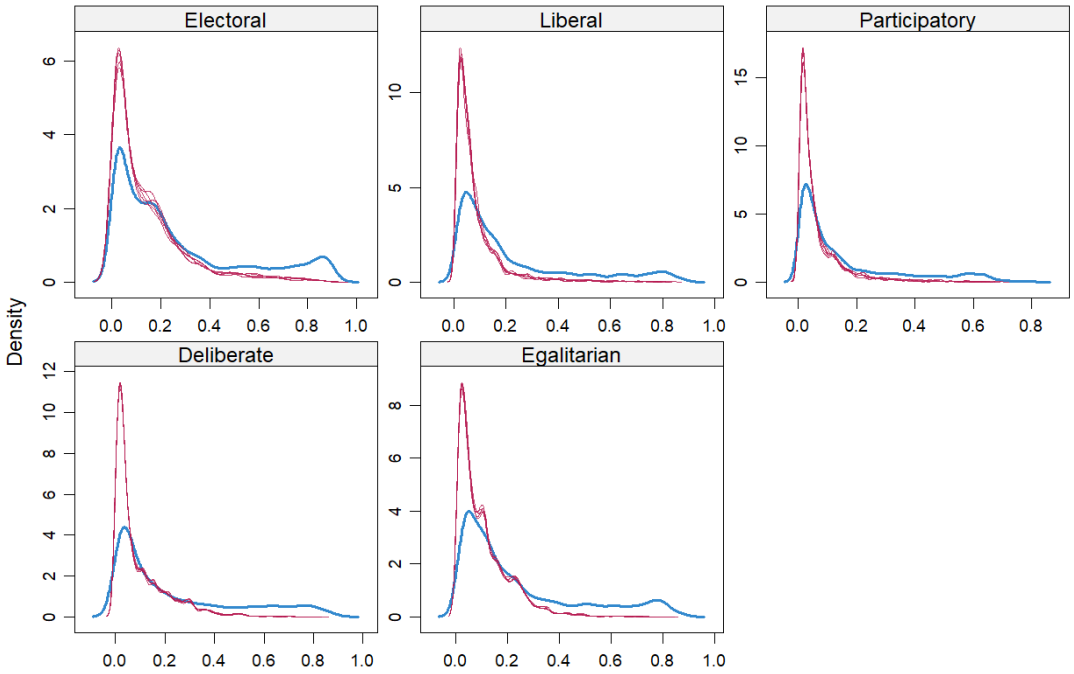
Imputation Method	Sum of Squared Error	Mean Absolute Error
Predictive Mean Matching	178	0.09
Random Forest	156	0.07
Bayesian Linear Regression	207	0.1

The following are density plots from the three imputation methods, and the plots compare the density of observed data with the ones of imputed data. We expect a higher accuracy model to have similar distributions between the original and imputed data.

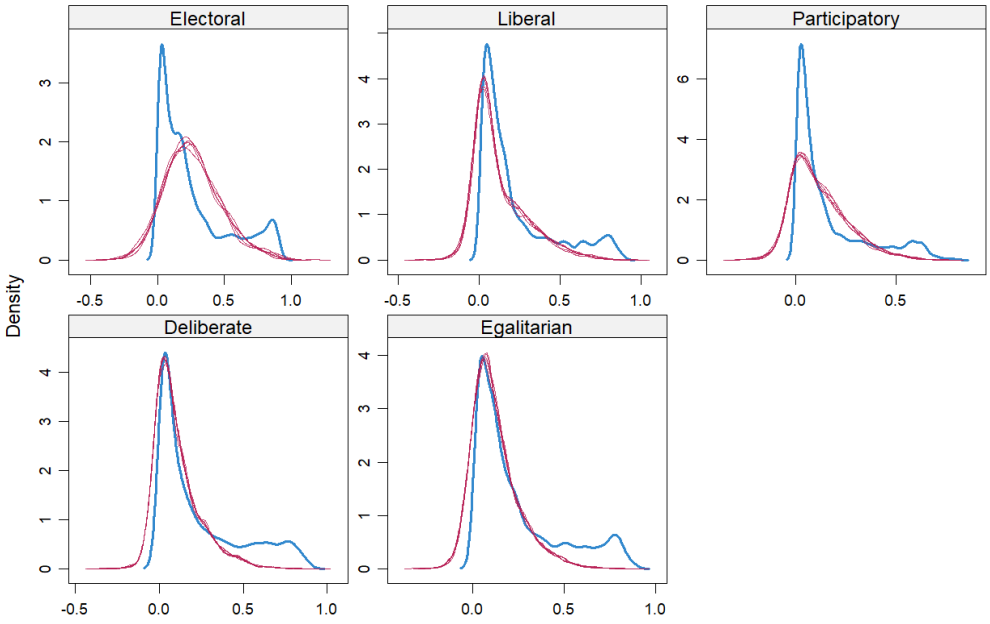
Predictive Mean Matching



Random Forest



Bayesian Linear Regression



Although all models are performing well, the PMM method has a better accuracy in terms of the distribution with random forest not far away. The Bayesian linear regression model isn't good at predicting the mean of distribution for 3 of the variables. However, it almost precisely matches the distribution for deliberate and egalitarian democracy, which is interesting.
To conclude, based on the MAE and the plots, we imputed the values using the Random Forest method.

Survival Analysis:

In this section, we aim to conduct a survival analysis to investigate the time a country takes to change its head of state. In many cases, the head of state will be the central ruler of the country, such as the President of the United States, while in other cases, the head of state will serve merely as a figurehead, such as the King of the United Kingdom. In this analysis, we will be performing a test not just to see how long heads of state last in general but also how long the first head of state of each country lasts. By conducting this analysis, we can gain insights into the complex factors that shape leadership transitions in newly independent nations across continents. But before we can start our investigation, some data cleaning is required.

Survival Analysis Data Cleaning:

We first categorized the countries by continent. This allows us to investigate potential differences in leadership transitions between countries of different regions. Next, to ensure that our analysis is focused on relevant countries, we subsetted the data to include only countries that

gained independence within the time interval covered by our dataset. This meant excluding countries that were already independent in 1789. We also did not want to include countries that no longer existed, so the data was subsetted not to have any nations that did not exist as independent states in 2022.

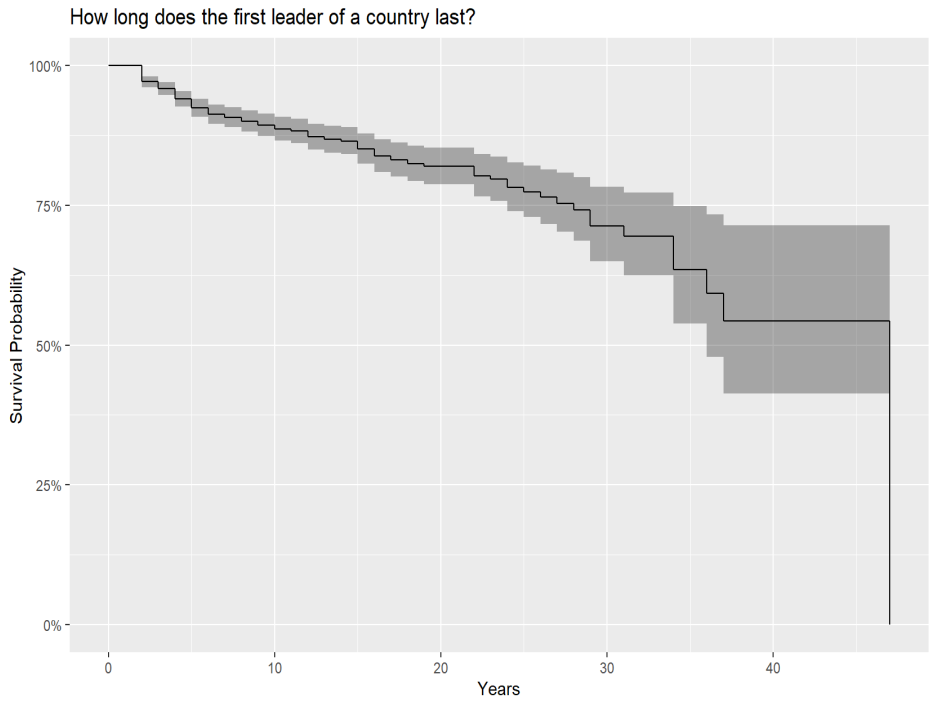
We then had to add a new variable to help identify each country's first year of independence in the dataset. This would allow us to get the years since independence for each country at a certain year. This year's independence variable would be used as the time variable in the survival analysis of newly independent countries. A new variable was also created to determine whether the current head of state was the first head of state, allowing for a more straightforward calculation of when the head of state changed. This was done by viewing the string in the Head of State variable.

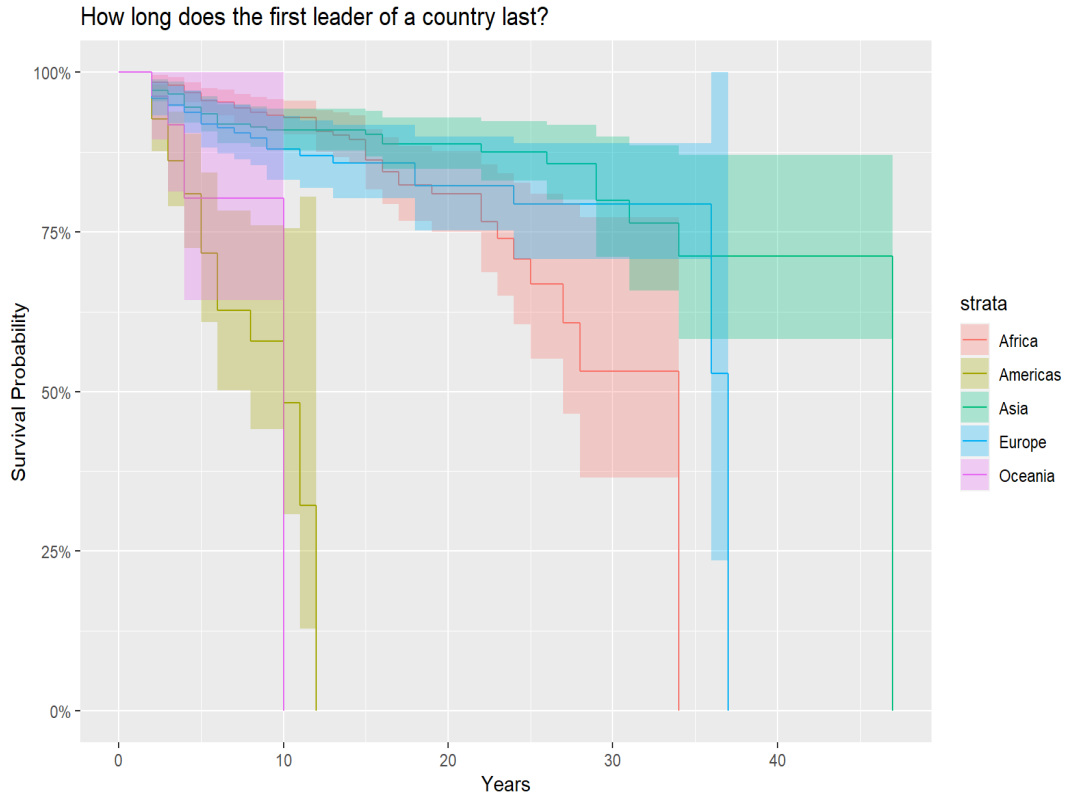
In the case of all heads of state, irrespective of the independence date, the 1789 restriction was removed, and we instead did calculations to find the years since the last head of state. This was again done by analyzing the strings. In this case, we did not care which countries were involved; we would only study the continents later.

Survival Analysis Visualization

To analyze how long the first Heads of State lasted, we created a survival object, where the first argument is the years since independence, and the second argument is whether or not they have changed their head of state yet. We used this survival object to create a Kaplan-Meier survival curve. In total, there were 148 events in this survival object, one for each country remaining in the dataset after the cleaning.

The following two plots are the Kaplan-Meier survival curve that shows how long the first Heads of State last. As we can see from the first plot, around 13% of the world's first heads of state will last about ten years, and around 80% of world heads of state will last about 20 years. The second plot shows that the first head of state from the Americas and Oceania generally lasts less time than the first head of state from Asia, Africa, or Europe.





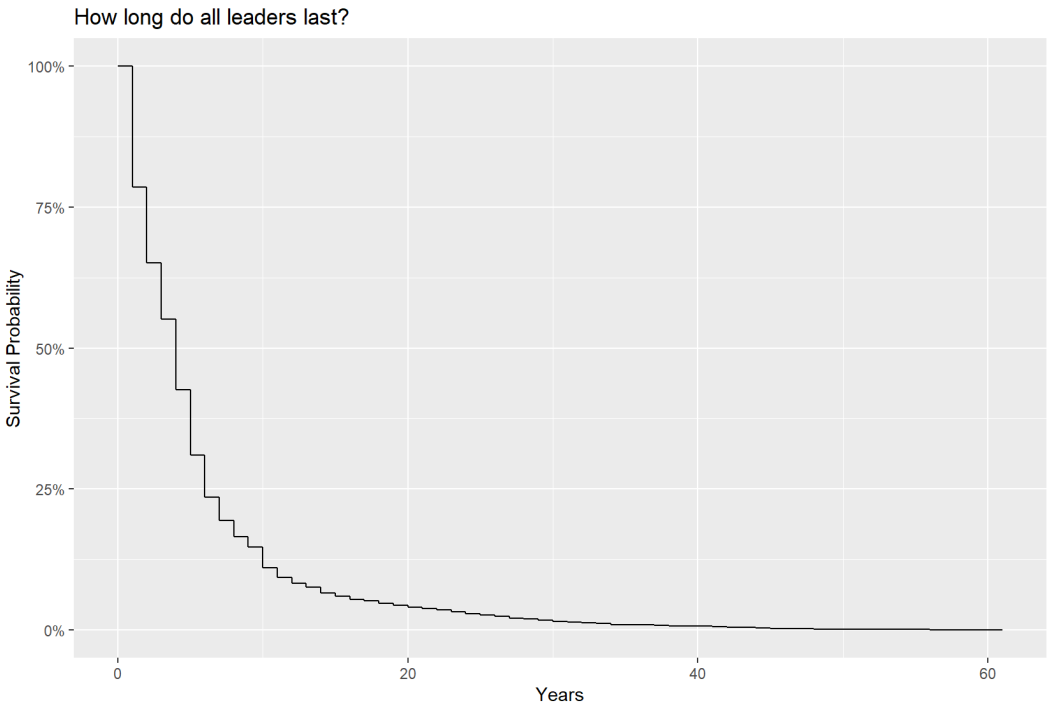
The theory behind this lies with the general government structures of every continent. The Americas and Oceania generally have had democratic governments upon independence, and thus the leaders are more likely to change more often. Africa has also generally had democratic governments upon independence, but these have been far less stable and far more prone to

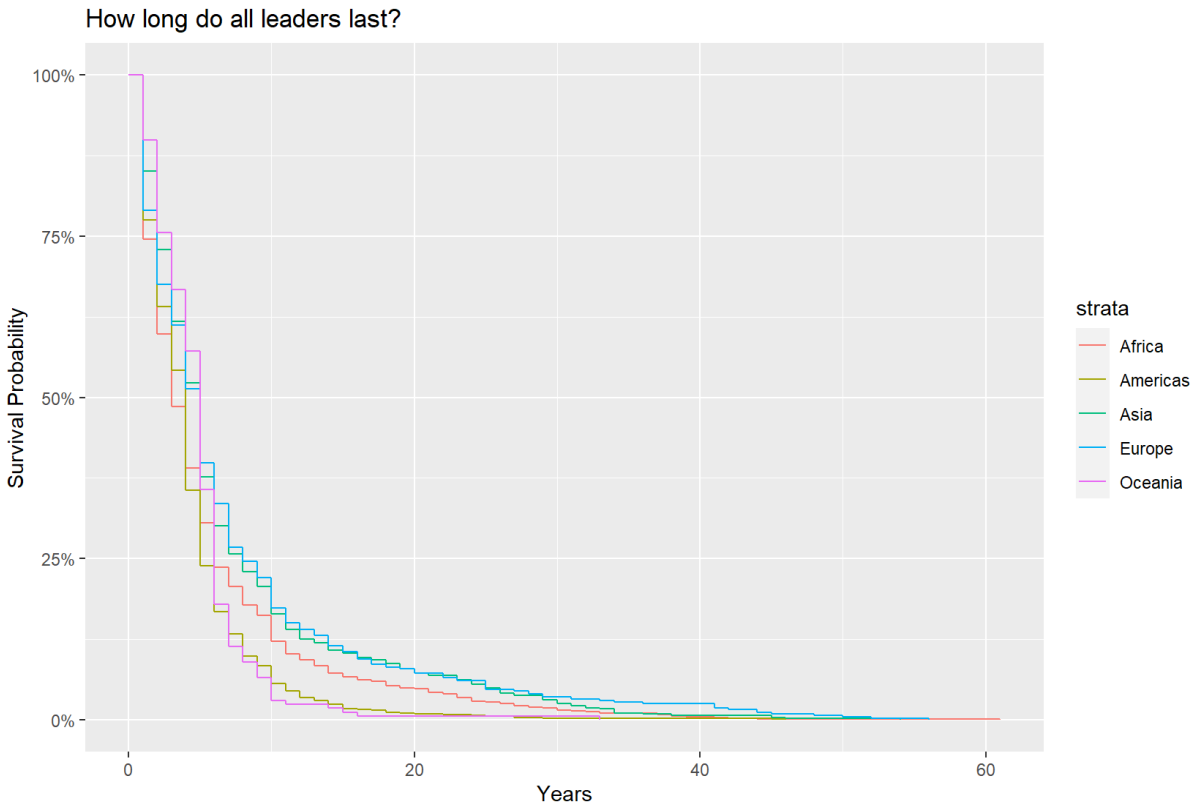
dictatorships, thus raising the term lengths. While Europe today tends to have democratic governments, many of the countries when they first gained independence were less democratic. Even so, regardless of the democratic nature of Europe, it is also home to many monarchs. While the Heads of Government may change over often, a Head of State that is a monarch generally reigns until their death, which can be a very long time. Asia faces a similar issue with Europe of having monarchies, but also is a generally less democratic continent, and so has more dictators upon independence that contribute to this high length of service for the first head of state.

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
continentAmericas  1.77554   5.90344  0.25069  7.083 1.41e-12 ***
continentAsia     -0.33620   0.71448  0.22521 -1.493 0.135483
continentEurope    0.05287   1.05429  0.23914  0.221 0.825033
continentOceania   1.45327   4.27709  0.43624  3.331 0.000864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
continentAmericas    5.9034     0.1694    3.6118     9.649
continentAsia        0.7145     1.3996    0.4595     1.111
continentEurope      1.0543     0.9485    0.6598     1.685
continentOceania     4.2771     0.2338    1.8190    10.057

Concordance= 0.617  (se = 0.029 )
Likelihood ratio test= 55.9  on 4 df,   p=2e-11
Wald test            = 72.12  on 4 df,   p=8e-15
Score (logrank) test = 93.41  on 4 df,   p=<2e-16
```

There only appears to be a statistically significant difference between Africa and Americas/Oceania, which makes sense given the graphs. This statistically significant difference also exists for Europe/Asia with the Americas/Oceania. Although Oceania does not have many data points, the difference is large enough between Africa and Oceania to deem it significant.

We also wanted to analyze how long heads of state generally last and not focus on the first one. For this, there were 3498 events, corresponding to the 3498 left in the dataset after the cleaning. The following two plots are the Kaplan-Meier survival curve showing how long the Heads of State last. As we can see from the first plot, around 50% of the world’s heads of state will last about 4 to 5 years, and around 13% of world heads of state will last about ten years. From the second plot, we observe that the head of state from Europe and Asia generally last longer than the head of state from the Americas.





These numbers are to be expected, as many countries in the world have either democratic or highly volatile dictatorships that cause most leaders not to last over ten years. However, the leaders that do last over ten years tend to last quite a while, as this subset is narrowed to either

monarchs or dictators that hold a firm grip over their countries. The continental relationships are similar to the ones described for first heads of state, as the general trends described can apply to a more extended period of time.

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
continentAmericas  0.17814   1.19499  0.04211  4.231 2.33e-05 ***
continentAsia     -0.22552   0.79810  0.05072 -4.446 8.75e-06 ***
continentEurope   -0.25421   0.77553  0.05597 -4.542 5.57e-06 ***
continentOceania   0.01473   1.01484  0.08286  0.178    0.859
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
continentAmericas    1.1950     0.8368    1.1003    1.2978
continentAsia        0.7981     1.2530    0.7226    0.8815
continentEurope      0.7755     1.2894    0.6950    0.8654
continentOceania     1.0148     0.9854    0.8627    1.1938

Concordance= 0.538  (se = 0.006 )
Likelihood ratio test= 91.69  on 4 df,   p=<2e-16
Wald test            = 90.47  on 4 df,   p=<2e-16
Score (logrank) test = 91.33  on 4 df,   p=<2e-16
```

In this case, because far more events occur, there are statistically significant differences between Africa and three other continents. The only exception is Oceania, where there is more overlapping with short-serving heads of state than in the first heads of state analysis.

6. Conclusion & Future Scope

Our project first aimed to address the problem of missing data in the V-Dem dataset by utilizing missing value imputation techniques. Through various visualizations and hypothesis testing, we determined that the missing data were not missing at random. This led us to select appropriate imputation techniques, including missing data imputation by chained equation using three different methods: PMM, Random Forest, and Bayesian Linear Regression. Our findings indicated that the random forest method performed the best among the three methods for imputing the missing data. By using missing value imputation techniques, we created a more complete and accurate dataset for subsequent analyses, ultimately enhancing the validity and reliability of our research findings.

Furthermore, our project also aimed to analyze the time it takes for a country to change its leader after achieving independence and throughout its history, using survival analysis to accomplish this goal. We created a Kaplan-Meier survival curve to analyze how long the first Heads of State and all Heads of State last in general. The results show that Heads of State generally last shorter in the Americas and Oceania than in Asia, Africa, and Europe.

Furthermore, around 50% of the world's heads of state tend to last about 4 to 5 years, and only about 13% tend to last at least ten years.

In a future project, we hope to combine the survival analysis and missing data imputation into a grander application by seeing how many years after independence it takes for a country to improve its democracy indexes by a certain percentage. By using missing data imputation to estimate the missing values, we could perform a survival analysis with more data points, albeit imperfect, due to the minor errors in missing data imputation.

7. Code

```
## For Missing data analysis and imputations:


library(vdemdata)
library(mice)
library(VIM)
library(naniar)
library(UpSetR)
library(tidyr)
library(missMethods)
data <- vdem




High_Level_Indices =
data[c("v2x_polyarchy","v2x_libdem","v2x_partipdem","v2x_delibdem","v2x_egaldem")]
colnames(High_Level_Indices) = c("Electoral","Liberal","Participatory","Deliberate","Egalitarian")
### v2x_polyarchy has 1282 Missing Values
### v2x_libdem has 2421
### v2x_partipdem has 1894
### v2x_delibdem has 8426
### v2x_egaldem has 8425


### Visualizing the Missing data
patt <- md.pattern(High_Level_Indices,rotate.names = TRUE)

marginplot(High_Level_Indices[c(2,5)])

vis_miss(High_Level_Indices)

gg_miss_upset(High_Level_Indices)
```

```
gg_miss_span(High_Level_Indices,Electoral,span_every = 1000)


### Testing the type of Missingness
mcar_test(High_Level_Indices)



### Imputing the values


Testing <- High_Level_Indices %>%
  drop_na()
Training <- delete_MCAR(Training,0.3,"Egalitarian")
sum(is.na(Training$Liberal))

?mice

# Run mice with different methods
imp1 <- mice(Training, method = "pmm", m = 5, maxit = 10)
imp2 <- mice(Training, method = "rf", m = 5, maxit = 10)
imp3 <- mice(Training, method = "norm", m = 5, maxit = 10)

# Completing the dataset using of the imputation
Training_1 <- complete(imp1,1)
Training_2 <- complete(imp2,1)
Training_3 <- complete(imp3,1)

# Calculating the Errors ==> Sum of Squared Error
sum((Training_1$Electoral-Testing$Electoral)^2+(Training_1$Electoral-Testing$Electoral)^2+(Tr
aining_1$Deliberate-Testing$Deliberate)^2+(Training_1$Participatory-Testing$Participatory)^2+(
Training_1$Liberal-Testing$Liberal)^2+(Training_1$Egalitarian-Testing$Egalitarian)^2)
sum((Training_2$Electoral-Testing$Electoral)^2+(Training_2$Electoral-Testing$Electoral)^2+(Tr
aining_2$Deliberate-Testing$Deliberate)^2+(Training_2$Participatory-Testing$Participatory)^2+(
Training_2$Liberal-Testing$Liberal)^2+(Training_3$Egalitarian-Testing$Egalitarian)^2)
sum((Training_3$Electoral-Testing$Electoral)^2+(Training_3$Electoral-Testing$Electoral)^2+(Tr
aining_3$Deliberate-Testing$Deliberate)^2+(Training_3$Participatory-Testing$Participatory)^2+(
Training_3$Liberal-Testing$Liberal)^2+(Training_3$Egalitarian-Testing$Egalitarian)^2)

# Calculating the Errors ==> Mean absolute error
mean(abs(Training_1$Electoral-Testing$Electoral)+abs(Training_1$Electoral-Testing$Electoral)
+abs(Training_1$Deliberate-Testing$Deliberate)+abs(Training_1$Participatory-Testing$Participa
tory)+abs(Training_1$Liberal-Testing$Liberal)+abs(Training_1$Egalitarian-Testing$Egalitarian))
```

```
mean(abs(Training_2$Electoral-Testing$Electoral)+abs(Training_2$Electoral-Testing$Electoral)
+abs(Training_2$Deliberate-Testing$Deliberate)+abs(Training_2$Participatory-Testing$Participa
tory)+abs(Training_2$Liberal-Testing$Liberal)+abs(Training_3$Egalitarian-Testing$Egalitarian))
mean(abs(Training_3$Electoral-Testing$Electoral)+abs(Training_3$Electoral-Testing$Electoral)
+abs(Training_3$Deliberate-Testing$Deliberate)+abs(Training_3$Participatory-Testing$Participa
tory)+abs(Training_3$Liberal-Testing$Liberal)+abs(Training_3$Egalitarian-Testing$Egalitarian))


# Now comparing the methods with plots
imp1 <- mice(High_Level_Indices, method = "pmm", m = 5, maxit = 10)
imp2 <- mice(High_Level_Indices, method = "rf", m = 5, maxit = 10)
imp3 <- mice(High_Level_Indices, method = "norm", m = 5, maxit = 10)


### density plot
densityplot(imp3)


### Strip plot
stripplot(imp3)

### Predictive model


model_fit <- with(data=imp3, exp=lm(Electoral ~ Liberal + Participatory + Deliberate +
Egalitarian))
# Pool the results
model_summary <- summary(pool(model_fit))

# Display the results
print(model_summary)



## Code for survival analysis


library(vdemdata)
library(dplyr)
library(countrycode)
library(ggplot2)
library(ggfortify)
library(survival)
library(tidyr)
```

```
data <- vdem
data$continent <- countrycode(data$country_text_id, "iso3c", "continent")

################################################################################
####
## SURVIVAL ANALYSIS DATA CLEANING
################################################################################
####

country_list <- unique(data$country_name[data$v2svindep == 1 & data$year == 2022])
country_notlist <- unique(data$country_name[data$v2svindep == 1 & data$year == 1789])

subset_data <- data[data$country_name %in% country_list & !(data$country_name %in%
country_notlist), ]

subset_data <- subset_data %>%
  group_by(country_name) %>%
  mutate(lastYearDependent = ifelse(any(v2svindep != 1),
                                max(year[v2svindep != 1]),
                                min(year[v2svindep == 1]) - 1),
         firstYearIndependent = ifelse(year > lastYearDependent,
                                  min(year[year > lastYearDependent]),
                                  NA))

subset_data$lastYearDependent = subset_data$firstYearIndependent - 1

subset_data$yearsSinceIndependence = subset_data$year - subset_data$lastYearDependent

unique(subset_data$country_name)

subset_data <- subset(subset_data, !is.na(yearsSinceIndependence))
subset_data$continent

noNA_data <- subset_data[, c("continent", names(subset_data)[colSums(is.na(subset_data)) ==
0])]

# Create a vector of column names to exclude
exclude_cols <- c("project", "historical", "codingstart", "codingend",
                  "codingstart_contemp", "codingend_contemp", "gap_index",
                  "COWcode", "v2xcl_rol", "v2xcl_rol_codelow", "v2xcl_rol_codehigh",
                  "v2xcl_rol_sd", "v2xeg_eqprotec", "v2xeg_eqprotec_codelow",
                  "v2xeg_eqprotec_codehigh", "v2xeg_eqprotec_sd", "v2xeg_eqaccess",
                  "v2xeg_eqaccess_codelow", "v2xeg_eqaccess_codehigh", "v2xeg_eqaccess_sd",
```

"v2elreggov", "v2ellocgov", "v2exrmhsol_1", "v2exrmhsol_2", "v2exrmhsol_3",
"v2exrmhsol_4", "v2exrmhsol_5", "v2exrmhsol_6", "v2exrmhsol_7",
"v2exrmhsol_nr", "v2ex_legconhog", "v2ex_legconhos", "v2juaccnt",
"v2juaccnt_codelow", "v2juaccnt_codehigh", "v2juaccnt_sd",
"v2juaccnt_osp", "v2juaccnt_osp_codelow", "v2juaccnt_osp_codehigh",
"v2juaccnt_osp_sd", "v2juaccnt_ord", "v2juaccnt_ord_codelow",
"v2juaccnt_mean", "v2juaccnt_nr", "v2juaccnt_ord_codehigh",
"v2cltort", "v2cltort_codelow", "v2cltort_codehigh", "v2cltort_sd",
"v2cltort_osp", "v2cltort_osp_codelow", "v2cltort_osp_codehigh","v2cltort_osp_sd",
"v2cltort_ord", "v2cltort_ord_codelow", "v2cltort_ord_codehigh", "v2cltort_mean",
"v2cltort_nr", "v2clslavef"
,"v2clslavef_codelow" ,"v2clslavef_codehigh" , "v2clslavef_sd" ,
"v2clslavef_osp"
,"v2clslavef_osp_codelow" , "v2clslavef_osp_codehigh" ,"v2clslavef_osp_sd" ,
"v2clslavef_ord"
,"v2clslavef_ord_codelow" , "v2clslavef_ord_codehigh" ,"v2clslavef_mean" ,
"v2clslavef_nr",
"v2clacjstm_codelow" , "v2clacjstm_codehigh" , "v2clacjstm_sd" ,
"v2clacjstm_osp"
,"v2clacjstm_osp_codelow" , "v2clacjstm_osp_codehigh", "v2clacjstm_osp_sd" ,
"v2clacjstm_ord"
,"v2clacjstm_ord_codelow" , "v2clacjstm_ord_codehigh" ,"v2clacjstm_mean",
"v2clacjstm_nr",
"v2clacjstw_codelow" , "v2clacjstw_codehigh" ,"v2clacjstw_sd" ,
"v2clacjstw_osp" , "v2clacjstw_osp_codelow" ,"v2clacjstw_osp_codehigh",
"v2clacjstw_osp_sd"
,"v2clacjstw_ord" , "v2clacjstw_ord_codelow" ,"v2clacjstw_ord_codehigh"
,"v2clacjstw_mean"
,"v2clacjstw_nr", "v2clacjust", "v2clacjust_codelow" ,
"v2clacjust_codehigh" ,
"v2clacjust_sd" , "v2clacjust_osp" , "v2clacjust_osp_codelow",
"v2clacjust_osp_codehigh",
"v2clacjust_osp_sd" , "v2clacjust_ord" , "v2clacjust_ord_codelow" ,
"v2clacjust_ord_codehigh",
"v2clacjust_mean" , "v2clacjust_nr" , "v2clsocgrp" ,
"v2clsocgrp_codelow" ,
"v2clsocgrp_codehigh" , "v2clsocgrp_sd" , "v2clsocgrp_osp",
"v2clsocgrp_osp_codelow",
"v2clsocgrp_osp_codehigh", "v2clsocgrp_osp_sd" , "v2clsocgrp_ord" ,
"v2clsocgrp_ord_codelow" ,
"v2clsocgrp_ord_codehigh" ,"v2clsocgrp_mean" , "v2clsocgrp_nr" ,
"v2clrgunev" ,
"v2clrgunev_codelow" ,"v2clrgunev_codehigh" , "v2clrgunev_sd" ,
"v2clrgunev_osp",

"v2clrgunev_osp_codelow" ,"v2clrgunev_osp_codehigh", "v2clrgunev_osp_sd" ,
"v2clrgunev_ord" ,
"v2clrgunev_ord_codelow" ,"v2clrgunev_ord_codehigh" ,"v2clrgunev_mean" ,
"v2clrgunev_nr",
"v2elsuffrage", "v2extithos", "v2exremhsp_codelow", "v2exremhsp_codehigh" ,
"v2exremhsp_sd" , "v2exremhsp_osp" , "v2exremhsp_osp_codelow",
"v2exremhsp_osp_codehigh", "v2exremhsp_osp_sd" , "v2exremhsp_ord" ,
"v2exremhsp_ord_codelow" ,
"v2exremhsp_ord_codehigh" ,"v2exremhsp_mean" , "v2exremhsp_nr" ,
"v2exhoshog",
"v2clslavem" ,"v2clslavem_codelow" , "v2clslavem_codehigh" ,
"v2clslavem_sd" , "v2clslavem_osp" ,
"v2clslavem_osp_codelow" , "v2clslavem_osp_codehigh", "v2clslavem_osp_sd" ,
"v2clslavem_ord" ,
"v2clslavem_ord_codelow" , "v2clslavem_ord_codehigh", "v2clslavem_mean" ,
"v2clslavem_nr",
"v2cldiscw" , "v2cldiscw_codelow" ,
"v2cldiscw_codehigh" , "v2cldiscw_sd" , "v2cldiscw_osp" ,
"v2cldiscw_osp_codelow" ,
"v2cldiscw_osp_codehigh", "v2cldiscw_osp_sd" , "v2cldiscw_ord" ,
"v2cldiscw_ord_codelow" ,
"v2cldiscw_ord_codehigh" , "v2cldiscw_mean" , "v2cldiscw_nr",
"v2clacfree" ,"v2clacfree_codelow" , "v2clacfree_codehigh" , "v2clacfree_sd" ,
"v2clacfree_osp" ,
"v2clacfree_osp_codelow" , "v2clacfree_osp_codehigh" ,"v2clacfree_osp_sd" ,
"v2clacfree_ord" ,
"v2clacfree_ord_codelow" , "v2clacfree_ord_codehigh" ,"v2clacfree_mean" ,
"v2clacfree_nr",
"v2juncind" , "v2juncind_codelow" ,
"v2juncind_codehigh" , "v2juncind_sd" , "v2juncind_osp" ,
"v2juncind_osp_codelow" ,
"v2juncind_osp_codehigh" , "v2juncind_osp_sd" , "v2juncind_ord" ,
"v2juncind_ord_codelow" ,
"v2juncind_ord_codehigh" ,"v2juncind_mean" , "v2juncind_nr",
"v2clrelig_codelow" , "v2clrelig_codehigh" ,
"v2clrelig_sd" , "v2clrelig_osp" , "v2clrelig_osp_codelow" ,
"v2clrelig_osp_codehigh" ,
"v2clrelig_osp_sd" , "v2clrelig_ord" , "v2clrelig_ord_codelow" ,
"v2clrelig_ord_codehigh" ,
"v2clrelig_mean" , "v2clrelig_nr" ,
"v2clfmove" , "v2clfmove_codelow" , "v2clfmove_codehigh" ,
"v2clfmove_sd" , "v2clfmove_osp" , "v2clfmove_osp_codelow",
"v2clfmove_osp_codehigh" ,

"v2clfmove_osp_sd", "v2clfmove_ord" , "v2clfmove_ord_codelow" , "v2clfmove_ord_codehigh" ,
"v2clfmove_mean" , "v2clfmove_nr" , "v2cldmovew" , "v2cldmovew_codelow" ,
"v2cldmovew_codehigh", "v2cldmovew_sd" , "v2cldmovew_osp" , "v2cldmovew_osp_codelow" ,
"v2cldmovew_osp_codehigh", "v2cldmovew_osp_sd" , "v2cldmovew_ord" , "v2cldmovew_ord_codelow" ,
"v2cldmovew_ord_codehigh" ,"v2cldmovew_mean" , "v2cldmovew_nr" , "v2clstown" ,
"v2clstown_codelow" , "v2clstown_codehigh" , "v2clstown_sd" , "v2clstown_osp" ,
"v2clstown_osp_codelow" , "v2clstown_osp_codehigh", "v2clstown_osp_sd" , "v2clstown_ord" ,
"v2clstown_ord_codelow" , "v2clstown_ord_codehigh" , "v2clstown_mean" , "v2clstown_nr" ,
"v2clprptym" , "v2clprptym_codelow" , "v2clprptym_codehigh" , "v2clprptym_sd" ,
"v2clprptym_osp" , "v2clprptym_osp_codelow" , "v2clprptym_osp_codehigh", "v2clprptym_osp_sd" ,
"v2clprptym_ord" , "v2clprptym_ord_codelow", "v2clprptym_ord_codehigh" ,"v2clprptym_mean" ,
"v2clprptym_nr" , "v2clprptyw" , "v2clprptyw_codelow" , "v2clprptyw_codehigh" ,
"v2clprptyw_sd" , "v2clprptyw_osp" , "v2clprptyw_osp_codelow" , "v2clprptyw_osp_codehigh",
"v2clprptyw_osp_sd" , "v2clprptyw_ord" , "v2clprptyw_ord_codelow" ,"v2clprptyw_ord_codehigh",
"v2clprptyw_mean" , "v2clprptyw_nr" ,
"v2svdomaut_codelow" , "v2svdomaut_codehigh" ,
"v2svdomaut_sd" , "v2svdomaut_osp" , "v2svdomaut_osp_codelow", "v2svdomaut_osp_codehigh",
"v2svdomaut_osp_sd", "v2svdomaut_ord" , "v2svdomaut_ord_codelow" , "v2svdomaut_ord_codehigh",
"v2svdomaut_mean" , "v2svdomaut_nr" , "v2svinlaut_codelow" ,
"v2svinlaut_codehigh", "v2svinlaut_sd" , "v2svinlaut_osp" , "v2svinlaut_osp_codelow" ,
"v2svinlaut_osp_codehigh", "v2svinlaut_osp_sd" , "v2svinlaut_ord" , "v2svinlaut_ord_codelow" ,
"v2svinlaut_ord_codehigh" ,"v2svinlaut_mean" , "v2svinlaut_nr" ,
"v2svstterr" ,
"v2svstterr_codelow" , "v2svstterr_codehigh" , "v2svstterr_sd" , "v2svstterr_mean" ,
"v2svstterr_nr" , "v2pepwrsoc" , "v2pepwrsoc_codelow" ,

```
            "v2pepwrsoc_codehigh" ,    "v2pepwrsoc_sd"    ,      "v2pepwrsoc_osp"   ,
"v2pepwrsoc_osp_codelow" ,
            "v2pepwrsoc_osp_codehigh", "v2pepwrsoc_osp_sd" ,     "v2pepwrsoc_ord"    ,
"v2pepwrsoc_ord_codelow" ,
            "v2pepwrsoc_ord_codehigh" ,"v2pepwrsoc_mean"    ,    "v2pepwrsoc_nr"      ,
"v2xnp_pres"            ,
            "v2xnp_pres_codelow"     , "v2xnp_pres_codehigh" ,    "v2xnp_pres_sd",
            "v2xnp_regcorr_codelow",  "v2xnp_regcorr_codehigh", "v2xnp_regcorr_sd" ,
            "v2x_clpol"     ,        "v2x_clpol_codelow"  ,    "v2x_clpol_codehigh" ,
"v2x_clpol_sd" ,
            "v2x_clpriv"       ,     "v2x_clpriv_codelow" ,    "v2x_clpriv_codehigh"  ,
"v2x_clpriv_sd" ,
            "v2x_gencl_codelow"    ,  "v2x_gencl_codehigh"    , "v2x_gencl_sd",
            "v2xcl_acjst"    ,       "v2xcl_acjst_codelow" ,   "v2xcl_acjst_codehigh",
"v2xcl_acjst_sd",
            "v2xcl_prpty"      ,     "v2xcl_prpty_codelow" ,   "v2xcl_prpty_codehigh"  ,
"v2xcl_prpty_sd", "v2xcl_dmove",
            "v2xcl_dmove_codelow",    "v2xcl_dmove_codehigh"  , "v2xcl_dmove_sd" ,
"v2xcl_slave"       ,
            "v2xcl_slave_codelow" ,   "v2xcl_slave_codehigh" , "v2xcl_slave_sd",
"v2xel_elecpres" ,
            "v2x_feduni", "e_v2x_clpol_3C"    ,     "e_v2x_clpol_4C"   ,     "e_v2x_clpol_5C"    ,
            "e_v2x_clpriv_3C"  ,      "e_v2x_clpriv_4C"  ,      "e_v2x_clpriv_5C"     ,
"e_v2x_feduni_3C"    ,
            "e_v2x_feduni_4C"   ,     "e_v2x_feduni_5C"   ,     "e_v2x_gencl_3C"       ,
"e_v2x_gencl_4C"      ,
            "e_v2x_gencl_5C"     ,    "e_v2x_suffr_3C"     ,    "e_v2x_suffr_4C"        ,
"e_v2x_suffr_5C"       ,
            "e_v2xcl_rol_3C"      ,   "e_v2xcl_rol_4C"      ,   "e_v2xcl_rol_5C"     ,
"e_v2xeg_eqprotec_3C"   ,
            "e_v2xeg_eqprotec_4C"  ,  "e_v2xeg_eqprotec_5C"  ,  "e_regiongeo"         ,
"e_regionpol"           ,
            "e_regionpol_6C"
)

# Subset the data to exclude the specified columns
potential_variables <- noNA_data[, !(names(noNA_data) %in% exclude_cols)]


##############################################################################
####
##  How long first Heads of State last in general
##############################################################################
####
```

```
# Subset the dataset to only include countries with more than 1 observation
potential_variables_subset <- potential_variables[duplicated(potential_variables$country_name)
| duplicated(potential_variables$country_name, fromLast = TRUE), ]

# Group the data by country_name
grouped_data <- potential_variables_subset %>% group_by(country_name)

# Get the value of v2exnamhos at yearsSinceIndependence = 1 for each country_name
v2exnamhos_at_1 <- grouped_data %>%
  filter(yearsSinceIndependence == 1) %>%
  select(country_name, v2exnamhos)

# Join the v2exnamhos_at_1 data with the original dataset
mutated_data <- potential_variables_subset %>%
  left_join(v2exnamhos_at_1, by = "country_name", suffix = c("", "_at_1")) %>%
  # Mutate the has_changed_histname variable to indicate whether a country has a different
v2exnamhos at yearsSinceIndependence = 1
  mutate(has_changed_histname = ifelse(v2exnamhos != v2exnamhos_at_1 &
yearsSinceIndependence != 1, 1, 0)) %>%
  ungroup()

# Remove rows where has_changed_histname is 1 but there is an earlier
yearsSinceIndependence for that country where has_changed_histname is also 1
final_data <- mutated_data %>%
  group_by(country_name) %>%
  filter(!cumsum(has_changed_histname) > 1) %>%
  ungroup()

# View the final data
final_data

# Count how many instances of each country_name there are in final data
table(final_data$country_name)

# Create a survival object
surv_obj <- Surv(final_data$yearsSinceIndependence, final_data$has_changed_histname)

# Fit a Cox proportional hazards model
cox_model <- coxph(surv_obj ~ continent, data = final_data)

# View the summary of the model
summary(cox_model)
```

```
# Compute the overall Kaplan-Meier survival curve
overall_km <- survfit(surv_obj ~ 1, data = final_data)

# Plot the Kaplan-Meier curve using ggplot2
ggplot2_km <- autoplot(overall_km, censor = FALSE, conf.int = TRUE, surv.scale = "percent") +
  ggtitle("How long does the first leader of a country last?") +
  xlab("Years") + ylab("Survival Probability")

# Display the plot
ggplot2_km

# Compute the overall Kaplan-Meier survival curve
overall_km <- survfit(surv_obj ~ continent, data = final_data)

# Plot the Kaplan-Meier curve using ggplot2
ggplot2_km <- autoplot(overall_km, censor = FALSE, conf.int = TRUE, surv.scale = "percent") +
  ggtitle("How long does the first leader of a country last by continent?") +
  xlab("Years") + ylab("Survival Probability")

# Display the plot
ggplot2_km

################################################################################
####
## How long Heads of State last in general
################################################################################
####

# Keep only rows where country_name is in country_list and not in country_notlist
subset_data <- subset(data, country_name %in% country_list & !country_name %in%
country_notlist)

# Remove any rows where continent is NA
subset_data <- subset(subset_data, !is.na(continent))

# Group the data by v2exnamhos and continent and count the frequency of each combination
freq_data <- subset_data %>%
  group_by(v2exnamhos, continent) %>%
  count()

# Rename the count column to "frequency"
freq_data <- rename(freq_data, frequency = n)
freq_data <- subset(freq_data, v2exnamhos != "[Collective Body]")
```

```
# Step 3: Fit Cox proportional hazards model
coxph_model <- coxph(Surv(frequency) ~ continent, data = freq_data)

# Step 4: Check model assumptions
cox.zph(coxph_model) # test for proportional hazards assumption

# Step 5: Interpret results
summary(coxph_model) # view coefficients, standard errors, p-values, etc.

# Compute the overall Kaplan-Meier survival curve
overall_km <- survfit(Surv(frequency) ~ 1, data = freq_data)

# Plot the Kaplan-Meier curve using ggplot2
ggplot2_km <- autoplot(overall_km, censor = TRUE, conf.int = TRUE, surv.scale = "percent") +
  ggtitle("How long do all leaders last?") +xlab("Years") + ylab("Survival Probability")

# Display the plot
ggplot2_km

# Compute the overall Kaplan-Meier survival curve
overall_km <- survfit(Surv(frequency) ~ continent, data = freq_data)

# Plot the Kaplan-Meier curve using ggplot2
ggplot2_km <- autoplot(overall_km, censor = TRUE, conf.int = FALSE, surv.scale = "percent") +
  ggtitle("How long do all leaders last by continent?") +xlab("Years") + ylab("Survival Probability")

# Display the plot
ggplot2_km
```