SKELETON (heading = website page)

Homepage/Contents + Brief Project Description

Our project explores whether correlations are observed between certain factors/characteristics and becoming/being a politician. Initially, we examine the influence of birth month, before extending our analysis to factors such as the first letter of politicians' names, gender, and even their zodiac signs! Our analysis spans across more than 200 countries allowing us to identify general trends and variances between country groups based on HDI, V-Dem, MYS, socioeconomic group, and hemisphere groupings. After this (given some data limitations), we look into the United Kingdom in greater depth, constructing a new dataset to delve deeper into the significance of University/Alma Mater and age as potential variables affecting political careers. Through this we hope to identify some interesting trends in characteristics amongst the politicians that represent and govern us.

Motivation

Politicians are the instrument of the government, they play an instrumental role in shaping the laws, policies, and decisions that govern our daily lives. The world has seen many politicians, from conservatives to liberals, dictators to democratic, aggressive to charismatic and efficient to futile, but what determines who becomes a politician? Traditional political science literature focuses mostly on race, gender, ethnicity and other identities, but are there other factors at play? Is it possible that underlying factors people completely overlook, such as birth month, have a bearing on who becomes a politician?

Our initial curiosity around birth month stems from various studies that claim to have found relationships between birth month and disease risks, likelihood of sporting success and even becoming a top politician! For example, Tukiainen et al. (2018) find correlations between being born earlier in the year and becoming a top politician in Finland, potentially suggesting a realisation of the Relative Age Effect. This describes the case where older peers have an inherent advantage over those born later due to factors potentially relating to physical/mental maturity and experience, leading to greater opportunities for development (Kiikka, 2018). So, with a wider lens, do we see a correlation between birth month and being a politician? Are January babies more likely to rule the world? Are there other factors we haven't found yet? Read further to find out …

Data Sources, Data Cleaning and Datasets

Data Sources:

EveryPolitician - for core analysis dataset
Wikipedia - For UK-specific analysis
UN data (UN Statistics Division) - to compare population births data (baseline)

World Bank - for income categories
UNDP - for HDI, GII, MYS
ONS - for UK births data for day of the week
TheyWorkForYou - Filling gaps in Wikipedia Data
International Institute for Democracy and Electoral Assistance - For gender quotas
V-Dem Institute- For Electoral Democracy Index
Google Public Data - For hemisphere classification of countries

Data Sources:

EveryPolitician - for core analysis dataset
Wikipedia - For UK-specific analysis
UN data (UN Statistics Division) - to compare population births data (baseline)
World Bank - for income categories
UNDP - for HDI, GII, MYS
ONS - for UK births data for day of the week
TheyWorkForYou - Filling gaps in Wikipedia Data
International Institute for Democracy and Electoral Assistance - For gender quotas
V-Dem Institute- For Electoral Democracy Index
Google Public Data - For hemisphere analysis

Data collection, limitations and cleaning

Need a paragraph on the process of getting the data from everypolitician

Data from EveryPolitician

The core dataset was obtained from EveryPolitician through a two-step process. EveryPolitician provides a machine-readable ‘index file’ in JSON format which contains details about and links to the data for each individual legislature. The links to all the datasets were collected and these were then extracted using the ‘get’ function of the module ‘requests’, similarly to web scraping. The data was obtained in JSON format, converted to a ‘pandas’ dataframe, and saved to a csv file for later use, since the data collection functions took a long time to run.

January Skew

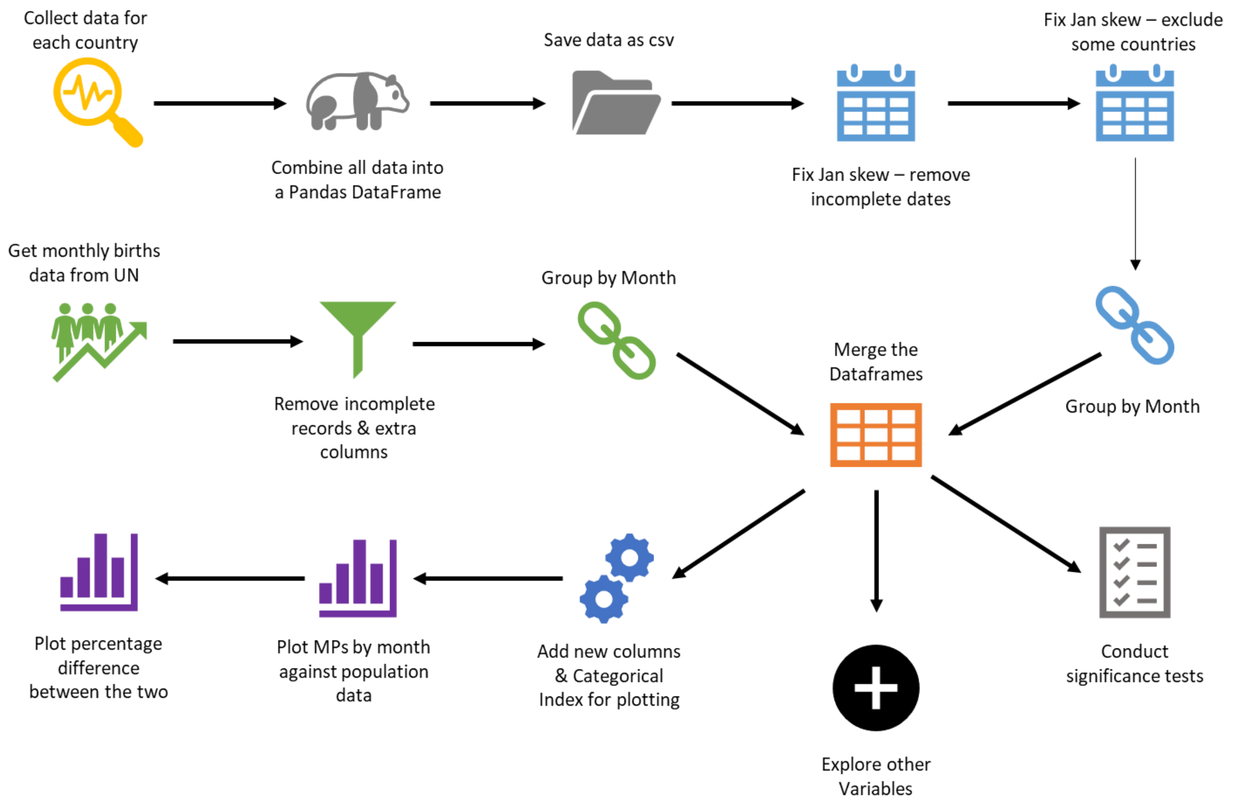
Beyond the initial data cleaning expected to make the dataset usable in analysis, we needed to perform some operations before the EveryPolitician dataset was fit for purpose. Initially, the proportion of politicians with a January birthday was *more than 4 times greater than any of the other months.* Upon deeper analysis, this was attributed to the standardised value of 1 January being assigned to politicians for which birth month data was not available (as the pd.to_datetime function defaulted to this when no day/month information is provided). We identified this phenomenon as the January skew. After exploring a few methods to address this, we chose to omit outlier countries that had an unreasonable number of 1 Jan values. The threshold for the omission was computed by excluding countries where the ratio of records on January 1st was more than 10 times the expected ratio of records for Jan 1st - seven countries for which this threshold was violated were excluded from the analysis, including Syria and Cameroon for which the proportion of people born on 1 January were 98.2% and 26.8% respectively. This differs greatly from the global proportion of 8.97% of births in January. Hence such results were removed from our birth month analysis to safeguard data quality, and then we were able to perform our analyses.

During the birth month analysis process, we noted (especially for some of the grouped analyses) that data for many less-developed countries often contained far fewer entries and was generally less complete. For instance many countries in Sub saharan Africa such as Eritrea and Sudan as well as some in Asia and Latin America like Myanmar and Venezuela. When performing gender analyses, we also noted the impact of data for historical MPs dampening the observed results. Therefore, due to these data quality issues (and facilitating deeper analysis of certain factors that the EveryPolitician dataset did not allow for, such as alma mater), we decided to perform a deep dive into the UK. From a data science perspective, focusing on one country allows us to address concerns about variations in data availability, quality and consistency across regions. From a personal perspective, we were the most familiar with universities from the UK given our attendance to one and hence thought it most interesting and reasonable to investigate.

For this UK-focused analysis, we used the Wikipedia API and BeautifulSoup-aided HTML-scraping to collect data on all sitting UK MPs (from each of the England, Wales, Scotland and Northern Ireland pages) - we chose this source as it provides alma mater and birth date in a scrapable fashion (although we did encounter some roadblocks in the process!)
However, Wikipedia allows for pub
lic contributions so to ensure data quality we cross-checked this Wikipedia data with another source, to limit errors in the dataset. This was indeed the case given that the data initially included MPs that had resigned/passed away and (once we accounted for this) was missing an MP!
To identify the missing MP, we cross checked using a website called TheyWorkForYou that sources its data directly from official parliamentary sources. Cross referencing produced 23 anomalies between the dataset, however, given the nature of similar names (eg. Jon Ashworth vs Jonathan Ashworth) we needed to find the missing MP manually from the anomaly list identified by the regex search command. After correcting for this the dataset was ready for use containing data on all 650 MPs in the UK (i.e across Wales, Scotland, England and Northern Ireland)

Analysis roadmap and process

Ojas’ Roadmap


Collect data for each country
Combine all data into a Pandas DataFrame
Save data as csv
Fix Jan skew – remove incomplete dates
Fix Jan skew – exclude some countries
Get monthly births data from UN
Remove incomplete records & extra columns
Group by Month
Merge the Dataframes
Group by Month
Plot percentage difference between the two
Plot MPs by month against population data
Add new columns & Categorical Index for plotting
Explore other Variables
Conduct significance tests

Wanted to explore the birth month trend was observable amongst a larger set of politicians, spanning as many countries as we could.
Identified a dataset from EveryPolitician that collected data for 233 countries across recent history.
Cleaned this, addressed issues with data quality (namely Jan skew).
Performed analysis to see whether we could observe any trends in birth month (comparing to UN births data) across all countries, and then performed the same analysis on subsets of countries (grouping by hemisphere, HDI, V-Dem score, Income Category and Mean years of Schooling) to see if different trends were observable between different country groups.
Then, given the dataset we had, we wanted to explore some other potential trends in MP characteristics - such as the first letter of name of MPs (to see if donkey voting may have had an effect historically), MP gender and zodiac sign (as a fun exploration!). Again, we explored these for the subsets previously identified.
Here we encountered issues with data quality and relevance (as not only did this dataset include historically sitting MPs, but also had poor data quality for a few countries) - as such we sought to do a deep dive into the UK (as this is most interesting to us).
Here we wanted to see if we observed a correlation between the university attended and becoming a politician, and also see if age had a correlation (as this was not possible with the historical dataset). We also repeated the birth month analysis to see if a different correlation was observed for the UK specifically.

Birth Month Results

During the birth month analysis process, we noted (especially for some of the grouped analyses) that data for many less-developed countries often contained far fewer entries and was generally less complete. For instance many countries like Afghanistan, Algeria, Iran, DRC, Bolivia, Chad etc. have <10 records of birth date .When performing gender analyses, we also noted the impact of data for historical MPs dampening the observed results. Therefore, due to these data quality issues (and facilitating deeper analysis of certain factors that the EveryPolitician dataset did not allow for, such as alma mater), we decided to perform a deep dive into the UK. From a data science perspective, focusing on one country allows us to address concerns about variations in data availability, quality and consistency across regions. From a personal perspective, we were the most familiar with universities from the UK given our attendance to one and hence thought it most interesting and reasonable to investigate.

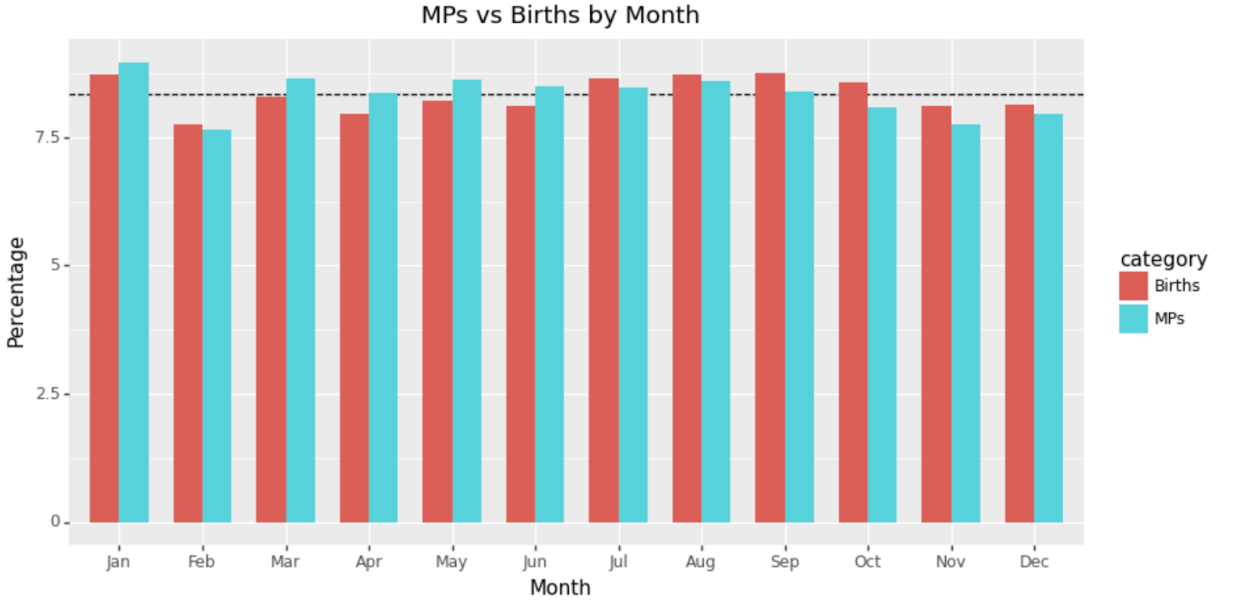
Data trends

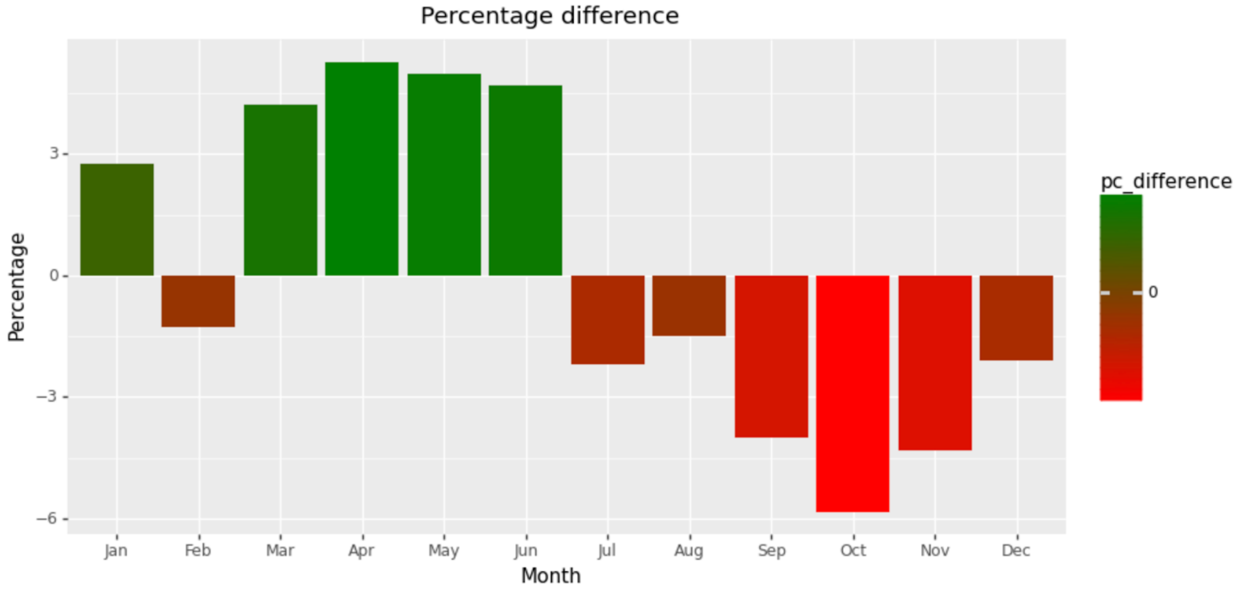
To explore the factors correlated with being a politician across the range of countries we decided to investigate trends in Birth Month, Gender, Zodiac (for fun) and Day of Birth. Beyond performing these analyses for the entire dataset, we also examined whether the trends changed between different groups of countries (based on high vs low Human Development Index, Mean Years of Schooling, Income Group, V-Dem Score and Hemisphere). The general trends observed were as follows:

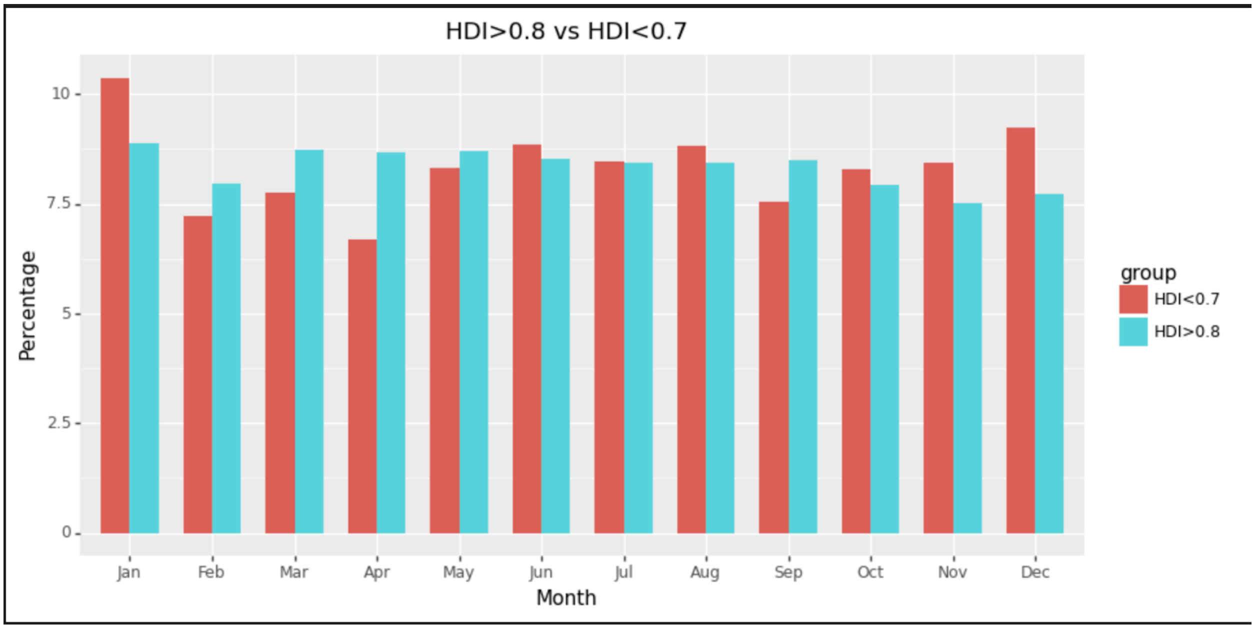
- Birth month: Being born in the first six months of the year is generally correlated with a slightly higher likelihood (6.7%) of being a politician. This is most prominently seen when looking at the chart for the percentage of MPs against the month of birth. As seen by the graph and substantiated by our table below there is a greater number of MPs relative to the number of births in each month before June, whereas afterwards the number of MPs was lower relative to births. This trend was noticeably more clear from the percentage difference graph which compares politician birth month data to population birth month data collected from the UN. It exhibits how on average we notice a 3.43% positive percentage difference across politician birth month data for the first six months of the year compared to on average a -3.32% negative percentage difference for the last. The relative age effect gives some insight into this, as politicians born in the first six months may have an advantage in education, leadership networking, and experience. This may potentially lead to increased confidence, abilities, and political aspirations heightening chances of becoming a politician.

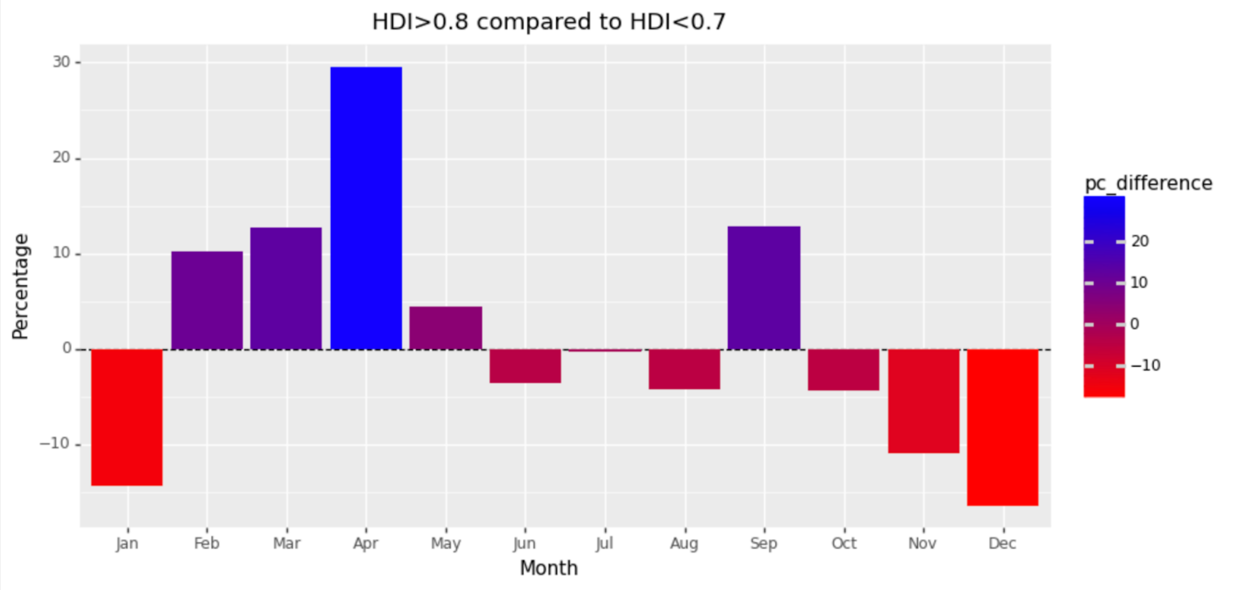
(Difference between months is significant, p-value=0.038)

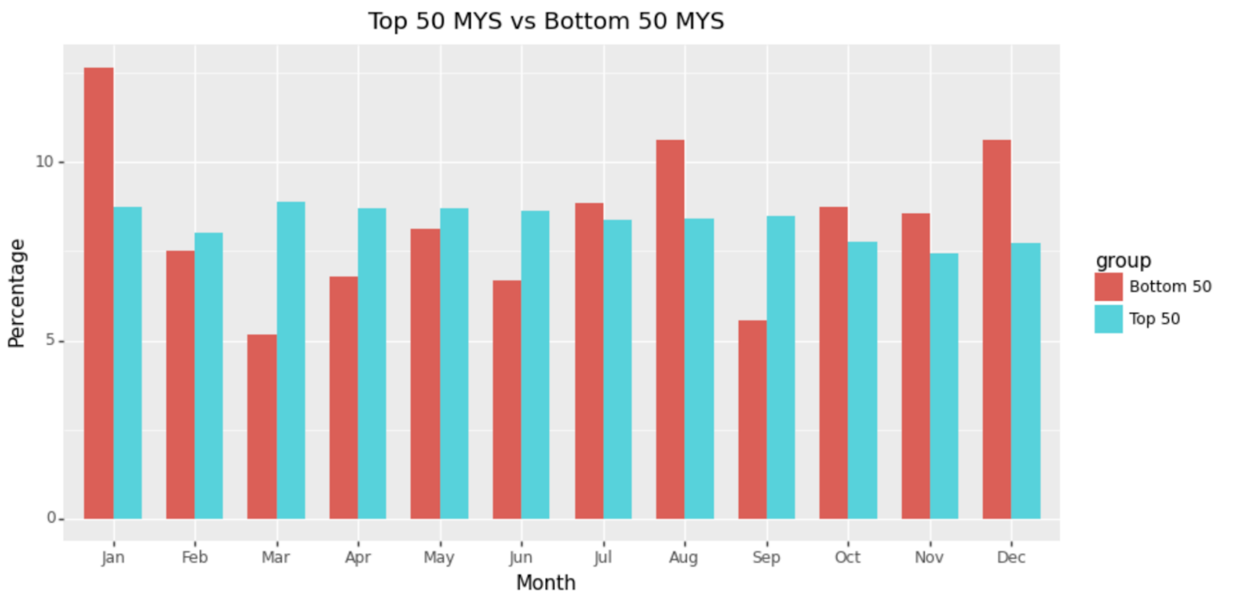
The subgroup analysis based on development indicators somewhat exhibits this but the results are relatively more pronounced for countries with lower levels of development such as the Democratic Republic of Congo and Kenya .(couldn't check dataset as couldn't get in right format based on intuition has to be cross checked) For instance the chances of becoming a politician given you are born in January for a country with low HDI are substantially much higher than the rest; 10.38% compared to the sample average of 8.33%. However the graph also shows a relatively high number of politicians from low HDI countries being born in December which invalidates the general trend. All 39 African countries in our dataset except Mauritius have a HDI lower than 0.7 and so make up the majority of the dataset for low HDI countries. The fact that the schooling year in most African countries starts in January could offer insight into why more politicians as seen on the graph are born in January or December. This is because relating to the relative age effect these individuals may have had an advantage in their education and career development due to their age meaning potentially greater likelihood of being a politician. These results are even more overstated for the Mean years of schooling graph where the disparity between being born in January for the first top 50 countries is more than 25% lower than the bottom 50 ones. Notwithstanding this, given that there are major concerns with data quality, specifically completeness to do with data entries for a number of low development countries, we are avoiding making any strong justifications for the patterns seen in the data. This also motivates our focus on the UK later on in our analysis.

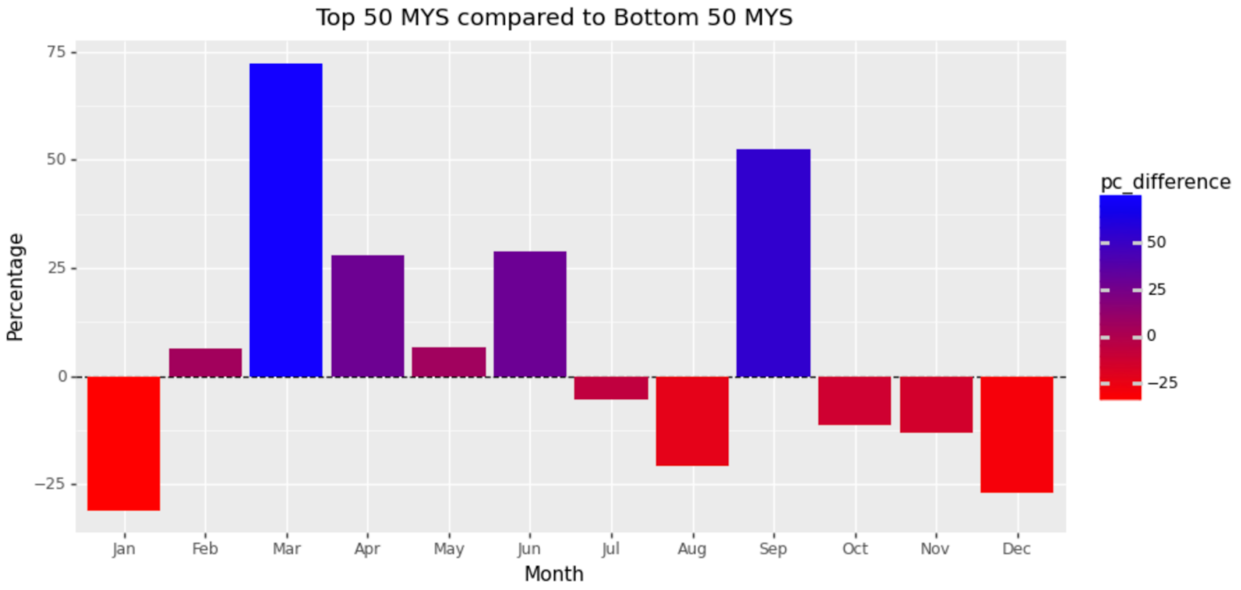


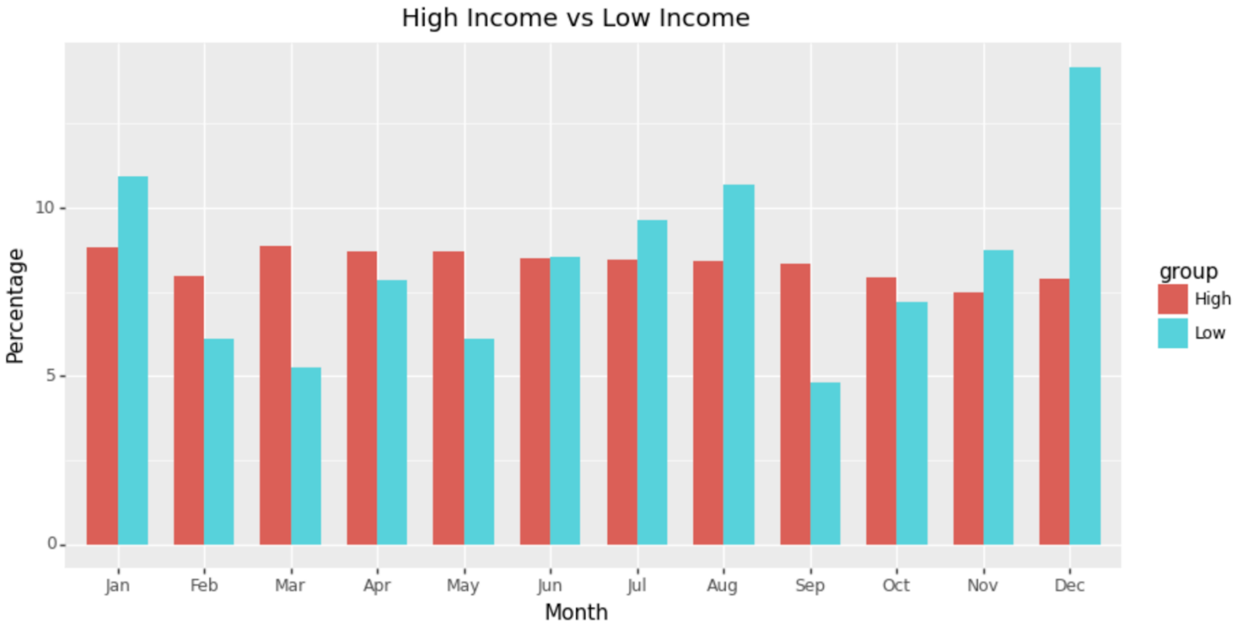


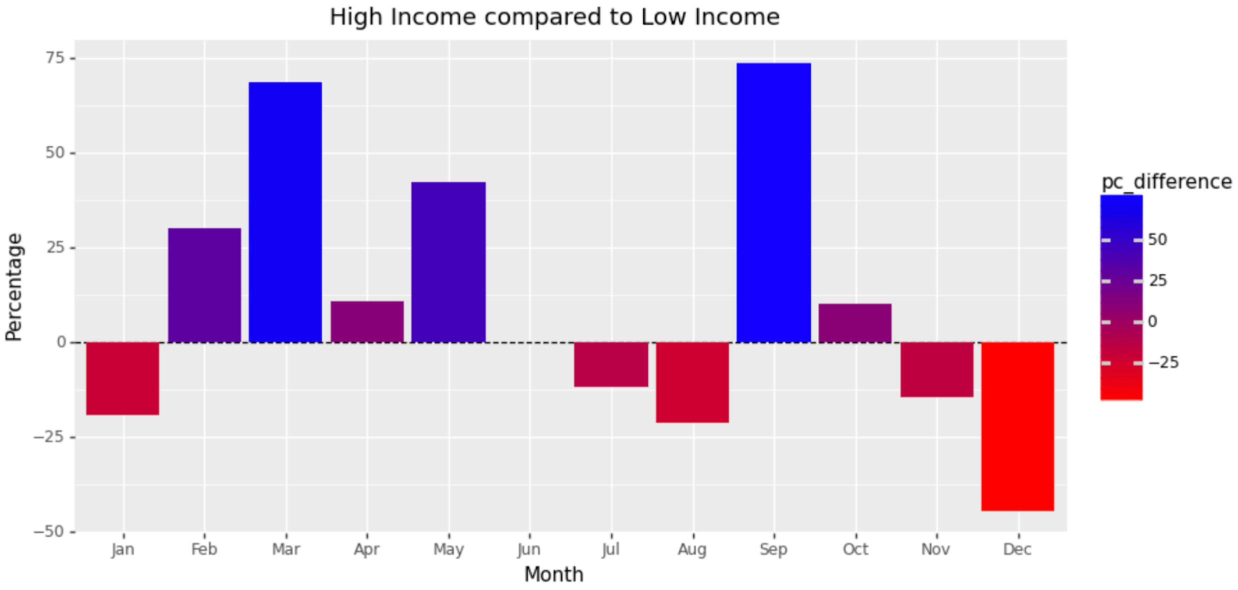

HDI>0.8 vs HDI<0.7
Percentage
10
7.5
5
2.5
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
group
HDI<0.7
HDI>0.8

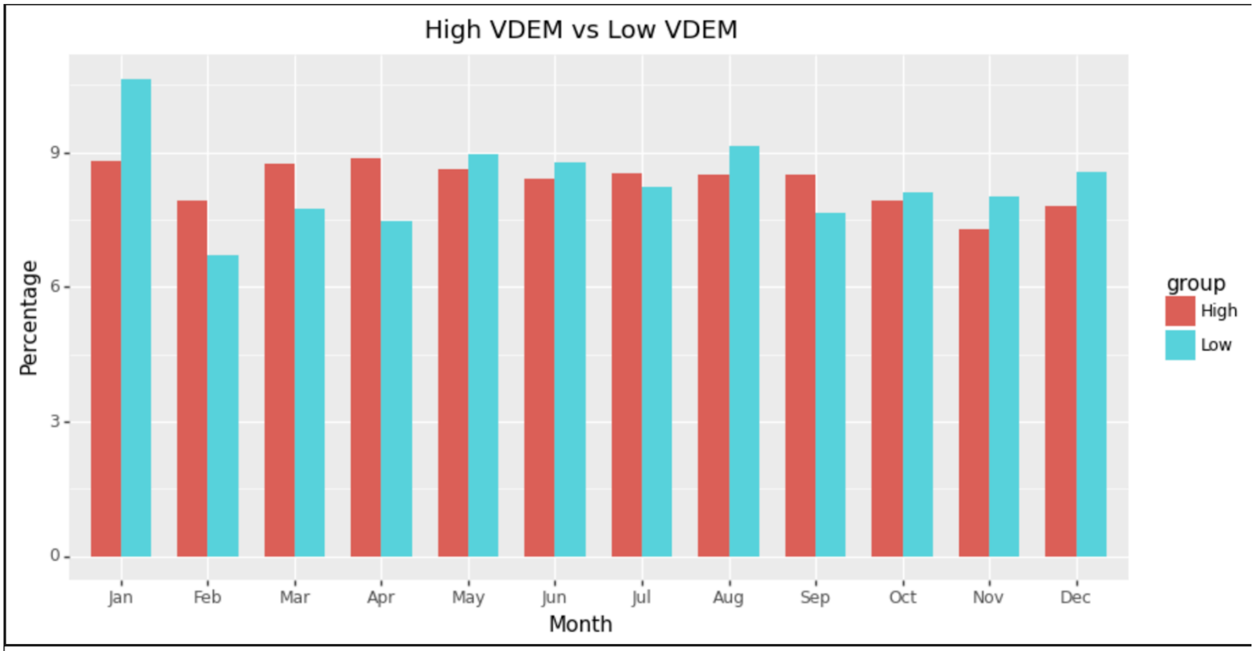

HDI>0.8 compared to HDI<0.7
Percentage
30
20
10
0
−10
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
pc_difference
20
10
0
−10


Top 50 MYS vs Bottom 50 MYS
Percentage
10
5
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
group
Bottom 50
Top 50

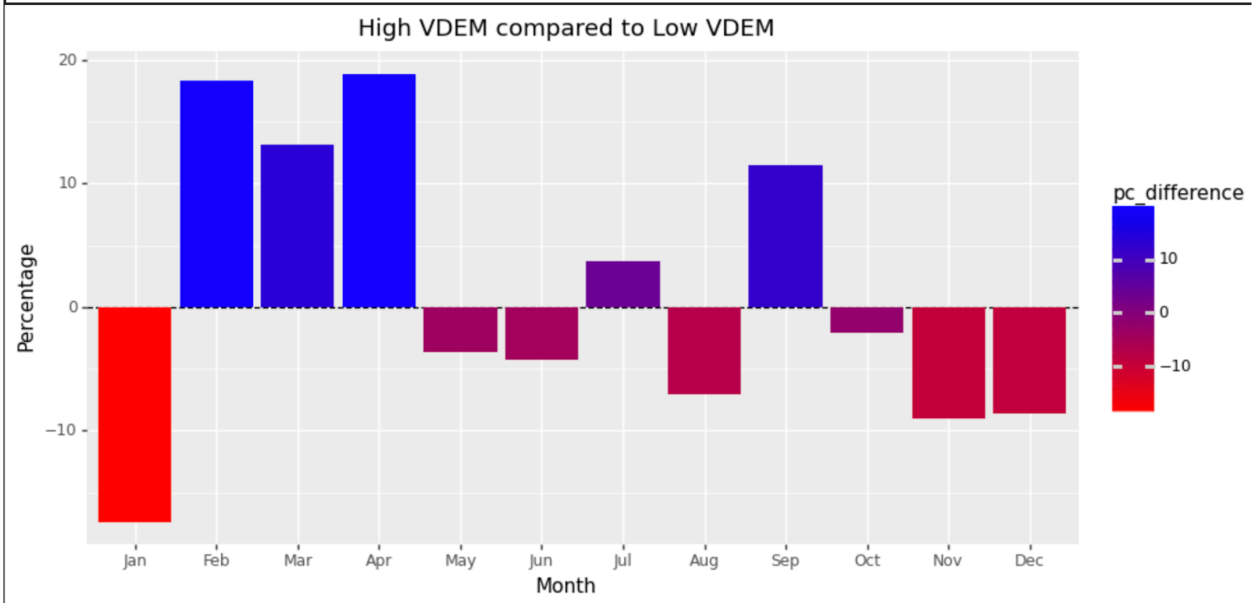

Top 50 MYS compared to Bottom 50 MYS
Percentage
75
50
25
0
−25
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
pc_difference
50
25
0
−25


High Income vs Low Income
Percentage
0
5
10
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
group
High
Low

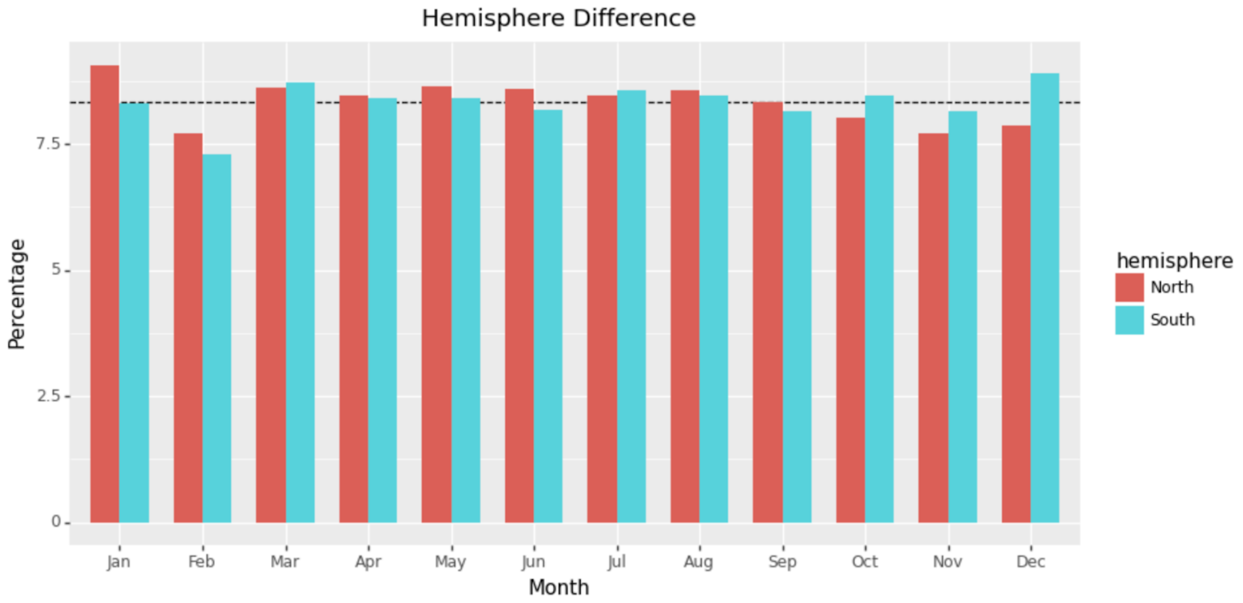

High Income compared to Low Income
Percentage
75
50
25
0
−25
−50
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
pc_difference
50
25
0
−25

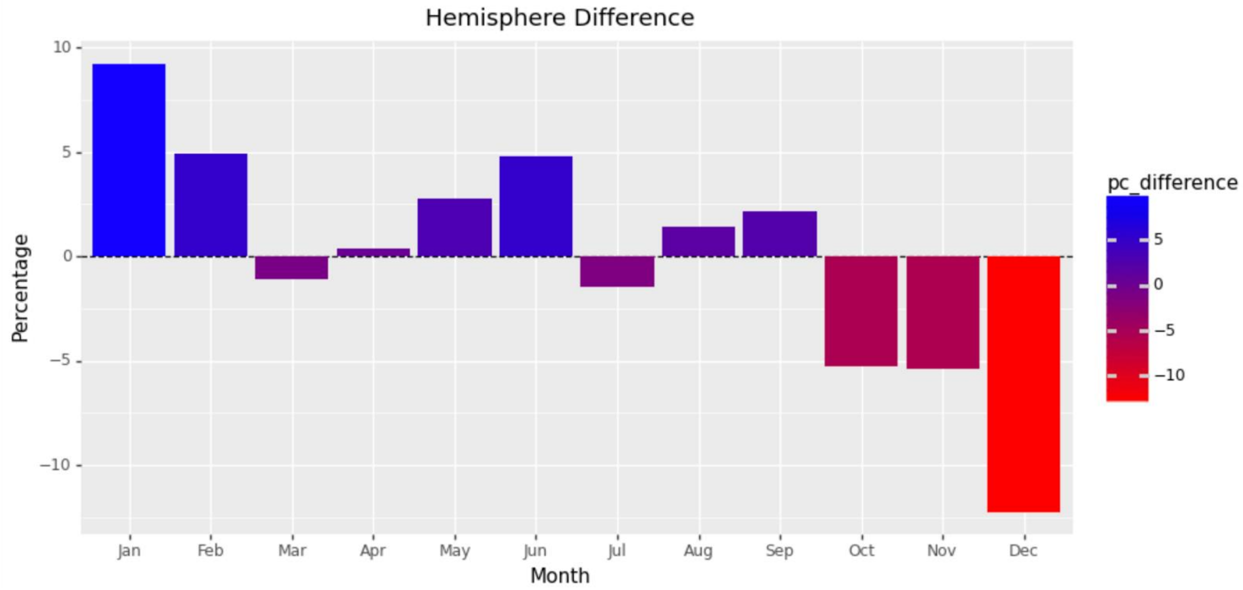


High VDEM vs Low VDEM



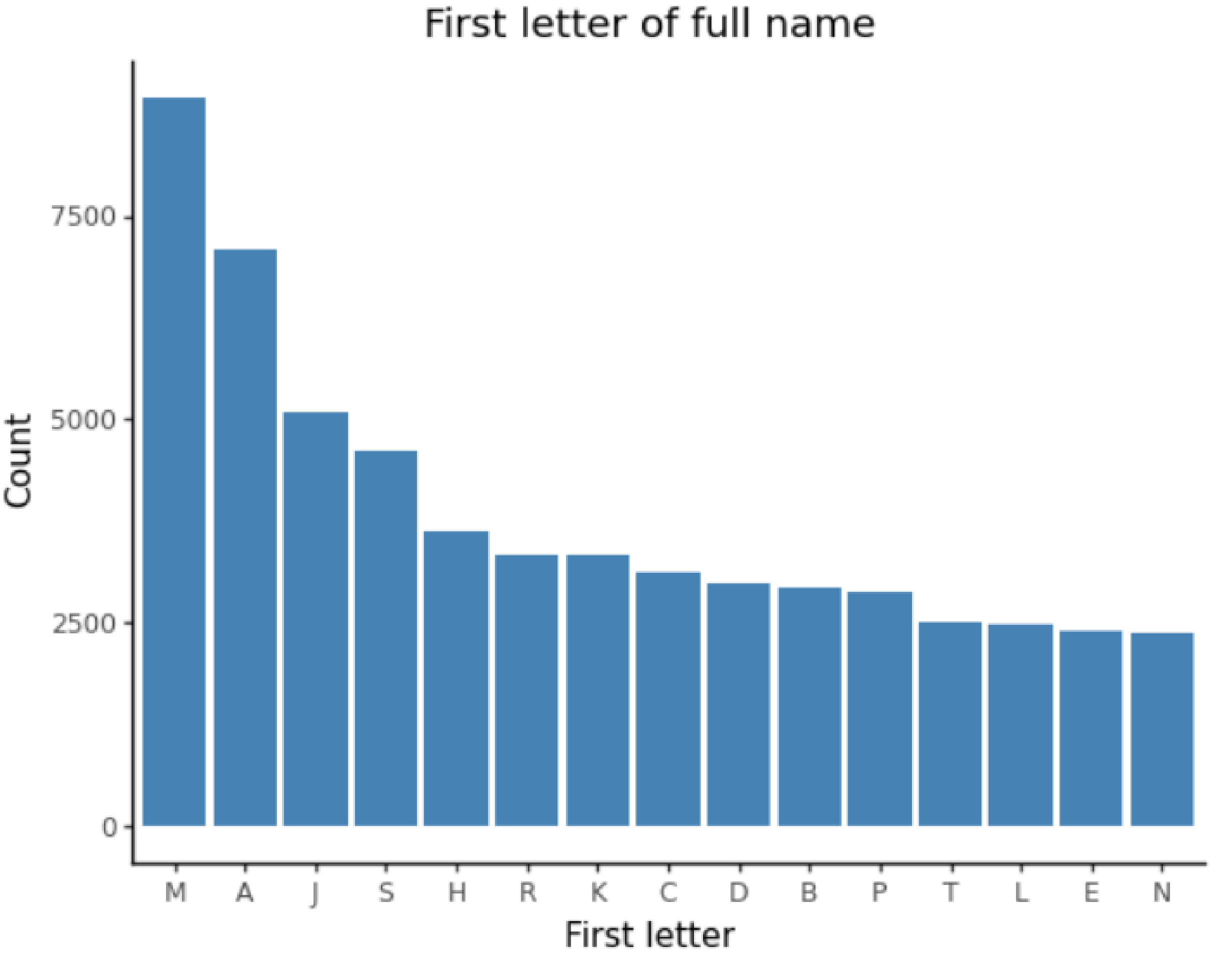
High VDEM compared to Low VDEM

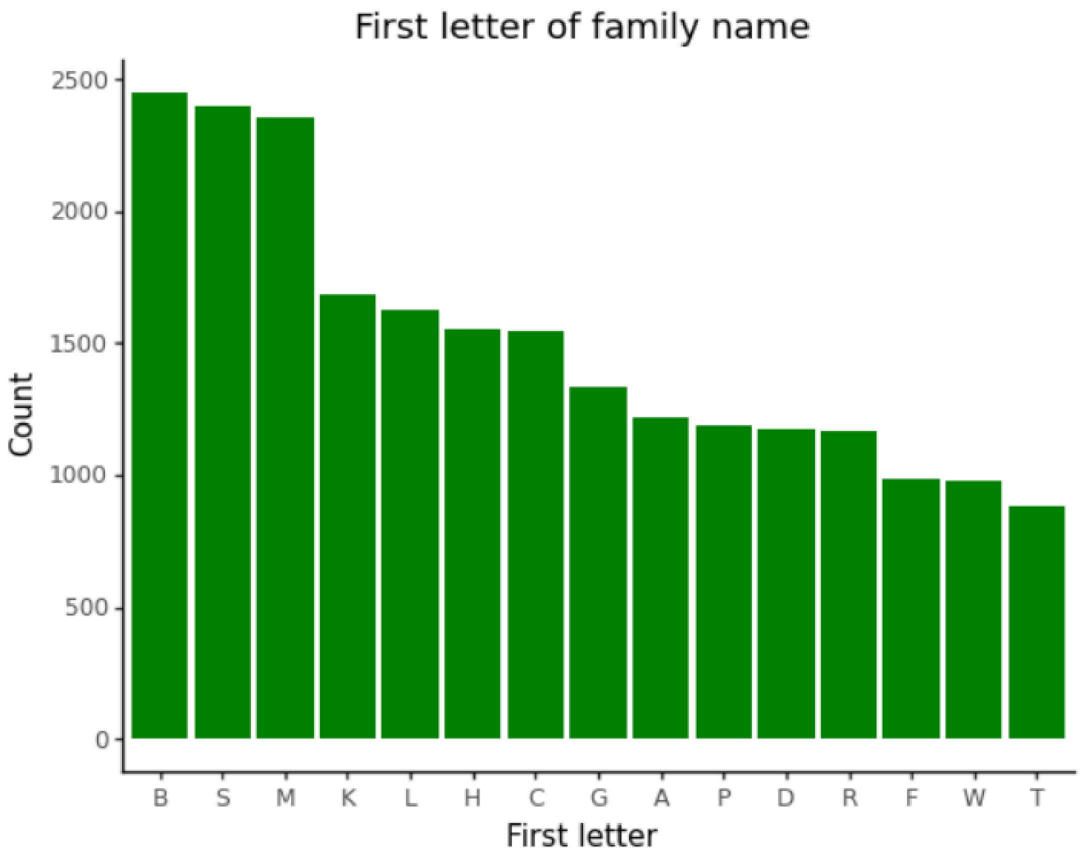




First letter of name

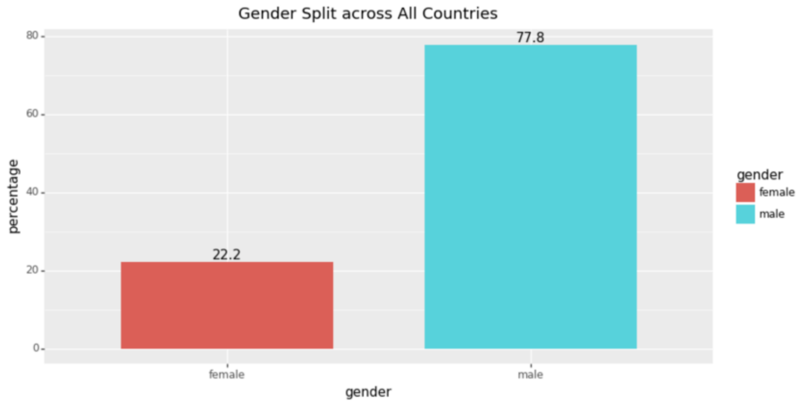
When analysing the bar chart for groupings to do with the first letter of the name there is no distinct or noticeable trend among the alphabets that most politicians’ names start with. This is substantiated by A; the first letter of the alphabet having the second highest percentage of political representatives while, the highest percentage of political representatives' names starting with M which is halfway down the alphabet. Similarly the highest frequency of politicians' names start with the letter B followed by S and M.

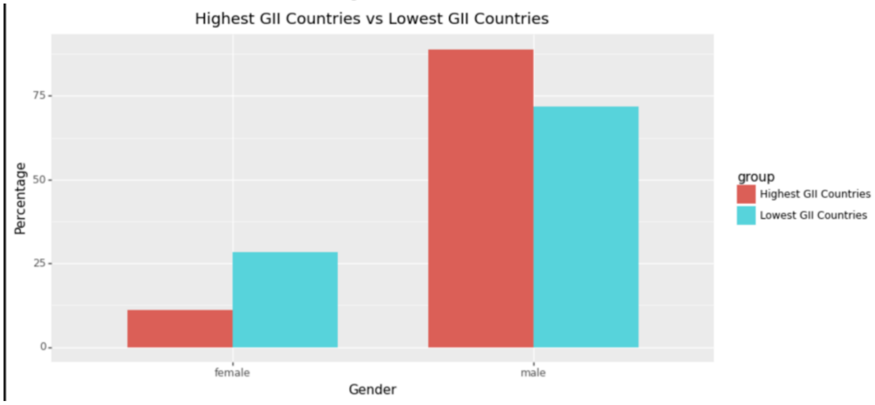

First letter of full name
Count
7500
5000
2500
0
M
A
J
S
H
R
K
C
D
B
P
T
L
E
N
First letter



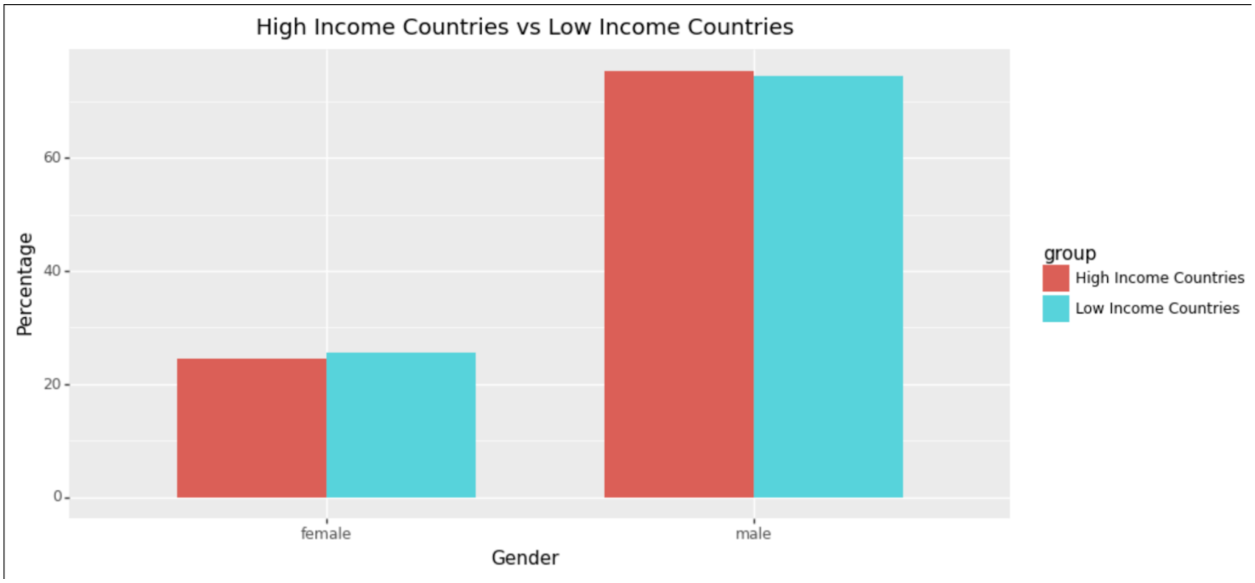
Gender

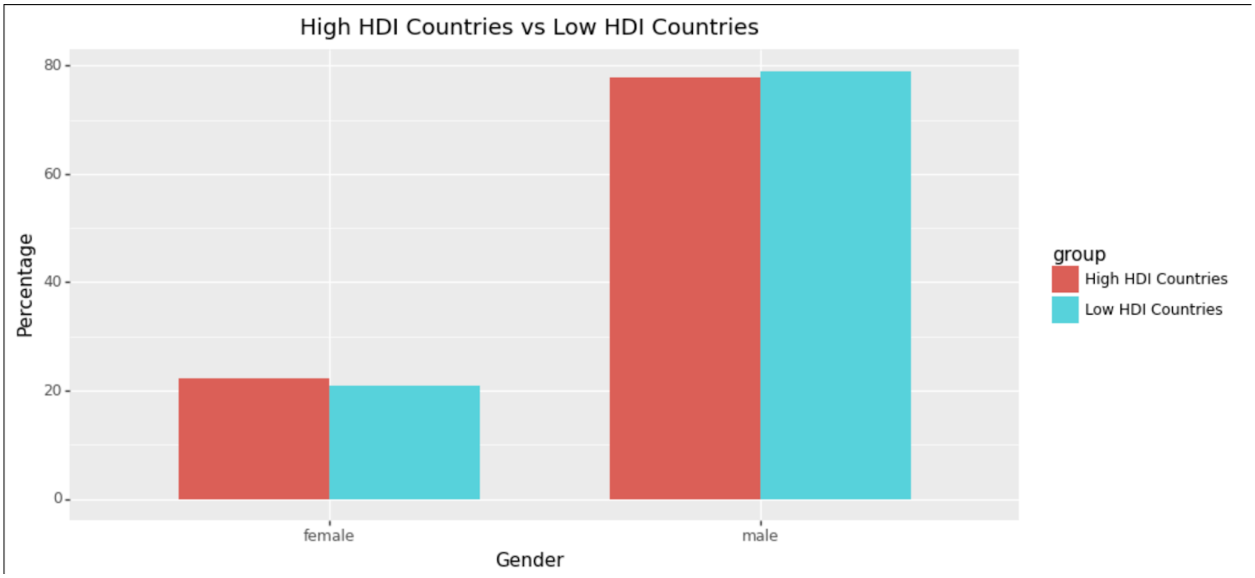
The gender gap in political representation is relatively robust to different groupings based on HDI, MYS or other development indicators. It certainly showcases a heavily positive correlation between being a politician and being male. This can be substantiated by historical and cultural norms, gender roles, structural barriers, and gender biases in the electorate. These factors play into limiting women's access to political opportunities thereby discouraging participation in politics.Although this result is intuitive it is worthwhile considering that the disparity between being a politician given you are male versus female is significantly more apparent for highly gender-unequal countries. This is shown most clearly by the percentage difference between being a politician given you are female for the highest Gender Inequality Index countries when compared to the lowest ones. So while the average split globally for gender representation is 22.2% female and 77.8% male, for the lowest GII countries it is 11.08% and 88.92%.



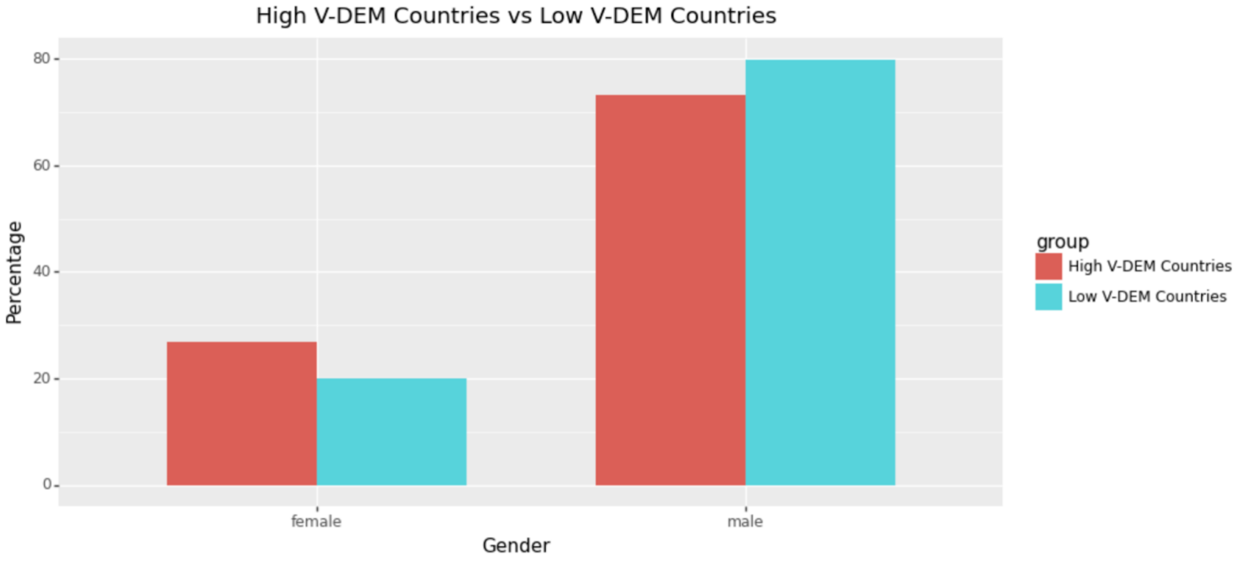


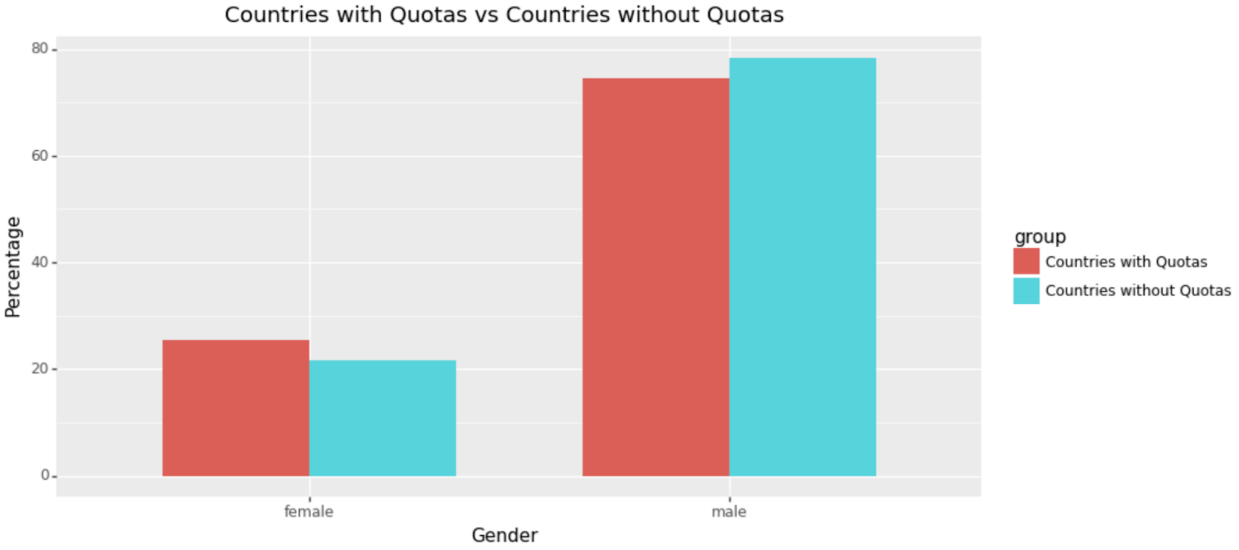
This difference is not as pronounced for other development indicators such as income and HDI as visualised in the graphs. This is unsurprising considering the construction of the HDI index and mean income is by nature focused more on overall economic factors whereas the GII is more suited to measuring gender disparities and biases.Moreover this explains why some countries like Rwanda that are classified as low development countries with an index of 0.543 actually have high political representation as accounted for by the low Gender Inequality value of 0.338.



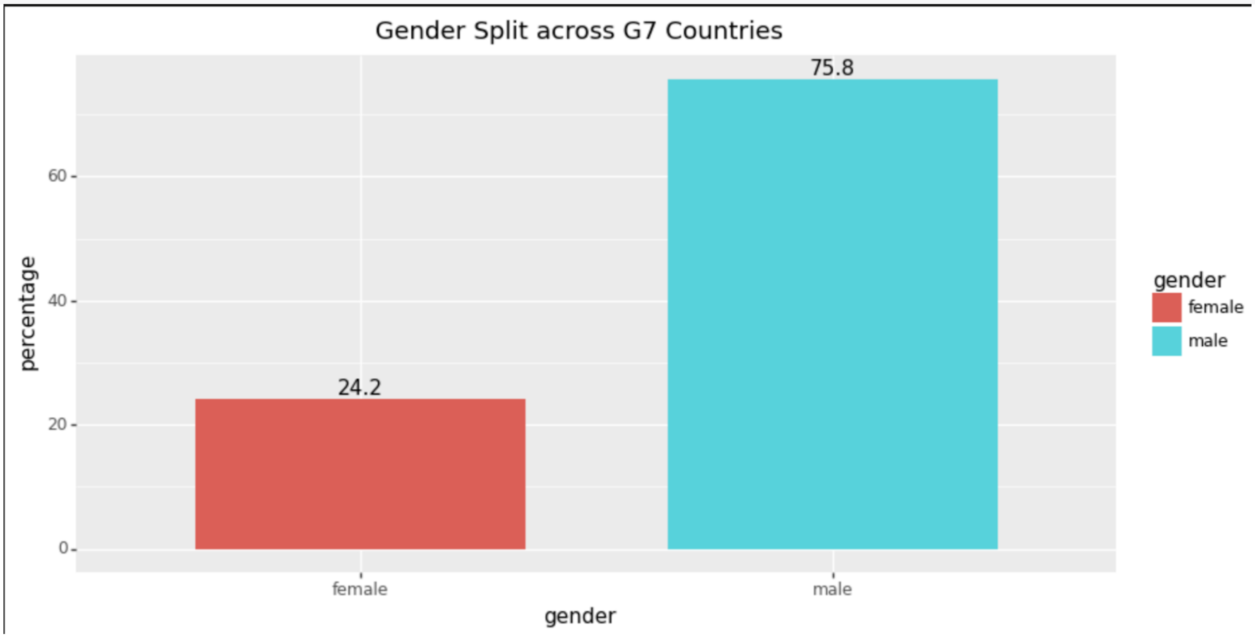


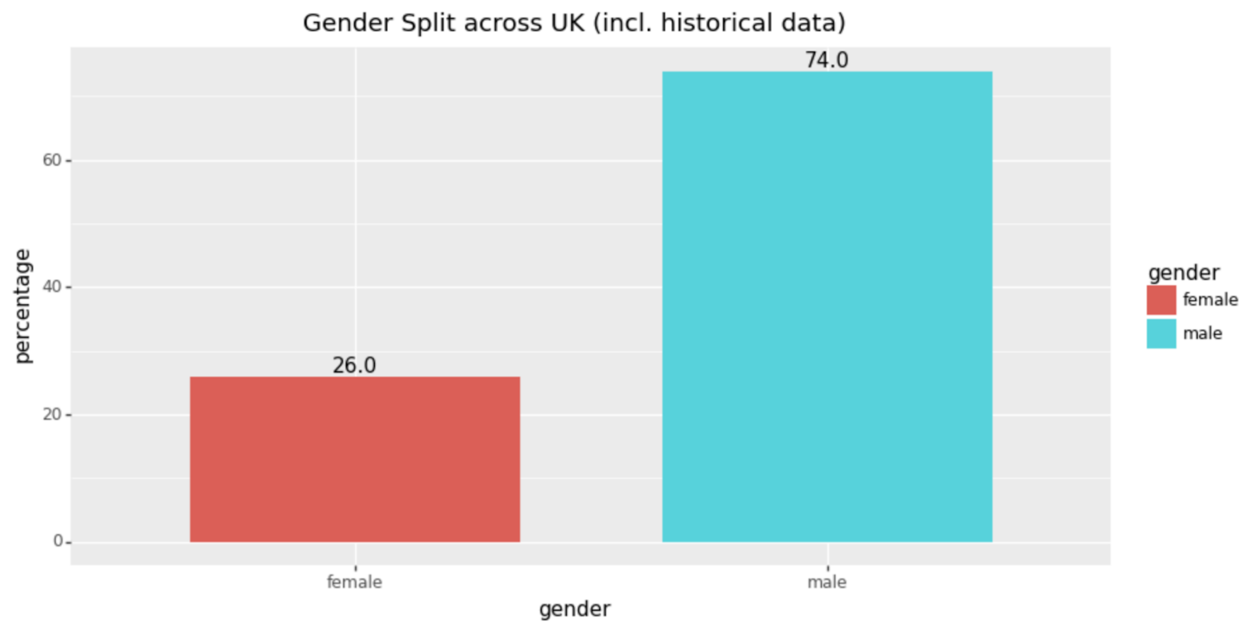
Another insightful comparison is between the bar charts for the V-DEM index which measures democracy quality worldwide, and the countries with & without quotas. Countries with quotas only have a 3.5 percentage point difference to countries that don't whereas countries with a high V-DEM versus low have a 6% percentage point difference. Results being more pronounced for V-DEM might be because quotas target specific gender imbalances while V-Dem index captures the overall democratic context, including civil liberties, electoral processes, and institutional functioning. Therefore it may produce more substantial results as it considers a more complex interplay of structural, cultural and institutional factors that influence political representation. This can be reflected by India‘s example which has a substantial number of quotas but still low political representation for women which is reflected by its low V-DEM value of 0.399.





When comparing gender disparities in political representation between G7 countries to the UK, while it is apparent that the gender split is clear in both figures, it is slightly lower for the UK. Our data collection has yielded the 26.4-74 gender split for political representation for the UK while a 24.2- 75.8 split for G7 countries. Although this can be attributed to the male in political representation. In comparison, the UK exhibited a historic data of 26.4% female and 74% male representation This could be attributable to the G7 countries having more data points and the inherent variance among those data points. For instance while Canada has made efforts to address gender disparities,another G7 country France has struggled with cultural norms and traditional gender roles. Moreover, while Germany has implemented policies promoting women's representation,Italy experiences lower levels of female representation due to cultural factors, and Japan faces challenges stemming from cultural norms and limited policy measures. Hence with more data points available for G7 countries than for the UK, there is a greater chance of encountering a wider range of gender disparities within the G7 dataset.



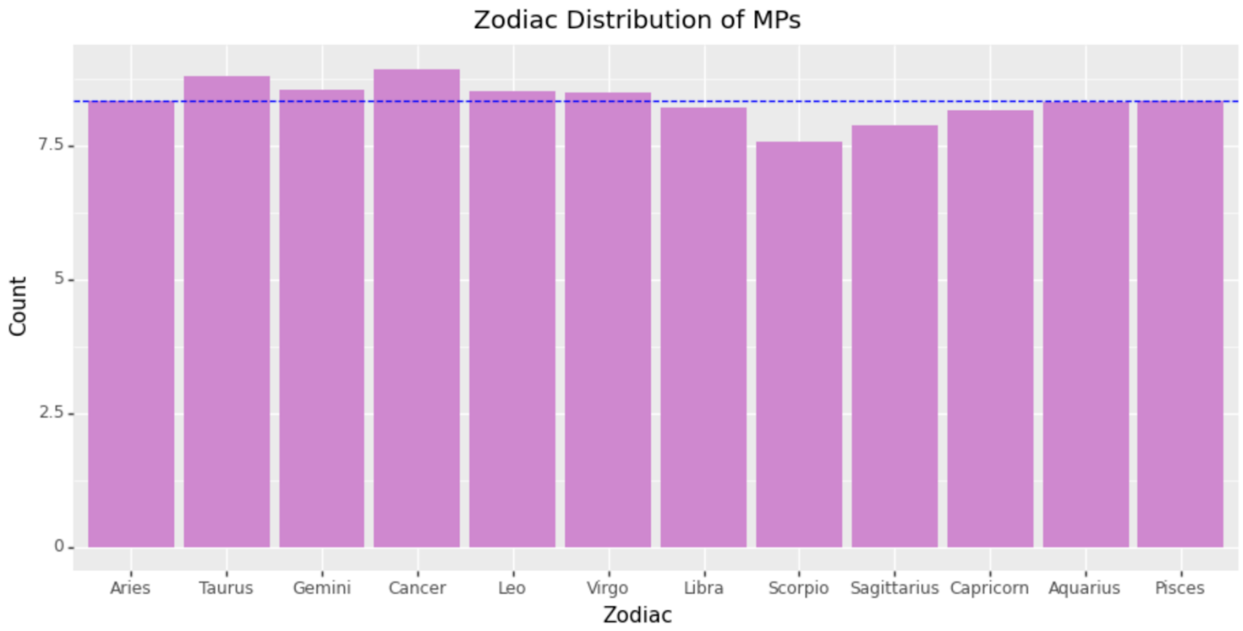


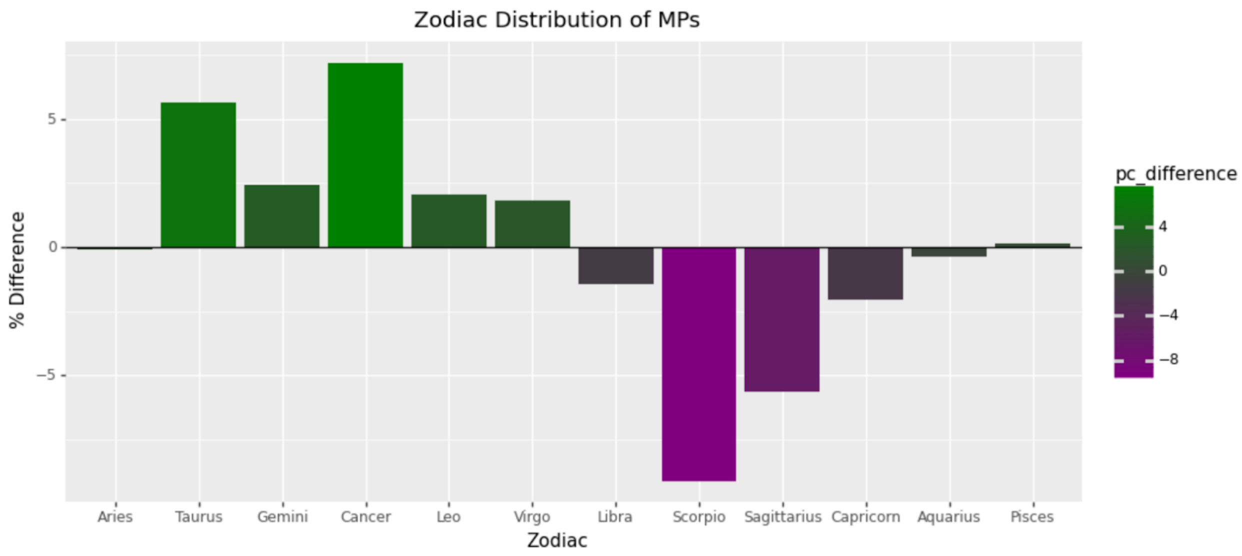
Zodiac

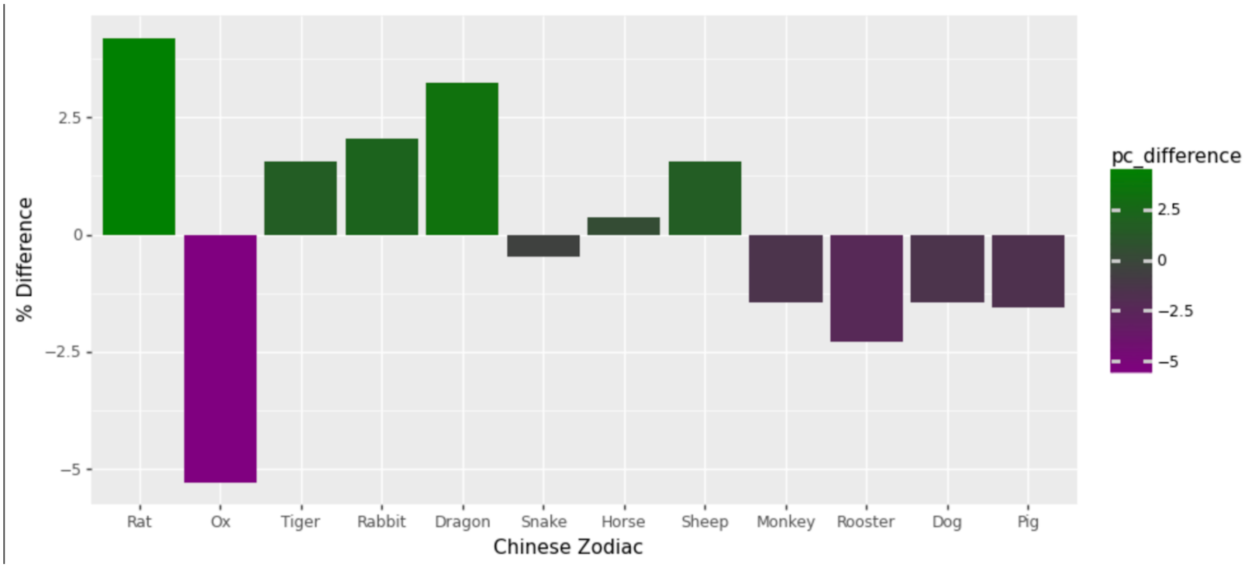
For entertainment purposes only we decided to also include zodiac signs and observed that people with the first 6 zodiac signs and by extension those born in the first 6 months have a relatively higher chance of becoming MP as visualised by the graph below. This was especially more pronounced for Cancerians and Taurians who were overrepresented relative to other signs. Being a Scorpio on the other hand …

For entertainment purposes only, we delved into the mysterious world of zodiac signs and uncovered an intriguing association between astrology and political careers. Individuals born under the first six zodiac signs appear to have an uncanny proclivity towards becoming Members of Parliament.Cancerians and Taurus almost have a celestial VIP access to the political arena! But, sadly, it appears that the stars have conspired against you, Scorpios. Perhaps they were engrossed in their own enigmatic allure, or were too preoccupied with scheming world dominance from the shadows. Don't be concerned, Scorpios; your cosmic destiny may have something even more fascinating in store for you, such as becoming a master spy or TV superstar.
I think ojas is the boss for this hahaha

As the mythological story in the Chinese zodiac goes, the Rat rode on the Ox but jumped off at the last moment to finish first in the race of animals, depriving the Ox of its victory. In political representation too the Rat seems to be enjoying success with a 4.2% advantage, at the expense of those born in the Year of the Ox (-5.3%). The prevalence of the remaining signs seems to follow their finishing order in the race…







Cancer and Rat are the best signs apparently…………… lol

UK deep dive

- University education (UK only)

There was a clear trend between going to Oxbridge and being a politician in the UK which is validated by the university and number of MP's percentage table. The graph; a visualisation of the table exhibits how 3 to two times more MP's attend Oxford and Cambridge than other top 10 universities. A substantially high number of politicians also attend LSE. This might be explained by the extensive political networks, influential alumni, and enhanced perceived competence gained from attending prestigious universities like Oxford, Cambridge, or LSE which positively influence the likelihood of becoming a politician. A Times Higher article based on testimonials of students from prestigious universities validates this stating how top-ranked universities tend to have stronger alumni networks as well as higher perceived degree value improving job prospects.

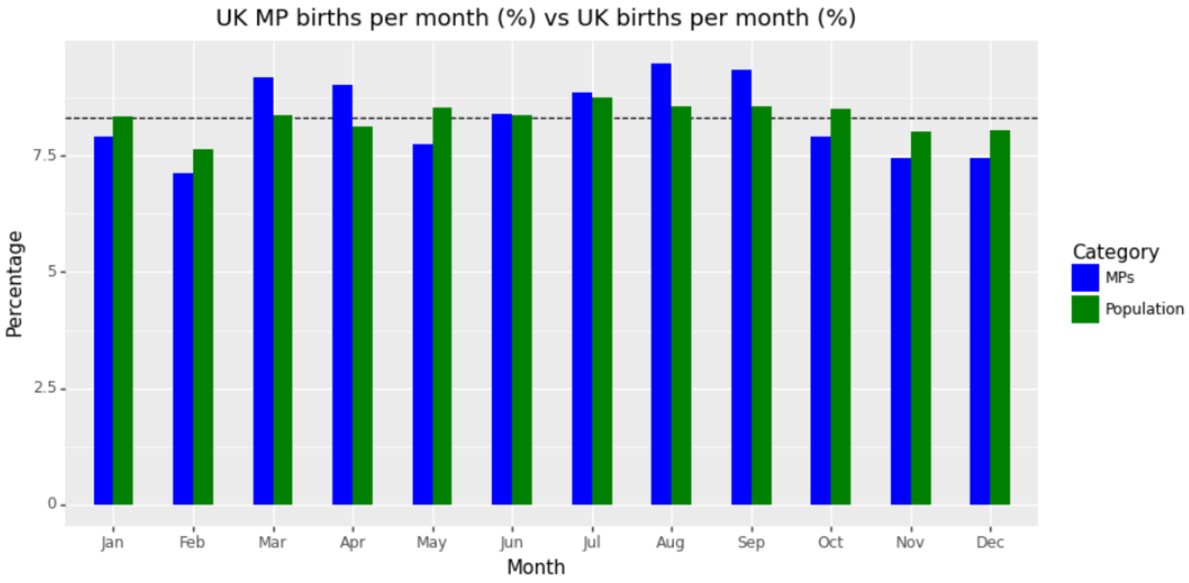
For instance

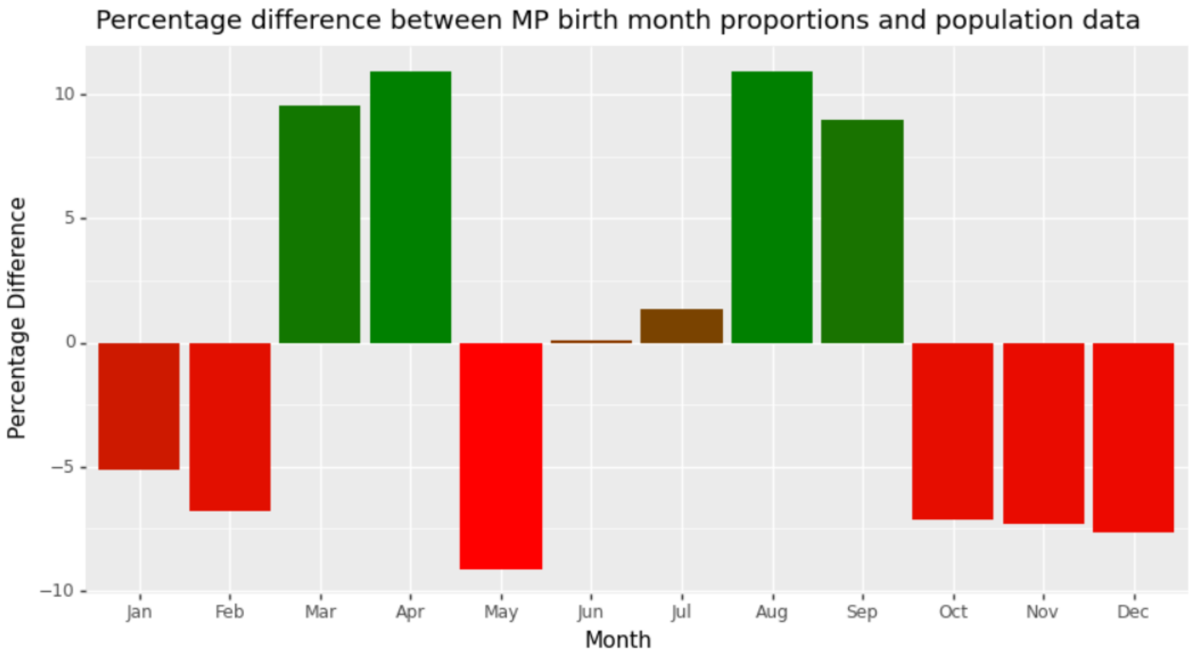
- Age (UK only)

The average age of a current UK MP was found to be 53.2 years…

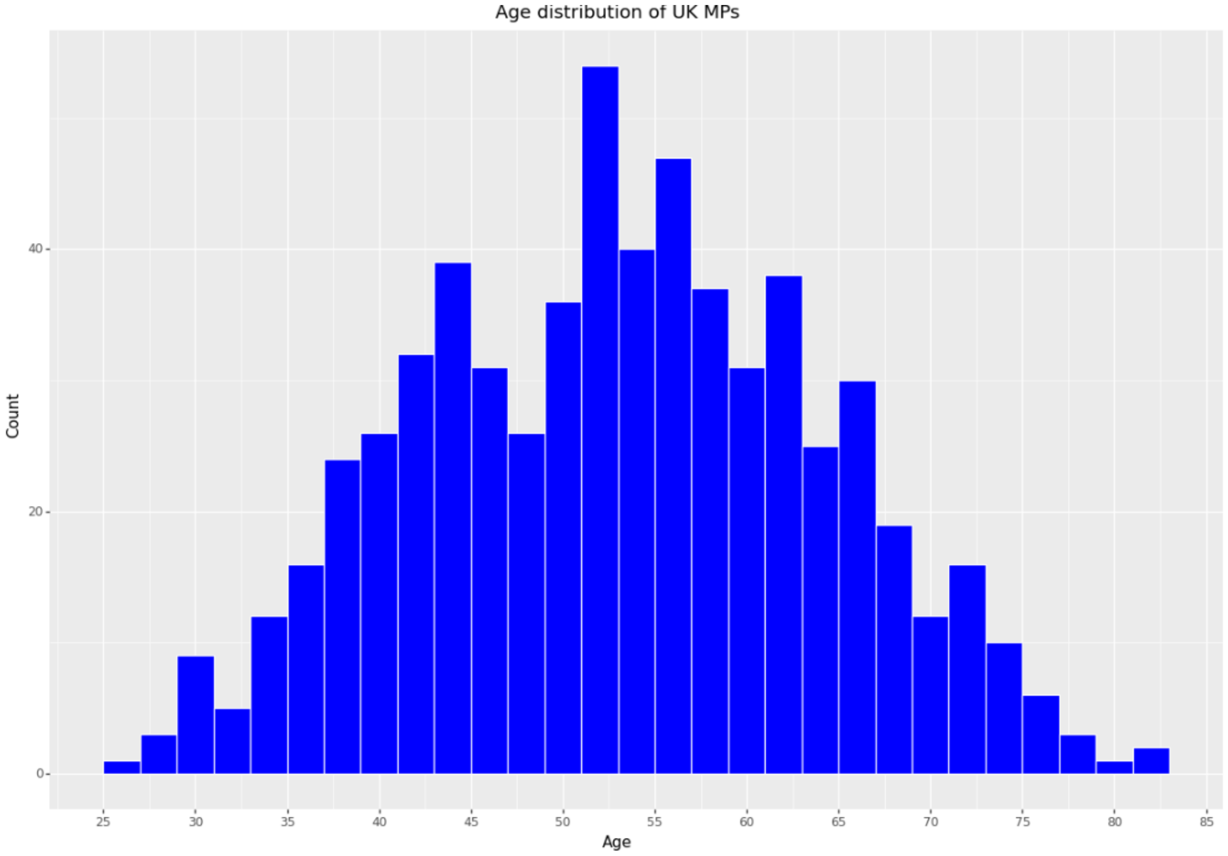
- Birth month for UK

Looking at the percentage difference between MP birth month proportions and population data, the general trend in the global dataset for being born earlier and becoming a politician doesn't hold. Although people born in March and April i.e initial months in the year are relatively higher than the rest, so are results for being born in September and August serving as a counterexample for a clear generalizable trend. However considering that the schooling system in Britain starts from September the earlier explanation of the relative age effect might still be at play. Specifically, the advantage of being born earlier in terms of maturity and development might improve the likelihood of future opportunities such as pursuing a career in politics, explaining why a much higher percentage of politicians 9.51 and 9.35 are born in August and September respectively. However, this is conflicting considering that the results are similar for March and April, 9.19 and 9.03 where the same reasoning doesn't hold. Overall, the results for birth month in the UK are not statistically significant (p-value=0.72).

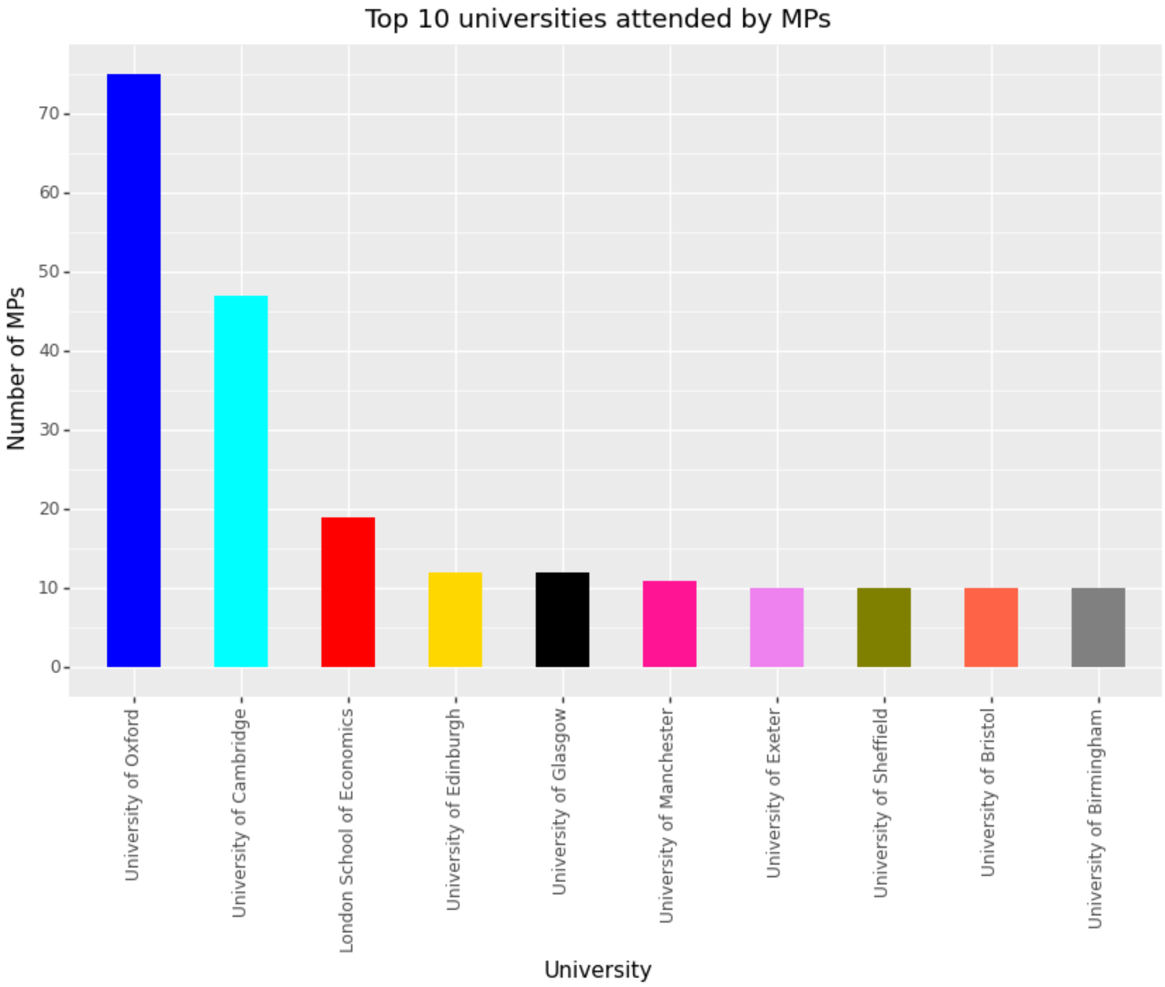




Percentage difference between MP birth month proportions and population data



Age distribution of UK MPs



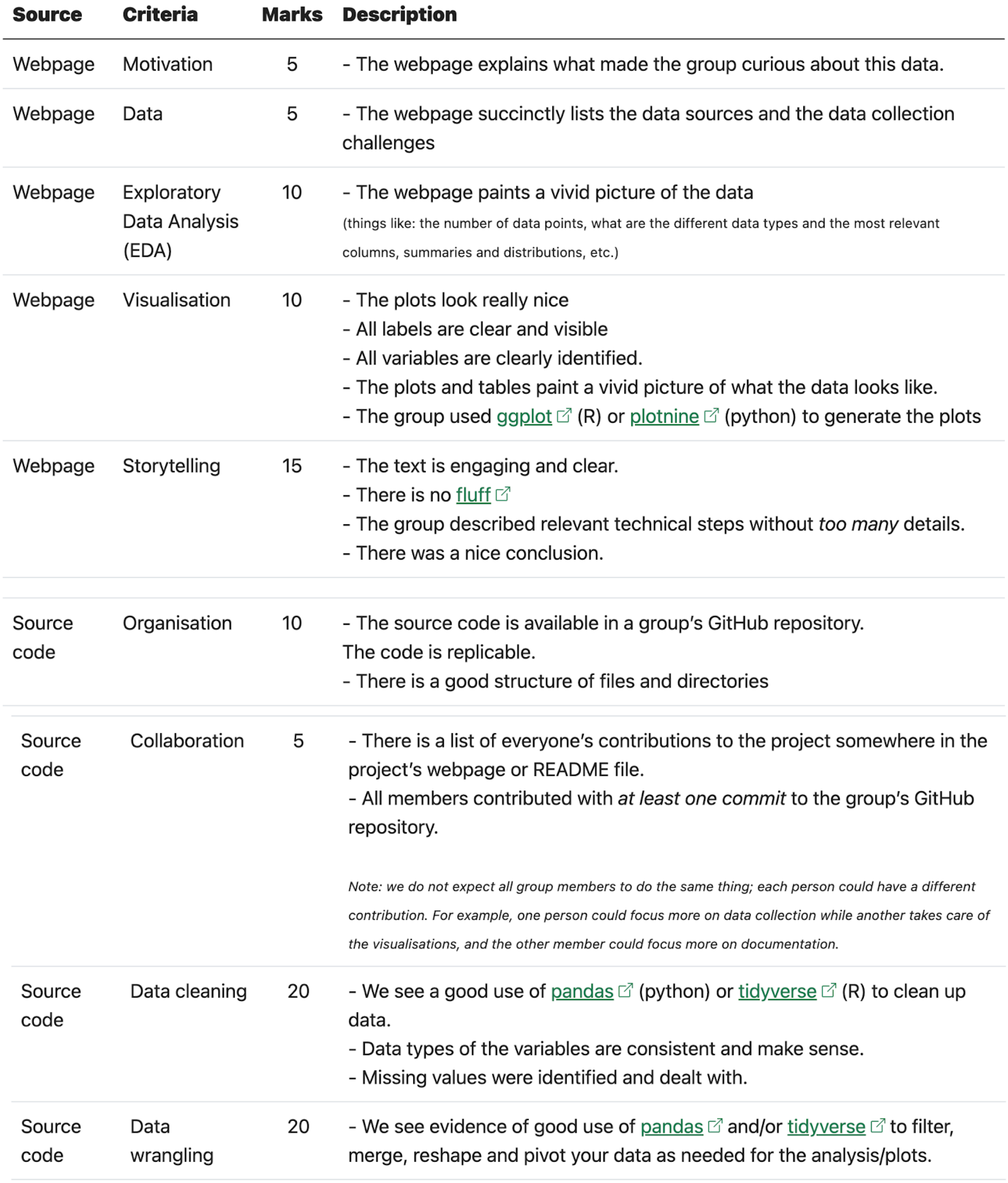
Summary / Conclusion / Key findings

Part of the reason for the literature identifying january skew in data could be data quality issues around missing values and strategies used to mitigate them

Advantage of being born in first half of the year
Strong gender gap robust to indicators

How to be a politician (or plan your kid to be a politician):
Birthday: 1st Jan (Or 29 Feb/1 Jun/20 Jan +32.63%)
Birth Year: 1958
Born on: Monday +2.73% (Sunday +16% for UK)
First letter of name: M
Length of name: 6
Name: Mehmet Wang
Gender: Male +55.6%
Zodiac: Cancer +7.2%, Year of the Rat (2032) +4.2%
University: Oxford if in UK

Mark scheme



Source	Criteria	Marks	Description
Webpage	Motivation	5	- The webpage explains what made the group curious about this data.
Webpage	Data	5	- The webpage succinctly lists the data sources and the data collection challenges
Webpage	Exploratory Data Analysis (EDA)	10	- The webpage paints a vivid picture of the data (things like: the number of data points, what are the different data types and the most relevant columns, summaries and distributions, etc.)
Webpage	Visualisation	10	- The plots look really nice - All labels are clear and visible - All variables are clearly identified. - The plots and tables paint a vivid picture of what the data looks like. - The group used ggplot (R) or plotnine (python) to generate the plots
Webpage	Storytelling	15	- The text is engaging and clear. - There is no fluff - The group described relevant technical steps without *too many* details. - There was a nice conclusion.
Source code	Organisation	10	- The source code is available in a group's GitHub repository. The code is replicable. - There is a good structure of files and directories
Source code	Collaboration	5	- There is a list of everyone's contributions to the project somewhere in the project's webpage or README file. - All members contributed with *at least one commit* to the group's GitHub repository. *Note: we do not expect all group members to do the same thing; each person could have a different contribution. For example, one person could focus more on data collection while another takes care of the visualisations, and the other member could focus more on documentation.*
Source code	Data cleaning	20	- We see a good use of pandas (python) or tidyverse (R) to clean up data. - Data types of the variables are consistent and make sense. - Missing values were identified and dealt with.
Source code	Data wrangling	20	- We see evidence of good use of pandas and/or tidyverse to filter, merge, reshape and pivot your data as needed for the analysis/plots.

Questions for the coders:

- The data sources we used
- Detail of the cleaning process
- The UK analysis error
- útiles error in the north month analysis

APPENDIX/OUTDATED CONTENT

Homepage/Contents + Motivation

Exploring whether certain factors/characteristics are correlated with being a politician - initially focusing on birth month, before extending to first letter of name, gender, and (comically) Zodiac sign. We initially focus our analysis on politicians across more than 200 countries, analysing general trends and differences between groups of countries (based on HDI, V-Dem, MYS, income group and hemisphere) before also providing a deeper dive into the UK to explore whether University/Alma Mater and Age seem to be relevant factors.
Observed correlation between birth month (relative age) on a range of outcomes such as:

Risk of diseases **(Winter -> Heart Diseases, Autumn -> Respiratory Diseases)**

Likelihood of succeeding in sports **(Jan > in AFL)**

Chance of becoming a top politician in Finland **(*thank you LSE*)**

Wanted to explore birth month trend was observable amongst a larger set of politicians, spanning as many countries as we could.
Identified a dataset from EveryPolitician that collected data for 233 countries across recent history.
Cleaned this, addressed issues with data quality (namely Jan skew).
Performed analysis to see whether we could observe any trends in birth month (comparing to UN births data) across all countries, and then performed the same analysis on subsets of countries (grouping by hemisphere, HDI, V-Dem score, Income Category and Mean years of Schooling)

Then, given the dataset we had, we wanted to explore some other potential trends in MP characteristics - such as the first letter of name of MPs (to see if donkey voting may have had an effect historically), MP gender and zodiac sign (as a fun exploration!). Again, we explored these for the subsets previously identified.

Here we encountered issues with data quality and relevance (as not only did this dataset include historically sitting MPs, but also had poor data quality for a few countries) - as such we sought to do a deep dive into the UK (as this is most interesting to us).

Here we wanted to see if we observed a correlation between the university attended and becoming a politician, and also see if age had a correlation (as this was not possible with the historical dataset). We also repeated the birth month analysis to see if a different correlation was observed for the UK specifically.

Analysis process

Ojas roadmap

Motivation

Politicians are the instrument of the government, they play an instrumental role in shaping the laws, policies, and decisions that govern our daily lives. The world has seen many politicians, from conservatives to liberals, dictators to democratic, aggressive to charismatic and efficient to futile, but what determines who becomes a politician? Is it possible that the explanation goes beyond the standard factors of education, party affiliations, and characteristic traits? Is it possible that underlying factors people completely overlook such as birth month have a bearing on who becomes a politician? Various studies have found relationships between birth month and disease risks or likelihood of succeeding in sports. For example, between those born in November and risk of ADHD, or those born earlier in the year and the chance of becoming a professional basketball player(Boland et al., 2015).The perks of positive birth month effects relate to a phenomenon called the relative Age effect. This describes the case where older peers have an inherent advantage over those born later due to factors potentially relating to physical/mental maturity and experience,leading to greater opportunities for development(Kiikka, 2018). So could this factor i.e birth month effects potentially have a correlation with becoming a politician? Are January babies more likely to rule the world?! Read further to find out …

Data Sources:

EveryPolitician - for core analysis dataset
Wikipedia - For UK-specific analysis
UN data (UN Statistics Division) - to compare population births data (baseline)
World Bank - for income categories
UNDP - for HDI, GII, MYS
ONS - for UK births data for day of the week
TheyWorkForYou - Filling gaps in Wikipedia Data

International Institute for Democracy and Electoral Assistance - For gender quotas
V-Dem Institute- For Electoral Democracy Index
Google Public Data - For hemisphere analysis

Data collection

To effectively explore the correlation between birth month and being a politician we needed a comprehensive dataset spanning multiple countries. After sifting across different datasets we decided to use EveryPolitician, which collects biographical data for over 78,000 legislators in 233 countries across recent history. Using a broad dataset that accounts for a wide range of countries and political systems provided us with a more comprehensive, global scope. Furthermore, because EveryPolitician is open-source and community-driven, the data is regularly updated and enhanced by volunteers, hence enhancing the data quality and completeness of our dataset. Moreover, Every Politician's data is in a structured and standardised format called Popolo where a person's birth date is represented as a string in the format YYYY-MM-DD. This standardised format makes web scraping easier because the birth date is always in the same format regardless of nation or politician, simplifying the data extraction process.

Data observations,initial limitations and fixes

After data collection, cleaning the data was necessary as, despite the advantage of using a more broad and complete dataset, there was still the possibility of working with incomplete or missing values which had the potential to confound results if they were clustered around a certain birth month. Following from this, issues with data quality lead to an initial overstatement of the likelihood of being a politician for January which was more than 4 times greater than the rest of the months. Upon deeper analysis, this was attributed to the standardised value of 1st Jan being assigned to politicians for which birth month data was not available. We identified this phenomenon as the January skew. To correct for this we omitted countries that had an unreasonable number of 1 Jan values. The threshold for the omission was computed by excluding countries where the ratio of records on January 1st was more than 10 times the expected ratio of records for Jan 1st and the number of such cases was at least 10. Seven countries for which this threshold was violated were excluded from the analysis, including Syria and Cameroon for which the proportion of people born on 1 January were 98.2% and 26.8% respectively. This differs greatly from the global proportion of 8.97% of births in January. Hence such results were removed from our birth month analysis to safeguard data quality, and then we were able to perform our analyses.

During the birth month analysis process, we noted (especially for some of the grouped analyses) that less-developed country data often contained far fewer entries and was generally less complete. When performing gender analyses, we also noted the impact of historical MPs/data dampening observed results.
Therefore, to account for these data quality issues (and facilitate deeper analysis of certain factors that the EveryPolitician dataset did not allow for, such as alma mater), we decided to perform a deep dive into the UK. From a data science perspective, focusing on one country also allows us to address concerns about variations in data availability, quality and consistency across regions.
For the UK, we wanted to see if we observed a correlation between university attended and being a politician, and also see if age had a correlation which wasn't possible with the historical dataset. We also repeated the birth month analysis to see if a different trend was observed for the UK specifically.

We collected the data for age and education from Wikipedia given its provision of alma mater and birth date in a scrapable fashion (although we did encounter some roadblocks in the process!) However, Wikipedia allows for public contributions, and to ensure data quality we cross-checked this Wikipedia data with another source, to limit errors in the dataset. This was indeed the case given that the data initially consisted of MP's that had resigned/passed away and (after accounting for this) was missing an MP! To identify the missing MP, we cross referenced using a website that sources its data directly from official parliamentary sources called TheyWorkForYou. Cross referencing produced 23 anomalies between the dataset however, given the nature of similar names (eg. Jon Ashworth vs Jonathan Ashworth) we needed to find the missing MP manually from the anomaly list identified by the regex search command. After correcting for this the dataset was ready for use containing data on all 650 MPs in the UK (i.e across Wales, Scotland, England and Northern Ireland)

Data trends

To explore the factors correlated with being a politician across the range of countries we decided to investigate trends in Birth Month, Gender, Zodiac (for fun) and Day of Birth. Beyond performing these analyses for the entire dataset, we also examined whether the trends changed between different groups of country (based on high vs low Human Development Index, Mean Years of Schooling, Income Group, V-Dem Score and Hemisphere). The general trends observed were as follows:

- Birth month

When looking at the chart for the percentage of MPs against month of birth, we noticed a slightly higher number of MPs relative to the number of births in each month before June, whereas afterwards the number of MPs was lower relative to births. This trend was noticeably more clear from the percentage difference graph, clearly illustrating the different trends pre and post June. We need to add analysis for each of the groupings

Gender;

The correlation between being a politician and being male is relatively robust to different groupings based on HDI,GII or other development indicators. It certainly showcases a heavily positive correlation between being a politician and being male. Although the disparity between being a politician given you are male versus female is significantly more apparent for economically less developed countries. This is shown most clearly by the percentage difference between being a politician given you are female for high GII countries when compared to low ones.

We need to add analysis for each of the groupingslook at page 8, this is skeleton of website - we need to put what we have so far into the relevant sections.

- First Letter of name: When looking at the first letter of name

Here

Im confused as to what this section is - we don't need this summary of results if we are going to go into them deeper in their respective sections. Think about how the website is organised - a home page then separate pages for each of these factors (where we will put every chart, and accompanying text describing trends and thoughts on them) - that is what the headings we put in this doc are (each of the website pages). Suggest we migrate these summaries into the relevant sections and build on them from there agreed wanted to ask about that as a group as well so do u suggest we simply list factors and 1-2 line trends? This

- I think the home page needs to be

- MOTIVATION
- DATA SOURCES & DATA
- ANALYSIS ROADMAP AND PROCESS
-
-
- Going by the marking scheme, then we go straight into the Actual analyses (on seperate pages in the website) okay so remove gen trends then i agree also for data sources because my first paragraph on the everypolician is quite extensive would u like something of similar length for other factors or half the length considering it spans subfactors i can just include the UN stuff as cross references but for other sources i mean
- Zodiac;

For entertainment purposes only we decided to also include zodiac signs and observed that people with the first 6 zodiac signs and by extension those born in the first 6 months have a relatively higher chance of becoming MP as visualised by the graph below. This was especially more pronounced for Cancerians and Taurians who were overrepresented relative to other signs. Being a Scorpio on the other hand …

- Day of birth;

There was a dip in MPs born on the weekend…
Specifically for the UK, when compared to population statistics for average number of people born on each day, MPs were more disportionately born on the weekends. This is visualised by the percentages table and graph which show positive values for those born on the weekends. This is significant considering that MP births are much lower compared to population average for all other days in the week shown by negative percentage difference values.

- University education (UK only)

There was a clear trend between going to Oxbridge and being a politician in the UK which is validated by the university and number of MP’s percentage table. The graph; a visualisation of the table exhibits how going to Oxford and Cambridge respectively correspond with a three and two times higher likelihood of becoming MP. Going to LSE also inflated the relative probability of becoming a politician :)))

- Age (UK only)

The average age of a current UK MP was found to be 53.2 years…

- Birth month for UK

Looking at the percentage difference between MP birth month proportions and population data, the general trend in the global dataset for being born earlier and becoming a politician doesn't hold. Although people born in March and April i.e initial months in the year are relatively higher than the rest, so are results for being born in September and August serving as a counterexample for a clear generalizable trend.

Deeper Exploratory data analysis

Birth month

Relative Age effect analysis - birth month + groupings

First letter of name and ~~length of name~~

Gender

- Groupings

Zodiac

Joke ;) Western + Chinese zodiac
There is a relatively weak trend between the zodiac of someone, which relates to birth month, and the distribution of MPs. Despite this, there potentially is a relationship for example Taurus, Gemini and Cancer all have a higher proportion of MPs, whereas Scorpio and Sagittarius, later on in the year, have lower proportions.

In terms of the 12-year Chinese zodiac cycle, politicians were born more in years of the Rat and Dragon, while less in the years of the Ox and Rooster.

UK deep dive

Data collection from Wikipedia
Cleanup & Fixing data
University, Birth Month, Age analysis

Summary / Conclusion / Key findings

Being born in the 1st half of the year vs 2nd half increases the likelihood of becoming a politician by 6.7%.