U.S. Presidential Predictions

Jacy Werner, Khaled Alketbi, Khalifa Alghaithi

2024-12-03

Introduction

Throughout U.S. history, presidents have been inaugurated into office from all backgrounds. While topically this may appear to be random, we set out to explore historical trends to analyze if there are common components that are key to a candidates success. This project provides a comprehensive analysis of U.S. presidential elections, focusing on various aspects of candidates and the factors influencing their credibility, public perception, and electoral success. This project analyzes two primary subsets that we believe contribute the most to a candidates success.

- Physical Characteristics
- Personal Credibility

Data collection has been dissected from historical records between the years of 1832 and 2024, in correlation between the Democratic and Republican parties. These party associations have been dominant throughout U.S. history which enacted the removal of separate parties for variable disturbances. Years prior to 1832 were found to be inconclusive when it came to data dissection, ultimately resulting in the decision to remove their incorporation.

Motivation

With the 2024 election occurring during this academic semester, we have found ourselves surrounded by electoral information. Which raised the question for this project, "is it possible to predict the outcome of presidential elections and identify factors that contribute to a candidate's success". The motivation for this project stems from the desire to understand the evolving characteristics and qualifications of U.S. presidential candidates and their relationship to electoral success.

Methology

For this project, we decided to analyze the data using Logistic Regression, SVM, Random Forrest, and Cross-Validation. To begin, we used the Logistic Regression Model as it is ideal for predicting binary outcomes. Throughout our data search, we have found that most of our data comes in binary variables such as YES and NO (1, 0), including facial hair, win, war, education, political history, and business owner. We used SVM Models for capturing complex relationships between features. This is beneficial for high-dimensional data as it works well with many predictors. Additionally, we were able to utilize Radial Basis Function (RBF) kernel to allow for flexible, non-linear decision boundaries. We used Random Forest models for analyzing feature importance. It handles categorical variables, over-fitting, and provided interpretative predictions. Cross-Validation was important for over-fitting as it ensured the model was evaluated across different subsets of data. It also provided performance metrics such as ROC, accuracy, sensitivity, and specificity for interpreting significance. When compared to other models such as KNN and Binary Regression, we found that these models consistently held more accurate performances.

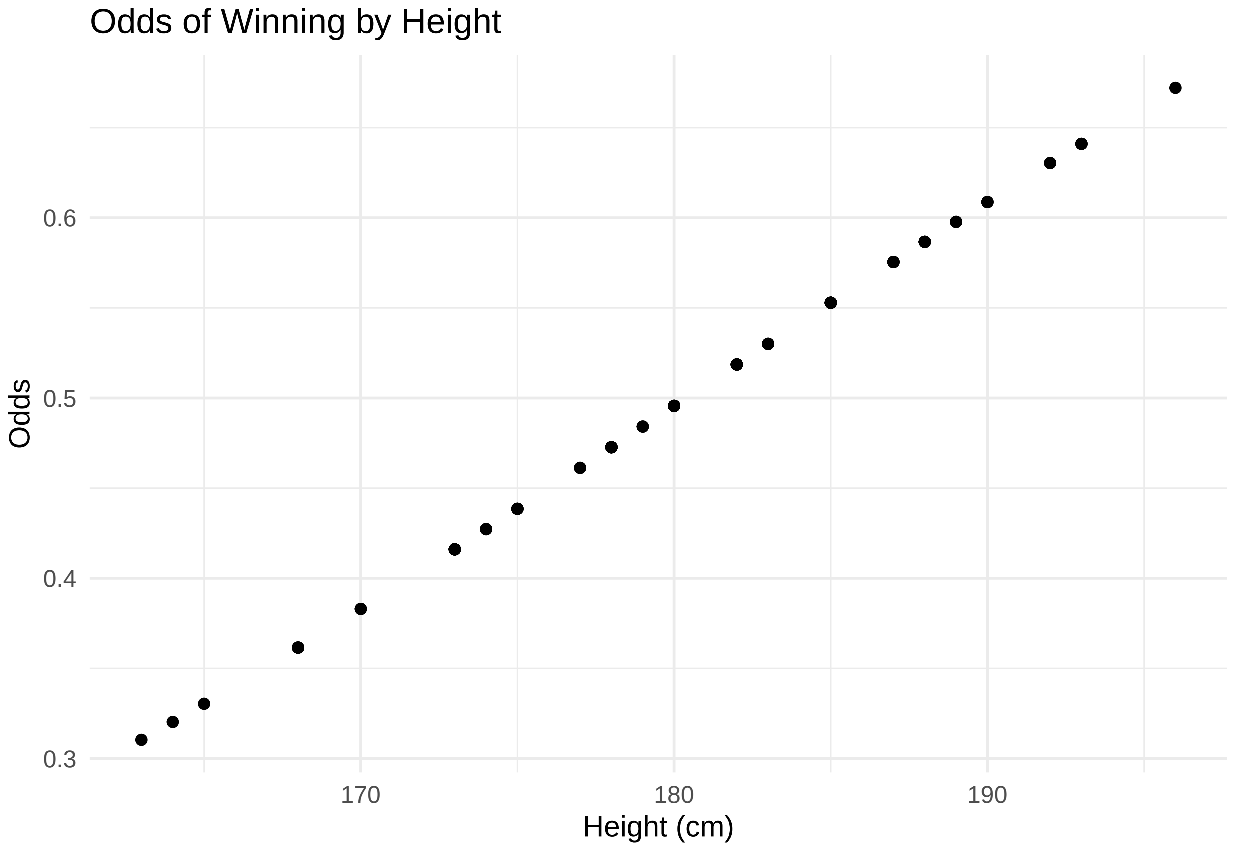
Constraints:

Through this project, we have found constraints that may damped the credibility of the data and the results. Politics are an opinionated topic, the direction of a voter can be influenced by their family nurture, wealth demographic, and geography. The ideologies of society have also changed within the past few centuries. For example, in history, obtaining facial hair was viewed as a symbol of "power" and "leadership". In modern times, clean-shaven individuals are viewed to be a more "professional" example of leadership and purity. Furthermore, education was not always important when inaugurating a president. We found that historical trends do not align with this ideology, as early presidents were often elected based upon their war leadership and public influence. Additionally, when retrieving the data, our group found retrieving data from over a century ago to be rather minimal and difficult. Throughout history, we have evolved in the way we capture/store data which can alter the data's significance.

Results

Candidates Height

Regarding a presidential candidates physical characteristics, we began by exploring their heights to explore if there is a correlation between height and their presidential inauguration. We found that height can play a symbolic role in public opinion as it is often viewed as a symbol leadership and authority, confidence, and visual representation. In order to do this, we implemented a logistic regression model as it best assessed binary classification between height and the target variable "Win". We first prepared the data by converting the target variable "Win" into a factor with two levels (No and Yes) which ensures the logisitic regression treats the outcome as binary classification.



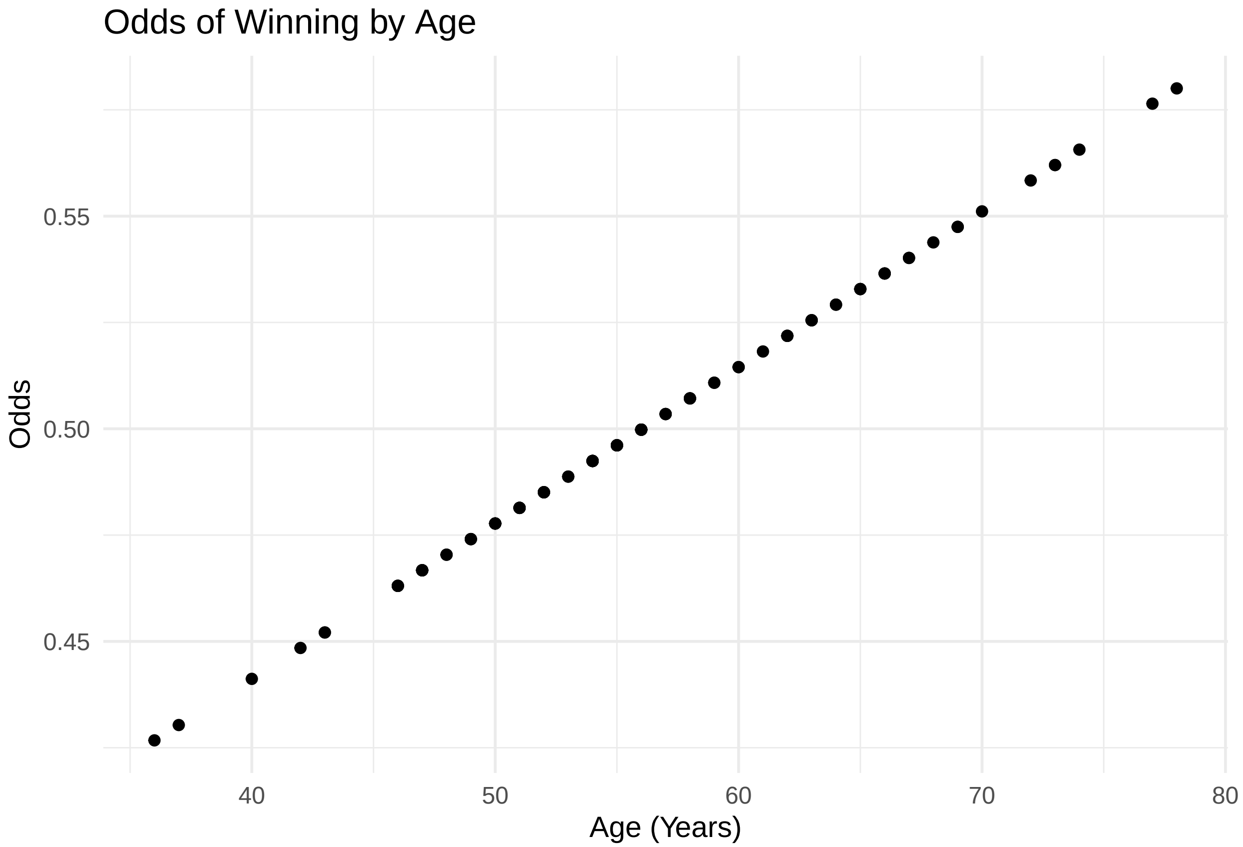
The logistic regression model shows a positive relationship between height and the likelihood of winning an election, with taller candidates having slightly higher odds of winning. For every 1 cm increase in height, the odds of winning increase by about 4.7%. However, the effect is not statistically significant as we obtained a moderately high p-value of 0.12, suggesting that height alone is a not a strong predictor of winning in the data set. Additionally, with a small reduction in deviance (133.36 to 135.86) and high AIC of 137.76, we see an indication of limited predictive power.

Candidates Age

The age of a president matters because it can influence public perception, governance style, and their physical and mental capacity to handle the demanding role of leading a nation. Age is often a topic for discussion within presidential elections because:

- Experience and Maturity
  - Young: Perceived as innovative and easily connectable to a younger audience
  - Older: Viewed as season, wise, and experience leaders
- Health
  - Health risks such as stamina, fitness, and acuite brain function
- Leadership style (caution vs risk)

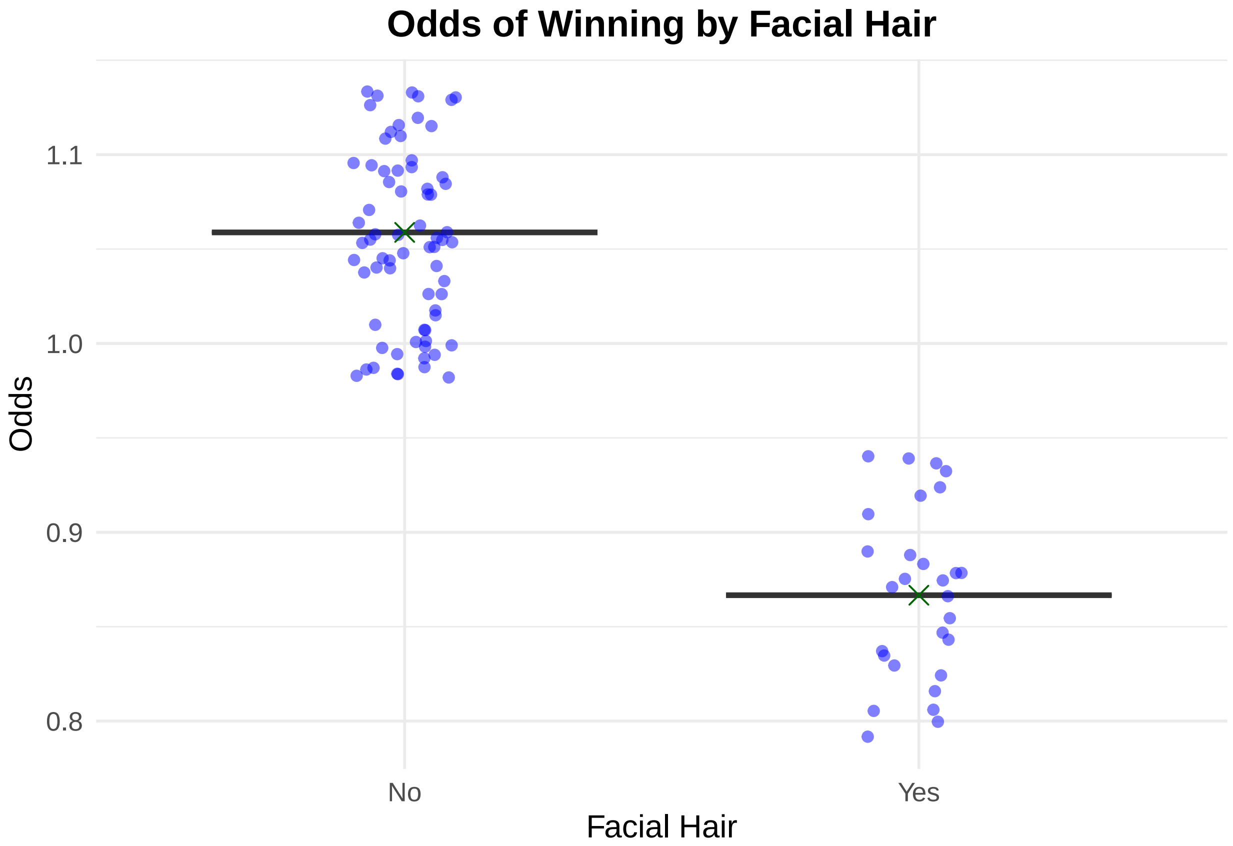
For this model, we decided to utilize the Logistic Regression Model with a model type of Binomial Logistic Regression.



The logistic regression model evaluating the effect of age on winning probability found that age alone is not a statistically significant predictor with a p-value of 0.538. While the model suggests that the odds of winning increase by approximately 1.5% per additional year, this effect was minimal and overshadowed by the statistic support. Additionally, the models AIC of 139.47 and low deviance reduction of 0.39 further confirms this analysis. Overall, age alone is not a meaningful predictor of winning elections in this dataset.

Facial Hair

While facial hair is not inherently important for a presidents ability to govern, it has played a symbolic role throughout history in which it shapes the public's perception. In times of war, it has been noted that a leader's beard gives them the perception of masculinity and power which has the opportunity of influencing the population for their vote. Additionally, in modern times, a clean-shaven appearance is often associated with professionalism and approach ability. Which can again give the perception of influence to a national audience. We continued using the logistic regression model with a model type of binomial logistic regression.

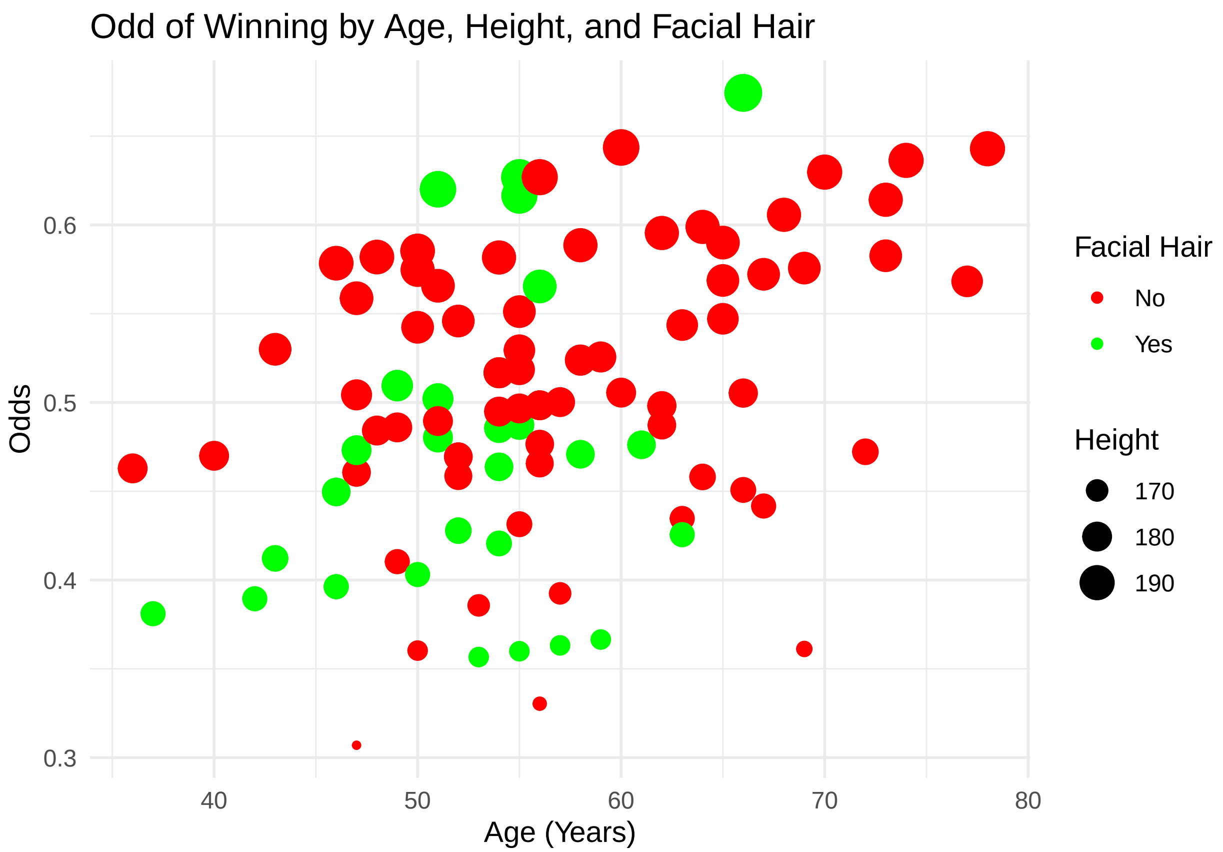


***

When analyzing the results, the calculated odds ratio for candidates with facial hair was approximately 0.82, indicating that their odds of winning were approximately 18% lower than candidates without facial hair. After converting the odds ratio to probability, we can see that the model predicted a winning probability of approximately 51.5% for candidates without facial hair and 46.5% for candidates with facial hair. However, with a high p-value of 0.655, we conclude that the effect was not statistically significant when analyzing for facial hair alone.

Combination of Physical Characteristics

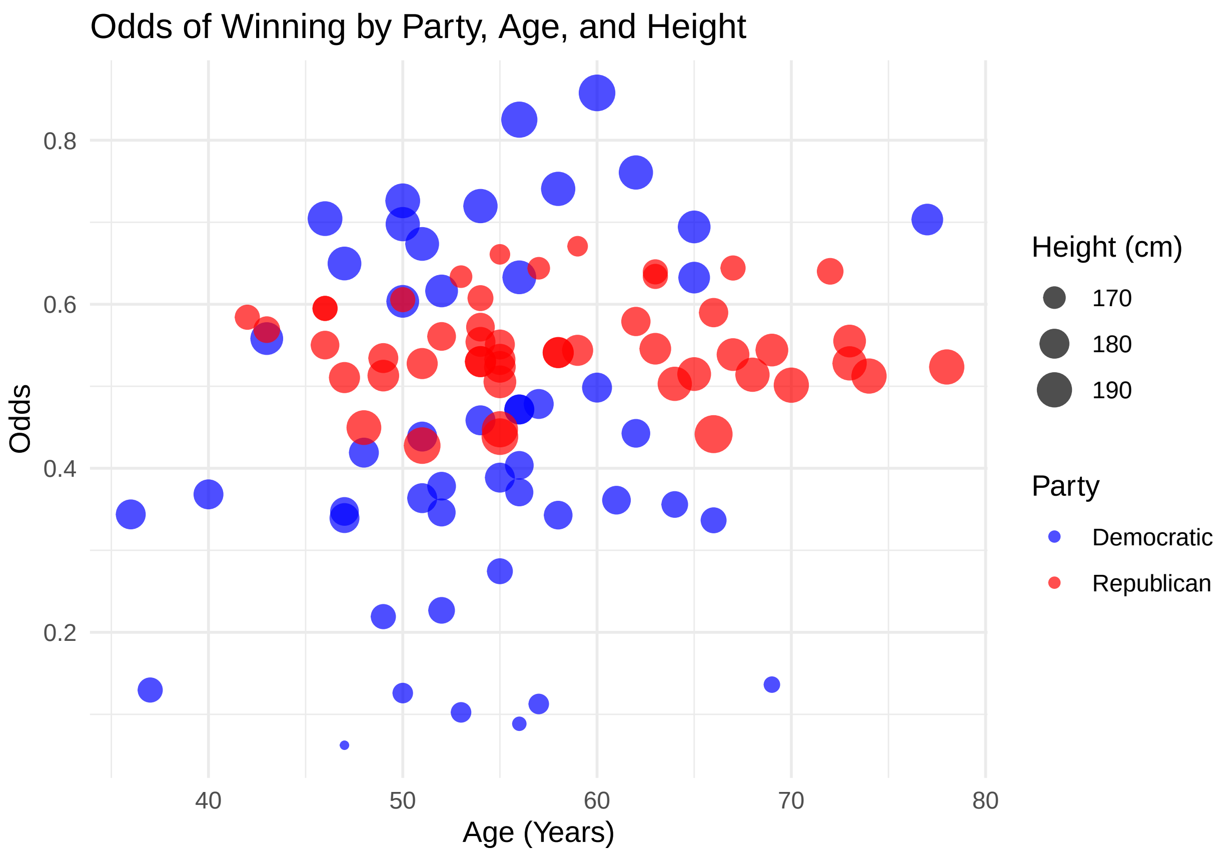
After collecting the results from the physical characteristics, we decided to incorporate the data simultaneously to continue analysis on whether these characteristics are statistically significant in the prediction of presidential winners. We again decided to use logistic regression model for combined predictors.



The combined model shows that height has the strongest positive association with winning, with odds increasing by 4.5% for every additional cm, but this effect is not statistically significant (p-value = 0.150). Age and facial hair contribute minimally, with no significant impact on the likelihood of winning. The model fit is weak, as indicated by a small improvement in deviance and an AIC of 141.26, suggesting physical traits alone are poor predictors of election outcomes. Overall, none of the traits significantly influence winning when combined, and other factors likely play a larger role.

Comparison of Party Preferences

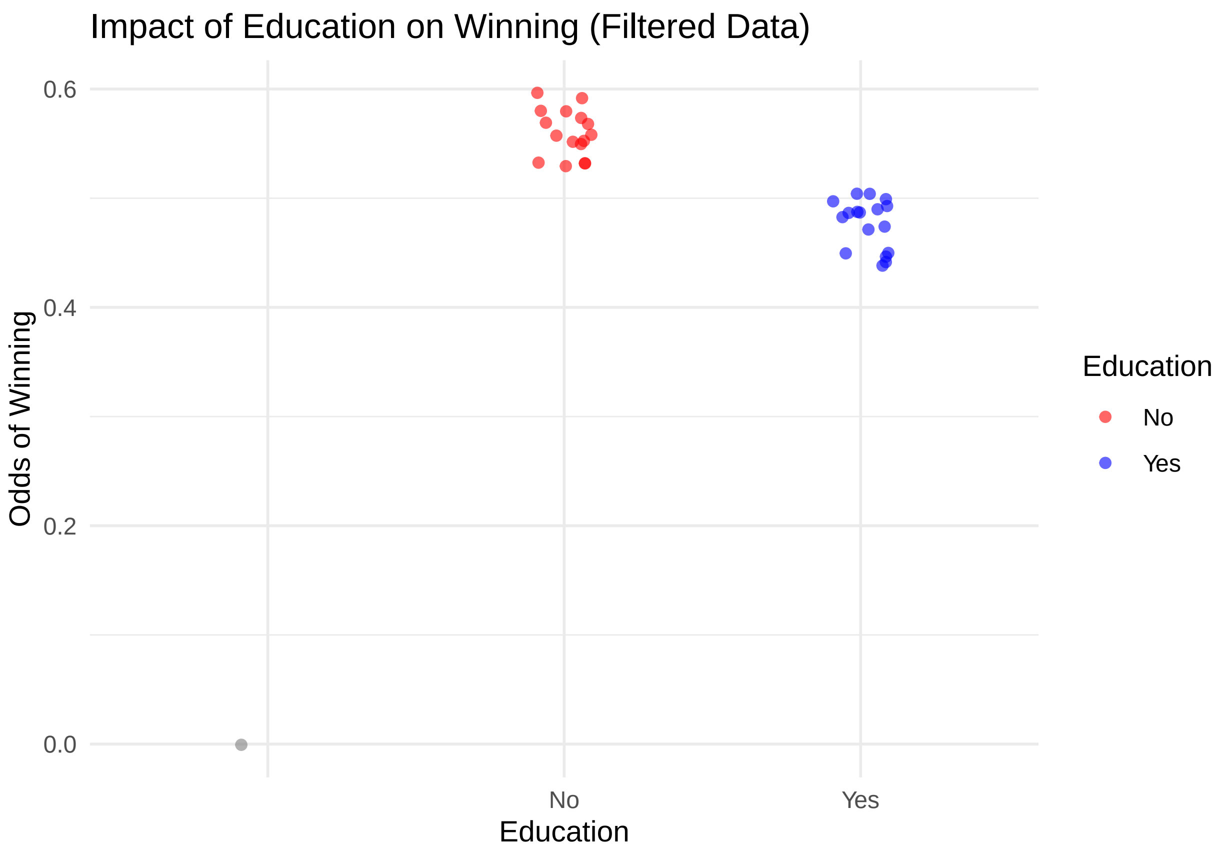
Additionally, we as a group wanted to analyze the importance of the physical traits- age, height, and facial hair to each political party. This analysis allows us to explore potential biases or trends within the candidates affiliated party (democratic and republican).By distinguishing these traits, we can explore whether physical attributes correlate with each party's nomination process and electoral outcomes. To prepare this data, we performed categorical conversions to ensure that the variables "party", "win", and "facial hair" are treated as factors, allowing for binomial logistic regression. We then set the target variable to "win" (binary) and the predictors of "height", "age", and "facial hair". We then separated the model summary to be an assortment by political affiliation (Democratic/Republican).



When analyzing the results, we can see that the logistic regression models built for Democrats and Republicans revealed different relationships between height, age, and facial hair. For Democrats, height significantly increases the odds of winning, with a 14.9% increase for every additional cm. With a p-value of 0.0125, we can see a statistically significant relationship between democratic height and political inauguration. In contrast, none of the traits (height, age, or facial hair) significantly influence winning for Republican candidates as there p-values were above 0.05. Overall, Democrats appear to care more about physical attributes, particularly height, than Republicans.

Candidate Education

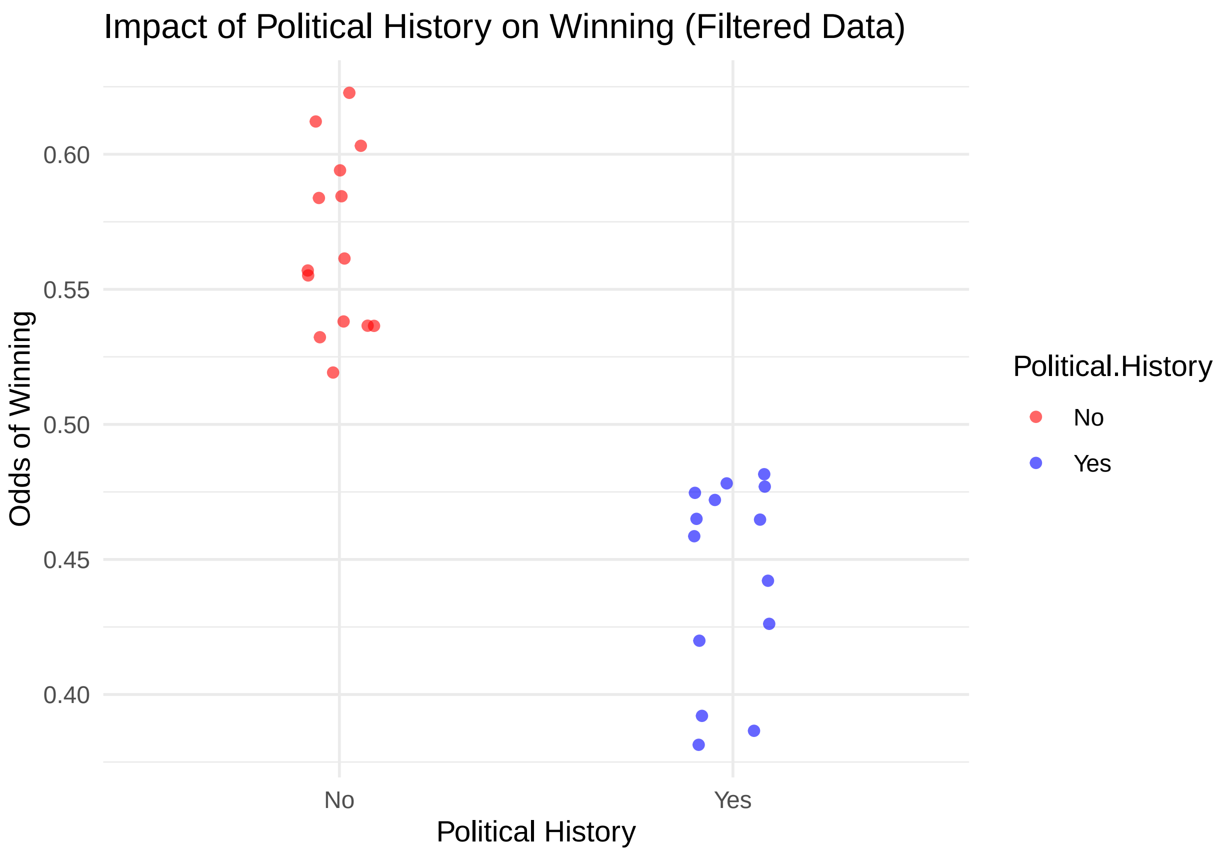
Moving forward, we as a group decided to explore how different different factors influence a candidates success in the presidential election. We first began by analyzing their personal credibility as a president. With components such as education, presidential experience, and business ownership, we believe that these are vital for public representation of knowledge in what it will take to run a nation. In modern political landscape, voters often look at educational achievement as a sign of intellect and preparedness for the complex responsibilities of governing. A well-educated president displays the qualities for informed decision-making, critical thinking, and strategic planning. We decided to use a Logistic Regression Model as the variable “EDUCATION” was binary. LRM models work effectively here as they are ideal for handling binary variables and their statistical inference on statistical significance. After exploring other models (such as SVM, Random Forrest, and Binary Regression), we found that the Logistic Regression Model gave us the most reliable data in terms of significance. Additionally, we decided to filter the data where the candidates held the same educational level to remove bias and provide more meaningful comparisons.



The logistic regression examined the impact of education on the likelihood of winning. The model estimated the log-odds of winning as a function of "Education". Using the model's predicted values, the predicted probabilities of winning were calculated and visualized. The resulting properties were approximately 0.55 for no education and 0.48 for education. While this highlights a difference, given the extremely high standard errors of 2399.54 and high p-values of 0.994-0.995, the model indicates that the level of education is not a statistically significant predictor of winning.

Candidate Political History

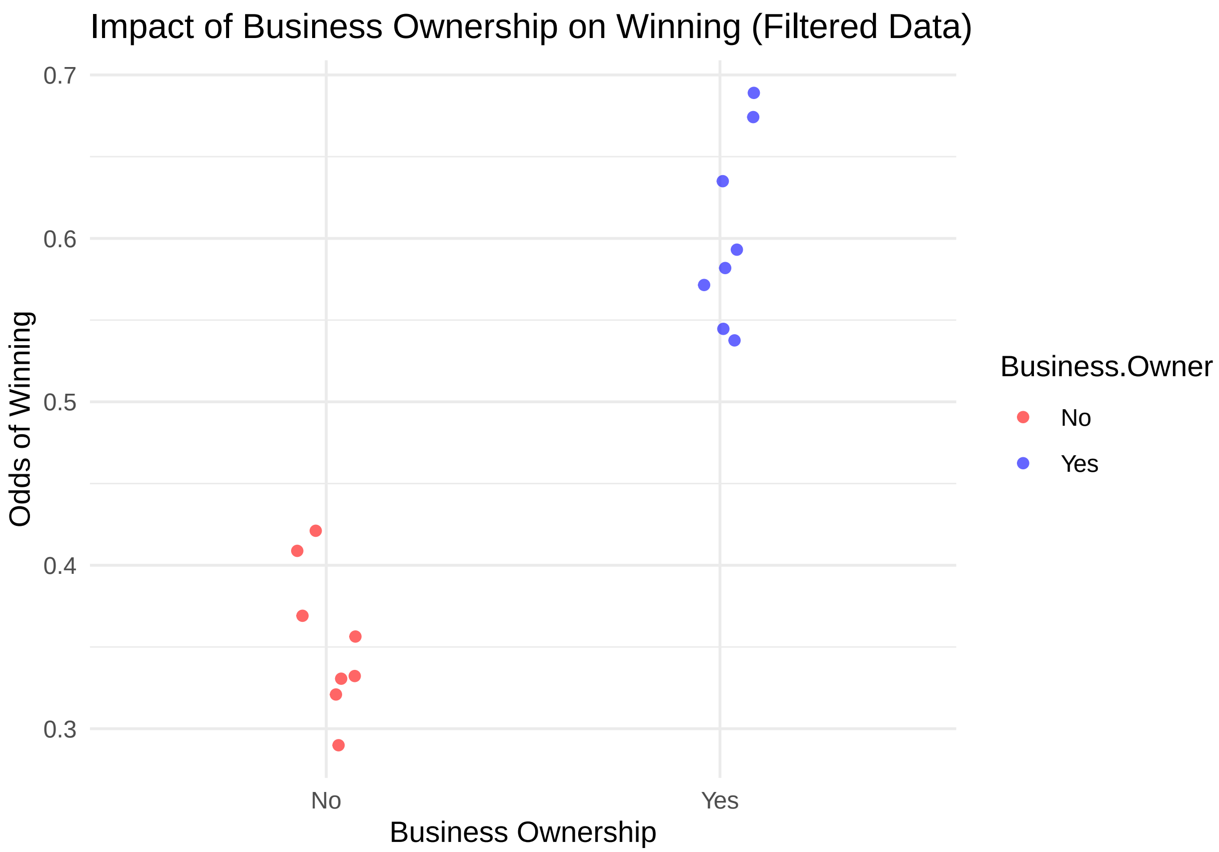
A candidates political history is also a significant contributor to an individuals credibility. By analyzing their political history, we can view how they have handled previous political turmoil. This can shape the public's perception as it displays their problem-solving, ideologies, and success. For this segment, we also decide to filter the data where the candidates held the same political history to remove bias and provide more meaningful comparisons. Again, we decided to use the logistic regression model as the variable "Political-History" is entirely binary. Additionally, this produced the most promising outcomes in terms of significance to other models.



The logistic regression examined the impact of political history on the likelihood of winning. The model estimated the log-odds of winning as a function of "Political-History". The log-odds of winning with no political history converted to predictive probability is approximately 0.57 while having political history is approximately 0.44. While there is a difference, there is a large p-value of 0.451 which indicates that the difference is not statistically significant. There is also a large standard error of 0.7638 which supports the idea that political history alone does not strongly predict election outcomes in the dataset.

Candidate Business Ownership

The public often views a presidential candidate with a business background as a strong contender due to their perceived leadership, decision-making skills, and economic expertise. Business owners are typically allocated with the characteristics of strategic-thinking, financial management and the ability to navigate complex negotiations. These factors can influence electoral success with economic stability in mind. For this segment, we decided to use logistic regression model to continue analyze binary variables.



The logistic regression model reveals that candidates who are business owners have an odds ratio of 2.78, meaning they are almost 3 times as likely to win compared to non-business owners. Candidates with no business ownership hold an odds ration of 0.60. When converted to probability, we can see that business owners have a 63% probability of winning compared to 38% to no. However, this effect is not statistically significant (p = 0.323), so the result could be due to random variation. Additionally, with the small reduction deviance of 1.011 and high p-values, we can see that business ownership alone is not a strong predictor of winning.

Comparison of Education, Political History, and Business Ownership

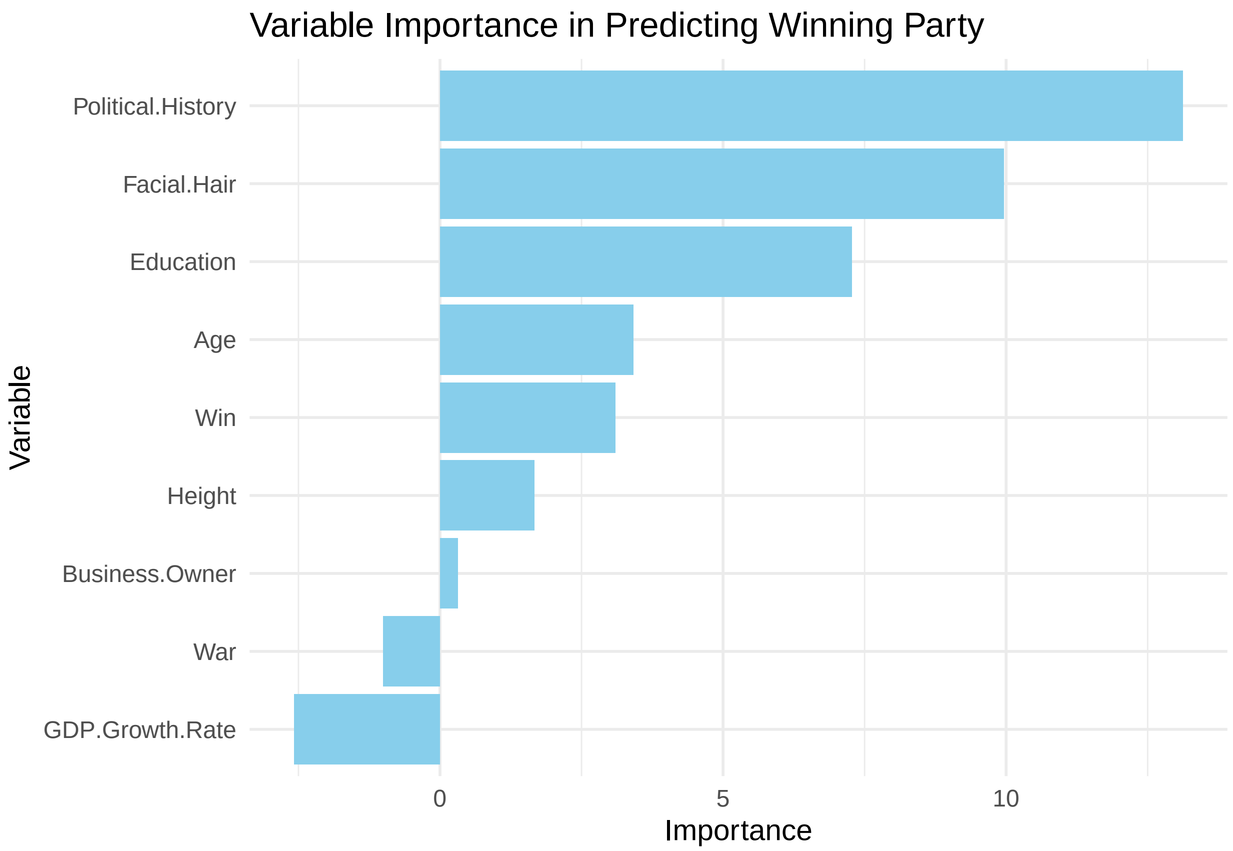
Moving forward, we decided to analyze the relationship of Education, Political History, and Business Ownership to understand their influence on elections. For this analysis, we decided to use a random forest model for a few reasons:

- Ability to Handle Categorical Data
- Non-Linear Relationships
- Predictive Power

This approach allowed us to collectively assess the candidates likelihood of winning. In order to do so, we held the target variable as "Win" with the predictive variables of "Education", "Political-History", and "Business Owner". We set the ntree to 500 which will help reduce the variance through averaging and building more stable predictions. When assessing the results, the OOB error rate is approximately 62.24% which suggests that the model classified the winning outcome 62% of the time. This high error rate suggests a poor model performance which can be a result of the small dataset. The confusion matrix made 49 predictions with a 34.7% error rate when guessing the non-winners compared to a 95.9% error rate for the winners. Again suggesting a struggling model to predict successful candidates.

Analysis of Variable Importance

Throughout this project, we have analyzed every variable and their influence on presidential predictions. We now explore how every variable together influences a candidates inaugurational success by party and examine their statistical significance. Again, we decided to use a random forest model because of the ability to handle categorical data non-linear relationships and its predictive power. We used a classification model type in predicting the political party (Democratic or Republican) with the number of trees being 500 (stability and accuracy).



When looking at the results, we can see that a candidates political history is the most important in contributing the random forest predictive power of 12.38%, while the GDP Growth Rate had the lowest with negative 3.10%. The model achieved an OOB (Out-of-Bag) error rate of 31.63%. Meaning that the model's overall prediction accuracy is approximately 68.37%, which is moderately good given the size of the dataset. Additionally, the model had a class error of 28.57% for Democrat and 34.69% for Republican. In conclusion, we can see a reasonably balanced classification performance for both parties and they both showed similar error rates. This indicates that the model does not favor one party over the other, suggessing a relatively fair and unbaised classification process.

The Perfect Candidate

After examining all the variables, we wanted to analyze if it would be possible to predict the outcome of an election given a candidates characteristics and traits. We first prepared the data by doing categorical conversion of the variables "Party", "Education", and "War" into factors. We then did numeric conversion of the variables "Height", "Age", and "GDP.Growth.Rate". We then decided to use a Support Vector Machine (SVM) Model Training as it is beneficial for classification tasks involving complex, non-linear relationships. SVM handles high-dimensional spaces efficiently and is effective with small to medium sized datasets. We used a Radial Basis Function (RBF) Kernel which is suitable for non-linear boundaries. Additionally, we used Cross-Validation to utilize a 10-fold cross-validation, ultimately splitting the data into 10 subsets for more reliable evaluation.

```
## [1] "Best Candidate Profile:"

##     Height Facial.Hair Age War GDP.Growth.Rate Education Political.History
## 442    165          No  50 Yes               2       Yes               Yes
##     Business.Owner      Party Predicted_Prob
## 442             No Democratic      0.7050638
```

With 98 samples (dataset), 9 predictors (variables) and a target variable of whether the candidate wins or not, we can see moderate performance in this model. The model recieved the highest ROC (Receiver Operating Characteristic) Score of 0.585 with an optimal Regularization Parameter (C) of 1.00. This suggests that the model was able to predict the winners from losers with a slightly better variance of 58.5%. The model better predicted winners with 60% correctly predicted compared to 52% in predicting losers. The model then predicted the most likely winning profile, which identified a male, 165cm tall, 50-years-old, clean shaven, an active war, GDP of 2%, Educated, Politically experienced, non-business owner, and Democrat which held a predicted probability of 70.51%.

Conclusion:

Throughout U.S. history, presidents have been inaugurated into office from all backgrounds. Which raised the question for this project, "is it possible to predict the outcome of presidential elections and identify factors that contribute to a candidate's success". More specifically, we decided to analyze a candidates physical traits ("Height", "Age" and "Facial Hair") and personal credibility ("Education", "Political History" and "Business Ownership"). Based upon the results from our explored SVM, Random Forest, Logistic Regression, and Cross-Validation models, we can see that there is minimal statistical significance that indicates accurate presidential predictability. Despite using advanced models, there was no single factor that definitively predicted electoral success. Regarding party-specific insight, we saw a statistically significant (p-value = 0.0125) effect on height, and a democratic candidates chance of winning. For every 1 cm increase of height, there was an increased odds of 14.9%. Additionally, using SVM modeling, we were able to predict the "perfect candidate profile" which predicted the most likely winning candidate to be:

- Height: 165cm
- Age: 50 Years
- Facial Hair: Clean-Shaven
- Education: Educated
- Political History: Experienced
- Business Owner: No
- Party Affiliation: Democratic

This candidate has a predicted probability of winning to be 70.51%. Which raised the question, why is the "perfect candidate" shorter, this can be explained by multivariate effects. The model considered multiple variables simultaneously, such as political history, education, and party affiliation. If shorter candidates in the dataset were more experienced, better educated, or affiliated with a winning party. The model could have over weighted their success despite the given height. We then decided to explore the constraints and how they could have influenced the results. Historical bias played a crucial role as cultural and political ideologies shifted over time. Affecting how traits such as facial hair and education were perceived. Additionally, data collection was proven to be very sparse throughout our research. With an election occurring once every four years, we had a smaller data set to analyze trends to make significant predictions. As data collection has evolved over time, data statistics from over a century ago were found to be turbulent between historical datasets. This infers that there is often misleading data that could provide a negative influence to the models results.

While the data and model results are predominately inconclusive, this data can be statistically significant in the next few centuries as there is more data to analyze and interpret. But in the meantime, we conclude that there is not enough data to successfully predict the outcome of presidential elections.

Sources

- Wikipedia - 1924 United States Presidential Election
- Wikipedia - Heights of presidents and presidential candidates of the United States
- Wikipedia - List of United States major party presidential tickets
- Coolidge Foundation - 1924 High Tide Conservatism
- Coolidge Foundation - Hoover vs Smith
- Hoover Blogs - A troubled relationship
- Britannica - Alf Landon
- Britannica - Wndell Willkie
- Empire State Plaza - Thomas E Dewey
- White House - Harry Truman
- Miller Center - Adlai E Stevenson II
- Columbia - Dwight D EisenHower
- White House - John F kennedy
- Nixon Foundation - Richard Nixon
- Warroom - Thanksgiving 1968
- Britannica - Barry Goldwater
- History - Hubert H. Humphrey
- Britannica - George McGovern
- Wikipedia - Jimmy Carter
- Wikipedia - Gerald Ford
- California Museum - Ronald Reagan
- Britannica - Walter Mondale
- North Eastern - Michael Dukakis
- Britannica - George H. W. Bush
- White House - William J Clinton
- Wikipedia - Bob Dole
- Bush Center - George W. Bush
- Britannica - John Kerry
- NPR - Obama
- New Yorker - John Mccain
- New Yorker - Mitt Romney
- Wikipedia - Donald Trump
- White House - Biden
- Britannica - Andrew Jackson
- Britannica - Henry Clay
- White House - Martin Van Buren
- Battle Fields - William Henry Harrison
- Britannica - Lewis Cass
- Britannica - James K. Polk
- White House - Zachary Taylor
- Wikipedia - Franklin Pierce
- Britannica - Winfield Scott
- White House - James Buchanan
- ThoughtCo - John C. Fremont
- NPS - Stephan A. Douglas
- White House - Abraham Lincoln
- Ohio Civil War - George B. McClellan
- MrLincoln - Horatio Seymour
- History Hit - Ulysses S. Grant
- Mental Floss - Horace Greeley
- Wikipedia - Samual J. Tilden
- White House - Rutherford B. Hayes

- Emerging Civil War - General Hancock
- White House - James Garfield
- White House - Grover Cleveland
- Wikipedia - James G. Blaine
- Wikipedia - Benjamin Harrison
- Wikipedia - William Jennings Bryan
- McKinley President
- Wikipedia - Alton B. Parker
- Britannica - Theodore Roosevelt
- NWHOF - William Taft
- 239 Days - Woodrow Wilson
- Wikipedia - Charles Evan Hughes
- Wikipedia - James M. Cox
- White House - Warren G. Harding
- Wikipedia - U.S. Presidential Election
- History - Abraham Lincoln
- Investopedia - Unemployment Rate
- NBER - Gross Nation Product
- Oxfordre - Economic Growth
- Wikipedia - Democratic Presidential Candidates
- Wikipedia - Republican Presidentail Candidates
- Wikipedia - Presidential Candidates
- CNN - 2024 Presidential Candidates
- Wikipedia - Mitt Romney
- Britannica - Hillary Clinton
- Wikipedia - Donald Trump
- Wikipedia - Joe Biden
- White House - Kamala Harris
- PLOS - Strength of Age
- PEW Research - Harris and Trump
- Elsevier - Facial Appearance and Leader Choice