# Effect of Gender Attitudes And Demographics on Female Leadership

Class : Statistics 112

Section : Discussion 1B

Group number : 2

Members : Luis Carias (806021988), Tyler Chia (105940878), Claire Nabours

(206014740),  Judy Zhu (305845639)

# Table of Contents

# Statement of the Problem

Does a person's demographics and their attitudes toward women shape voting behavior, particularly in terms of their support for women in public office and their political party preference?
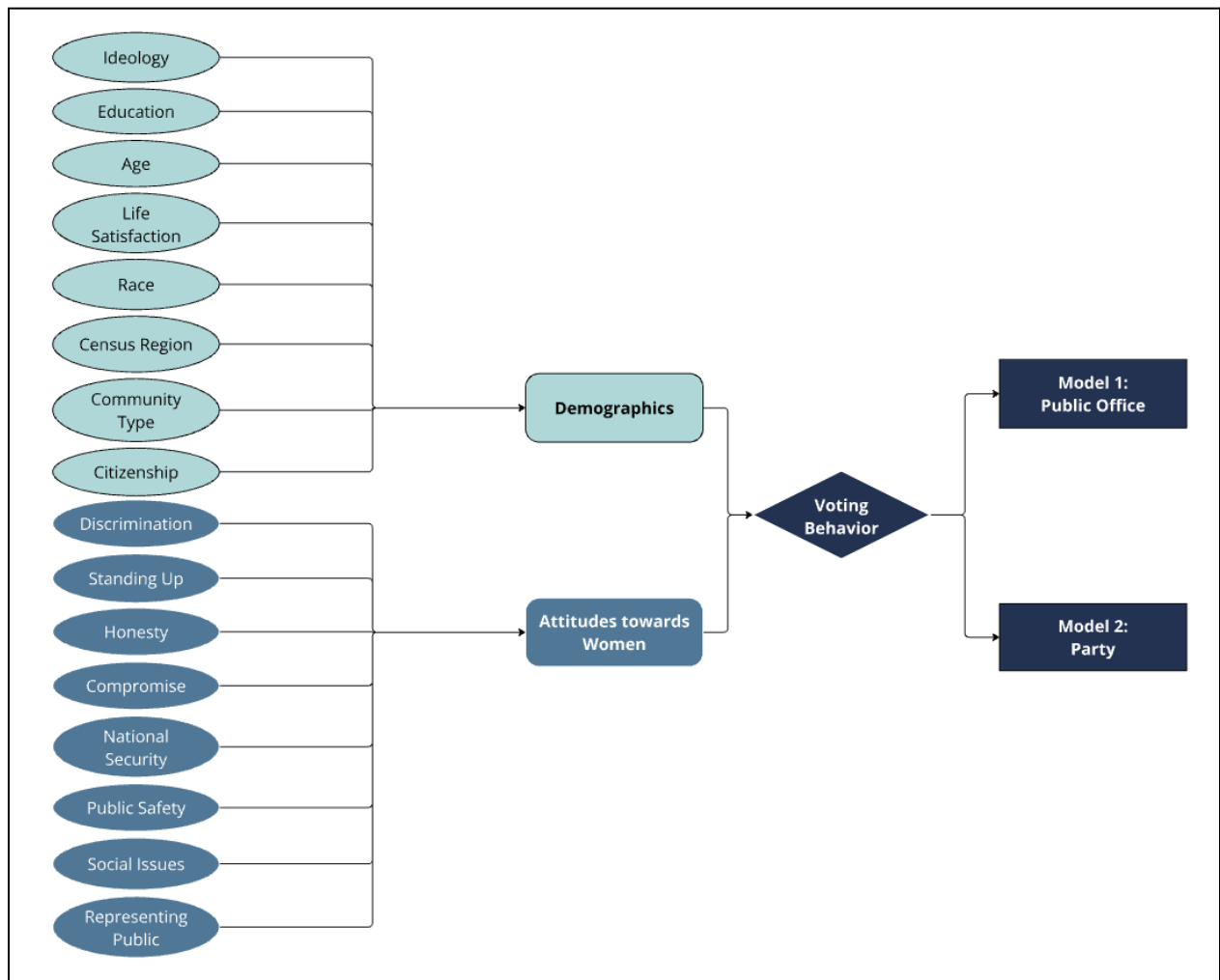


*Figure1*. Path Diagram

**Data Set:** *sexdiscrimination.csv* provided by Professor Esfandiari through BruinLearn.

# Abstract

In the wake of the recent 2024 election, the objective of our project is to uncover how demographics and attitudes toward women influence voting behavior, specifically in supporting women for public office and political party preference, and to find which group of predictors are more significant predictors. For public office outcomes, we used multiple linear regression, refined regression models, and random forest analysis. Attitudinal measures such as perceptions of honesty, national security, and leadership were consistently found to be significant predictors, whereas demographic factors like race and education were insignificant predictors. For political party preference, logistic regression and random forest models highlighted life satisfaction, age, and attitudes toward government honesty, and other attitudinal measures as key predictors. Our models and EDA revealed that individuals with favorable. or neutral perceptions of women exhibited the strongest support for women in leadership roles, despite their demographics. Challenges in our project included addressing overfitting in regression models and overcoming missing values within our dataset. Overall, our project found voting behavior to be determined by attitudes towards women more than demographics.

# Model Variables

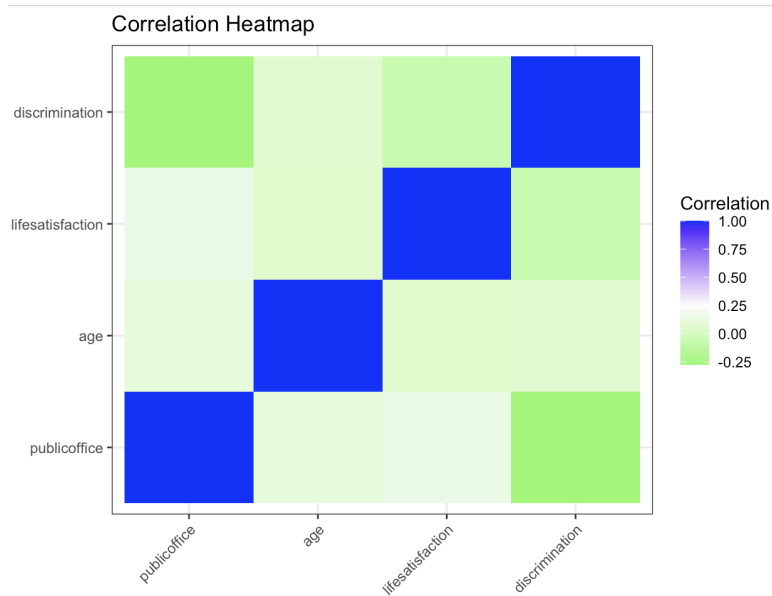| Variables | Variable Type | Measurement | Description |
|---|---|---|---|
| Public Office (publicoffice) | Outcome | Numerical (0–100) | Positive attitude toward recommending women for public office; high scores indicate strong endorsement. |
| Party | Outcome | Categorical (Republican, Democrat) | Respondent's political party affiliation. |

**Demographics**

| Variables | Variable Type | Measurement | Description |
|---|---|---|---|
| Ideology | Predictor | Categorical (Conservative, Very Conservative) | Respondent's political ideology. |
| Education (edu) | Predictor | Categorical (High school or less, Two-year college, Four-year college, Graduate) | Respondent's level of education. |
| Age | Predictor | Numerical (years) | Respondent's age in years. |
| Life Satisfaction (lifesatisfaction) | Predictor | Numerical (0–100) | Respondent's overall life satisfaction; higher scores indicate greater satisfaction. |
| Race | Predictor | Categorical (Black, White, Others) | Respondent's racial identity. |
| Census Region (censusregion) | Predictor | Categorical (Midwest, South, Northeast, West) | Geographic region where the respondent resides. |
| Community Type | Predictor | Categorical (Urban, | Type of community in which the |

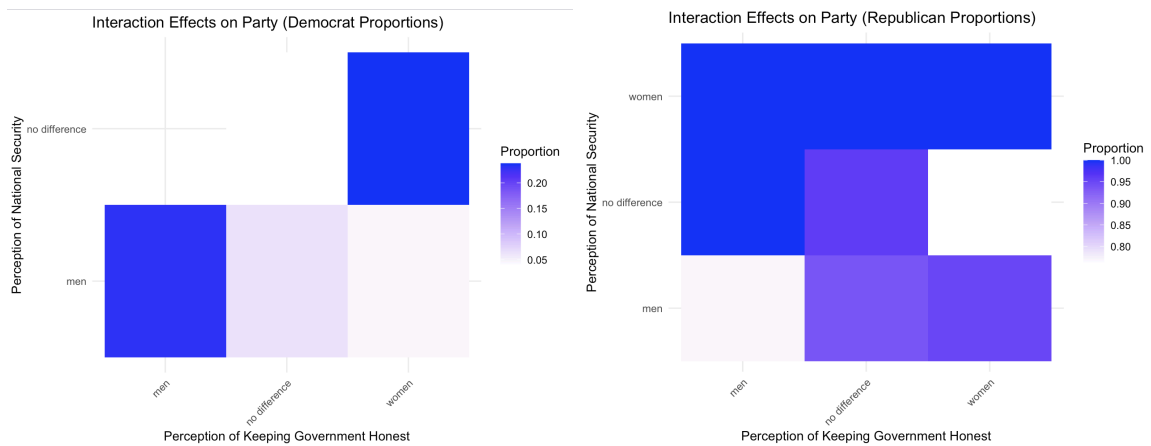| | | | |
|---|---|---|---|
| (communitytype) | | Suburban, Rural) | respondent lives. |
| Citizenship | Predictor | Categorical (US citizen, Not US citizen) | Respondent's citizenship status. |

**Attitudes Toward Women**

| Variables | Variable Type | Measurement | Description |
|---|---|---|---|
| Discrimination | Predictor | Numerical (0–100) | Attitudes toward discriminating against women; higher scores indicate more discriminatory attitudes. |
| Standing Up (standingup) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is more likely to stand up. |
| Honesty (keepgovhonest) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is more likely to keep the government honest. |
| Compromise | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is more likely to compromise. |
| National Security (nationalsecurity) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is more effective at ensuring national security. |
| Public Safety (publicsafety) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is more likely to maintain public safety. |
| Social Issues (dealsocialissue) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is better at addressing social issues. |
| Representing Public (representpublic) | Predictor | Categorical (Men, Women, No difference) | Perception of which gender is better at representing the public. |

# Exploratory Data Analysis



Correlation Heatmap

The heatmap of correlations among continuous variables provides valuable insights into relationships within the dataset. High correlations were observed between attitudinal predictors, such as discrimination and life satisfaction, and the outcome variable public office. These relationships indicate that individuals who report higher life satisfaction and less discriminatory attitudes tend to exhibit stronger support for women in public office. Notably, the correlation between demographic factors, such as age, and attitudinal predictors was relatively weak, reinforcing earlier findings that demographic factors are less influential in shaping attitudes toward gender equity in leadership. The visual representation of these correlations offers a comprehensive understanding of the interplay between attitudinal and demographic factors, guiding subsequent modeling efforts.

Interaction Effects on Party (Democrat Proportions) — Interaction Effects on Party (Republican Proportions)

These plots comparing publicoffice scores across attitudinal predictors, such as keeping government honest and national security, revealed distinct patterns in support for women in leadership. Respondents who perceived "No Difference" between genders in these predictors consistently displayed the highest median publicoffice scores, suggesting that gender-neutral evaluations of leadership traits strongly correlate with greater support for women. While respondents favoring women in traits such as honesty or compromise also showed high scores, the emphasis on gender equality appeared to have the strongest association. This reinforces the importance of promoting unbiased perceptions of leadership qualities in fostering equitable attitudes toward female representation.

To explore how interactions between predictors influence party preference, a heatmap of proportions for the Democrat category was created, focusing on keepgo honest and national security. The results demonstrated that respondents who perceived women as better at ensuring honesty in government and national security were more likely to identify as Democrats. Conversely, those with more traditional or male-dominant views in these areas showed a weaker association with Democratic affiliation. This visualization underscores the nuanced ways in which specific combinations of attitudes toward leadership traits can predict political alignment. These findings provide a foundation for understanding the complex relationships between attitudes and political behavior.

# Summary Of Models

**Model 1A:** Multiple Linear Regression for Public Office Outcome

Model 1A Displays the support for women in public office based on demographic factors and

attitudes towards women.

```
Residuals:
      Min         1Q      Median         3Q        Max
-0.0063835 -0.0013140 -0.0001317  0.0016754  0.0049752

Coefficients:
                               Estimate Std. Error   t value Pr(>|t|)
(Intercept)                   -4.917e-04  1.778e-03    -0.277    0.782
raceOthers                     1.750e-03  1.288e-03     1.359    0.176
racewhite                      7.268e-04  7.888e-04     0.921    0.358
communitytypeS                -1.280e-04  4.763e-04    -0.269    0.788
communitytypeU                -5.126e-05  5.141e-04    -0.100    0.921
age                            1.316e-05  1.145e-05     1.149    0.252
censusregionNorth East         2.521e-04  5.244e-04     0.481    0.631
censusregionsouth             -1.409e-04  4.346e-04    -0.324    0.746
censusregionwest              -9.248e-05  5.399e-04    -0.171    0.864
partyRepublica                -1.831e-05  6.731e-04    -0.027    0.978
lifesatisfaction              -6.291e-06  1.087e-05    -0.579    0.563
edugraduate                   -7.186e-04  4.903e-04    -1.466    0.144
eduHS or less                 -1.419e-04  4.862e-04    -0.292    0.771
edutwo year college           -6.969e-04  4.823e-04    -1.445    0.150
discrimination                 5.668e-06  1.007e-05     0.563    0.574
standingupno difference        1.429e+01  6.551e-04 21805.076   <2e-16 ***
standingupwomen                7.142e+00  6.653e-04 10735.539   <2e-16 ***
keepgovhonestno difference     1.429e+01  8.414e-04 16980.102   <2e-16 ***
keepgovhonestwomen             7.144e+00  8.154e-04  8761.728   <2e-16 ***
compromiseno difference        1.429e+01  5.884e-04 24280.880   <2e-16 ***
compromisewomen                7.143e+00  5.226e-04 13666.302   <2e-16 ***
nationalsecurityno difference  1.429e+01  4.707e-04 30352.724   <2e-16 ***
nationalsecuritywomen          7.141e+00  9.472e-04  7539.598   <2e-16 ***
publicsafetyno difference      1.429e+01  5.198e-04 27483.890   <2e-16 ***
publicsafetywomen              7.144e+00  6.367e-04 11220.072   <2e-16 ***
dealsocialissueno difference   1.429e+01  1.020e-03 13999.830   <2e-16 ***
dealsocialissuewomen           7.144e+00  9.724e-04  7347.338   <2e-16 ***
representpublicno difference   1.428e+01  6.404e-04 22305.665   <2e-16 ***
representpublicwomen           7.140e+00  6.463e-04 11048.108   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002493 on 201 degrees of freedom
Multiple R-squared:      1,  Adjusted R-squared:      1
F-statistic: 1.003e+09 on 28 and 201 DF,  p-value: < 2.2e-16
```
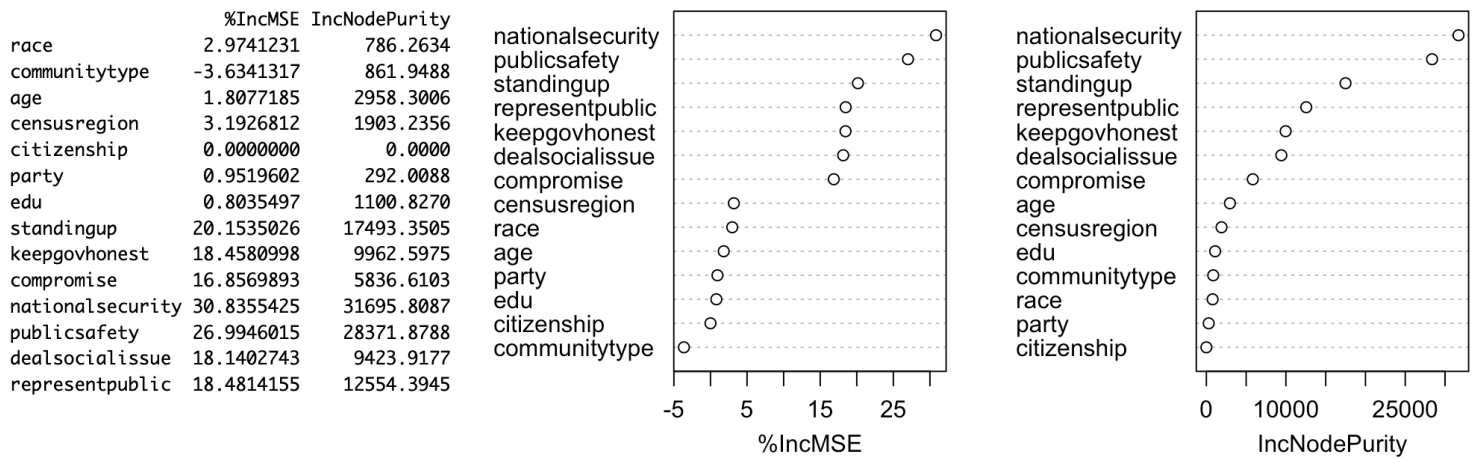
The first step into analyzing the support for women in public office was building a multiple linear regression model that assessed both the potential demographic and attitudinal variables highlighted in our path diagram. The dependent variable (public office score) was continuous, ranging from 0 to 100, where higher scores indicated a positive attitude towards recommending women for public office; leading us down the use of a linear model due to its numeric nature.

The regression results showed that attitudinal measures were far more significant predictors than demographic factors. For example, variables like standing up, keeping government honest, compromise, and national security had large positive coefficients and p-values below 0.001. This shows that people who believed women were better or as equally capable in these areas as men were much more likely to support women in public office. On the other hand, demographic factors such as race, age, and education were not statistically significant, suggesting they had minimal influence on the outcome.

The R-squared value of 1 suggested the model fit the data well, though there might be some potential overfitting; leading us to refining our model down the line (Model 1C). This initial linear regression helped us identify the strongest predictors, but in order to get further reassurance, we further pursued a non-linear approach using random forests.

**Model 1B:** Random Forest for Public Office Outcome

| | %IncMSE | IncNodePurity |
|---|---|---|
| race | 2.9741231 | 786.2634 |
| communitytype | -3.6341317 | 861.9488 |
| age | 1.8077185 | 2958.3006 |
| censusregion | 3.1926812 | 1903.2356 |
| citizenship | 0.0000000 | 0.0000 |
| party | 0.9519602 | 292.0088 |
| edu | 0.8035497 | 1100.8270 |
| standingup | 20.1535026 | 17493.3505 |
| keepgovhonest | 18.4580998 | 9962.5975 |
| compromise | 16.8569893 | 5836.6103 |
| nationalsecurity | 30.8355425 | 31695.8087 |
| publicsafety | 26.9946015 | 28371.8788 |
| dealsocialissue | 18.1402743 | 9423.9177 |
| representpublic | 18.4814155 | 12554.3945 |

To complement the linear regression and look for potential non-linear relationships, we built a random forest model. This model ranked predictors based on their significance in public office scores without assuming linearity or independence among variables.

The variable importance plot from the random forest, similar to Model 1A, highlighted attitudinal measures as the most influential predictors. Variables like national security, public safety, and standing up had the highest %IncMSE (percentage increase in mean squared error), showing their significance in the model's performance. However, demographic variables such as citizenship and community type came as unimportant, reinforcing the idea that support for women in public office is driven more by attitudes than background characteristics.

## Model 1C: Refined Linear Regression for Public Office Outcome

```
Residuals:
      Min        1Q     Median        3Q        Max
-0.0063875 -0.0011645 -0.0001494  0.0017111  0.0053480

Coefficients:
                                Estimate Std. Error    t value Pr(>|t|)
(Intercept)                   -0.0001689  0.0009046     -0.187    0.852
nationalsecurityno difference 14.2862333  0.0004543 31446.332   <2e-16 ***
nationalsecuritywomen          7.1414332  0.0009084  7861.543   <2e-16 ***
publicsafetyno difference     14.2856255  0.0004967 28761.960   <2e-16 ***
publicsafetywomen              7.1439359  0.0006136 11642.939   <2e-16 ***
standingupno difference       14.2855667  0.0006125 23323.888   <2e-16 ***
standingupwomen                7.1422704  0.0006240 11445.210   <2e-16 ***
keepgovhonestno difference    14.2864266  0.0007965 17936.698   <2e-16 ***
keepgovhonestwomen             7.1440137  0.0007658  9328.371   <2e-16 ***
dealsocialissueno difference  14.2865559  0.0009692 14739.895   <2e-16 ***
dealsocialissuewomen           7.1444836  0.0009286  7694.037   <2e-16 ***
representpublicno difference  14.2835978  0.0005975 23906.207   <2e-16 ***
representpublicwomen           7.1400784  0.0005960 11980.838   <2e-16 ***
compromiseno difference       14.2863125  0.0005654 25265.684   <2e-16 ***
compromisewomen                7.1427576  0.0005076 14071.710   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002461 on 215 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.059e+09 on 14 and 215 DF,  p-value: < 2.2e-16
```
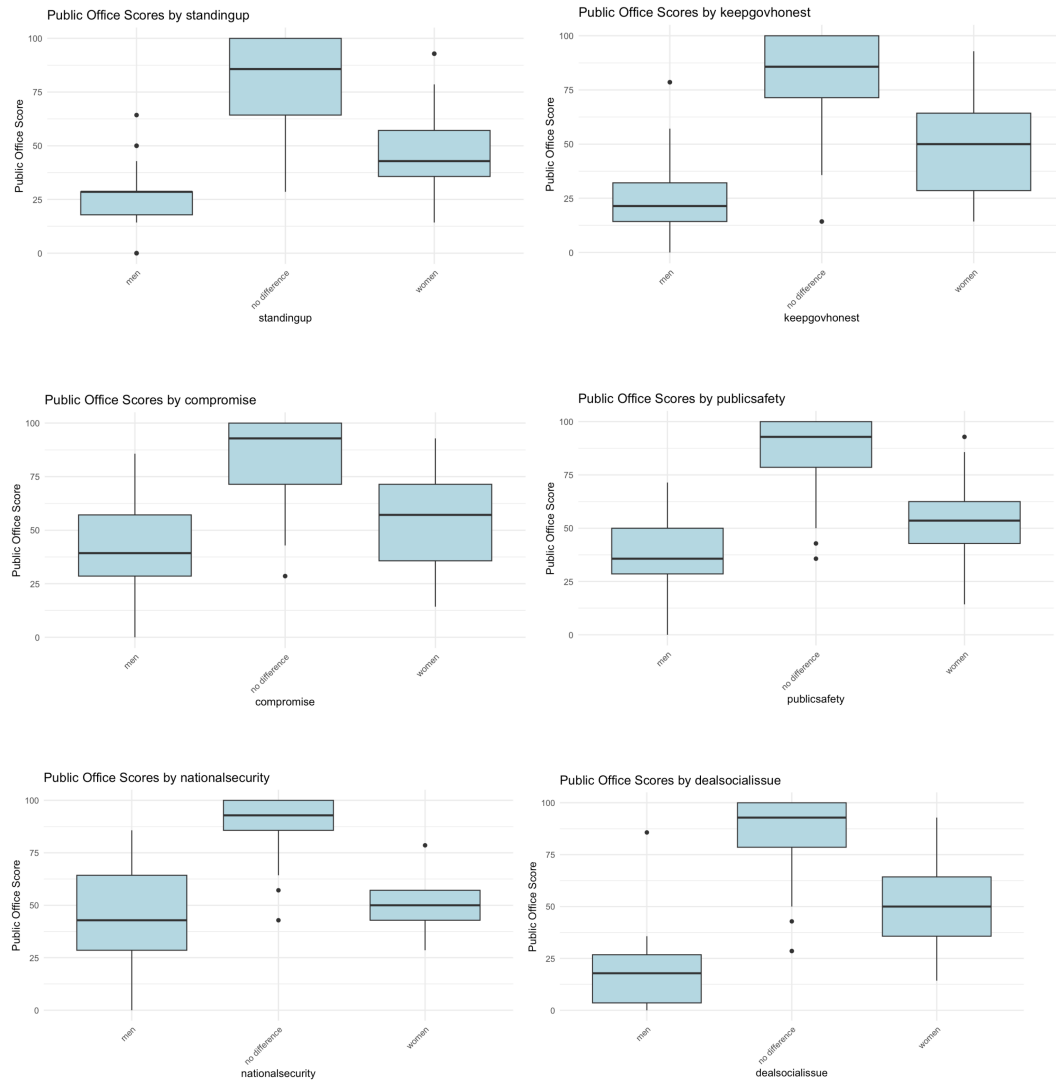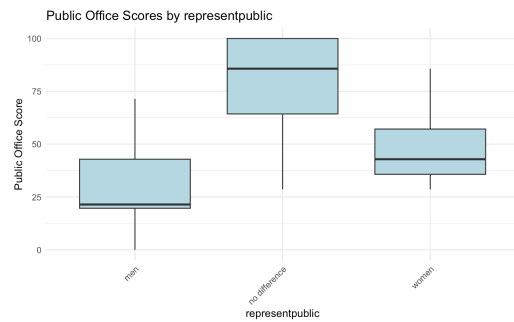
Seeing the results of both the initial linear regression and the random forest model, we refined

the clean data's linear regression by excluding insignificant predictors. Variables such as race,

education, and citizenship, which had high p-values in the original regression, were taken out to

improve interpretability and simplify the model in attempts to get rid of any potential overfit.

This approach focused on the most significant attitudinal measures.


The refined regression model reaffirmed that attitudes toward women's abilities in areas like

standing up, keeping government honest, and compromise were the strongest predictors of

support for women in public office. Additionally, the model maintained a high R-squared value,

indicating that the exclusion of insignificant variables did not reduce the model's explanatory strength.

**Model 1D:** Boxplots – Variance Levels Across Predictors

Public Office Scores by representpublic

Lastly, in order to visualize the distribution of public office scores for each level of the significant categorical predictors we defined, we built boxplots to define variability. Interestingly, participants who perceived "no difference" in attitudinal traits between men and women consistently demonstrated the highest public office scores in each predictor (median was highest). This suggests a strong correlation between neutral perceptions of gender and higher support for women in public office. This finding seems to show that individuals who believe leadership skills are not gender-specific are likely to follow a more merit-based or unbiased perspective, translating into increased support for women in public office roles. While direct preference for "women" leads to a higher median public office score than men in all predictors, the forefront held by the "no difference" category suggests that promoting gender-neutral evaluations of leadership qualities may be an effective strategy for increasing public support for women in leadership positions.

**Model 2A:** Full Logistic Regression Model for Political Party Preference

Our second model aims to predict the political party of any given individual based on a multitude of predictors.

We first started off by developing a logistic regression model that includes every single predictor, including age, race, education, life satisfaction, census region, etc. Based on the summary of the model, nearly all of the predictor variables are insignificant. However, we see that for age, education (high school or less), who is more likely to stand up for women, and preference for which gender to represent the public are partially significant when it comes to predicting the political party as their respective p-values are between 0.05 and 0.1.

```
Call:
glm(formula = party ~ race + communitytype + age + censusregion +
    lifesatisfaction + edu + discrimination + standingup + keepgovhonest +
    compromise + nationalsecurity + publicsafety + dealsocialissue +
    representpublic, family = "binomial", data = data_clean)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.875e+00  2.864e+00   0.655   0.5127
raceOthers                    1.697e+01  3.253e+03   0.005   0.9958
racewhite                     4.085e-01  1.289e+00   0.317   0.7512
communitytypeS               -1.499e+00  1.001e+00  -1.498   0.1340
communitytypeU               -8.565e-01  1.099e+00  -0.780   0.4356
age                           4.269e-02  2.366e-02   1.804   0.0712 .
censusregionNorth East       -3.301e-01  1.021e+00  -0.323   0.7465
censusregionsouth            -3.703e-02  7.382e-01  -0.050   0.9600
censusregionwest              1.004e+00  1.272e+00   0.789   0.4299
lifesatisfaction              7.477e-03  1.959e-02   0.382   0.7027
edugraduate                  -1.565e-01  1.096e+00  -0.143   0.8865
eduHS or less                -1.455e+00  8.603e-01  -1.691   0.0909 .
edutwo year college          -2.447e-01  1.004e+00  -0.244   0.8075
discrimination               -1.833e-02  1.849e-02  -0.991   0.3216
standingupno difference      -1.505e+00  1.402e+00  -1.073   0.2832
standingupwomen              -2.143e+00  1.205e+00  -1.779   0.0752 .
keepgovhonestno difference    1.882e+00  1.249e+00   1.507   0.1319
keepgovhonestwomen            6.614e-01  1.186e+00   0.558   0.5771
compromiseno difference       3.700e-01  1.048e+00   0.353   0.7241
compromisewomen              -2.835e-01  8.926e-01  -0.318   0.7508
nationalsecurityno difference -8.534e-01  8.693e-01  -0.982   0.3262
nationalsecuritywomen         1.809e+01  2.751e+03   0.007   0.9948
publicsafetyno difference    -1.022e+00  1.011e+00  -1.010   0.3124
publicsafetywomen             1.688e+01  1.843e+03   0.009   0.9927
dealsocialissueno difference  2.044e-01  1.663e+00   0.123   0.9022
dealsocialissuewomen          2.698e-01  1.560e+00   0.173   0.8626
representpublicno difference  1.728e+00  1.049e+00   1.647   0.0995 .
representpublicwomen          2.223e+00  1.248e+00   1.781   0.0750 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 121.277  on 229  degrees of freedom
Residual deviance:  86.376  on 202  degrees of freedom
AIC: 142.38
```
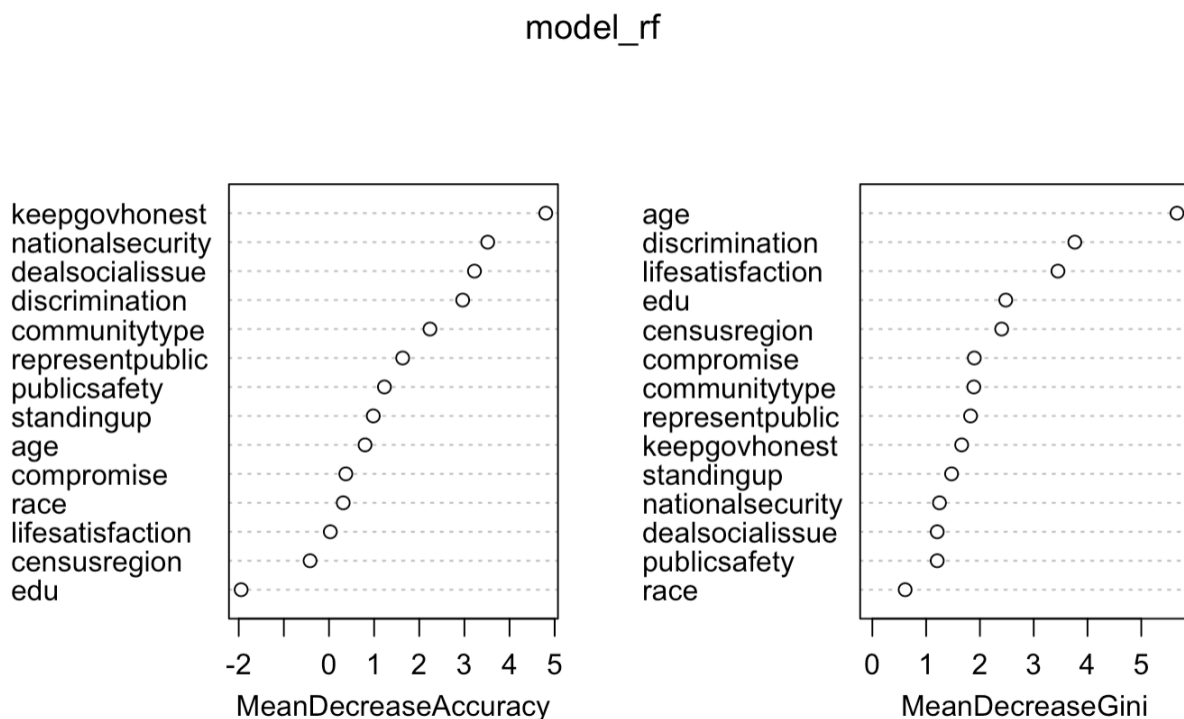
**Model 2B:** Random Forest for Political Party Outcome

Following the original model with all terms, a random forest and Variable Importance plot was constructed in order to select the most important variables in predicting the political party for individuals. Based on the results of the mean decrease in accuracy and the mean decrease in gini, we see that age, lifesatisfaction, discrimination, keepgovhonest, nationalsecurity, and dealsocialissue were the most significant predictors in contributing to the model performance.

model_rf



**Model 2C:** Reduced Logistic Regression Model for Political Party Preference

After running the random forest and variance importance plot, we used feature selection to create a new reduced logistic regression model with only the significant predictors seen above. Based on the summary of the reduced model, we see that age and the preference of people who want to

keep the government honest are the only partially significant predictors with p-values between

0.05 and 0.1. In addition, we see that the null deviance remained the same as the original model,

showing that the baseline level of variability is the same for the full and reduced models.

However, we see that the residual deviance is lower for the full model compared to the reduced

model, showing that the full model fits the data better than the reduced model. Lastly, we see that

the AIC for the full model is higher compared to the reduced model. This shows that the reduced

model provides a better tradeoff between simplicity and model fit, even though the full model

fits the data better (as seen by lower residual deviance). In summary, the full model provides a

better fit for the data, but it is more complex and includes more variables, versus the reduced

model offers a more simple representation that should be used if the improvement from the full

model is not statistically significant. In order to do this, we conducted a Likelihood Ratio Test

between the full and reduced models. We see that the p-value for the Likelihood Ratio Test is

greater than 0.05, showing that the improvement in fit is not statistically significant and we

should use the reduced model for simplicity and interpretability.

```
Call:
glm(formula = party ~ age + lifesatisfaction + discrimination +
    keepgovhonest + nationalsecurity + dealsocialissue, family = "binomial",
    data = data_clean)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  3.652e-01  1.678e+00   0.218   0.8278
age                          3.388e-02  1.821e-02   1.861   0.0628 .
lifesatisfaction             4.894e-03  1.538e-02   0.318   0.7503
discrimination              -4.454e-03  1.592e-02  -0.280   0.7796
keepgovhonestno difference   1.890e+00  9.644e-01   1.960   0.0500 .
keepgovhonestwomen           8.438e-01  8.874e-01   0.951   0.3417
nationalsecurityno difference -6.730e-01 6.851e-01  -0.982   0.3259
nationalsecuritywomen        1.480e+01  1.288e+03   0.011   0.9908
dealsocialissueno difference -5.211e-01  1.361e+00  -0.383   0.7019
dealsocialissuewomen        -2.628e-01  1.270e+00  -0.207   0.8360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 121.28  on 229  degrees of freedom
Residual deviance: 109.64  on 220  degrees of freedom
AIC: 129.64

Number of Fisher Scoring iterations: 16
```
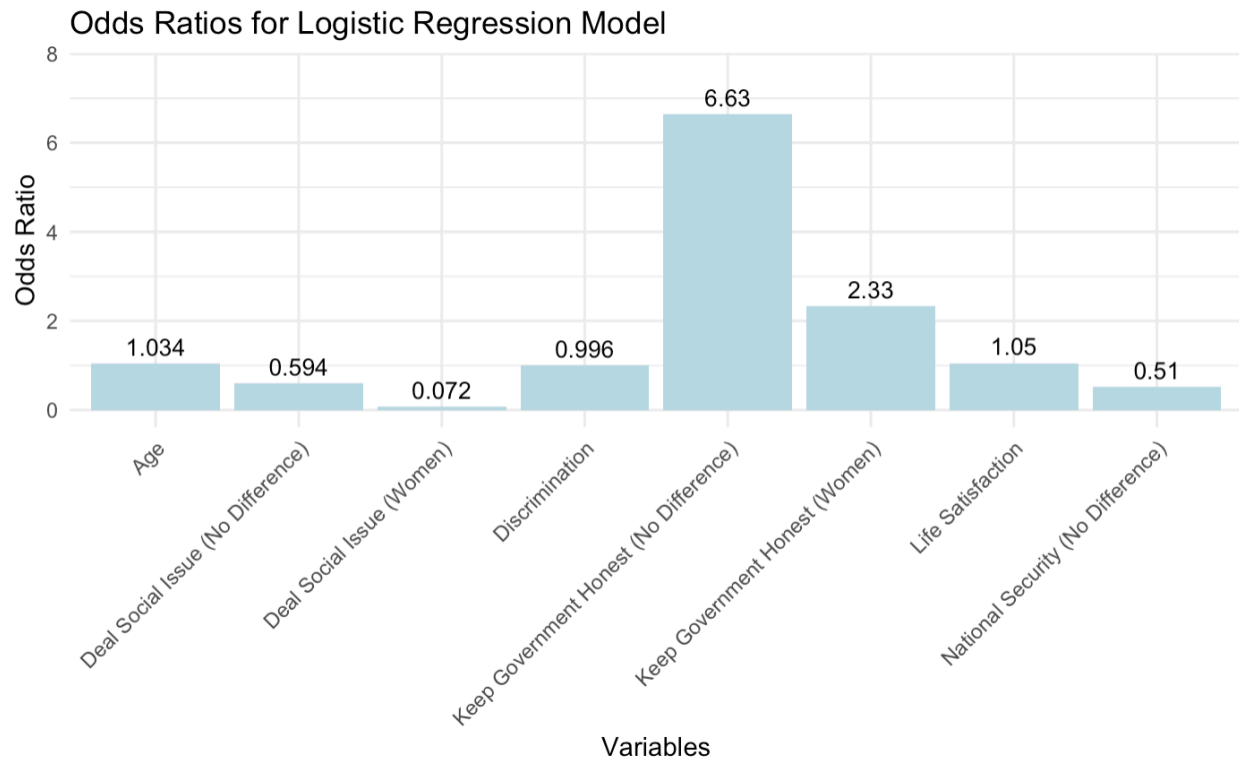
17

```
Analysis of Deviance Table

Model 1: party ~ age + lifesatisfaction + discrimination + keepgovhonest +
    nationalsecurity + dealsocialissue
Model 2: party ~ race + communitytype + age + censusregion + lifesatisfaction +
    edu + discrimination + standingup + keepgovhonest + compromise +
    nationalsecurity + publicsafety + dealsocialissue + representpublic
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       220    109.641
2       202     86.376 18   23.265   0.1806
```

**Model 2D:** Odds Ratio of Reduced Logistic Regression Model for Political Party

Preference



The reduced logistic regression model highlights the significant predictors of political party preference. The most influential factor is the belief that there is "no difference" in genders' ability to keep the government honest, with an odds ratio of 6.63. This means individuals with this perception are over six times more likely to align with the reference party, demonstrating the strong role of gender-neutral attitudes in shaping political alignment. Similarly, those who perceive women as better at keeping the government honest (odds ratio of 2.33) are also more likely to identify with the reference party, albeit to a lesser extent.

Age and life satisfaction show minimal influence on party preference, with odds ratios close to 1, indicating these factors have little impact. Meanwhile, perceptions like "deal social

issue (women)" and "national security (no difference)" show negative associations with party alignment, suggesting individuals with these views are less likely to affiliate with the reference party. For example, believing there is no difference in gender for national security reduces the odds of affiliation by almost half.

Overall, the findings emphasize the dominant role of gender-neutral evaluations, particularly regarding honesty in governance, over demographic factors in predicting political party preference. This suggests that fostering unbiased perceptions of leadership traits may play a key role in shaping political behaviors.

Note: All models were ran with interaction terms, but AIC and deviances were extremely high compared to the models without interaction terms.
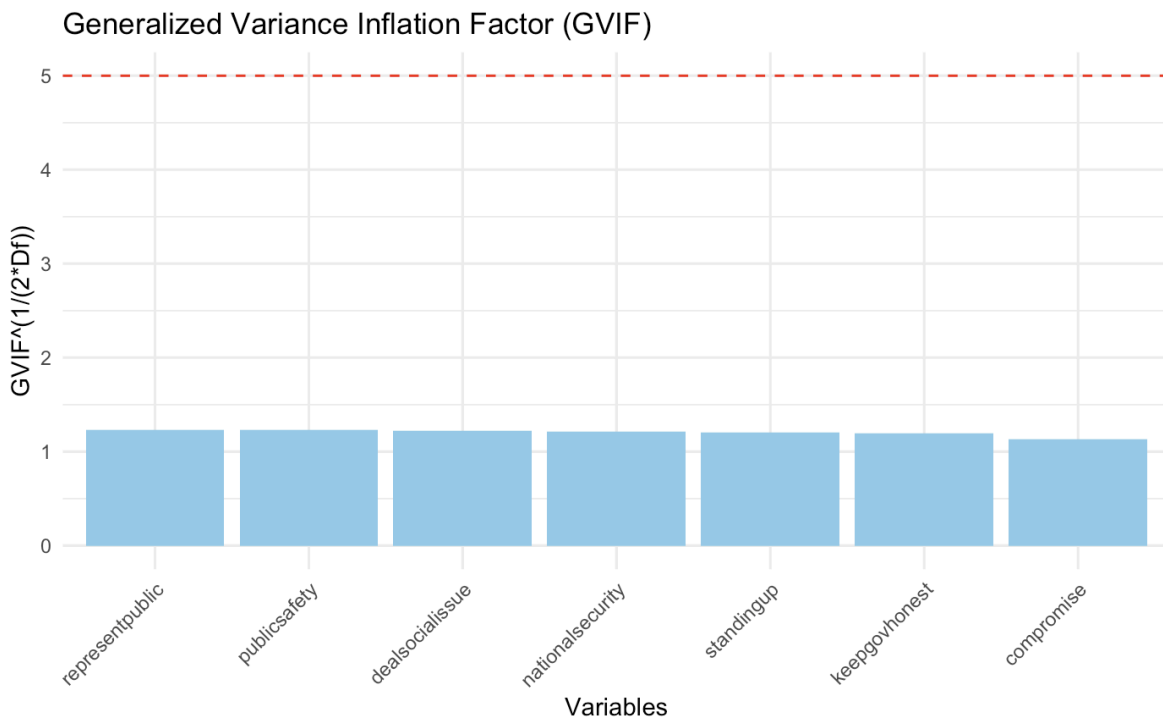
# Interaction Effects

Because both of our models, the multiple linear regression and logistic regression model, did not include interaction terms, there are no interaction effects to be plotted and analyzed. This prevented overfitting in our model.
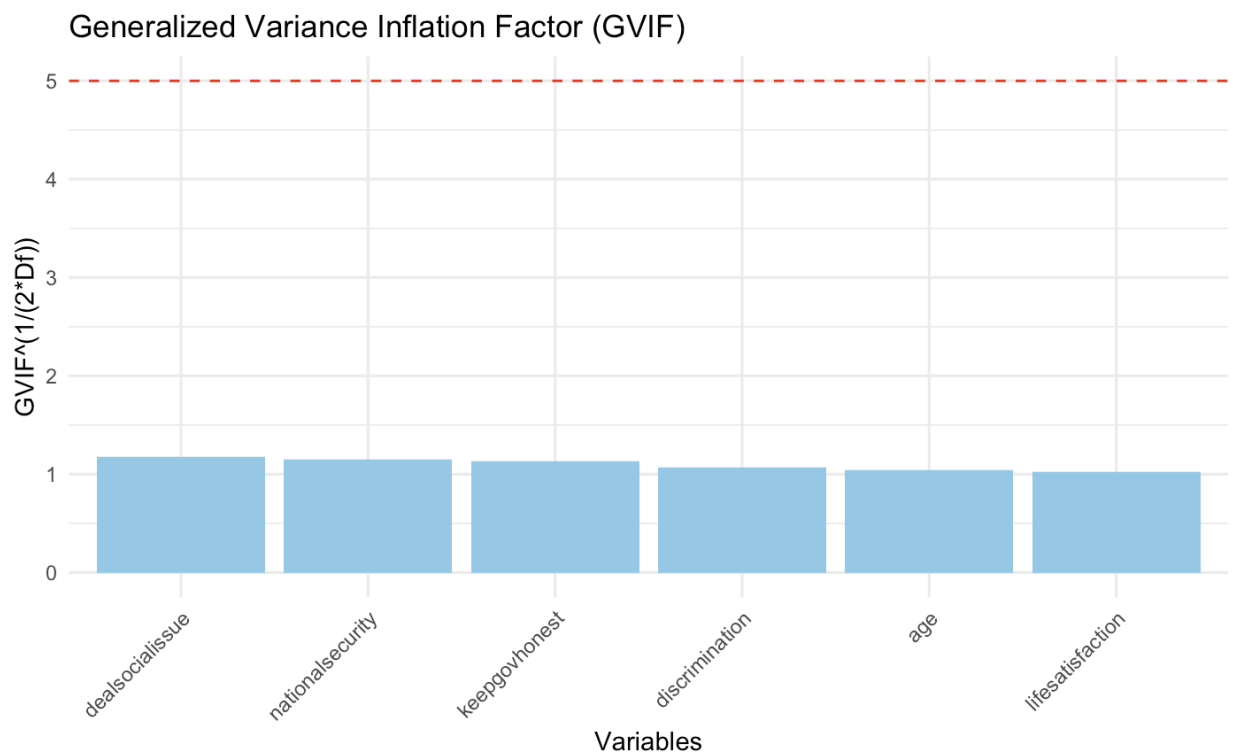
# Checking for Multicollinearity

**Multiple Linear Regression Model:**

| Predictor | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| **nationalsecurity** | 2.168497 | 1 | 1.472582 |
| **publicsafety** | 2.264183 | 1 | 1.504720 |
| **standingup** | 2.109391 | 1 | 1.452374 |
| **keepgovhonest** | 2.066816 | 1 | 1.437643 |
| **dealsocialissue** | 2.198323 | 1 | 1.482674 |
| **representpublic** | 2.268210 | 1 | 1.506058 |
| **compromise** | 1.645284 | 1 | 1.282686 |

## Generalized Variance Inflation Factor (GVIF)

**Logistic Regression Model:**

|  | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| **age** | 1.076365 | 1 | 1.037480 |
| **lifesatisfaction** | 1.053373 | 1 | 1.026339 |
| **discrimination** | 1.141699 | 1 | 1.068503 |
| **keepgovhonest** | 1.656004 | 2 | 1.134398 |
| **nationalsecurity** | 1.725559 | 2 | 1.146126 |
| **dealsocialissue** | 1.905344 | 2 | 1.174880 |



Looking at both models' variance inflation factors, we see that they are all below 5, demonstrating that both of our models do not have any issues with multicollinearity.

# Relevant Model Assumptions

## Multiple Linear Regression Model (1A & 1C):

**QQ Plot for Public Office Model**

Linearity:

- There is a **linear relationship** between the independent variables and the dependent variable (public office score). This assumption was confirmed in Model 1A and Model 1C

Independence:

- **The observations are independent of each other.** In the context of the study, this means that one individual's response does not influence another's.

Homoscedasticity:

- The **variance of errors is constant across all levels of the independent variables**. If the model were to be visualized with residuals plotted against fitted values, they should appear randomly scattered.

Normality

- The **residuals of the regression model should be approximately normally distributed**. This is confirmed in our QQ plot.

## Random Forest Assumptions (Model 1B and 2B):

Independence

- **Observations are still assumed to be independent**, as the random forest method relies on independent data points to construct trees.
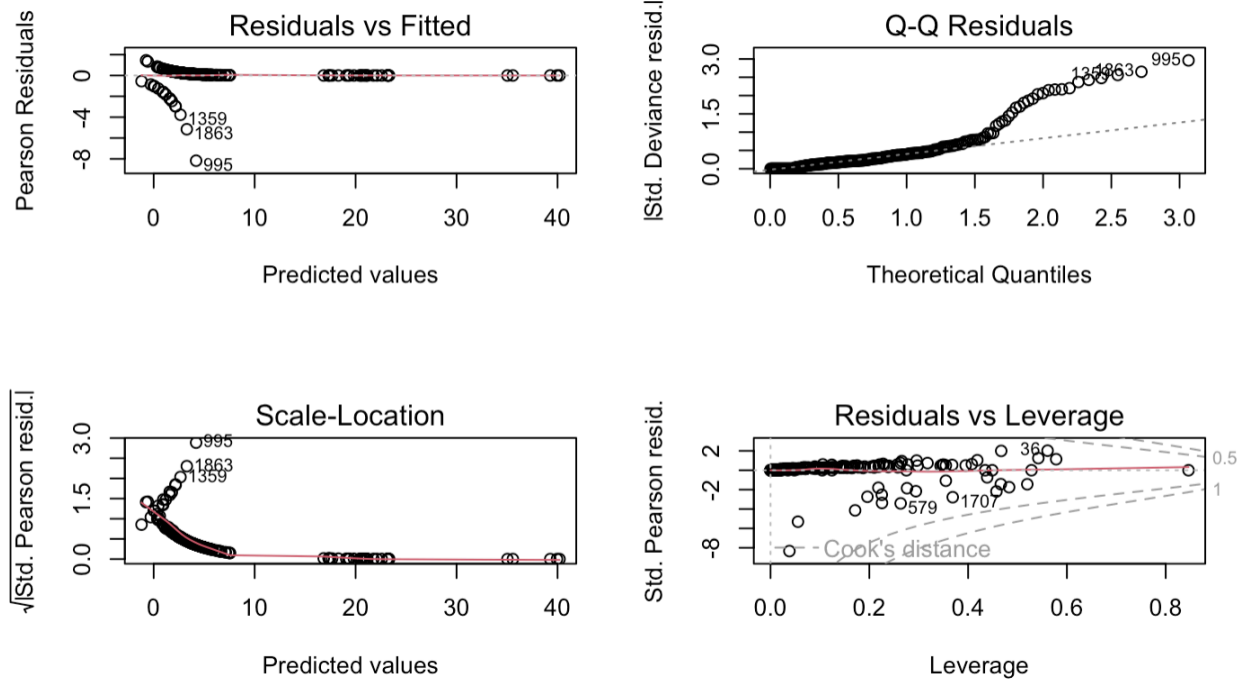
Non-Linearity

- We **do not assume a linear relationship between predictors and the dependent variable.** This allows for the modeling of non-linear relationships between predictors and outcomes.

Sampling:

- We assume our **sample is large and random,** which will provide diverse, unbiased training sets for each tree.

## Logistic Regression Model (2A & 2C):

We see above from the plots of the residuals that the assumptions of the model may not hold as the Q-Q plot is skewed and the residuals are extremely high. The model does not have normality of the error terms and there is no constant variance.

Independence:

- **Each observation is independent of the others**. This assumption ensures that each data point, each individual survey response, does not influence one another.

Binary Outcome

- The dependent variable in logistic regression is binary (political party preference). The logistic regression model assumes that the outcome variable is categorical with two categories (e.g., Democrat vs. Republican).

Linearity of Log-Odds

- There is a **linear relationship between the predictors and the log-odds of the outcome**. This means the model predicts the log-odds of the outcome variable, which can then be transformed back to probabilities.

Sampling:

- We assume the samples come from a **sufficiently large sample, with outliers identified and addressed, with no perfect separation of data**. Ensuring the model's stability and generalizability by minimizing the influence of anomalies and preventing distortion from outliers.

# Conclusion

Our project aimed to explore the factors influencing support for women in public office and political party preferences using linear regression, random forest models, and logistic regression. Key attitudinal variables consistently emerged as the most significant predictors, highlighting their importance over demographic factors in shaping support for women and political party preferences.

Findings from Model 1:

Overall, model 1 confirmed that attitudes toward women were more important than demographics in determining support for women in public office.

- **Significant Predictors:** Attitudinal factors, particularly Discrimination, Standing Up, Honesty, Compromise, National Security, Public Safety, Social Issues, and Representing Publicwere significant predictors of support for women in public office ($p < 0.001$ for these variables).

- **Partially Significant Predictors:** None

- **Insignificant Predictors**: Demographic factors such as Ideology, Education, Age, Life Satisfaction, Race, Census Region, Community Type, and Citizenship had no significant impact on the outcome, with p-values indicating minimal influence.

The R-squared value was 1, indicating a perfect fit to the data, although potential overfitting was noted, suggesting the need for model refinement (as seen in Model 1C).

Findings from Model 2:

Overall, Model 2 suggests that political party preference is primarily influenced by attitudes toward women, rather than demographic characteristics, with an exception for age.

- **Significant Predictors**: None

- **Partially Significant Predictors:** The partially significant predictors in the reduced logistic regression model include age, education (high school or less), who is more likely to stand up, and preference for which gender to represent the public.

- **Most Significant According to Odds ratio & Variable Importance plots:** age, life satisfaction, discrimination, keeping the government honest, national security, and dealing with social issues

The Likelihood Ratio Test between the full and reduced models showed no statistically significant improvement in fit (p-value > 0.05).  Overall, Model 2 shows that attitudes towards women, along with the demographic of age, are most significant in predicting an individual's party affiliation.


Overall:

Based on the findings from both Model 1 and Model 2, it is evident that overall voting behavior is primarily influenced by predictors related to attitudes towards women, rather than demographic characteristics. Specifically, variables such as who is more likely to stand up for women and the preference for government honesty were significant predictors of political party preference, while demographic variables like age, race, and education had a weaker or insignificant impact. These findings suggest that fostering equitable and unbiased perceptions of leadership traits may be a key strategy for increasing support for women in public office.

# Limitations and Recommendations

**Limitations:**

1. Small Sample Size:

   - The original dataset contained 2,250 observations, but after removing missing values (NAs), the cleaned dataset was reduced to only 230 observations. This increases the risk of sampling bias and reduces the statistical power of the models, making it more challenging to detect subtle but meaningful relationships.

2. Missing Data Bias:

   - Removing NA's may have introduced bias if the missing data was not random. This could skew results and reduce the validity of the conclusions.

3. Overfitting Concerns:

   - The high R-squared value in some models, particularly in Model 1A, suggests potential overfitting. While adjustments were made to simplify the models, the small sample size increased the risk of overfitting, thus the results might only work well for this limited data.

4. Limited Demographic Representation:

   - Many demographic predictors, such as race, education, and region, were found to be insignificant, potentially due to limited variability within the reduced dataset. This may prevent a deeper understanding of how demographic factors influence political preferences.

**Recommendations:**

1. Expand and Enhance the Sample:

- Future research should aim to address the issue of missing data by possibly collecting additional data to ensure a larger and more representative sample and complete observations with no missing responses.

2. Handle Missing Data in a Different Strategy:

- Rather than removing NAs outright, we could use advanced methods, such as machine learning approaches, to estimate missing values and retain more observations for analysis.

3. Integrate External Data Sources:

- Future studies could integrate external datasets, such as census data or public opinion polls, to enrich the analysis. This would help contextualize the results and validate findings against larger, more diverse populations.