Analysis of Political Affiliation Determinants

Introduction

Understanding the factors that shape political affiliation is crucial for gaining insights into the diverse dynamics of American society. I chose this topic because of its relevance to current political discourse and its potential to shed light on how demographic and socio-economic factors influence voter behavior. The goal of this project is to explore the determinants of political affiliation across U.S. states and counties. Using data sourced from the United States Census Bureau and the Harvard Dataverse, I compiled a dataset comprising 3,114 observations. This dataset includes key determinants such as population characteristics and socio-economic conditions that may influence political preferences. By analyzing the dataset using Python, this project aims to uncover patterns and relationships that shape the political dynamics across different regions, providing valuable insights into the factors driving political behavior.

The dataset contains a wide range of covariates about U.S. counties, including population size, racial distribution, age demographics, education levels, income, unemployment rates, and more. The primary objectives are to conduct a comprehensive exploration of the dataset to understand its structure, uncover patterns, and identify relationships, and to apply statistical techniques to derive meaningful insights.

Data Exploration and Cleaning

The analysis began with combining data from two datasets. The first dataset, `pop`, is based on the 2019 American Community Survey and includes demographic and socio-economic characteristics across U.S. counties. The second dataset, `votes`, provides political affiliation statistics for 2019. The data preparation process focused on aligning and cleaning these datasets to ensure consistency. Key steps included standardizing column formats and merging the voting data with population data to create a unified dataset for analysis. The majority of variables in the dataset are expressed as percentages to facilitate meaningful comparisons across counties with varying population sizes.

To maintain consistency, the political affiliation data was also transformed into percentages.

Dataset Summary

The combined `pop` dataset contains 3,114 entries across 20 columns, with data types including floats and objects. Most columns are numeric, while categorical data like state, county, and leading political party are stored as objects.

```
RangeIndex: 3114 entries, 0 to 3113
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   State                        3114 non-null   object
 1   County                       3114 non-null   object
 2   Population                   3114 non-null   float64
 3   White                        3114 non-null   float64
 4   African American             3114 non-null   float64
 5   Asian                        3114 non-null   float64
 6   Hispanic                     3114 non-null   float64
 7   Foreign born                 3113 non-null   float64
 8   Less Than HighSchool         3114 non-null   float64
 9   At Least High School         3114 non-null   float64
 10  Bachelors Degree and Higher  3114 non-null   float64
 11  Poverty                      3114 non-null   float64
 12  Median age                   3114 non-null   float64
 13  Homeownership                3113 non-null   float64
 14  Household income             3113 non-null   float64
 15  Crime                        2946 non-null   float64
 16  Unemployment                 3114 non-null   float64
 17  Democrat                     2949 non-null   float64
 18  Republican                   2949 non-null   float64
 19  Leading Political Party      2949 non-null   object
dtypes: float64(17), object(3)
```

Figure 1: Dataset Summary

Categorical column `Leading Political Party` was converted into a binary format by mapping "Democrat" to 1 and "Republican" to 0, creating a new column called `Political Binary`. This transformation simplifies the representation of political affiliation and facilitates statistical and machine learning techniques that require numerical input.

Summary Statistics

The next step involved using the `describe()` method to summarize the central tendencies and overall distribution of the dataset. The `describe()` method computes various statistics, including the count, mean, standard deviation, minimum, maximum, and 25th, 50th, and 75th percentiles for each numeric column.

	count	mean	std	min	25%	50%	75%	max
Population	3114.0	97754.3	312385.3	81.5	11268.6	25952.8	66157.1	9801950.0
White	3114.0	79.0	19.4	2.5	68.0	86.5	94.4	99.2
African American	3114.0	8.8	14.4	0.0	0.4	2.0	10.1	86.1
Asian	3114.0	1.1	2.3	0.0	0.2	0.4	1.0	42.7
Hispanic	3114.0	7.9	13.0	0.0	1.4	3.0	7.9	97.2
Foreign born	3113.0	4.3	5.4	0.0	1.2	2.4	5.3	51.1
Less Than HighSchool	3114.0	16.9	7.3	0.7	11.5	15.4	21.6	52.1
At Least High School	3114.0	83.0	7.5	29.9	78.4	84.6	88.5	99.3
Bachelors Degree and Higher	3114.0	19.0	8.7	3.7	13.1	16.8	22.5	71.0
Poverty	3114.0	15.5	6.4	0.0	11.0	14.7	19.0	52.2
Median age	3114.0	39.9	4.9	21.7	37.1	39.9	42.8	62.5
Homeownership	3113.0	73.4	7.8	0.0	69.6	74.6	78.5	91.3
Household income	3113.0	44133.7	11451.2	19351.0	36921.0	42388.0	48958.0	115574.0
Crime	2946.0	2.6	2.1	0.0	1.1	2.0	3.4	19.9
Unemployment	3114.0	7.7	2.8	0.0	5.8	7.5	9.3	28.3
Democrat	2949.0	38.2	14.6	3.4	27.7	37.0	47.0	93.4
Republican	2949.0	59.8	14.7	6.0	50.5	60.9	70.4	95.9

Figure 2: Summary Statistics Table

The formulas used to calculate summary statistics are as follows:

Sample Mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Where $\bar{x}$ is the sample mean, n is the number of observations, and x_i is the value of the i-th observation.

Sample Standard Deviation:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Where s is the sample standard deviation, n is the number of observations, x_i is the value of the i-th observation and $\bar{x}$ is the sample mean

Minimum and Maximum:

$$min(x) = \text{smallest value in the data}$$

$$max(x) = \text{largest value in the data}$$

Percentiles:

$$P_k = x_{k \cdot \frac{n+1}{100}}$$

Where P_k is the k-th percentile, calculated based on the position $k \cdot \frac{n+1}{100}$ in the ordered dataset.

Overall, the summary statistics highlight significant heterogeneity across counties, with some variables like population and income showing extreme ranges.

Missing Values

Handling missing values is crucial to ensure the integrity and reliability of the analysis. For numeric columns, missing values were replaced with their respective mean values. However, for the critical categorical variable `Leading Political Party`, rows with missing values were removed to maintain the accuracy of the analysis.

```
In [9]: missing_values = pop.isnull().sum()
   ...: print(missing_values)
State                            0
County                           0
Population                       0
White                            0
African American                 0
Asian                            0
Hispanic                         0
Foreign born                     1
Less Than HighSchool             0
At Least High School             0
Bachelors Degree and Higher      0
Poverty                          0
Median age                       0
Homeownership                    1
Household income                 1
Crime                          168
Unemployment                     0
Democrat                       165
Republican                     165
Leading Political Party        165
dtype: int64
```

Figure 3: Missing Values Summary

```
In [11]: missing_values = pop.isnull().sum()
    ...: print(missing_values)
State                            0
County                           0
Population                       0
White                            0
African American                 0
Asian                            0
Hispanic                         0
Foreign born                     0
Less Than HighSchool             0
At Least High School             0
Bachelors Degree and Higher      0
Poverty                          0
Median age                       0
Homeownership                    0
Household income                 0
Crime                            0
Unemployment                     0
Democrat                         0
Republican                       0
Leading Political Party          0
dtype: int64
```

Figure 4: Missing Values Summary After Removing Missing Values

Data Visualization

Data visualization is a powerful tool for exploring the data further. Using the `seaborn` package, separate pairplots were created for different subsets of variables, such as population characteristics, education levels, and economic indicators, in relation to the `Leading Political Party`. The variables were also visualized using box plots and bar plots.

Variables such as `Bachelors Degree and Higher`, `Homeownership`, `Crime`, `Median age`, and variables related to racial composition show differences in the distribution patterns depending on political affiliation. In contrast, variables like `Less Than Highschool` and `Household income` show similar distributions across political affiliations, indicating weaker distinctions. Many variables exhibit skewed distributions, which could impact the analysis. To address this, transformations such as logarithmic scaling or normalization can be applied to reduce skewness and improve interpretability.

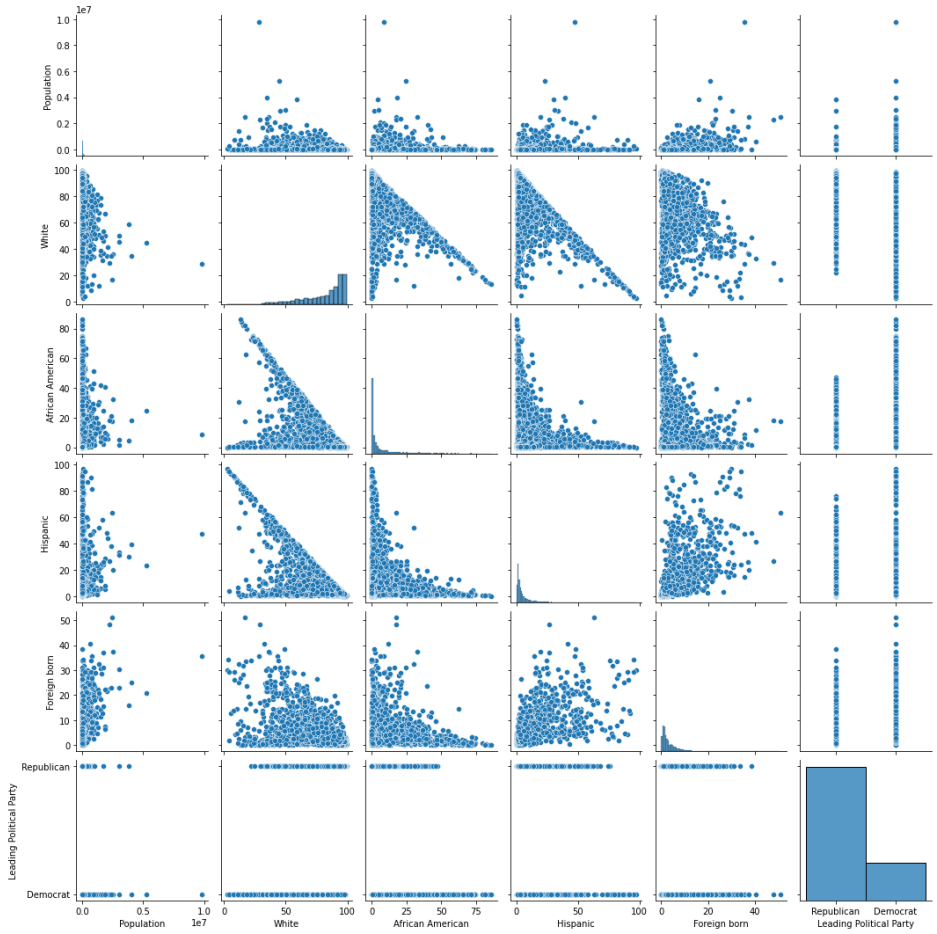

Figure 5: Pairplot for Population, Leading Political Party and variables that represent different racial groups

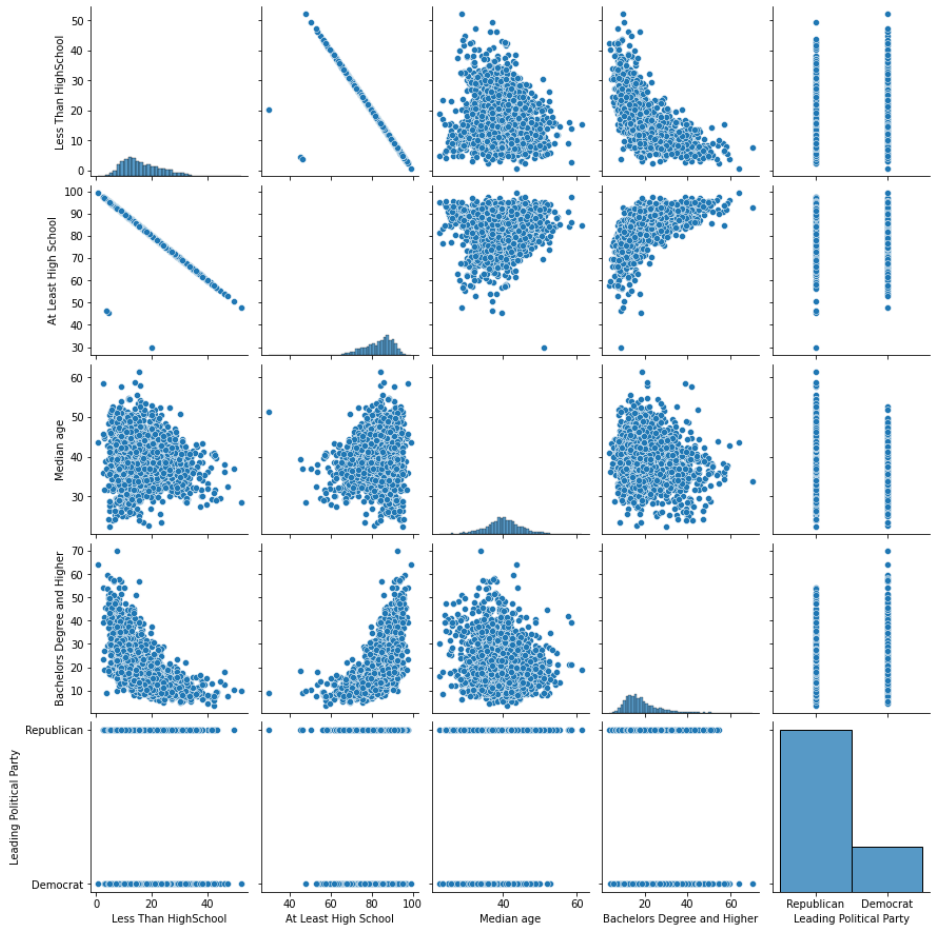

Figure 6: Pairplot for Leading Political Party and variables that represent different education levels

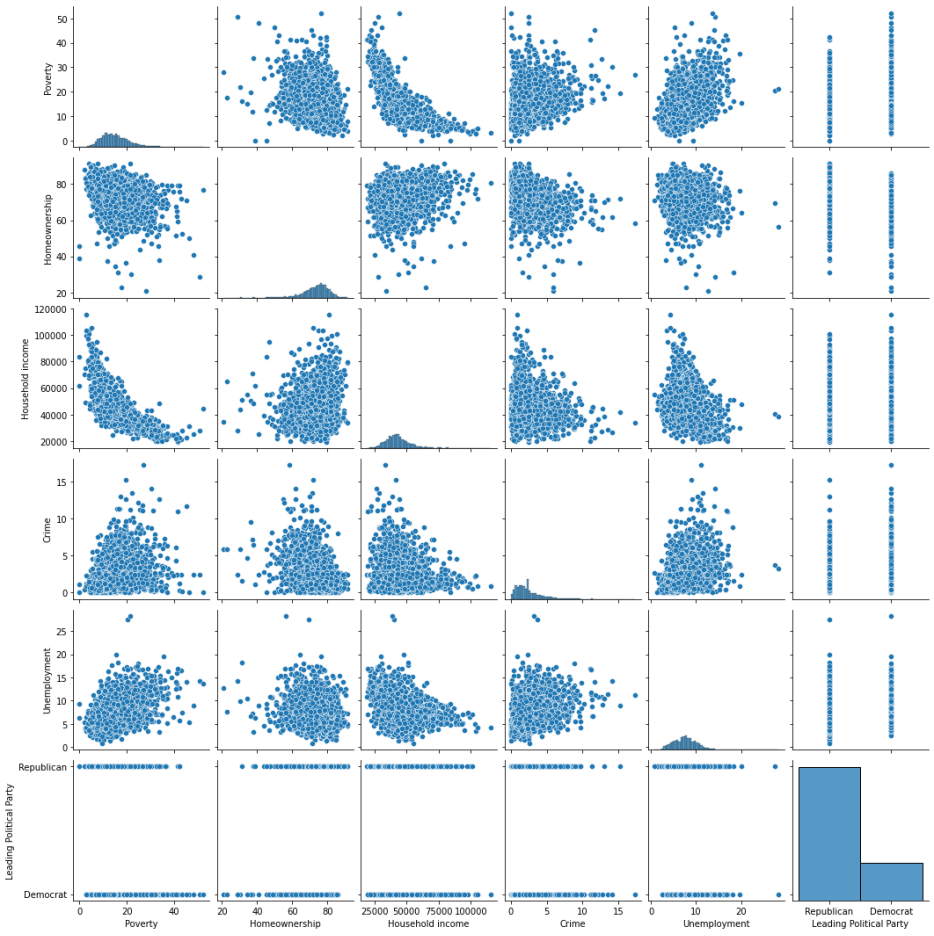

Figure 7: Pairplot for Leading Political Party and socio-economic variables

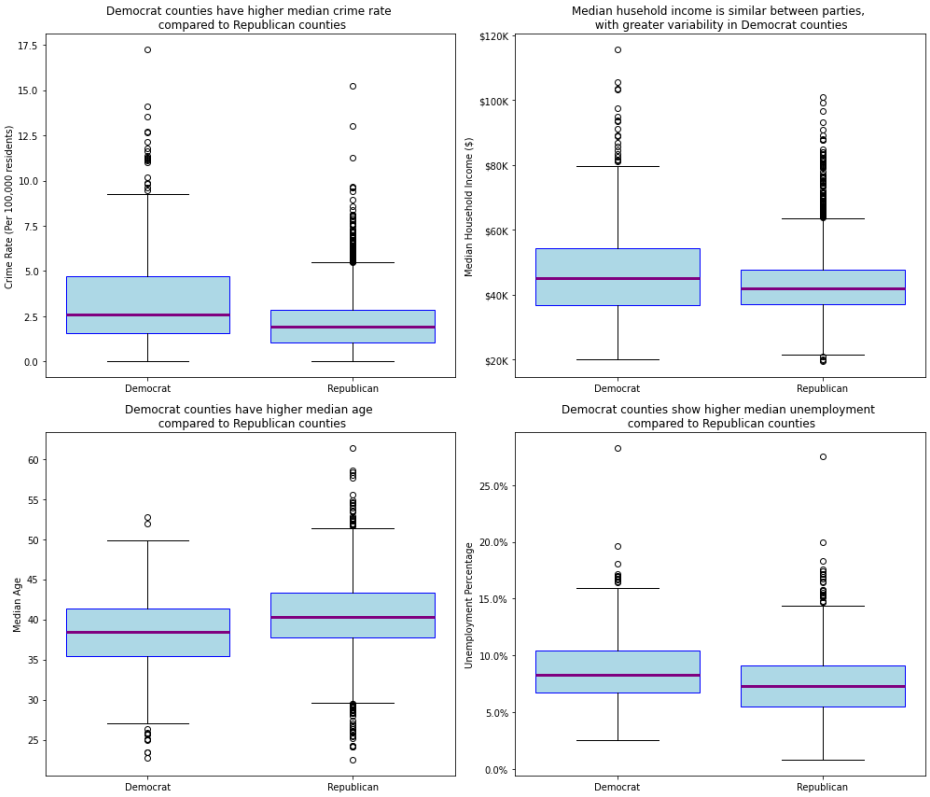

Figure 8: Box plots

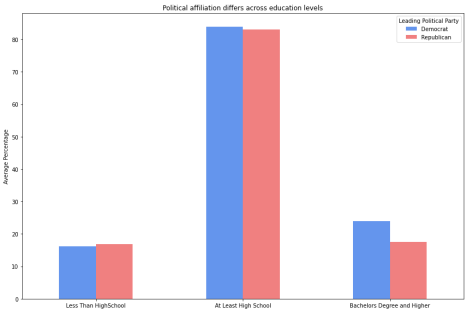

Figure 9: Political affiliation differs across education levels

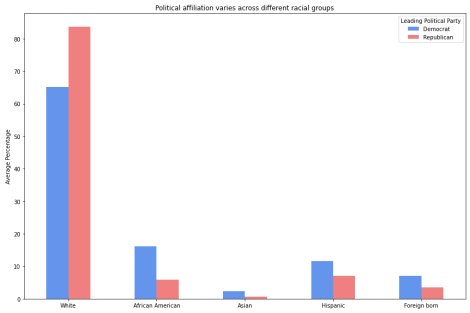

Figure 10: Political affiliation varies across different racial groups

Heatmap Analysis

The heatmap visualizes correlations between `Political Binary` and other numeric variables, calculated using Pearson's correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- r_{xy}: Pearson's correlation coefficient between variables x and y.
- x_i, y_i: Individual data points of x and y.
- $\bar{x}, \bar{y}$: Mean values of x and y, respectively.
- n: Number of data points.

The heatmap revealed that all correlations fall within a moderate range or less. Although the correlation coefficients are relatively low, this is typical for studies involving social behaviors or preferences, where multiple factors interact in complex ways. Even moderate correlations can provide meaningful insights into trends, as no single variable fully explains political affiliation. Variables such as `African American`, `Asian`, and `Bachelors Degree and Higher` showing stronger positive correlations with Democratic affiliation, while `White` and `Homeownership` exhibited stronger negative correlations.

Data Analysis

Hypothesis Testing

From exploration stage, we can notice that the crime rate has a slight positive correlation with Democratic affiliation and Democrat counties have higher median crime rate compared to Republican counties in general. However, it is worth considering whether crime rate is a meaningful indicator of political beliefs. Crime may not directly influence political leanings but could instead be indirectly associated with other factors, such as population size, which is also positively correlated with Democratic affiliation, with average population being significantly larger in Democratic counties. Larger populations often face more complex social dynamics, which can contribute to higher reported crime rates.

Figure 11: Correlation Heatmap

To better understand this relationship, a hypothesis test was conducted to determine whether an increase in population size is associated with an increase in crime. Since we saw from summary statistics that population is showing extreme ranges, logarithmic transformation was be applied to help normalize the data.

$H_0 : \beta = 0$ (Population size has no effect on crime rate)

$H_1 : \beta \neq 0$ (Population size has a significant effect on crime rate)

Using `statsmodels.formula.api`, linear regression was ran to test the hypothesis. Linear regression models the relationship between a dependent variable and one or more independent variables, estimating the effect of predictors on the outcome.

General equation for linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Where

- y: dependent variable (outcome being predicted).
- β_0: intercept, the value of y when all predictors are zero.
- $\beta_1, \beta_2, \ldots, \beta_k$: coefficients representing the effect of each independent variable $x_1, x_2, \ldots, x_k$ on y.
- $x_1, x_2, \ldots, x_k$: independent variables (predictors).
- ϵ: error term, accounting for variability not explained by the predictors.

The regression equation used to test this hypothesis was:

$$\text{Crime} = \beta_0 + \beta_1 \cdot \log(\text{Population}) + \epsilon$$

Based on the results of the linear regression, we reject the null hypothesis. The p-value for `log(Population)` is 0.000, which is below the standard significance level of 0.05. This indicates that population has a statistically significant effect on the crime rate. The coefficient for `log(Population)` is 0.477. For every 1% increase in population size, the crime rate increases by approximately 0.477 on average, assuming a log-linear relationship.

Having established that population size affects the crime rate, another question arises: is the crime rate inherently higher in Democratic counties, or is it correlated with Democratic affiliation primarily because these counties tend to have larger populations?

```
In [18]: formula = ("Crime ~ np.log(Population)")
    ...: # Fit the linear regression model
    ...: model = smf.ols(formula=formula, data=pop).fit()
    ...: print(model.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Crime   R-squared:                       0.133
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     454.0
Date:                Thu, 28 Nov 2024   Prob (F-statistic):           8.08e-94
Time:                        15:05:29   Log-Likelihood:                -5871.8
No. Observations:                2949   AIC:                         1.175e+04
Df Residuals:                    2947   BIC:                         1.176e+04
Df Model:                           1
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                  -2.4128      0.232    -10.401      0.000      -2.868      -1.958
np.log(Population)          0.4770      0.022     21.308      0.000       0.433       0.521
==============================================================================
Omnibus:                     1268.207   Durbin-Watson:                   1.477
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7869.718
Skew:                           1.943   Prob(JB):                         0.00
Kurtosis:                       9.997   Cond. No.                         74.3
==============================================================================
```

Figure 12: OLS Regression Results

More than half of all residents live in just 143 big counties (in terms of the number of residents), according to U.S. Census Bureau county estimates analysis, with the median population for these counties being approximately 821,725. To focus on these highly populated regions, counties with populations of 800,000 or more were selected for the analysis. Next, we test the following hypotheses:

H_0 : The mean crime rate in large Democratic counties is equal to the mean crime rate in large Republican counties.

H_1 :The mean crime rate in large Democratic counties is greater than the mean crime rate in large Republican counties.

$$H_0 : \mu_D = \mu_R$$

$$H_1 : \mu_D > \mu_R$$

Using `ttest_ind` from `scipy.stats`, a two-sample t-test was performed to compare the mean crime rates between large Democratic and Republican counties, assuming unequal variances. A one-tailed test was used to assess whether the mean crime rate in Democratic counties was greater. The for-

mula for the t-statistic is:

$$t = \frac{\bar{X}_D - \bar{X}_R}{\sqrt{\frac{s_D^2}{n_D} + \frac{s_R^2}{n_R}}}$$

Where:

- $\bar{X}_D$: average crime rate in Democratic counties.
- $\bar{X}_R$: average crime rate in Republican counties.
- s_D^2, s_R^2: variances of crime rates for Democratic and Republican counties.
- n_D, n_R: number of Democratic and Republican counties with large populations.

```
In [19]: population_threshold = 800000
    ...: large_population_data = pop[pop['Population'] >= population_threshold]
    ...: group_democrat = large_population_data[large_population_data['Political Binary'] == 1]['Crime']
    ...: group_republican = large_population_data[large_population_data['Political Binary'] == 0]['Crime']
    ...:
    ...: t_stat, p_value_two_tailed = ttest_ind(group_democrat, group_republican, equal_var=False)
    ...: p_value_one_tailed = p_value_two_tailed / 2 if t_stat > 0 else 1
    ...:
    ...: print(f"One-Tailed P-Value: {p_value_one_tailed:.4f}")
One-Tailed P-Value: 0.2186
```

Figure 13: Two-sample t-test

The resulting p-value of 0.2186 exceeds the common significance level of 0.05, indicating insufficient evidence to reject the null hypothesis. Thus, the data does not support the claim that large Democratic counties have higher crime rates than large Republican counties.

Linear Regression

Before proceeding to linear regression, it is necessary to check for multicollinearity among the variables to ensure the reliability of the regression coefficients. Variance Inflation Factor (VIF) was calculated by importing `variance_inflation_factor` from `statsmodels` and used to quantify multicollinearity, where VIF values above 10 indicate severe multicollinearity. Initially, multicollinearity was extremely high, prompting a refinement of the

variables. Some variables that showed high VIF and no correlation with political affiliation in earlier steps of the analysis were removed, and related variables were grouped (racial and foreign-born proportions were combined into a `Diversity` variable). After this adjustment, the VIF values improved, although they were not perfect. The remaining multicollinearity was addressed through transformations during the linear regression process.

```
In [21]: variables = pop[['Population','White','African American','Asian','Hispanic',
    ...: 'Foreign born','Less Than HighSchool','At Least High School',
    ...: 'Bachelors Degree and Higher','Median age','Household income',
    ...: 'Homeownership','Poverty','Crime','Unemployment']]
    ...:
    ...: vif = pd.DataFrame()
    ...: vif["VIF"] = [variance_inflation_factor(variables.values, i)
    ...:               for i in range(variables.shape[1])]
    ...: vif["Features"] = variables.columns
    ...: vif
Out[21]:
          VIF                    Features
0    1.641344                  Population
1  177.352505                       White
2    7.350367            African American
3    2.719825                       Asian
4    9.026395                    Hispanic
5    6.416882                Foreign born
6   44.566420         Less Than HighSchool
7  407.997023         At Least High School
8   20.015902  Bachelors Degree and Higher
9  110.873411                  Median age
10  70.012950             Household income
11 215.303652                Homeownership
12  32.283604                     Poverty
13   4.045073                       Crime
14  13.174703                Unemployment
```

Figure 14: VIF values for all the variables

```
In [25]: variables = pop[['Population','White','Diversity',
    ...: 'Bachelors Degree and Higher','Unemployment']]
    ...: vif = pd.DataFrame()
    ...: vif["VIF"] = [variance_inflation_factor(variables.values, i)
    ...:               for i in range(variables.shape[1])]
    ...: vif["Features"] = variables.columns
    ...: vif
Out[25]:
         VIF                    Features
0   1.339781                  Population
1  12.433380                       White
2   3.551309                   Diversity
3   7.378440  Bachelors Degree and Higher
4   9.266697                Unemployment
```

Figure 15: VIF values for selected variables

For the linear regression, the dataset was split into training and testing sets, with the training data used to build the model and testing data reserved for evaluating its performance. Variables were added to the model incrementally, which progressively increased the R^2, indicating better explanatory power. Transformations like logarithms, square root and adding interaction terms were applied, further improving both R^2 and adjusted R^2, which accounts for the number of predictors to avoid overfitting. Variables and transformations were added iteratively until R^2 no longer improved and adjusted R^2 began to decrease. The resulting formula represents the model that provided the best balance between complexity and predictive power.

```
In [28]: formula = ("Democrat ~ np.log(Population)*Q('Median age')*Homeownership + White * Diversity +"
    ...:     "Q('Bachelors Degree and Higher') * Q('Less Than HighSchool') + np.sqrt(Unemployment)")
    ...: # Split the dataset into training and testing sets
    ...: train_data, test_data = train_test_split(pop, test_size=0.2, random_state=365)
    ...: # Fit the model on the training data
    ...: train_model = smf.ols(formula=formula, data=train_data).fit()
    ...: print(train_model.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:               Democrat   R-squared:                       0.542
Model:                            OLS   Adj. R-squared:                  0.539
Method:                 Least Squares   F-statistic:                     198.1
Date:                Thu, 28 Nov 2024   Prob (F-statistic):               0.00
Time:                        23:38:45   Log-Likelihood:                -8757.6
No. Observations:                2359   AIC:                         1.755e+04
Df Residuals:                    2344   BIC:                         1.763e+04
Df Model:                          14
Covariance Type:            nonrobust
================================================================================================================================
                                                                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------------
Intercept                                                     283.7422     76.717      3.699      0.000     133.302     434.182
np.log(Population)                                            -22.6898      7.603     -2.984      0.003     -37.599      -7.780
Q('Median age')                                                -8.0314      2.063     -3.893      0.000     -12.077      -3.986
np.log(Population):Q('Median age')                              0.8370      0.206      4.071      0.000       0.434       1.240
Homeownership                                                  -2.5283      1.057     -2.392      0.017      -4.601      -0.456
np.log(Population):Homeownership                                0.2176      0.104      2.085      0.037       0.013       0.422
Q('Median age'):Homeownership                                   0.0843      0.028      3.038      0.002       0.030       0.139
np.log(Population):Q('Median age'):Homeownership               -0.0082      0.003     -2.942      0.003      -0.014      -0.003
White                                                          -0.3815      0.031    -12.117      0.000      -0.443      -0.320
Diversity                                                       0.2518      0.034      7.398      0.000       0.185       0.319
White:Diversity                                                -0.0085      0.000    -19.810      0.000      -0.009      -0.008
Q('Bachelors Degree and Higher')                                0.6243      0.053     11.750      0.000       0.520       0.728
Q('Less Than HighSchool')                                      -0.1308      0.069     -1.901      0.057      -0.266       0.004
Q('Bachelors Degree and Higher'):Q('Less Than HighSchool')     -0.0261      0.004     -5.984      0.000      -0.035      -0.018
np.sqrt(Unemployment)                                           7.6009      0.511     14.865      0.000       6.598       8.604
==============================================================================
Omnibus:                       21.315   Durbin-Watson:                   2.036
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               21.775
Skew:                           0.231   Prob(JB):                     1.87e-05
Kurtosis:                       2.913   Cond. No.                     1.16e+07
```

Figure 16: OLS Regression Results

According to the regression summary, the model explains 54.2% of the variability in Democratic affiliation ($R^2 = 0.542$), with an adjusted $R^2 = 0.539$, indicating a good fit while accounting for model complexity. When dealing with human behavior and preferences, R^2 value between 0.10 and 0.50 is generally considered acceptable, as human behavior is often complex and difficult to predict with high accuracy, meaning a relatively low R-squared can still be meaningful if the explanatory variables are statistically significant. Most predictors and interaction terms have significant p-values ($p < 0.05$). The F-statistic 198.1 confirms the model's overall significance.

Next step was predicting the outputs of the regression by creating a plot comparing $\hat{y}$ (predicted values) against y_{train} (actual values) to visualize the behavior and alignment of the model's predictions with the actual values.

```
In [40]:
    ...: y_train = train_data['Democrat']
    ...: # Predict using the training data
    ...: y_hat = train_model.predict(train_data)
    ...: # Plot the actual vs predicted values
    ...: plt.figure(figsize=(16, 12))
    ...: plt.scatter(y_train, y_hat, alpha=0.7, s=150)
    ...: plt.plot([y_train.min(), y_train.max()],
    ...:          [y_train.min(), y_train.max()],
    ...:          color='red', linestyle='--')
    ...: plt.xlabel('Targets (y_train)', size=18)
    ...: plt.ylabel('Predictions (y_hat)', size=18)
    ...: plt.show()
```

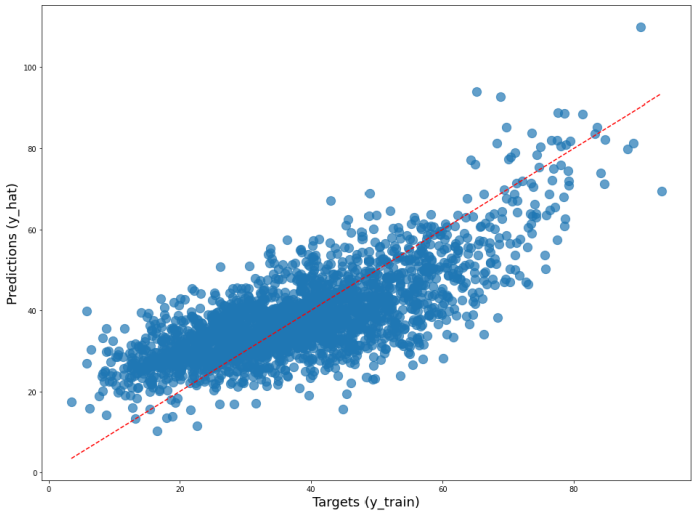

Figure 17: Predicted vs Actual Values (Training Data)

Observing the plot, the predictions $\hat{y}$ and targets y_{train} generally follow the 45-degree line, indicating a reasonable alignment between the regression predictions and the actual target values. However, some deviations from the line are visible, particularly at lower and higher values, suggesting the model may struggle with extreme cases. Despite this, the overall pattern demonstrates that the model captures the underlying trends in the data reasonably well.

Finally, the model was applied to predict outcomes using the y_{text} data by creating a scatter plot with the test targets and the test predictions. This testing phase assesses how well the model generalizes to new, unseen data.

```
In [42]:
    ...: y_test = test_data['Democrat']
    ...: x_test = test_data
    ...: # Predict on the test dataset
    ...: y_hat_test = train_model.predict(x_test)
    ...: # Plot actual vs predicted values
    ...: plt.figure(figsize=(16, 12))
    ...: plt.scatter(y_test, y_hat_test, alpha=0.7, s=150)
    ...: plt.plot([y_test.min(), y_test.max()],
    ...:          [y_test.min(), y_test.max()],
    ...:          color='red', linestyle='--')
    ...: plt.xlabel('Targets (y_test)', size=18)
    ...: plt.ylabel('Predictions (y_hat_test)', size=18)
    ...: plt.show()
```

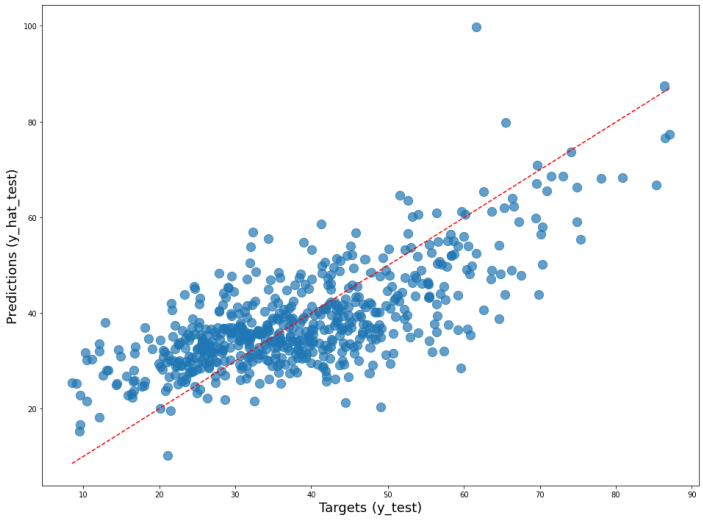

Figure 18: Predicted vs Actual Values (Test Data)

The plot shows that most points cluster around the line, indicating a reasonable level of accuracy in the model's predictions.

In conclusion, the alignment between predicted and actual values in both training and testing phases highlights the model's ability to capture key patterns in the data. While some discrepancies exist, particularly for extreme values, the analysis offers a solid foundation for understanding the relationship between predictors and political affiliation.

Logistic Regression

Logistic Regression is a machine learning classification algorithm that is used to predict the probability of a categorical dependent variable.

General equation for logistic regression:

$$y = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Where:

- x: input value
- y: predicted output
- β_0: intercept term
- β_1: coefficient for input (x)

In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts $P(Y = 1)$ as a function of X. The model is particularly suitable for this analysis, as it can predict whether a county is Democratic (1) or Republican (0) based on various socio-economic and demographic factors.

Logistic regression was applied to model the relationship between political affiliation (Democrat or Republican, encoded as a binary variable) and variables that were selected based on insights from the data visualization step, where box plots and bar graphs highlighted differences in key features between Democratic and Republican counties. The `LogisticRegression` class from `sklearn` and the `Logit` function from `statsmodels` were utilized to fit the model, with logarithmic scaling applied to the population variable

to address its wide range and compress its values. The dataset was split into training and testing sets, and a 15-fold cross-validation was conducted to assess the model's robustness. The mean cross-validation accuracy was 84.53%, with a standard deviation of 2.15%, indicating consistent performance across folds.

The logistic regression summary showed that most predictors had significant coefficients ($p < 0.05$). The pseudo R-squared value of 0.3024 suggests that the model explains about 30.24% of the variability in political affiliation, which is acceptable given the complexity of human behavior and the multifaceted nature of political preferences.

This method demonstrates strong potential for analyzing political affiliation, as it identifies significant relationships between demographic factors and party alignment. Further refinements could improve the model's performance and provide deeper insights into political behavior.

```
#Logistic Regression
columns_to_use = ['Population', 'White', 'Diversity', 'Less Than HighSchool','Median age',
                  'Bachelors Degree and Higher', 'Homeownership', 'Unemployment']
# Subset the data
X = pop[columns_to_use]
y = pop['Political Binary']

# Apply log transformation
log_transform = 'Population'
X[log_transform] = np.log(X[log_transform])

# Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
logit_model_reg = LogisticRegression(max_iter=1000)
logit_model_reg.fit(X_train, y_train)

X_train_sm = sm.add_constant(X_train)
X_test_sm = sm.add_constant(X_test)

# Fit the logistic regression model
logit_model = sm.Logit(y_train, X_train_sm).fit()
print(logit_model.summary())

cv_scores = cross_val_score(logit_model_reg, X_train, y_train, cv=15, scoring='accuracy')
```

Figure 19: Logistic Regression Code

```
                           Logit Regression Results
==============================================================================
Dep. Variable:       Political Binary   No. Observations:                 2359
Model:                          Logit   Df Residuals:                     2350
Method:                           MLE   Df Model:                            8
Date:                Fri, 29 Nov 2024   Pseudo R-squ.:                  0.3024
Time:                        11:28:19   Log-Likelihood:                -872.35
converged:                       True   LL-Null:                       -1250.6
Covariance Type:            nonrobust   LLR p-value:                5.084e-158
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                          0.8082      1.147      0.705      0.481      -1.439       3.055
Population                     0.2526      0.055      4.589      0.000       0.145       0.360
White                         -0.0719      0.008     -8.994      0.000      -0.088      -0.056
Diversity                     -0.0087      0.007     -1.212      0.226      -0.023       0.005
Less Than HighSchool          -0.0836      0.017     -4.822      0.000      -0.118      -0.050
Median age                     0.0745      0.016      4.600      0.000       0.043       0.106
Bachelors Degree and Higher    0.0605      0.011      5.474      0.000       0.039       0.082
Homeownership                 -0.0435      0.010     -4.443      0.000      -0.063      -0.024
Unemployment                   0.1614      0.026      6.107      0.000       0.110       0.213
==============================================================================================
Cross-Validation Accuracy Scores: [0.83544304 0.87974684 0.82911392 0.85443038 0.82802548 0.84076433
 0.85987261 0.84713376 0.81528662 0.89171975 0.85987261 0.8089172
 0.84076433 0.8343949  0.85350318]
Mean Cross-Validation Accuracy: 0.8453
Standard Deviation of CV Accuracy: 0.0215
```

Figure 20: Logistic Regression Results