# Group 4’s Project

Riley Hawley, Eddie Kuang, Jince Shi, Eric Tashji

# Problem Statement

“Republican voters tend to reside more in rural areas, whereas Democrat voters tend to reside more in urban areas. Republicans want to increase their voter turnout in urban areas by 5% and Democrats want to increase their turnout in rural areas by 5%.”

This is an interesting problem because it explores the values that rural and urban residents have that influence the ways that they vote, and can point to significant differences in the values between rural and urban residents in America.

# Dataset Description

# Sources


- https://www.nj.gov/state/elections/election-information-ballots-cast.shtml
- https://www.citypopulation.de/en/usa/ua/NJ__new_jersey/
- https://www.nj.gov/health/fhs/primarycare/rural-health/DesignatedRuralAreasinNewJersey

# Official Primary Election Turnout

**June 4, 2024**

| COUNTY | Total Number of Persons Entitled to Vote | Total Number Democratic Ballots Cast | Total Number Republican Ballots Cast | Total Number of Ballots Cast | Percent of Eligble Voters | Total Number of Ballots Rejected | Number of Election Districts |
|---|---|---|---|---|---|---|---|
| Atlantic | 200,594 | 13,621 | 12,816 | 26,437 | 13% | 124 | 151 |
| Bergen | 665,551 | 55,814 | 33,466 | 89,280 | 13% | 590 | 562 |
| Burlington | 330,231 | 40,288 | 23,942 | 64,230 | 19% | 324 | 366 |
| Camden | 381,605 | 49,401 | 15,249 | 64,650 | 17% | 429 | 343 |
| Cape May | 74,186 | 5,691 | 10,535 | 16,226 | 22% | 47 | 127 |
| Cumberland | 94,093 | 6,333 | 5,281 | 11,614 | 12% | 70 | 92 |
| Essex | 572,278 | 53,528 | 9,187 | 62,800 | 11% | 1,358 | 575 |
| Gloucester | 216,050 | 20,858 | 13,588 | 34,446 | 16% | 116 | 229 |
| Hudson | 400,344 | 48,133 | 6,980 | 55,113 | 14% | 571 | 466 |
| Hunterdon | 105,460 | 8,095 | 11,216 | 19,311 | 18% | 88 | 117 |
| Mercer | 260,380 | 27,596 | 7,591 | 35,187 | 14% | 335 | 243 |
| Middlesex | 570,166 | 56,671 | 21,793 | 78,464 | 14% | 1,043 | 615 |
| Monmouth | 293,405 | 31,319 | 34,473 | 65,792 | 22% | 702 | 474 |
| Morris | 385,282 | 27,037 | 30,445 | 57,482 | 15% | 312 | 396 |
| Ocean | 463,218 | 22,563 | 40,801 | 63,364 | 14% | 403 | 414 |
| Passaic | 326,713 | 29,192 | 15,491 | 44,683 | 14% | 266 | 283 |
| Salem | 48,346 | 2,918 | 3,201 | 6,119 | 13% | 105 | 43 |
| Somerset | 252,431 | 20,851 | 16,259 | 37,110 | 15% | 218 | 267 |
| Sussex | 117,421 | 5,717 | 13,754 | 19,471 | 17% | 229 | 119 |
| Union | 379,658 | 33,221 | 12,052 | 45,273 | 12% | 642 | 434 |
| Warren | 85,561 | 4,765 | 8,883 | 13,648 | 16% | 61 | 86 |
| **TOTAL** | **6,222,973** | **563,612** | **347,003** | **910,700** | **15%** | **8,033** | **6,402** |

**Population Density by County and Municipality: New Jersey, Census 2020 and Estimates 2022**

| FIPS County | FIPS MCD | AreaName | Land Area (square miles) | Total Population Census 2020 | Total Population Estimates 2022 | Persons per Square Mile Census 2020 | Persons per Square Mile Estimates 2022 |
|---|---|---|---|---|---|---|---|
| ***000*** | | ***New Jersey*** | ***7,354.8*** | ***9,289,031*** | ***9,261,699*** | ***1,263.0*** | ***1,259.3*** |
| ***001*** | ***00000*** | **Atlantic County** | *555.5* | *274,536* | *275,638* | *494.2* | *496.2* |
| 001 | 00100 | Absecon city | 5.5 | 9,136 | 9,155 | 1,670.7 | 1,674.2 |
| 001 | 02080 | Atlantic City city | 10.8 | 38,501 | 38,561 | 3,578.0 | 3,583.6 |
| 001 | 07810 | Brigantine city | 6.5 | 7,718 | 7,665 | 1,183.3 | 1,175.2 |
| 001 | 08680 | Buena borough | 7.6 | 4,507 | 4,500 | 595.2 | 594.2 |
| 001 | 08710 | Buena Vista township | 41.1 | 7,037 | 7,079 | 171.3 | 172.3 |
| 001 | 15160 | Corbin City city | 7.7 | 473 | 481 | 61.5 | 62.5 |
| 001 | 20290 | Egg Harbor township | 67.0 | 47,844 | 47,946 | 713.6 | 715.1 |
| 001 | 20350 | Egg Harbor City city | 10.9 | 4,396 | 4,408 | 405.1 | 406.2 |
| 001 | 21870 | Estell Manor city | 53.4 | 1,667 | 1,677 | 31.2 | 31.4 |
| 001 | 23940 | Folsom borough | 8.3 | 1,807 | 1,810 | 218.7 | 219.0 |
| 001 | 25560 | Galloway township | 88.7 | 37,815 | 37,870 | 426.5 | 427.1 |
| 001 | 29280 | Hamilton township | 110.9 | 27,483 | 28,155 | 247.8 | 253.9 |
| 001 | 29430 | Hammonton town | 40.7 | 14,710 | 14,833 | 361.0 | 364.0 |
| 001 | 40530 | Linwood city | 3.8 | 6,956 | 6,962 | 1,825.0 | 1,826.6 |
| 001 | 41370 | Longport borough | 0.4 | 897 | 884 | 2,269.9 | 2,237.0 |
| 001 | 43890 | Margate City city | 1.4 | 5,312 | 5,216 | 3,751.1 | 3,683.3 |
| 001 | 49410 | Mullica township | 56.4 | 5,815 | 5,820 | 103.1 | 103.2 |
| 001 | 52950 | Northfield city | 3.6 | 8,437 | 8,451 | 2,355.1 | 2,359.0 |
| 001 | 59640 | Pleasantville city | 5.7 | 20,629 | 20,662 | 3,605.9 | 3,611.6 |
| 001 | 60600 | Port Republic city | 7.5 | 1,102 | 1,111 | 147.7 | 148.9 |
| 001 | 68430 | Somers Point city | 4.0 | 10,474 | 10,495 | 2,611.7 | 2,616.9 |
| 001 | 75620 | Ventnor City city | 2.0 | 9,206 | 9,246 | 4,708.3 | 4,728.8 |
| 001 | 80330 | Weymouth township | 11.8 | 2,614 | 2,651 | 221.1 | 224.2 |

# Key Components of the Dataset

- Includes data on demographics, family size, income, and voting patterns.
- Variables: Age, Gender, Marital Status, Number of Children, Occupation, Salary, Standard of Living, Rate of Voter Turnout, Most Important Issue, Engagement with Campaign, Votes for Democrat (Last 8 Years), Votes for Republican (Last 8 Years), Votes for Other Candidate (Last 8 Years).
- Analyzed across rural and urban counties.

# Addressing Data Issues

- Missing values in "Number of Children" filled using the median.
- Outliers in "Salary" capped to prevent skewed results.
- Ensured data remained representative and reliable for analysis.

# Data Transformation for Analysis

- Categorical data (e.g., Gender, Marital Status) converted into numerical values.
- Continuous features (e.g., Age, Salary) normalized.
- Ensured equal contribution from all features during analysis.

# Linear Regression

3 models to analyze voter turnout and party preference in urban and rural areas

# Linear Regression

- Linear regression models the relationship between multiple input features and a continuous output variable.
- It assumes a linear relationship between the features and the target variable.
- The goal is to find the best-fitting linear equation that minimizes the sum of squared residuals.

# Model 1: Voter Turnout Prediction Implementation details

- Used sklearn's LinearRegression class.Features included both numeric and categorical variables.Categorical variables were one-hot encoded.Numeric features were standardized using StandardScaler.Used a pipeline to combine preprocessing and model fitting.
- Metrics used: Mean Squared Error (MSE) and R-squared (R2) score.Performed cross-validation to assess model stability.

# Model 1: Voter Turnout Prediction output

Dataset Information: Urban dataset: 789,865 samples, 14 features. Rural dataset: 120,835 samples, 14 features

Model Performance:

Urban Model:MSE: 0.3637 R2 Score: 0.3116 Mean CV R2 score: 0.3081

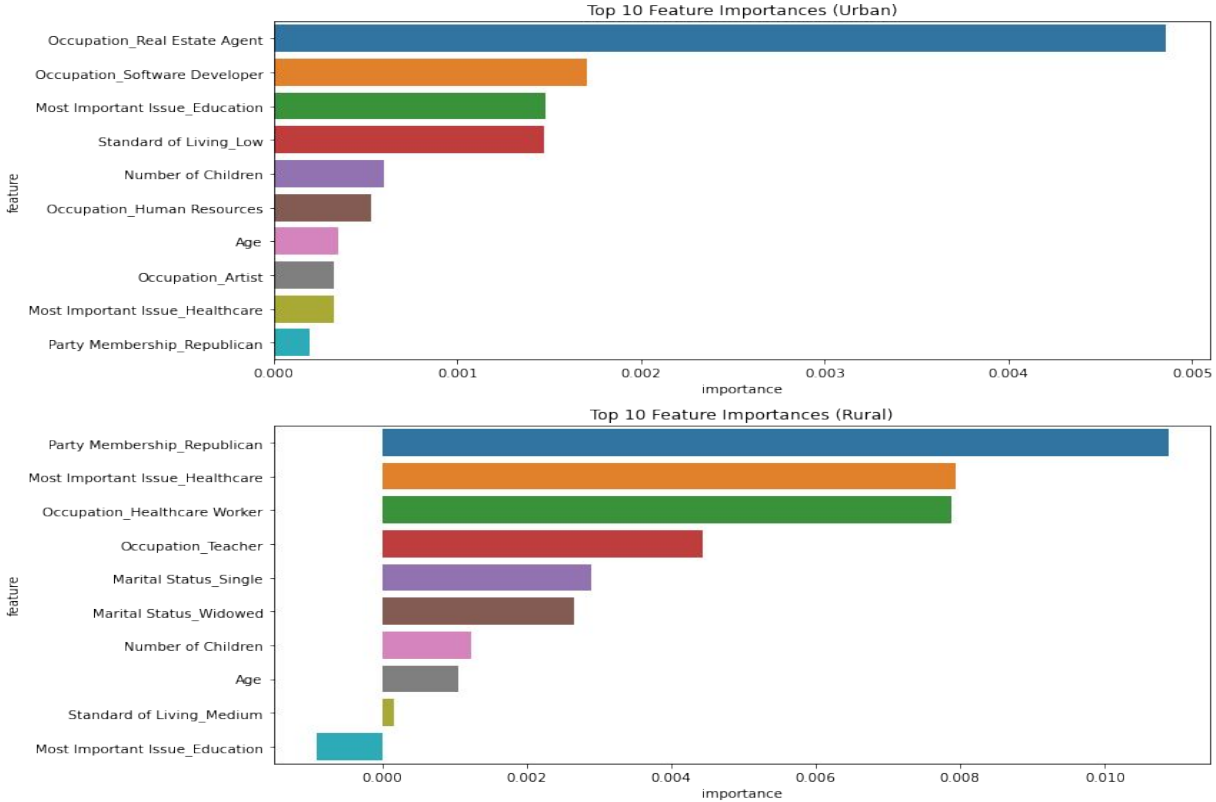
Rural Model:MSE: 0.3657 R2 Score: 0.3214. Mean CV R2 score: 0.3081

Urban Top 5:Occupation_Real Estate Agent (0.004858) Occupation_Software Developer (0.001703)

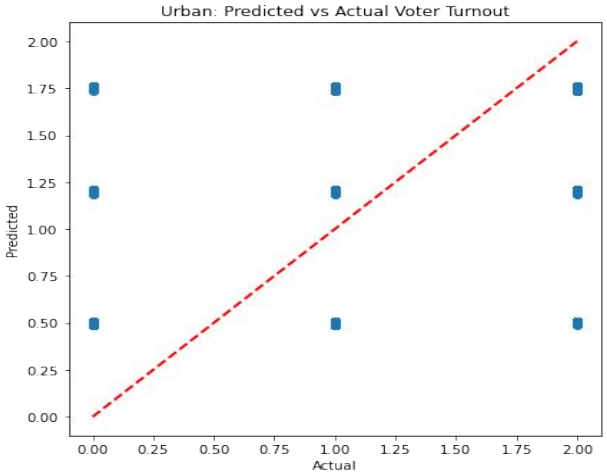
Most Important Issue_Education (0.001477). Standard of Living_Low (0.001471) Number of Children (0.000605)

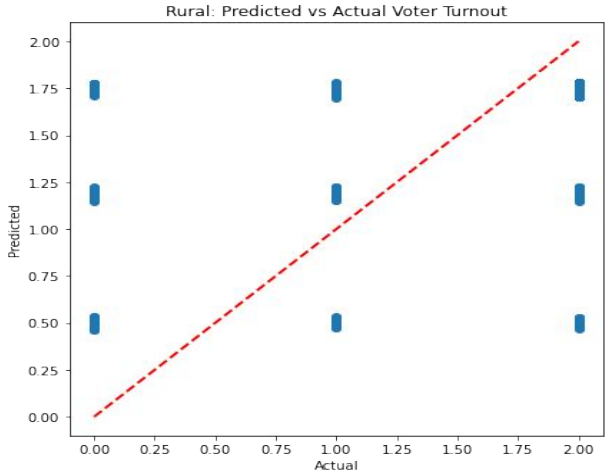
Rural Top 5:Party Membership_Republican (0.010878). Most Important Issue_Healthcare (0.007936).

Occupation_Healthcare Worker (0.007879). Occupation_Teacher (0.004436). Marital Status_Single (0.002902)

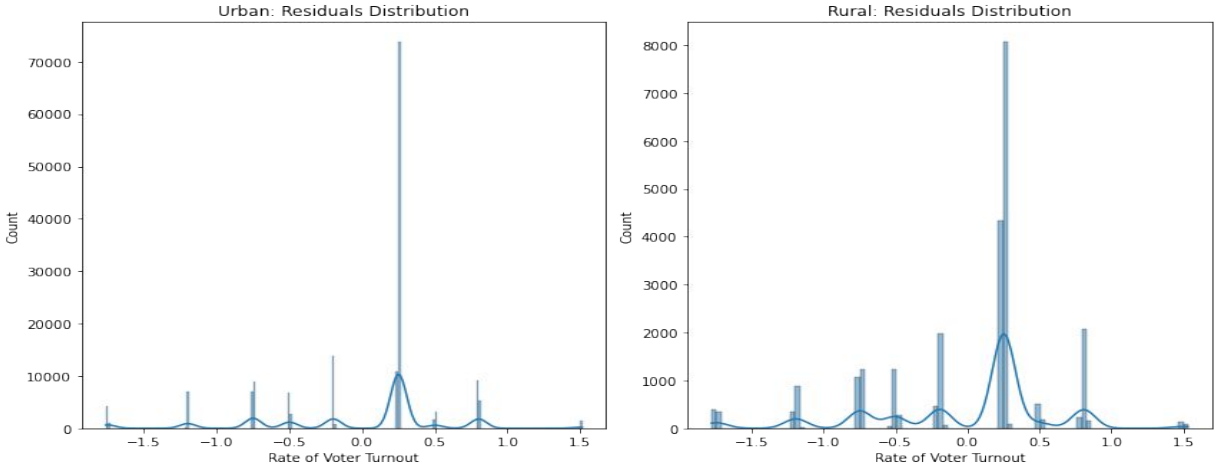



Implementation: The model used multiple features to predict voter turnout in urban and rural areas.
Results: In urban areas, Occupation_Real Estate Agent was most important (0.004858), while in rural areas, Party Membership_Republican had the highest impact (0.010878).




Implementation: The model's predictions were compared to actual voter turnout for both urban and rural areas. Results: Both urban and rural models showed moderate predictive power, with R-squared scores of 0.3116 and 0.3214 respectively.




Implementation: The differences between predicted and actual voter turnout were plotted for urban and rural models. Results: Both models showed roughly symmetric residual distributions, indicating unbiased predictions with mean CV R2 scores of about 0.308 for both urban and rural areas.

# Model 1 conclusion

- These results indicate that the models have moderate predictive power for voter turnout, with R2 scores around 0.31-0.32. The feature importances reveal different factors influencing turnout in urban and rural areas, with occupations playing a significant role in urban areas and party membership being crucial in rural areas.

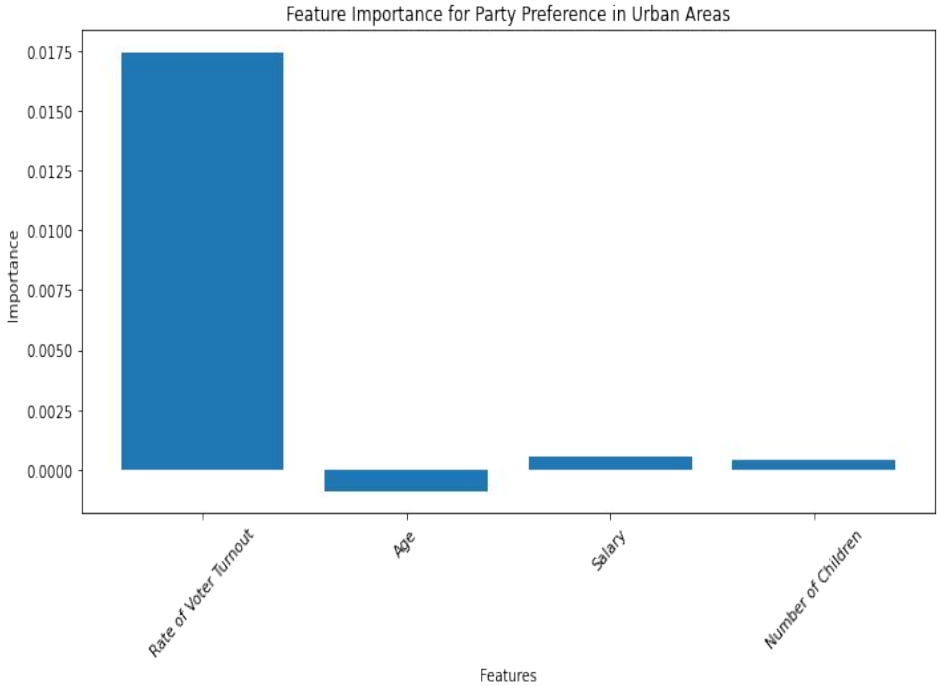
# Model 2: Party Preference Prediction (Limited Features) Implementation details

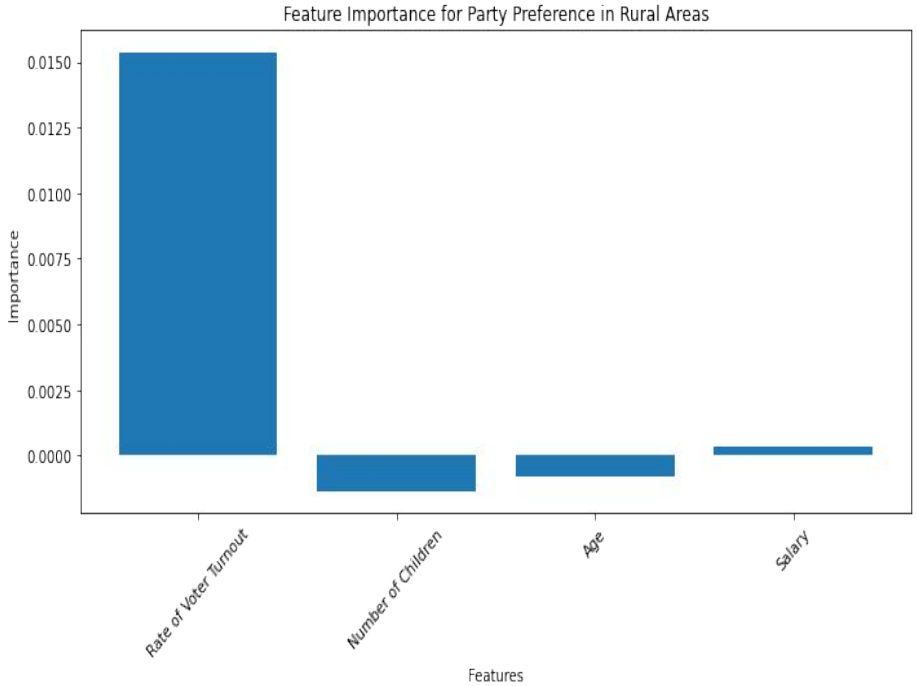
- Used a subset of features: Age, Number of Children, Salary, Rate of Voter Turnout.Features were standardized using StandardScaler.No categorical variables were used in this model.
- Used Mean Squared Error (MSE) and R-squared (R2) score.

# Model 2: Party Preference Prediction (Limited Features) output

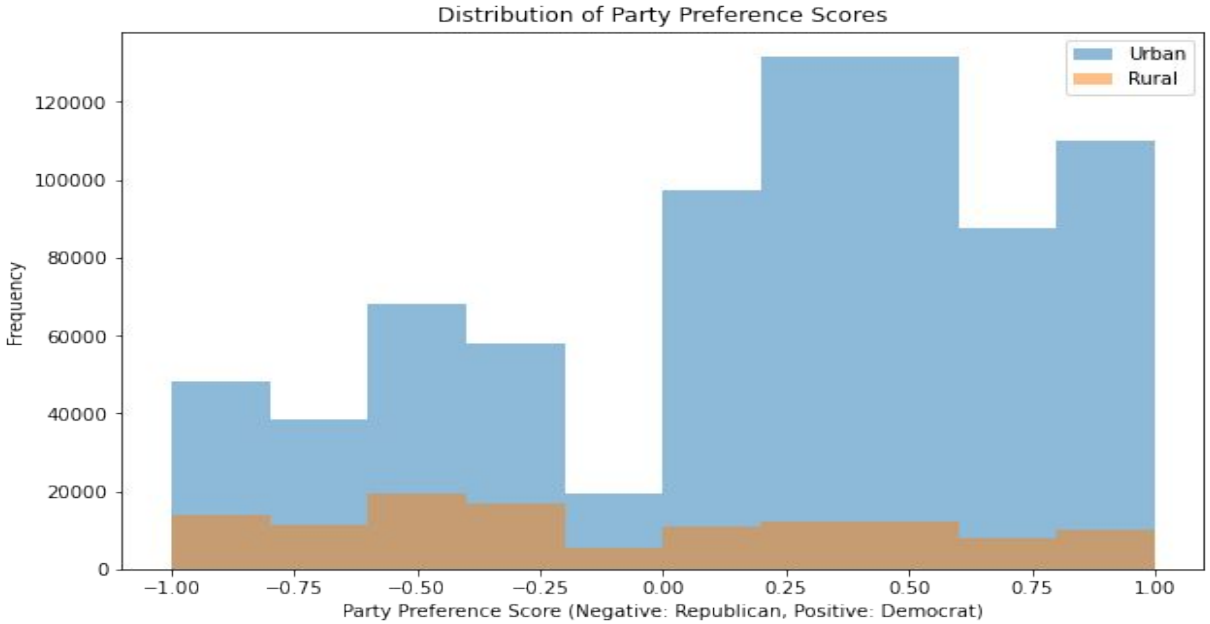
- Model Performance: Urban Model:Mean Squared Error: 0.3366
- R-squared Score: 0.0007
- Rural Model:Mean Squared Error: 0.3747
- R-squared Score: 0.0009

1. Feature Importance: Urban:Rate of Voter Turnout (0.017442)
2. Age (-0.000892)
3. Salary (0.000550)
4. Number of Children (0.000410)
5. Rural:Rate of Voter Turnout (0.015370)
6. Number of Children (-0.001355)
7. Age (-0.000775)
8. Salary (0.000343)




Implementation: The model used four features to predict party preference in urban areas. Results: Rate of Voter Turnout had the highest importance (0.017442), while Age had a slight negative impact (-0.000892).




Implementation: The same four features were used to predict party preference in rural areas. Results: Rate of Voter Turnout was also the most important feature (0.015370), with Number of Children having the second highest impact (-0.001355).

# Model 2: conclusion

- Key Findings:Very low R-squared scores indicate poor predictive power
- Rate of Voter Turnout is the most important feature for both urban and rural models
- The model performs slightly better for rural areas, but the difference is negligible
- Comparison of Voting Patterns:Average Urban Party Preference Score: 0.2088 (leans Democratic)
- Average Rural Party Preference Score: -0.0836 (leans Republican)

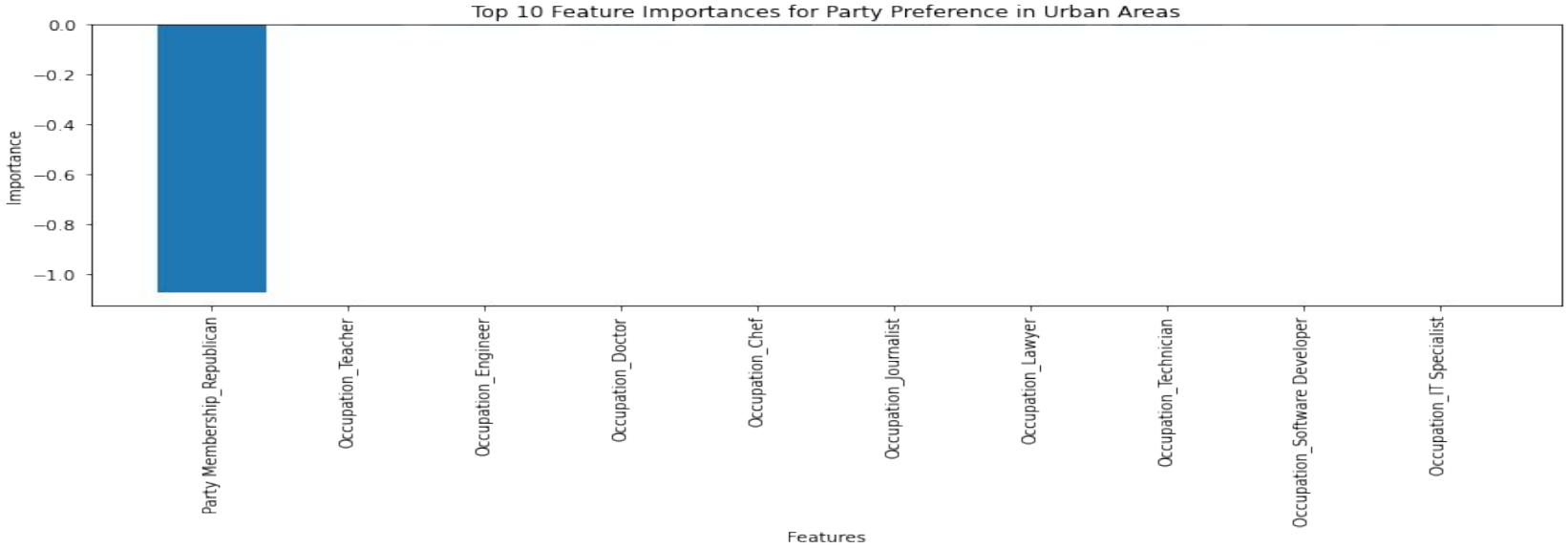
- This model demonstrates that using only these limited features (Age, Number of Children, Salary, Rate of Voter Turnout) is insufficient to predict party preference accurately, as evidenced by the very low R-squared scores.

# Model 3: Party Preference Prediction (Extended Features) Implementation details

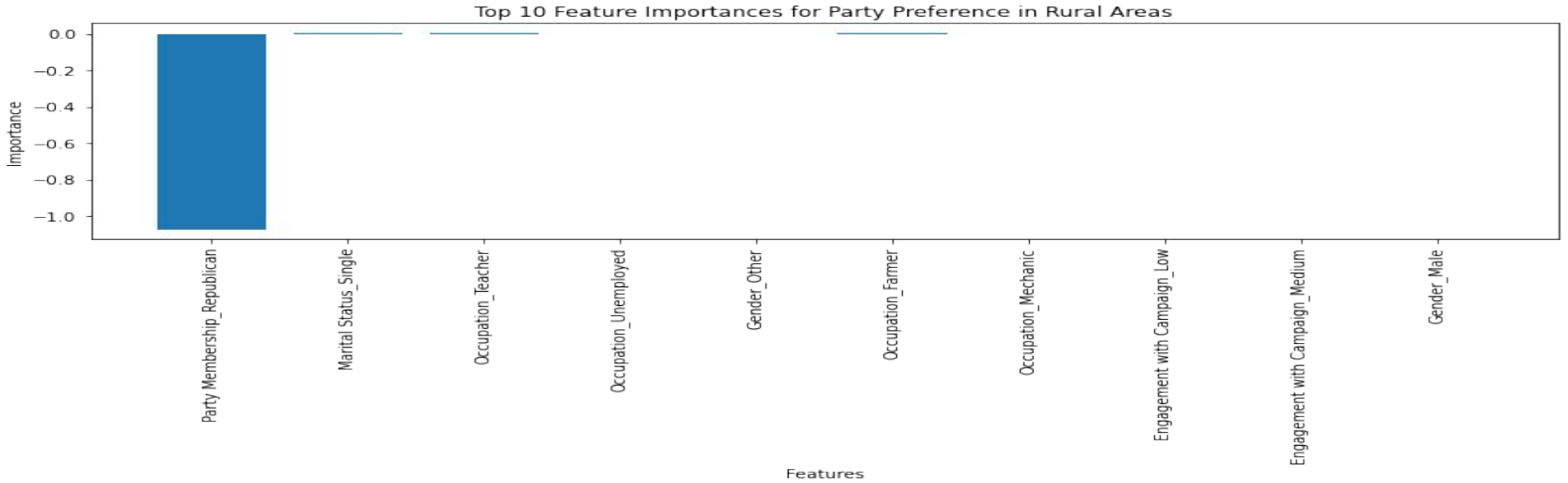
- Included both numeric and categorical features.Used ColumnTransformer for preprocessing:StandardScaler for numeric features.
- OneHotEncoder for categorical features.
- Implemented in a sklearn Pipeline for streamlined preprocessing and model fitting.
- Used Mean Squared Error (MSE) and R-squared (R2) score.Analyzed feature importances based on model coefficients.

# Model 3 Output:

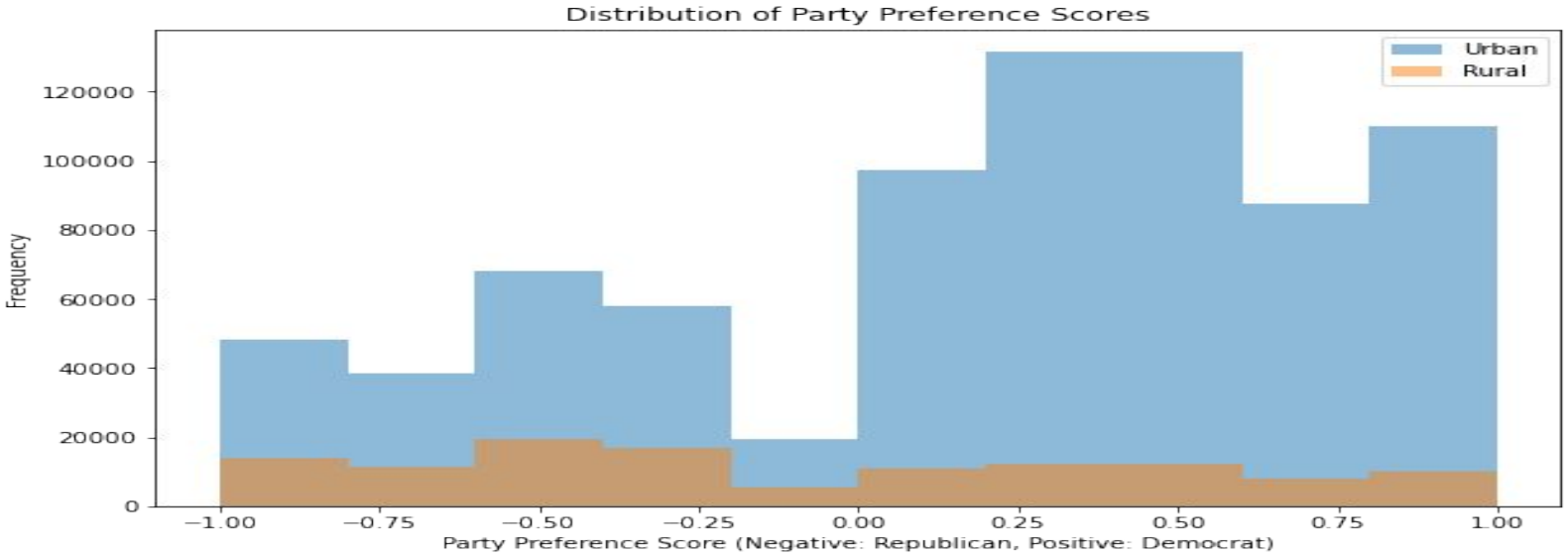
- Urban Area Model Performance: Mean Squared Error: 0.09124208383412899 R-squared Score: 0.7290889850234468
- Rural Area Model Performance: Mean Squared Error: 0.09141329487440927 R-squared Score: 0.7562878399850786
- Comparison of Voting Patterns: Average Urban Party Preference Score: 0.2088 Average Rural Party Preference Score: -0.0836




Implementation: The model used an extended set of features, including categorical variables, to predict party preference in urban areas. Results: Party Membership_Republican had the strongest negative impact (-1.073949), followed by various occupations with smaller influences.




Implementation: The same extended feature set was used to predict party preference in rural areas. Results: Party Membership_Republican again had the strongest negative impact (-1.073616), with Marital Status_Single showing the highest positive influence (0.005097).




Implementation and Results: The histogram shows that urban areas lean Democratic (average score 0.2088) and rural areas lean Republican (average score -0.0836), with the model achieving high R-squared scores of 0.7291 for urban and 0.7563 for rural areas.

# Model 3: conclusion

**Urban Areas:**

Occupation plays a significant role, with teachers, engineers, doctors, chefs, and journalists showing some impact on party preference.

**Rural Areas:**

Marital Status (being single) has a positive impact on Democratic preference.

Occupation (especially teachers and farmers) influences party preference.

Gender and Engagement with Campaign also show some importance.

**Urban vs Rural Differences:**

There is a clear difference in party preference between urban and rural areas:

- Urban areas show an average party preference score of 0.2088, indicating a lean towards the Democratic Party.
- Rural areas show an average party preference score of -0.0836, indicating a lean towards the Republican Party.

This confirms the initial assumption that urban areas tend to favor Democrats while rural areas tend to favor Republicans.

# Logistic Regression

# Using Logistic Regression for Voter Analysis

- Logistic regression was employed to address two critical tasks:
  - Predicting Voter Turnout
  - Predicting Party Affiliation
- Chosen for its effectiveness in binary classification problems.

# Advantages of Logistic Regression

- Ideal for binary classification tasks, such as voter turnout and party affiliation.
- Offers easily interpretable results through model coefficients.
- Compatible with both numerical (e.g., age, salary) and categorical data (e.g., marital status, party membership) after proper encoding.

# Focus Areas for Logistic Regression

- Age, Salary, Number of Children, Marital Status, Occupation
- Party Membership, Engagement with Ground Campaigns
- Data preprocessing included one-hot encoding for categorical variables and normalization for continuous features.

# Logistic Regression (Implementation)

# Data Processing

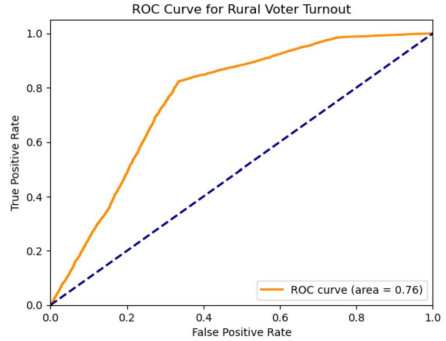
- Imputed missing values in "Number of Children" using the median.
- Capped extreme outliers in "Salary" to prevent skewed results.
- Categorical data (e.g., gender, marital status) converted via one-hot encoding.
- Continuous features like age and salary normalized.

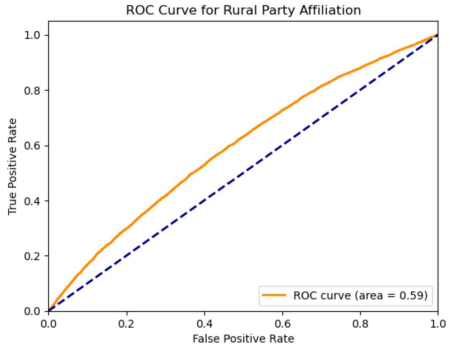
# Training the Logistic Regression Model

- Dataset split: 80% for training, 20% for testing.
- Model built using SGDClassifier with logistic regression loss.
- Evaluation on testing set provided insights into accuracy and reliability.

# Model Performance and Evaluation

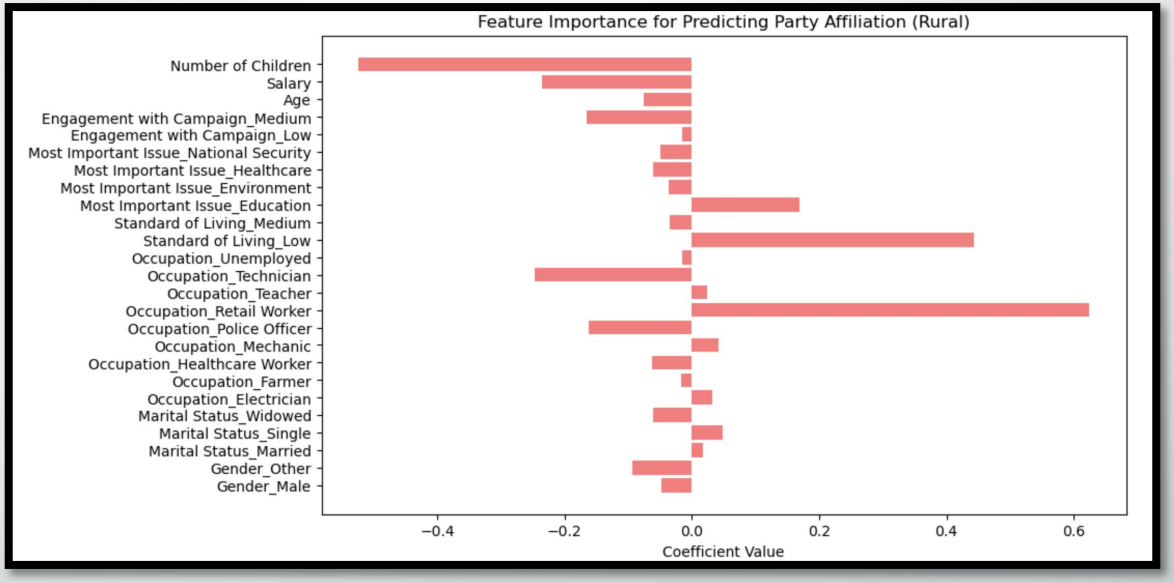
- Metrics used: Precision, Recall, F1-Score, and Accuracy.
- Rural Data: 59% accuracy for party affiliation, 76% for voter turnout.
- Urban Data: 69% accuracy for party affiliation, 77% for voter turnout.
- ROC AUC Scores: Indicated model reliability, especially for voter turnout.



# Understanding the Logistic Regression Coefficients

- Coefficients provide insights into how features influence outcomes.
- Rural Data: Retail workers linked to Democrat affiliation, police officers to Republican.
- Urban Data: High-salary occupations like finance linked to Democrat affiliation.
- Voter Turnout: Higher salaries negatively associated with turnout, engagement positively associated.

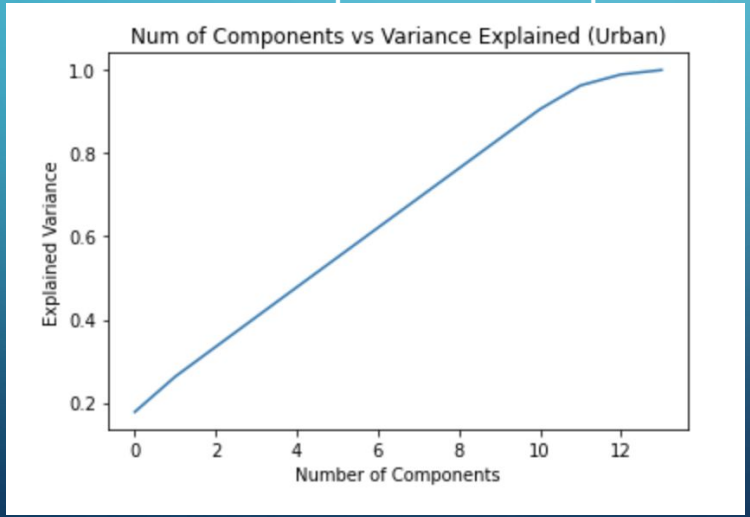


# GAUSSIAN MIXTURE MODEL

-USE TO EXPLORE, FIND PATTERNS AND INFORMATION ABOUT URBAN AND RURAL VOTERS.

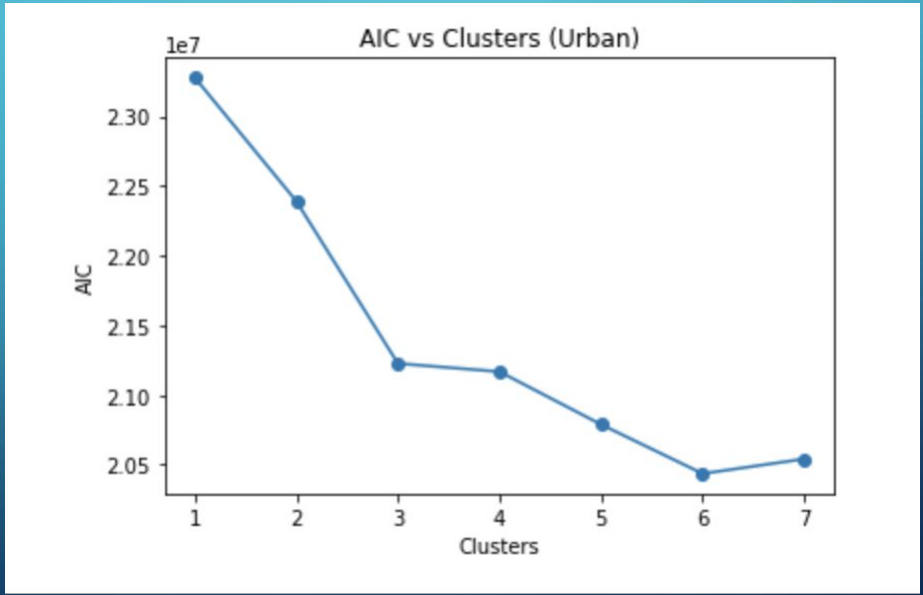
# URBAN DATA -PCA

- PCA was used to reduce number of dimensionalities.
- We want the number of components to explain about 90%-95% of variance.

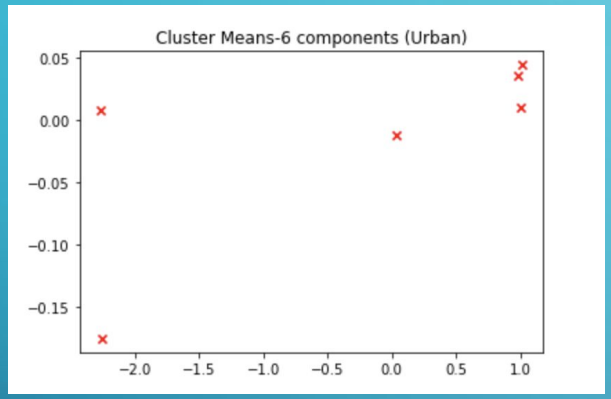


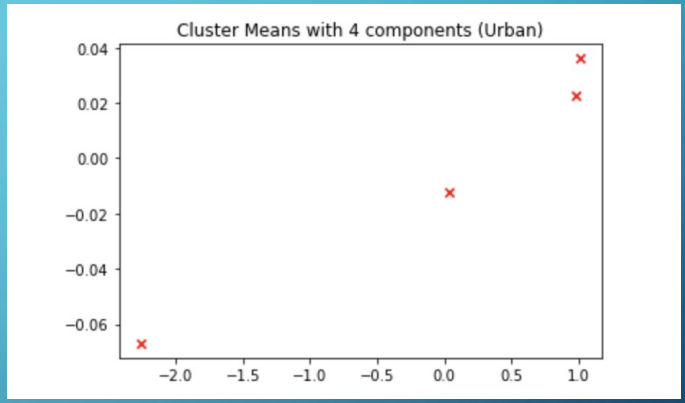
# URBAN DATA-AIC

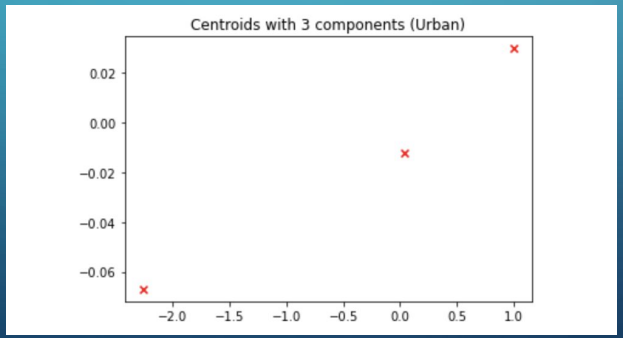
- Checking optimal cluster using AIC.



# URBAN DATA-CENTROIDS



# URBAN DATA-RESULTS

| Cluster | Age | Gender | Marital Status | Number of Children | Occupation |
|---|---|---|---|---|---|
| 0 | 53.987892 | 0.574418 | 1.501406 | 2.470703 | 9.431493 |
| 1 | 54.048862 | 0.576017 | 1.498727 | 2.473116 | 9.641882 |
| 2 | 54.303936 | 0.576438 | 1.483829 | 2.477594 | 9.482920 |

| Cluster | Salary | Standard of Living | Party Membership |
|---|---|---|---|
| 0 | 49961.077672 | 1.001485 | 1.000000 |
| 1 | 49994.875508 | 0.999745 | 0.000000 |
| 2 | 502959.490814 | 0.994415 | 0.702819 |

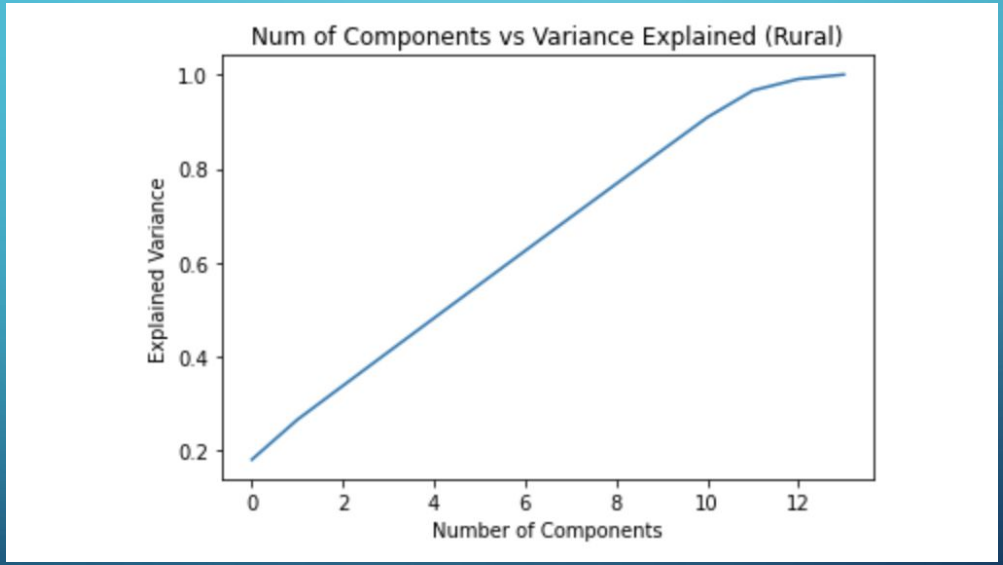
| Cluster | Rate of Voter Turnout | Most Important Issue |
|---|---|---|
| 0 | 1.080062 | 2.001836 |
| 1 | 1.090456 | 2.002413 |
| 2 | 1.083907 | 2.017665 |

| Cluster | Engagement with Campaign | Votes for Democrat (Last 8 years) |
|---|---|---|
| 0 | 1.100196 | 6.003769 |
| 1 | 1.199827 | 2.003160 |
| 2 | 1.124692 | 4.791401 |

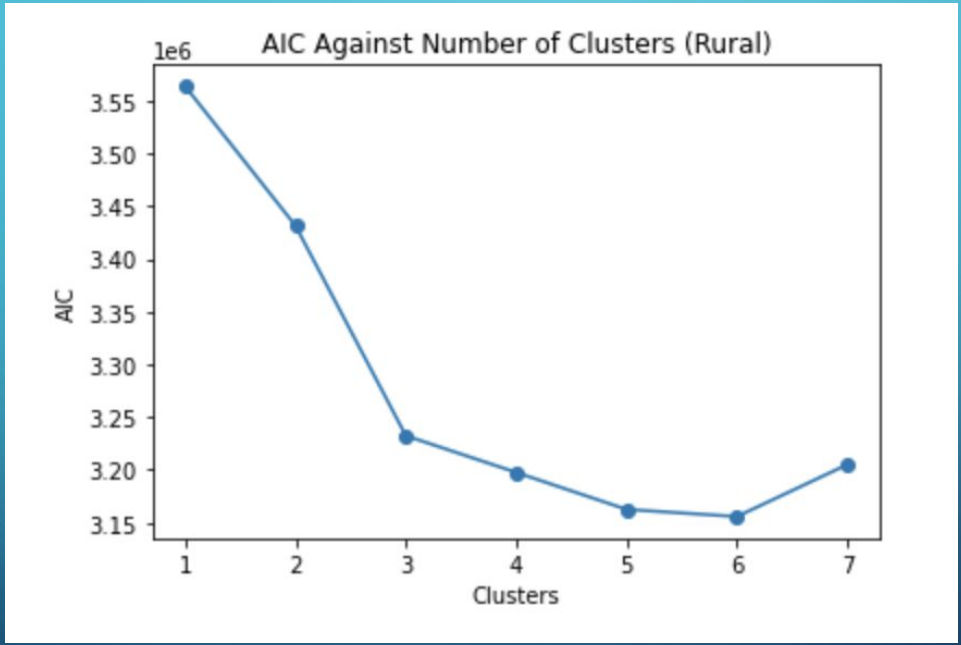
| Cluster | Votes for Republican (Last 8 years) |
|---|---|
| 0 | 1.998375 |
| 1 | 5.999474 |
| 2 | 3.175347 |

| Cluster | Votes for Other Candidate (Last 8 years) |
|---|---|
| 0 | 0.999309 |
| 1 | 1.000200 |
| 2 | 0.997272 |

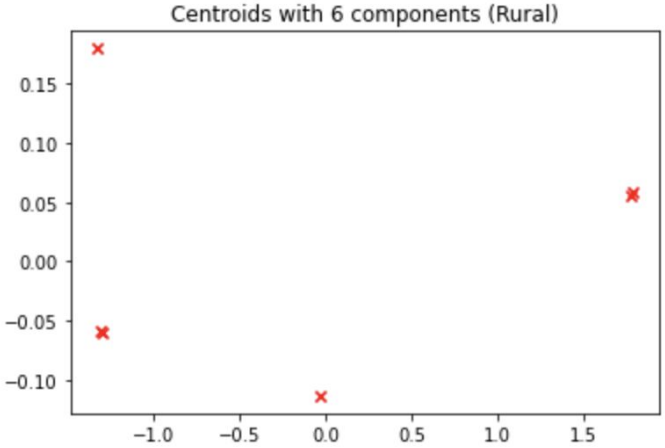
# RURAL DATA-PCA

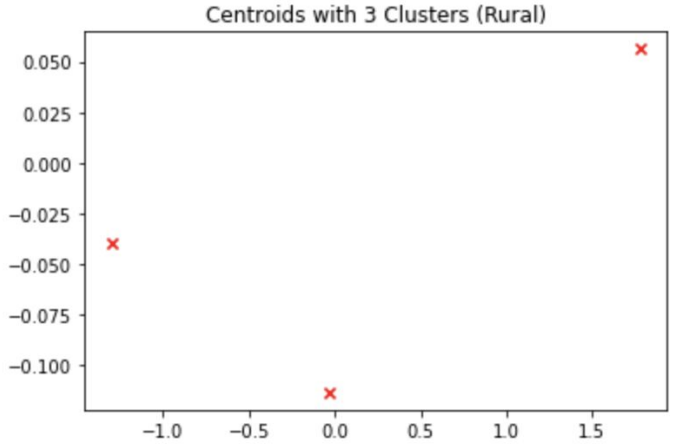


# RURAL-AIC


AIC Against Number of Clusters (Rural)
1e6
3.55
3.50
3.45
3.40
3.35
3.30
3.25
3.20
3.15
AIC
1
2
3
4
5
6
7
Clusters

# RURAL-CENTROIDS



| Cluster | Age | Gender | Marital Status | Number of Children | Occupation |
|---|---|---|---|---|---|
| 0 | 54.078143 | 0.573636 | 1.507458 | 2.468495 | 4.368491 |
| 1 | 54.026521 | 0.569804 | 1.496035 | 2.466210 | 4.661689 |
| 2 | 53.814719 | 0.577489 | 1.519481 | 2.520346 | 4.559307 |

| Cluster | Salary | Standard of Living | Party Membership |
|---|---|---|---|
| 0 | 50029.160131 | 0.997685 | 0.000000 |
| 1 | 49899.602572 | 0.997844 | 1.000000 |
| 2 | 500266.003307 | 1.051082 | 0.415584 |

| Cluster | Rate of Voter Turnout | Most Important Issue |
|---|---|---|
| 0 | 1.090918 | 2.002734 |
| 1 | 1.076499 | 1.997528 |
| 2 | 1.083983 | 2.024242 |

| Cluster | Engagement with Campaign | Votes for Democrat (Last 8 years) |
|---|---|---|
| 0 | 1.194381 | 1.995660 |
| 1 | 1.100172 | 6.004133 |
| 2 | 1.153247 | 3.661472 |

| Cluster | Votes for Republican (Last 8 years) |
|---|---|
| 0 | 5.998626 |
| 1 | 2.002887 |
| 2 | 4.344589 |

| Cluster | Votes for Other Candidate (Last 8 years) |
|---|---|
| 0 | 1.004066 |
| 1 | 1.001009 |
| 2 | 0.993939 |

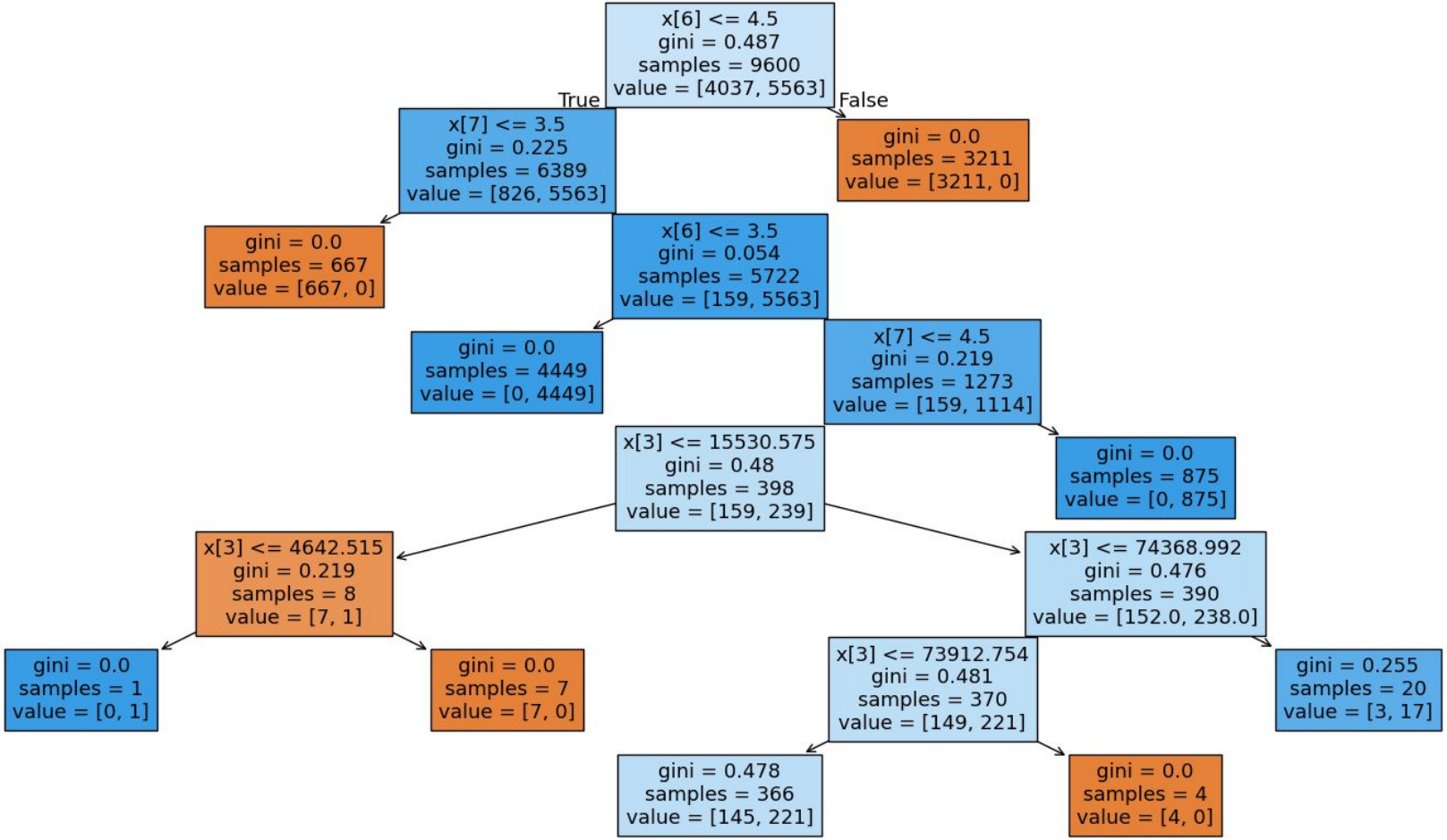
# Decision Tree

-Used to determine trends among party affiliation and voter turnout.

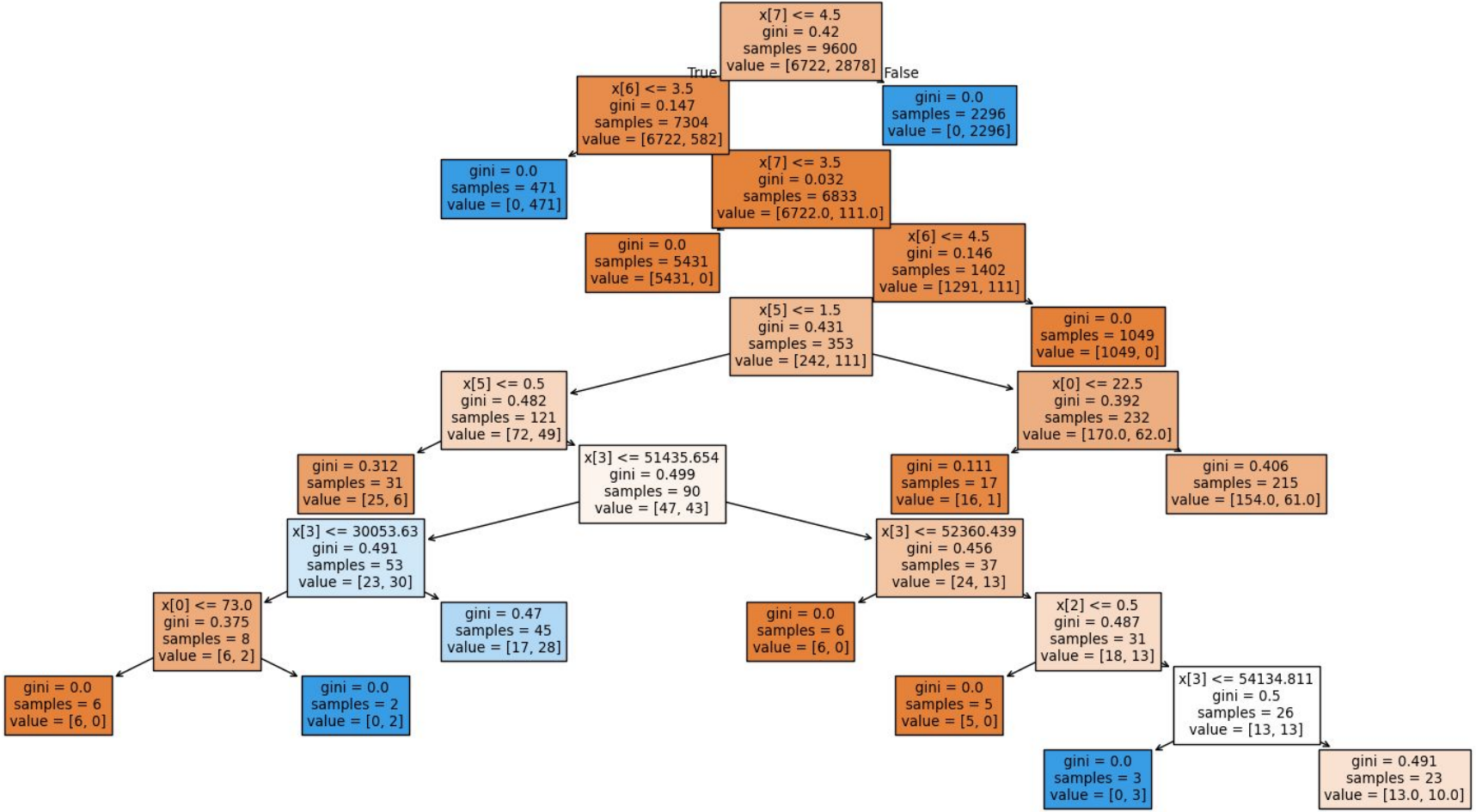
# Decision Tree

- I decided to implement a decision tree in order to examine potential trends among voters in urban and rural areas.
- I began by organizing the data into qualitative, quantitative, and boolean data.
- I then chose 1200  random samples from the urban dataset and 1200 random samples from the rural data set.
- 80% of each sample was set out for training data and 20% of each set was set aside for testing data.
- I implemented the decision tree by adjusting for missing data, and then including only the quantitative and boolean data in the decision tree model.
- I then made a prediction model using the decision tree and tested it with the test data.

# Rural Decision Tree



# Urban Decision Tree



# Comparison among these machine learning algorithms

- Our project employed four distinct machine learning algorithms to analyze voter behavior and party preferences in urban and rural areas: Linear Regression, Logistic Regression, Gaussian Mixture Model (GMM), and Decision Trees. Each algorithm provided unique insights and had its own strengths and limitations.

# Linear Regression vs. Logistic Regression:

Performance: Linear regression models for voter turnout showed moderate predictive power (R2 scores of 0.31-0.32) in both urban and rural areas. Logistic regression performed better for voter turnout prediction (accuracy of 76-77%) but was less effective for party affiliation prediction (59-69% accuracy).Insights: Linear regression highlighted the importance of occupation and key issues (e.g., education in urban areas, healthcare in rural areas) in predicting voter turnout. Logistic regression provided more nuanced insights into how specific occupations relate to party affiliation.Strengths: Linear regression offered easily interpretable coefficients, while logistic regression was better suited for binary classification tasks like predicting turnout.

# Gaussian Mixture Model (GMM) vs. Decision Trees:

- Clustering vs. Classification: GMM provided a clustering approach, identifying three main voter groups in both urban and rural areas, while decision trees offered a classification approach with very high accuracy (98%).Insights: GMM revealed that salary, party membership, and voting history were the most significant factors in cluster formation. Decision trees identified specific voter groups that could be targeted to increase turnout (e.g., 7% of rural Democrats and 6% of urban Republicans who voted infrequently).Strengths: GMM excelled at uncovering underlying patterns in the data, especially among high-income voters. Decision trees provided clear, actionable insights for increasing voter turnout.

# Overall Comparison:

- Accuracy: Decision trees achieved the highest accuracy (98%) for classification tasks, followed by logistic regression (59-77%), then linear regression (R2 of 0.31-0.32 for voter turnout).Interpretability: Linear regression and decision trees offered the most easily interpretable results, with clear coefficients and decision paths respectively. Logistic regression provided interpretable odds ratios, while GMM required more complex interpretation of cluster characteristics.Versatility: Logistic regression and decision trees were versatile, handling both voter turnout and party affiliation predictions.
- Linear regression was limited to continuous outcome variables, while GMM provided a unique clustering perspective.Insights: Each model contributed unique insights:Linear regression: Identified key demographic factors influencing turnout.
- Logistic regression: Revealed occupation-specific party affiliations.
- GMM: Uncovered distinct voter clusters, especially among high-income voters.
- Decision trees: Pinpointed specific voter groups for targeted turnout efforts

# CONCLUSION

- Models has pro and cons.
- Some areas where models did well.
- GMM, find rural high earners almost split and urban high earners lean democrat, but need more compute power to check optimal clusters.
- Logistic regression perform well in predicting voter turn out but not as good predicting party.
- Decision tree predict well in increasing voter turnout by at least 5%.
- Linear regression perform well in some areas and bad at others as well.