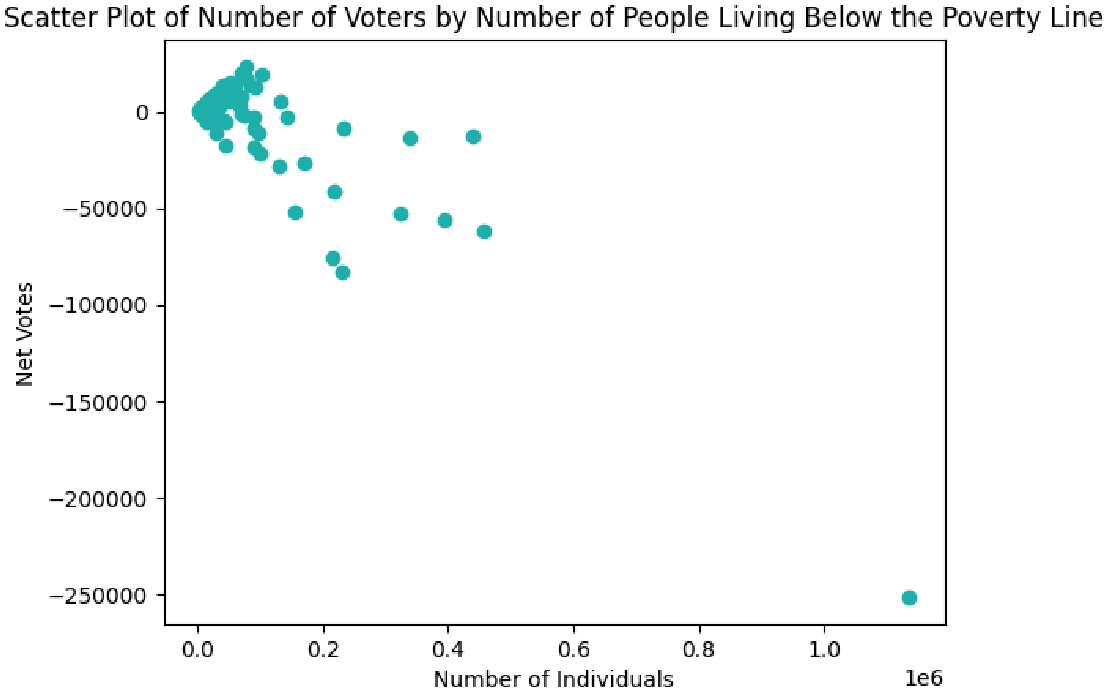
Summary:

The goal of the project is to build a predictive model that can predict the outcome of the 2024 presidential election in Virginia. With the model, the goal of the model is to accurately predict the net difference between the Republicans and Democrats in each county (Republican Votes - Democrat Votes). The county data, generated from the US Census, explores multiple variables such as income, education, poverty, transportation time, etc.. Of these variables, sex, education, median income, and poverty status by county were explored. These variables were merged with the net vote count using each county zip code and then analyzed to determine the relationships between each other. The scatterplots indicated that all of the variables appear to have relationships with net votes except for poverty status, indicating that the predictor should not be used to fit a model. After creating multiple models including a multiple linear regression model, a decision tree, and a k-nearest neighbor regression, I found that the KNN regression model was the best model out of the 3 based on metrics such as the R^2 value, SSE, and RMSE. The KNN model had a high accuracy as the predicted values of the testing data set and the actual values of the testing data set were quite similar. Since the values were similar, the model does a good job of predicting the election results given the data.

Data:

Most of the data cleaning was done in the df_melt.csv file. In the file, each column refers to a specific question and the associated levels of that variable. For instance, I decided to look at the number of individuals in each county by sex. The sex variable was split into two categories (AV1AA, and AV1AB). AV1 refers to a specific question, while the latter part of the column name (AA or AB) refers to the levels the question can take. AA refers to individuals who are male while AB refers to individuals who are female. Similarly, the education variable was split into 3 columns (B69AA, B69AB, and B69AC) referring to three education levels: no high school education, some high school or college education, or those with a bachelor's degree or higher. The household median income (B79AA) was recorded for each county; this variable has no levels to it. Similarly, the number of individuals in poverty for each county was stored in the column (AX6AA). These variables were initially selected for analysis because they seem to be related to voting trends. Sex, education, income, and poverty are key markers that help determine what issues are important to an individual and thus will likely determine how an individual votes. Net vote counts were determined by grouping by each county and subtracting the total votes for the Democrats from the total votes for the Republicans (Total Republican – Total Democrat). A negative net vote indicates that the Democrats had more votes in that county compared to the Republicans, while a positive net vote indicates that Republicans had more votes than the Democrats. Since some counties have a high population count, the net votes were transformed using the inverse sine function in order to make the data "more manageable" for analysis and model building. The net votes and transformed net votes were merged with the county data using the zip codes and then used for analysis. Initially, there were some missing values in each column of the county data; however, when the two data frames were merged via an inner join, these missing values disappeared and thus there was no cleaning required. The hardest challenge of this data aspect was understanding what the data represented. Since the column names were so vague, the code-book was quite helpful in determining how each column was related to the questions asked in the Census. Looking at the exploratory data analysis, Figures 1, 2, and 3 all indicate that Sex, Median Income, and Education Level all have a relationship with the net votes. For instance, in figures 1 and 2, there is a clear difference in the number of voters by each sex and by each education level for the net votes in each county in Virginia. Similarly, there appears to be a non-linear trend associated with net votes and income (it appears that as the income level increased, the more likely that voters of a higher income level were to vote for the Democrats). However, there appears to be no strong relationship between the number of individuals living in poverty and net votes as it appears that most of the data points are clustered near the top right corner.


Scatter Plot of Number of Voters by Number of People Living Below the Poverty Line

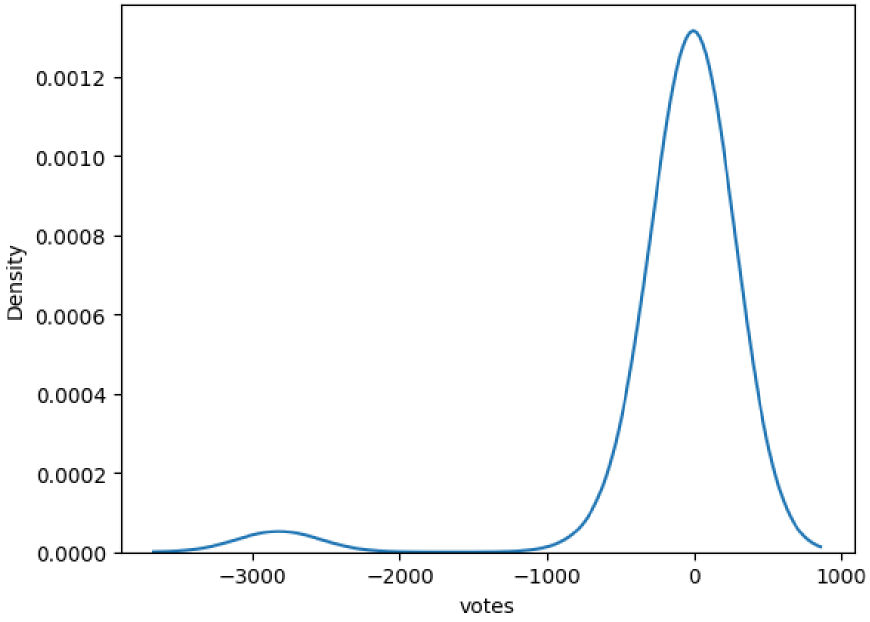
With this in mind, I decided to proceed with the model building process only using the variables Sex, Education Level, and Income.

Results:

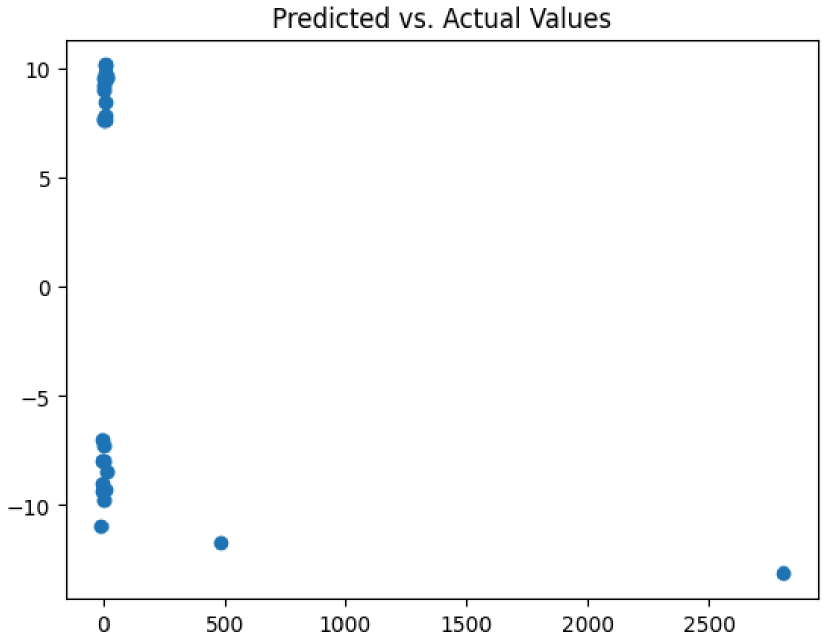
In order to effectively predict, I first decided to build multiple models and then test them using various criteria. I first started off with a multiple linear regression model. Using the three variables I identified, the model I got was:

$$\hat{y} =- 0.29893 + 0.000067(AV1AA) - 0.00105(AV1AB) + 0.001403(B69AA) + 0.000765(B69AB) + 0.000496(B69AC) + 0.00008(B79AA) .$$

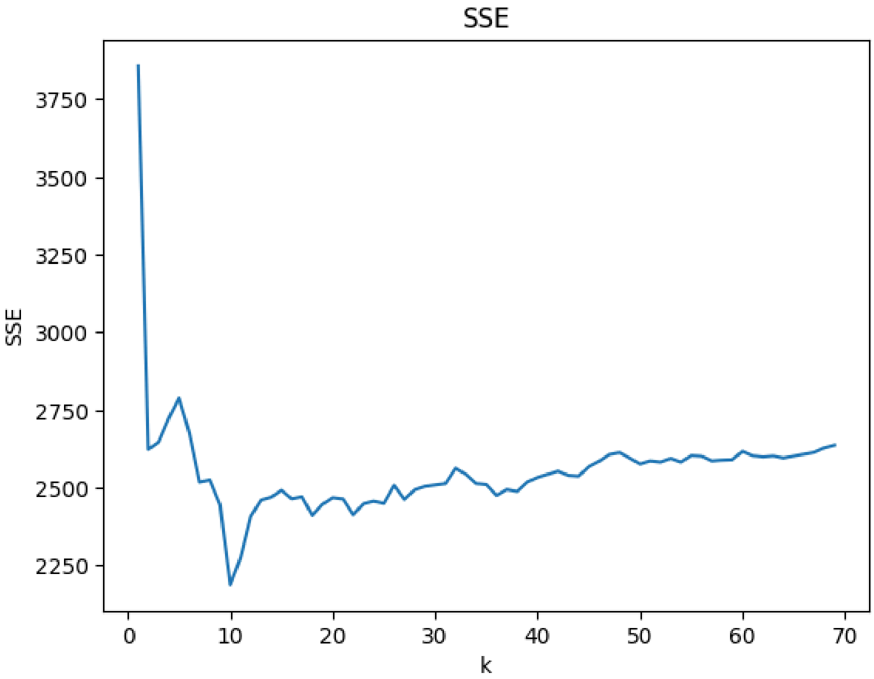
This model meets assumptions because the residuals appear to be normally distributed since the sample size is large enough and the assumption is robust.



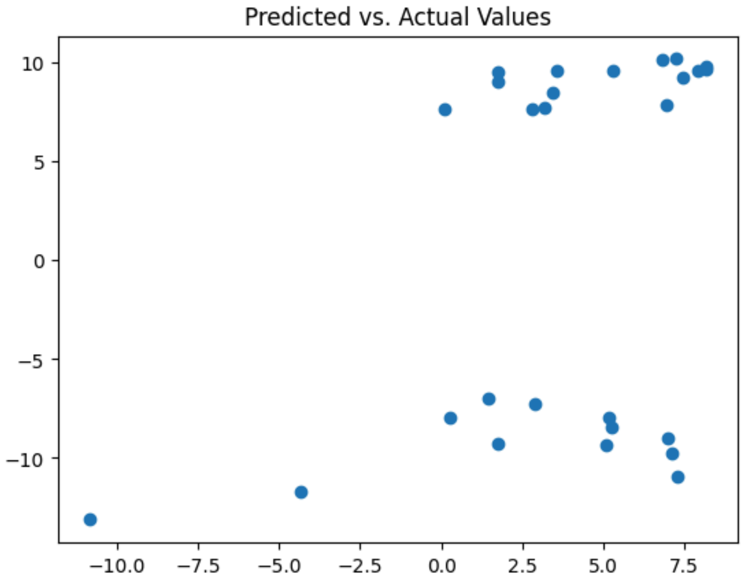
The model does not do an effective job at predicting because the graph of predicted versus actual values does not show a linear relationship. If the model was effective, then the predicted values should equal the actual values, meaning we should see a linear line with the slope of 1. However, in the graph shown down below, it appears that the model does not do an effective job because there is no slope of 1.



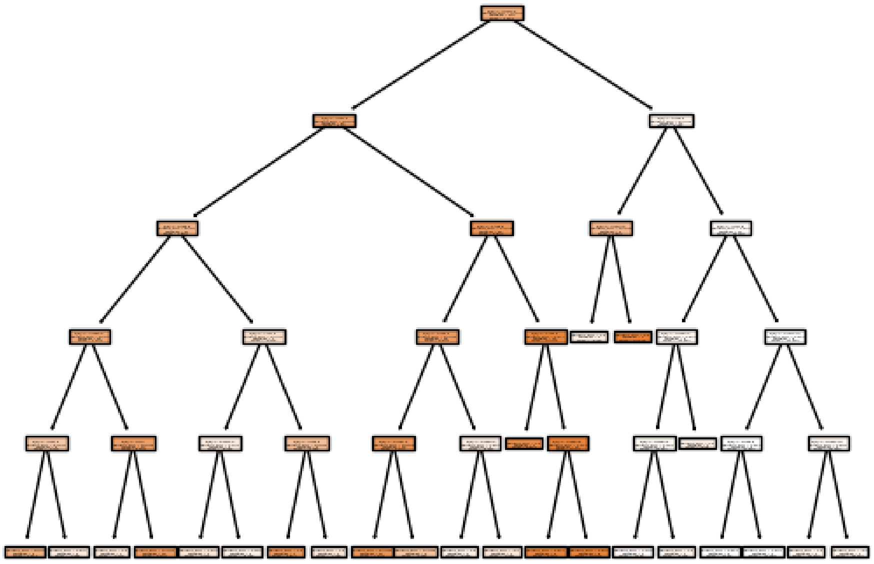
After the linear regression model, I decided to build a KNN model. I first looked at the SSE in order to determine which k value would be the best to use to build the model. Looking at the SSE plot, a k-value of 10 minimizes the SSE the greatest, indicating that k-star should be set to 10. After the k-star value was determined, the KNN regression model was built to fit the most optimal model.



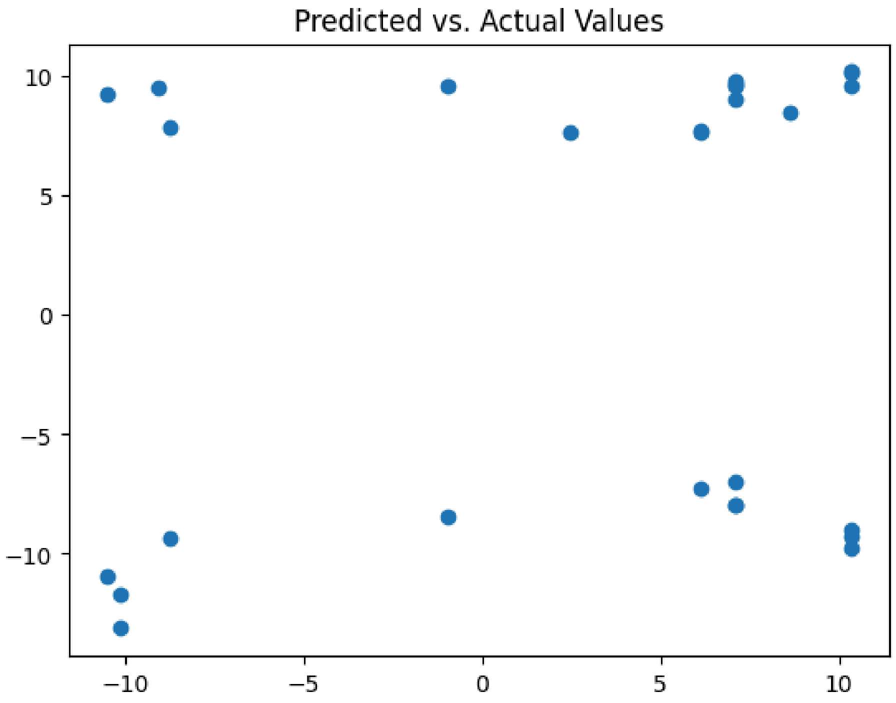
The predicted values vs actual values were then graphed. Looking at the plot, this model appears to be doing better in terms of prediction as high values of the inverse sine of net votes also have high predicted values. However, there are instances in which the model under-predicts (ie: meaning the model predicts that Republicans won in a county, when in real life the Democrats won) because there is a cluster of points to the bottom right of the plot, indicating that those predictions are incorrect.



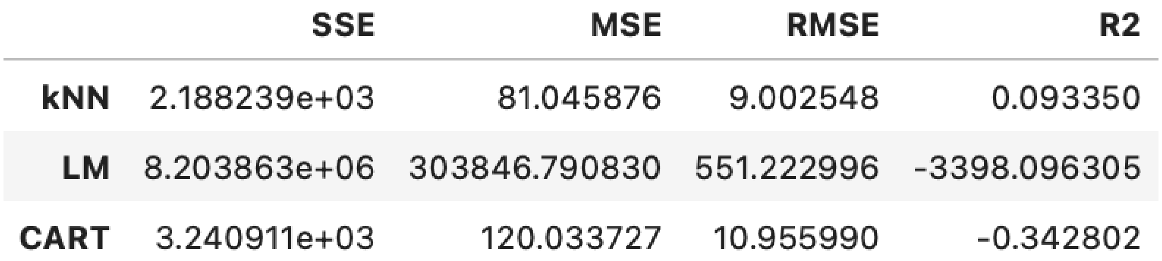
Finally, a decision tree model was built to compare with the other 2 models.



The decision tree model appears to have normally distributed residuals, indicating that the normality assumption is met for the model. The predictions appear to be better than the linear regression model, but appear to do worse than the KNN model. The predicted versus actual values plot indicates that the decision tree model sometimes accurately predicts the outcome or inverse sine of net votes. However, there are points that are clustered at the top left and bottom right (predictions that deviate from the actual values) that indicate that the model is not the best for prediction, but might still be useful.

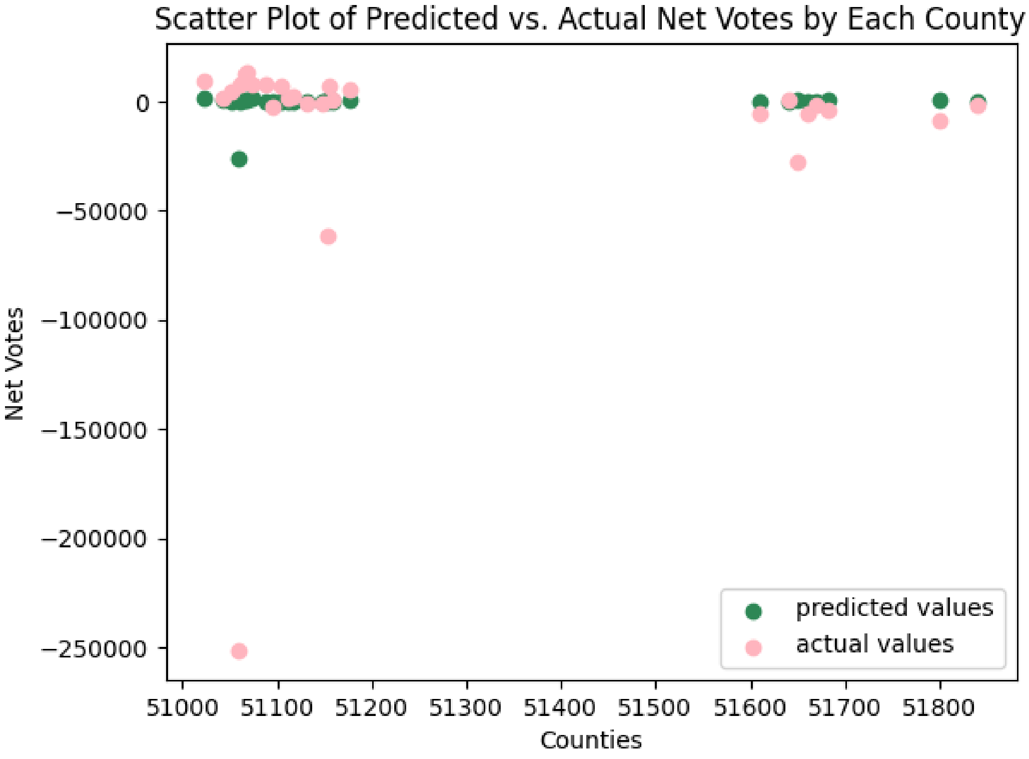


In order to compare all 3 models, I decided to look at the SSE, MSE, RMSE, and R^2 of all 3 models to determine which model is the most effective at predicting.



	SSE	MSE	RMSE	R2
kNN	2.188239e+03	81.045876	9.002548	0.093350
LM	8.203863e+06	303846.790830	551.222996	-3398.096305
CART	3.240911e+03	120.033727	10.955990	-0.342802

Of the 3 models, it appears that the KNN model is the best followed by the decision tree model. The SSE, MSE, and RMSE should all be minimized, while the R^2 value should be closer to 1 for an effective model. The KNN model has the smallest SSE, MSE, and RMSE compared to the other two models, while also having the highest R^2 values at 0.09 (not a large R^2 value, but higher than the other models). Given these criteria, the KNN model will be the best model to use for predicting the 2024 election. Finally, looking at the scatterplot of predicted and actual net votes by each county, it appears that the KNN model does accurately predict the net votes because the predicted and actual values are clustered together for the various counties in the testing dataset. Given the data for 2023, we can make predictions regarding the outcome of the 2024 election in Virginia by totalling the net-votes for each county in Virginia. If the total of the net-votes are positive, then the Republicans won in Virginia, but if they are negative, then Democrats won in Virginia



Conclusion:

In all, a KNN model appears to be the best model for prediction. This was determined by first calculating the net vote count (taking the total votes for Democrats and subtracting the total votes for Republicans). The net vote was transformed by taking the inverse sine of it in order to make the data more manageable. A couple of variables including sex, education, poverty status, and income were then plotted against net-vote count to see whether or not these variables had relationships with the response variable. Of the 4 variables, only poverty status appeared to have a weak relationship with net-vote count and thus was not used to build the models. After the exploratory data analysis, multiple models were built and assessed to determine which one has the highest accuracy for prediction. A multiple linear regression model, a KNN model, and a decision tree model were the 3 models considered and the KNN model appeared to be the best model in which criteria such as SSE, MSE, and RMSE were minimized, while the R^2 value was maximized (indicating it is the model that performs the best). To assess the model's accuracy, the predicted net votes and the actual net votes were plotted against each other. Since the two values were clustered together, the model appears to be an effective model at predicting the outcome of the election. Although the models are very rudimentary and do not have high R^2 values, the high accuracy values indicate that the models will be effective at predicting for the most part. Furthermore, even though it appeared that some of the variables appeared to have a curve-linear relationship with the response variable, these variables were not transformed. The model might have a greater R^2 value if these explanatory variables were transformed, indicating that the model will not always be accurate. In the future, it is important to consider other predictors. There were many predictors that were found in the ts county data sets. While most of them were not considered, it should be looked at more closely as there is a high chance that some predictors were missed since they were not looked at in the beginning. Furthermore, it would be better to mess around with models and see if any transformations are required. Even though there were no outliers present in the response and explanatory variables, since the values of the explanatory variables are so large (in the thousands or more), it might be important to evaluate whether these variables should be transformed as well.

Appendix:

Figure 1: Scatterplot of Voters by Sex

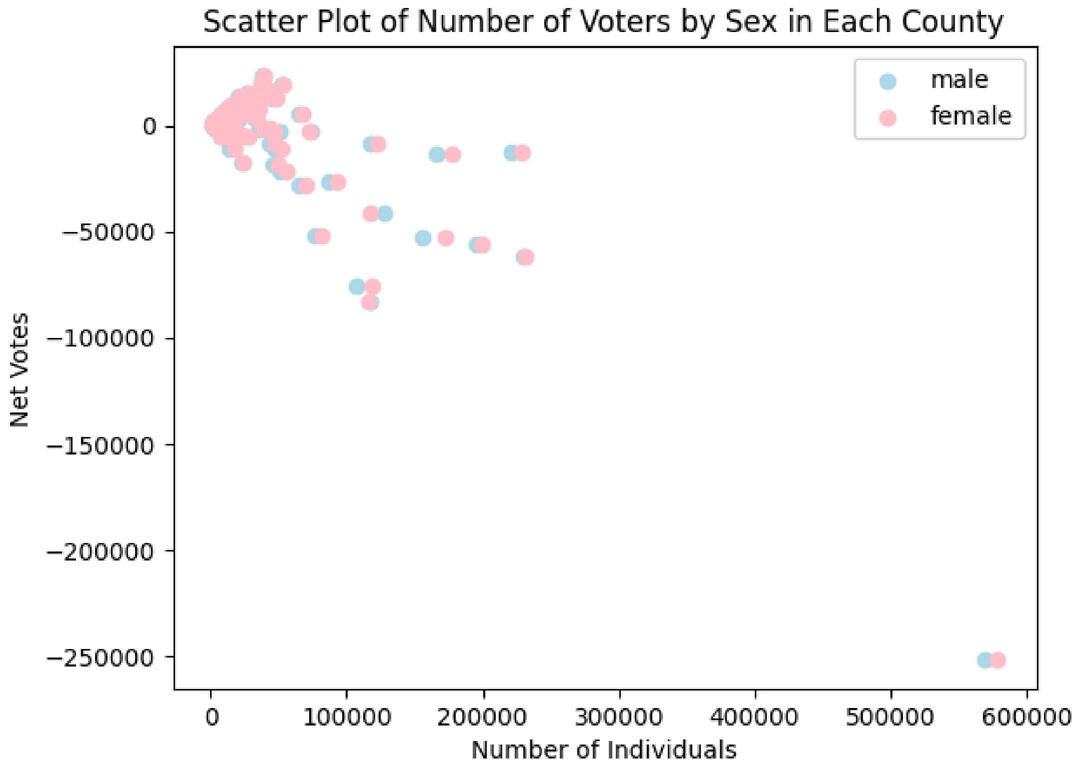


Figure 2: Scatterplot of Voters by Education Level

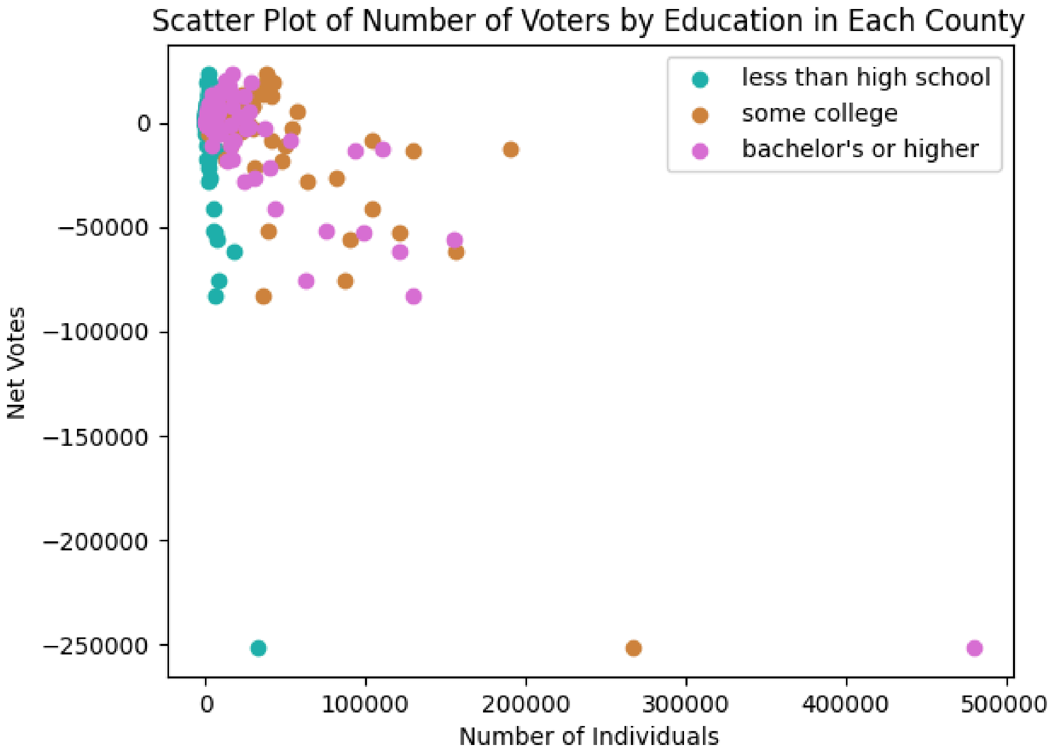
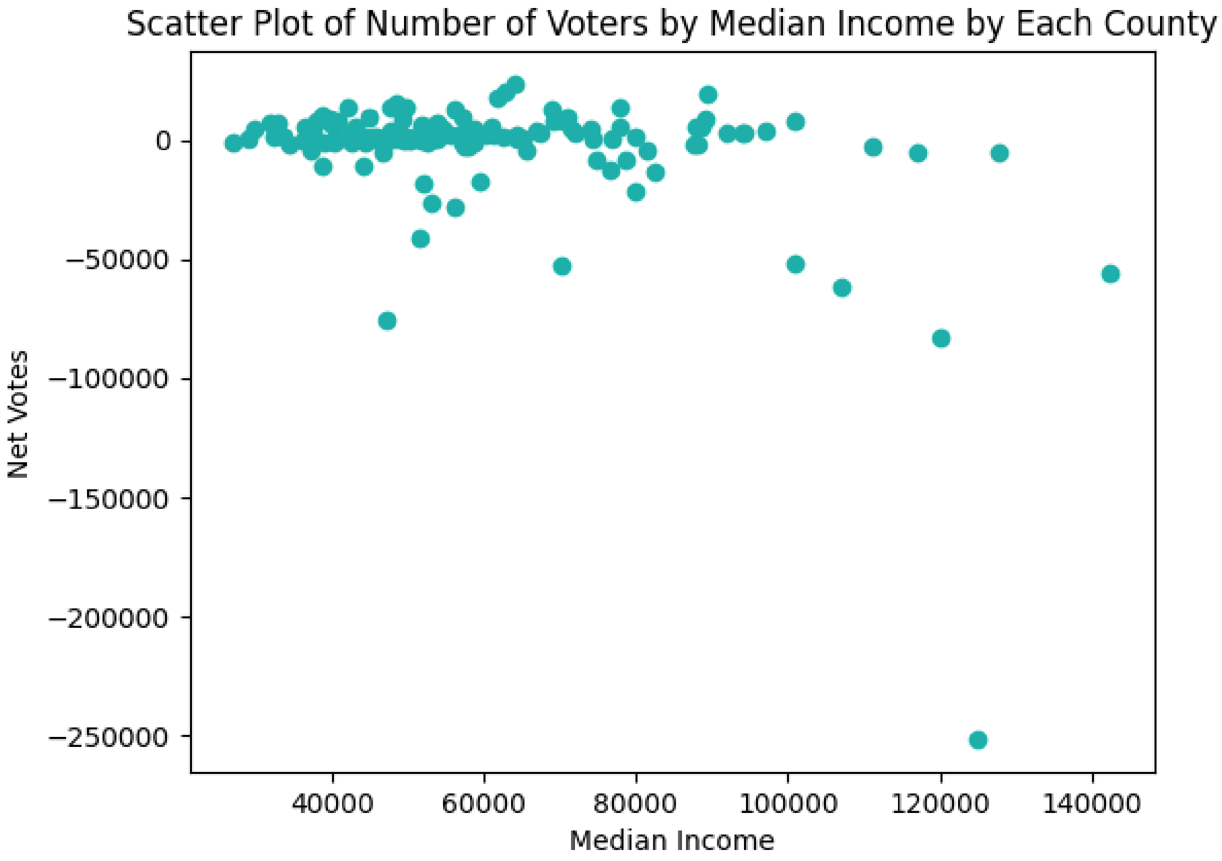


Figure 3: Scatter plot of Voters by Median Income


Scatter Plot of Number of Voters by Median Income by Each County
Net Votes
Median Income