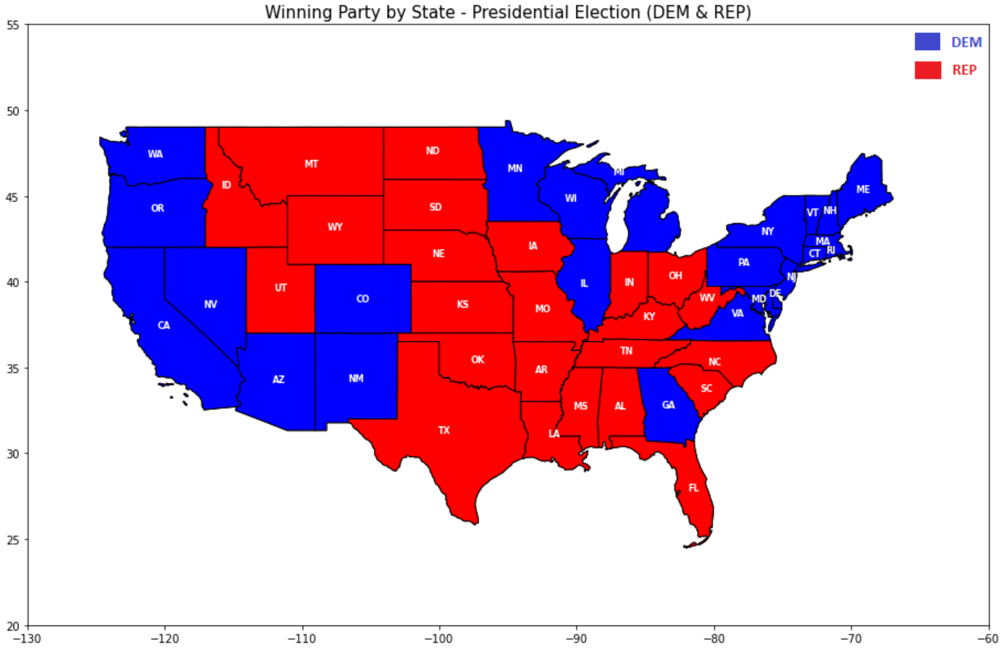
US Election 2020 Analysis

Term Project



Objective

Analyze the publicly available *voting results and demographic statistics* at the county level to try and identify what kind of demographic categories could be used to infer the voter turnout in the 2020 US presidential election.

Goal:

Following the 2016 US presidential elections, there were some indications that voter turnout decreased for certain demographics of the US population and that many of those same demographics voted predominately in favour of the Democratic Party. The argument followed that if those communities had voted at the same rate as the general population, the outcome of the election would have turned in favour of the Democrats.

With the 2020 US elections still fresh in our memories, we chose the available voting results and various demographic categories that could be used to infer the voter turnout in the 2020 US presidential election. Additionally, we analyzed the data to determine to what extent these demographic categories could account for the variations in voter turnout in the election. These questions will be investigated using single and multiple linear regression models.

Besides linear regression models, logistic regression models were also performed to predict president wins using median income and ethnic groups. In addition, a mapping method, an election result versus polling prediction and an ANOVA hypothesis testing and various visualizations were completed to provide insights into the datasets.

Questions to answer:

- *What kind of trends can be seen in governor and presidential elections?*
- *What kind of county demographics categories could be used to infer the voter turnout?*
- *What are the election results versus polling predictions?*
- *How can we predict electoral wins based on median income?*
- *How can we estimate the presidential winning party based on ethnic groups?*

Data Preparation

Our data is a combination of multiple source datasets that were downloaded in CSV format and ingested into Python along with various listings of municipalities by county, in which case the data was web scraped from the various websites by copying and pasting into a CSV spreadsheet.

Data sources:

The following table summarizes these datasets.

Name	Source	Description
US Election 2020 Presidential Votes by County	Kaggle, Raphael Fontes	A table listing each candidate in each county or municipality in the USA, the number of votes they received, and whether the candidate won in that riding.
2019 County Estimates by Age, Sex, Race, and Hispanic Origin	US Census Bureau	An estimate of demographic data for each county in the United States, including Age, Sex, Race, and Hispanic origin

Name	Source	Description
2019 Population estimates for US counties	US Dep't of Agriculture	Urban vs. rural indicators and population and population change (migration) details for each county in the US.
ERS 2015 County Typology Codes	US Dep't of Agriculture	A classification of all US counties according to six mutually exclusive categories of economic dependence and six overlapping social categories include low education, low employment, persistent poverty, persistent child poverty, population loss, and retirement destination.
Listings of municipalities by county	Multiple (see notebook)	Mappings indicating what county various municipalities in the states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.
Presidential Polls 2020	FiveThirtyEight	The dataset contains an entry for each poll conducted in the United States for the 2020 election
Income and employment by State and County	U.S. Dep't Agriculture	"Unemployment and median household income for the U.S., States, and counties, 2000-19"
Maps Data Set (USA - States and Counties)	Cartographic Boundary Files	Cartographic Boundary Files (.shp files) were used to create a dataset to plot the map of the USA including its state and county boundaries.".shp" files contain this information with the "geometry" column that shows us its respective spatial data. Also, latitude and longitude of various states and counties were obtained from the uscities.csv file.
2020 Polling dataset	FiveThirtyEight	Presidential_polls_2020 file contains the data of polling answers of each state from various pollsters with the date and sample size.

Data preparation process:

A number of preparatory steps were taken to clean and prepare the final voting data set for analysis by combining county data sets using state name and county name. To perform the join of various data sets, it was desirable to have each data set contain only one row per county, however the voting and county estimate data sets both contained multiple rows per county. The voting data set broke down each combination of candidate and county in separate rows and the county estimate data set had separate entries for each combination of age group and county. Multiple rows for the candidates and age groups were pivoted into columns of interest in order to arrive at a single row per county in these two data sets.

Another significant action that had to be taken before joining the data sets was required as the voting data for the states of *Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island and Vermont* listed each municipality instead of the county, where multiple municipalities belonged to a single county. To adjust this, a mapping between municipalities and counties in each of those states were created and multiple municipalities in the voting data were combined to create county-level data. For states such as *Alaska and the District of Columbia*, it was not evident if the same mapping could be performed so we opted to remove those states from the analysis.

Geopandas was used in our analysis to plot the state and county map of the United States of America. This package combines geometry objects of shapely. Geometries are stored in a column called geometry that is a default column name for storing geometric information in geopandas. This allows us to easily manipulate, plot and analyze data.

The last step in terms of data preparation was to choose which input columns were of interest to the analysis. In some cases, this involved excluding parameters as they were not featured as a part of the questions being asked and in other cases, this involved performing calculations on the columns in order to allow for the desired analysis. Examples of these calculations were the combination of eighteen age groups into four, the conversion of counts (votes, populations, etc) into proportions (percentages). Minor actions were also taken for aligning capitalization and spelling of counties to ensure that the join of various tables gave accurate results. The resultant data set had 3111 US counties and the corresponding voting and county characteristics columns.

Data Analysis

1. Data Visualization for Governor and Presidential Elections

In this section, governor and presidential datasets were compared to see how both elections differed in terms of vote count and how race affected the Democratic (DEM) and Republican (REP) party choice.

When vote count is taken into consideration, Republican Party votes in states like Indiana, West Virginia and New Hampshire were almost 50% higher compared to democratic votes but in North Carolina and Washington, the Democratic Party won the governor election. For the presidential election, Suffolk County had the highest number of REP votes (381,021) as opposed to Los Angeles County which had about 3,028,885 DEM votes.

For presidential elections, winning counties for both DEM and REP parties had a higher ratio of White Americans compared to any other race. Also, we can see that counties that had a higher population of Hispanics voted for the REP party. When we review governor election results, winning counties for both DEM and REP parties also had White Americans who topped the race distribution and clearly counties with higher population of Black Americans chose the DEM party over the REP party.

An interesting observation from the visualizations (*Figures – 1.1 & 1.2*) is that when we look at county level results for both parties, we see that more number of counties voted for the REP party. Even though the number of counties were higher, the highly populated counties voted for the DEM party and hence, the winning party in presidential and governor elections.

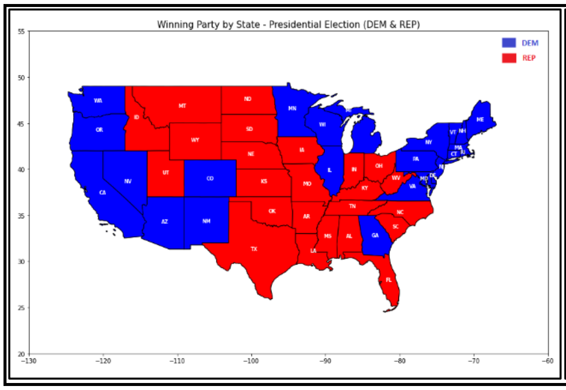


Figure – 1.1

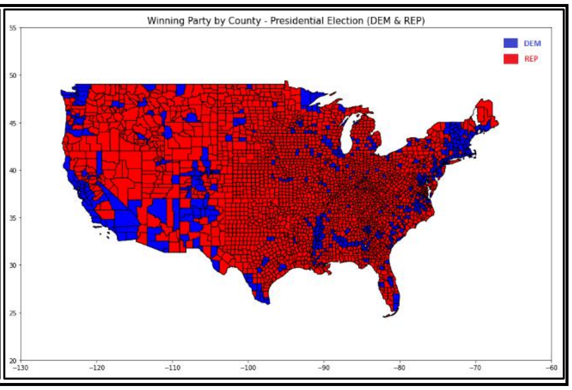
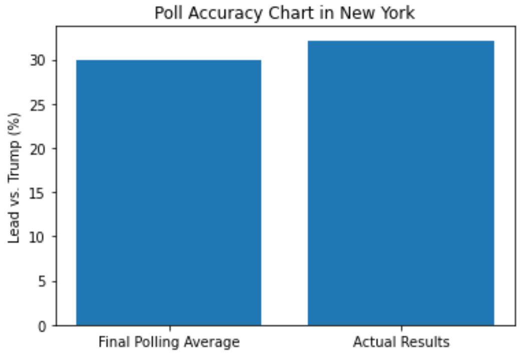
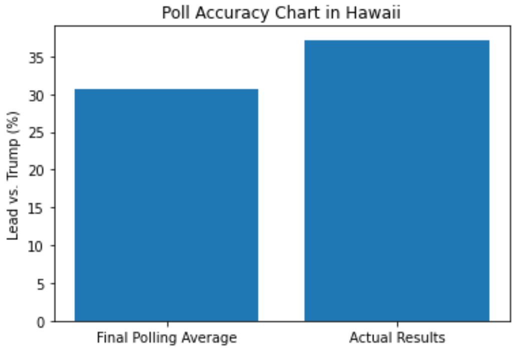


Figure – 1.2

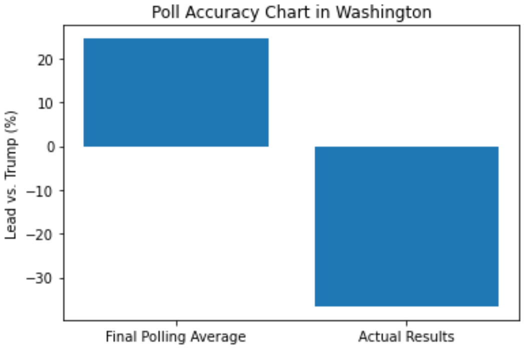
2. Election Result vs Polling Prediction

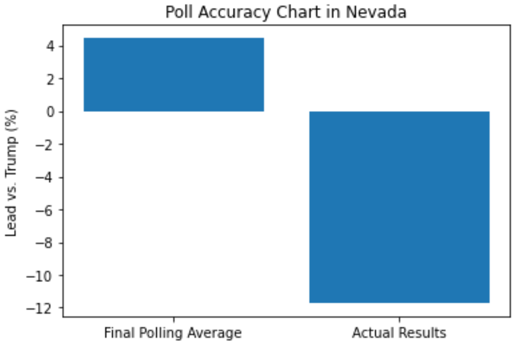
In this analysis, we compared the pollster's polling results with the 2020 actual election results to find out how accurate the polling result matched the final one. To calculate the polling accuracy, we counted the percentage average of Republican and Democratic Party winning chances in each state from the polls dataset, measured each county's votes, decided the two candidates winning cases and compared the candidates' results.



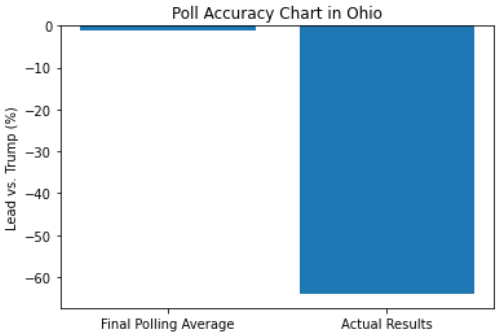


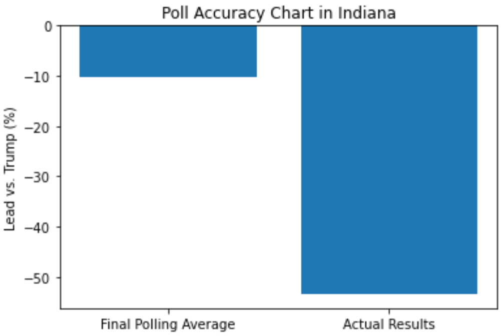
The above plots describe the accuracy of the polls where Biden (DEM party) won both New York and Hawaii states. The percentage point of REP party (Trump) underestimation in New York and Hawaii is 2.14% and 6.44% respectively. The US presidential election is elected by the Electoral College and not the national popular votes so it would be helpful to consider the swing states.





Trump's (REP party) chances in Washington were underestimated by huge margins despite Biden (DEM party) having primarily positive outcomes. Nevada is yet another state that polls missed to predict accurately. The percentage point of Trump (REP party) underestimation in Washington and Nevada is 61.11% and 16.15% respectively.





Ohio and Indian's submission samples were not enough to predict accurate data. However, the percentage point of Trump (REP party) underestimation in Ohio and Indiana is 63% and 44.36%.

From the polling results of all 50 states and taking the average winning percentage of pollsters' election candidates, we found that the 2020 polls' result was accurate for eight states.

3. Correlation and Visualization based on Race Distribution, Age Groups and Gender

In this section, the Republican Party won 24 states and Democratic Party won 25 states. Data is divided into two groups namely, republican states and democrat states. We will look into how race, age, and gender correlated with the final election results.

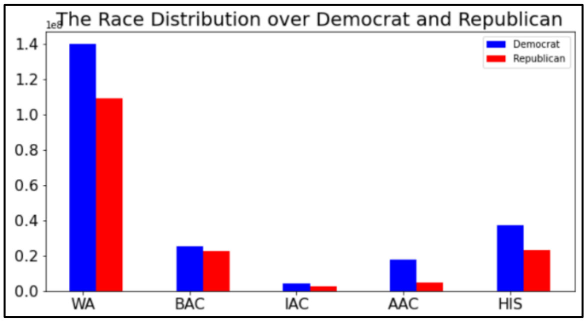


Figure 3.1 Race Distribution

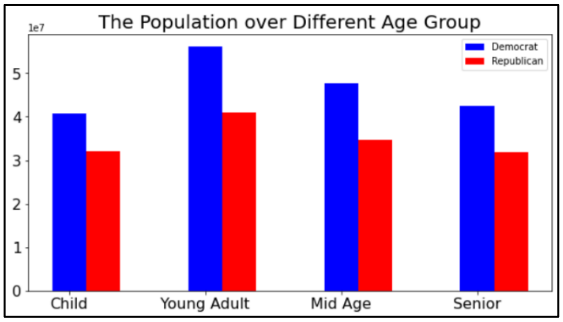


Figure 3.2 Different Age Groups

As we can see from *Figure 3.1*, the democrats won the elections having more White American, Asian American, and Hispanic populations in those states. *Figure 3.2* shows that the largest population in the United State is that of young adults. There is not much difference between these different age groups.

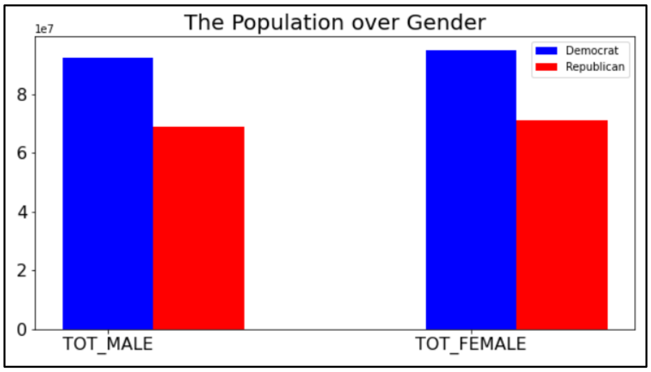


Figure 3.3 Gender Distribution

In terms of gender, there doesn't seem to be much difference between male and female population of different states that choose democrat and Republican Party.

	Low_education	Low_employment	poverty
DEM	87	131	118
REP	379	770	233

The above table compares the number of counties having low education, low employment and poverty levels between democrat and republic states. The states that republicans won had more people with low education and low employment. So, poverty rate is much higher in the counties of those states as well. Besides looking at the demographic data, a hypothesis test was applied as well.

Null Hypothesis : The median of the population for each age group from each county in New Jersey is the same across all groups. Only age groups above 18 is considered. The three age groups are young adult, mid age & senior.

Alternative Hypothesis : At least one median is different.

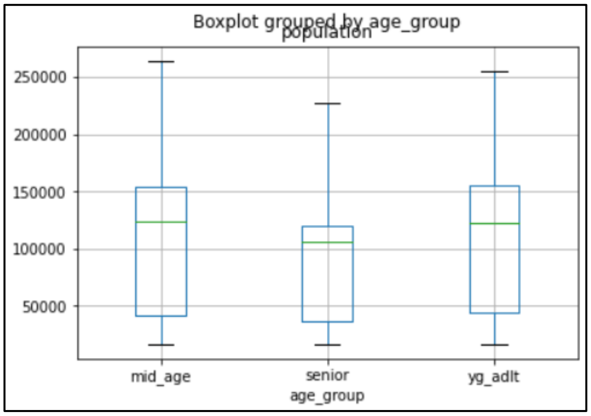
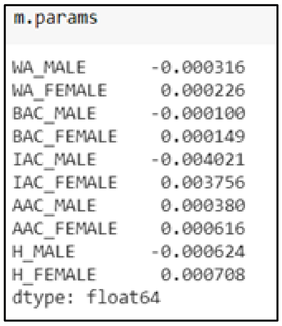


Figure 3.4

As shown in *Figure 3.4*, data was plotted into a boxplot to make sure that data is not skewed. As the data wasn't normally distributed, ANOVA test didn't seem feasible in this case. Hence, **Kruskal-Wallis H-test** was applied and the p-value is 0.664. Therefore, we failed to reject the null hypothesis and we accept that the population median of all of the groups are equal.

4. Logistic Regression: Estimate President Win Based On Ethnic Groups

A logistic regression model was fitted using ten ethnic-gender groups including White Male, While Female, Black Male, Black Female, Native Male, Native Female, Asian Male, Asian Female, Hispanic Male, and Hispanic Female. In each county, the percentage of each group is known as well as the winning candidate. The model calculated resulting regression parameters as shown below in *Figure 4.1*:

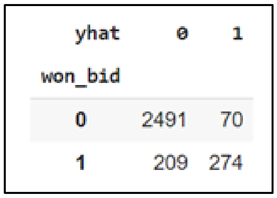


```
m.params

WA_MALE      -0.000316
WA_FEMALE     0.000226
BAC_MALE     -0.000100
BAC_FEMALE    0.000149
IAC_MALE     -0.004021
IAC_FEMALE    0.003756
AAC_MALE      0.000380
AAC_FEMALE    0.000616
H_MALE       -0.000624
H_FEMALE      0.000708
dtype: float64
```

Figure 4.1

From the above list, we can see that these parameters have very low correlation to the voting results in each county. The confusion matrix (*Figure 4.2*) was calculated and it showed that Biden (DEM party) would lose in 2491 counties and win in 274 counties. The remaining 209 + 70 estimates were incorrect predictions.



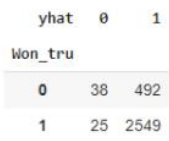
yhat	0	1
won_bid		
0	2491	70
1	209	274

Figure 4.2

As we all know, Biden (DEM party) won in the presidential election last year. However, the logistic model based on the five ethnic groups as mentioned earlier predicted otherwise (DEM party would not win). So, we can conclude that the predictors in this model (five groups with genders) were not enough for predicting election results. The reason could be due the fact that election results are determined based on Electoral College votes and not the popular national votes. This data set is for popular national votes, so it is not enough to determine the results.

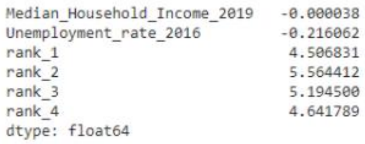
5. Logistic Regression: Predicting Trump (REP party) wins using Median Income

In this section, logistic regression was applied to income and employment variables by grouping them into quantile buckets in our data set to determine whether they can be used as reliable predictors of a win or loss for Trump (REP party) in each county.



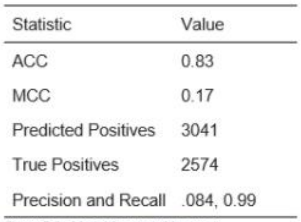
yhat	0	1
Won_tru		
0	38	492
1	25	2549

Figure 5.1- Confusion Matrix



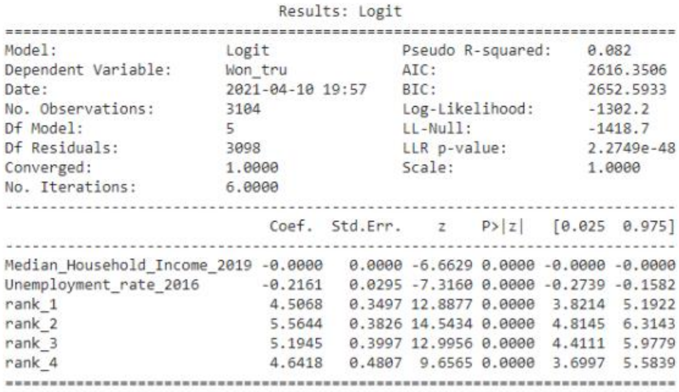
```
Median_Household_Income_2019    -0.000038
Unemployment_rate_2016          -0.216062
rank_1                           4.506831
rank_2                           5.564412
rank_3                           5.194500
rank_4                           4.641789
dtype: float64
```

Figure 5.2 – Regression Parameters



Statistic	Value
ACC	0.83
MCC	0.17
Predicted Positives	3041
True Positives	2574
Precision and Recall	.084, 0.99

Figure 5.3 - Model Accuracy Measures



```
                              Results: Logit
============================================================================
Model:                   Logit              Pseudo R-squared:    0.082
Dependent Variable:      Won_tru            AIC:                 2616.3506
Date:                    2021-04-10 19:57   BIC:                 2652.5933
No. Observations:        3104               Log-Likelihood:      -1302.2
Df Model:                5                  LL-Null:             -1418.7
Df Residuals:            3098               LLR p-value:         2.2749e-48
Converged:               1.0000             Scale:               1.0000
No. Iterations:          6.0000
----------------------------------------------------------------------------
                              Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
----------------------------------------------------------------------------
Median_Household_Income_2019 -0.0000   0.0000 -6.6629 0.0000 -0.0000 -0.0000
Unemployment_rate_2016       -0.2161   0.0295 -7.3160 0.0000 -0.2739 -0.1582
rank_1                        4.5068   0.3497 12.8877 0.0000  3.8214  5.1922
rank_2                        5.5644   0.3826 14.5434 0.0000  4.8145  6.3143
rank_3                        5.1945   0.3997 12.9956 0.0000  4.4111  5.9779
rank_4                        4.6418   0.4807  9.6565 0.0000  3.6997  5.5839
============================================================================
```

Figure 5 4 - Model Statistics

While the model showed an accuracy level of 0.83 (*Figure 5.3*) and a clear ability to predict wins, the model predicted more wins than actual wins when based on median income. This is seen by comparing the ACC value of 0.83 and the MCC value of 0.17 (*Figure 5.3*) which indicates that this model is more accurate than chance but the lower MCC value indicates that the model isn't very good at discriminating between wins and losses due to a weak correlation between the predicted and true class. This is further supported by high precision and high recall (*Figure 5.3*) which indicates that the model is good at finding wins but not good at discriminating a win from a loss. Another consideration to think of is that the data provided is imbalanced. Population density might be a good variable to include as a predictor.

6. Linear Regression: Voter Turnout

In order to identify what county demographics categories could be used to infer the voter turnout in the counties across the US, we first measured the correlation between the target variable and 27 input variables and eliminated those features that had a Pearson correlation of less than approximately 0.2. A simple (single input feature) linear regression model using the *Ordinary Least Squares algorithm* was created for each of the remaining inputs and *Figure 6.1* shows the R2 value for the top six results. The strongest results were age related, with R^2 values of 0.23 and 0.22. Of the variables listed in *Figure 6.1*, only Percentage of Seniors was positively correlated with the percentage of voter turnout. All of the other 5 variables listed namely, Percentage of Young Adults, Low Education Flag, Percentage of Hispanics, Child Poverty Flag, and the Low Employment Flag were all negatively correlated with the output. In other words, they had negative coefficients with the rate of voter turnout.

Notably absent from this top tier of predictors are the input variables related to the gender ratio in the county, the rural/urban (and metro) status of the county as well as the economic drivers of the county (farming, mining, manufacturing, government, recreation, and non-specialized).

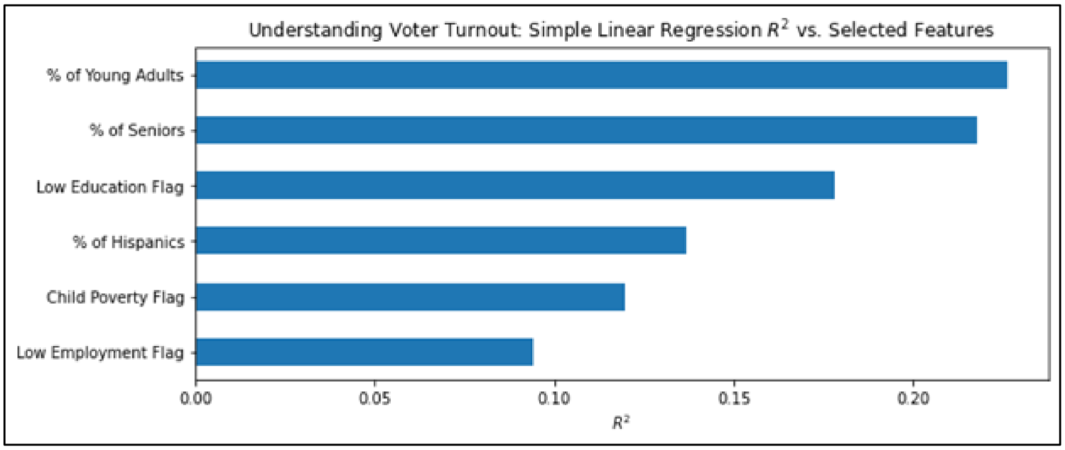
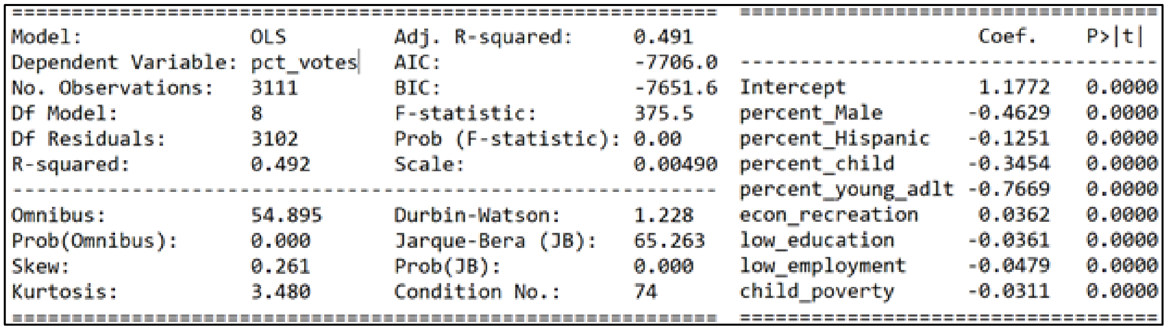


Figure 6.1

The final component of this analysis was the creation of a multiple linear regression model. To do so, the cross correlation of the remaining 14 features was then assessed, and if any pair of features had a correlation greater than 0.5, one was eliminated to reduce the collinearity of the model. Through this step, the percentage of females in a county, the percentage of White Americans, the percentage of seniors, and the rate of poverty in each county were all eliminated because they had high correlation with other variables. From the resulting linear regression model, the Government Economic influence flag had to be removed from the model as it was not statistically significant, and the percentage of black Americans was also removed as its coefficient had switched from being negative to positive.

The following table summarizes the resulting model statistics, including the adjusted R^2 of 0.49. The skew and kurtosis indicate that the residuals are indicative of a normal distribution. The Durbin-Watson statistic shows that the residuals are not independent but are within acceptable levels. The analysis in the Jupyter notebook indicates that each variable is linearly related to the target variable, and that a single value exceeds normal influence measures but it was assessed that it should not be removed from the analysis.



```
============================================================ ==================================
Model:              OLS         Adj. R-squared:      0.491                              Coef.    P>|t|
Dependent Variable: pct_votes   AIC:                 -7706.0 ----------------------------------
No. Observations:   3111        BIC:                 -7651.6 Intercept           1.1772  0.0000
Df Model:           8           F-statistic:         375.5   percent_Male       -0.4629  0.0000
Df Residuals:       3102        Prob (F-statistic):  0.00    percent_Hispanic   -0.1251  0.0000
R-squared:          0.492       Scale:               0.00490 percent_child      -0.3454  0.0000
------------------------------------------------------------ percent_young_adlt -0.7669  0.0000
Omnibus:            54.895      Durbin-Watson:       1.228   econ_recreation     0.0362  0.0000
Prob(Omnibus):      0.000       Jarque-Bera (JB):    65.263  low_education      -0.0361  0.0000
Skew:               0.261       Prob(JB):            0.000   low_employment     -0.0479  0.0000
Kurtosis:           3.480       Condition No.:       74      child_poverty      -0.0311  0.0000
============================================================ ==================================
```

Conclusion

The linear regression analysis of voter turnout of various metrics (such as age, ethnic/language, social, and economic factors) that were analyzed indicated that age was the strongest indicator followed by social factors like low education, child poverty, and low employment. Counties with a higher percentage of Hispanic people were also indicative of lower voter turnout. Independently, these factors could account for approximately 10-20% of the variation in voter turnout between US counties and when combined with other statistically significant variables, it could account for almost half of the variation between counties.

The regression model based on median income predicted more wins than true wins. It was not very good at discriminating between wins and losses as the correlation was weak between the predicted and true class.

The logistic model based on ethnic groups appeared to be incorrect in predicting election results. It could be due to the lack of predictors and the fact that election results are determined based on Electoral College votes, not the popular national votes.

After looking at our polling comparison, the 2020 pre-election polling data can measure which candidate is prevalent in a particular county or state; however, if no candidate achieves an absolute majority, Elector College or other factors can significantly change the final election result.

Looking at demographic data, the democratic states have much more population than republican states. In the U.S., there is more population in young adults than mid-age adults and senior adults which means a young adult group is an important group in the election. Also, Republican states have a greater number of counties with low education, low employment, and poverty compared with democratic states.

Visualizations of county and state maps also shows us that even more number of counties voted for the Republican party, the highly populated counties voted for the Democratic party thereby increasing the number of votes per state for the Democratic party and hence, the winning party in the presidential and governor elections.