# **Final Project Write Up**

**Questions:** Does primary election turnout predict general election turnout?

How do certain characteristics such as # of eligible voters, # of registered voters, election type

(re-election vs new election) and # of voters by state have an impact on general elections?

**Hypotheses:** We believe the voter turnout of the general election will be impacted by these characteristics, but we believe primary turnout will have the most impact.

**Finding data**

**Data – Links:** Raw data was sourced from the United States Election project and the US Census Data website files

* United States Election project <http://www.electproject.org/home/voter-turnout/voter-turnout-data>
* US Census Data website files <https://www.census.gov/topics/public-sector/voting/data/tables.html>

1. 1980-2014 Turnout Rates,

2. 2000, 2004, 2008, 2012, 2016 Primary Election data,

3. 2000, 2004, 2008, 2012, 2016 General Election data

4. 1996-2016 Registered Voter data.

**Cleaning before python import**

* This data was restructured and standardized by importing it into XL.
* Any differences in format were corrected and any missing data was derived from rate of Population values and average rates. For example, if we were missing one year of turnout rate, we took the average of available data.
* We decided there was no need for overseas data for general elections, since we did not have primary data to compare to, nor did we have which states those votes should be applied to.
* Due to missing data for primary rates per state, we selected (5) yrs. 2000, 2004, 2008, 2012, 2016 to maximize usefulness (presidential election data on 4yr cycle) and to ensure a manageable amount of data.
* It was then passed to subsequent python processes.

**Cleaning after Python Import**

* Imported 2000-2016 General Election turnout rate and counted the number of rows to insure 255 rows.
* Had some spacing issues with key data, where spaces were present after words and making it difficult to refer to specific columns. For this reason, we removed any blank spaces in key words (i.e.. Rhode Island, New Hampshire, etc.).
* Renamed columns and filtered out unneeded columns
* Read in Primary Election Turnout data from CSV file and followed above cleaning process.
* Read in Registered Voter by State data from CSV file and followed above cleaning process.

**Jupyter Notebook: Voter\_Turnout\_JA**

* Imported 2000-2016 General Election turnout rate and counted the number of rows to insure 255 rows.
* We noticed we were having spacing issues with the data, where spaces were present after words and making it difficult to refer to specific columns. For this reason, we looked for a removed any spacing.
* We renamed columns and filtered out unneeded columns
* Read in Primary Election Turnout data from CSV file and followed above cleaning process.
* Read in Registered Voter by State data from CSV file and followed above cleaning process.
* We merged data with two step process. First, we merged General Election Turnout data with Registered Voter by State data using Left merge on “Year” and “State” because the Registered data had the most data, and General election had the 2nd highest amount of data. We check rows and columns to make sure we did not lose anything.
* Initially we lost some rows due to missing state values in Primary data that was missed during initial clean up process. We were missing Total Turnout Rate, Voter Eligible Population. We decided removing the data could affect our machine learning and decided to fill in blanks for state turn out based on average turnout of other primary years in the same state. We also filled in VEP numbers based on VEP from general election.
* Second and final merge was done with Primary Election Results data.
* We checked data but counting number or rows and columns and looked for null values.
* Once we saw all data was appearing as it should, we exported data frame into CSV file Voter Final Clean.

**Jupyter Notebook: ML Linear Regression**

* Import Voter Final Clean Data
* Imported from sklearn.preprocessing import LabelEncoder in order to label code states column
* We wanted to see if voter turnout was impacted by new election year vs re-election year, so we added a new column called “Re-election” that tracked with 1 (reelection year) and 0(not a reelection year) whether a specific election year was re-election or not.
* Further cleaning of data was required to fit data into ML model such as removing “%” and “,” symbols as well as converting data type into float or int.
* We assigned X and y values, and reshaped y.
* We Split the data into training and testing and imported from sklearn.preprocessing import StandardScaler to scale data for model.
* We transformed the training and testing data using the X\_scaler and y\_scaler models to Normalize data.
* We did not want normalize year, state, and re-election values, since we were not looking into the numeric values themselves but looking to see if patterns existed with election year, state, or election type.
* Because we did not want to normalize all x values, after scaling the columns we wanted, we reset x train and x test to scaled values. I am still unsure if this was the best approach.
* After Creating a Linear Regression model and fitting it to the scaled training data, we used X\_test\_scaled, y\_test\_scaled, and model.predict(X\_test\_scaled) to calculate MSE and R2
* **ML Linear Regression Model 2:**
  + Features: Election Year, State, Voter Eligible Population, Voters Registered, Primary Voter Turnout Count, Type of Election
  + **MSE: 0.05744684732134782 R2: 0.9914012628851123**
  + The MSE tells shows that our data values are dispersed closely to its mean, and our R2 shows a high positive correlation.
* We made predictions using a fitted model and plotted the difference between the model predicted values and actual y values, versus the model predicted values.
* We also created a data frame comparing actual values, predicted values, training residuals and test residuals.
* **ML Linear Regression Model 2:**
  + Features: Election Year, State, Voter Eligible Population, Primary Voter Turnout Count, General Turnout Count, Type of Election   
    (Removed Registration)
  + **MSE: 0.05744684732134782 R2: 0.9914012628851123**
  + The MSE shows that our data values are dispersed midway between that of Model 1 and Model 2 relative to its mean, and our R2 shows a positive correlation midway between that of Model 1 and Model 2.
* **ML Linear Regression Model 2:**
  + Features: Election Year, State, Voter Eligible Population, Primary Voter Turnout Count, General Turnout Count   
    ( Removed Election Type & Registration)
  + **MSE: 0.09559354853215604 R2: 0.9856914028874513**
  + As we reduce the # of features, we see the error rate continue to increase and our R2 correlation decrease.

**Machine Learning Conclusions:**

Relations, Correlations, Predictions

* While our R2 indicated high correlation in all 3 models, we saw a notable increase in prediction error rate as we reduced the number of features.
* Correlation is often interpreted as causation which is a big misconception. Correlation between variables does NOT indicate causation. Any highly correlated variable should be examined and thought of carefully.  Correlations are very useful in many applications, especially when conducting regression analysis. However, it should not be mixed with causality and misinterpreted in any way. You should also always check the correlation between different variables in your dataset and gather some insights as part of your exploration and analysis.
* We found General Election Voter Turnout was affected by the characteristics that we identified. However, Primary Turnout did not have the greatest impact, instead, Voter Registration had the greatest correlation with General Voter Turnout.
* If we had more time, we would investigate the impact in Primary turnout based on Primary/Caucus, Open/Close Primary, and many other factors that can give us more insight.

**Tableau Visualization Conclusions**

**Bar Chart/ Line Chart**

* The trend overall all for all variables is going up (VEP, Registered, General, and Primary)
* Re-elections years have lower primary turnout, which makes sense, since incumbent is not being challenged. The turn out for general during re-election, however, continues to increase with registered voter increase and does not seem to be affected by primary turn out.
* The one outlier is 2008 and 2012.

**Scatter Plot**

* As registrations increase, General Voter Turnout increase as well in most cases.
* Interestingly, as registrations increased, it did not seem to have a dramatic affect on primary turnout. When looking at the high registration states with low primary turnout we did notice they were all caucus states. After research, we realized all the extremely low primary voter turnouts were mostly caucus states.
* It would be interesting to dig deeper to see if open/close primary affects turnout as well, we were not able to get that data.