Capstone Project 1 Milestone Report:

Goal:

Tens of millions of Americans vote for their House and Senate representatives every two years and their president every four years. The last presidential election saw 128 million people cast their votes. Polling organizations like Gallup, the Pew Research Center, and YouGov attempt to poll America and determine in advance which way these elections will go. Due to the heterogenous nature of the country and cost of conducting wide, far-reaching polls, poll quality can vary greatly. This project seeks to examine a broad slice of polling data against election results to determine which polls are trustworthy and, more importantly, what makes a poll trustworthy.

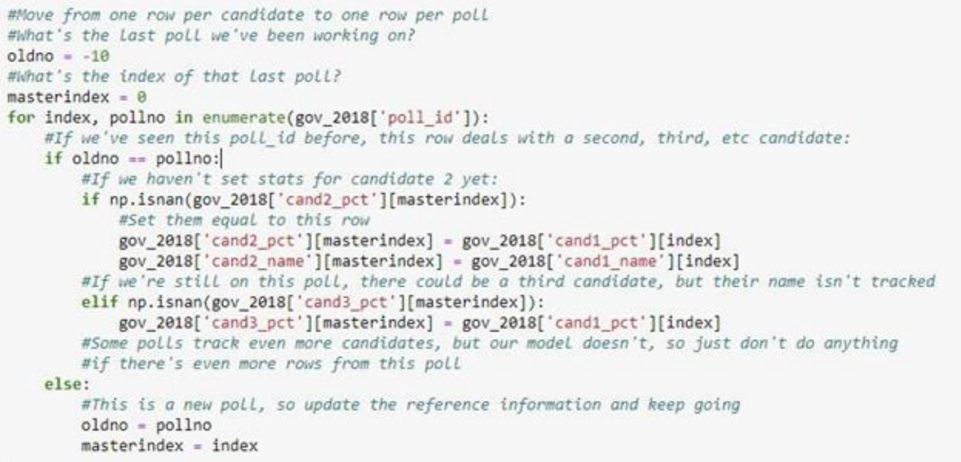
Polling organizations conduct a wide variety of polls, from automated phone calls to online surveys to in-person interviews. These methods differ wildly in their costs to conduct and their accuracy. As phone scams and spam calls become more frequent, people answer the phone less, and the costs of polling increase. Responses to phone polls are down by around 70%. As more and more Americans find politics a source of interest and concern, it is necessary to know just what a poll means and how much it can be trusted.

This project seeks to forecast the quality of a poll by examining its methodology – how many people were surveyed, where the poll took place, how the pollster has performed in the past, et cetera. This is of interest to any political candidate who wishes to know how they are performing and how likely they are to win an election. The target audience is polling companies such as Gallup Poll and Harris Interactive, who would be interested in determining how to best improve their polling strategies with the least cost to themselves.

Data:

The website FiveThirtyEight collects and analyzes polling data from hundreds of different pollsters. The data are curated to remove low-quality and outright falsified polls. The data are publicly available via a GitHub page and were compared against final election results, taken from the New York Times and Wikipedia. Candidate party affiliation was also taken from the New York Times and Wikipedia where necessary.

The 2018 data were formatted differently than the rest of the data, with one candidate per poll per row, instead of the prior convention of one poll per row. The 2018 data were cleaned to match and the data were merged. Several variables of interest, such as time between poll and election, or state the poll was conducted in, were also added to the dataset.



```python
#Move from one row per candidate to one row per poll
#What's the last poll we've been working on?
oldno = -10
#What's the index of that last poll?
masterindex = 0
for index, pollno in enumerate(gov_2018['poll_id']):
    #If we've seen this poll_id before, this row deals with a second, third, etc candidate:
    if oldno == pollno:
        #If we haven't set stats for candidate 2 yet:
        if np.isnan(gov_2018['cand2_pct'][masterindex]):
            #Set them equal to this row
            gov_2018['cand2_pct'][masterindex] = gov_2018['cand1_pct'][index]
            gov_2018['cand2_name'][masterindex] = gov_2018['cand1_name'][index]
        #If we're still on this poll, there could be a third candidate, but their name isn't tracked
        elif np.isnan(gov_2018['cand3_pct'][masterindex]):
            gov_2018['cand3_pct'][masterindex] = gov_2018['cand1_pct'][index]
        #Some polls track even more candidates, but our model doesn't, so just don't do anything
        #if there's even more rows from this poll
    else:
        #This is a new poll, so update the reference information and keep going
        oldno = pollno
        masterindex = index
```

Figure One: Code used to clean and reformat the 2018 dataset.

The major issues that arose when tracking and wrangling the data were the intermittent presence of third-party candidates, named candidates, and single-party races. FiveThirtyEight stores polls and results as the percent of the vote received by ‘candidate 1’, followed by ‘candidate 2’. The convention is that, in partisan races, candidate 1 is the Democrat and candidate 2 is the Republican. However, in races such as the California 2018 senate election, both candidates were Democrats.

The cand_name columns attempt to clarify party affiliation, but these columns can contain names, such as ‘Kerry’, names followed by party affiliation, such as ‘Lamont (D)’ or only party affiliation, such as ‘Republican’, making an analysis of polling error by party much more difficult, as there was no concrete way to identify party affiliation from the data. To solve this, candidate affiliations were found on Wikipedia and added to the dataset.

The primary dataset was cleaned and ready to be analyzed: Over 10,000 polls taken over 20 years of polling, tracking such factors as year, sample size, methodology, error and bias in the poll, state the poll was taken in, and many more.

Statistical Data Analysis:

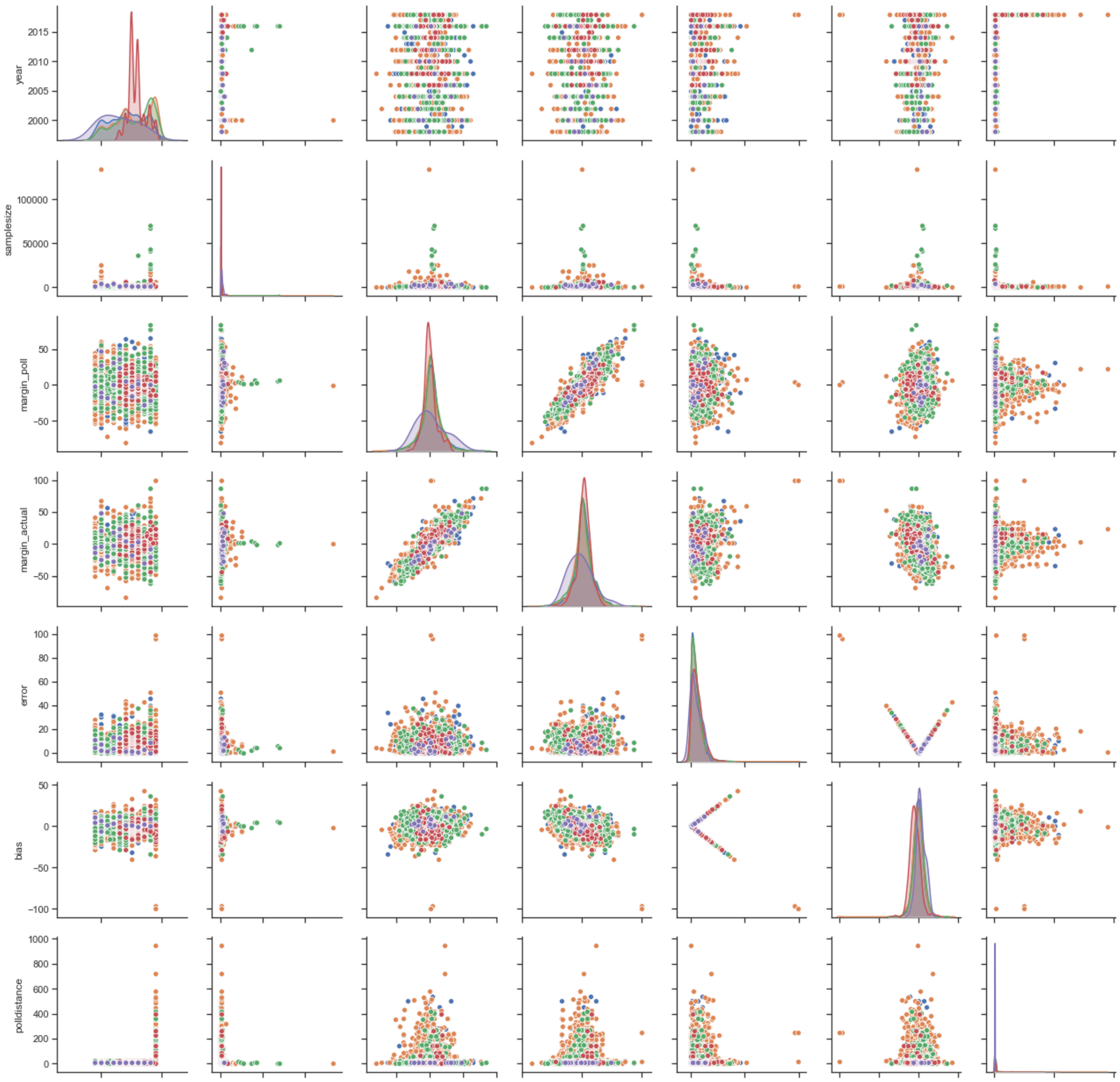
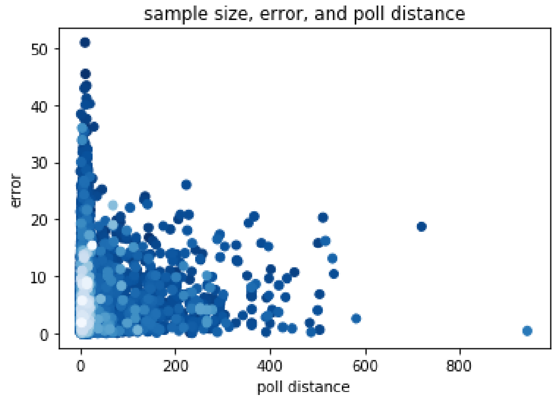


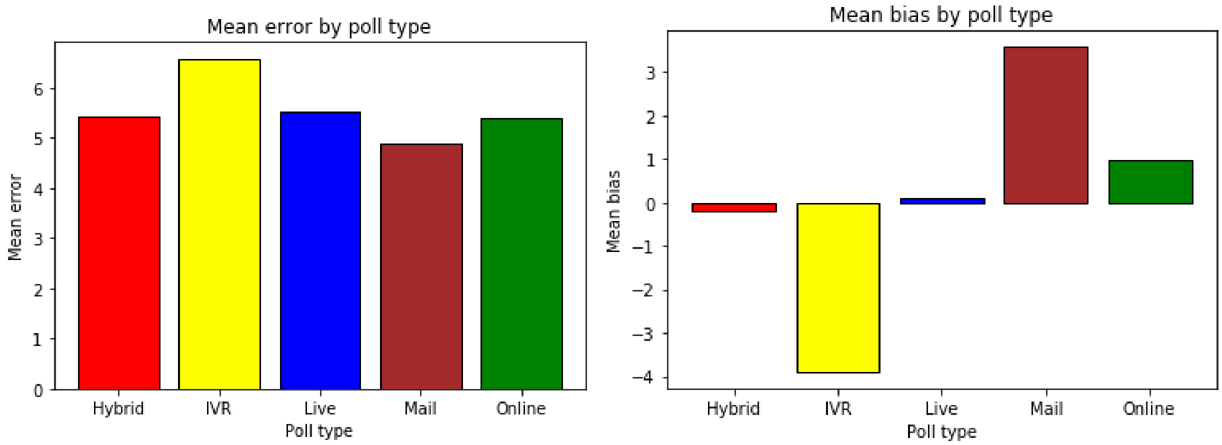
Figure Two: Pairplot of the most relevant statistics in the capstone dataset. Points are colored by poll type.

The goal was to find the variable that exerted the most influence on the error and bias of a poll. ‘Error’ here means the absolute value of the difference between the predicted result of the election and the final result of the election. ‘Bias’ is like error, but can be negative. Per FiveThirtyEight convention, in partisan races a positive bias indicates the Democratic candidate outperformed their polls, and a negative bias indicates the Republican candidate outperformed their polls.

Two of the first tested hypotheses tested were that polling data improves as the election approaches, and the type of poll strongly impacts the error of the poll. Neither proved to be exactly correct.



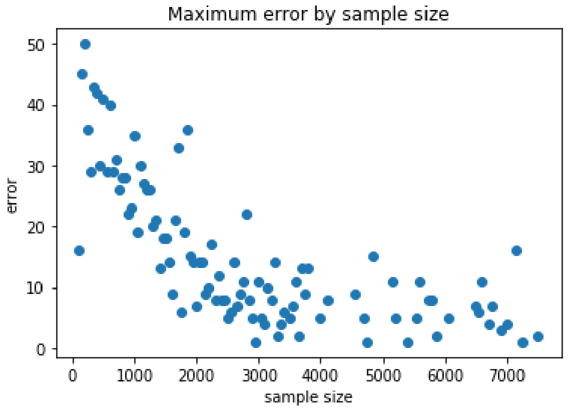
As this graph indicates, we observe a spike in extremely inaccurate polling in the days immediately before the election. This is likely because many of these last-minute polls have low sample sizes. Polling error appears to negatively corellate with time to election, but sample size is a confounding variable here.

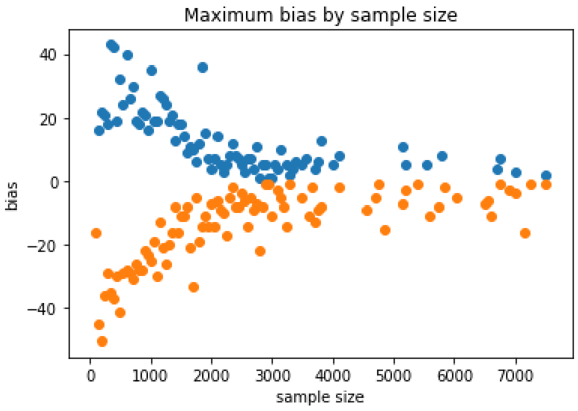


Poll type strongly predicts bias, but not error: All polling methodologies are likely to miss their mark by the same margin, but some polls tend to overrate Democratic candidates, some overrate Republican candidates, and some show no consistent bias.

It was immediately clear that sample size correlated negatively with the maximum error of a poll. Determining the relationship between error/bias and polling sample size was complicated by the fact that sampling size described a maximum error, but did not restrict the minimum. In other words, it was possible to sample 50 people and make an error-free poll, but a poll with 7500 respondents was very unlikely to be badly wrong. To account for this, I binned the data by sample size in 50-person intervals and took the maximum error of each.

This produced the following plots:



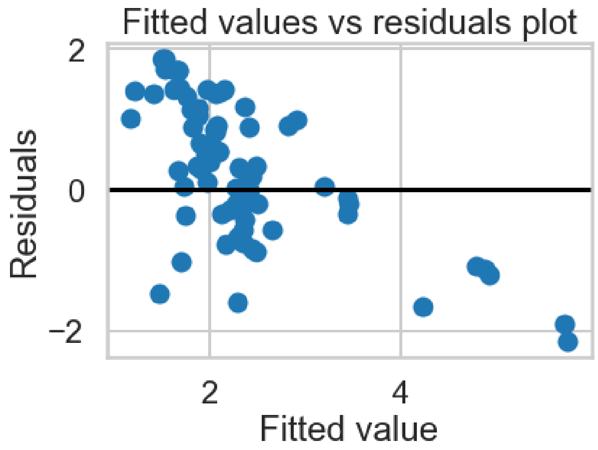


We can see that maximum bias and error have a roughly linear relationship with sample size up to around 4000 respondents. Increasing sample size beyond this threshold does not significantly improve accuracy.

I examined several models and found the statsmodels OLS was a good fit for the data. A simple linear regression of sample size up to 4000 respondants against error yielded an adjusted R-squared of 0.655 and F-statistic of 142. With a p-value of zero, it was clear that these factors were correlated.

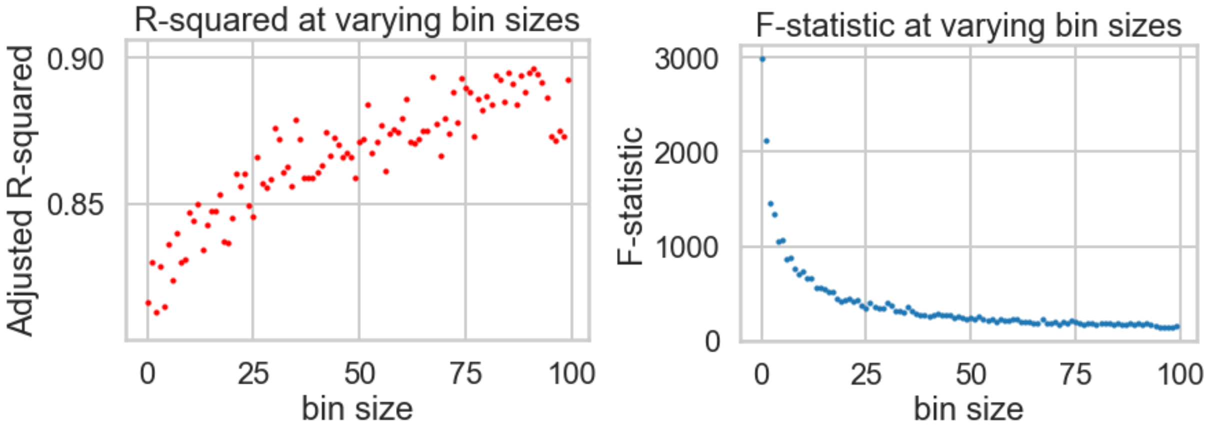
Another relevant variable was the prior performance of the pollster that performed the poll. After some testing, the best metric to use to incorporate this data was found to be the standard deviation of their polling errors. Adding this variable into the OLS regression produced a model with an adjusted R-squared of 0.832 and F-statistic of 184. P-values were 0.020 for sample size and 0 for pollster standard deviation.

This model was further improved by taking the natural logarithm of the error, improving the OLS regression to an adjusted R-squared of 0.859 and F-statistic of 226. P-values were zero for both sample size and pollster.

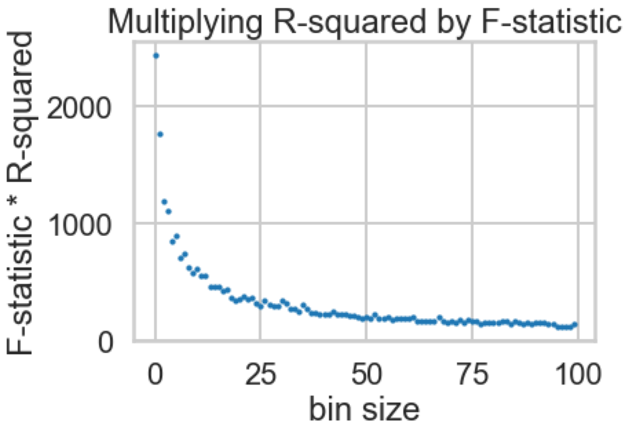


As this residuals plot indicates, the residuals are not perfectly evenly distributed, indicating some bias remains.

This analysis was done with an arbitrarily chosen binning of 50. After determining that there was strong correlation between sample size, past pollster performance, and error, the model was re-tested for a range of bins up to 100 in size.



Running the model with smaller bin sizes reduces the R-squared value but increases the F-statistic. The gains in F-statistic are significantly larger than the losses in R-squared, as shown here:



It appears that smaller binnings produce better results with this model. The added computational cost is not prohibitive: we will use smaller binnings to model maximum error.

This model does well for predicting maximum error, but fares significantly less well for predicting exact error of a given poll. When applied to the dataset of all polls with no binning for error maximization, R-squared fell to 0.516. Additional modeling is needed to accurately predict the actual error of a poll, not just the maximum error.

Next steps:

The statsmodels OLS of log(error) against sample size and past pollster performance was a strong predictor of maximum error, but a more robust model is needed to predict actual error. One promising avenue of approach is to incorporate methodology and track bias instead of error: This will introduce significant dimensionality into the results, as we will divide the error dataset in half and add in several new methodology variables, but we have seen that methodology is a strong predictor of bias.

Another option is to examine other methods to track past pollster performance: mean error proved to be a poor fit for modeling maximum error, but it may fare better for predicting exact error.

A major priority is developing a high-quality way to visualize the model and act on its information. An interactive model showing the range of error possible for a submitted sample size, pollster company, and poll methodology would clearly and effectively communicate all findings.