To-Do

ML Proposal

1. Back Ground

## **Proposal Outline**

#### **How this problem is solved today/previously**

* *What interventions exist/will exist*
  + *Not many for local level*
  + *Typically assess using social media data*
* *What data do you have and what additional data will you need?* 
  + *We have census data*
  + *(Maybe) have precinct level outcomes data*
* *Important*: You should do data exploration and provide descriptive stats to show that you have enough relevant data to solve this problem

#### **What analysis are you proposing to do? (pause)**

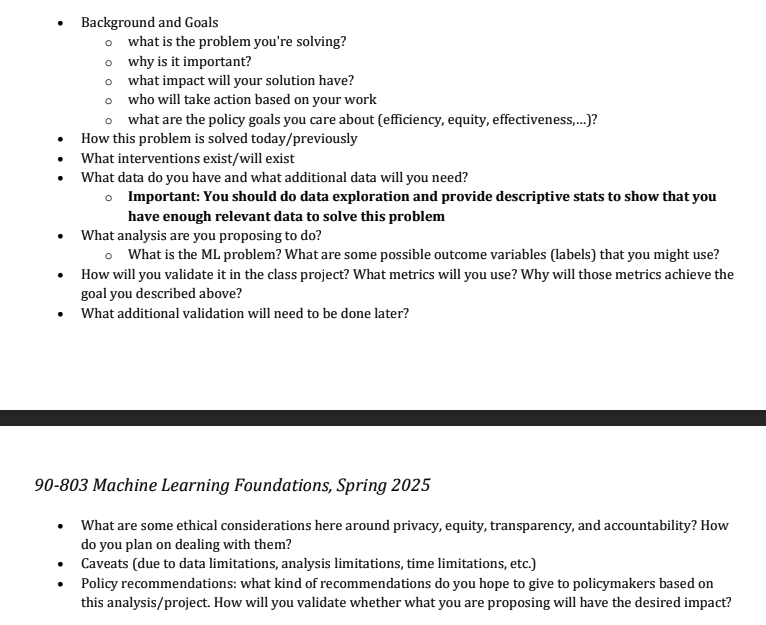
* *What is the ML problem? What are some possible outcome variables (labels) that you might use?*
  + *Prediction of Dem winning or Republican*
* *How will you validate it in the class project? What metrics will you use? Why will those metrics achieve the goal you described above?*
  + *Accuracy*
* What additional validation will need to be done later?
  + Different context (cities, or years…)

#### **Caveats and Concerns**

* What are some ethical considerations here around privacy, equity, transparency, and accountability? How do you plan on dealing with them?
* Caveats (due to data limitations, analysis limitations, time limitations, etc.)

##### **Policy recommendations:**

* What kind of recommendations do you hope to give to policymakers based on this analysis/project?
* How will you validate whether what you are proposing will have the desired impact?



Just for reminder for myself

* Granularity city v precinct // multi-city
  + Mayoral electoral results may only be found on county or city’s websites, instead of state’s websites
  + Precinct: San Francisco, Pittsburgh

|  | **City** | Precinct |
| --- | --- | --- |
| Data Accessibility | **High** | Low |
| Sample Size | **Small (large # of cities)** | Large |
| Salient | **Higher** | Lower |
| Policy Implication | **May also redistricting** | Redistricting? |
| Coverage | **Cover more regions** | Cover only few regions |

* What is the ML model prediction (e.g., vote share or just party)
  + Vote share of just one party (Democrats): vote share of the other party is usually 1 - Democrats
    - If we focus more on the blue cities, DV could be the vote share of Republicans
    - Alternative: Vote share of independents: we focus on what factor may boost independents’ popularity
      * Better to use City-level data instead of precinct. Each precinct may has only very few votes of independent
  + Who wins?
    - Less policy implication
    - Prediction of vote share can also predict who will win
  + Voter turnout
    - Could be the DV or IV
    - If IV, how salience this election is

City-level vote share (Vote share)

**3 regression models** each predicting the vote share of the different political groups (i.e., dem, rep, and others)

* Last 3 elections from the largest city of the 50 states
* Pittsburgh
* WHat’s targeted set of number of cities how far back do we want to go

Each row is an election - one label is vote share of one party

* Zoe will look into how to pull census data

Proposal Bullet Points

Background and Goals

* Most approaches center on the federal level
* Aiming to establish more awareness and importance on local elections

Current Approaches

* Most online polls center around presidential elections, with HoR and Senate elections being somewhat prevalent ([link](https://elections2024.thehill.com/forecast/2024/president/pennsylvania/))
* One study by the University of Illinois uses Bayesian model to predict probabilities ([link](https://ea-staging.web.engr.illinois.edu/site/file/Rigdon2009.pdf))
* “Who are you voting for?” vs. “Who do you think will win?” ([link](https://www.brookings.edu/articles/forecasting-elections-voter-intentions-versus-expectations/))

Datasets

* Data focuses around select cities known for highly competitive political atmosphere
* ***TDB - name specific cities and reasons***
* Precinct census for select cities with competitive party affiliation
* Ballotpedia provides outcomes of municipal elections for various cities
* ***TDB - any more datasets? Sentiment analysis?***

Proposed Analysis

* ***TBD - discuss and decide model***
* ***TBD - how will we validate?***
* ***Clustering***
  + Which cities are most similar in terms of the ability to successfully challenge an incumbent? Which cities are most likely to be motivated to unseat an incumbent due to changes in economic conditions?
* ***Regression analysis***
  + Predicting the vote margin for challenger and incumbent in elections
* ***Decision Tree // Random Forest*** 
  + Given certain attributes of an election and city, predict if the challenger or incumbent will win the election.
  + Can see how the salient features change over time
* ***Validation***
  + Near term:
  + Long term: Elections are relatively low-frequency events; however, the model can be tested and updated before and after elections. The model’s ability to accurately predict the outcome can be easily validated once the results have been certified.
  + **Metrics**: Given the penalty for false positives is the same as that for false negatives in this scenario, we’ll be primarily assessing the models’ accuracy. :
    - What are the consequences of saying an incumbent will win if the challenger is supposed to win (FP)?
    - Consequences for saying a challenger will win but incumbent(FN)?

Ethical Concerns and Limitations

* Privacy:
  + Voting and census data has been previously collected by the U.S. government
  + Datasets and information are available publicly online, can be accessed by anyone
  + No data on any one specific individual, most is aggregated numbers
* Data ownership:
  + Data is under the ownership of the state/federal governments, thus public domain
  + The dataset has imputed information on race and gender of the candidates. The dataset creators did not get the candidates' consent for such predictions.
  + ***TBD - run through each data source and discuss***
* Bias, equity, and fairness:
  + Potential to lean data in one direction or another, attempting to fix this by using data from several cities with varying degrees of party affiliation
  + Parties aside from Democrats and Republicans are largely ignored
  + Predicting that one candidate will win may bias voters in one way or another
  + ***The dataset has imputed values for political party affiliation and demographic characteristics of the candidates. Racial / ethnic group predictions were estimated based on Bayesian name predictions, and image classification. These methods have the potential for bias.***

Transparency:

* + We maintain that all people can use our model, regardless of party affiliation
  + Can also share which features are most influential in our predictions once we’ve completed the analysis
* Trustworthiness and accountability:
  + No intentions of purposely feeding
* Limitations:
  + Little-to-no prior knowledge when scaling down to state level
  + Considered funding as a variable, but not many candidates have sent their expenditures
* Incumbents often have a huge advantage over challengers in part driven by fundraising, name recognition, and \_\_\_\_\_\_(reference). By helping uncover the features that support challenger wins, we can provide challengers and incumbents with insights to help create competitive elections, inform voters, and potentially incentivize mayors to provide better governance to their constituents while in office.
* Misc
  + This is largely the status quo given federal and state level election forecasts but…
    - Election results are already being challenged, adding these predictions could further fuel rumors that elections are rigged if the model predicts incorrectly (e.g., how did x candidate win if y was predicted to win?)
    - Making a prediction in one way or another may sway people not to come out to vote – (e.g., the incumbent is supposed to win so why bother)

Future Work

* Extend our results to other cities

Proposal Draft

Zoe Baker (zbaker)

Simon Corpuz (scorpuz)

Shawn Lu (shaogul)

## **Predicting challenger vs. incumbent wins in municipal elections 1980-2020**

**Purpose**

Every few years, presidential elections take the country by storm. Whether it be monitoring the race between candidates, debating the effectiveness of the current system, or throwing around accusations of rigged votes, most if not all American citizens have some awareness of federal elections. However, the same cannot be said about those that occur at city-wide levels. While the presidential election, and other federal elections (e.g., House of Representatives and the Senate), are largely scrutinized by the public, many elections for city-level positions do not have the same impact or level of competition. Incumbents often have a huge advantage over challengers in part driven by fundraising and name recognition[[1]](#footnote-0), even some research demonstrates the declining effect of incumbent advantages. It is our belief that placing more emphasis on city-level voting and incumbent versus challenger wins, we will assist both candidates and voters alike. Candidates for municipal elections would naturally benefit from a tool that allows them to predict their favorability within each municipality. Additionally, by helping uncover the features that support challenger wins, we can provide insights to help create competitive elections, inform voters, and potentially incentivize mayors to provide better governance to their constituents while in office.

**Current Approaches**

As a concept, election forecasting is not new. A simple Google search will display countless prediction models and polling sites that have been established over years of campaigning. Though as mentioned above, the majority of these online resources focus solely on federal elections. One online forecasting tool run by The Hill places a major focus on states such as Pennsylvania.[[2]](#footnote-1) The Keystone State, along with several others, play a pivotal role in deciding presidential election results and control of the House and Senate due to their competitive political landscape. These kinds of sites rarely have data pertaining to city-level voters.

Many federal election forecasts primarily focus on the either-or victory of one candidate over another, while others focus on the probabilities of each candidate’s success. To come up with such predictions, many models rely heavily on sentiment analysis of social media data. One study conducted by the University of Illinois created a model that utilizes Bayesian probabilities and sentiment analysis of Tweets to predict the chances of each candidate’s victory.[[3]](#footnote-2) Other models heavily rely on polling data which some critics have called more of a detection method as opposed to prediction. Furthermore, many studies frame their analysis with a focus on political party. By focusing on challenger versus incumbent outcomes, our analysis will provide greater insight into the incumbents’ advantage, allowing all candidates to tailor their campaign strategies accordingly.

Other studies have delved into the psychology of election forecasting. Experts from the Brookings Institute have claimed that regardless of whoever someone actually votes for, their expectations of the election results are much more likely to yield accurate predictions.[[4]](#footnote-3) This is likely because voters can make their own small-scale predictions based on the opinions of friends and family, or polling data they’ve encountered. While this study still focuses on federal elections rather than municipal elections, it does show that value can be found when reducing the overall scale. We will consider this for future modeling.

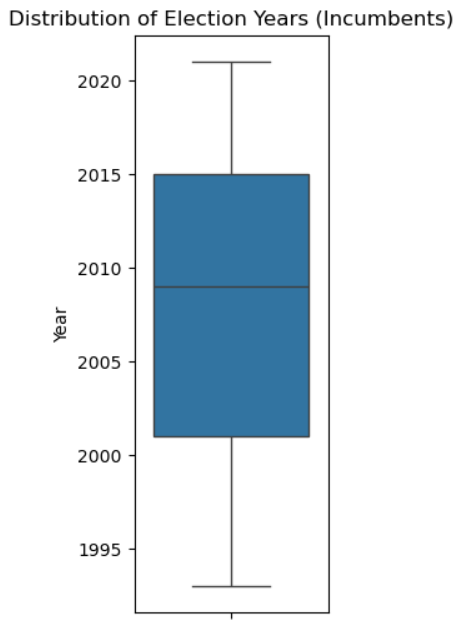
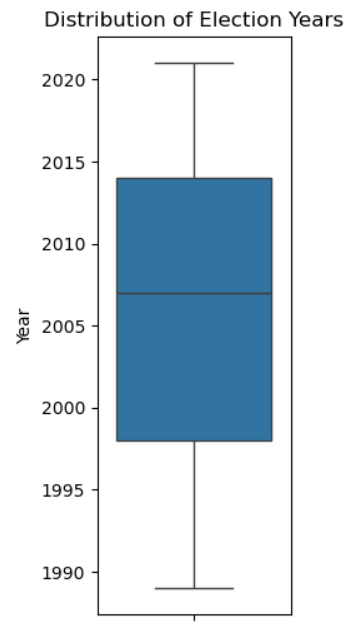
Our project contributes to the existing literature in several important ways. First, it centers on local elections, an area that remains relatively underexplored compared to national contests. Second, rather than forecasting which party will win, our model estimates the probability of victory for incumbents and challengers, offering a more nuanced perspective on electoral dynamics. Finally, unlike many predictive models that rely on polling data—often subject to sampling bias and representativeness concerns—our approach is independent of such data, enhancing both its robustness and applicability across different contexts.

**Datasets**

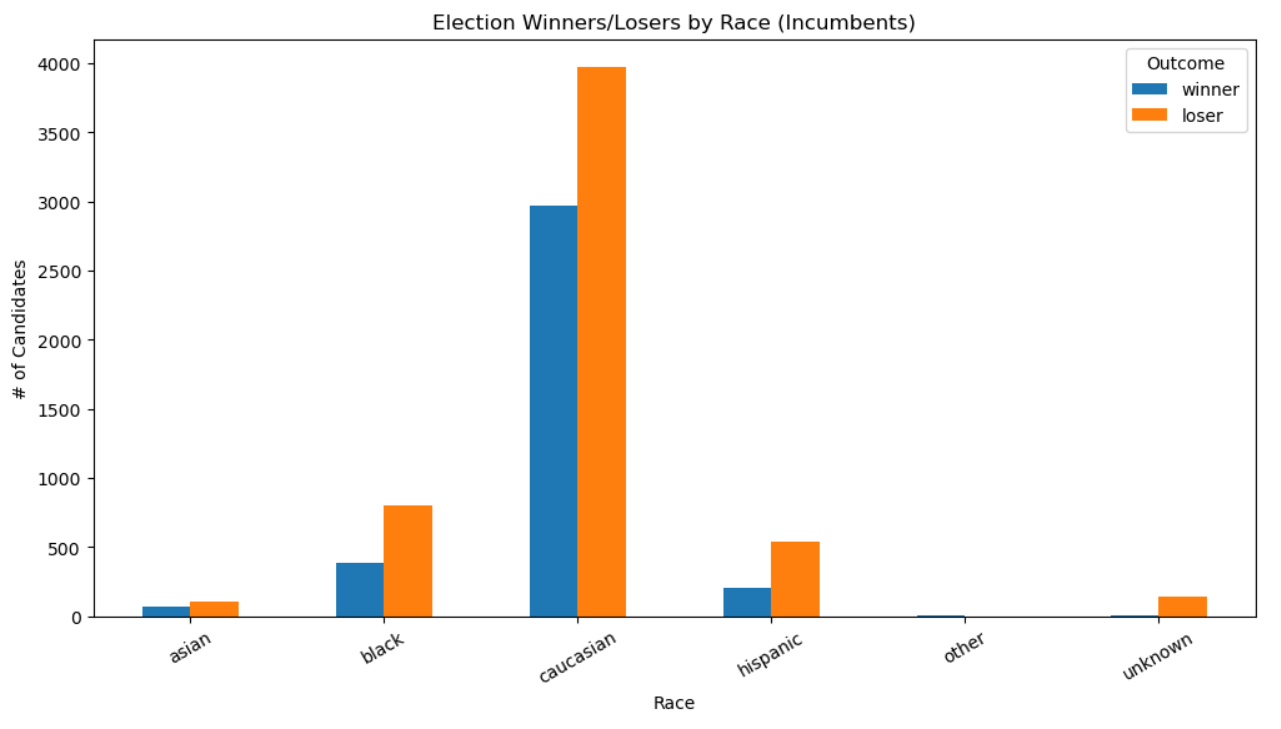
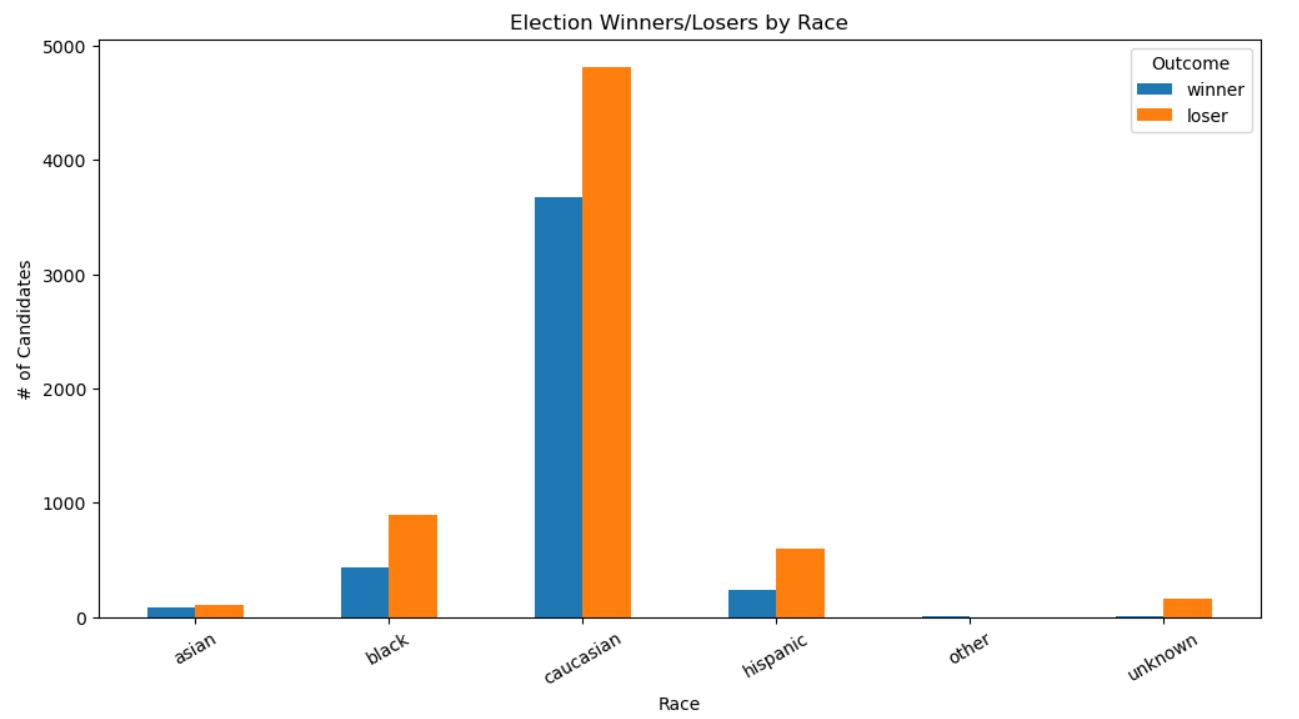
As stated previously, information on city-level elections is largely absent from online databases. However, we have managed to locate one extremely robust dataset hosted by the [Open Science Foundation](https://osf.io/mv5e6/). Each row in the dataset represents a candidate who ran for a local position, with information on their election year; the percentage of votes they received, and whether they won; political and demographic information; and whether they were an incumbent, or involved in an election with one. There are 11,020 records in total, with 9,215 involving an incumbent candidate in some form. In analyzing our data, we have made it a point to consider both cases separately.

We have largely addressed N/A values by replacing them with more fitting information. For example, one field provides an approximation of the candidate’s race based on various numerical estimations. When a result is unavailable, we decided to replace the null value with a new “unknown” category. This allows us to use many more variables without losing information. There are, however, some features such as a monetary contribution rating that are largely unable to be imputed with reason, and contain more than 5,000 null values. We decided to remove these variables.

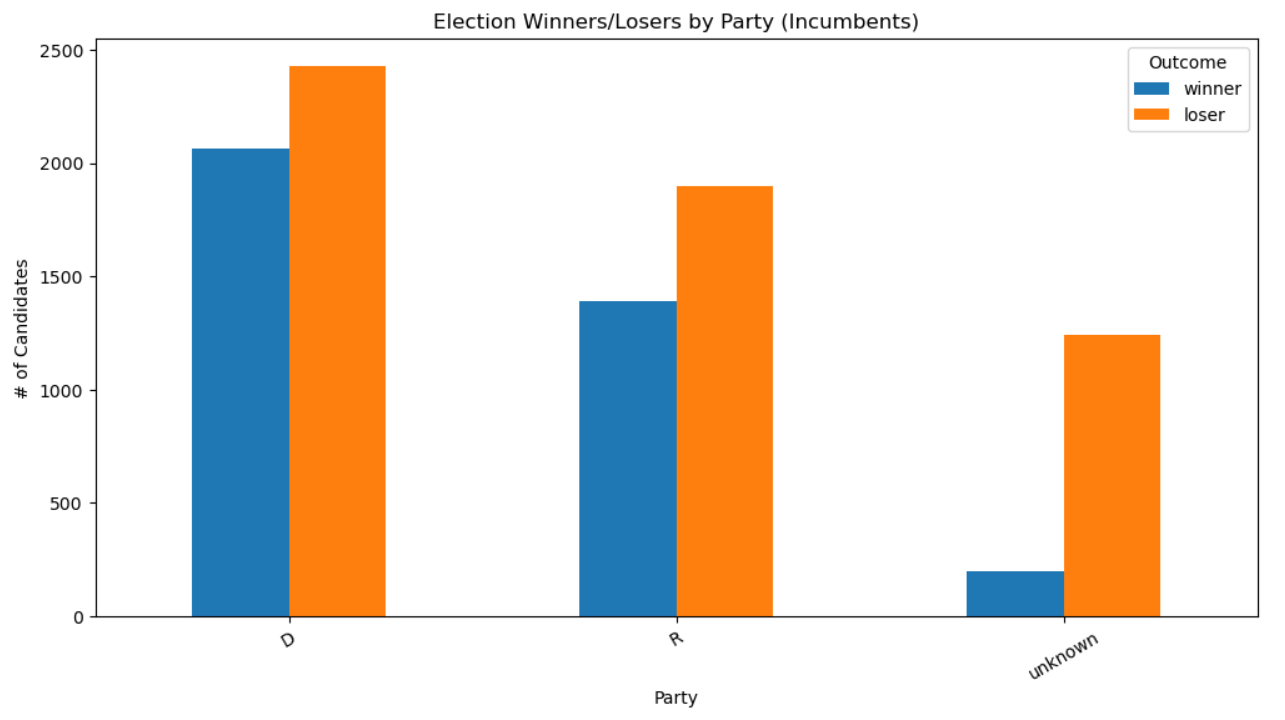
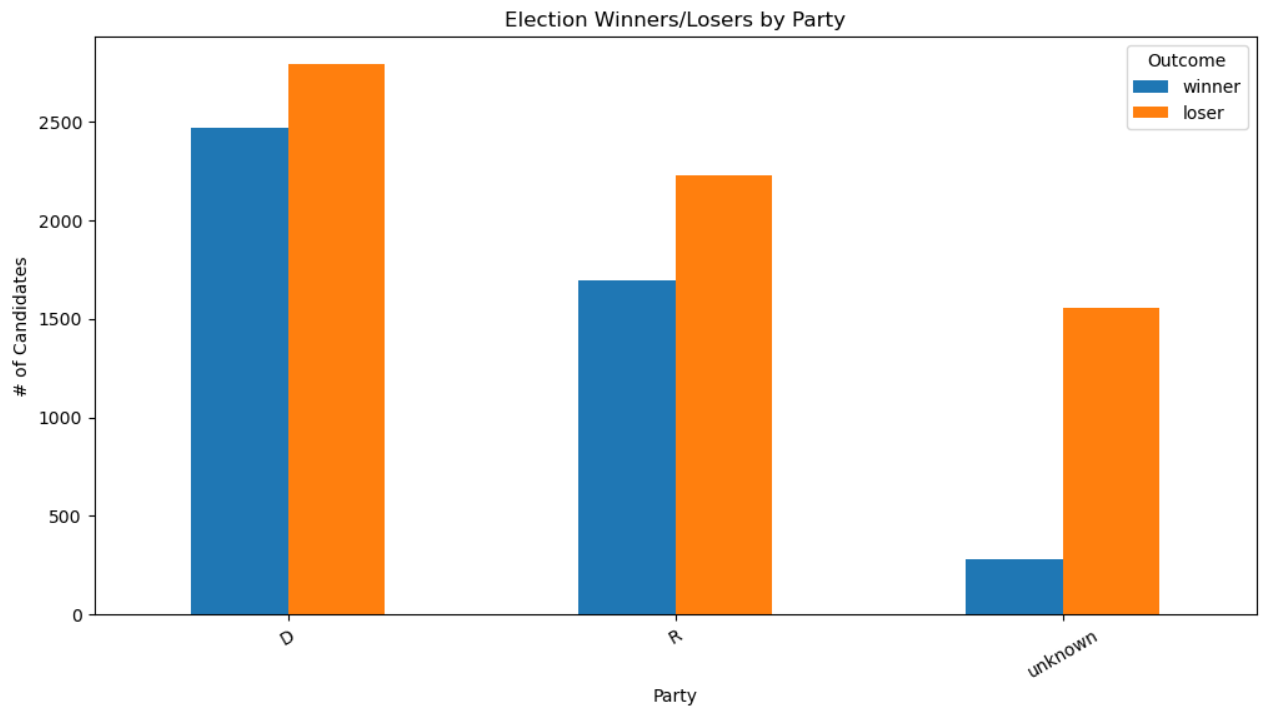
The boxplots below show that we have a relatively balanced range of election years, among both all elections and just those with incumbents. Both distributions are equally spread throughout the approximately 30-year range of 1990 to 2020.



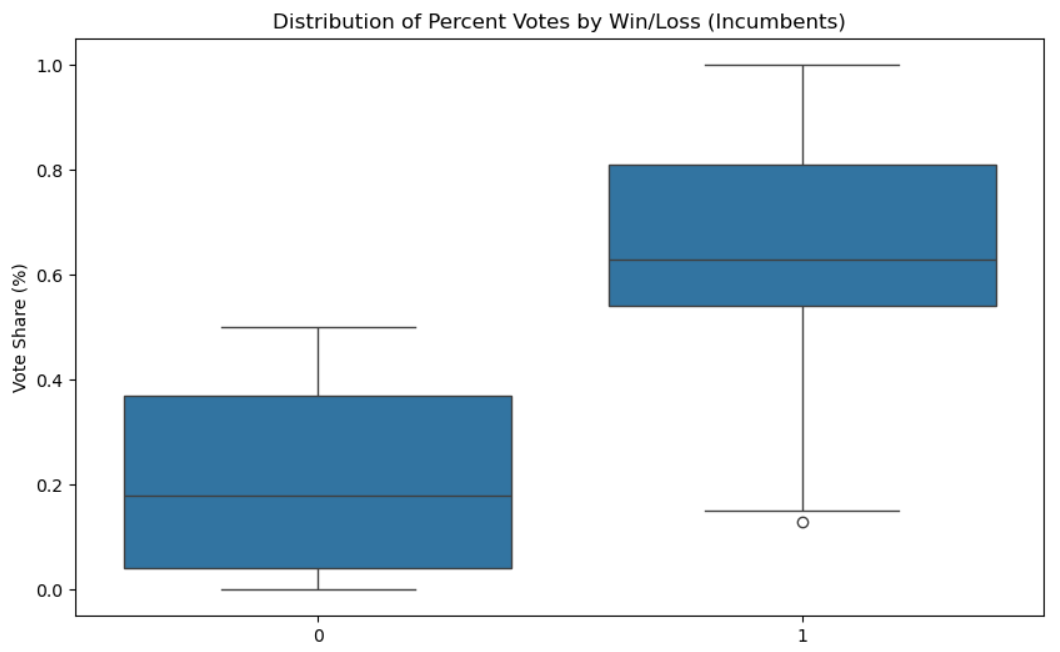
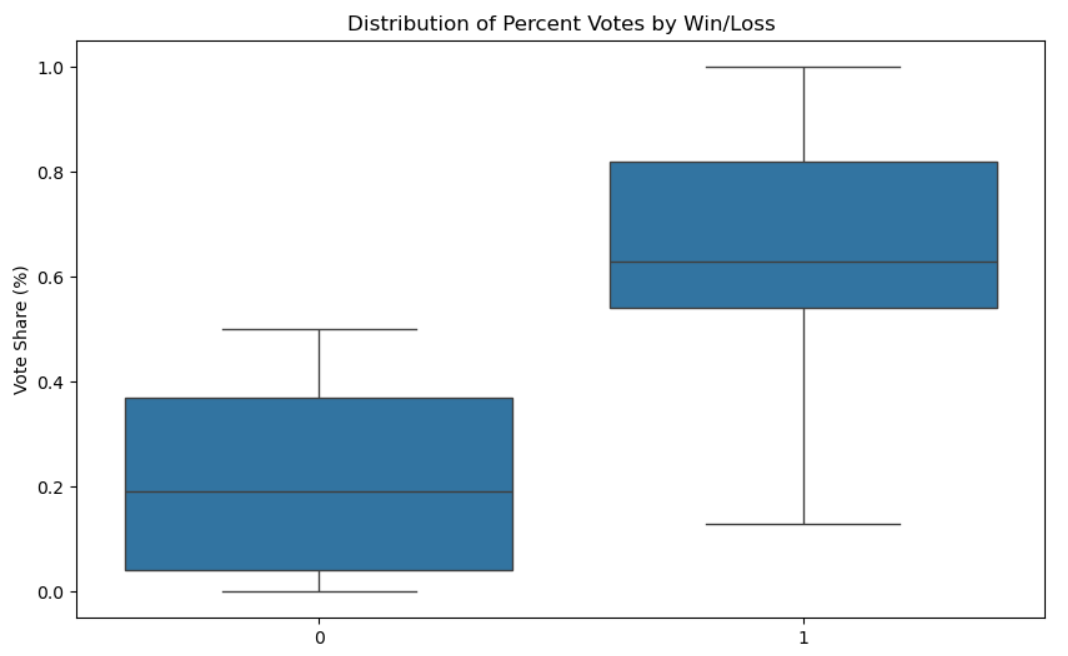
The data consists of a greater portion of people who lost elections. When we disaggregate the win/loss results by race, we see that a majority of candidates are white. The second largest group of candidates are Black. While they barely make up one-fourth the amount of caucasian individuals, they still number somewhere in the 500-1,000 range for both overall elections and incumbent elections. This has assuaged us slightly on whether there is enough data to accurately represent the smaller demographic groups.



We see a similar trend upon splitting the win/loss results by political party. There appear to be more Democrats than Republicans involved in city elections, which makes sense considering large cities lean more towards liberal policies. However, there are still at least 3,500 instances of Republican candidates being involved in incumbent elections. This provides us with enough data to balance political affiliation and avoid bias.



Lastly, there is a noticeable difference between the average percentage of votes that the incumbent has received when they win versus when they lose. This is to be expected, though the distributions are surprisingly similar between overall elections and incumbent elections. We would’ve expected incumbents to see a much larger gap between winners and losers, considering they are more experienced and well-known. This could potentially lead to our model working the same on both instances of elections.



**Proposed Analysis**

Our team will begin by predicting whether the incumbent or challenger will win the municipal election. To do so, we propose utilizing decision trees and random forest to classify the election results (i.e., will a challenger or incumbent win) To achieve the highest performance, we will will test multiple models (e.g., boosting, bagging) and different specifications. Finally, if we have time, we will use regression to predict the vote share each candidate will receive to complement our classification analysis.

The preliminary set of variables we have identified to inform feature generation includes:

* Election characteristics: location (e.g., state or region), year, # of candidates, voter turnout, results of the previous election
* Candidate characteristics: incumbent status, political affiliation, race, and gender
* Socioeconomic demographics: economic conditions (e.g., median income, # of residents on SNAP benefits), population demographics (e.g., size, race, sex, education-level),
* Geographic characteristics: region, population size

For classification tasks, our model will be optimized to accurately predict challenger wins because the incumbent advantage is so strong. For example, if we decide to assign a challenger win as a ‘positive’ event, then we’d aim to minimize false negatives with recall. When predicting the vote share for each candidate, we will measure performance with RMSE. Regarding long-term validation, elections are relatively low-frequency events; however, the model can be tested and updated before and after municipal elections. The model’s ability to accurately predict the outcome can be easily validated once the results have been certified.

**Ethical Concerns and Limitations**

When examining ethical considerations and limitations, we must primarily focus on the data utilized and the potential unintended consequences of our analysis. The dataset we use for the elections contains predictions on candidates’ race/ethnicity and gender based on Bayesian name predictions and image classification. Regardless of the accuracy of the predictions, candidates have not consented to have their race or gender predicted, and this could be considered a violation of their agency. However, when someone enters an election for public office, they forfeit aspects of traditional privacy. Additionally, predicting the election results has the potential to bias the outcome or fuel misinformation if the prediction is inaccurate. Election results are already challenged, and incorrect predictions could further fuel rumors that elections are rigged. Still, this risk is generally acceptable and is largely considered the status quo given federal and state-level forecasts.

The other data we plan on using does not present substantial considerations related to ethical concerns. The Census, American Community Survey, and election results data are public and are published by the U.S. or municipal government. This information has been made public domain and can be freely used for our model.

Our project aims to enhance transparency in the municipal electoral process. As previously mentioned, our analysis can identify the factors that contribute to the success of challenger candidates, thereby promoting more competitive elections, especially in cities without term limits. Additionally, by providing voters with information about key aspects of challenger victories, such as voter turnout, we hope to encourage greater civic participation and increase public awareness.

**Future Work**

We have identified a few places that we can improve in the future work.

1. Differentiate each city based on their electoral system
   1. For example:
      1. Nashville: two-round system
      2. Salt Lake City: ranked-choice system
      3. Dallace: first-past-the-post system
   2. Some electoral systems may provide more chance for challengers, such as ranked-choice
2. Clustering to identify similar types of cities and candidates:
   1. Challengers may have higher chance to win in certain types of Cities
   2. Certain types of candidates may have higher willingness to run as challengers
   3. What types of challengers have higher chances of winning? It’s the populist outsiders or the party’s insiders
3. Sentiment analysis of social media and newspaper discussions of scandal and campaign issues (add as features)
   1. Campaign issues: environment, Medicare, homelessness, safety, etc.
   2. Scandal: corruption, criminal record, etc.

1. Smith, B. (2023, December 20). Why do incumbents have such a big advantage in elections? *IPB News*.<https://www.wfyi.org/news/articles/why-do-incumbents-have-such-a-big-advantage-in-elections> [↑](#footnote-ref-0)
2. “Pennsylvania Presidential Forecast,” The Hill, November 5, 2024, <https://elections2024.thehill.com/forecast/2024/president/pennsylvania/>. [↑](#footnote-ref-1)
3. Steven E. Rigdon, et. al., “A Bayesian Prediction Model for the U.S. Presidential Election,” American Politics Research 37, no. 4 (2009): 700-24, <https://ea-staging.web.engr.illinois.edu/site/file/Rigdon2009.pdf>. [↑](#footnote-ref-2)
4. Justin Wolfers and David Rothschild, “Forecasting Elections: Voter Intentions versus Expectations,” Brookings Institution, November 1, 2012, <https://www.brookings.edu/articles/forecasting-elections-voter-intentions-versus-expectations/>. [↑](#footnote-ref-3)