For project 2 we propose to focus on creating descriptive statistics and visualizations from Political TV Ad Data and additional political demographic datasets. We hope to summarize which types of ads are being aired by region, time, party and current election results. We will then draw some correlations between certain types of ads and the votes received by candidates as well as visualize the patterns in ad buys. In order to limit the scope of our analysis somewhat, we will treat states as the unit of analysis (as opposed to county or city) and we will limit the dataset to only presidential primary election data, excluding Senate and House races.

Owen will be focusing on the relationship between television advertisements and receiving votes. There are several key variables which describe important characteristics of each advertisement: message (pro/con/mixed), type of television program it was aired on, time of day, sponsor type (candidate committee, Super PAC, etc.), political party, and state. First we will summarize the frequency of ads run in each of these categories. Next, we will link the ad data the election results dataset in order to examine the correlation between number of ads run in a state by groups supporting a particular candidate and the percentage of votes that candidate received. In order for the merge between these datasets to work properly, an ad will count as supporting a candidate in the ads dataset regardless of whether it is run by the campaign itself or by an outside organization. This means that all ads that are categorized as supporting a candidate in the ads dataset, regardless of whether the message of the ad was positive for the supporting candidate or negative towards another candidate, will be factored into our correlations with that candidate’s vote percentage. Once the overall correlation between ads and votes has been calculated, the ads dataset will be further broken down by the key variables in order to calculate correlations at a more granular level.

Elizabeth will focus on creating visualizations to detail how data from the Political TV Ad data and County Facts data clusters over a map of the United States. The visualizations should support the majority of the analyses completed by Owen. The frequency of ads run in each category will be overlaid on a map of the United States, while deep dives into individual states or more granular analysis will be displayed either by state or tabulation that clearly supports our findings. To complete these visualizations Elizabeth will use the Geopy library for coordinates, Bokeh, Matplotlib, D3 and possibly Seaborn. We will additionally focus on density of each clustered variable and the states and/or overlapping states where the least dense and densest clusters exist. Elizabeth also hopes to provide visualizations showing bucketed groups of time that ads will be displayed taking into consideration time zones, day of the week and corresponding descriptive statistics, such as min, max, sd… to be displayed at the bottom of each visualization that relate to Owen’s analysis. She will use the datetime module along with the Arrow Library. For the additional measures and parameters detailed above, such as groupings created for aggregation and filter logic to ensure that the dataset is manageable for the project size, a data dictionary will be made for the final datasets created and all aggregated datasets will be published on the GitHub repo.

QUESTIONS:

1. What is the positive/mixed/negative ad distribution for each candidate and how does it change by state and over time?
2. For which states and candidates do the number of ads run correlate the strongest with the share of the vote received?  More generally, what is the overall correlation between ads and vote share?
3. How does the distribution of outside group vs candidate committee support (in the form of ads run) change as election day approaches? Does this vary by state or candidate?
4. What day and what time of the day are the most ads run? Does this vary by state?
5. Which candidate had the most tv ads, did this vary by state? Is there a visual decrease in number of ads run as the election proceeds?
6. Did the number of ads aired for a candidate and the total number of votes received have a positive or negative correlation? How to show this in a visualization.

Structure of paper:

* Introduction:
  + Describe the dataset and parameters
  + Describe measures created, graphing and tabulation packages used
* Research Question 1:
  + Describe methodology
  + Present findings (tables, graphs, maps)
  + Discuss results
* Research Question 2:
  + SAME
* RQ 3
* RQ 4
* RQ 5
* Conclusion

Links to additional datasets and packages with small description of dataset and how it will be used in the analysis:

1. Geopy
   * 1. Geopy has a library with coordinates for counties and states within the US and additional datasets for coordinates within the world
     2. <https://github.com/geopy/geopy>
2. 2016-US-Election Dataset
   * 1. <https://www.kaggle.com/benhamner/2016-us-election>
3. Bokeh
   * 1. A visualization library
     2. <http://bokeh.pydata.org/en/latest/>
4. Matplotlib
5. Arrow
   * 1. http://crsmithdev.com/arrow/
6. Seaborn
   * 1. <https://stanford.edu/~mwaskom/software/seaborn/installing.html>
     2. Still haven’t decided if will use
7. D3
   * 1. D3py
     2. Elizabeth hopes to use the d3py module to generate xml pages, they will be added to the report and the link will be provided

Please visit our repo for further information and to review each dataset, developing data dictionaries and basic profiling that has been completed.

Link: <https://github.com/CHABOBO/project_ad_el_archive>