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 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.

While Elizabeth would like to 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 all analyses completed by Owen and will additionally focus on density of each cluster and the states and/or overlapping states. Elizabeth also hopes to provide visualizations showing bucketed groups of time that ads will be displayed taking into consideration time zones and day of the week and corresponding descriptive statistics, such as min, max, sd,… for each visualization. For any additional measures made a data glossary will be provided and a data dictionary will be made for the final datasets used.

|  |
| --- |
| Who are the people that put up the advertisement and do they support or oppose that person? |
|  |

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. 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 glossaries.

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