Project 2 Proposal

Team Members
Bronte Baer, Jean-Luc Jackson, Christian Montecillo, and Richard Robbins

GitHub Repository
Project2_Baer_Jackson_Montecillo_Robbins

Primary Dataset
The Political T.V. Ad Archive, a project of the Internet Archives, archives political ads and tracks ad airings in select markets in the 2016 election. This dataset is one of the pre-approved datasets, and it appears at: http://politicaladarchive.org/data/.

We refer to this dataset as the "advertising dataset" in this proposal.

Each row in the advertising dataset corresponds to a single airing of an ad. We will use the following data from the advertising dataset:

- *location* (name of T.V. broadcast market, based on Nielsen Market names)
- *start_time*
- *end_time*
- *sponsors* (organization sponsoring the political ad, as it appears in the ad)
- *subjects* (subjects covered in the ad; subject index from PolitiFact).
- *candidates* (candidates named in the ad)
- *message* (pro, con, mixed; input by Internet Archive researchers)

The television ad data is specific to thirteen regional markets in the United States. Each of the thirteen markets is a Nielsen Designated Market Area. Each of the markets includes more than one city, none of the markets cover an entire state, and several include portions of more than one state. Each market comprises several counties.

Additional Dataset
MIT Election Data and Science Lab, County Presidential Returns 2000-2020, published by the Harvard Dataverse. We will use this dataset to consider how our analysis of advertising patterns relates to the voting results in the relevant markets, and it appears at: https://doi.org/10.7910/DVN/VOQCHQ.

We refer to this dataset as the "votes dataset" in this proposal.

Each row in the votes dataset corresponds to the number of votes received by a specific presidential candidate in a particular county of a state in the general election for president. The votes dataset includes information for each presidential election from 2000 through 2020. We will use the data from the 2016 election.

Additional Supplemental Information
We found maps and lists that identify which counties are in the various broadcast markets on the internet. We will rely on information at https://en.wikipedia.org/wiki/List_of_United_States_television_markets for that purpose.

Overview
We will focus on the advertising dataset referenced above and supplement the data from that dataset with data from the votes dataset, also referenced above. Our goal is to identify patterns and trends in the advertising dataset and consider how our observations relate to the election results.

We will focus on the 2016 presidential general election as opposed to the primary elections. Moreover, we will look at the data related to the Democrat and Republican candidates, Hillary Clinton and Donald Trump, and do not anticipate analyzing data about third-party candidates. Our focus is on the general election, not the primary election). Accordingly, we plan to analyze the data for television ads aired after the end of the primaries, so we do not intend to consider information concerning ads that aired before July 2016.

We will take the advertising data at the market level, identify the relevant counties in each market, and aggregate county-level voting data from the votes dataset to investigate relationships between political advertising and election results.

We will focus on four primary areas of interest.
1. Volume of ads
2. Ad subject matter
3. Ad message type (positive vs. negative)
4. Timing of ads (i.e., day of the week, time of day, etc.)

First, we intend to look at the volume of ads per candidate.
Questions:
- Did one candidate air more ads than the other in the covered markets in the aggregate and each of the thirteen markets individually?
- Was the candidate who aired more ads in a market the candidate who won the vote in that market?
- Did a particular candidate begin airing more ads than the other in specific markets as election day drew nearer?

Visualizations:
- We intend to create a map visualization displaying the T.V. ad density in each of the thirteen markets using the advertising dataset. We plan to merge the votes dataset with the advertising dataset on the *"location"* fields.
- We plan to exhibit the change in ad volume by the candidate over time until election day. We will use the *"candidates"* and the *"date_created"* variables from the advertising dataset. We derived the *"date_created"* variable from the *"start_time"* data.

Second, we plan to analyze the subject matter identified in the ads.
Questions:
- Are there certain subjects each candidate covered more than others?

- Were certain subject matters covered in specific markets more than others?

Visualizations:

- We expect to display the number of ads for each candidate by subject by creating a stacked, horizontal bar chart. We plan to obtain this information by referencing the *"candidates"* and the *"subjects"* columns in our advertising dataset.

Third, we will look at the breakdown between positive and negative messages presented in the ads.
Questions:

- How did the candidates use positive and negative advertising?
- Did their use of positive and negative advertising vary by the market?

Visualizations:

- We plan to present the flow of message types, both negative and positive, by candidate, by subject matter. In the advertising dataset, we intend to obtain this information using the *"message"*, *"candidates"*, and *"sponsors"* as our variables.
- We could also create a bubble chart to show the proportions of positive versus negative ads, by candidate, by region. To do so, we will rely on the *"candidates"*, *"messages"*, and *"location"* variables from the advertising dataset.

Lastly, we will look at ad timing.
Questions:

- Did the candidates air more ads during a specific time of day?
- Were more ads shown on weekends or weekdays?
- Does the aired timing vary across markets?

Visualizations:

- We intend to generate line charts to exhibit the timing of ads airing on television. The *"start_time"* and *"end_time"* fields from the advertising dataset will be crucial for this analysis.
- A scatter plot could also be interesting to look at the time-related data. Again, the *"start_time"* and *"end_time"* columns in the advertising dataset include the information we need to analyze ad timing.

We will consider how any differences between the candidates in that analysis relate to the voting results in the relevant markets. The votes dataset will play a significant role in the insights we hope to procure from the data.

Initial Data Exploration

Below are a few value count results from our early exploration of the advertising dataset.

```
df.subjects.value_counts()
```

```
Women, Candidate Biography, Children                        4184
Candidate Biography, Children                               4015
Taxes, Candidate Biography, Terrorism, Jobs                 3902
Economy, Jobs, Federal Budget, Taxes, Families              3216
Candidate Biography                                         3139
                                                            ...
Campaign Finance, Candidate Biography, Legal Issues            1
Candidate Biography, Military, Foreign Policy                  1
Candidate Biography, Taxes, Terrorism, Jobs, Housing           1
Immigration, Children                                          1
Candidate Biography, Civil Rights, Immigration, Islam          1
```

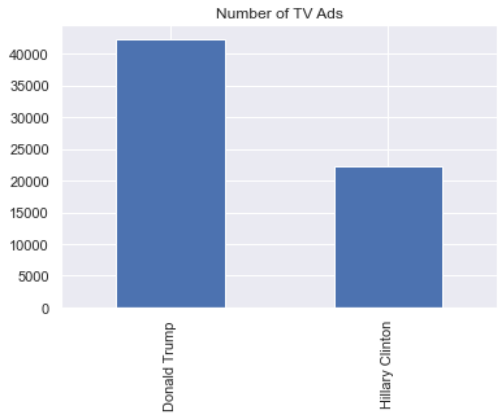
```
df.day_of_week.value_counts()
```

```
Tuesday      15883
Wednesday    14365
Thursday     14262
Friday       14241
Monday       13497
Saturday      7758
Sunday        7145
```

```
df.message.value_counts()
```

```
con          47752
mixed        22370
pro          15176
unknown       1853
```

We expect to have plenty of data to work with from the Political Ad Archive if we only focus on Hillary Clinton's and Donald Trump's television ads.



The thirteen regional markets are listed below. Additionally, below is a table snapshot of a preliminary version of our votes dataset that demonstrates how we map election results to the markets covered by the advertising dataset.

```
df.location.value_counts()
```

```
Tampa-St. Petersburg Region                13248
Philadelphia Region                        12412
Las Vegas Region                           10954
Cleveland Region                           10604
San Francisco-Oakland-San Jose Region       9025
Raleigh-Durham-Fayetteville Region          6733
Cedar Rapids-Waterloo-Iowa City Region      5295
Boston-Manchester Region                    5016
Denver Region                               4712
Milwaukee Region                            4101
Phoenix-Prescott Region                     3247
Washington DC-Hagerstown Region             1731
New York City Region                          48
```

```
df[df.year == 2016]
```

county_fips	office	candidate	party	candidatevotes	totalvotes	version	mode	location
4001.0	PRESIDENT	HILLARY CLINTON	DEMOCRAT	17083.0	27661.0	20191203	TOTAL	Phoenix-Prescott Region
4001.0	PRESIDENT	DONALD TRUMP	REPUBLICAN	8240.0	27661.0	20191203	TOTAL	Phoenix-Prescott Region
4001.0	PRESIDENT	OTHER	OTHER	2338.0	27661.0	20191203	TOTAL	Phoenix-Prescott Region
4005.0	PRESIDENT	HILLARY CLINTON	DEMOCRAT	32404.0	59784.0	20191203	TOTAL	Phoenix-Prescott Region
4005.0	PRESIDENT	DONALD TRUMP	REPUBLICAN	21108.0	59784.0	20191203	TOTAL	Phoenix-Prescott Region
...	...	...	...	...	...	...	...	...
56031.0	PRESIDENT	DONALD TRUMP	REPUBLICAN	3437.0	4529.0	20191203	TOTAL	Denver Region
56031.0	PRESIDENT	OTHER	OTHER	373.0	4529.0	20191203	TOTAL	Denver Region
8014.0	PRESIDENT	HILLARY CLINTON	DEMOCRAT	19731.0	37689.0	20191203	TOTAL	Denver Region
8014.0	PRESIDENT	DONALD TRUMP	REPUBLICAN	14367.0	37689.0	20191203	TOTAL	Denver Region
8014.0	PRESIDENT	OTHER	OTHER	3591.0	37689.0	20191203	TOTAL	Denver Region

The plot below shows there may be additional insights from analyzing secondary variables such as ad durations.