Final Report:2016 Presidential ad campaign data analysis

Explorers: Anamika Sinha, Chet Gutwein, Nathaniel Velarde

Introduction

The 2016 presidential race was unique in many regards. We had many non-conventional factors in play such as the candidates themselves, the significant impact of social media, and the advent of fake news to name a few. Having said that, we still cannot rule out the impact of conventional factors like gender, education, immigration, race, income gap, etc. Political parties and sponsors poured huge amounts of money into TV advertisements to influence voters in addition to spending significant dollars on social media ads, rallies and grassroot outreach. The 2016 race also left its mark as one of the most negative campaigns in history. Given the tremendous amount of emotions and passion generated by one of the most contentious elections in our country's history, we set out to dig deeper into one aspect of this election. Could we identify some key insights by looking into the TV ad trends and analyzing if the ad spends had any bearing on the outcome of the election?

We set out to explore the TV ad campaign data for the 11 battleground states of Arizona, Colorado, Iowa, Nevada, New Hampshire, North Carolina, Florida, Ohio, Pennsylvania, Virginia & Wisconsin.

Research Questions

Our group worked on 3 major lines of inquiry with differing levels of perspective:

1) At a high level, we tried to understand the data in terms of ad volume by the two parties, type of messages and what these messages said.

a. We did a comparison of ad volume in each of the battleground states
b. Which candidate aired the most ads in a state and whether that had a correlation with the final outcome?
c. What was the mix of positive and negative ads at the state level by both parties?
d. What were the key topics mentioned in these ads?

2) State by State Time Series Analysis – Combine Dataset with RCP State Poll of Polls Data

a. We looked at the evolution over time of ad volume and message mix (% negative ads) in response to shifts in poll data in each state. This was done to get some insights into the ad strategy adopted by both parties and how it changed over time.

b. We looked at evolution of who sponsored the ads and what type of message those sponsors conveyed in their ads.

3) At the opposite end of the spectrum, we looked at individual ad level data by taking advantage of the fact that each ad in the archive has a unique ID, to analyze the following:

a. Which ads were featured most by each candidate? What made these ads more frequently aired in comparison to other ads?

b. How unique ads were aired in different markets? How did airings of unique ads evolve throughout the campaign?

c. We also analyzed whether ads being fact-checked mattered by leveraging the fact-check grade data provided by the archive.

Our data:

Primary Dataset: 2016 Election Data (http://politicaladarchive.org/data/)

The dataset had information about political ads airings, with details about when and where the ads aired in select markets, which were primarily battleground states, in the 2016 elections. Below is the interpretation of some key fields:

Field_name	Description	Sample value	Interpretation
airing_id	This is the unique identifier for a specific airing of a political ad	1	To get count of ads(top ads aired)
market	Name of TV broadcast market, based on Nielsen Market names.	Raleigh-Durham-Fayetteville, NC	To get state info by splitting
start_time, end_time	Date/time ad aired, start.("coordinated universal time) Date/time ad aired, end.	2016-09-09 00:12:59 | 2016-09-09 00:13:29	To get ad duration, weekday of ad airing and filter (8/1/2016
archive_id	A unique alphanumeric id for each ad identified	PolAd_HillaryClinton_f1h3j	To join with unique ads dataset
embed_url	Url for embedding ad.	https://archive.org/embed/PolAd_HillaryClinton_f1h3js	To watch some of the unique ads
sponsor_types	Candidate committee, Super PAC, 501(c), 527 etc	PAC	To find out the sponsor type to ad count ratio
race	The federal race the ad is targeted toward.	"PRES" "COS1"	Filter by "PRES" as senate race was also present
cycle	Election cycle, i.e. 2016 = the 2015-2016 elections.	2016	Filter by 2016
subjects	Subjects covered in ad; subject index from PolitiFact	"Energy, China, Jobs","Terrorism, Veterans"," Military"	Get a frequency count of subjects

candidates	Candidate(s) named in ad	Jeb Bush Marco Rubio	Data validity check.
type	Campaign ad, issue ad, unknown	"campaign", "unknown"	Filter by campaign
message	Describes how the ad mentions the candidate.	"pro","con", "mixed"	To find out if ad was negative, positive

Additional datasets

1. Get victory margins for republicans and democrats for 2012 & 2016 in battleground states http://uselectionatlas.org/RESULTS/data.php?year=2016&datatype=national&def=1&f=0&off=0&elect=0
2. Get state RCP polling data http://www.realclearpolitics.com/epolls/2016/president/oh/ohio_trump_vs_clinton_vs_johnson_vs_stein-5970.html
3. Unique Ad Data - also from the Political Ad Archive http://politicaladarchive.org/api/v1/ads?output=csv

Initial Data Exploration and Cleansing

Cleaning Primary Dataset:

The original dataset had 364,718 rows and 19 columns. Each row represented a unique ad airing. Some of the key fields that had missing values were location, race, cycle, subjects, candidates, sponsors, sponsor types.

Race and Date fields:

The dataset contained ad airing dates from September 2014 onwards starting with state senate, presidential primaries race and finally the presidential race in November 2016. We decided to limit the scope of our inquiry to the presidential general election. Thus, we excluded (using boolean selection) all ads that aired before the end of the Republican and Democratic National Conventions. As the Democratic National Convention concluded on July 28, 2016, we set August 1, 2016 as the start date for our dataset. Excluding dates before August 1, 2016 reduced the number of observations by 163,273 or ~45%.

As a sanity check, we calculated the minimum and maximum air dates. Surprisingly, we found ads that aired after November 8, 2016 somehow existed within the dataset. We believe this was due to contests that ran beyond Election Day (e.g., special run-off elections). In order to exclude these we put an end date of November 8, 2016 as the upper limit to our filter. This removed an additional 162 observations.

At this point in the cleaning process, our dataset still included ads for state-level races. We used the race field to only select ads designated as 'PRES'. This mask removed another 116,131 observations. We then did a sanity check that ad end date/time was always greater than ad start date/time.

"Location" field:

We noticed locations like 'Boston, MA/Manchester', ' DC/Hagerstown, MD', 'New York City, NY' and 'San Francisco-Oakland-San Jose, CA' that were not located in battleground states. A little digging led us to the following conclusions. Ads aired in the Boston media market was most likely used to target southern New Hampshire voters. Similarly, Washington, DC media buys were likely used to reach voters in northern Virginia. We were not sure why the dataset had ads aired in New York City. These ads could have been targeted at eastern Pennsylvania voters. We chose to ignore these airings as there were only 47 instances. We also changed CA to "National" as the dataset documentation indicated that these ads represented national network buys.

"Candidate" field:

We were using this field to determine who the ad was for until we did a sampling of the data. The data revealed Donald Trump and Hillary Clinton together as values in certain scenarios which made us realize that this field could not be used to determine which candidate was behind the ad. To get past this issue, we researched each of the 45 unique values in the sponsor field using OpenSecrets.org and Wikipedia to determine which candidate they were supporting. Surprisingly, we found that some of the sponsors for non-presidential races (eg., Strickland for

Senate). We classified these sponsors as 'Other' so that we could remove these observations (1,778 in total) using boolean selection. We created a mapping dictionary where the keys were the sponsors and the values our 'sponsor classes' - Clinton Campaign, Democratic PAC, Democratic National Committee, Trump Campaign, Republican PAC, Republican National Committee, Johnson, Stein and Other. We used the '.map()' method in conjunction with our mapping dictionary to create a new column called 'sponsor_class'. We used the same general process to create a higher level variable called 'party'. Using these new columns, we were able to associate every ad with one of the two contestants. For validity check, some of the ad content was watched to see if the candidate field had the opposing candidate's name if the message was "con" and the sponsoring candidate name if the message was "pro".

Subjects:
72 unique topics were mentioned in ads. Top three subjects mentioned in ads were candidate biography, jobs and children.

After removing irrelevant data, the dataset that we worked with had 83,303 rows. Now our working dataset was about 23% of the size of the original dataset.

Cleaning the Unique_ads dataset:

Reference_count:

The reference_count attribute (aka fact-check score) was assigned to each unique advertisement per the Political Ad Archive:

"number of fact or source checks from our partner organizations for this particular ad. For example, the claim that Donald Trump once supported impeaching former President George W. Bush, contained in this ad sponsored by Our Principles PAC, a super PAC opposing Trump, was fact checked by PolitiFact, which rated it as "True." The PolitiFact story is embedded on the Political TV Ad Archive page displaying the ad."

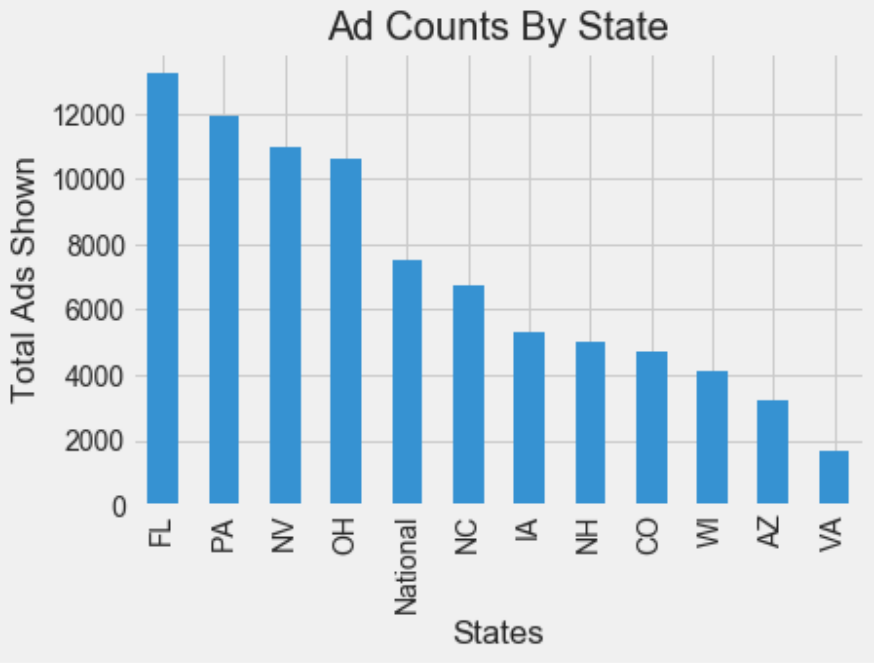
The following steps were taken to quantify whether or not ads were held accountable for being factual:

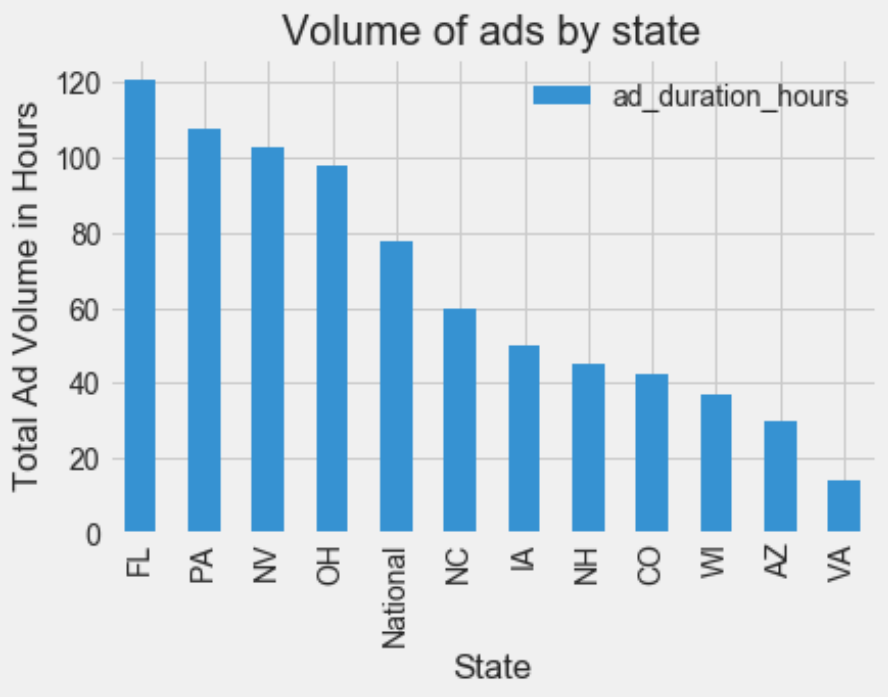
- Apply the unique ad fact-check grade to each airing from the master ad dataset
- Create a new column in the master ad dataframe 'fc_min' for "fact-check-minutes" by multiplying the fact-check grade by the ad length in seconds and dividing by 60

Our Data Story

1) Summary Analysis

a. We started by looking at the summary level of ad counts in the battleground states. How did the states line up in terms of ad count numbers and sheer volume of hours? Given the wide variability in ad length, we took into account ad volume in addition to ad count for each state.

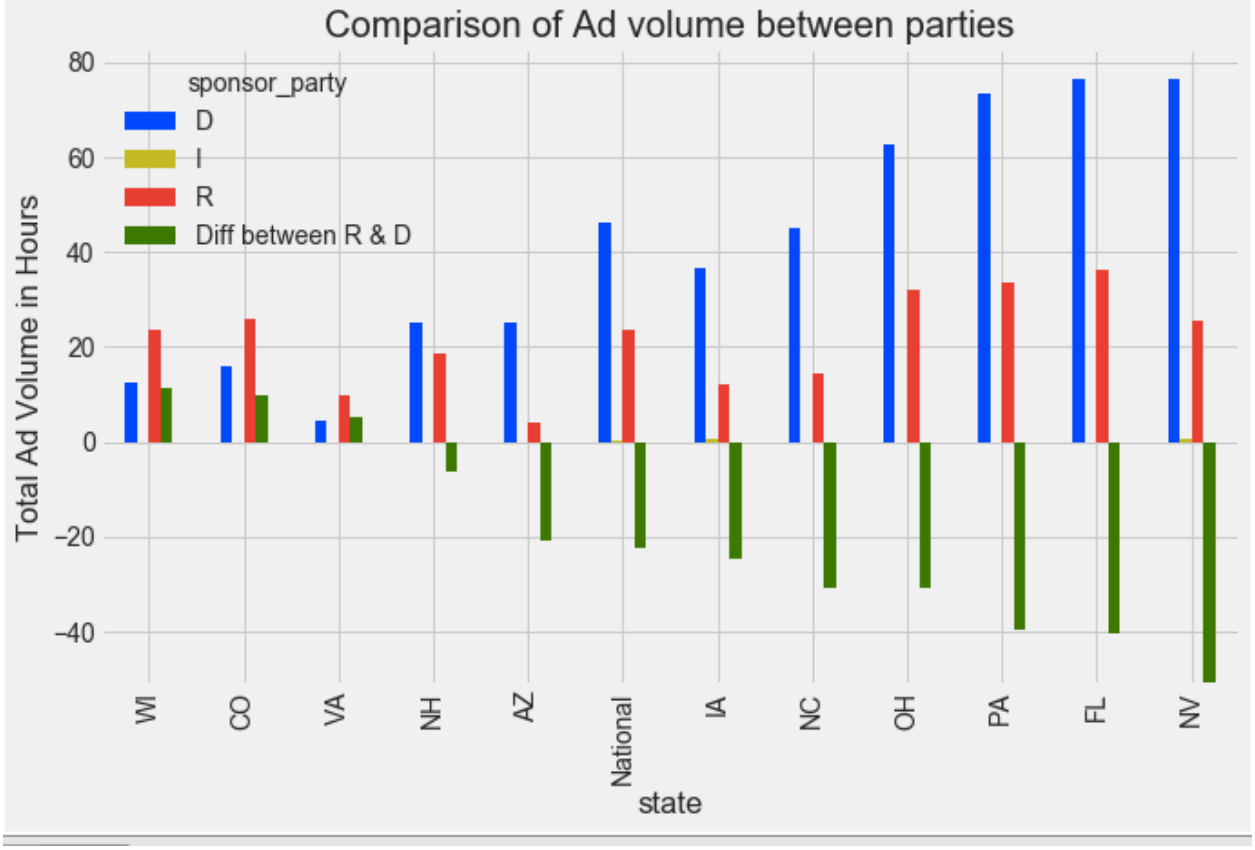




Takeaway: It was interesting to note that the states lined up exactly in the same order.

Florida was subjected to most ads while Virginia was spared for the most part. Given the swing nature of the state, the high number of electoral votes it has and the decisive role it has played in past elections (we all remember the hanging chad from the Al Gore vs George Bush campaign), it seems to have rightly earned the most attention for ads.

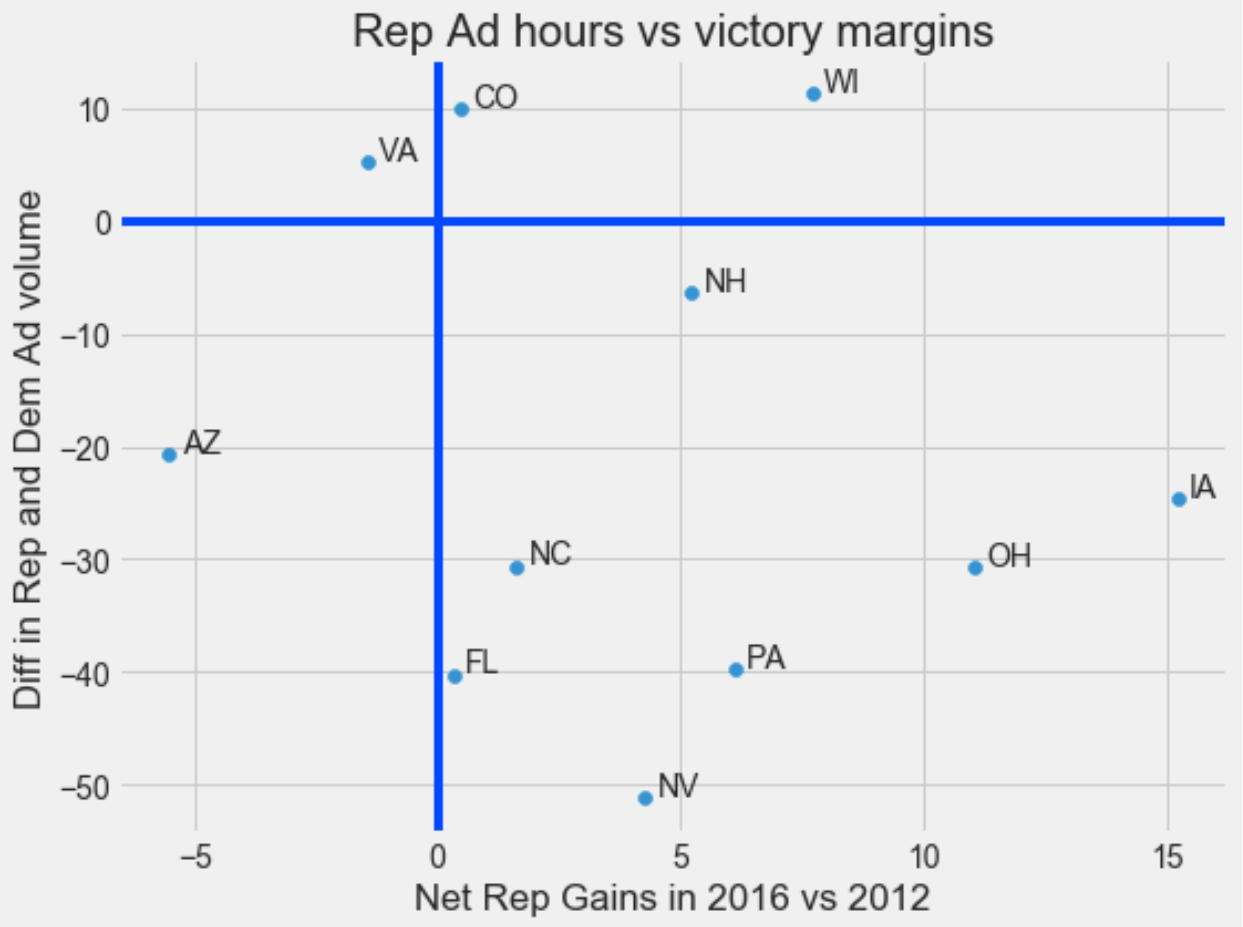
b. We decided to dig further by dissecting the ad volume by party.



Takeaway - Interestingly, democratic party outspent the republican party in 8 out of the 11 battleground states.

Was there any correlation to victory margins from 2012 elections?

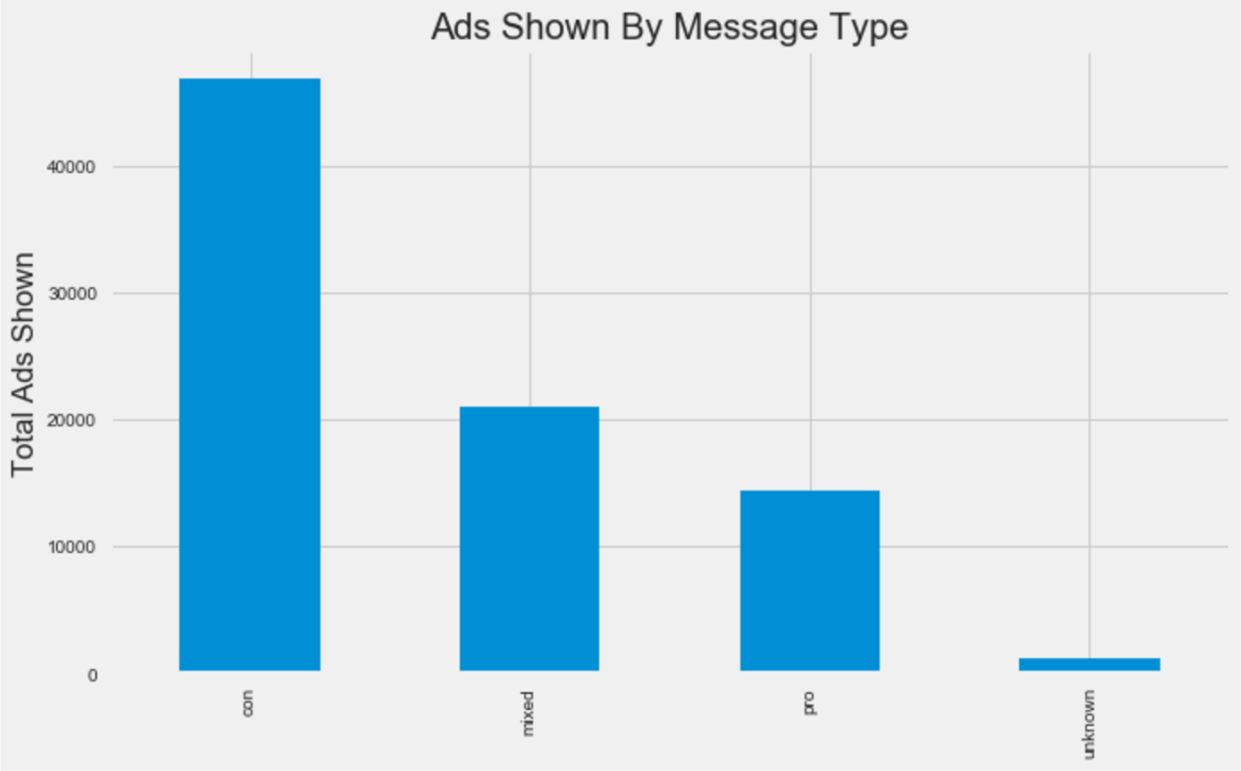
In order to analyze this question, we referred to an external dataset and collected data for 2012 presidential elections for each of these states and compared it to the 2016 presidential election results for the same states. We plotted the the net change in republican margins from 2012 to 2016 (Rep margin of victory/defeat in 2012 - Rep margin of victory/defeat in 2016 with a positive number indicating a net improvement in Rep margins, whether in victory or defeat for a given state) - and compared it to the ad volume difference (Rep-Dem for ad volume with a negative number indicating that the Dem had more ad volume than Rep in a given state).



We saw that in seven states the Republican net margins increased despite lower ad volumes compared to Democrats. In two states, their net margins increased with more ad volume. In Virginia, their victory margins fell despite more ads. In Arizona, the margins decreased with less ad volume. This led us to conclude that there isn't a strong correlation between ad volume and election results. One can make an argument that there was a negative correlation given that net margins in seven states increased despite lower ad volume. Using the pearson correlation method, we got a correlation factor of -0.14. We could have made a stronger case if we had access to 2012 ad campaign data.

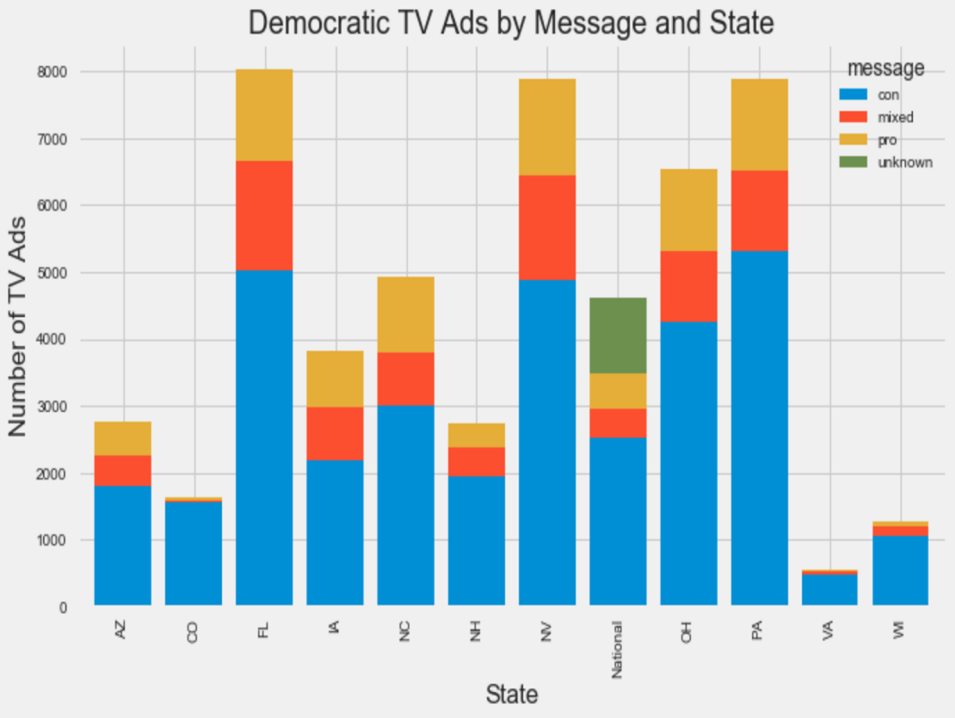
c. To look at the data from another perspective, we turned our attention to the tone of message in these ads.

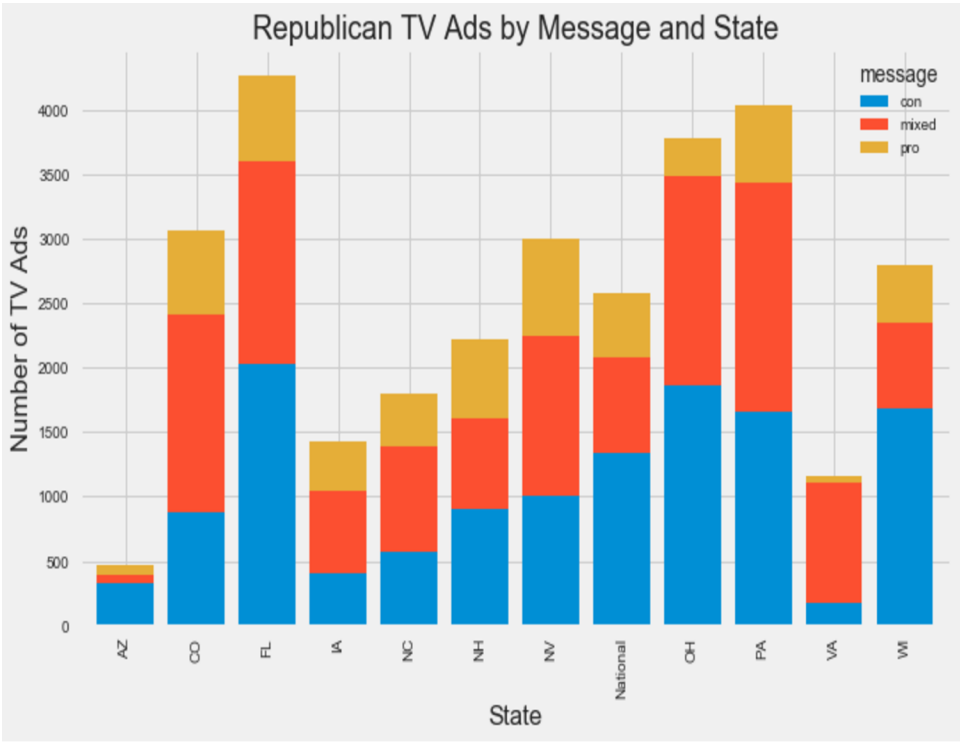
Ads Shown by Message Type



We saw a very negative tone to campaign. This chart shows ads by message type for the dataset as a whole. We took it to the next level by state and party.

Democratic & Republican TV Ads by State and Message

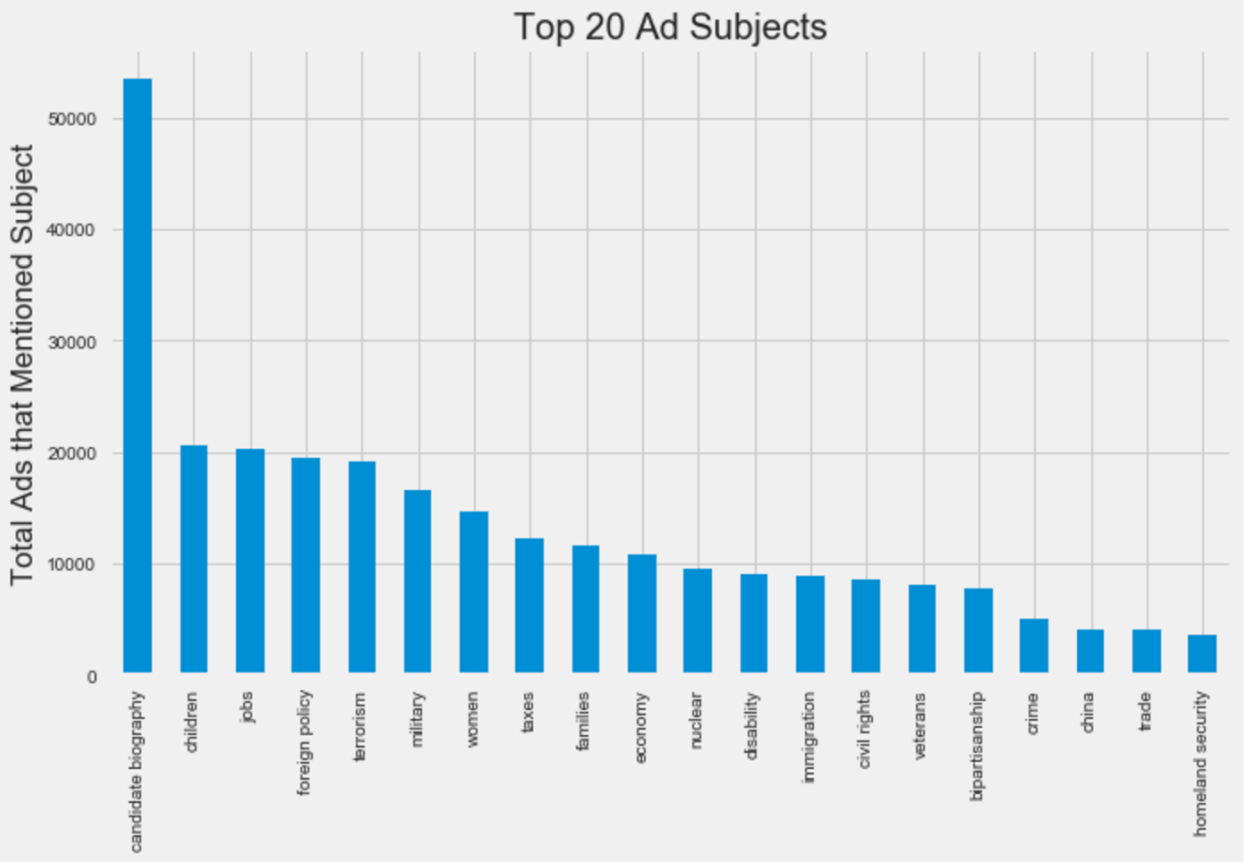




Takeaway: We notice a high proportion of negative messages in Democratic ads compared a more balanced proportion of mixed and negative messages in Republican ads.

Top 20 Ad Subjects

We enquired into the content of the ads

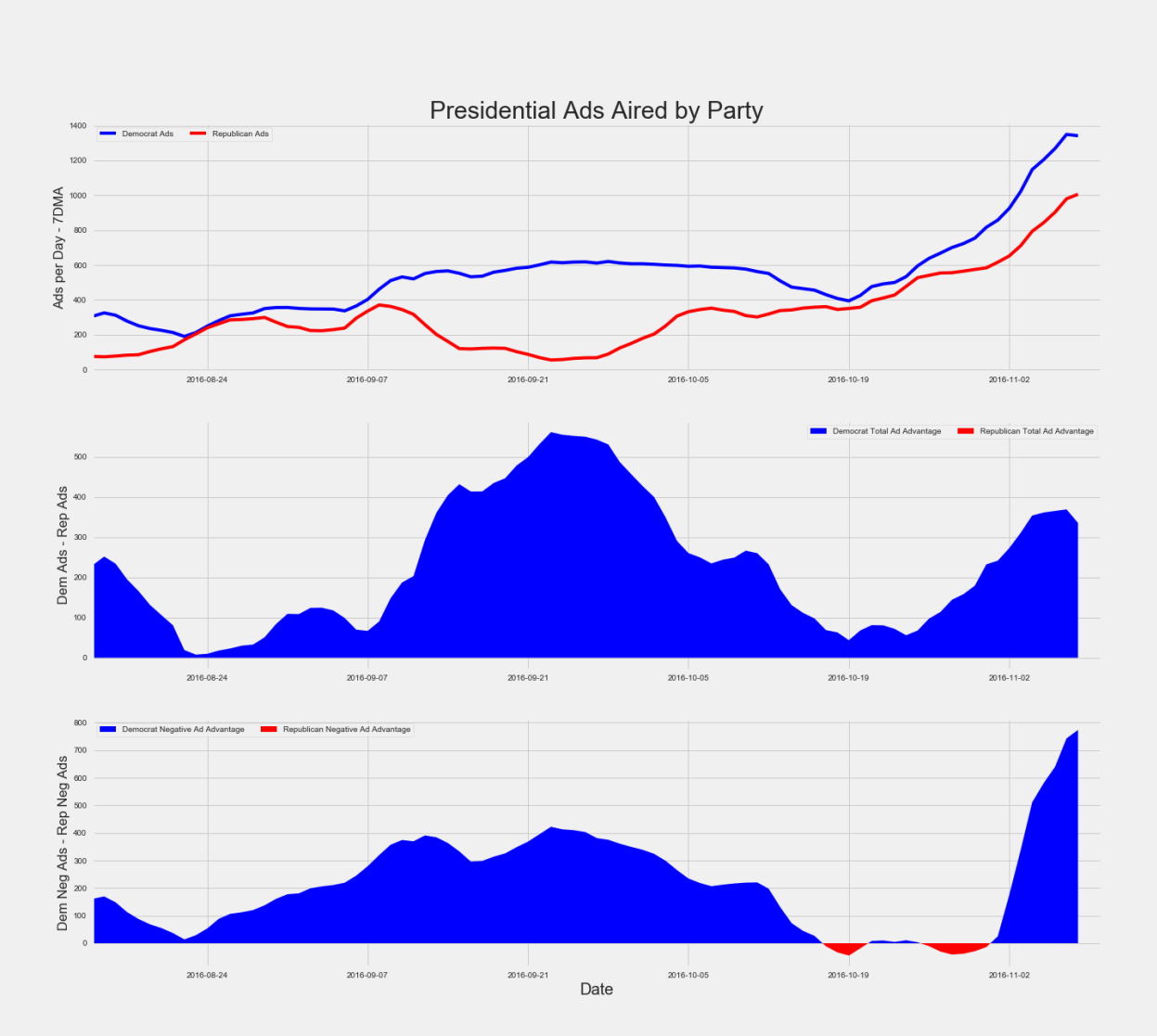


Takeaway: Given their share of press coverage, it is interesting how far down the list Immigration/Trade/China were. From a methodology standpoint, most ads mentioned multiple subjects (on average, 3). The chart shows the total number of ads where the subject/topic was mentioned. This ad subject ranking is likely skewed towards topics mentioned in Democratic ads given their advantage in number of airings.

2) Second level Analysis - Time Series

The next stage in our analysis was to conduct time series analyses on our cleaned dataset to see how the candidates adjusted their ad volumes and message mix to shifts in polls and major events (eg., debates) during the campaign. We looked at time series trends on the dataset as a whole as well as on a state-by-state level.

a. How Ad Volume Evolved During the General Election



This chart shows the average ads per day (7 day moving average) aired by the Republicans and Democrats from August 1st to Election Day in the states/media markets covered by the dataset.

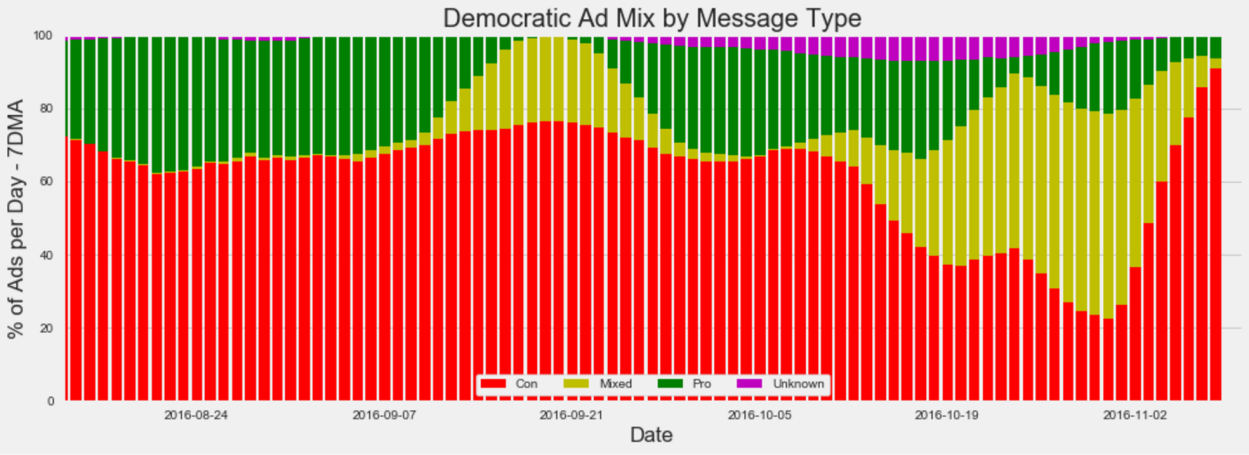
Some takeaways:

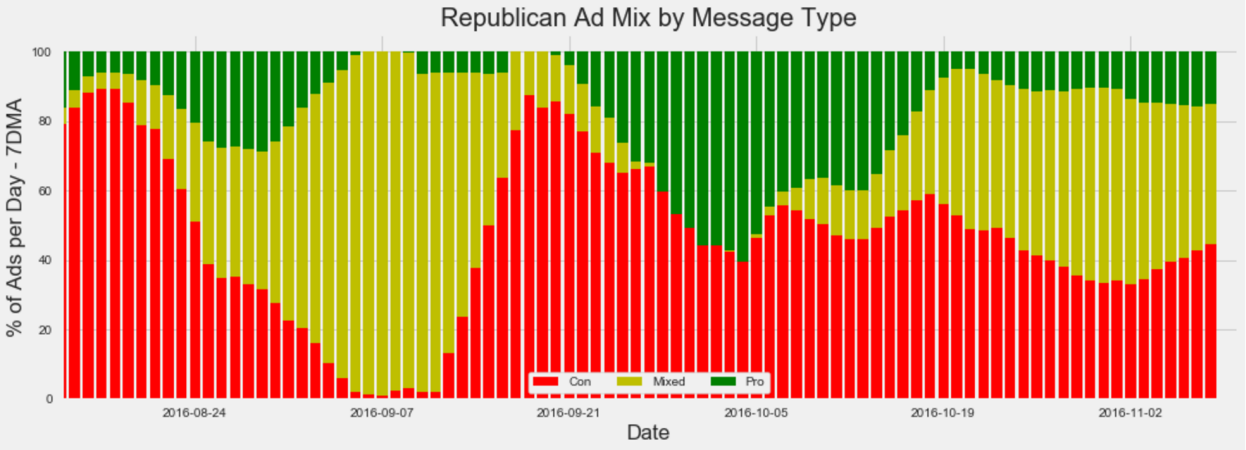
- The top panel of the chart shows how ad frequency increased, especially in the lead-up to the election.
- The middle panel shows the advantage Clinton had in terms of ad volume versus Trump – at no point from August 14, 2016 did Clinton air less ads than Trump.
- The bottom chart shows the negative ad advantage that Clinton had in terms of ad volume and how it peaked towards the end of the campaign. Trump did have a small advantage in negative ads twice during the campaign cycle.

This graph leads to some interesting questions:

- Were Trump's ads really that much more effective to make up for the disparity in ad numbers and/or did Clinton allocate too much campaign resources to TV in general (and in the wrong states in particular), when other channels (social media, internet ads, radio) would have been more effective (interesting to know if we had access to both campaign budgets).
- Is TV's yesterday's way of reaching a broad audience given the well-documented fragmentation of television audiences as evidenced by the shift in power and viewership from networks, to cable, to streaming services, YouTube, etc?

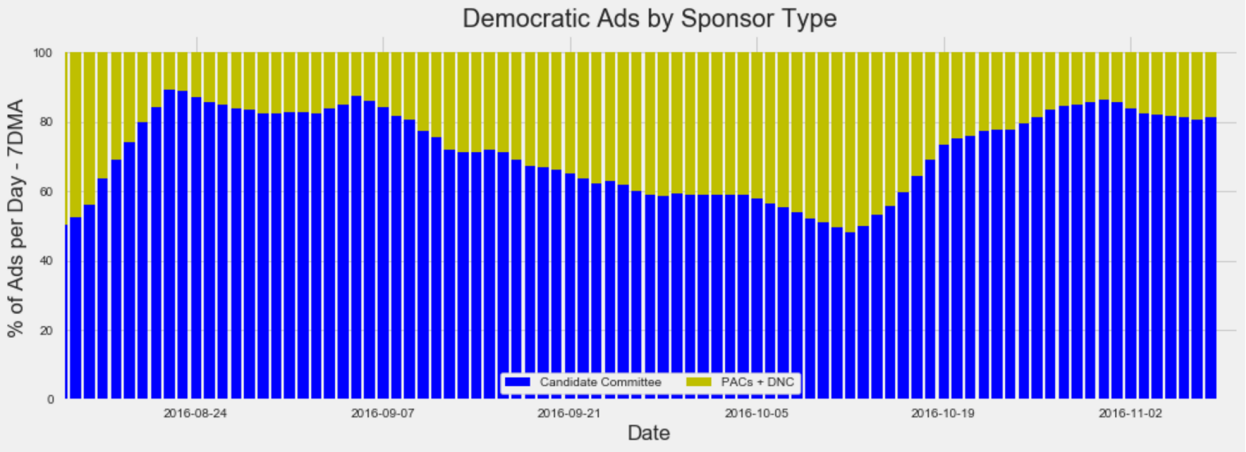
How Ad Message Mix Evolved During the General Election

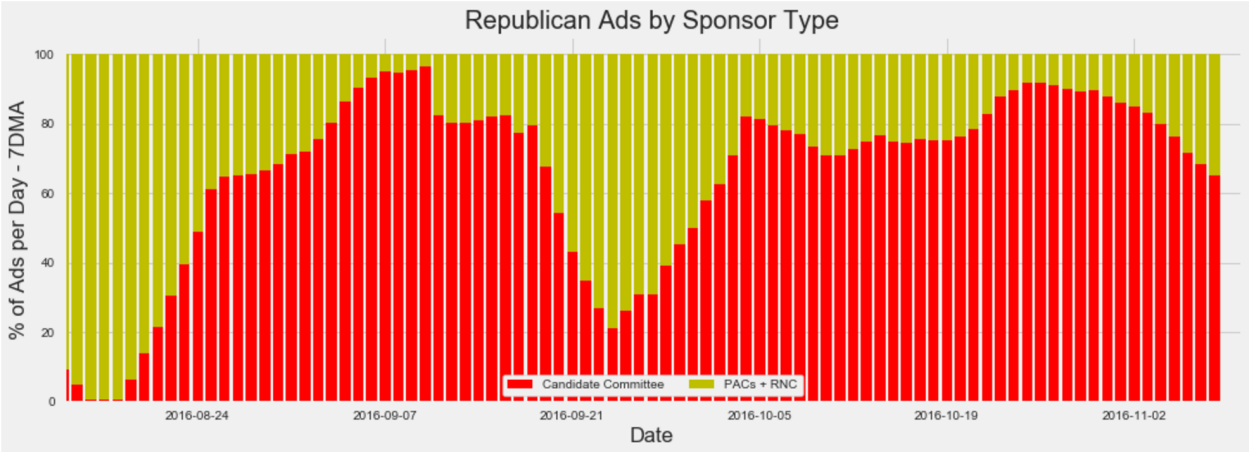




These charts show the evolution of message mix for each candidate. Relative to Republicans, Democratic ads skewed more negative (61% on average versus 47%), which was contrary to the Democratic campaign rhetoric of "When they go low, we go high." Democratic ads were extremely negative during the last week of the campaign.

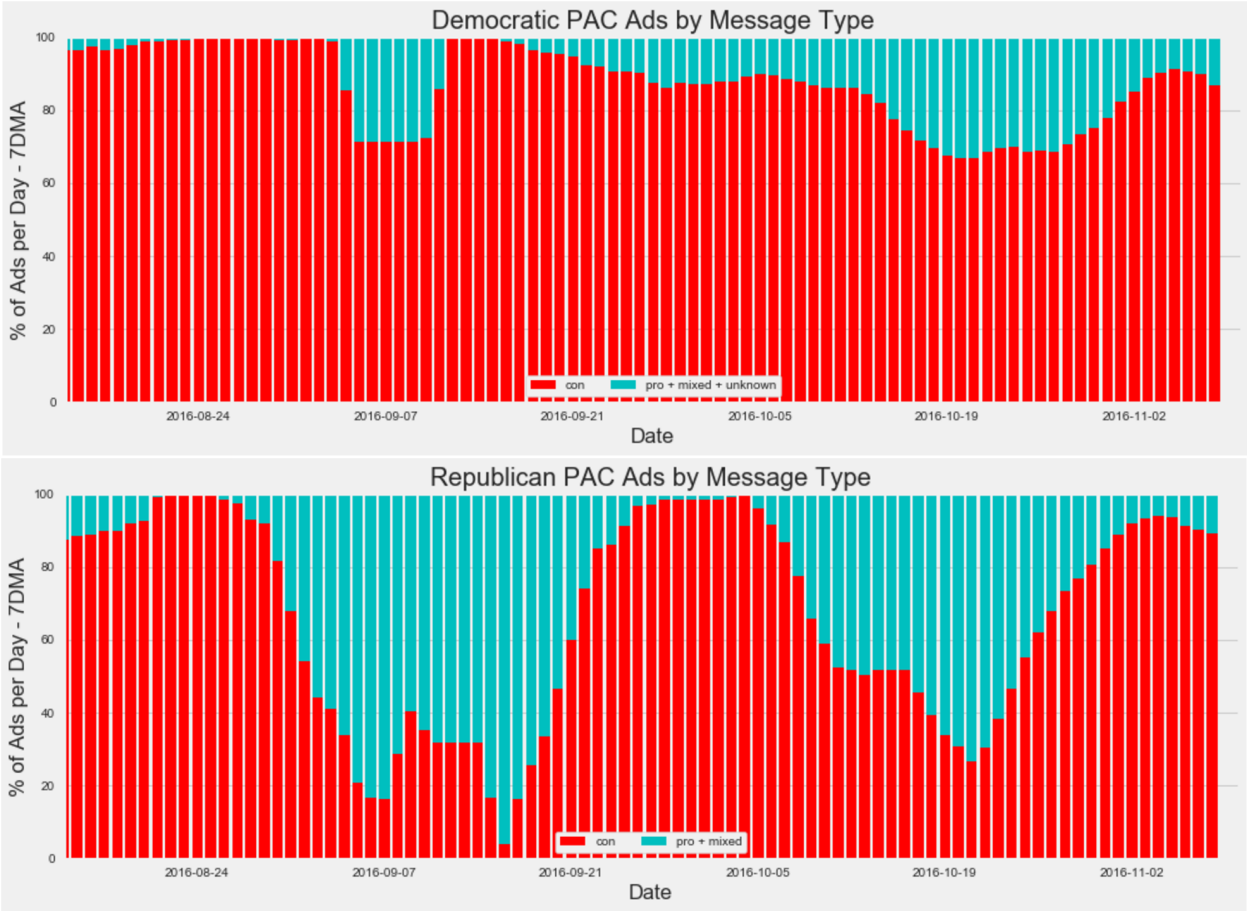
b. How Important Were PAC-Sponsored Ads to the Campaigns?





Republican PACs sponsored more of their candidate's ads than their Democratic counterparts (35% to 28%). Republican PAC support for Trump was also more volatile.

PAC Ads = Attack Ads



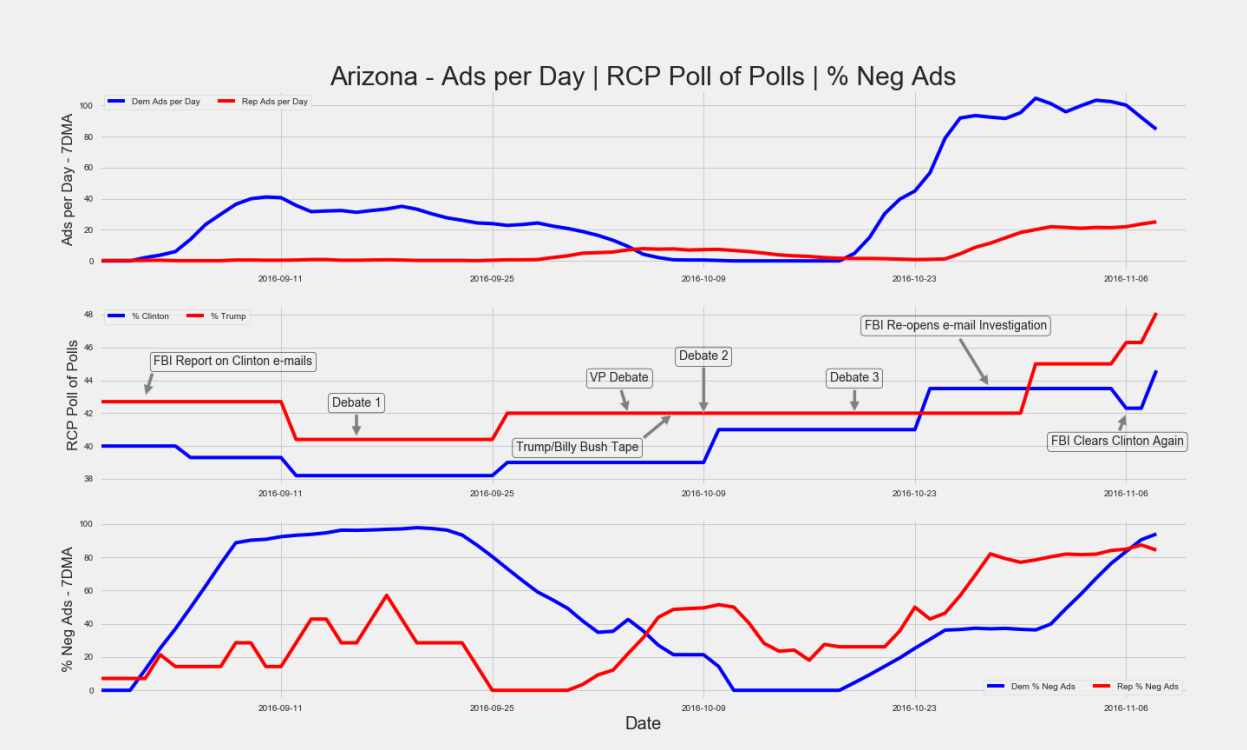
87% of Democratic PAC-sponsored ads were negative versus 67% of Republican PAC ads. PACs accounted for 42% of all negative Democratic ads aired versus 62% for all negative Republican ads. Broadly speaking, Clinton dished the dirt herself, whereas Trump relied more on his PAC partners to throw mud. As measured by ad mix, the Clinton's TV ad strategy could be simply defined as a vote for Clinton was a vote against Trump. People are more motivated to vote for something rather than against. Perhaps this is a reason why Clinton's seemingly commanding lead in the polls did not translate into victory on election day.

c. State-by-State Time Series Analysis

While we conducted time series analysis on all the battleground states in our dataset, Arizona and Wisconsin represented the most interesting case studies as the campaigns respective strategies were strikingly different in these states. They each had very different interpretations

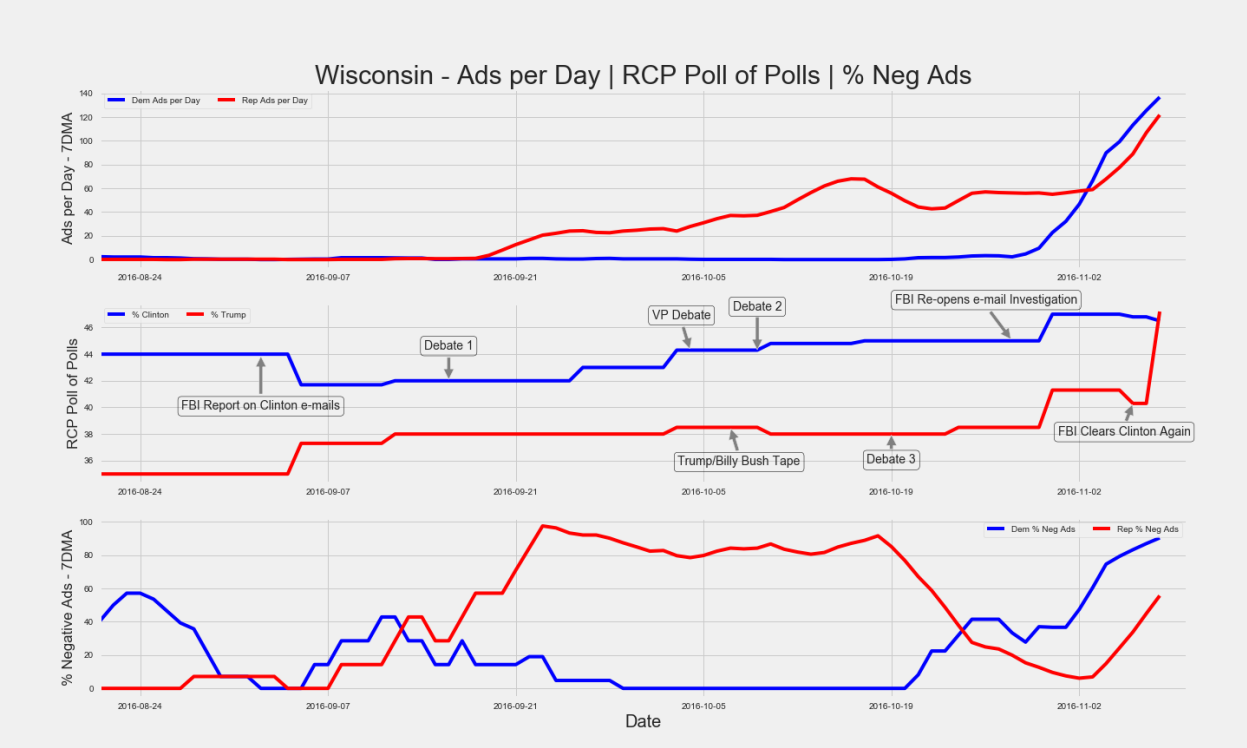
of the same polling data as evidenced by the pattern of their TV ad airings. Clinton dedicated time and resources to win a state, Arizona, that was not required for victory while taking a state that was essential, Wisconsin, for granted for much of the campaign.

Arizona



The top panel of the chart shows the surge in Clinton ads during the last weeks of the campaign in response to polls that indicated that she could swing a reliably red state to blue. Assuming the pre-election polls in the other battleground states were accurate, without Arizona, Trump would have no route to 270 Electoral College votes even if he won the key states of Ohio and Florida. The most interesting aspect of this chart was that there was no commensurate response from the Trump campaign to the late Clinton TV ad surge. Either they had confidence in their own internal polling that said that the state was safe despite Clinton taking a lead in the public polls or they lacked the resources to respond.

Wisconsin



Wisconsin was one of only three states where Trump aired more TV ads than Clinton. From the top panel of the chart, it appears that Clinton took Wisconsin for granted, not airing many ads until the last stage of the campaign. After all, Wisconsin was a key brick in the Democrats "Blue Wall" that made a Republican victory in the Electoral College an uphill battle since 1992. The polls predicted a comfortable Clinton win (the final Wisconsin RCP poll of polls on November 2nd had Clinton winning by 6.5% versus an average margin of error of 3.7%) indicating that the Democrats' Wisconsin strategy was sound and their "Blue Wall" intact. Obviously, the pollsters and the Clinton campaign missed a significant block of voters that Trump eventually captured to win Wisconsin by 0.7%. A similar pattern occurred in the other "Blue Wall" states of Michigan and Pennsylvania, where Clinton had poll leads of 3.5% and 1.5%, respectively, going into the last days of the campaign (important to note that these leads were within the average margin of error).

3) The Third level of Analysis - Individual Ads

"Why would we look at the ads at a granular level?" one might ask. The trends and insights that we've gleaned thus far may very well be consistent with individual ads, however, it's important to see if the analysis is heavily weighted by just a handful of individual advertisements, and if so why is that the case. On a lighter note, since most, if not all of us are part of the target audience of these advertisements, it's just interesting to see why we ended up on the viewing end of certain ads.

Most Featured Ads

A table below shows a table of the top 10 advertisements aired by each candidate along with key attributes of the ads:

Democratic Ad Summary:

	sponsors	subjects	message	date_ingested
1	Hillary for America	Candidate Biography, Children	con	2016/08/08 5:26:26 UTC
2	Hillary for America	Candidate Biography, Terrorism, Military, Fore...	con	2016/08/22 10:57:06 UTC
3	Hillary for America	Children, Candidate Biography	pro	2016/09/22 4:27:55 UTC
4	Priorities USA Action	Disability, Candidate Biography	con	2016/06/27 7:39:07 UTC
5	Hillary for America	Nuclear, Candidate Biography, Military, Foreig...	con	2016/10/08 11:11:39 UTC
6	Hillary for America	Disability, Bipartisanship, Children	mixed	2016/10/11 4:12:11 UTC
7	Hillary for America	Women, Candidate Biography, Children	con	2016/09/23 4:51:21 UTC
8	Priorities USA Action	Military, Nuclear, Terrorism	con	2016/09/07 3:24:48 UTC
9	Hillary for America	Energy, China, Jobs	pro	2016/09/12 2:12:03 UTC
10	Hillary for America	Women, Candidate Biography, Children	con	2016/11/01 10:20:20 UTC

Certain points that this table reiterates:

- The high concentration of negatives ads
- "Candidate Biography" dominated as subject in 80% of these ads
- Hillary Clinton sponsored most of her ads

Republican Ad Summary:

	sponsors	subjects	message	date_ingested
1	Donald J Trump For President	Taxes, Candidate Biography, Terrorism, Jobs	mixed	2016/10/20 12:24:34 UTC
2	Donald J Trump For President	Economy, Jobs, Federal Budget, Taxes, Families	mixed	2016/09/07 10:22:13 UTC
3	Donald J Trump For President	Candidate Biography, Economy, Jobs, Crime, For...	mixed	2016/11/02 6:21:59 UTC
4	Donald J Trump For President	Candidate Biography, Terrorism, Islam, Civil R...	con	2016/10/25 4:01:22 UTC
5	Donald J Trump For President	Iran, Terrorism, Nuclear, Foreign Policy, Cand...	con	2016/11/04 3:13:48 UTC
6	Donald J Trump For President	Candidate Biography, Workers, Nuclear, Foreign...	con	2016/10/13 5:24:35 UTC
7	Donald J Trump For President	Immigration, Terrorism, Social Security, Crimi...	mixed	2016/08/19 5:19:51 UTC
8	Donald J Trump For President	Families, Taxes, Women, Workers, Children, Sma...	pro	2016/10/07 7:12:16 UTC
9	Donald J Trump For President	Homeland Security, Terrorism, Immigration, Fam...	pro	2016/08/26 2:56:07 UTC
10	Rebuilding America Now	Candidate Biography	con	2016/08/08 11:30:04 UTC

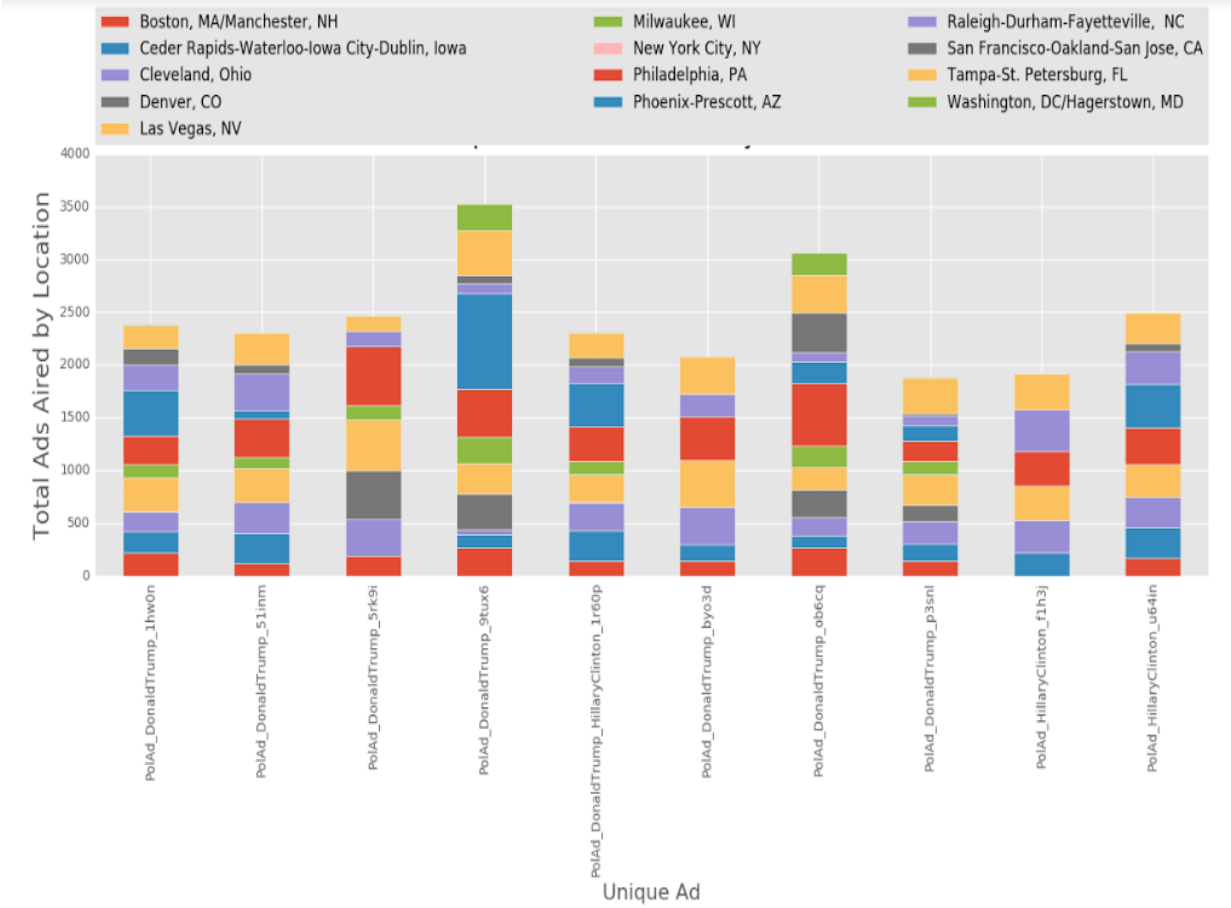
Takeaways from the table:

- The message of Republican ads seems to be more balanced with a nice blend of mixed, negative and positive as found in time series analysis of message type.

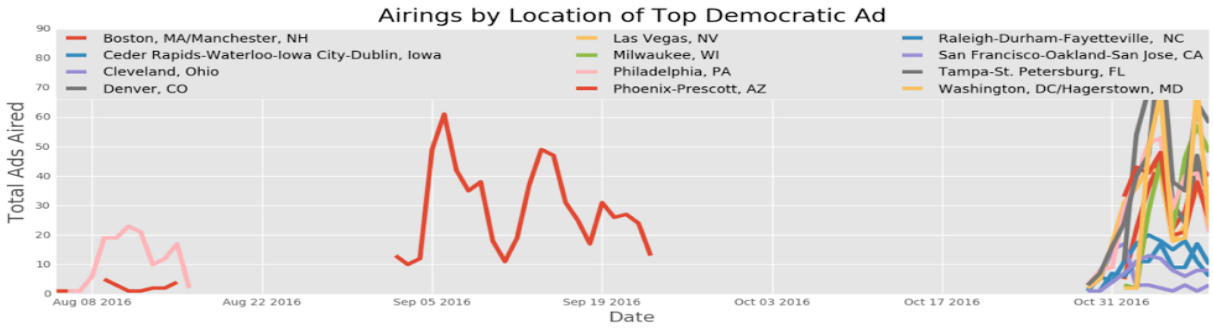
Unique Ads Aired by Location

Using the market data for each individual ad we looked at how the top ten ads were aired differently across different locations. This analysis yields some insight as to what each campaign's strategy was as they targeted voters in different areas of the country.

First, we looked at the **Democratic campaign**:



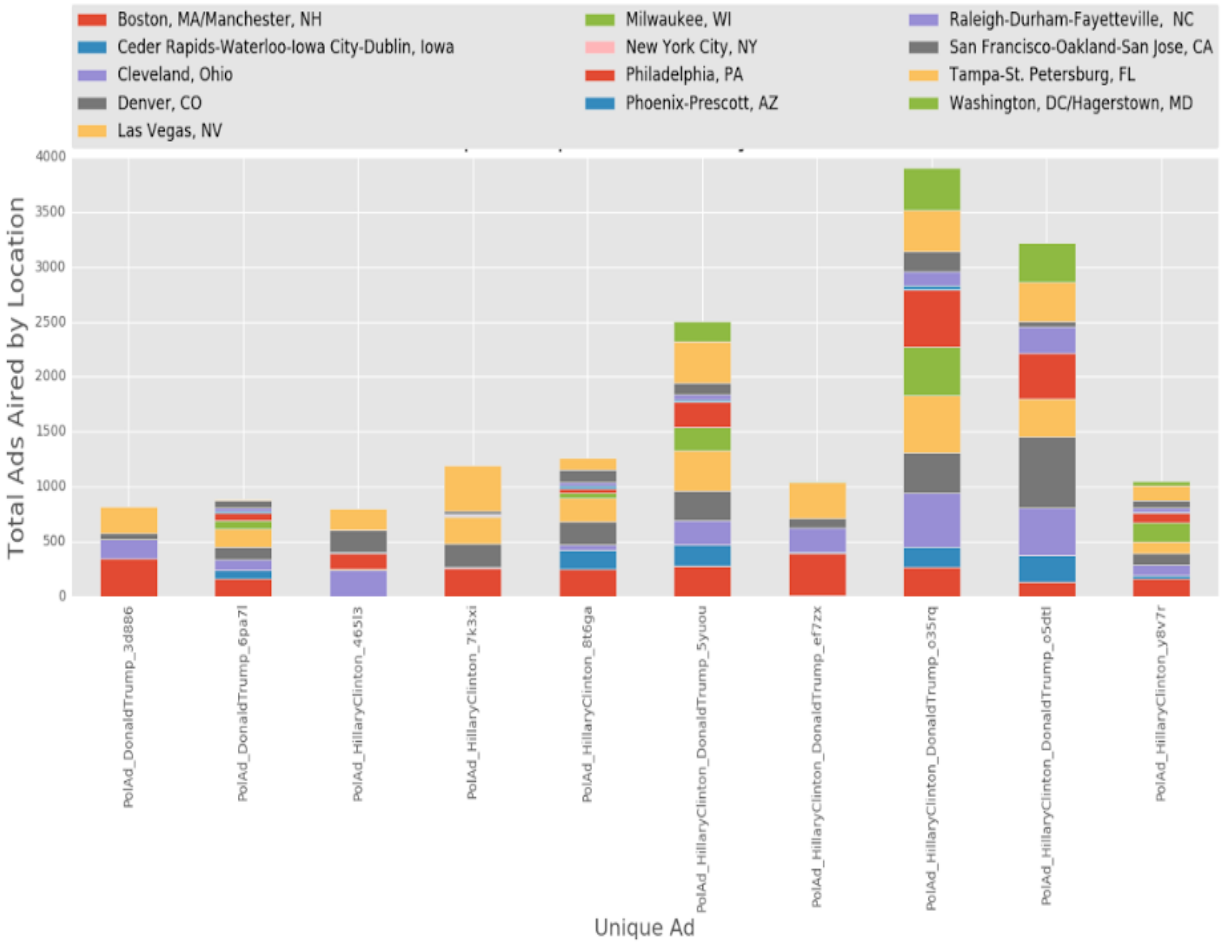
Takeaway: We saw that the top ad (PolAd_DonaldTrump_9tux6) was aired far more frequently in Arizona compared to other ads. We were intrigued by this and looked further into this advertisement to search for answers by doing a time series plot of this ad by location.



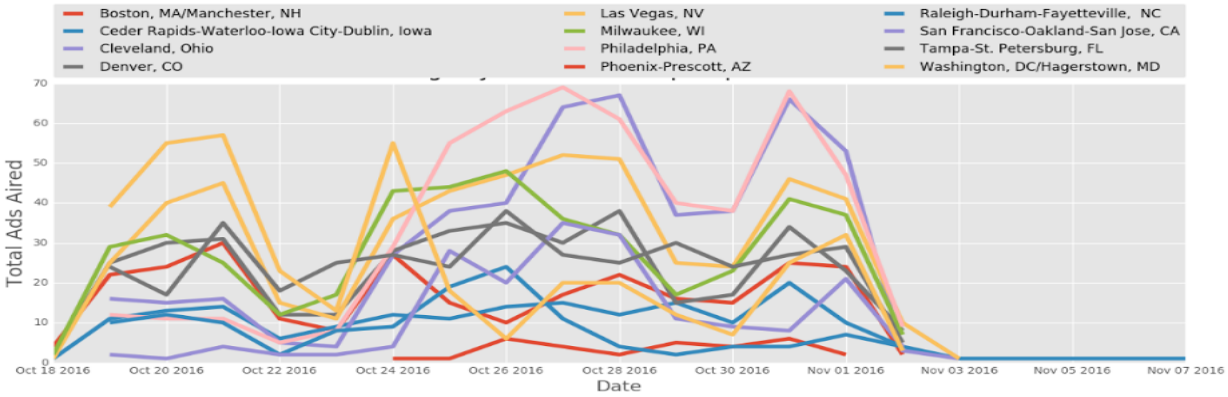
Finding: For about a three-week period in September, this ad was aired heavily in Arizona and nowhere else. Furthermore, the only other time it was aired was for a two-week period in August in the Philadelphia market and then much of airings were during the final week of the campaign before election day. Based on the ad usage, we suspect the advertisement was being

tested in the Philadelphia and Arizona markets to determine how strong of a response it received by voters prior to ramp up usage later in the campaign. When we looked at the polling data for this period in Arizona, we found that Hillary's ratings fell. We are not sure what was the democratic party strategy in using this ad as the top ad during the final week of the campaign.

Looking at the **Republican campaign**



For the Republican campaign, we saw that the most frequently aired ad (PolAd_HillaryClinton_DonaldTrump_o35rq) was one of only two ads that had significant airings in Wisconsin. We investigated further by using a time series plot as well:



Interestingly, the time series plot showed this ad was aired heavily up until the final week leading up to the election and then it was dropped in favor of different ads. The main advertisement for the Republican campaign over the last two weeks was the third most aired advertisement (PolAd_HillaryClinton_DonaldTrump_5yuou). This ad was ONLY aired during the month of November.

Bottomline, we did not find a clear strategy behind the top ad and what was the reason for its high usage. The Republican strategy seemed to have more uniform distribution over time and location unlike the Democratic strategy which had a huge spike in all locations for a one week period.

Unique Ads Fact-checking Analysis

Looking at reference_count

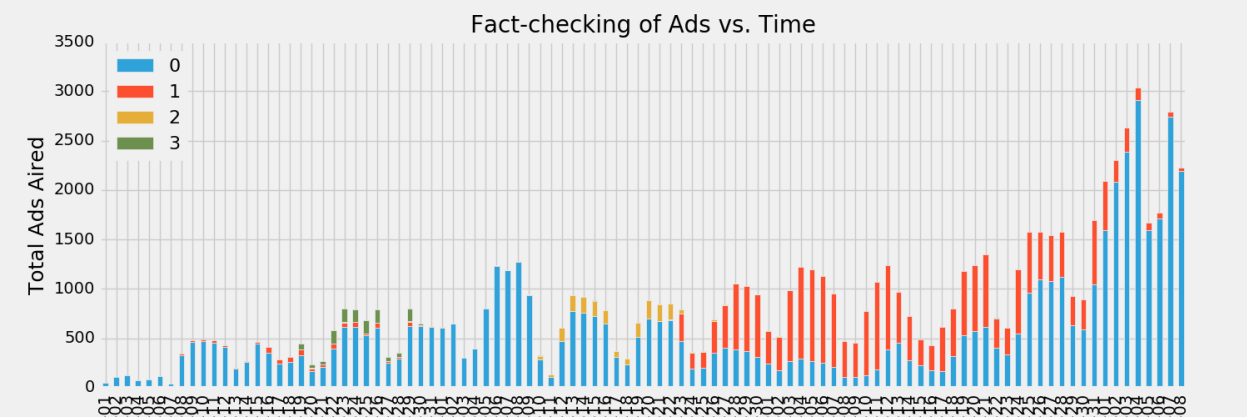
Running the pandas 'describe' command on the full dataset of unique ads shows us that there are many ads and very few have any fact check references as evidenced by the mean of 0.0468 references per unique ad.

We can see that when we filter for only ads that contain at least one fact check reference, there are 120 unique ads that were fact-checked, amounting to 4.22% of all unique ads.

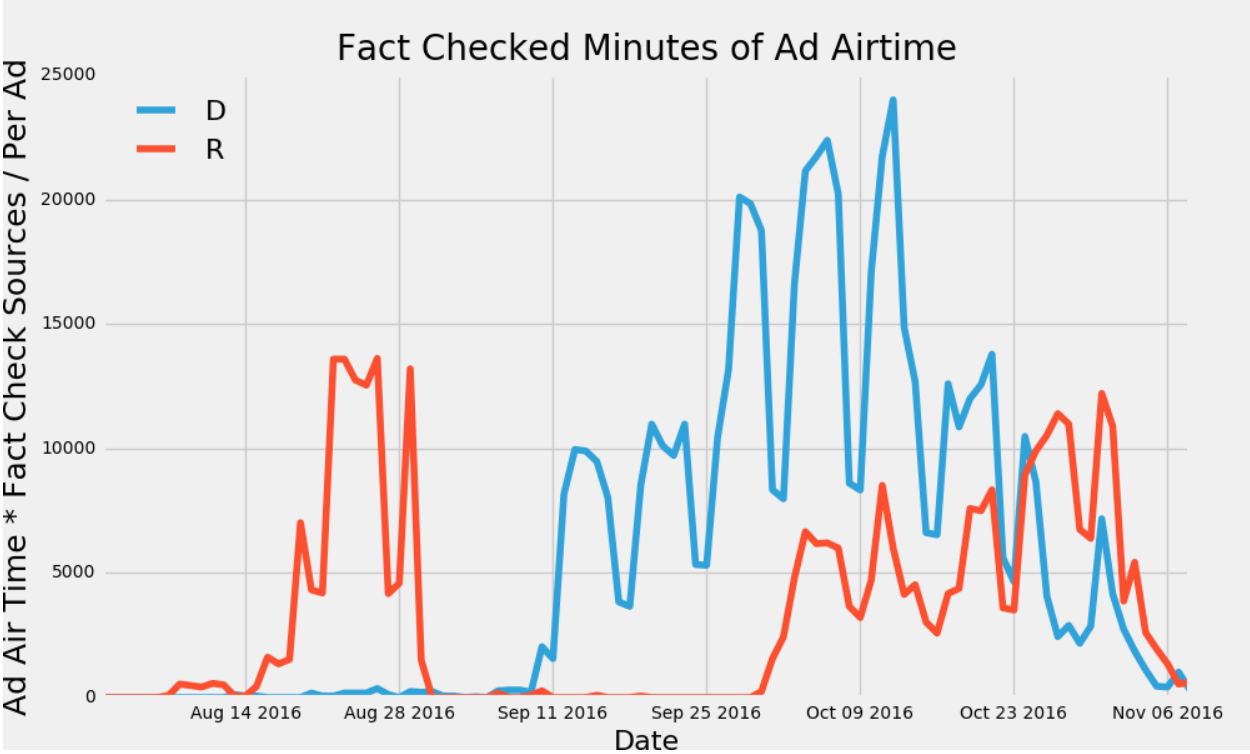
Fact Checked Ads in the Presidential Campaign

There was a total of 245 unique ads run in the 2016 presidential campaign. Of these 245, 20 of them were fact-checked, resulting in 8.16% of unique advertisements were fact checked.

There was a total of 83,349 advertisements aired during the 2016 presidential campaign. Only 24,971 aired advertisements were fact-checked, resulting in approximately 30% of ads.
Number of ads that were fact checked over time



Takeaway: As the blue bars highlight, a huge proportion of ads aired during the campaign were not fact checked.



Takeaway: The graph shows that more republican ads were fact checked during the earlier part of the campaign. This trend fell to gain some momentum in the month of October before declining again. For the democratic party, the ads that aired in September and beginning of October reached had more fact checking. Fact checking has been instrumental in holding politicians accountable but given the small percentage of ads that were fact checked in the last presidential race, we seem to be moving away from checks and balances.

Conclusion

Our analysis of the TV ad data for the battleground states helped answer some of the questions we set out to explore.

The summary analysis gave us a bird's eye view of the ad volume in the battleground states and a broad understanding of message sentiment and content. Our analysis of the data, which included comparison of both political party's TV spending in each state to their respective election results, did not show a strong direct correlation between the ad volume and final election outcome. The derogatory tone that we felt as consumers of these ads was confirmed by the high percentage of negative ads and the preponderance of "candidate biography" as subject in these ads. Putting on our consumer hat, we are of the opinion that several additional factors like TV news shows, social media, etc also contributed to the negative perception.

Further analysis by doing time series plot comparison of ad count by party shed light on trends during the campaign with democratic party maintaining a lead in ad counts throughout. The Democratic party strategy seemed to be to stick to their game plan while the Republican party strategy seemed to have more twists and turns. Perhaps, these were signs of an unconventional campaign. The party may have invested heavily in other forms of media coverage which may have propelled them to victory or maybe the free Twitter outreach by Donald Trump helped in mobilizing enough support. With various conflicting factors in play, it is difficult to say how much of a role TV ads played in the 2016 presidential elections.

One can use this analysis in conjunction with number of tweets, TV news coverage to make a credible argument that alternate mediums of engagement with target voters potentially eclipsed TV ads.