DS4A Team 53 FINAL REPORT

Prianka Ball
Nikki Carroll
Raques McGill
Sydney Pearsall
Ripunjay Singh

Impact of Ad Spending on U.S Presidential Election Outcomes

INTRODUCTION AND BUSINESS CONTEXT

Advertising spending for candidates running for federal office has reached unprecedented amounts, totaling at least $2.5 billion on TV ads so far (PBS, 2020). New political election spending projections for 2020 will now hit $10.8 billion, according to an estimate from the Center for Responsive Politics—50% higher than the 2016 presidential election period (MediaPost, 2020). Advertising spending does not inherently indicate who will win any election. For instance, although Hillary Clinton spent $1,184.1M for the 2016 presidential election, she did not win more electoral votes than Donald Trump although he only spent $616.5M (Bloomberg, 2016).

Problem Statement

Political campaign spending has been increasing over time, making it increasingly important for political candidates to allocate funds effectively. The key questions we hope to answer are: 1) how effective is presidential ad spending at influencing voting outcomes; and 2) how does this ad spending vary in effectiveness based on the location?

PROJECT DEFINITION AND SCOPE

With optimal marketing strategies, presidential candidates will be able to better market their campaigns in a way that translates to votes, allocating their marketing budget in a way that reaches the audiences most critical to the success of presidential candidates. We also anticipate that this can inform presidential campaigns about where marketing spend would be ideal for the success of their campaigns.

DATA

Key Data Sources

Dataset #1: Federal Election Commission Campaign Finance Data

This dataset includes 1.8 million observations of 26 columns, including 5 key fields. FEC data was pulled from BigQuery FEC public dataset using SQL. The data pulled are:

- Transaction Amount *(float/ int)*: Dollar amount of the expenditure
- Transaction Date *(str*): Date of transaction
- Disbursement Category Code *(str)*: Code to expenditure type with different codes for House/Senate vs President
- Disbursement Category Code Description *(str)*: Spending categories (e.g advertising expenses)
- Purpose *(str)*: Filtered data from Category Code description to accommodate missing categories

One advantage of this dataset is that it provides extensive information on the amount of money spent in specific categories for campaigning. A second advantage is that this dataset allows us to see the transactions by state. However, one disadvantage is that the data, as accessed through Google BigQuery, only dates back to 2004. Another complication is that this dataset contains inconsistent naming/ usage of categories related to marketing/ advertising expenses. To overcome the complication of missing categories, we also filter in data in the “Purpose” column containing keywords related to media and advertising.

Dataset #2: Historic Election Result

This dataset contains election outcome data down to the county level by candidate by party. However, for the purpose of our analysis, we decided to aggregate all data by state level. Key fields are as follows:

- State *(str)*: Name of the state
- Year *(str)*: Election year
- Office *(str)*: Name of the office the candidate is running for (e.g US President)
- Party *(str)*: Party name on the election ticket
- Candidatevotes *(int)*: Total number of votes the candidate received
- Total votes *(int)*: Total votes cast in that state for that election race

This dataset will provide a comprehensive view of the election outcomes at the level of detail desired (e.g. state vs county). This will be complementary to the FEC data which only has expenditures by candidate.

Dataset #3: Swing State Analysis by Jon Clayton

This dataset contains swing state data, based on elections between 1992 and 2016, defined as states being in the top 12 in each of 3 criteria:

- Shift: number of times the majority of the state voted for the opposite party compared to previous election
- Battleground: states having most narrow voting margins
- Bellwether: states with best records for voting for the party that wins the presidential election

This dataset contained the following key fields:

- Bellwether_bool *(bool)*: whether state was in top 12 for the bellwether criteria
- Swing *(bool)*: whether state in top 12 lists for all 3 criteria (battleground, shift, and bellwether)

Combining Datasets

All three datasets were merged based on states, Presidential party and election year (Exhibit 1). For this analysis we decided to use Presidential election campaigns from 2004 -2020 only.

Data Cleansing Methodology

The end goal was to obtain advertising expenditure by party by state and the election outcome for a given election year. The most comprehensive data set is the FEC data which had operating expenditures by committee id (cmte_id) and party. The election results data has the party name with which to link to the spending data.

Feature Selection:

We have identified key fields that will be useful in building a model to explain advertising expenditures most predictive of a successful election outcome.

The following considerations were taken into account with respect to the data:

1. The disbursement categories determined to be inclusive of marketing are advertising expenses, campaign materials, campaign event materials, media expenditures, and expenditures for mass mailings and other campaign materials, thus all other expenses will be excluded.
2. The disbursement purposes determined to be inclusive of marketing are those which contain one of the following stem words: “media,” “digital,” “advertis,” “/ads,” and “promo”.

Key Steps:

- Generate master dataset: Use SQL to pull relevant fields and join key FEC datasets from Google BigQuery
    a. Datasets included the 2004-2020 tables for the following tables:
        i. Operating Expenditures (oppexp)
        ii. Committee Master (cm)
        iii. Candidate Committee Linkage (ccl)
        iv. Candidate Master(cn)
- Filter the data for rows where
    a. "cand_office" field denotes presidential
    b. "Disbursement Category Code" is in a list of select codes related to advertising
    c. "Purpose" field consists of select stem words related to advertising
    d. State is in a list of 50 states and D.C., excluding U.S. territories
        i. While looking at states, we have noticed data to have more than 50 states. That comprises countries outside the US where candidate supporters are raising money for candidates. We decided to leave out values that do not fall within the 50 states in the US as it reflects less than 0.05% of the data. It is possible for each committee id to be spending in multiple states.
- Aggregate data by year, state, and party
    a. Transaction column had a significant portion of rows with negative numbers reflecting voids in expenditures, we aggregated them to reflect the net transaction amounts.
- Engineering features (from existing columns):
    a. Party_simplified: made 3 party categories: Democrat, Republican, Other
    b. Election year: Combined all spending into the election year for which transactions were made (e.g., spending in 2010 for the 2012 election)
    c. Percent_of_total_spend: percent of total spend made in each region by a party in a given year out of the overall national spend by the party for that year
    d. region: state categorized into one of seven US regions (i.e. Midwest, Southeast, Southwest, Mid-Atlantic, Pacific Coast, Rocky Mountain and Northeast)
    e. Won: boolean; per year, state, party
    f. Percentvote: percent of total votes received by party per state, per year
    g. Spend_share: % spend per party by year and state

EXPLORATORY DATA ANALYSIS

Total transactions have seen a large growth over the years, reaching unprecedented highs in the latest Presidential Elections.

However, spend share by party has been mostly consistent since 2008, with non-major parties collectively spending as much as any major party.

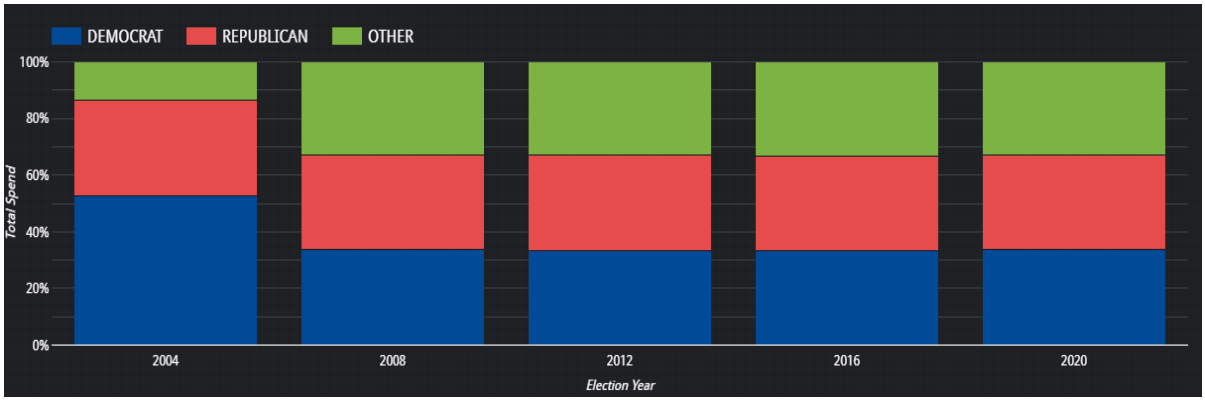


Figure 1: Total spend by parties over years

The greatest variation in spend share exists by state.

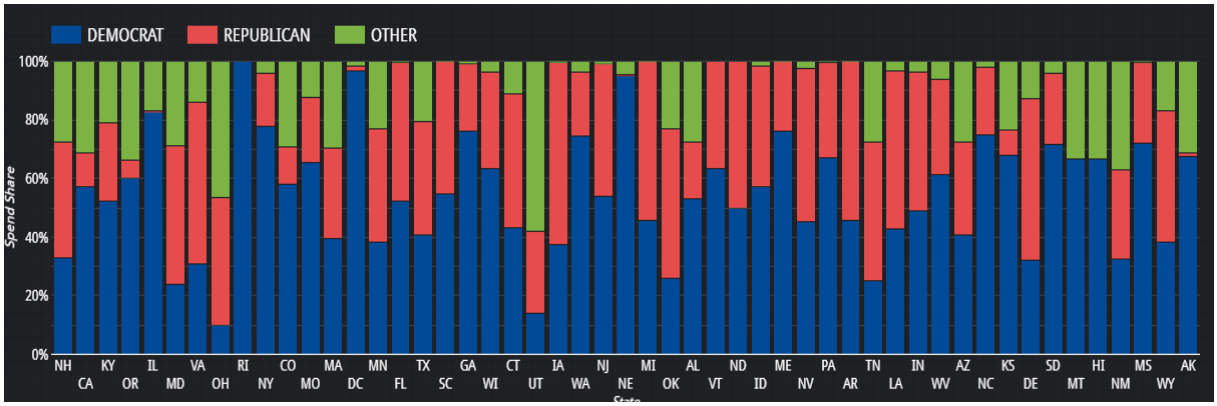


Figure 2: Spend share by state

States with the larger populations (California, Florida, and Texas) garner the most votes due to population size, while smaller states (D.C., Virginia, and New York) garner the largest ad spend.

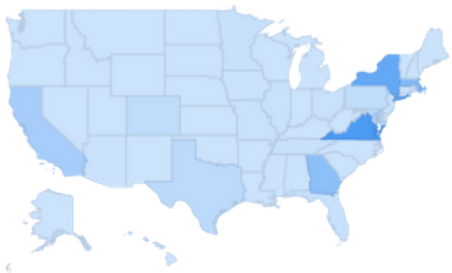


Figure 3.1: Votes by state

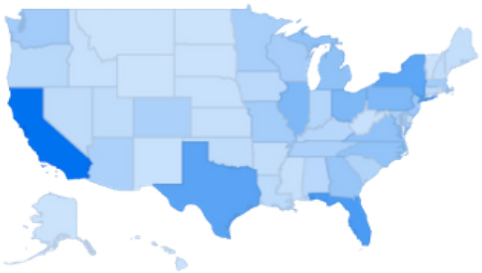


Figure 3.2: Ad spend by state

When looking at spend and vote data, we found the data to be highly skewed. We have used log transformation to manipulate the data and make it easier to do some of our analysis and graphs.

Log transformation normalizes the data and makes it easier to read some of the graphs. Statistical analysis of log data was also helpful to understand the skewed dataset and how much we can rely on it to interpret our results. We have tested out log manipulated data in the regression model but decided to not use it as a final model as it had similar results to non-log data.

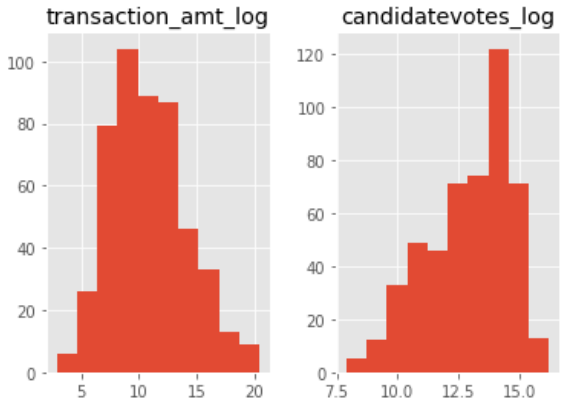


Figure 4.1: Histogram for transaction_amt_log and candidatevotes_log

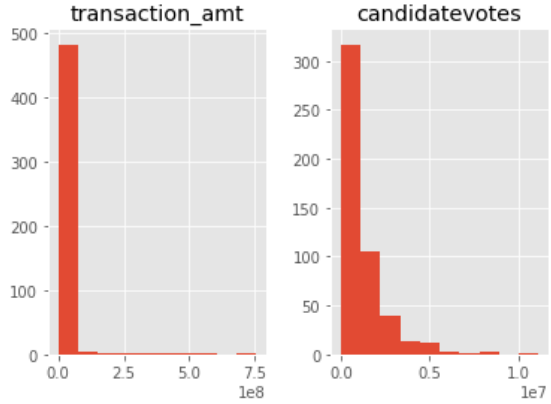


Figure 4.2: Histogram for transaction_amt and candidatevotes

As scatter plot graphs could not find correlated variables, we have sliced the data in different ways to understand the different trends. In the graphs below, we use log data as it gives a better visual representation of the data and easier to understand relationships between different datasets visually.

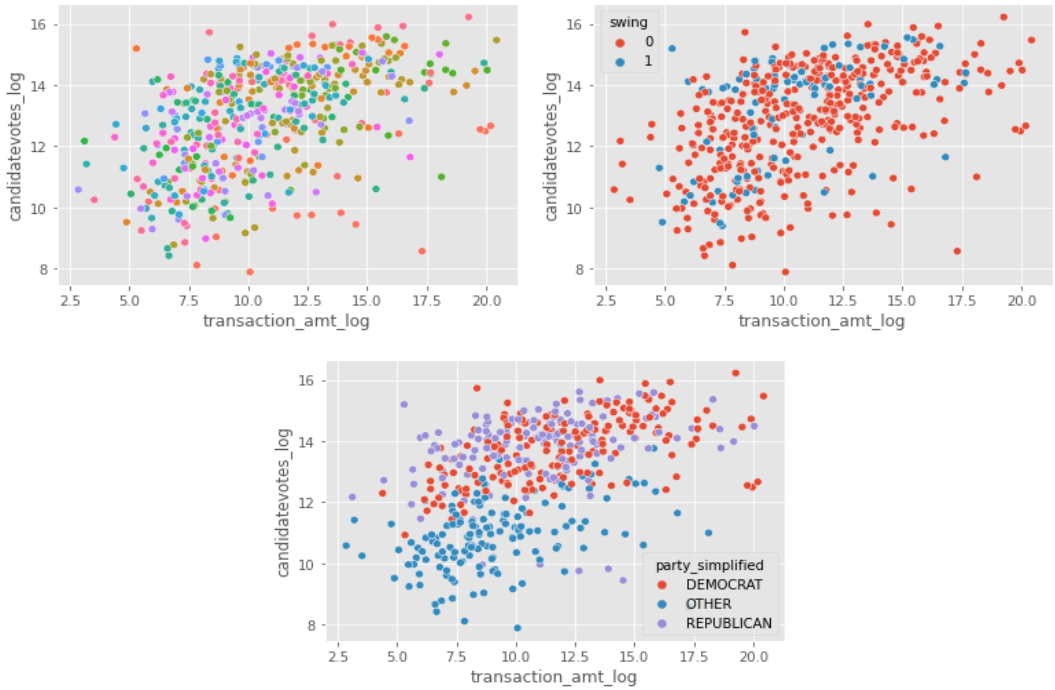


Figure 4.3: Scatter Plots showing relation between candidate votes and transaction amount for (from left to right) swing states, parties and regions

In 2004, the voting share was heavily distributed, and while the Democrats were able to gather votes in five of the seven regions, Republicans saw support in only two regions (Midwest and Southeast).

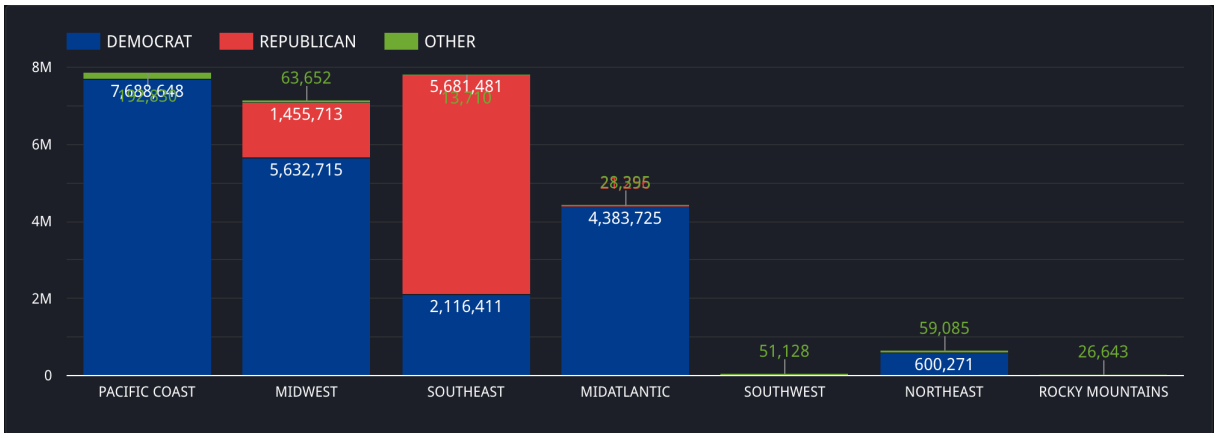


Figure 5: Vote share by region 2004

From 2012 onwards, both the Democratic and Republican parties were involved in a close competition in some states, whereas there were clear winners in the others. Compared to the two big parties, other parties didn't gather many votes.

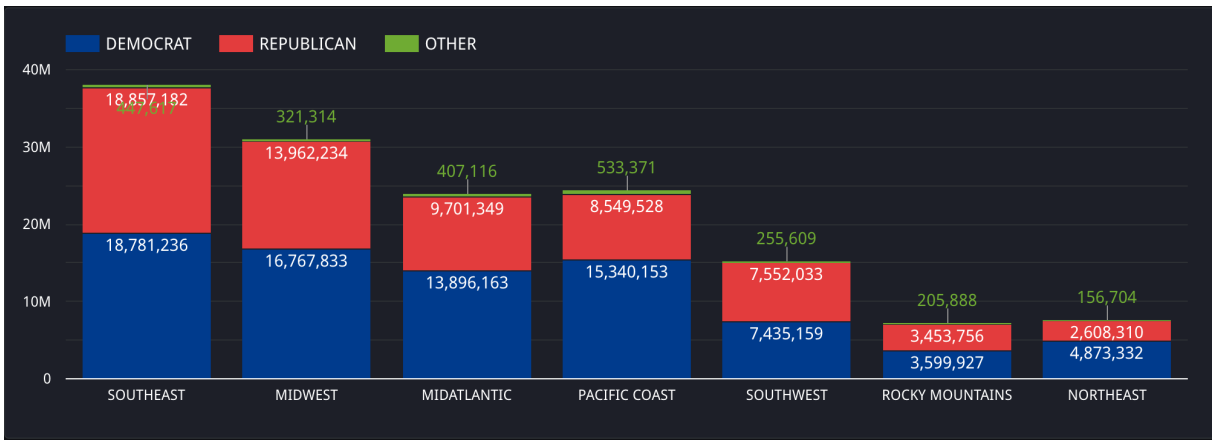


Figure 6: Vote share by region 2020

Democratic and Other candidates allocate greater proportions of their spending to the Northeast than what Republicans allocate.

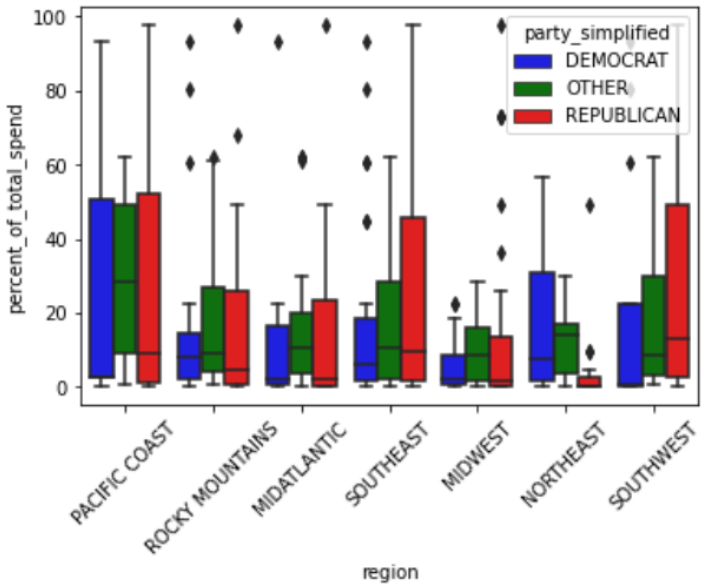


Figure 7: Percent of total spend by region

Democratic, Republican, and other presidential candidates do not allocate significantly greater proportions of their spending on either bellwether states or swing states.

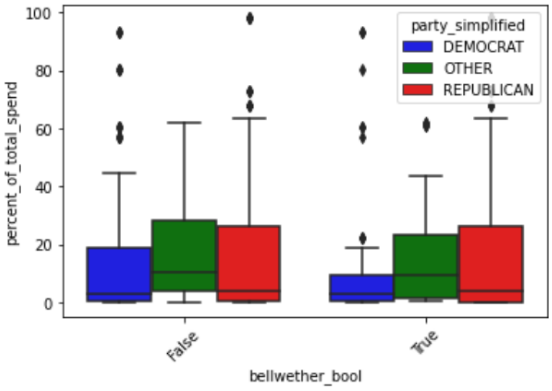


Figure 8.1: Percent of total spend by Bellwether status

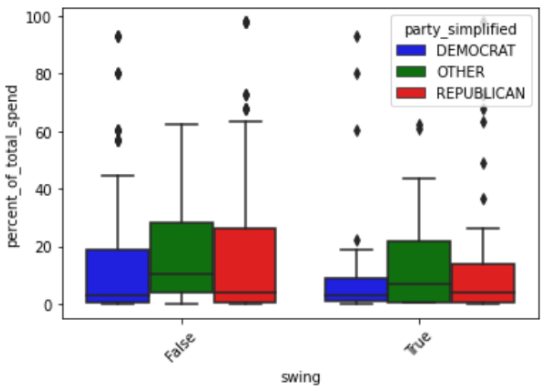
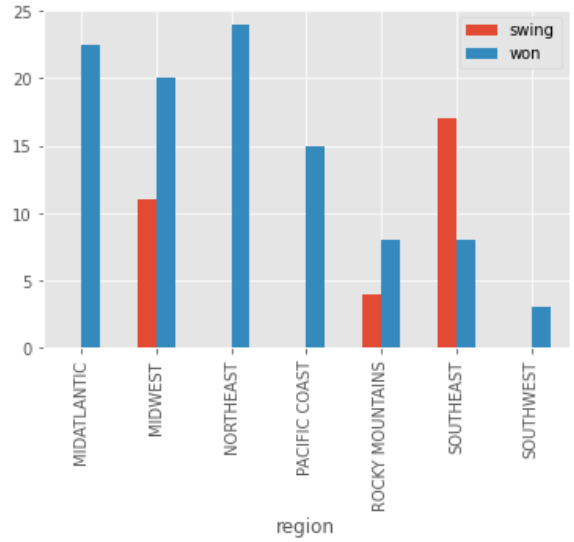


Figure 8.2: Percent of total spend by Swing status

Democrats had the most wins in Mid-Atlantic, Northeast and Pacific Coast; while Republicans won the most in Southeast and Southwest. Midwest and Rocky Mountain had equal wins across both major parties. Midwest and Southeast have the most number of swing states.



Figureure 9.1: Democratic wins in swing states

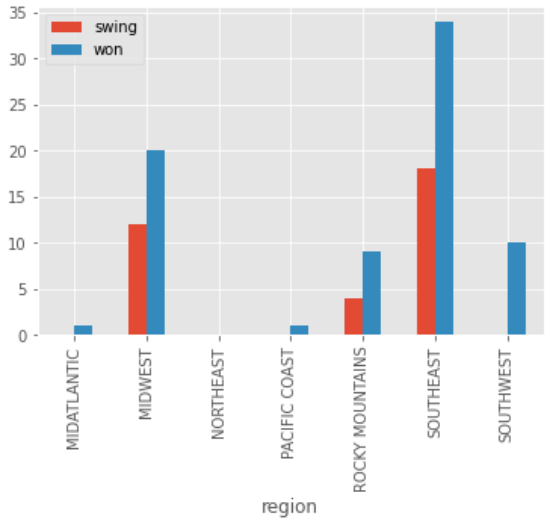
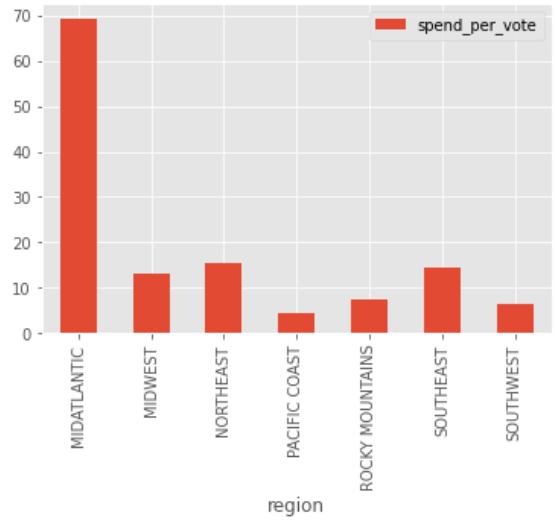


Figure 9.2: Republican wins in swing states

Median spend per vote in Mid-Atlantic is high for both parties, which may just be due to this being an expensive region. Republicans tend to spend more in Republican leaning regions. The analysis also showed that median spend per vote for Republicans in Southwest is quite high.



Figureure 10.1: Democratic median spend per vote

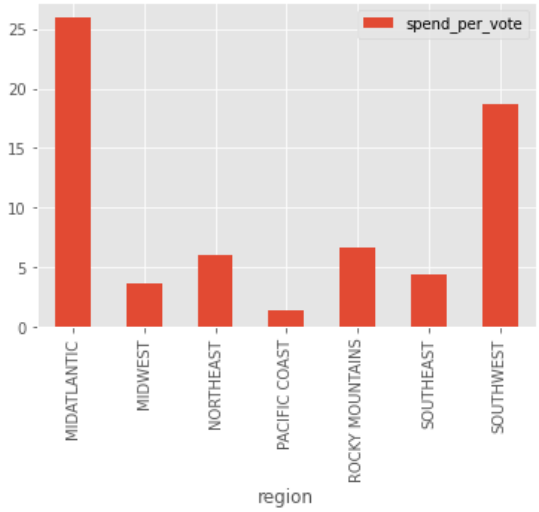


Figure 10.2: Republican median spend per vote

Statistical Analysis

Correlation between votes and spending

Democrats have a more positive correlation between spend per vote and percentage of final vote compared to Republicans who have a slightly negative correlation. This indicates that Democrats have a stronger need to spend more money to win an election.

Looking at regions specifically for Democratic candidates, the median spend per vote in the Mid-Atlantic is high in addition to there being a negative correlation between spend and votes. However, the Southwest and Rocky Mountains have higher correlations even with Democratic candidates historically losing in the Southwest. The best shift would be to move marketing spend from the Mid-Atlantic region to the Rocky Mountains.

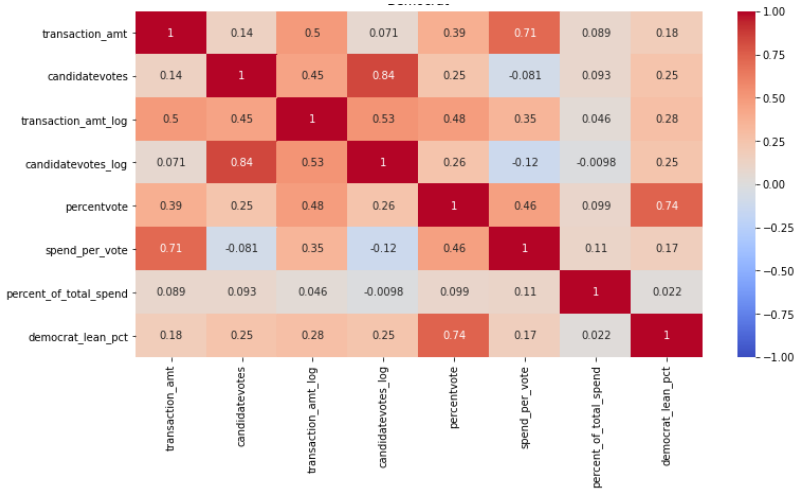


Figure 11.1: Correlation matrix for Democrats

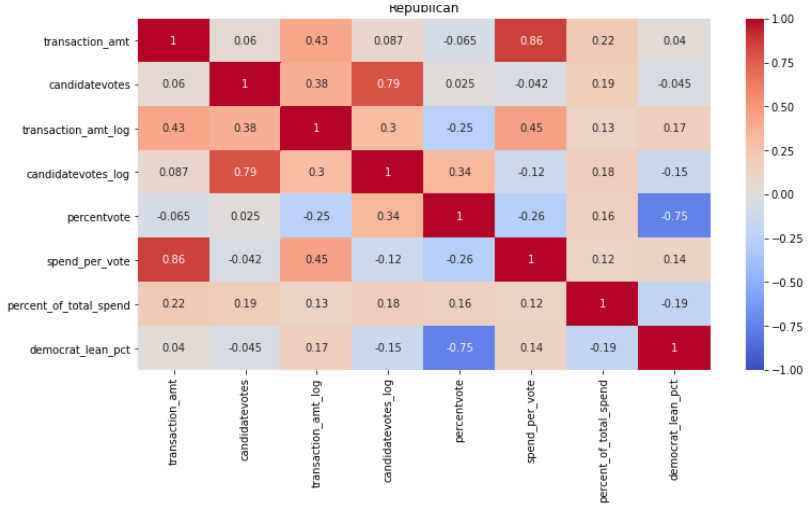


Figure 11.2: Correlation matrix for Republicans

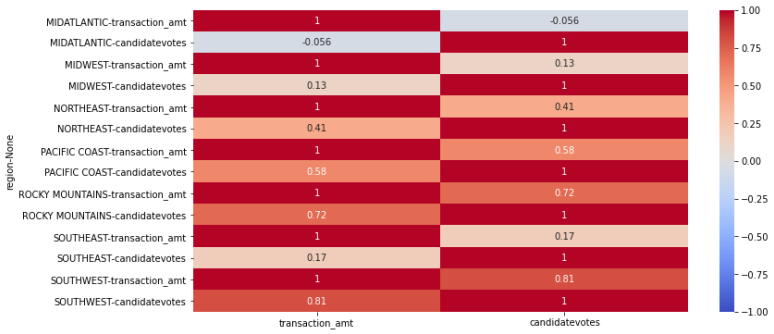


Figure 12.1: Correlation table for region, spend and votes for Democrats

While people that tend to vote for Republicans are less responsive to a higher marketing spend, the Pacific Coast, Northeast, and Southwest all have the highest correlations between spend and votes. Therefore, it would behoove them to spend the most in these regions. When log function is used, correlation between votes and spend is high for both parties.

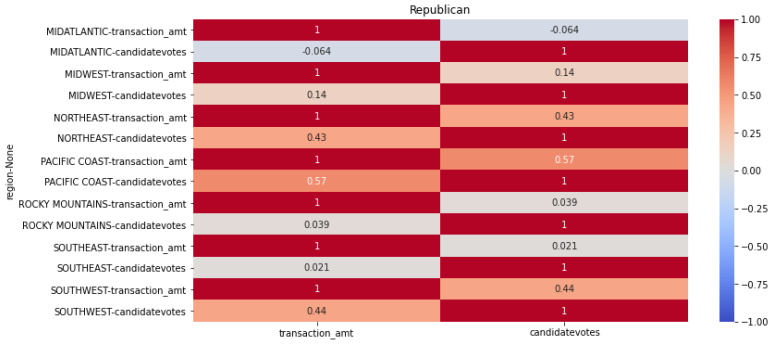


Figure 12.2: Correlation table for region, spend and votes for Republicans

Regional Spending Trends

region	transaction_amt	candidatevotes	percentvote	spend_per_vote	percent_of_total_spend	democrat_lean_pct
MIDATLANTIC	1475590.920	2135532.0	60.242755	69.298136	2.180	100.00
MIDWEST	100300.000	1374039.0	46.453841	13.177525	2.180	14.29
NORTHEAST	46825.600	377193.5	58.377901	15.467099	7.635	100.00
PACIFIC COAST	66603.320	1541550.5	56.601062	4.487713	2.870	100.00
ROCKY MOUNTAINS	35316.080	327670.0	40.549369	7.451394	8.250	14.29
SOUTHEAST	140393.375	868318.0	42.209144	14.328231	6.000	28.57
SOUTHWEST	32948.830	1034707.0	44.589767	6.538893	0.680	0.00

Figure 13.1: Democratic Median Statistic by Region

region	transaction_amt	candidatevotes	percentvote	spend_per_vote	percent_of_total_spend	democrat_lean_pct
MIDATLANTIC	170345.540	1227251.0	36.710777	25.962165	2.275	100.00
MIDWEST	33659.350	1445814.0	48.822437	3.608099	1.670	28.57
NORTHEAST	53861.695	363195.5	40.355705	6.026447	0.320	100.00
PACIFIC COAST	10000.000	1290670.0	38.766978	1.325952	9.150	100.00
ROCKY MOUNTAINS	54648.000	554119.0	47.666247	6.639163	4.480	0.00
SOUTHEAST	63618.815	1470754.0	55.016407	4.352458	9.710	28.57
SOUTHWEST	471178.440	1233654.0	53.635248	18.681930	13.100	0.00

Figure 13.2: Republican Median Statistic by Region

Median spend per vote in Mid-Atlantic is high for both Democrat and Republican. The Mid-Atlantic might just be an expensive region. Median spend per vote for Republicans in Southwest is high. We also noticed that Republicans tend to spend more in Republican leaning regions (e.g., Southwest). It is more expensive for Democrats to get votes in the Midwest than for Republicans.

Model

Linear Regression

After running several regression models, we concluded that spend and states are important factors when trying to determine the number of votes. For the final model, we used:

$$votes = B0 + B1 * spend + B2 * states$$

There is enough evidence that if spend increases by 1, then votes are expected to increase by 0.0017 given all states are equal. The p-value was less than 0.05 meaning that it was statistically significant.

The model has r^2 values of 0.914. The r^2 value was very high which suggests we might be overfitting the data and would like to use more years of data in the analysis.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         candidatevotes   R-squared:                       0.914
Model:                            OLS   Adj. R-squared:                  0.899
Method:                 Least Squares   F-statistic:                     61.95
Date:                Sat, 06 Feb 2021   Prob (F-statistic):          5.53e-131
Time:                        18:05:23   Log-Likelihood:                -5059.1
No. Observations:                 351   AIC:                         1.022e+04
Df Residuals:                     299   BIC:                         1.042e+04
Df Model:                          51
Covariance Type:            nonrobust
==============================================================================
```

Figure 14: Linear Regression Model Results

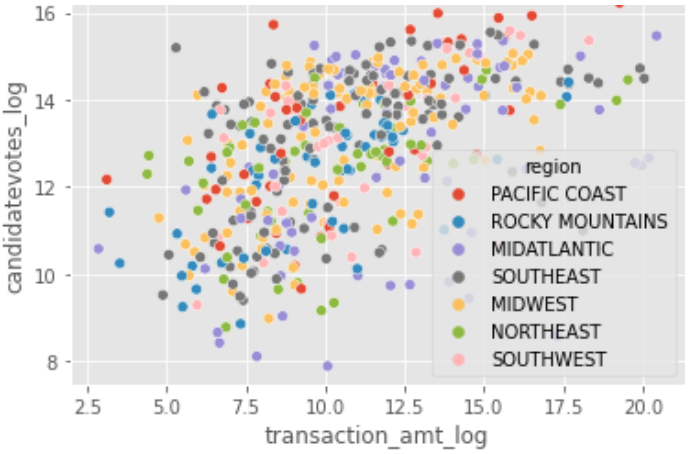


Figure 15: Linear Regression Model

DASHBOARD

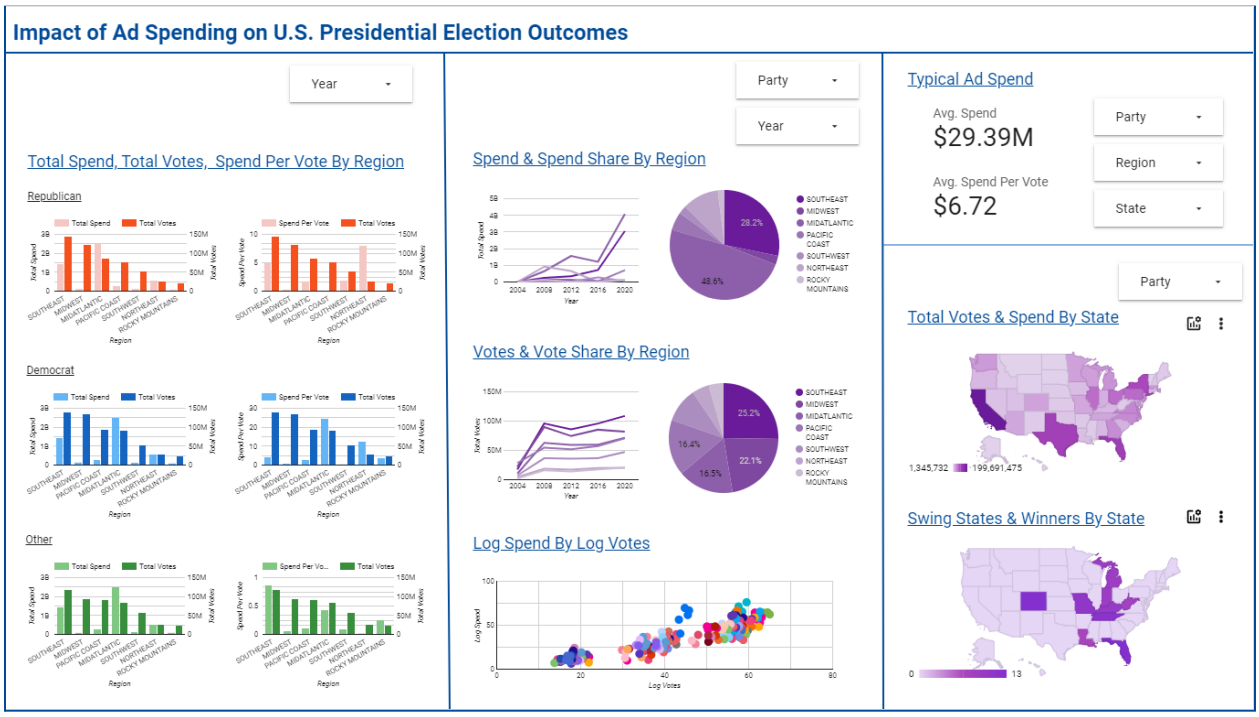


Figure 16: Tool for candidates to determine typical ad spend based on party, region, and state

Also accessible here: https://datastudio.google.com/reporting/8d39e63c-edf8-46b7-8e19-f5e0515c05f1

RESULTS

Taking all the insights and analysis into consideration, we have concluded the following insights that might be useful for future presidential candidates:

- Median spend per vote is very high for Democrats in Mid-Atlantic. There is also a negative correlation between spend and vote in this region for Democrats. Democratic candidates should evaluate their spend in this region as they can spend the money in other regions that is likely to have more effect due to increase in spend.
- Democratic candidates can increase spending in Southwest and Rocky Mountain instead, as their regions have shown to have high correlation, meaning high spend can lead to more votes. However, historically Democrats had fewer wins in Southwest. They have a better chance of winning in Rocky Mountain areas if they spend more.
- Pacific coast, Northeast, Southwest have a high correlation between spend and votes for Republicans. This shows that high ad spend can help increase the number of the votes for the party. Historically, Republicans also had the most wins in Southwest so it would be a good region to spend money to get votes.
- Other parties (not Democrats or Republican) should consider merging to form a competitive third major party, since Other parties collectively have been spending as much as the major parties since 2008.
- MidAtlantic shows to be an expensive region.
- Both parties don't allocate greater proportions of spending to swing or bellwether states.

NEXT STEPS

Due to time constraints, we were able to look at a small segment of US elections. We would like to take this analysis further in the following ways:

- Analyzing
    - Quality of different fundraising channels
    - Effectiveness of different marketing channels
    - Local/ state election impact
    - Electoral votes as outcomes
    - Effectiveness of timing of monthly/ yearly spending
    - Effect of marketing on voter turnout rates by year, state, party
    - Independent spend as it would have super pac spending.
        - Citizen United act was passed in 2016, which leads to large increase in campaign budget. Data set does not contain sending from super pacs as they will be considered as independent spending
- Deploying model into web application to:
    - Predict election outcomes based on expected budget, party, fundraising channels, and candidate office
    - Optimize budget allocation across localities and marketing channels

APPENDIX

Exhibit 1. Entity Relationship Diagram