Assignment 2

The goal of Mass Alliance is to attain meaningful, to them, information about the trends of voters throughout the years. Our project involves providing them with an insight of the answers of the Ballot Questions asked in the last 17 years. We have available 24 Ballot Questions. The data we retrieve from these questions is the amount of [yes, no, blank, total] votes for each of the available questions. Our goal is to combine those results from the lowest area denominator, which is a precinct, to the higher one which would be a district for the state of Massachusetts. After combining the results, we need to create an average score for each of these localities depending on the answers given. The metrics and the scores for each question depending on the answer are provided to us by Mass Alliance.

A. Analysis of our algorithms

fetchCSVLink :

In order to achieve the final result, we needed to scrape the available information about the questions. We access the provided link with all the questions and for each question we get the ulr that contains a “csv” with the data we need. These urls are stored in mongodb and accessed later.

questionsAllYears:

Accesses the stored csv urls and retrieves the data for each of the available questions. Each question is stored as a separate collection in mongodb.

questionsMajorityResult

Accesses each question and aggregates the data of each question to the highest available denominator. After the aggregation is completed a majority and a weighted score is computed for that question according to the number of yes’s and no’s , which in turn gets aggregated to the scores of the the previous questions to create an average. The file creates two datasets both with the format of [Locality, Score]. One of them reflects a score based on the majority of the answers (majority) and the one reflects the score after weighting the amount of yes’s and no’s (weighted). These two results are then stored in mongo.

ballotQuestions

Transforms the original ballot questions csv to provide the scores assigned to each question asked over the years. It also calculates a corresponding “NoScore” which is the inverse of the given score to help determine if a given district/precinct is Progressive or Conservative.

correlationCoeff

Retrieves the two available scoring datasets , “majority” and “weighted” and computes the correlation coefficient between these two datasets.

B. Constraints Problem

Being able to access the amount of votes can give us insight on the trends of the population. We specific focused on one year, 2016, and used averages of the answers to formulate a simple constraints problem that could reflect if an area would be good for more conservative or progressive advertising. Using the amount of yes/no/blank votes we can determines these kinds of trends by filtering out the records unsatisfy the constraints.
More specifically:

We have applied three constraints:

1. Find the locality where more people answer questions with 'yes' than people answer with 'no'. This constraint simple show the inclination of political attitude at a certain area. It only output the place that could possible be a conservative one.
2. Find the locality where the number of people answering questions with 'yes' is greater than one and a half times of the number of answering with 'no'. Since the places found have much more conservative people than progressive, we can claim that putting any conservative advertisement could have positive effect for these localities.
3. Find the locality where the number of respondent is greater than two times of the number of people that gives up answering. From this constraint, we are able to predict how much impact the advertisement are likely to have. If there are a lot respondents to the ballot questions, we can assume most people at this spot are likely to react with political stuffs because they are not indifferent to give up answering questions.

According to these constraints we remove areas from our dataset and we only keep the areas that reflect a more conservative attitude and have positive results of any conservative advertising.

C. Statistical Analysis

Our main results decide the score of a locality based on the majority of yes's or no's. Apart from that we also create a weighted version of the scores based on the actual amount of answers instead of just the majority. Computing the correlation coefficient between these 2 datasets yield a result around 0.5. While this is not a relatively close to 1, it's still a positive number close to 0.8. We can't strongly say that the weighted results and the majority are highly correlated but there is a trend there that means that the majority is usually dominating in determining the result rather than taking into account the actual amount of votes.