Case Study: Data Analytics Study Design

(EPORTFOLIO ASSIGNMENT)

November 13, 2022

Ronald Chum
Northwestern University

IDS 452: Intro to Data and Analytics

Table of Content

Why is this an important study / research question?

From 1972 to 1988, Republicans have a track record of dominating elections in those years, with their candidates frequently receiving more than 75% of the electoral votes cast across all 50 states. In recent years, from 2004 to 2020, the presidential election has started to change. The entire presidential election was no longer won by a single party in a landslide. In modern society, the odds of winning the election for both the Democratic party and the Republican party are about even (based on the research findings).

This study focuses on predicting the 2024 presidential election outcome. Even though the data set is limited in volume and variety, the predicted outcome seems reasonable based on the research findings. The research is important because it gives the company or business an earlier advantage to produce a product or service that is compatible and specialized for the winning group. It is not cost-effective to produce and transport products related to the Republican Party and sell them in areas that are dominated by the Democratic Party. It is more efficient and effective to be the first to win in any competition.

Following the completion of a data analysis on the 2024 presidential election. There are some essential findings and modifications that will improve the prediction and forecast in a more accurate manner. The college degree education rate is deemed an effective variable to predict the election outcome for each state.

What should be the null and alternate hypothesis for the study / research question?

The updated hypothesis in this research will be that states with a high college degree (CD rate) education percentage of the population are correlated with the Democrats winning the state election in general.

The updated null hypothesis
H0: The Democratic Party will NOT win state-level elections because of a high percentage of the college-educated population.

The updated alternative hypothesis
Ha: The Democratic Party will win state-level elections because of a high percentage of the college-educated population.

Originally, the hypothesis (original hypothesis) tests the increase in education rates (Δ education rate), However, the evidence is not significant enough to conclude that the Δ education rate has a correlation with the state level election result. Therefore, the hypothesis (new/updated hypothesis) was then adjusted and modified based on this original hypothesis' findings.

The result from original hypothesis has revealed an interesting information after categorizing states into tiers. Based on 2020 values, 20 states (plus the DC special district) revealed a 9.77% difference between states that vote for the Democratic Party and states that vote for the Republican Party. The blue states have an average educated population (education rate) of 37.83%, while the red states have an average educated population (education rate) of 28.06%. In Tier 3, the calculated upper and lower confidence for blue and red states do not overlap each other. This represented that they use this as a tool to predict the election of a specific state if it falls within those confidence bands. The updated hypothesis and findings are

reliable enough to determine the possible election outcome based on the college degree education rate as it is deemed to have correlation.

What type of team will be needed for this study?

Under normal circumstances, a team of four would be in charge of the research from the beginning until the end. The team usually consists of a project manager, a data wrangler, a data analyst, and a presentation designer.

The project manager's main duty is to sell the project, sell its potential value to the management, obtain a budget for the team, manage the entire project as a whole, and report the progress to management on a regular basis. It is important to demonstrate that the team has an essential purpose because that is what keeps the team alive (or else the team will be terminated or disbanded). It is the middle person between the management and the analytical team members (taking the stress, pressure, and feedback from management and breaking it down for the team to work on), and it is the leader of the analytical team.

The data wrangler and the data analyst (scientist) are the main labor-intensive positions in the analytical team. They work together to build the necessary usable structured data sets, algorithms, and other statistical engines. Then they use those tools to test the hypothesis, find correlations and messages hidden among the data, and apply the models to fit the business needs.

The presentation designer needs to produce a visualized version of the research for the internal and external stakeholders to understand in a simple way (the management would prefer the message to be delivered like speaking to a young child; they would prefer easy-to-understand graphics and plain English rather than in a technical way). The goal for the presentation designer is to deliver messages, trying to make the presentation interesting enough for the management or shareholder to digest the information but not too difficult to understand.

What are the time constraints for the study – e.g. is the need immediate or can this be a longer-term study?

This is a long-term study that needs to be completed prior to the 2024 presidential election. In fact, the data set includes election results for each state from 1960 through 2022. However, the data set was trimmed down as "noise," and nonessential data were removed to maintain focus on the 2012–2022 data. The data was collected from various sources – mostly U.S. government sites such as the census bureau, election commissions, and departments of state and commerce.

Since the dataset was trimmed down to only contain education data related to 2020 and 2022, 2012 and 2016 education data were obtained from the National Center for Education Statistics and added to the dataset to increase the sample size and enhance the accuracy of the prediction outcome.

The data set is more balanced as it includes education data from 2012, 2016, and 2020, as well as election results from 2012 to 2020 (by state). These education data have minimal bias towards the political parties because they only focus on the education rate in each state, and the agency that operates under the U.S. Department of Education specializes in collecting, analyzing, and publishing statistics related to nation-wide education.

What type of study design will be utilized?

In this case study, nominal data (historical outcomes: election results in D or R) will provide a linear progression prediction. The historical data will provide a surface prediction of the election outcome, and it was used to form hypotheses. Ratio data (such as the high school and college degree education rates) were utilized in the study. This ratio data is essential to the case study. It provides numerous connections and correlations among various attributes. Ratio data from 2012 to 2020 was the primary focus.

What methods should be used to remediate any data issues – e.g. outlier tests, an attempt to structure data from unstructured data, accounting for data accuracy, etc.

Outlier tests are a great tool to remove outliers, confine central tendency values, and maintain research focus. However, there are many ways to remediate data issues. In this case study, the data was scrubbed and cleaned prior to the analysis. The data was audited and deemed to contain no errors.

Extra data is a two-way street. Excessive and nonessential data will be prone to creating noise along the way during the analytic process. It is possible that extra data can also increase the accuracy and reliability of the data. Irrelevant data was removed, which reduced the data scrubbing and cleaning time.

The only error that would occur in the analysis is that the data itself is fraudulent or erroneous; it is possible that the values themselves are inaccurate. However, the data was obtained from the government, and the data was compared to prior years' estimations and seemed reasonable and reliable to be used for analysis.

What types of statistical tests should be conducted and why?
How should the statistical tests be used in a manner to support each other or find ‘hidden’ meaning in the data?

Linear regression was also used to compare calculated values among selected states. Linear regression is used because the predictor variable is expected to be continuous and the election outcome should follow the previous trend, which allows the study to project the outcome based on prior data (in a continuous linear manner).

A correlation test was used to confirm the correlation between the college education rate and the state-level election outcome. It has been established that a high rate of college education correlates with Democratic victories at the state level.

What are the possible risks and errors associated with the data outcomes and how might this affect the usefulness of the data outcomes?

The data that was used to predict the election outcome was mainly obtained in 2022. It is possible that the values will have changed by 2024 and that the data will no longer be valid at that time. The data set is always going to be a lagging indicator and should be used with caution. By acknowledging the possible change in 2024 data (that will be obtained in the future), the data scientist should include this as a threat or a wild card that can change the outcome of the prediction.

What threats to statistical validity (internal and external) exist?
How might the identified threats affect the study design?

There are ancillary environmental factors that could affect the research, such as party candidates' popularity, scandals, or domestic or foreign policies. This factor was not considered in the research so as to reduce the complexity and the time spent on research. There is a possibility that some ancillary factors might change the variables that we have tested in the research. Some might even be more important than the ones that were under consideration. The ancillary factors could potentially increase or reduce the significance of the variables in the dataset, making them more or less influential. In order to mitigate such risks, it is necessary to quantify the threats by rating the threats. Threats with a higher potential for influence would receive a higher score, and the same goes for threats with a low influence potential. These ratings can be included in a logistic regression to find out the impact of those threats.

Q & A section

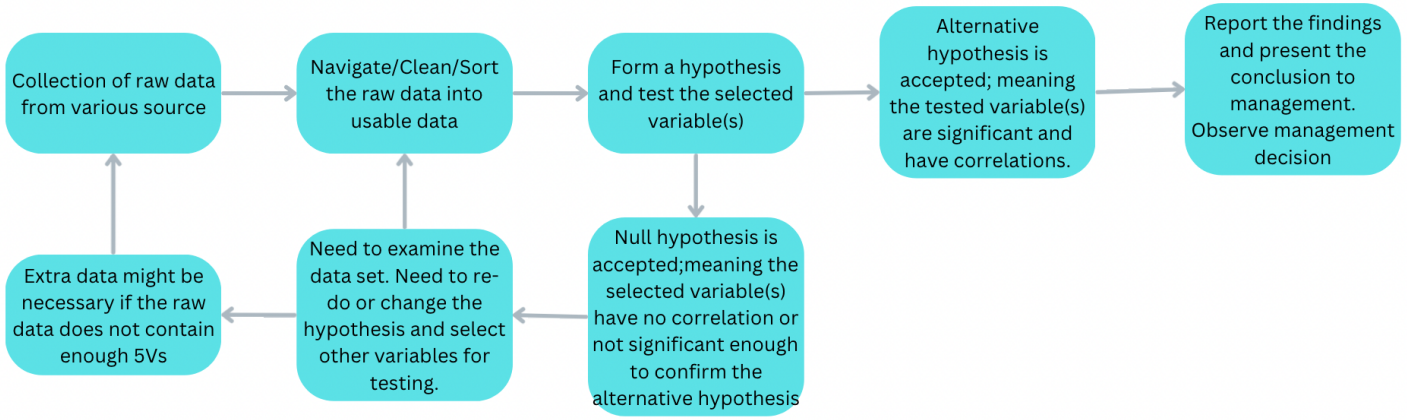
Why did you adjust or modify your hypothesis?

I have accepted my original null hypothesis in my case study. Originally, I tried to fit the data into supporting my alternative hypothesis. I was hoping to make my HA strong enough so that I could conclude my result with a strong correlation and a strong significance level. I realized that I shouldn't beat the data up to say or support my perspective. I later accepted the fact that the data was not significant enough to support my hypothesis. I found some new information during my analysis. I found out that the blue states usually have a 9% higher education rate than the red states, on average.

What is the prediction outcome?

The Democratic Party will win the election with 293 electoral votes, and the Republican Party will have 245 electoral votes. These calculations are solely based on the college degree education rate. The assumption is that any states with CD rates above 30% will vote for Democrats, and any states with CD rates below 30% will vote for Republicans.

Theoretical Research Flowchart



Actual Research Flowchart in this case study