Team 2 (Jessica, Chelsea, William, Graeme, Patrick)

M2 – Voting Analysis Exercise

Overview:

- *This assignment addresses techniques for answering the following research question:*

    - ***Question: Could the majority of the popular vote for U.S. President in 2012 have been cast for Romney rather than Obama because of voter fraud?***

    - ***Answer**: The basic summation of the total votes earned by each candidate indicates that Obama clearly won the popular vote, even after three separate experiments.*

1. Take a Look at the Data

In bullet points, list your team's responses to the questions below. While these bullets can be brief, please make sure you *identify the data sources*, *the data field*, and the *issue clearly* in your responses.

- What data fields might be useful in answering the research question and how could you use them?
    - Public Interest Legal Foundation (Over-registration by county Data)
        - County, State, Registration Rate, Combined State/County field
        - Source: http://publicinterestlegal.org/county-list/

    - 2012 U.S. Presidential Election Data (Guardian - U.K)
        - County, State, FIPS Code, Total Votes Cast, Votes (Obama), Votes (Romney), Votes (Johnson), Votes (Stein), Votes (Goode)
        - Source: http://image.guardian.co.uk/sys-files/Guardian/documents/2012/11/14/US_elect_county.xls

- How is your data formatted (what type of files, string versus numerical data)?
    - Data is in Excel format with a combination of string and numerical data.

- Did your team find any missing data?

    - No column for # of registered voters by county / FIPs in the Guardian – U.K. data. This makes us assume that the winner of the counties with fraud may have affected close votes to sway in the opposite direction

    - The Obama vs. Romney comparison seems to be missing approximately 500+ counties compared to the list found on the FULL DATA worksheet. These counties appear to be from the states of CO, CT, DC, FL, GA, SC, UT, and WY.

        - The Obama vs. Romney comparison is missing 12 counties that are on the Public Interest Data of U.S. counties with more voter registrations than people alive file.

- Did your team find any data that could be erroneous?

    - Yes, there is a discrepancy in the percentage of the vote taken between Obama and Romney in the states of MA, ME, NH, and VT between the Romney vs. Obama worksheet and the FULL DATA in the Guardian – U.K. data.

    - The Public Interest data shows that there are more than 100% registered voters which is not possible. “Counties with more registered voters than people of voting age that are alive:” There were accusations that dead people had voted. Franklin County, Illinois had 190% and Pulaski County, Illinois had 176% registered voters. Even 100% is likely not to be possible.

    - Some of the FIPS are not reported correctly. For example, Alaska has 0 and another for 2000 when Alaska does not have a county FIPS of 2000 (they range from 2013-2290) – it’s also counted twice. DC has 0 and is counted twice with the same amount for each data entry (222,332). Alabama at 0 has only one entry; it has 10 counties that have over 100% registered voters.

    - Missing data is included in the winner field, which does not display data for losses. This data may need to be transformed into a binary format.

    - The Winner column appears to mark the row with State/County results (by FIPS) erroneously. X’s are placed for the candidate that won the state across all counties in that state vs. marking X’s appropriately based on the true winner in the county. Some “winner” fields do not filter correctly.

- What unusual facets of the data must the analysis take into account?

    - The FIPS Code of 0 appears to carry state-level votes cast. While not erroneous, we must consider this in order to make sure we do not overstate vote counts.

    - Columns have similar names, which pre-empts the question in how this data was organized/compiled.

    - Lots of unnecessary columns, which could be deleted to minimize confusion

    - Spaces and additional text used to identify the county. For example:

        - NM De Baca vs. NM DeBaca;
        - Additional Text: Parish for LA (i.e. LA Tensas Parish, LA St. Helena Parish, LA Cameron Parish)

- National Politician ID for third-party candidates – there appears to be 9 separate fields with this name. The data values appear to be inconsistent.

- There are a lot of unused columns that could be deleted and there are some that could be combined or should be clarified more.

- It at first appears as if each candidate has their own columns (first, middle, last name, etc.) but when you scroll down it switches between candidates. If each had their own column it would make it easier to keep a more accurate track of the counts and mark the "winner (x)" within their respective column.

2. Analysis Design

Answer the following question.

- Give a formula for a computation your team would use for this analysis. Describe your computation specifically and how you assess it relative to the criteria in the video.

  Adjusted Total Votes
  - =IF(REGISTRATION RATE > 1, Total Reported Votes * (1 - (REGISTRATION RATE - 1)), Total Reported Votes)
  - Purpose:
    - Identifies counties with > 100% registered voters relative to voters alive
    - Readjusts the total votes cast down by the proportion that is over 100% in order to provide a more realistic representation of the voter population.

  Correlated Obama Percentage
  - =IF(REGISTRATION RATE > 1, Obama % * (1 + Obama Correlation Modifier), Obama %)
  - Purpose:
    - Adjusts the proportion of total votes earned by Obama down based on the positive correlation coefficient.

  Correlated Romney Percentage
  - =IF(REGISTRATION RATE > 1, Romney % * (1 + Romney Correlation Modifier), Romney %)
  - Purpose:
    - Adjusts the proportion of total votes earned by Romney up based on the negative correlation coefficient.

- Explain why this computation makes sense, in a few short sentences.

  - For a given county in the voting data, if the registration % is greater than the total population, determine if there is a correlation between over-registration and votes earned by each candidate. If so, modify the votes earned by that candidate by the inverse of the correlation. For example, if Obama took 72% of the votes in a county where voter registration was 130% of the population, where a correlation analysis indicates that

Obama votes have a correlation of 0.13, reduce the total votes in the county by 30%, and reduce Obama's voter % by 13%.

  - ***In our first experiment**, we modified the total votes earned by each candidate based on a correlation analysis on over-registered counties and votes earned. We found a positive correlation of 0.1303 for Obama and -0.1319 for Romney. We then adjusted the total eligible votes in each county based on its percentage over-registered to get a more realistic representation of the eligible population.*

  - ***Next**, we applied the correlation coefficients to the modified voter turnout to adjust the proportion of each county's total votes earned by each candidate. This first experiment resulted in fewer eligible votes for Obama (roughly 300,000), and slightly more eligible votes for Romney (roughly 100,000), but there was still a clear victory for Obama.*

  - ***Finally**, we modified Obama's voter turnout by not only the correlation coefficient, but instead by his proportion of total voters being based on a modifier. We completely removed those voters from the possible voters Obama could have earned. While this had the greatest effect at bringing Obama and Romney closer together, it still did not grant Romney enough votes to win the popular vote.*

  - ***In each scenario**, Obama beat out Romney by a minimum of 2.9 million votes.*

3. Marrying Data Sources

In brief bullet points, list your team's responses to the following prompts.

- What data fields would your team use to tie together the two data sources?

  - Given the data from Public Interest is more limited, we'd use the state/county to tie the two data sources. Specifically, we'd need to combine the State/County fields in the Public interest data to match the format in the 2012 U.S. presidential election data. (i.e. ST County).

  - The National Politician ID (NPID) can also be used as an option to tie candidate-level votes between the Obama vs. Romney worksheet and the FULL DATA worksheet.

- What software might you use in tying the data sources together?

  - Python, by using a unique identifier such as State County (given what was provided in the data sets from the two sources).

  - Power Query and PowerPivot is an easy way to tie these data sources together.

- What issues do you see in correlating those fields due to, for example, differences in formatting?
  - The ordering of election data by candidate is out of alignment throughout the FULL DATA worksheet. For example, beginning with column O in the FULL DATA worksheet, candidate level info is inter-mingled. This makes marrying data very difficult.
  - Combining the various data types into Python may need additional time and steps to be sure that conversion of the type of data is consistent (i.e. numbers in Excel are shown as integers or float in Python).
  - Minor formatting issues seen across the data sources. For example:
    - Spacing: NM De Baca vs. NM DeBaca
    - Additional Text: LA Tensas vs. LA Tensas Parish
  - FIPS is another method to tie the data sources together but will require an additional step to add the FIPS column to the over-registration data.

4. Propose a Better Question

Respond to the following question:

- What improved research question would your team suggest?
  - What contributes to a county having more registered voters relative to the number of actual, non-deceased registered voters who are eligible in the county?
- Why is your question an improvement on the original question?
  - The data available in the two data sources cannot answer the original question posed. These two data sources should have included population data and voter turnout data so that a more accurate estimate of the relevant population (living) can be determined. By estimating this proportion of a county's population, it would be easier to determine if there is statistically significant voter fraud occurring. For example, if the actual voter turnout in a given state is found to be 62% with standard deviation 1.5, states with abnormal turnouts, such as 80%, could be reviewed to see whether an investigation into voter fraud is necessary.