2/19/2019

Shubham handa

8638369

ASSIGNMENT 2

MATHEMATICS FOR DATA ANALYSIS

Q1)

1. The main problem with 3 tables is that the no of values differs in LPP and the rest of the 2 tables. In order to solve this problem, I had to manipulate data by selecting and merging it through constituency number. The 1st part of this question I summarized the data by each constituency so it would be easy to merge.
2. In order to summarize each district by row I had to implement reshape package which can transform the whole summarize table. I then had to split the values of candidates from the parties as they are in the same column. I did this using the split operation and the separator was ‘/’.
3. All the 3 tables were merged using the vector operation in r language i.e by selecting all the columns 1 through the no of columns in a vector.
4. In order to find the weather my party lost or won the election. I calculated the max vote

Getter from parties other than mine and subtracted from the number of votes from my

Party in that constituency. If the value comes out to be negative that shows the party lost and if the value was positive it means that our party won.

1. This was the indicator used and through this another column was added that shows weather we won or not.

Q2)

1. After merging and manipulating the data the next step was to summarize data by using measures of central tendency and dispersion. For lpp2015 the data dictionary was varying for many variables so calculating mean was ineffective so in that case the mode was calculated i.e the value which occurred the most. The measures of dispersion were that I used was IQR and standard deviation. The reason for choosing this was for further analysis in future these methods are relevant in calculating other factors such as outliers etc.

2) A bar plot represents the relative comparison of each member in the data , so a bar plot

was used to represent how conservative party stood as comparison to other parties by

taking the total number of votes (in all constituency) my party got.

Q3)

1. In order to calculate and categorize the parties for next election, I calculated the

percentage by which the margin of victory occurred if the values were 12% I categorized

them as safe for upcoming election if the values were >-12 it was categorized as in play

for next election. Otherwise the results were showed no chance.

1. The number of districts lying in the above 3 categories were summarized in the form of a bar plot stating how many numbers occur in each category and comparing them to each other. You can see as the only one party can win the safe category has the least number. Proving the algorithm used to analyze could be implemented to evaluate results for next election.
2. The rational that I chose was I divided the margin of victory(number of votes my party won or lost in the district) by the total of number of votes(in that district) this gives us the total percentage or gain or loss. if this gain or loss is significant

I categorized them as respectively.

Q4)

1. In order to display the correlation plot a package named “corrplot” was installed. Correlation and variance was calculated between two variables and for most of the variables listed in appendix 2.
2. In order to display the co-relation between performance and voter turnout I used a correlation matrix which displays the dependency between each of them.
3. Also, a summary of variables in a histogram format was displayed with the frequency being a line histogram.

Q5)

1) All the parties and their respective leaders were analyzed and their dependence

on each other. The list of variables that were analyzed can be seen in the script and related

graphs were plotted.

2) The reason for choosing this hypothesis is the fact I wanted to analyze how an individual

Candidate affects the likeness of party. The leaders were confiting with the party that is

parties\_2 is conservative but leaders\_1 (Stephen harper) is the leader for conservative and

not leaders\_2(Justin Treadeau). This was done through analyzing data. Through calculating

and analyzing I found that the Variables are insanely related to each other (more than 0.8 for

each case). Also, I wanted to calculate how population and voter turnout and voter turnout

and performance are affected. The results for the same were also displayed.