Project 4 DS 1300 sp25

This project will be presented like project 2 and you will be using the same data set as problem 2. You can use the clusters that you found in project 2 or not, it is your choice.

COMPONENT: You need to conduct at least 3 tests for relationships (correlation, independence, independence for ordinal data or ANOVA). This can be done on variables that are currently in the dataset or with ones that you create from current variables or ones that you add to the data.

COMPONENT: You need to bring something new to the table.

One option is to add more variables to your data set. For some of the data sets, I have variables you can add.

- For the real estate data, you could add a column that classifies the neighborhoods with something like distance to schools, or average income of residents
- For the hate crime data, you could add information about the states such as if they are considered red, blue, or swing states. You could add information about the average cost of living for the state.
- For the Bob Ross data you could add the average rating for the episodes.

Another option is to add more rows to your data set (for this option you need to add at least 10 more rows).

- For the real estate data, you can look on sites such as Zillow and find the same information that you currently have on the most recently sold 15 homes.
- For the recent graduate's data, you could see if you can get the same types of information about SMU graduates (or other Texas Universities) to compare or maybe you can find some information about majors that are not included in that dataset.
- For the Washington State electric vehicle data, you can find some current information on newer cars that have come out since the dataset was compiled.

You can use an analysis that is new for the class. I can help you with this.

- For the tweet data, predict the probability of the unknowns being in one location as opposed to the US.
- For the diabetes dataset, predict the probability of a person having heart disease given they are diabetic with other risk factors.

INCLUDE IN PRESENTATION: Make sure variables are treated correctly for their types with the graphs and tests. Make sure all graphs have titles and axis labels. Double check the slides for accuracy.

- Include an explanation of what the added information.
- Include at least 5 graphical displays. One that relates to each of the 3 tests you conduct and 2 that relate to the additions to your data. Make sure to mention any key details from the graphs such as patterns or outliers.
- At least one of your graphs should have an "extra" such as an emphasized group, labelled outliers, arrow pointing to a point, or be a different type of graph such as a map.

- Each test needs an explanation of what you are testing if the results are significant and what significance means in this case.
- You should summarize along the way but also have a summary at the end with key points/results.

GRADING: This is your final and will be graded a little harsher than the other projects. You need to show what you have learned. Tests and graph choices should have a reason of why you chose those methods/variables.

- 15 points for the added variables, rows and/or analysis including reasoning for choice.
- 35 points for the 5 graphs
- 30 points for the three tests, including explanations
- 10 points for the summary
- 10 points for the overall presentation