DAB501 - Project #2 – Univariate Analysis with Data Transformation

Instructions

1. The data for this project should come from the data sets used in Project #1.
2. Each group should use a **single dataset**.
3. **Univariate Analysis**
    a. There should be 1 numeric and 1 categorical variable explored per group member.
    b. All variables should be unique
    c. See below for what to include for each analysis in this section
4. Remember that this project is about understanding a particular dataset. Each submission should reflect this objective and not be a sequence of isolated exercises where no knowledge has been shared between group members.
5. All plots should be properly labeled but do not need to have a caption.

Project Submission

1. There should be 1 submission per group.
2. Each submission should be a **zip** (NOT .rar) file which includes the following:
    a. a well-organized and well-formatted HTML version of an R Notebook that:
        i. includes all necessary code for the project;
        ii. all answers to the project questions
        iii. any references used in the completion of the project
    b. the associated .Rmd file
    c. the data sets used in **.csv** format
    d. a completed version of the **academic integrity** statement shown below

Academic Integrity

Replace the underscores below with each group member's name acknowledging that each person has read and understood the statement in the context of St. Clair College's Academic Integrity policies.

We, ____________________, ____________________, and ____________________, hereby state that we have not communicated with or gained information in any way from any person or resource that would violate the College's academic integrity policies, and that all work presented is our own. In addition, we also agree not to share our work in any way, before or after submission, that would violate the College's academic integrity policies.

Univariate Analysis

For each numeric variable:

1. Create an appropriate plot to visualize the distribution of this variable. (4 marks)

2. Consider any outliers present in the data. If present, specify the criteria used to identify them and provide a logical explanation for how you handled them. (4 marks)
3. Describe the shape and skewness of the distribution. (2 marks)
4. Based on your answer to the previous question, decide if it is appropriate to apply a transformation to your data. If no, explain why not. If yes, name the transformation applied and visualize the transformed distribution. (This video and this video may help.) (4 marks)
5. Choose and calculate an appropriate measure of central tendency. (3 marks)
6. Explain why you chose this as your measure of central tendency. Provide supporting evidence for your choice. (4 marks)
7. Choose and calculate a measure of spread that is appropriate for your chosen measure of central tendency. Explain why you chose this as your measure of spread. (2 marks)

For each categorical variable:

1. Create an appropriate plot to visualize the distribution of counts for this variable. (4 marks)
2. Create an appropriate plot to visualize the distribution of proportions for this variable. (4 marks)
3. Discuss any unusual observations for this variable? (2 marks)
4. Discuss if there are too few/too many unique values? (2 marks)