PHP2510

EXPLORATORY DATA ANALYSIS WEEK 5

DATA DETECTIVE CHALLENGE

This activity challenges you to use visualizations to uncover interesting aspects of a dataset. The objective is not to find one specific answer, but to improve your ability to visualize and interpret interesting patterns and relationships within data.

DIRECTIONS

Select one of the following datasets to investigate.

- Option 1: MASS::epil
    - Description: Data from a clinical trial on epilepsy
    - The Question: What factors are most associated with a reduction in seizures?
- Option 2: MASS::Pima.te
    - Description: A health dataset on Pima Indian women who were tested for diabetes
    - The Question: What are the most significant health indicators of diabetes in this population?
- Option 3: MASS::OME
    - Description: A study of auditory perception in children with OME.
    - The Question: Which children had the poorest performance on the test?

Notes:

- You are not expected to answer the question - that just motivates the data collection
- Not inspired by the choices above? Feel free to choose a different dataset.

CODING REQUIREMENTS

Use R to create and interpret a series of plots to answer the following questions.

1. Univariate Analysis

Your first task is to understand the distribution of key individual variables.

- Choose a few variables that you think are central to the question. Create a histogram (or other appropriate visualizations) of the variables. Describe the shapes. Are there any obvious outliers? What distributions seem reasonable to model your data?

2. Bivariate Analysis

Now, begin to look for relationships between variables.

- Choose numeric variables from your dataset. Create a scatterplot (or other appropriate visualizations) to determine their relationship. Is there a positive, negative, or no correlation? Is the relationship linear or non-linear?
- Choose a numerical variable and a categorical variable from your dataset. Create a boxplot to compare the distribution of the numerical variable across the different categories. What does this tell you about the difference between the groups?

3. Exploration

Explore the data freely to find what you believe are interesting and informative relationships.

- Create additional visualizations that you believe are compelling or unexpected. This could be a different plot type or a plot that explores a relationship not covered in the previous steps. Explain what you see in words.

4. Final Report

Make a slide to summarize your findings. Include relevant R code in the speaker notes. Submit your slide on Canvas as part of your participation grade.