Assignment 1

2023-08-31

Overview

This assignment will give you a chance to apply the techniques discussed in lecture. You will find an external dataset, provide summary statistics and visualizations for the data and create a linear model for that dataset. You are required to submit the assignment using RMarkdown (HTML, PDF or Word output).

Guidelines

1. You may acquire an **external** dataset using any method. The dataset must be included with submission. If you'd like to download the data directly from the internet, the function requries you to specify the URL, which is sufficient evidence. If you're reading in a `.csv` file, attach the file with your submission and include the source in the assignment. Do not use a dataset discussed in class (Latte, Big Mac Index, Wine). Some sources for data include:

   - UCI Machine Learning Repository (https://archive.ics.uci.edu/)
   - Kaggle datasets (https://www.kaggle.com/datasets) - the repository can filter based on features (such as just returning `.csv` files)
   - Google dataset search (https://datasetsearch.research.google.com/)

   The dataset requires a minimum of 300 observations and at least 4 variables. 3 of the variables need to be numerical variables (preferably continuous numerical variables) and 1 variable needs to be a variable that can be used as a factor (character, boolean, or an integer variable with levels to it). You don't have to use ALL the variables in the dataset. If the dataset has more than 4 variables, just pick 4 that you want to explore and/or provide the most insight into the data. If the dataset has many variables, you can reduce it down to the variables you are going to use but include a step in the code where the dataset is reduced to the working variables.

2. Provide a summary of the data. This includes the `str()` and `summary()` functions.

3. Pick a numerical column of your dataset and identify outliers. Use the three-sigma, Hampel or boxplot methods for identification and be sure to identify the method utilized.

4. Provide visualizations for some variables. At a minimum, include:

   - One scatterplot between two variables
   - One boxplot between two variables
   - One barplot providing the count of your factor variable

   Additional plots (Q-Q, histogram, density) are encouraged. Label the X and Y axes with an appropriate name for the variables. Include a discriptive title for each visualization.

5. Create a linear model for two of the numerical variables. Summarize the model using the `summary()` function. Feel free to use least-squares regression or robust regression. Include a scatterplot of your data with the regression line included (axes labeled, title as well)