Project 2

In this project you will use one of the datasets available from Kaggle which contain numerous entries of used car sales:

https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

Part 1.

You should first start with EDA to understand the factors affecting the pricing of cars in the market for used cars. The final goal is to build a (linear regression) model that will take in features of a car and predict the price.

Follow the following general steps:

Step 1: Reading and Understanding the Data

- Importing the data
- Understanding the structure of the data (identify the numerical and categorical variables)

Step 2: Data Cleaning and Preparation

- Checking and removing duplicates
- Checking and removing entries with missing values
- Fixing spelling errors, etc.

Step 3: Visualizing the data

- Histograms
- Boxplots
- Scatter plots

State the initial inferences based on your visual analysis such as:

- Toyota is a preferred car company
- Sedan is preferred car type
- Convertible type has higher average price than hatchback

Make observations from scatter plots about variables that correlate with car prices and generate a list of such variables. Some of the (numerical) variables that should correlate well with price are mileage and age of the car and as well as categorical variables such as Make and Model.

Step 4: Perform Hypothesis testing using the initial inferences from the previous step

- Clearly state the questions you are interested in (choose 3-4 questions based on the visual analysis)
- State them using Hypothesis Testing framework
- Perform the test using R and explain the results

Part 2

Step 5: Model Building

Pick several (3 - 5) categories from your dataset (e.g. ‘Toyota, Sedan’, ‘Honda SUV’) and 2 numerical variables, mileage and car age. Follow the example of ‘LinearRegressionExample.Rmd’ to generate a linear model for price using mileage and age as explanatory variables. Display the boxplot of residuals for each selected category on the same plot for visual comparison of the performance between selected categories.

Step 6: Residual Analysis of Model

Multiple regression methods for linear model generally depend on the following four assumptions:

1. the residuals of the model are nearly normal,
2. the variability of the residuals is nearly constant,
3. the residuals are independent,
4. each variable is linearly related to the outcome.

Follow the logic of Open Intro Stat book, Section 8.2 (included in BB posting) to check your model.

Part 3

Step 7: Prediction and Evaluation

(will be outlined later)