Assignment 2

Please submit two documents: Your answers to the questions in .pdf or .doc format, and your R script for the questions, in .R format. In your document with answers, **please do NOT respond with R output only**. While it is okay to include R output in that document, please make sure you spell out the response to the question asked. Please submit your assignment through Canvas and name your files using the convention LastName_FirstName_AssignmentNumber. For example, Yazdi_Mohammad_2.pdf and Yazdi_Mohammad_2.R.

In this assignment, you will work on "AutoLoss.csv" data which is posted on Canvas. This data file is adapted from a data set of loss payments made by insurance companies. An explanation of the variables in this data set is included below. If you wish to know more about this data set, see the following url: https://archive.ics.uci.edu/ml/datasets/Automobile.

Description of the AutoLoss data set

Each row represents one particular type of vehicle. The columns in the data set are as follows:

- Losses: The losses covered by the insurance company for this vehicle
- Fuel type: whether the vehicle has gas or diesel engine
- Aspiration: shows if the vehicle is standard or turbo
- NumDoors: whether the vehicle has two or four doors
- BodyStyle: whether the vehicle is convertible, hardtop, hatchback, sedan, or station
- DriveWheels: whether the vehicle is front-wheel drive, rear-wheel drive, or four-wheel drive
- Length: length of the vehicle
- Width: width of the vehicle
- Height: height of the vehicle
- Weight: weight of the vehicle
- EngineSize: Engine size of the vehicle
- Horsepower: Horsepower of the vehicle
- PeakRPM: The peak RPM the vehicle can reach
- Citympg: The mpg of the vehicle in city driving
- Price: The price of the vehicle

To start your work on this assignment, read the data in AutoLoss.csv to a data frame called AutoLoss. This data set has missing values, marked as "?" in the data file. Run the following two lines of code to first replaces ?s with NA while reading the data from the .csv file, and to remove all the observations with any NA.

AutoLoss = read.csv("AutoLoss.csv", na.strings = "?", stringsAsFactors = TRUE)

AutoLoss = na.omit(AutoLoss)

Question 1)

In this question, you are going to explore the relationships among the losses for an automobile, the automobile's body style (sedan, hatchback, etc.), number of doors (two or four), and the automobile's drive wheel (whether it is front-wheel, rear-wheel, or four-wheel drive).

a) What is the number of automobiles in each body style?

b) What is the proportion of automobiles in each body style?

c) What is the average loss for each body style?

d) What is the average loss for each combination of body style and number of doors?

e) Based on parts (c) and (d), what are your observations regarding the relationships among losses, body style, and number of doors?

f) Create a subset of the data that includes only automobiles with rear-wheel drive. Focusing on this subset of the data, what is the average loss for each body style?

g) Based on parts (c) and (f), what are your observations regarding the relationships among losses, body style, and the drive wheels?

Question 2)

In this question, you will practice with some plotting commands.

a) Display a bar chart, which shows the average loss for cars with two doors versus four doors. What do you conclude from the graph?

b) Determine the 10 cars with the lowest losses and obtain a table, which shows the average loss for each body style, focusing on these 10 cars only. What are your conclusions from this table?

c) Display a boxplot, which shows the losses for each possible type of drive wheel. What do you conclude from the graph?

d) Generate a graph that represents each car in the Auto data set as a dot, with the x-coordinate being "Price" and the y-coordinate being "Losses." What can you say about the relationship between the price and losses, based on the graph only?