ESI 4606: Analytics I - Foundations of Data Science
Homework 1
Due: September 14st (11:00AM), 2022

Problem 1 (1 points)

You will now think of some real-life applications for classification, regression and clustering.

(a) Describe two real-life applications in which classification might be useful. Describe the response, as well as the predictors. Predictors need to include both qualitative and quantitative predictors.

(b) Describe two real-life applications in which regression might be useful. Describe the response, as well as the predictors. Predictors need to include both qualitative and quantitative predictors.

(c) Describe two real-life applications in which clustering might be useful.

Problem 2 (2.5 points)

This exercise involves using R to analyze the "Auto" data. Variable descriptions can be found in Table 1

Table 1: Variable descriptions for "Auto" data

Variable	Variable Description
mpg	miles per gallon
cylinders	Number of cylinders between 4 and 8
displacement	Engine displacement (cu. inches)
horsepower	Engine horsepower
weight	Vehicle weight (lbs.)
acceleration	Time to accelerate from 0 to 60 mph
year	Model year
origin	Origin of car (1. American, 2. European, 3. Japanese)
name	Vehicle name

(a) Import "Auto-HM1-2022.txt" into R. What is the sample size of the data set? Which variables are quantitative, and which are qualitative?

(b) Calculate the sample mean, the sample variance and the sample standard deviation of variable "weight".

(c) Draw histogram with relative frequency of variable "horsepower". Based on the histogram, describe the shape of the data (e.g., unimodal or bimodal; symmetric, left-skewed or right-skewed). Note: Use "breaks=10".

(d) Draw scatter plot between variable "horsepower" and variable "weight". Based on the scatter plot, describe the relationship between these two variables.

(e) Draw scatter plot between variable "weight" and variable "year". Based on the scatter plot, describe the relationship between these two variables.

(f) Draw a side-by-side boxplot by comparing "horsepower" of vehicles under different origin groups. Based on the boxplot, compare central location and dispersion among different origin groups by looking at "median" and "interquartile range".

Note: (i) Download "Auto-HM1-2022.txt" from CANVAS/Files/Assignments
(ii) **To get full points, include R codes in the appendix sections**

Problem 3 (1.5 points)

Considering a sample data with observations $x_1, x_2, ..., x_n$ and suppose that the values of the sample mean $\bar{x}$, the sample variance s_x^2 and the sample standard deviation s_x have been already calculated.

(a) Let $y_i = x_i - \bar{x}$ for $i = 1, 2, ..., n$. What are the values of the sample mean $\bar{y}$, the sample variance s_y^2 and the sample standard deviation s_y for the centered data observations y_i's? Using analytical derivation to justify your answer.

(b) Let $z_i = x_i / s_x$ for $i = 1, 2, ..., n$. What are the values of the sample mean $\bar{z}$, the sample variance s_z^2 and the sample standard deviation s_z for the scaled data observations z_i's? Using analytical derivation to justify your answer.