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) Applications of classification:

1. Determine flower species using its features:
   Response - Species (setosa, versicolor, virginica...).
   Predictors - Color (red, purple, yellow...), Sepal Length (in cm), Sepal Width (in cm), Petal Length (in cm), Petal Width (in cm).
2. Diabetes Risk Estimation:
   Response - Risk Levels (high, medium, low).
   Predictors - Body Weight (in kg), Age, Sex (Female, Male), Family history (Yes, No).

(b) Applications of regression:

1. House pricing:
   Response - Price (in dollars).
   Predictors - Area (sq.ft), Number of Bedrooms, Local Population, House age, House Type (single house, townhouse, condo...).
2. Product online sales forecast:
   Response - Product sales volume through online shopping.
   Predictors - Number of Being viewed, View Duration (in min), Number of Being Added to Cart, Number of regular buyers, Discount, Product Review.

(c) Applications of clustering:

1. Construct the phylogenetic tree:
   Cluster according to the edit distances between DNA sequences (DNA similarities)

2. Clustering of river pollution types:
   Cluster according to the Euclidean distance of multiple detected indicators (such as total phosphorus, total nitrogen, oxygen content, etc.) of each river channel

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?

1. Sample size = 350.
2. Quantitative Variables: mpg, displacement, horsepower, weight, acceleration.
   Qualitative Variables: origin, name.
   Quantitative/Qualitative: cylinders, year.

```
# Import file
> df <- read.table('Auto-HM1-2022.txt', header = T, sep = '\t')
# Sample size of the data set = the number of rows
> dim(df)[1]
[1] 350
# Which variables are quantitative, and which are qualitative?
> str(df)
'data.frame':   350 obs. of  9 variables:
 $ mpg          : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders    : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower   : int  130 165 150 150 140 198 220 215 225 190 ...
```

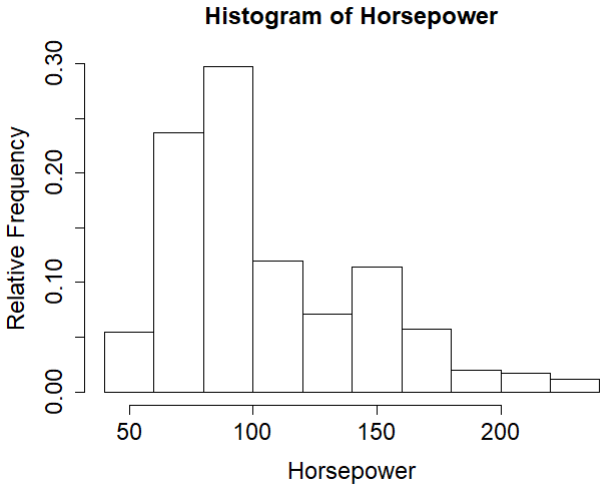
```
 $ weight       : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year         : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name         : chr  "chevrolet chevelle malibu" "buick skylark 320"...
```

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

1. Sample mean = 3024.266
2. Sample variance = 764272.9
3. Sample standard deviation = 874.227

```
> mean(df$weight, na.rm = T)
[1] 3024.266
> var(df$weight, na.rm = T)
[1] 764272.9
> sd(df$weight, na.rm = T)
[1] 874.227
```

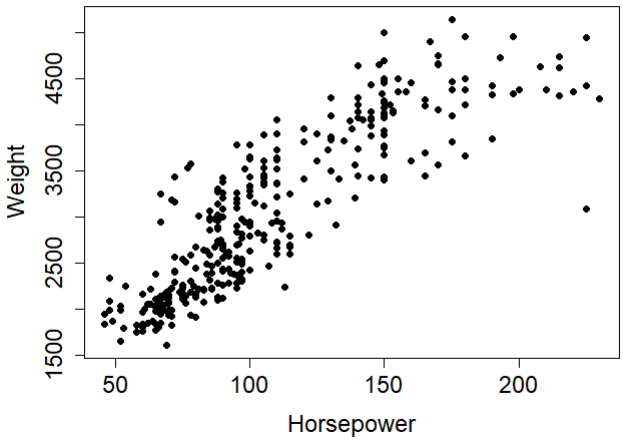
(c) Draw histogram with relative frequency of variable "horsepower". Based on the histogram, describe the shape of the data (e.g., unimodal or bimodal; symmetric, left-skewed or right-skewed). Note: Use "breaks=10".



According to the histgram, the data is bimodal, right-skewed.

```
> h = hist(df$horsepower, breaks = 10)
> h$counts = h$counts/sum(h$counts)
> plot(h, ylab = "Relative Frequency", xlab = "Horsepower",
       main = "Histogram of Horsepower", cex.lab=1.5, cex.axis=1.5,
       cex.main=1.5, cex.sub=1.5)
```

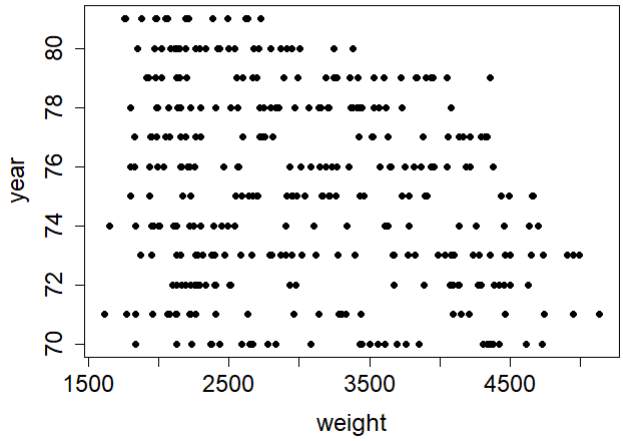
(d) Draw scatter plot between variable "horsepower" and variable "weight". Based on the scatter plot, describe the relationship between these two variables.



Based on the scatter plot, there is approximately linear relationship between variable "horsepower" and "weight".

```
> plot(df$horsepower, df$weight, pch = 19, xlab = "Horsepower",
       ylab = "Weight", cex.lab=1.5, cex.axis=1.5, cex.main=1.5,
       cex.sub=1.5)
```

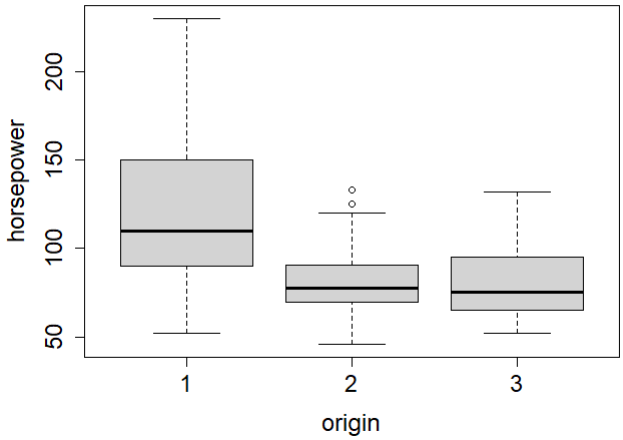
(e) Draw scatter plot between variable "weight" and variable "year". Based on the scatter plot, describe the relationship between these two variables.



There is no explicit relationship between variable "weight" and "year".

```
> plot(df$weight, df$year, pch = 19, xlab = "weight", ylab = "year",
       cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
```

(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".



According to the boxplot, the median of housepower of cars from origin 1 > origin 2 > origin 3. The dispersions are: origin 1 > origin 3 > origin 2.

```
> boxplot(horsepower ~ origin, data = df, cex.lab=1.5, cex.axis=1.5,
       cex.main=1.5, cex.sub=1.5)
```

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 $x_1, x_2, \ldots, x_n$ 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.

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n} \cdot n \cdot \bar{x} \\
&= \bar{x} - \bar{x} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= s_x^2
\end{aligned}$$

$$s_y = \sqrt{s_y^2} = \sqrt{s_x^2} = s_x$$

(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.

$$\begin{aligned}
\bar{z} &= \frac{1}{n}\sum_{i=1}^{n} z_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{s_x} \\
&= \frac{1}{s_x}(\frac{1}{n}\sum_{i=1}^{n} x_i) \\
&= \frac{\bar{x}}{s_x}
\end{aligned}$$

$$\begin{aligned}
s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \bar{z})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} (\frac{x_i}{s_x} - \frac{\bar{x}}{s_x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_x^2} \\
&= s_x^2 \cdot \frac{1}{s_x^2} \\
&= 1
\end{aligned}$$

$$s_z = \sqrt{s_z^2} = 1$$