Assignment 4 - 20%

This assignment consists of 2 sections. There are total of 40 marks, with 20 marks per section.
Instructions: Provide answers to the questions, take screenshots of the R code and the output produced. Submit as a single PDF or Word file.

Section 1:

In this section you will apply the holdout method to a dataset. There are 20 marks total for this section.

Problem:

To classify the species of Iris by using their Sepal Length, Sepal. Width, Petal. Length, and Petal. Width.

Dataset:

Iris dataset is available with R installation in the dataset packages. Run the following commands to get the dataset.

```
library(datasets)
ir_data<- iris
```

Section 1: 20 Marks

1a) How many rows and columns are there in ir_data ? (1 marks)

There are 5 columns and 150 rows.

1b) How many cases of each of the species are there in the dataset? (1 marks)

There are three species in the iris dataset.

2a) Partition the dataset into training dataset (60%), validation dataset (20%) and test dataset (20%). Use simple random number generator in this case. (4 marks)

See R code attached to assignment.

2b) How many observations are there in each of the datasets? (1 marks)

The test and validation datasets have 30 observations respectively. The training dataset has 90 observations.

2c) How many cases of each of the species are there in each dataset? (3 marks)

There are 27 cases of setosa, 34 cases of versicolor and 29 cases of virginica in the training dataset.

There are 11 cases of setosa, 8 cases of versicolor and 11 cases of virginica in the validation dataset.

Tthere are 12 cases of setosa, 8 cases of versicolor and 10 cases of virginica in the test dataset.

3a) Partition the dataset into training dataset (66%) and test (34%) using caret package and related functions. (4 marks)

3b) How many observations are there in each of the datasets? (1 marks)

The validation dataset has 60 observations while the training set has 90 observations.

3c) How many cases of each of the species are there in each dataset? (2 marks)

There are 17 cases of setosa, 17 cases of versicolor and 17 cases of virginica in the training dataset.

There are 33 cases of setosa, 33 cases of versicolor and 33 cases of virginica in the test dataset.

4a) What is the difference between the datasets obtained by random sampling in 2c and using stratified random sampling using caret package in 3c. (3 marks)

In example 3c, a stratified random sample is used. The **createDataPartition** function subsets the data while respecting important groupings – in this case, species. As a result, the caret package divides the population into smaller groups, or strata, based on this shared characteristic. Since we set the partition threshold as 66% of the dataset, caret pools 50 setosa observations, 50 versicolor observations and 50 virginica observations (species), then randomly selects 66% of the observations from each of these strata, adding them to the training dataset. As a result, we end up with 17 observations within each of these strata.

In dataset 2c, a simple random sample is used. As a result, the dataset represents the entire data population and randomly selects individuals from the population without any other consideration based on the thresholds in place. In this case, we set a threshold of 60% for the training dataset using base R so the training dataset contains 30 observations of flowers (regardless of species). This means that unlike the stratified sample, we have unequal proportions of species e.g. we may have 10 cases of setosa, 9 versicolor cases and 11 virginica cases.

Section 2: 20 marks

In this section you will formulate a linear programming model using R to solve the business problem. There are 20 marks total for this section.

Business Problem:

A computer company produces laptop and desktop computers. The data analytics team in collaboration with the marketing team developed a predictive model that forecasts the expected demand for laptops to be at least 1,000, and for desktops to be at least 800 per day. (Assumption: production is in accordance with the demand)

The production facility has a limited capacity of no more than 2,000 laptops and 1,700 desktops per day. The Sales Department indicates that contractual agreements of at most 2,000 computers per day must be satisfied.

Each laptop computer generates $600 net profit and each desktop computer generates $300 net profit.

The company wants to determine how many of each type should be made daily to maximize net profits.

1. Identify which type of analytics/modeling/solution can be used to solve this problem? (2 marks)

For this case, linear programming (optimization), the most popular method of operations research, is the best way to solve this problem. Linear optimization is a type of prescriptive modelling that focuses on choosing inputs that will result in the best possible outputs, under given circumstances. It is well suited for a situation, such as this one, when the goal is to maximize the utilization of available resources.

2. Formulate a linear programming model that represents the preceding business challenge.
    a. Summarize the provided figures in the form of a table. (3 marks)

Computer	Lower Bound	Upper Bound	Profit
Desktop	800	1700	300
Laptop	1000	2000	600

    b. Identify the decision variables. (1 mark)

A decision variable in linear programming problems affects the quantity being optimized. The objective of solving a linear programming problem is to find a set of decision variables that will produce the optimal output. For the above example, the total number of units for laptops and desktops denoted by X & Y respectively are my decision variables.

$$X = Number\ of\ laptop\ units\ produced$$

Y=*Number of desktop units produced*

c. What is our objective function in terms of these decision variables, write down its mathematical equation.

The decision variables can be arranged in a mathematical equation to give us the objective function. An **objective function** in linear programming defines the quantity that we wish to optimize—maximize or minimize. It is expressed as a linear equation in terms of the decision variables.

In the above problem, we wish to maximize the profit, P. Our objective function is as follows:

$$Max\ P = 600X + 300Y$$

d. What are the constraints for this problem, write down their respective mathematical equations. (3 marks)

Constraints:

Expected demand x>=1000 and y>=800

Production capacity x<=2000, y<=1700

Contractual agreements x + y <=2000

Non-negative restrictions x>=0 and y>=0

3. Solve the problem using lpSolveAPI library in R to come up with the most optimum solution for this problem
   a. Write down the R program to get the most optimal solution (4 marks)

See R code attached to assignment.

b. What is the optimal number of laptop and desktop computers to be made each day (2 marks)

The factory should produce 1200 laptops and 800 desktops per day.

c. What is the value of the objective function (total net profit) for the solution? (2 marks)

The maximum achievable profit is $960,000.