Practice Questions – I

Q1.)

Load the data "Fraud Detection" data and write the R code for the following:

a.) How many rows and columns does the data have?

b.) How many claims are reported for the brand Honda?

c.) Create a bar plot for b.)

d.) Create a histogram and kernel density plot for the variable "Deductible" and comment.

e.) Create a correlation plot for the data and comment.

f.) Based on the column "Year" create a new column "Years_Till_Date" by calculating the years past till the current year.

g.) View the top and bottom 10 rows of the data to ensure the new column is created appropriately.

h.) Ensure that the minimum value in "Years_Till_Date" is not less than 0.
Calculate mean, median, variance, quantiles of the same.

i.) Create a train – test (70:30) split of the data.

j.) Create a Linear Regression model to predict "Deductible".

k.) Comment on the goodness of fit of the model.

l.) Perform forward selection on the model created in h.)

m.) Create a Logistic Regression and Decision Tree model to predict "Fraud".

n.) Perform oversampling if required.

o.) Create a Confusion Matrix for the predictions made on Train and Test data for both models.

p.) Comment on the AUC value for both the models. Which model has performed better?

Q2.)
Create the following matrix in R.

	Band	Oi	Ei
1	1	95787	96807
2	2	98888	103144
3	3	103690	102264
4	4	103484	101882
5	5	102061	102810

a.) Add a new column named "Cal" which calculates:

$$\frac{(Oi - Ei)^2}{Ei}$$

b.) Calculate:

$$\sum \frac{(Oi - Ei)^2}{Ei}$$

<u>Key Points</u>:

- Know how to import files like ".csv", ".txt", ".xlsx".
- At start, perform the questions which you know.
- Write down the question no. for which the answer is written. Use same question no. as mentioned in the question paper.
- Attempt remaining subparts of a question even if you don't know the answer for the initial subpart.
- Paste the R code in your answer script in case you don't get an output.
- All required libraries will be made available during the exam.
- Practice all visuals from ggplot and well as base package:
    - Bar Plot
    - Histogram
    - Scatterplot
    - Correlation Plot (corrplot)
    - Correlation Chart (PerformanceAnalytics)
    - QQ Plot
    - Kernel Density Plot
    - Pie chart
    - Box Plot