Graded Project

Excel Project – Investment Firm

You have been hired as a Data Analyst in an investment firm and your work is to do research about 700 firms so as to help the company in investing more consciously. You are provided with the dataset containing the sales and other attributes of these 700 firms. (Refer to **700 firms level** dataset to answer the questions)

Data Dictionary for 700 Firms level dataset:

1. sales: Sales (in millions of dollars).
2. capital: Net stock of property, plant, and equipment (in millions of dollars).
3. patents: Granted patents.
4. randd: R&D stock (in millions of dollars).
5. employment: Employment (in 1000s).
6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States
7. tobinq: tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.
8. value: Stock market value.
9. institutions: Proportion of stock owned by institutions.

Project Questions:

1. Understand the Data (15 marks)

a. Data Preparation: Find if there are any null values (blank cells) in the data. If found, replace those cells with the mean value of that variable. (Note: Name the variable/s which have null values in the business report).

Also, convert categorical variables into numeric variables.
(Hint: Categorical variable is sp500 ; convert values of yes as 1 & no as 0.) – (3 marks)

b. Generate summary statistics for each of the variables and note observations. – (3 marks)

c. Plot the histogram of – (3 marks)

i) Sales Variable (original data) and write your inferences (bin size 500)

d. Compute the covariance matrix and share your observations. (Should only include continuous variable) – (3 marks)

e. Create a correlation matrix of all the variables. (Should only include continuous variable) – (3 marks)

2. Simple Linear Regression with one variable (10 Marks)

a. Build an initial regression model with Sales as the y or the Dependent variable and Value variable as the Independent Variable. (4 marks)

b. Interpret the Regression Summary Output in terms of adjusted R-square, variance explained, coefficient value, Intercept and the Residual plot (3 marks)

c. Is Value variable significant for the analysis based on your model? (HINT: Significant variables are those whose p-values are less than 0.05. If the pvalue is greater than 0.05 then it is insignificant) – (3 marks)

3. Multilinear Regression Model with all variables (10 marks)

a. Build a Regression model with all variables. Sales shall be the Dependent Variables-

(4 marks)

b. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to Sales. Explain.- (3 marks)

c. Find all the significant variables present in the model. (3 marks)

4. Best fit Model (10 Marks)

a. Make another new instance of the Regression model using only the significant variables selected in Question 3 - (4 marks)

b. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to Sales. – (3 marks)

c. Is the performance of this model better than the previous model you built in Question 2 & 3? (Hint: Compare adjusted R-square) – (3 marks)

5. Share some business recommendation based upon the analysis done on the models which are built in above questions? (5 marks)