Capstone Project

Graded Project – 90 marks

Part I – Descriptive Statistics and EDA (30 Marks)

Problem Statement:

Cold Storage started its operations in Jan 2016. They are in the business of storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavored Milk Drinks. To ensure that there is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 º - 4 º Centigrade. In the first year of business, they outsourced the plant maintenance work to a professional company with stiff penalty clauses. It was agreed that if it was statistically proven that the probability of temperature going outside the 2 º – 4 º Centigrade during the one-year contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance case). In case it exceeded 5% then the penalty would be 25% of the AMC fee. The average temperature data at the date level is given in the file “Cold_Storage_Temp_Data_.csv”
Domain: Supply Chain management

Data Description: There are 4 features

- Season – season categories
- Month – Month of the year
- Date – Date of tracking the data
- Temperature – Temperature in degrees Centigrade

Tasks/ Questions to be Answered:

Use EDA methodologies and relevant plots where ever necessary. Please ensure comments at every step and also inferences. Please note the comments and inferences needs to be present in both code file and business report.

Data Summary (4 Marks)

1. Read the data set, check shape and info, and get familiar with the data. (1 point)
2. Check the summary statistics of the data-frame and comment on your findings. (1 point)
3. Check for duplicates, unique and null values and clean the data using appropriate values. Provide comments on your approach for data imputation (Hint: Use appropriate plot) (2 point)

Descriptive Statistics (24 Marks)

4. Find mean cold storage temperature for Summer, Winter, and Rainy Season. (hint: use appropriate plot) (4 marks)
5. Find the overall mean temperature for the full year. (4 marks)
6. Find Standard Deviation of temperature for the full year. (4 marks)
7. Check for distribution,

---

a. Assuming Normal distribution, what is the probability of temperature having fallen below 2º C? (4 marks)

b. Assume Normal distribution, what is the probability of temperature having gone above 4º C? (4 marks)

8. What will be the penalty for the AMC Company? (Hint: Total probability of temperature being below 2 degree or above 4 degree) (4 marks)

Business Report & Inferences (2 Marks)

9. Business report, Inferences and observations for all the Tasks
    a. Each step needs to be executed clearly separately gradually. Also ensure you provided proper comments at each step and provide respective inferences or observations
    b. Please note the hypothesis inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
    c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
    d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
    e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis

Please Note: your optimization of code, usage of functions and packages shall be considered along with business report, inferences and suggestions during evaluation.

Part II – Inferential Statistics (20 Marks)

Problem Statement:

In Mar 2018, Cold Storage started getting complaints from their clients that they have been getting complaints from end consumers of the dairy products going sour and often smelling. On getting these complaints, the supervisor pulls out data of the last 35 days' temperatures. As a safety measure, the Supervisor has been vigilant to maintain the mean temperature 3.9º C or below.

Assume 3.9º C as the upper acceptable mean temperature and at alpha = 0.1 do you feel that there is a need for some corrective action in the Cold Storage Plant or is it that the problem is from the procurement side from where Cold Storage is getting the Dairy Products. The data of the last 35 days is in "Cold_Storage_Mar2018_.csv"

Domain: Supply Chain management

Data Description: There are 4 features

- Season – season categories
- Month – Month of the year

- Date – Date of tracking the data
- Temperature – Temperature in degrees Centigrade

Tasks/ Questions to be Answered:

Data Summary (2 Marks)

1. Read the data set, check shape and info, and get familiar with the data. (1 point)
2. Check the summary statistics of the data-frame and comment on your findings. (1 point)

Inferential Statistics

3. Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer. (6 marks) (Descriptive)
4. Perform the Hypothesis Testing
    a. State the Hypothesis (2 Marks)
    b. Perform necessary calculations to accept or reject the corresponding null hypothesis. (6 marks)

Business Report & Inferences (4 Marks)

5. Business report, Inferences and observations for all the Tasks
    a. Each step needs to be executed clearly separately gradually. Also ensure you provided proper comments at each step and provide respective inferences or observations
    b. Please note the hypothesis inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
    c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
    d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
    e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis

Please Note: your optimization of code, usage of functions and packages shall be considered along with business report, inferences and suggestions during evaluation.

Part II – Inferential Statistics (40 Marks)

Problem Statement:

You are a part of an investment firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

Domain: Insurance Claims

Data Description: The firm Level Data set has the following features

1. sales: Sales (in millions of dollars).
2. capital: Net stock of property, plant, and equipment.
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**.**

Tasks/ Questions to be Answered:

Data Summary & Exploratory Data Analytics (15 Marks)

1. Read the data set, check shape and info, and get familiar with the data. (Default requirement Marks will be deducted if not performed)
2. Check the summary statistics of the data-frame and comment on your findings. (Default requirement Marks will be deducted if not performed)
3. Describe the data briefly. (Check the null values, data types, shape, EDA). (5 Marks)
4. Perform Univariate Analysis (Hint: use all possible methods) (5 Marks)
5. Bivariate Analysis. (Hing: use all possible methods and also check for multicollinearity) (5 Marks)

Data Preparation for Model Building (10 Marks)

6. Impute null values if present? (3 Marks)
7. Try test Scaling options and confirm if you think scaling is necessary in this case? (2 Marks)
8. Encode the data (having string values) for Modelling. (3 Marks)
9. Data Split: Split the data into test and train (Hint: 70:30). (2 Marks)

Model Building & Model Performance (10 Marks)

10. Apply Linear regression. (2 Marks)
11. Check the performance of Predictions on Train and Test sets using performance metrices (Eg: MSE, R-square, RMSE etc). (4 Marks)
12. Check for important features that impact the predictor and list them down (2 Marks)
13. Drop unnecessary features and build a regressor for the best fit line (2 Marks)

Business Report & Inferences (5 Marks)

6. Business report, Inferences and observations for all the Tasks

a. Each step needs to be executed clearly separately gradually. Also ensure you provided proper comments at each step and provide respective inferences or observations
b. Please note the hypothesis inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis

Please Note: your optimization of code, usage of functions and packages shall be considered along with business report, inferences and suggestions during evaluation..