Exploratory Data Analytics Project

Graded Project – 70 Marks

Problem Statement:

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

Domain: Wholesale sales and retail trade

Data Description: You have been given a dataset of 440 distribution outlets of a wholesale store in Boston. Please refer to the data dictionary below. The data set contains the following fields

Buyer/Spender: Customer Number

Channel: Type of customer

Region: Geographical region

Fresh: Spends on the fresh products required/ ordered (in $)

Milk: Spends on the Milk and dairy Products (in $)

Grocery: Spends against Grocery Products(in $)

Frozen: Spends against (in $)

Detergents_Paper: Spends against Detergent Paper and cleaning items (in $)

Delicatessen: Spends against detergent for delicates (in $)

Tasks/ Questions to be Answered: Your key job is to analyse the extent and magnitude of each variable and its impact. For this, you have the following deliverables to execute.

Basic working on Data (10 points)

1. Import Necessary Libraries. (1 points)
2. Display a sample of five rows of the data frame. (1 point)
3. Check the shape of the data (number of rows and columns). (2 point)
4. Check the percentage of missing values in each column of the data frame. (2 point)
5. Check if there are any duplicate rows and Remove duplicates (2 point)
6. Work on Missing values and replace missing values with appropriate methods (2 Points)

Univariate Analysis (10 Points)

7. Use descriptive measure against each variable (6 given Variable) and answer the following questions

a. Plot a count Plot for the categorical variables and comment on your understanding (2 points)

b. Plot distribution of each continuous variable (6 continuous variables) for provide your comments (Eg: Left tail, right tail or normal distribution etc) (2 Points)

c. Plot Box plots for all the 6 given variables and provide inferences on the behaviour (Which variable is very in-consistent) (2 Points)

d. Provide the descriptive measures of each of these variables such as Mean, Median, Standard deviation (2 Points)

e. Calculate the Coefficient of variance for each of these continuous variables (2 Points)

Note: Provide the inferences on which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

Outlier Treatment & New Feature Creation (15 Points)

8. Outlier treatment (5 Points)
    a. Are there any outliers in the data?
    b. Use Appropriate measures for Outlier treatment
9. Create a new feature “Total Spends” and use the Pivot tables to get the aggregation of Spends as given below (10 Points)
    a. Plot Total spends against “Channels”
    b. Plot Spends against “Regions”
    c. Plot spends against with both “Channels” and “Regions” combined
    d. Which Region and which Channel seems to spend more?
    e. Which Region and which Channel seems to spend less?

Bi-variet Analysis (30 Points)

10. There are 6 different varieties of items use appropriate plots and compare the aggregate spends (Sum aggregation) of each item by category and region. Answer the following Questions based on the 6 Plots generated (Hint: Use Box Plots with channels on X axis, Value of the Items in Y axis and Region as hue) (20 Points)

    a. Use Pivot Table and find aggregate spends on fresh items (1 Points)

    b. Plot the spends on fresh products against channels and regions by colour (2 Points)

    c. Use Pivot Tables and find aggregate spends on Milk items (1 Points)

    d. Plot Spends on milk products against channels and regions by colour (2 Points)

    e. Use Pivot Tables and find aggregate spends on Groceries (2 Points)

    f. Plot Spends on Groceries against Channels and Regions by colour (2 Points)

    g. Use Pivot Tables and find aggregate spends on Frozen items (2 Points)

    h. Plot Spends on Frozen against Channels and Regions by colour (2 Points)

    i. Use Pivot Tables and find aggregate spends on Detergent Paper (2 Points)

j. Plot Spends on Detergent Paper against Channels and Regions by colour (2 Points)

k. Use Pivot Tables and find aggregate spends on Delicatessen (2 Points)

l. Plot Spends on Delicatessen against Channels and Regions by colour (2 Points)

m. Do all varieties show similar behaviour across Region and Channel? (1 Points)

11. Understand the Correlation between variables and provided insights (10 Points)
    a. Plot correlation using heat map
    b. Identify if there are highly correlating variables
    c. Identify Multi collinearity

Business Report, Inferences & Suggestions (5 Points)

12. Business report, Inferences and observations for all the Tasks
    a. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem?
    b. Each step needs to be executed clearly separately step by step. Also ensure you provided proper comments at each step and provide respective inferences or observations
    c. Please note the inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
    d. All the answers to the questions to be submitted in a sequential manner as part of the business repot
    e. 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
    f. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis