03 August 2021

SMDM Project

PGP-DSBA June-Batch

SooREJ KUNNOOL BALAN

2021

Table of Contents

[1. Problem 1 4](#_Toc78808221)

[Data Description 4](#_Toc78808222)

[Sample of the dataset: 4](#_Toc78808223)

[Exploratory Data Analysis: 4](#_Toc78808224)

[Let us check the type of variables in the data frame 5](#_Toc78808225)

[Check for missing values in the dataset 5](#_Toc78808226)

[Correlation Plot 5](#_Toc78808227)

[Pairplot 6](#_Toc78808228)

[1.1 Questions-Part 1 7](#_Toc78808229)

[1.1.1 Use methods of descriptive statistics to summarize data. 7](#_Toc78808230)

[1.1.2 Which Region and which Channel spent the most? 12](#_Toc78808231)

[1.1.3 Which Region and which Channel spent the least? 13](#_Toc78808232)

[1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer. 13](#_Toc78808233)

[1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior? 21](#_Toc78808234)

[1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 22](#_Toc78808235)

[1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective. 23](#_Toc78808236)

[2. Problem 2 23](#_Toc78808237)

[Data Description 23](#_Toc78808238)

[Sample of the Dataset: 24](#_Toc78808239)

[Exploratory Data Analysis: 24](#_Toc78808240)

[Let us check the type of variables in the data frame. 24](#_Toc78808241)

[Check for missing values in the dataset. 24](#_Toc78808242)

[Correlation Plot 25](#_Toc78808243)

[Pairplot 26](#_Toc78808244)

[2.1 For this data, construct the following contingency tables (Keep Gender as row variable). 27](#_Toc78808245)

[2.1.1 Gender and Major 27](#_Toc78808246)

[2.1.2 Gender and Grad Intention 27](#_Toc78808247)

[2.1.3 Gender and Employment 27](#_Toc78808248)

[2.1.4 Gender and Computer 28](#_Toc78808249)

[2.2 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions. 28](#_Toc78808250)

[2.2.1 What is the probability that a randomly selected CMSU student will be male? 28](#_Toc78808251)

[2.3 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions. 29](#_Toc78808252)

[2.3.1 Find the conditional probability of different majors among the male students in CMSU? 29](#_Toc78808253)

[2.3.2 Find the conditional probability of different majors among the female students in CMSU? 29](#_Toc78808254)

[2.4 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions. 30](#_Toc78808255)

[2.4.1 Find the probability that a randomly chosen student is a male and intends to graduate? 30](#_Toc78808256)

[2.4.2 Find the probability that a randomly chosen student is a female and does not have a laptop? 30](#_Toc78808257)

[2.5 Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions. 31](#_Toc78808258)

[2.5.1 Find the probability that a randomly chosen student is a male or has full time employment? 31](#_Toc78808259)

[2.5.2 Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 31](#_Toc78808260)

[2.6 Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now, and the table is a 2x2 table. Do you think the graduate intention and being female are independent events? 32](#_Toc78808261)

[2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data. 33](#_Toc78808262)

[2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 33](#_Toc78808263)

[2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more. 33](#_Toc78808264)

[2.8 Question Header 33](#_Toc78808265)

[2.8.1 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. 33](#_Toc78808266)

[2.8.2 Write a note summarizing your conclusions. 36](#_Toc78808267)

[3. Problem 3 37](#_Toc78808268)

[Data Description 37](#_Toc78808269)

[Sample of the Dataset: 37](#_Toc78808270)

[Exploratory Data Analysis: 37](#_Toc78808271)

[Let us check the type of variables in the data frame. 38](#_Toc78808272)

[Check for missing values in the dataset. 38](#_Toc78808273)

[Descriptive Statistics 38](#_Toc78808274)

[3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps. 38](#_Toc78808275)

[3.2 Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed? 39](#_Toc78808276)

[THE END! 39](#_Toc78808277)

List of Figures List of Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure-1.1 Correlation Heatmap………………………………………..………… | 5 |  | Table-1.1 Dataset Sample…………………………………………… | | 4 |
| Figure-1.2 Pairplot………………………………………………………………… | 6 |  | Table-1.2 Concise data summary…………………………………… | | 4 |
| Figure-1.3 Histogram of Fresh……………………………………………….…… | 8 |  | Table-1.3 Summary of the data……………………………………… | | 7 |
| Figure-1.4 Histogram of Milk…………………………………………………….. | 9 |  | Table-1.4 Region-wise Spending…………………………………… | | 12 |
| Figure-1.5 Histogram of Grocery…………………………………………………. | 9 |  | Table-1.5 Channel-wise Spending…………………………………… | | 12 |
| Figure-1.6 Histogram of Frozen…………………………………………………… | 10 |  | Table-1.6 Region-wise spending across product varieties…………… | | 18 |
| Figure-1.7 Histogram of Detergents\_Paper…..…………………………………... | 11 |  | Table-1.7 Channel-wise spending across product varieties…………. | | 20 |
| Figure-1.8 Histogram of Delicatessen...................................................................... | 12 |  | Table-1.8 Coefficient of Variation with summary of data…………… | | 21 |
| Figure-1.9 Region vs Total spend bar plot............................................................... | 12 |  | Table-1.9 CV of product varieties…………………………………… | | 22 |
| Figure-1.10 Channel vs Total Spend bar plot.......................................................... | 13 |  | Table-2.1 Dataset Sample……………………………………………. | | 24 |
| Figure-1.11 Region vs Spending in Fresh bar plot.................................................. | 13 |  | Table-2.2 Concise data summary……………………………………. | | 24 |
| Figure-1.12 Channel vs Spending in Fresh bar plot................................................. | 14 |  | Table-2.3 Gender vs Major contingency table………………………. | | 27 |
| Figure-1.13 Region vs Spending in Milk bar plot.................................................... | 14 |  | Table-2.4 Gender vs Grad Intention contingency table……………… | | 27 |
| Figure-1.14 Channel vs Spending in Milk bar plot.................................................. | 15 |  | Table-2.5 Gender vs Employment contingency table……………….. | | 27 |
| Figure-1.15 Region vs Spending in Grocery bar plot............................................... | 15 |  | Table-2.6 Gender vs Computer contingency table………………….. | | 28 |
| Figure-1.16 Channel vs Spending in Grocery bar plot............................................. | 15 |  | Table-2.7 Male Students vs Major contingency table………………. | | 29 |
| Figure-1.17 Region vs Spending in Frozen bar plot ................................................ | 16 |  | Table-2.8 Female students vs Major contingency table…………….. | | 29 |
| Figure-1.18 Channel vs Spending in Frozen bar plot............................................... | 16 |  | Table-2.9 Gender vs Grad Intention contingency table……………… | | 30 |
| Figure-1.19 Region vs Spending in Detergents\_Paper box plot............................... | 17 |  | Table-2.10 Gender vs Computer contingency table…………………. | | 30 |
| Figure-1.20 Channel vs Spending in Detergents\_Paper box plot............................. | 17 |  | Table-2.11 Gender vs Employment contingency table……………… | | 31 |
| Figure-1.21 Region vs Spending in Delicatessen box plot....................................... | 17 |  | Table-2.12 Gender vs Major contingency table……………………… | | 31 |
| Figure-1.22 Channel vs Spending in Delicatessen box plot.................................... | 18 |  | Table-2.13 Gender vs Grad Intention contingency table……………. | | 32 |
| Figure-1.23 Pie Chart of region-wise spending across product varieties................. | 18 |  | Table-3.1 Dataset Sample…………………………………………… | | 37 |
| Figure-1.24 Bar plot of region-wise spending across product varieties................... | 19 |  | Table-3.2 Concise data summary……………………………………. | | 37 |
| Figure-1.25 Pie Charts of channel-wise spending across product varieties............. | 20 |  | Table-3.3 Summary of the data……………………………………… | | 38 |
| Figure-1.26 Bar plot of region-wise spending across product varieties................... | 21 |  |  |  | |
| Figure-1.27 Box plot of product varieties................................................................ | 22 |  |  |  | |
| Figure-2.1 Correlation Heatmap.............................................................................. | 25 |  |  |  | |
| Figure-2.2 Pairplot................................................................................................... | 26 |  |  |  | |
| Figure-2.3 Histogram of GPA.................................................................................. | 34 |  |  |  | |
| Figure-2.4 Histogram of Salary................................................................................ | 34 |  |  |  | |
| Figure-2.5 Histogram of Spending........................................................................... | 35 |  |  |  | |
| Figure-2.6 Histogram of Text Messages.................................................................. | 36 |  |  |  | |

# Problem 1

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).



## Data Description

1. Buyer/Spender : nominal identifier for buyer in numerical format from 1 to 439.
2. Channel : types of sales channel (Hotel/Retail).
3. Region : Name of region(Lisbon, Oporto, Other).
4. Fresh : continuous from 3 to 112151.
5. Milk : continuous from 55 to 73498.
6. Grocery : continuous from 3 to 92780.
7. Frozen : continuous from 25 to 60869.
8. Detergents\_Paper : continuous from 3 to 40827.
9. Delicatessen : continuous from 3 to 47943.

## Sample of the dataset:

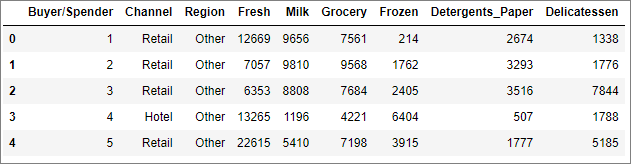


Table-1.1 Dataset Sample

Dataset has 9 variables with spend across 3 regions and 2 channels for 6 product varieties.

## Exploratory Data Analysis:

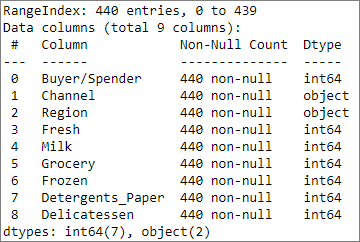


Table-1.2 Concise data summary

### Let us check the type of variables in the data frame

There are a total of 440 rows and 9 columns in the dataset. Of the 9 columns, 2 are of object type and rest 7 are integer data type.

### Check for missing values in the dataset

From Table-1.2 we can see that all the columns have 440 non-null values and hence we have no missing values in the dataset.

### Correlation Plot

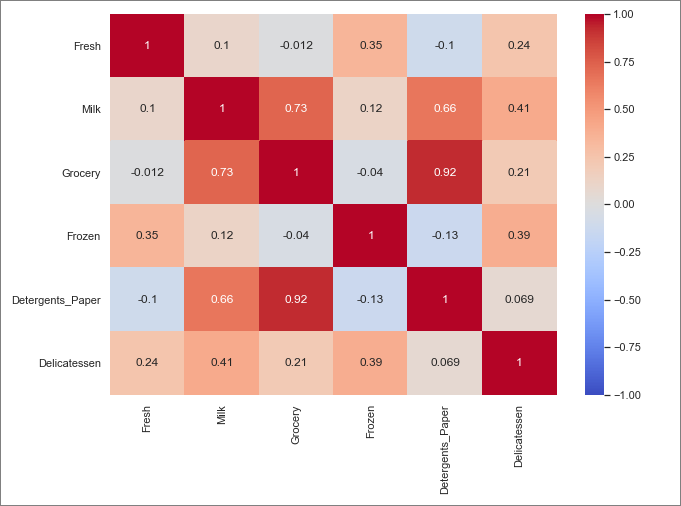


Figure-1.1 Correlation Heatmap

From the correlation plot (Figure-1.1), we can see that there is a strong positive correlation between ‘Grocery’ and ‘Deteregents\_Paper’. Also, we have moderately strong positive correlation between ‘Grocery’ and ‘Milk’ and ‘Milk’ and ‘Detergents\_Paper’. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

### Pairplot

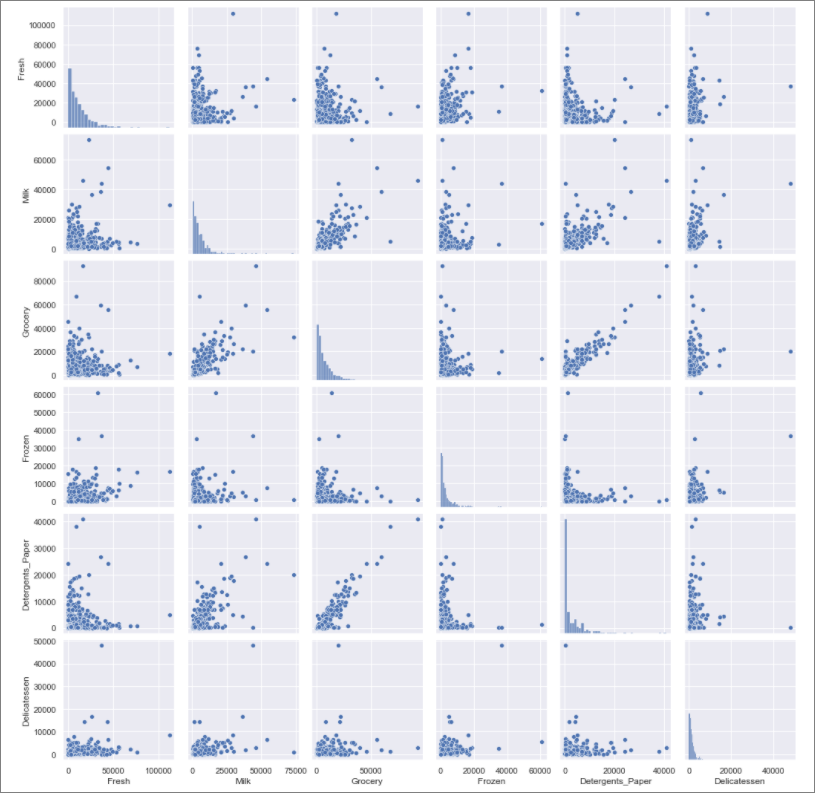


Figure-1.2 Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

From the graph, we can see that there is positive linear relationship between variables like ‘Grocery’ and ‘Deteregents\_Paper’. From the histograms we can see that the spend on all the 6 varieties of products are right skewed.

## Questions-Part 1

### Use methods of descriptive statistics to summarize data.

Descriptive statistics help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of central tendency (mean, median, and mode), and measures of variability/spread (standard deviation, variance, minimum and maximum variable) which are used at almost all levels of math and statistics.

Let us delve into the data to derive the descriptive statistics:

The below table(Table-1.3) shows descriptive statistics for all 9 columns.

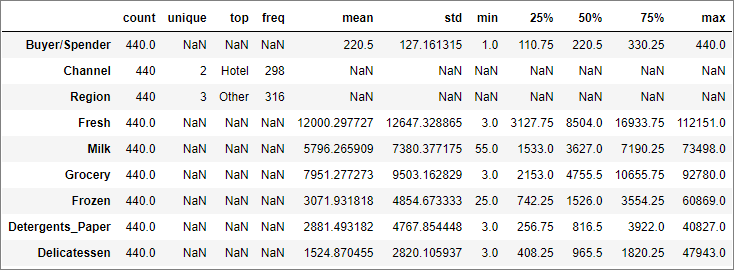


Table-1.3 Summary of the data

Let’s examine each data column individually:

* Buyer/Spender

Even though this column of integer type, its nominal in nature.

Hence the only interesting statistics is the count, which is the total number of observations – 440

* Channel

This is a categorical value and there are no missing data as per the count.

The column consists of 2 unique values, with one value ‘Hotel’ having the maximum frequency of 298, which means (298/440) 67.73% of retailers use channel “Hotel”.

* Region

This is a categorical value and there are no missing data as per the count.

The column consists of 3 unique values, with one value ‘Other’ having the maximum frequency of 316, which means (316/440) 71.82% retailers are from region “Other”.

* Fresh

This is a numerical continuous variable with no missing data as per the count.

Mean – 12000.298

Standard Deviation – 12647.329

Q1(1st Quartile) – 3127.75

Q2(2nd Quartile)/Median – 8504.00

Q3(3rd Quartile) – 16933.75

IQR(Inter-Quartile Range) = Q3- Q1 = 13806.00

Minimum value = 3.00

Maximum value = 112151.00

Range = 112151.00 – 3.00 = 112148.00

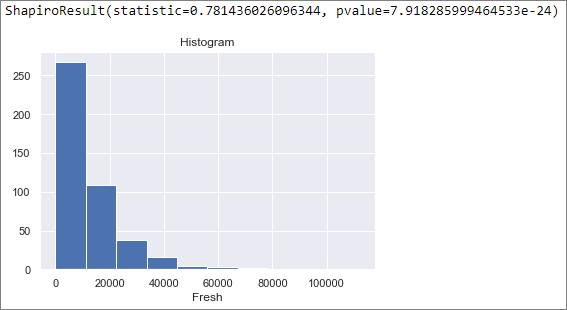


Figure-1.3 Histogram of Fresh

Figure-1.3 depicts the histogram of “Fresh” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 7.92e-24 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Fresh” column data does not have a normal distribution.

* Milk

This is a numerical continuous variable with no missing data as per the count.

Mean – 5796.266

Standard Deviation – 7380.377

Q1(1st Quartile) – 1533.00

Q2(2nd Quartile)/Median – 3627.00

Q3(3rd Quartile) – 7190.25

IQR(Inter-Quartile Range) = Q3- Q1 = 5657.25

Minimum value = 55.00

Maximum value = 73498.00

Range = 73498.00 – 55.00 = 73443.00

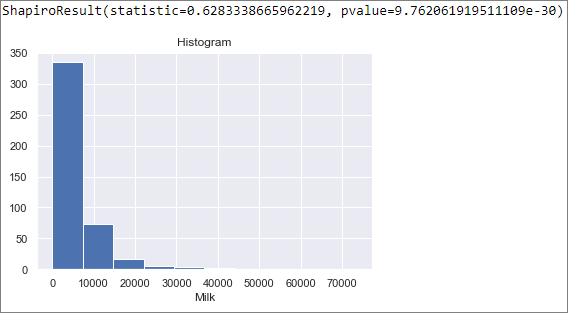


Figure-1.4 Histogram of Milk

Figure-1.4 depicts the histogram of “Milk” which appears to be normal in distribution but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 9.76e-30 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Milk” column data does not have a normal distribution.

* Grocery

This is a numerical continuous variable with no missing data as per the count.

Mean – 7951.277

Standard Deviation – 9503.163

Q1(1st Quartile) – 2153.00

Q2(2nd Quartile)/Median – 4755.50

Q3(3rd Quartile) – 10655.75

IQR(Inter-Quartile Range) = Q3- Q1 = 8502.75

Minimum value = 3.00

Maximum value = 92780.00

Range = 92780.00 – 3.00 = 92777.00

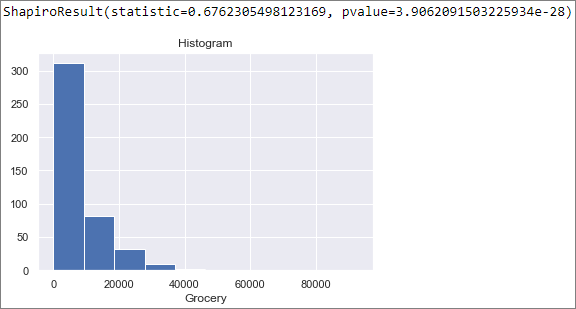


Figure-1.5 Histogram of Grocery

Figure-1.5 depicts the histogram of “Grocery” which appears to be normal in distribution but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 3.91e-28 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Grocery” column data does not have a normal distribution.

* Frozen

This is a numerical continuous variable with no missing data as per the count.

Mean – 3071.932

Standard Deviation – 4854.673

Q1(1st Quartile) – 742.25

Q2(2nd Quartile)/Median – 1526.00

Q3(3rd Quartile) – 3554.25

IQR(Inter-Quartile Range) = Q3- Q1 = 2812.00

Minimum value = 25.00

Maximum value = 60869.00

Range = 60869.00 – 3.00 = 60844.00

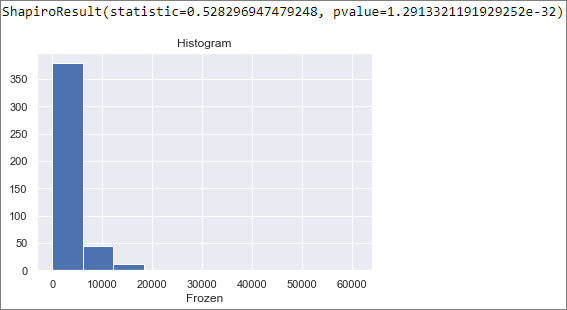


Figure-1.6 Histogram of Frozen

Figure-1.6 depicts the histogram of “Frozen” which appears to be normal in distribution but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.29e-32 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Frozen” column data does not have a normal distribution.

* Detergents\_Paper

This is a numerical continuous variable with no missing data as per the count.

Mean – 2881.4931

Standard Deviation – 4767.854

Q1(1st Quartile) – 256.75

Q2(2nd Quartile)/Median – 816.50

Q3(3rd Quartile) – 3922.00

IQR(Inter-Quartile Range) = Q3- Q1 = 3665.25

Minimum value = 3.00

Maximum value = 40827.00

Range = 40827.00 – 3.00 = 40824.00

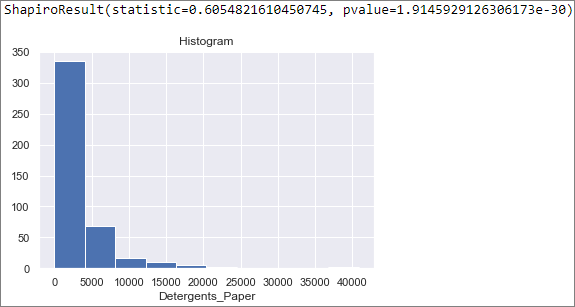


Figure-1.7 Histogram of Detergents\_Paper

Figure-1.7 depicts the histogram of “Detergents\_Paper” which appears to be normal in distribution but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.91e-30 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Detergents\_Paper” column data does not have a normal distribution.

* Delicatessen

This is a numerical continuous variable with no missing data as per the count.

Mean – 1524.870

Standard Deviation – 2820.106

Q1(1st Quartile) – 408.25

Q2(2nd Quartile)/Median – 965.5

Q3(3rd Quartile) – 1820.25

IQR(Inter-Quartile Range) = Q3- Q1 = 1412

Minimum value = 3.00

Maximum value = 47943.00

Range = 47943.00 – 3.00 = 47940.00

Figure-1.8 depicts the histogram of “Delicatessen” where we don’t have enough data to comment on the normality, but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.75e-36 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Delicatessen” column data does not have a normal distribution.

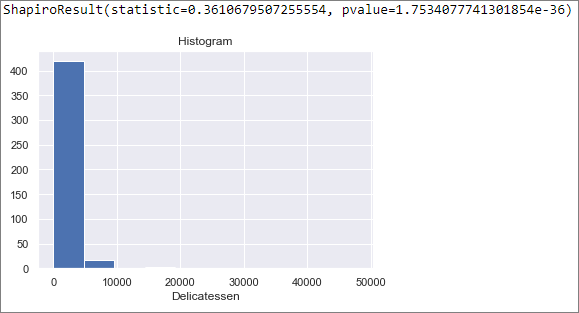


Figure-1.8 Histogram of Delicatessen

### Which Region and which Channel spent the most?

The total spending for each retailer is the sum spent on all the 6 different varieties of product.

Grouping this total expenditure across Region provides the below information:

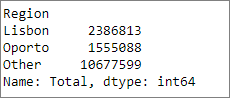


Table-1.4 Region-wise Spending

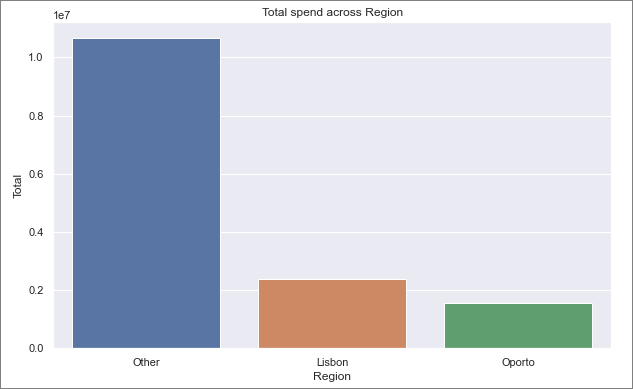


Figure-1.9 Region vs Total spend bar plot

From Table-1.4 and Figure-1.9 we can see that among regions, “Other” has spent the maximum of 10,677,599/-.

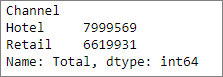


Table-1.5 Channel-wise Spending

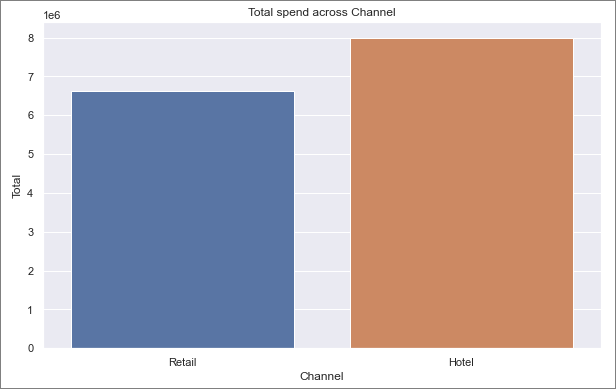


Figure-1.10 Channel vs Total Spend bar plot

From Table-1.5 and Figure-1.10 we can see that among channels, “Hotel” has spent the maximum of 7,999,569/-.

### Which Region and which Channel spent the least?

From Table-1.4 & Figure-1.9 we can see that among regions, “Oporto” has spent the least of 1,555,088/- . And from Table-1.5 and Figure-1.10, among channels, we can see that “Retail” has spent the least of 6,619,931/-.

## There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Product across Region & Channel.

Let us look at the expenditure on each product across Region and across Channel.

* Fresh

Fresh across Region shows the following data:

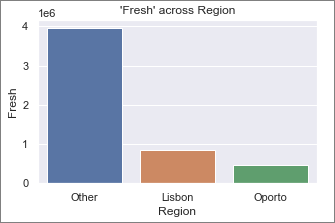


Figure-1.11 Region vs Spending in Fresh bar plot

We can see that the order of spending for Fresh across regions is in the following order:

Other > Lisbon > Oporto

Fresh across Channel shows the following data:

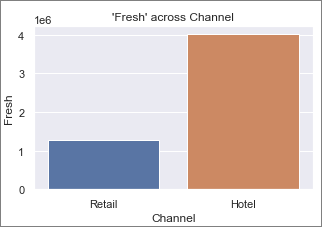


Figure-1.12 Channel vs Spending in Fresh bar plot

We can see that the order of spending for Fresh across channels is in the following order:

Hotel > Retail

* Milk

Milk across Region shows the following data:

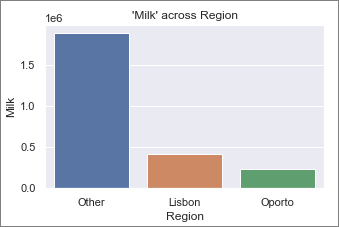


Figure-1.13 Region vs Spending in Milk bar plot

We can see that the order of spending for Milk across regions is in the following order:

Other > Lisbon > Oporto

Milk across Channel shows the following data:

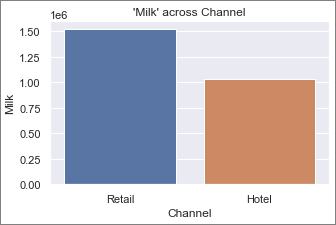


Figure-1.14 Channel vs Spending in Milk bar plot

We can see that the order of spending for Milk across channels is in the following order:

Retail > Hotel

* Grocery

Grocery across Region shows the following data.

We can see that the order of spending for Grocery across regions is in the following order:

Other > Lisbon > Oporto

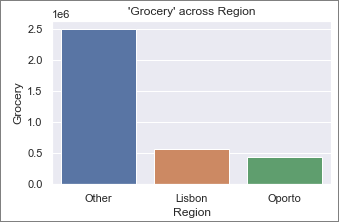


Figure-1.15 Region vs Spending in Grocery bar plot

Grocery across Channel shows the following data:

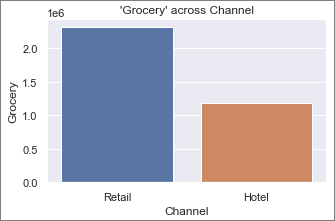


Figure-1.16 Channel vs Spending in Grocery bar plot

We can see that the order of spending for Grocery across channels is in the following order:

Retail > Hotel

* Frozen

Frozen across Region shows the following data:

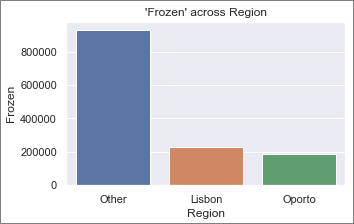


Figure-1.17 Region vs Spending in Frozen bar plot

We can see that the order of spending for Frozen across regions is in the following order:

Other > Lisbon > Oporto

Frozen across Channel shows the following data:

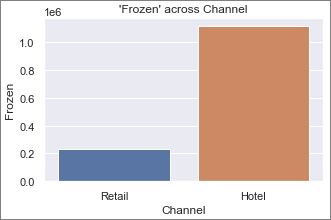


Figure-1.18 Channel vs Spending in Frozen bar plot

We can see that the order of spending for Frozen across channels is in the following order:

Hotel > Retail

* Detergents\_Paper

Detergents\_Paper across Region shows the following data:

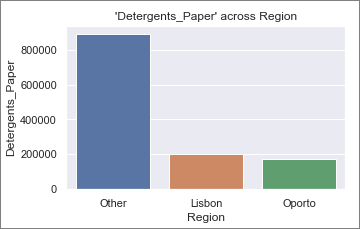


Figure-1.19 Region vs Spending in Detergents\_Paper box plot

We can see that the order of spending for Detergents\_Paper across regions is in the following order:

Other > Lisbon > Oporto

Detergents\_Paper across Channel shows the below data:

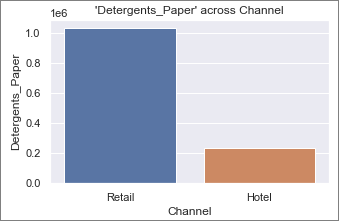


Figure-1.20 Channel vs Spending in Detergents\_Paper box plot

We can see that the order of spending for Detergents\_Paper across channels is in the following order:

Retail > Hotel

* Delicatessen

Delicatessen across Region shows the following data:

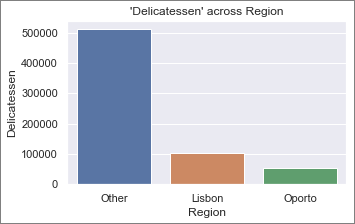


Figure-1.21 Region vs Spending in Delicatessen box plot

We can see that the order of spending for Delicatessen across regions is in the following order:

Other > Lisbon > Oporto

Delicatessen across Channel shows the below data:

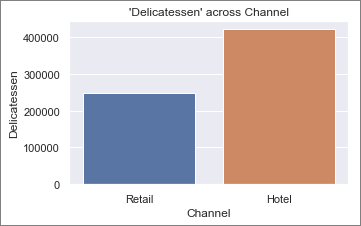


Figure-1.22 Channel vs Spending in Delicatessen box plot

We can see that the order of spending for Delicatessen across channels is in the following order:

Hotel > Retail

Products across Region

Let us look at the spending on all product varieties across regions.

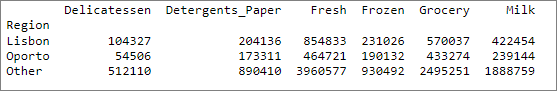
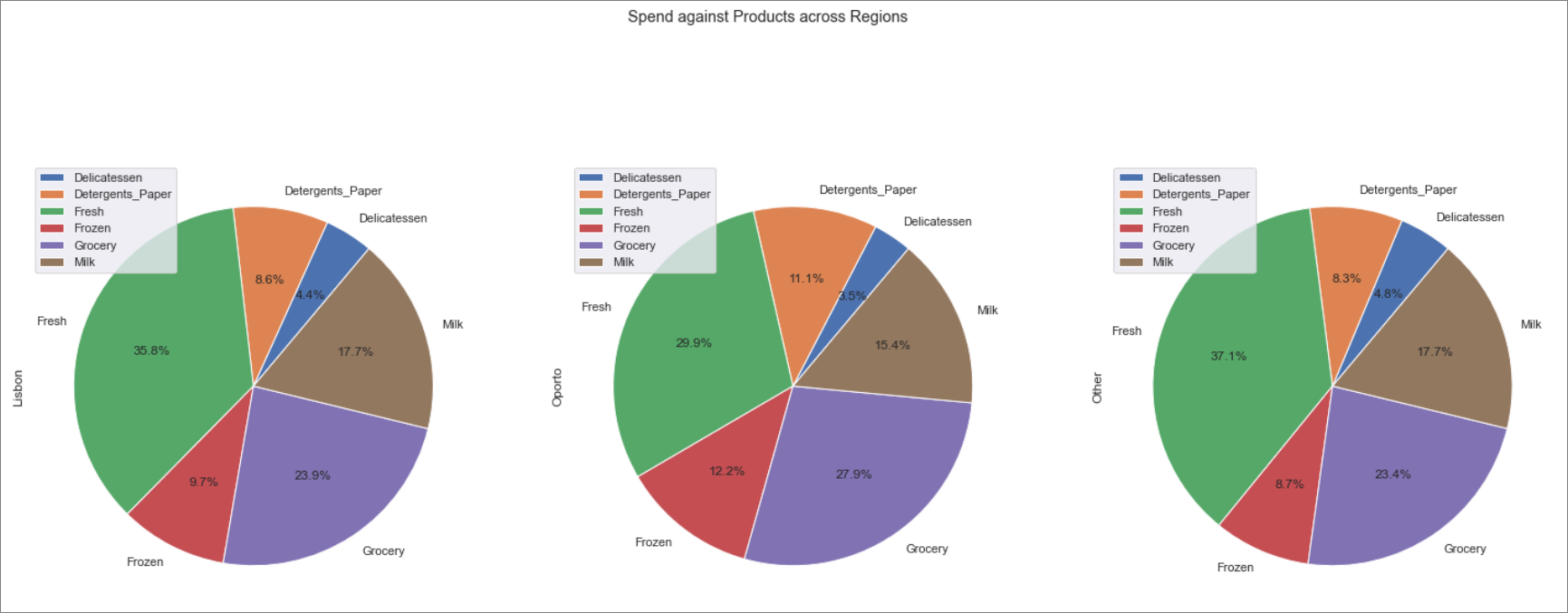


Table-1.6 Region-wise spending across product varieties

The above data is visually represented below as a pie chart & bar plot.

 Figure-1.23 Pie Charts of region-wise spending across product varieties

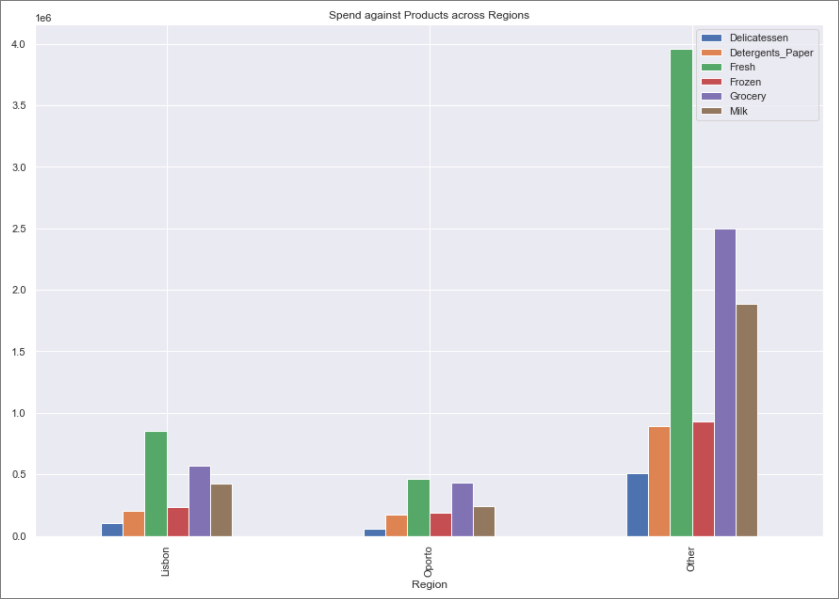


Figure-1.24 Bar plot of region-wise spending across product varieties

From Figure-1.23 and Figure-1.24, we can see spending on ‘Fresh’ product variety is highest among all the 3 regions and spending on ‘Delicatessen’ is the least in all the 3 regions.

In ‘Lisbon’, the expenditure is in the below order:

Fresh > Grocery > Milk > Frozen > Detergents\_Paper > Delicatessen

In ‘Oporto’, the expenditure is in the below order:

Fresh > Grocery > Milk > Frozen > Detergents\_Paper > Delicatessen

In ‘Other’, the expenditure is in the below order:

Fresh > Grocery > Milk > Frozen > Detergents\_Paper > Delicatessen

As we can see above, the order of expenditure over the 6 product varieties is the same across the regions.

Products across Channel

Let us look at the sale of all products across Channels.

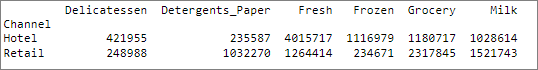


Table-1.7 Channel-wise spending across product varieties

The above data is visually represented below as a pie chart and bar plot:

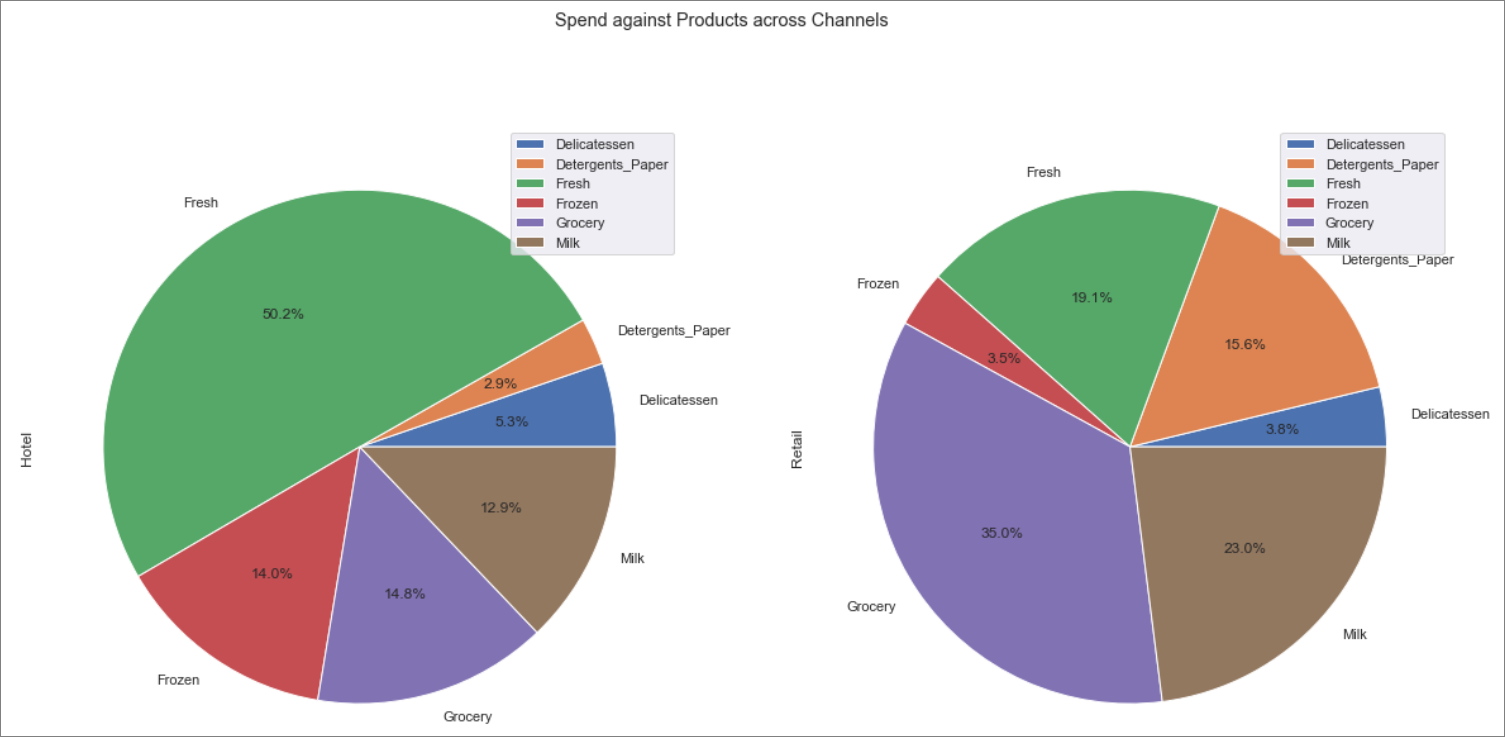


Figure-1.25 Pie Charts of channel-wise spending across product varieties

From Figure-1.25 and Figure-1.26, we can see that in channel ‘Hotel’ spending on, ‘Fresh’ product variety is highest and ‘Detergents\_Paper’ is least whereas in channel ‘Retail’, highest spend is on ‘Grocery’ and least is on ‘Frozen’.

In ‘Hotel’, the expenditure is in the below order:

Fresh > Grocery > Frozen > Milk > Delicatessen > Detergents\_Paper

In ‘Retail’, the expenditure is in the below order:

Grocery > Milk > Fresh > Detergents\_Paper > Delicatessen > Frozen

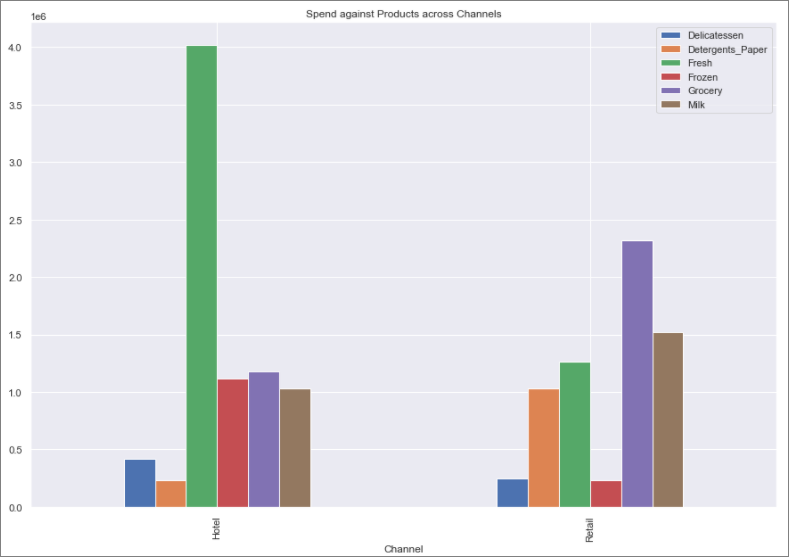


Figure-1.26 Bar plot of region-wise spending across product varieties

## On the basis of a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

Coefficient of variation (ratio of standard deviation to mean in %) gives us an insight to the level of dispersion around the mean. The lower the coefficient of variation, lesser is the dispersion around the mean which indicates higher consistent behavior.

For the given data, Coefficient of variation is calculated as column CV.

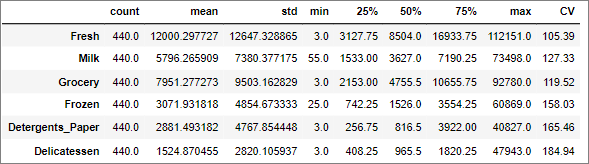


Table-1.8 Coefficient of Variation with summary of data

Arranging the CV value as per magnitude across the 6 varieties, we get the following data:

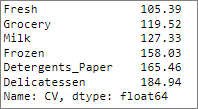


Table-1.9 CV of product varieties

Based on the above information, ‘Delicatessen’ displays the most inconsistent behavior with a CV value of 184.94% and ‘Fresh’ displays the least inconsistent behavior with a CV value of 105.39%.

## Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

To check the presence of outliers in the 6 varieties of products (which constitutes the continuous variables, the other columns are categorical in nature), we can check the boxplot of the products.

The below figure (Figure-1.27) shows the box plot of the 6 varieties of products and we can see that outliers are present in all the 6 varieties of products. Outliers are denoted by black dots in the graph for each product, for product “Fresh” the outliers have been highlighted in yellow.

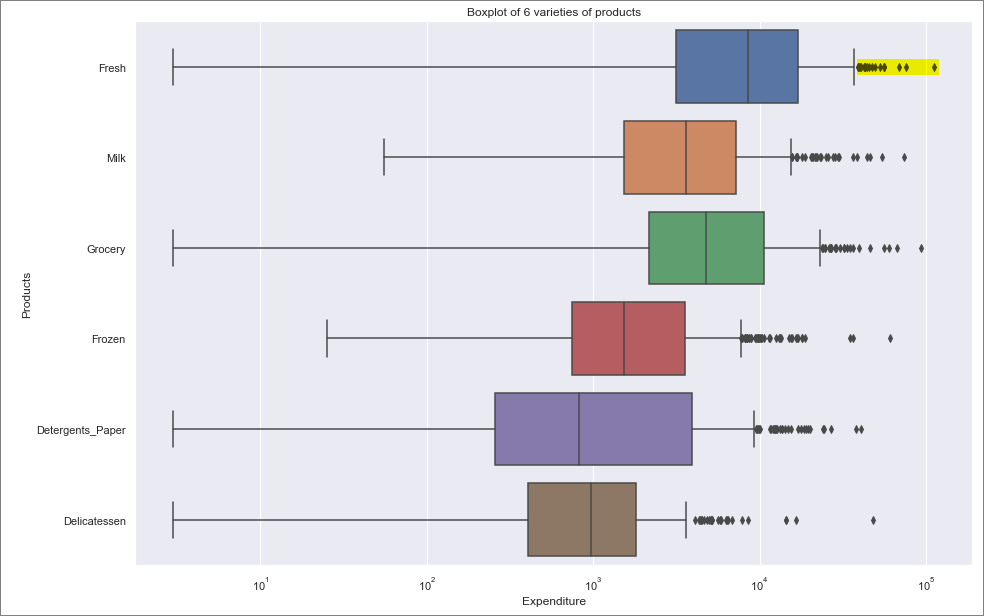


Figure-1.27 Box plot of product varieties

Outliers are values that don’t conform to most values in the dataset, and they can cause deviations in statistical analysis. Outliers could be introduced in the dataset because of data entry error or sampling error, which if it was the case should be omitted. In this scenario, I would assume that the data provided is correct and outliers should not be omitted and is a legitimate observation.

## On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

From the data exploration, we can suggest/recommend below suggestions to the business:

1. There is a strong positive correlation between ‘Grocery’ and ‘Deteregents\_Paper’. Also, we have moderately strong positive correlation between ‘Grocery’ and ‘Milk’ and ‘Milk’ and ‘Detergents\_Paper’. Increasing sales in any of these 3 product varieties could show corresponding increase in the sales of the other 2 product varieties.
2. The 3 product varieties on which buyers are spending maximum are Fresh, Grocery and Milk across regions, hence the company should ensure that maximum profit is being derived from these 3 items and hence all sales strategy needs to be pivoted on these 3 varieties.
3. From the data we have seen that product ‘Fresh’ has the highest consistency among the varieties, hence the business should look towards maximizing sales of this product variety.

# Problem 2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).



## Data Description

1. ID : nominal identifier for students in numerical format from 1 to 62.
2. Gender : type of gender(Female, Male)
3. Age : continuous from 18 to 26.
4. Class : type of class (Junior, Senior, Sophomore)
5. Major : type of major (Other, Management, CIS, Economics/Finance, Undecided, International Business, Retailing/Marketing, Accounting)
6. Grad Intention : type of intent(Yes, Undecided, No)
7. GPA : continuous from 2.3 to 3.9
8. Employment : type of employment(Full-Time, Part-Time, Unemployed)
9. Salary : continuous from 25 to 80
10. Social Networking : continuous from 1 to 4
11. Satisfaction : continuous from 1 to 6
12. Spending : continuous from 100 to 1400
13. Computer : type of computer (Laptop, Tablet, Desktop)
14. Text Messages : continuous from 0 to 900

## Sample of the Dataset:

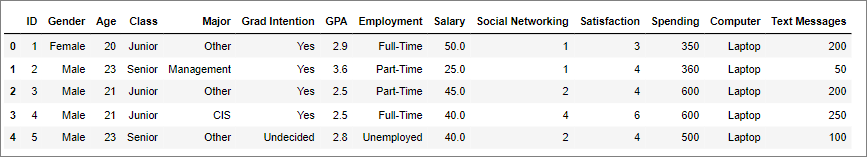


Table-2.1 Dataset Sample

Dataset has 14 columns with students from different class and across Majors.

## Exploratory Data Analysis:

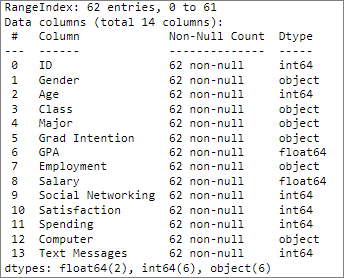


Table-2.2 Concise data summary

### Let us check the type of variables in the data frame.

There are a total of 62 rows and 14 columns in the dataset. Of the 14 columns, 6 are of object type, 2 are float type and rest 6 are integer type.

### Check for missing values in the dataset.

From Table-2.2 we can see that all the columns have 62 non-null values and hence we have no missing values in the dataset.

### Correlation Plot

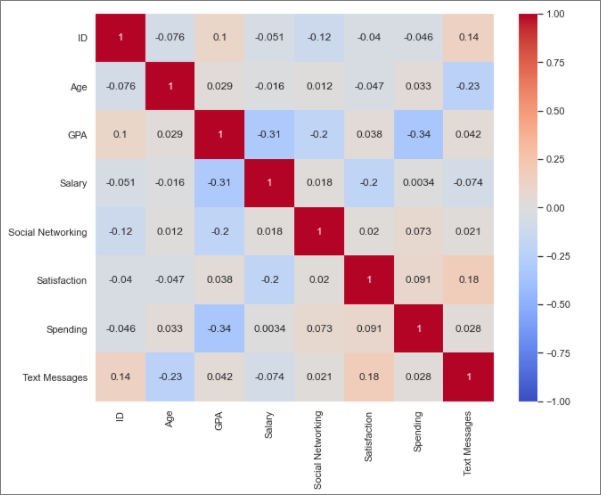


Figure-2.1 Correlation Heatmap

From the correlation plot, we can see that there is no strong correlation between the lists. A weak negative correlation exists between ‘GPA’ - ‘Spending’ and between ‘GPA’ – ‘Salary’.

Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

### Pairplot

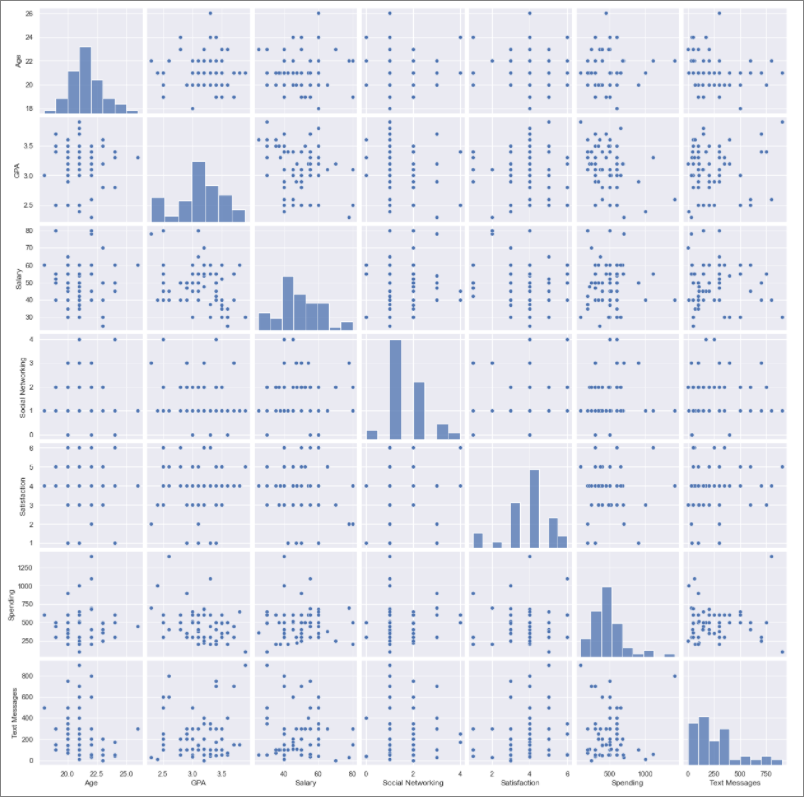


Figure-2.2 Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

From the graph, we cannot see any positive or negative linear relationship between variables. From the histograms we can see that Age, GPA, Salary, Spending and Text Messages seems to have a normal distribution. Spending and Text Messages display a right skew in their histograms.

## For this data, construct the following contingency tables (Keep Gender as row variable).

### Gender and Major

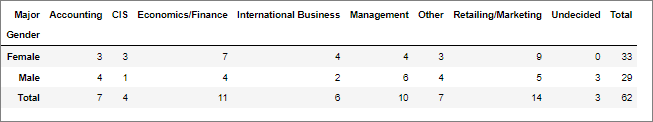


Table-2.3 Gender vs Major contingency table

The above table shows the contingency table between Gender and Major.

### Gender and Grad Intention

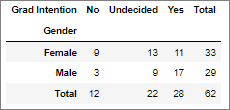


Table-2.4 Gender vs Grad Intention contingency table

The above table shows the contingency table between Gender and Grad Intention.

### Gender and Employment

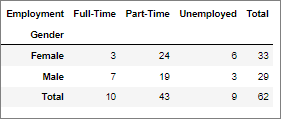


Table-2.5 Gender vs Employment contingency table

The above table shows the contingency table between Gender and Employment.

### Gender and Computer

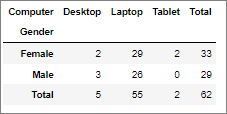


Table-2.6 Gender vs Computer contingency table

The above table shows the contingency table between Gender and Computer.

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions.

### What is the probability that a randomly selected CMSU student will be male?

Number of male students = 29

Total number of students = 62

Probability of Male Student = (Number of male students / Total number of students)

= 29/62

= 0.46774193548387094

= 46.77 %

Hence the probability that a randomly selected CMSU student will be male is 46.77 %.

* + 1. What is the probability that a randomly selected CMSU student will be female?

Number of female students = 29

Total number of students = 62

Probability of female Student = (Number of female students / Total number of students)

= 33/62

= 0.532258064516129

= 53.23%

Hence the probability that a randomly selected CMSU student will be female is 53.23 %.

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions.

### Find the conditional probability of different majors among the male students in CMSU?

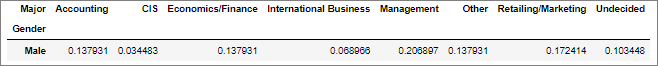


Table-2.7 Male Students vs Major contingency table

The above table shows the conditional probability of different majors among male students, the same is noted below in percentage.

Conditional probability of major in Accounting among male students is 13.79 %.

Conditional probability of major in CIS among male students is 3.45 %.

Conditional probability of major in Economics/Finance among male students is 13.79 %.

Conditional probability of major in International Business among male students is 6.90 %.

Conditional probability of major in Management among male students is 20.69 %.

Conditional probability of major in Other among male students is 13.79 %.

Conditional probability of major in Retailing/Marketing among male students is 17.24 %.

Conditional probability of male students who haven’t yet decided major (‘Undecided) is 10.34 %.

### Find the conditional probability of different majors among the female students in CMSU?



Table-2.8 Female students vs Major contingency table

The above table shows the conditional probability of different majors among female students, the same is noted below.

Conditional probability of major in Accounting among female students is 9.09 %

Conditional probability of major in CIS among female students is 9.09 %

Conditional probability of major in Economics/Finance among female students is 21.21 %

Conditional probability of major in International Business among female students is 12.12 %

Conditional probability of major in Management among female students is 12.12 %

Conditional probability of major in Other among female students is 9.09 %

Conditional probability of major in Retailing/Marketing among female students is 27.27 %

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions.

### Find the probability that a randomly chosen student is a male and intends to graduate?

The contingency table for ‘Gender’ and ‘Grad Intention’ is given below:

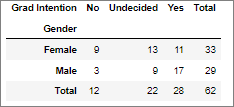


Table-2.9 Gender vs Grad Intention contingency table

From the above contingency table, we can see that number of male students with ‘Grad Intention’ as ‘Yes’ is 17.

Probability of male & intends to graduate = No. of male students with intent to graduate / Total students

= 17 / 62

= 0.27419354838709675

= 27.42 %

The probability that a randomly chosen student is a male & intends to graduate is 27.42 %.

### Find the probability that a randomly chosen student is a female and does not have a laptop?

The contingency table for ‘Gender’ and ‘Computer’ is given below:

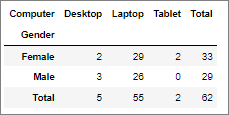


Table-2.10 Gender vs Computer contingency table

From the above contingency table, we can see that number of female students without Laptop as (Female students with ‘Desktop’ + Female students with ‘Tablet’) 4.

Probability of female & does not have a laptop = No. of Female students without laptop / Total Students

= 4 / 62

= 0.06451612903225806

= 6.45 %

The probability that a randomly chosen student is a female & does not have a laptop is 6.45 %.

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following questions.

### Find the probability that a randomly chosen student is a male or has full time employment?

The contingency table for ‘Gender’ and ‘Employment’ is given below:

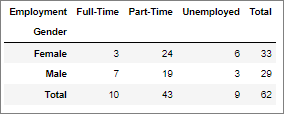


Table-2.11 Gender vs Employment contingency table

We know the probability addition rule, P(A or B) = P(A) + P(B) – P(A and B). Let FTE denote full time employment.

Probability of male = 29 / 62

Probability of FTE = 10 / 62

Probability of male student and FTE = 7/62

Probability of male or FTE = Probability of male +Probability of FTE–Probability of male & FTE

= ( 29 / 62 ) + ( 10 / 62 ) – ( 7 / 62 )

= 0.5161290322580645

= 51.61 %

The probability that a randomly chosen student is a male or has full time employment is 51.61 %.

### Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

The contingency table for ‘Gender’ and ‘Major’ is given below:

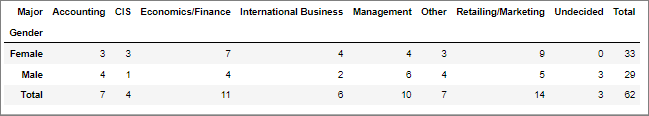


Table-2.12 Gender vs Major contingency table

P(x|y) denotes conditional probability of x given y

Total number of female students = 33

Number of female students majoring in ‘International Business’ = 4

Number of female students majoring in ‘Management’ = 4

P(International Business | Female) = 4 / 33

P(Management | Female) = 4 / 33

P(International Business or Management | Female) = P(International Business | Female) +

P(Management | Female)

= ( 4 / 33 ) + ( 4 / 33 )

= 0.24242424242424243 = 24.24 %

The conditional probability that given a female student is randomly chosen, she is majoring in International Business or Management is 24.24 %.

## Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now, and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

The contingency table Gender and Grad Intention is given below:

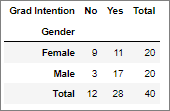


Table-2.13 Gender vs Grad Intention contingency table

As can be seen, the ‘Undecided’ Grad Intention students have not been considered.

For independent events A and B, we know that P(A and B) = P(A) \* P(B).

From the above table,

Probability of female students = 20 / 40

Probability of students with graduate intention = 28 / 40

Probability of female students and having graduate intention = 11/ 40

Probability of female students \* Probability of students with graduate intention = ( 20 / 40 ) \* ( 28 / 40 )

= 0.35 = 35 %

Probability of female students and having graduate intention = 11 / 40

= 0.275 = 27.5 %

Since probability of female with graduate intention (27.5 %) is not equal to product of probabilities of female and graduate intention (35 %), the events are not independent events.

## Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data.

### If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

Number of students with GPA < 3 = 17

Total Number of students = 62

Probability of students with GPA < 3 = Number of students with GPA < 3 / Total Number of students

= 17 / 62

= 0.27419354838709675 = 27.42 %

The probability that a randomly chosen student has GPA less than 3 is 27.42 %.

### Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

P(x|y) denotes conditional probability of x given y

Number of male students = 29

Number of male students earning 50 or more = 14

Number of female students = 33

Number of female students earning 50 or more = 18

P(Earns > 50 | male) = Number of male students earning 50 or more / Number of male students

= 14 / 29

= 0.4827586206896552 = 48.28 %

P(Earns > 50 | female) = Number of female students earning 50 or more / Number of female students

= 18 / 33

= 0.5454545454545454 = 54.55 %

The conditional probability that a randomly selected male earns 50 or more is 48.28%.

The conditional probability that a randomly selected female earns 50 or more is 54.55 %.

## Question Header

### Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution.

GPA

The histogram for GPA is given below:

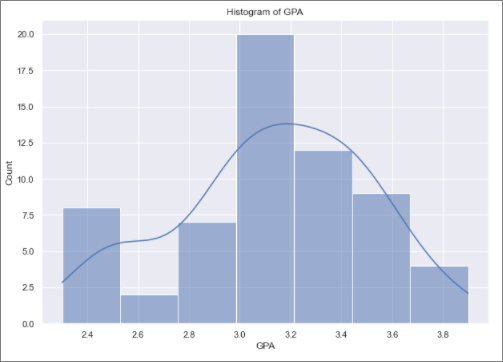


Figure-2.3 Histogram of GPA

The histogram seems to be Gaussian in nature. Since the observations are only 62 and this is a continuous variable let’s use Shapiro-Wilk test to check if GPA has a normal distribution.

Shapiro Result (statistic=0.9685361981391907, pvalue=0.11204058676958084)

Since pvalue is greater than 0.05, we don’t have enough evidence to reject the null hypothesis for

Shapiro-Wilk test, which shows that GPA does follow a normal distribution.

Salary

The histogram for Salary is given below:

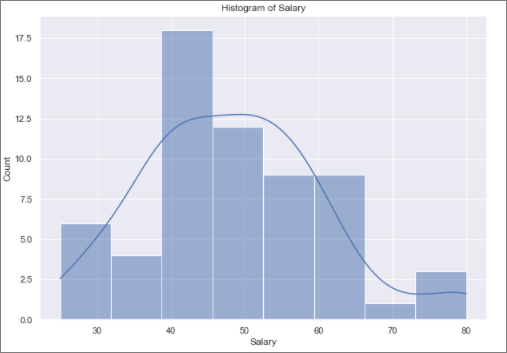


Figure-2.4 Histogram of Salary

The histogram seems to be Gaussian in nature. Since the observations are only 62 and this is a continuous variable let’s use Shapiro-Wilk test to check if Salary has a normal distribution.

Shapiro Result (statistic=0.9565856456756592, pvalue=0.028000956401228905)

Since pvalue is lesser than 0.05, we have enough evidence to reject the null hypothesis for

Shapiro-Wilk test, which shows that Salary does not follow a normal distribution.

Spending

The histogram for Spending is given below:

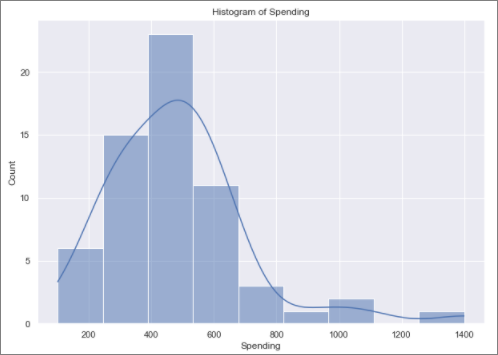


Figure-2.5 Histogram of Spending

The histogram seems to be Gaussian in nature, though we have missing values. Since the observations are only 62 and this is a continuous variable let’s use Shapiro-Wilk test to check if Salary has a normal distribution.

Shapiro Result (statistic=0.8777452111244202, pvalue=1.6854661225806922e-05)

Since pvalue is lesser than 0.05, we have enough evidence to reject the null hypothesis for

Shapiro-Wilk test, which shows that Spending does not follow a normal distribution.

Text Messages

The histogram for Spending is given below (Figure-31):

The histogram seems to be Gaussian in nature. Since the observations are only 62 and this is a continuous variable let’s use Shapiro-Wilk test to check if Salary has a normal distribution.

Shapiro Result (statistic=0.8594191074371338, pvalue=4.324040673964191e-06)

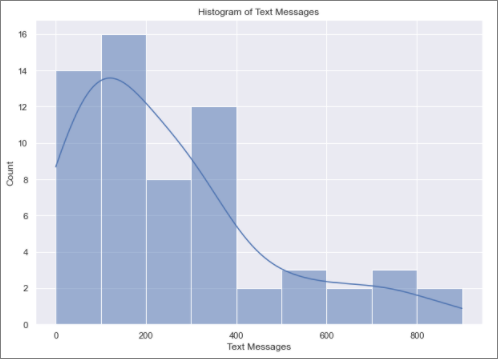


Figure-2.6 Histogram of Text Messages

Since pvalue is lesser than 0.05, we have enough evidence to reject the null hypothesis for

Shapiro-Wilk test, which shows that Spending does not follow a normal distribution.

Conclusion

In conclusion, of the four numerical (continuous) variables in the data set - GPA, Salary, Spending and Text Messages, only GPA follows a normal distribution and the other 3 variables do not follow a normal distribution as per Shapiro-Wilk test.

### Write a note summarizing your conclusions.

From the data exploration, we can summarize the following:

1. There are no strong positive or negative correlation between variables in the dataset. There is a moderate negative correlation between ‘GPA’ and ‘Salary’ and between ‘GPA’ and ‘Spending’. CMSU could ponder on providing additional course support to working students.
2. There are more female undergraduate students than male, which indicates that CMSU can run campaigns at high school levels to attract more male students, unless the demographics of the area where CMSU is situated in, is skewed towards more female teenage population than male.
3. The probability of female students’ graduate intention is much lesser than male students. CMSU can look for ways to provide more encouragement for female undergraduate students to graduate.
4. ‘Retailing/Marketing’ among female students and ‘Management’ among male students are the popular Major choice. CMSU should ensure to have more support staff for these subjects.
5. Around 15% of students are not employed, unless it is a matter of choice, CMSU can look out for ways to provide part time/full time employment for students to assist them financially.

# Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.¶

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.



## Data Description

1. A : mean moisture content of A type shingles, continuous from 0.13 to 0.72
2. B : mean moisture content of B type shingles, continuous from 0.1 to 0.58

## Sample of the Dataset:

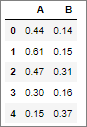


Table-3.1 Dataset Sample

Dataset has 2 columns which denotes the mean moisture content of A and B type shingles.

## Exploratory Data Analysis:

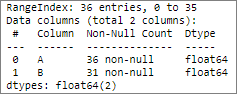


Table-3.2 Concise data summary

### Let us check the type of variables in the data frame.

There are a total of 36 rows and 2 columns in the dataset. Both the columns are float type.

### Check for missing values in the dataset.

From Table-24 we can see that column A has 36 non-null values, but column B has 31 non-null columns. Hence, we have 5 null observations for column B.

### Descriptive Statistics

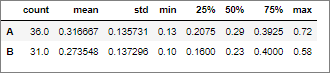


Table-3.3 Summary of the data

The above table displays the mean, standard deviation, Quartile information of both columns.

## Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

‘A’ type shingle

Let us first state the hypothesis:

Ho (null hypothesis) 🡪 mean moisture of ‘A’ = 35 🡪 µA = 0.35 (permissible limit)

Ha (alternate hypothesis) 🡪 mean moisture of ‘A’ > 35 🡪 µA > 0.35

Since α, significance of the test is not provided, let’s assume it as 0.05. The level of significance 0.05 is related to the 95% confidence level.

One sample T Test on the ‘A’ shingle data, with population mean as 0.35, provides the following result:

Statistic value = - 1.4735046253382782

pvalue = 0.07477633144907513

Since pvalue > α , we do not have enough evidence to reject the null hypothesis. The level of significance 0.05 is related to the 95% confidence level.

Hence the mean moisture of ‘A’ is within the permissible limits.

‘B’ type shingle

Let us first state the hypothesis:

Ho (null hypothesis) 🡪 mean moisture of ‘B’ = 35 🡪 µB = 0.35 (permissible limit)

Ha (alternate hypothesis) 🡪 mean moisture of ‘B’ > 35 🡪 µB > 0.35

Since α, significance of the test is not provided, let’s assume it as 0.05. The level of significance 0.05 is related to the 95% confidence level.

One sample T Test on the ‘B’ shingle data, with population mean as 0.35, provides the following result:

Statistic value = - 3.1003313069986995

pvalue = 0.0020904774003191826

Since pvalue < α , we have enough evidence to reject the null hypothesis.

Hence the mean moisture of ‘B’ is NOT within the permissible limits.

## Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Let us first state the hypothesis:

Ho (null hypothesis) 🡪 mean moisture of ‘A’ = ‘B’ 🡪 µA = µB

Ha (alternate hypothesis) 🡪 mean moisture of ‘A’ <> ‘B’ 🡪 µA <> µB

Since α, significance of the test is not provided, let’s assume it as 0.05. The level of significance 0.05 is related to the 95% confidence level.

On performing the Independent Sample T Test on ‘A’ and ‘B’ data, we get the following result:

Statistic value = 1.2896282719661123

pvalue = 0.2017496571835306

Since pvalue > α , we do not have enough evidence to reject the null hypothesis.

Hence the population mean for shingles ‘A’ and ‘B’ are equal.

The assumptions for test of equality of means are as follows:

1. Observations are independent.
2. Data for the samples is obtained via a random sample of the population.
3. The data is from a population which follows normal distribution.
4. The variances of the population are equal.

# THE END!