SMDM PROJECT REPORT

KIRAN.N

GREAT LEARNING

Contents

List Of Figures

List of Tables

Problem 1

Executive Summary

A wholesale distributor operates with several large retailers across different regions of Portugal. The dataset consists of annual spending information of several large retailers on several varieties of products across different regions and different channels. Based on different channel and region spends on a product varies. In this problem statement we will explore the spends made by large retailers on different varieties of products at different regions and different channels.

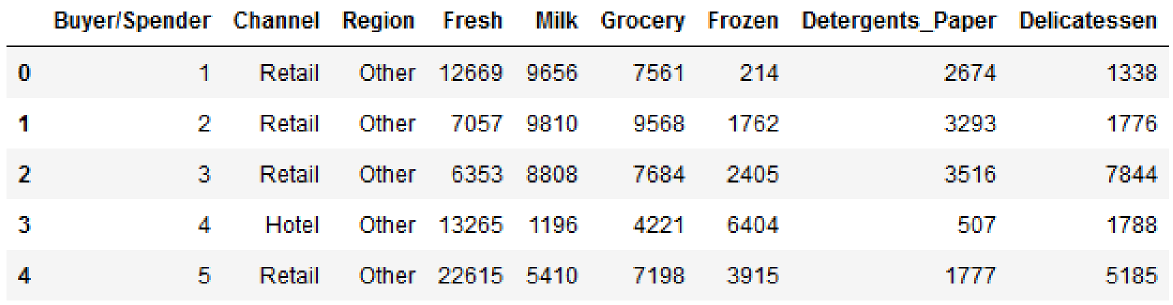
Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 440 Large retailers spending information on 6 different varieties of products in 3 different regions and across different sales channel. This report helps us in exploring the summary statistics.

Data Description

1. Buyer/Spender: Serial number / ID for retailer.
2. Channel: Retail / Hotel
3. Region: Lisbon, Oporto, Other
4. Fresh: Annual amount spent on Fresh products.
5. Milk: Annual amount spent on Milk products.
6. Grocery: Annual amount spent on Grocery products.
7. Frozen: Annual amount spent on Frozen products.
8. Detergents_Paper: Annual amount spent on Detergents Paper products.
9. Delicatessen: Annual amount spent on Delicatessen products.

Sample of the dataset



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

Table 1 - Dataset Sample

Dataset has 9 variables with 6 different varieties of products in 3 different regions and across 2 different sales channels. Annual spends on a specific product varies across different regions and across different sales channel.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
Buyer/Spender     int64
Channel           object
Region            object
Fresh             int64
Milk              int64
Grocery           int64
Frozen            int64
Detergents_Paper  int64
Delicatessen      int64
```

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

Check for missing values in the dataset

```
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
Buyer/Spender     440 non-null    int64
Channel           440 non-null    object
Region            440 non-null    object
Fresh             440 non-null    int64
Milk              440 non-null    int64
Grocery           440 non-null    int64
Frozen            440 non-null    int64
Detergents_Paper  440 non-null    int64
Delicatessen      440 non-null    int64
```

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot

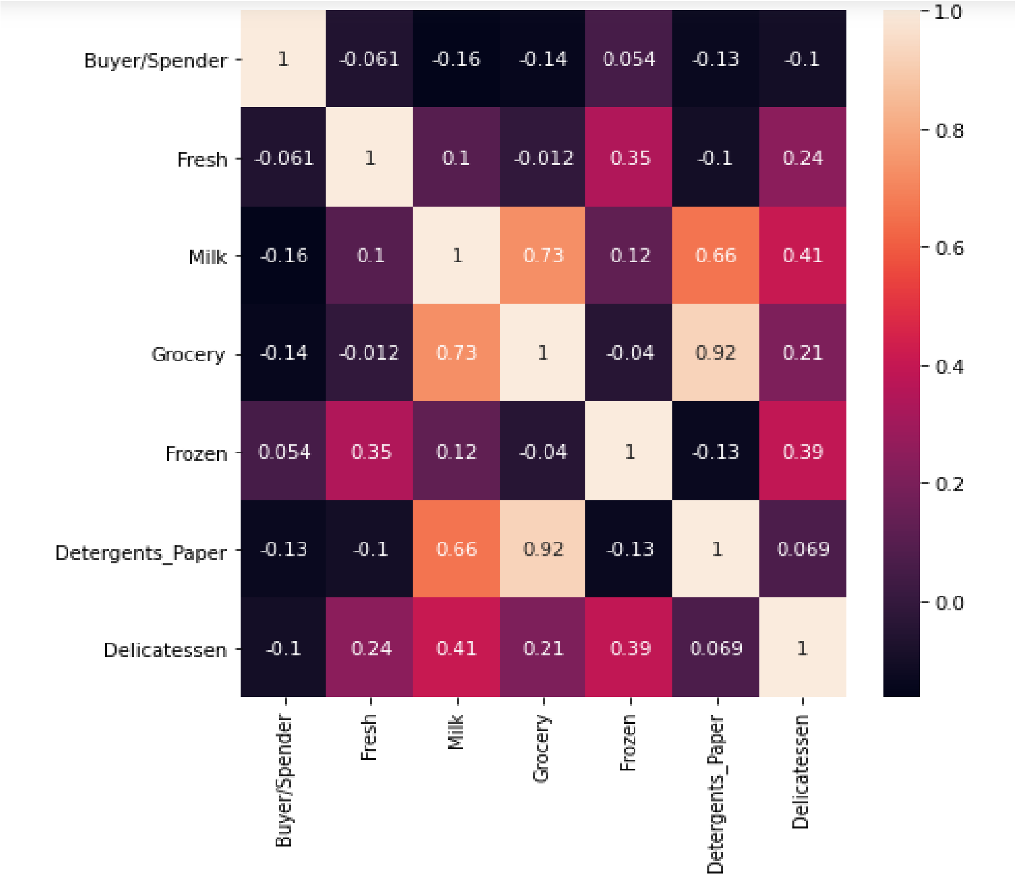
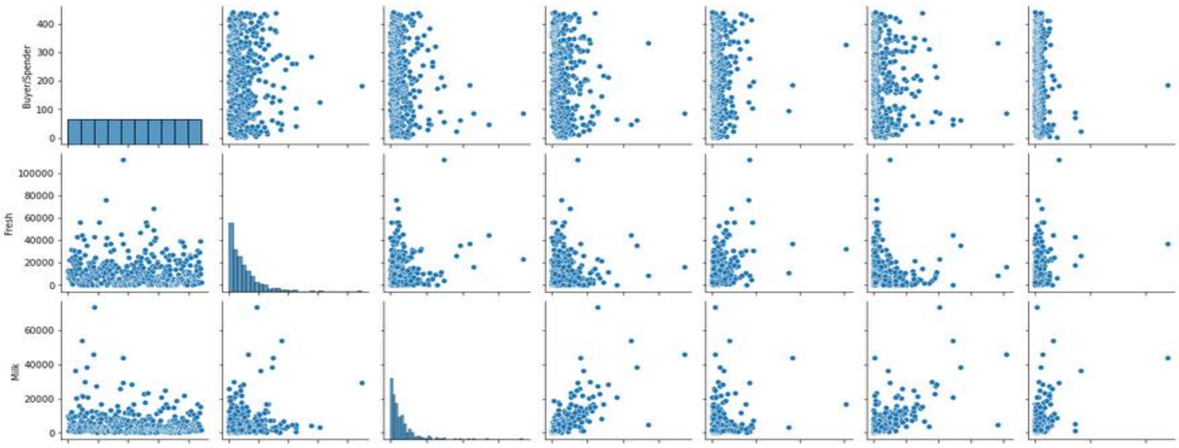


Figure 1 - Correlation Heatmap

From the correlation plot, we can see the correlation among different variables. 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

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



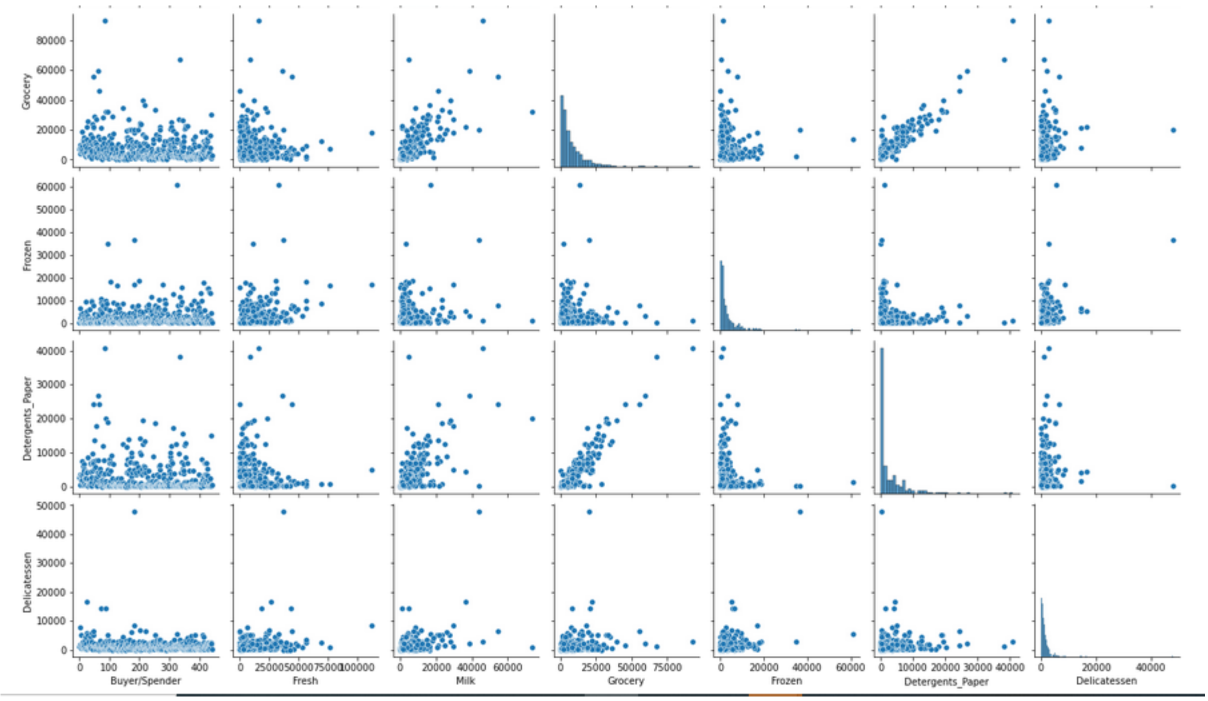
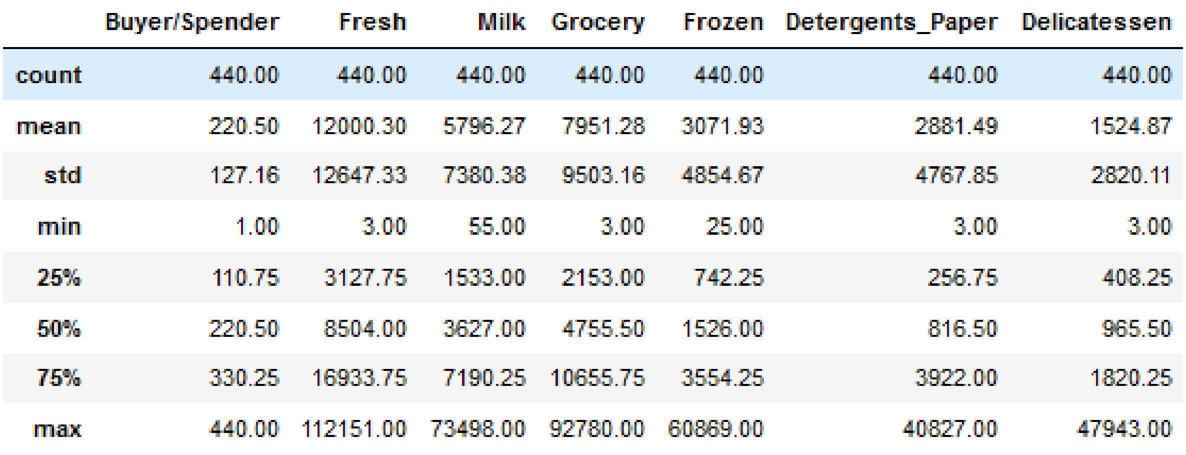


Figure 2 - Pairplot

Q 1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Descriptive statistics helps to describe and understand the features of a specific dataset by giving short summaries about the sample and various measures of the data. The different components of descriptive statistics are:

1) Measures of Central Tendency: mean, median, and mode.
2) Measures of Dispersion: Range, Inter Quartile Range, Standard Deviation

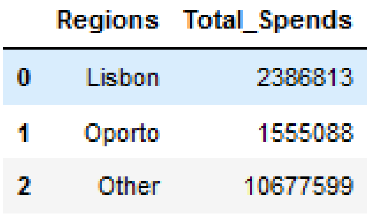


	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.00	440.00	440.00	440.00	440.00	440.00	440.00
mean	220.50	12000.30	5796.27	7951.28	3071.93	2881.49	1524.87
std	127.16	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
min	1.00	3.00	55.00	3.00	25.00	3.00	3.00
25%	110.75	3127.75	1533.00	2153.00	742.25	256.75	408.25
50%	220.50	8504.00	3627.00	4755.50	1526.00	816.50	965.50
75%	330.25	16933.75	7190.25	10655.75	3554.25	3922.00	1820.25
max	440.00	112151.00	73498.00	92780.00	60869.00	40827.00	47943.00

Table 2 - Summary Of Data

From the descriptive statistics, we see that the average annual spends on Fresh products is 12000.30 euros, Milk products is 5796.27 euros, Grocery products is 7951.28 euros, Frozen products is 3071.93 euros, Detergent's paper is 2881.49 euros and Delicatessen products is 1524.87 euros.

Calculating the total spends Region wise.



	Regions	Total_Spends
0	Lisbon	2386813
1	Oporto	1555088
2	Other	10677599

Table 3 - Total Spends Region wise

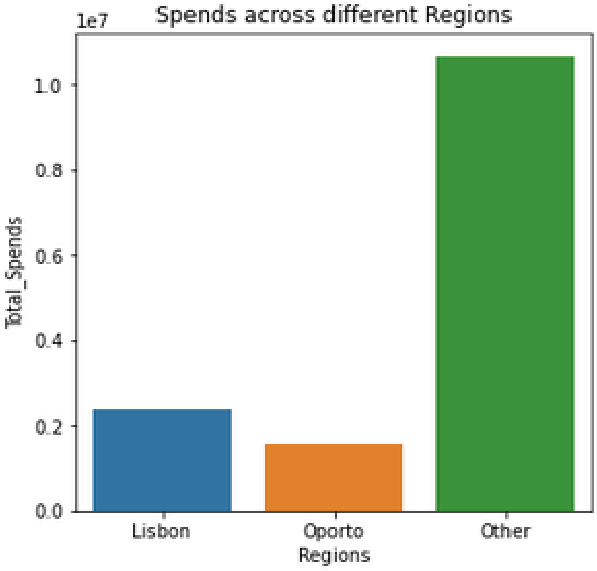
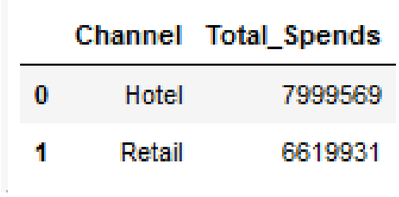


Figure 3 - Regions vs Total Annual Spends

After calculating the total spends region wise, we found that Other region has spent more and Oporto region has spent least.

Calculating the total spends Channel wise.



	Channel	Total_Spends
0	Hotel	7999569
1	Retail	6619931

Table 4 - Total Spends Channel wise

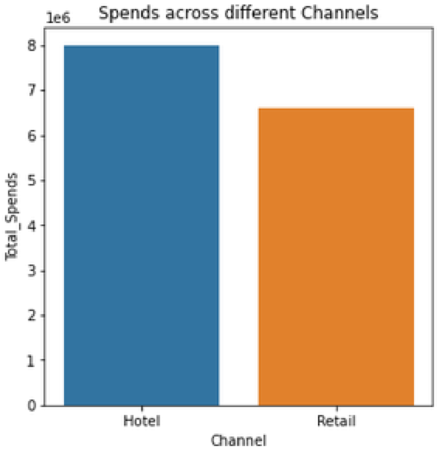
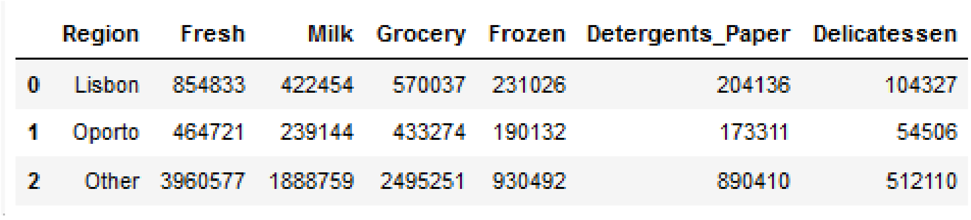


Figure 4 - Channel vs Total Annual Spends

After calculating the total spends region wise, we found that Hotel channel has spent more and Retail channel has spent least.

Q 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.

Calculating the Spends on each Variety of Item Region wise.



	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Lisbon	854833	422454	570037	231026	204136	104327
1	Oporto	464721	239144	433274	190132	173311	54506
2	Other	3960577	1888759	2495251	930492	890410	512110

Table 5 - Spends on Different Variety of Items Region wise

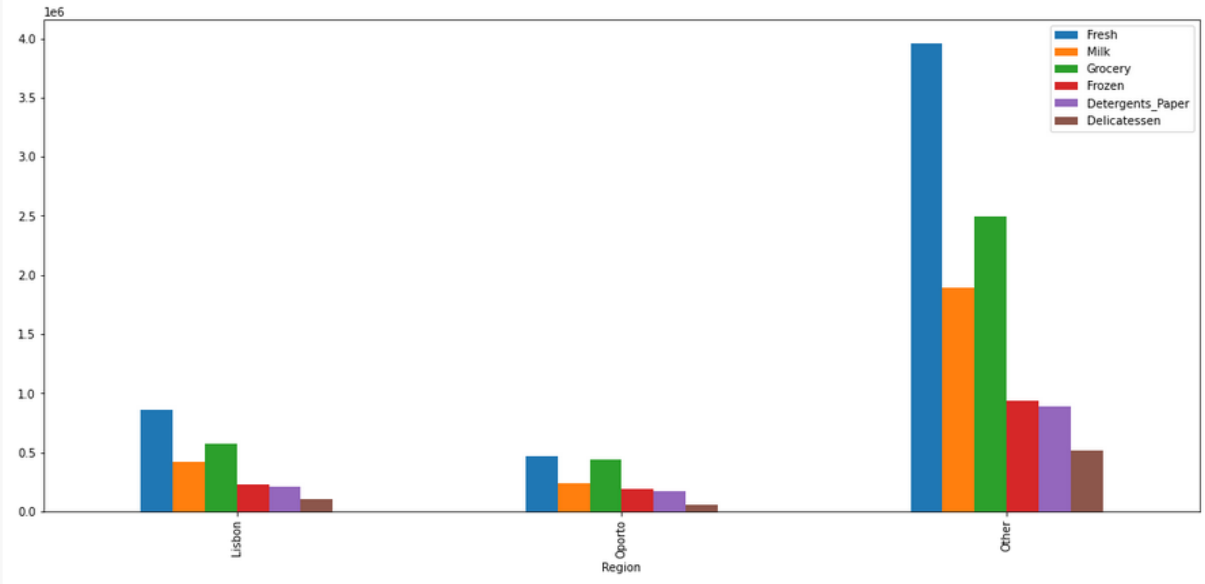
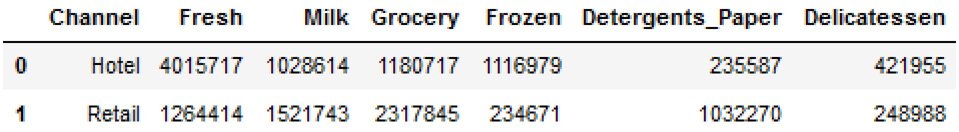


Figure 5 - Spends vs Different Variety of Items Region wise

The spends on the different varieties of Items varies from region to region. Overall spends are more in Other region when compared to other two regions. Considering the different varieties of Items, highest spends are made on Fresh products, lowest spends are made on Delicatessen products in all regions. Spending on Grocery products stands second highest followed by Milk products, Frozen products and then comes Detergent Paper products in all regions.

Calculating the Spends on each Variety of Item Channel wise.



	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	Hotel	4015717	1028614	1180717	1116979	235587	421955
1	Retail	1264414	1521743	2317845	234671	1032270	248988

Table 6 - Spends on Different Variety of Items Channel wise

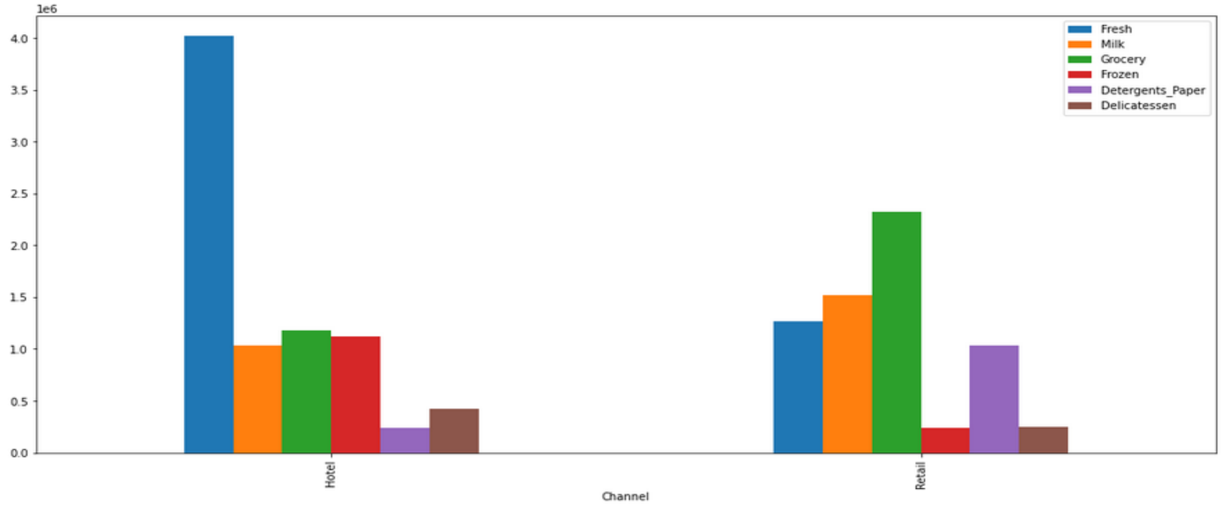


Figure 6 - Spends vs Different Variety of Items Channel wise

The spends on the different varieties of Items varies from channel to channel. Overall spends are more in Hotel channel when compared Retail channel.

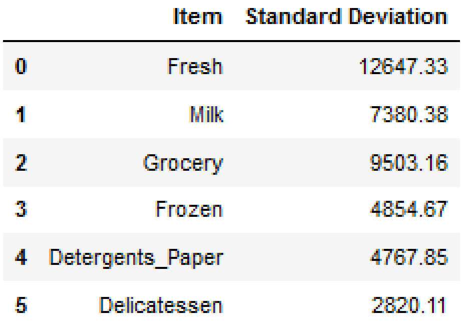
In Hotel channel highest spends are made on Fresh products and lowest spends on Detergent Paper products. Spending on Grocery products stands second highest followed by Frozen products, Milk products and then comes Delicatessen products.

In Retail channel highest spends are made on Grocery products and lowest spends on Frozen products. Spending on Milk products stands second highest followed by Fresh products, Detergent Paper products and then comes Delicatessen products.

Q 1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

Standards Deviation which is part of measures of dispersion talks about the inconsistent behaviour of data.

Calculating the Standard Deviation of Annual Spends of each Variety of Item.



	Item	Standard Deviation
0	Fresh	12647.33
1	Milk	7380.38
2	Grocery	9503.16
3	Frozen	4854.67
4	Detergents_Paper	4767.85
5	Delicatessen	2820.11

Table 7 - Standard Deviation of Spends of each Variety of Item

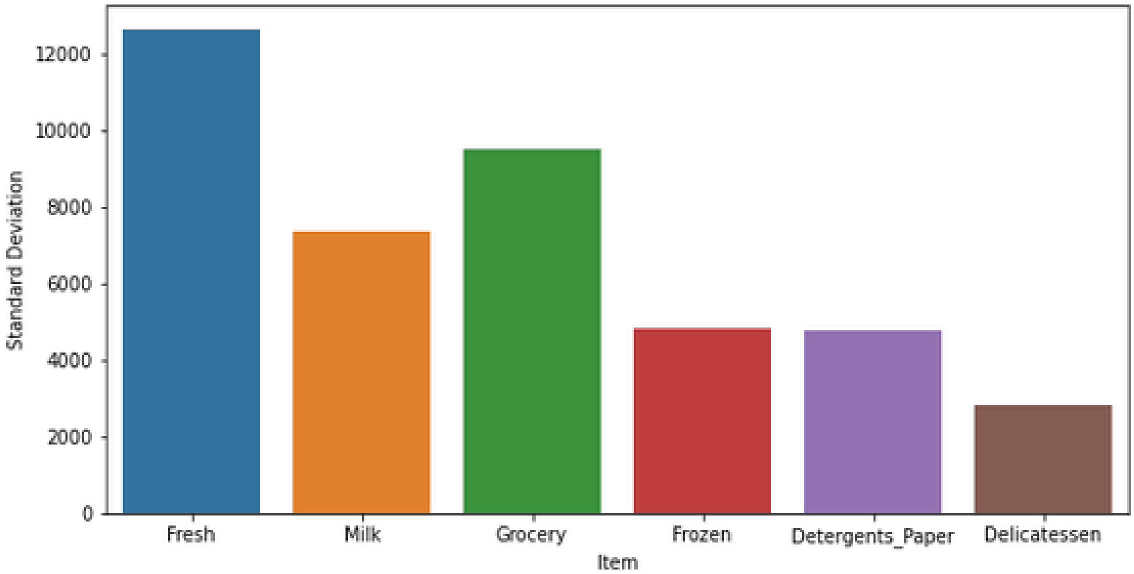


Figure 7 - Item vs Std Deviation of Annual Spends

From the above data, Fresh products has highest standard deviation of spends indicating it as the most inconsistent behaviour and Delicatessen products has lowest standard deviation of spends indicating it as the least inconsistent behaviour.

Q 1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

An outlier is a data point that differs significantly from other observations. Boxplots are very useful in indicating the presence of outliers in data.

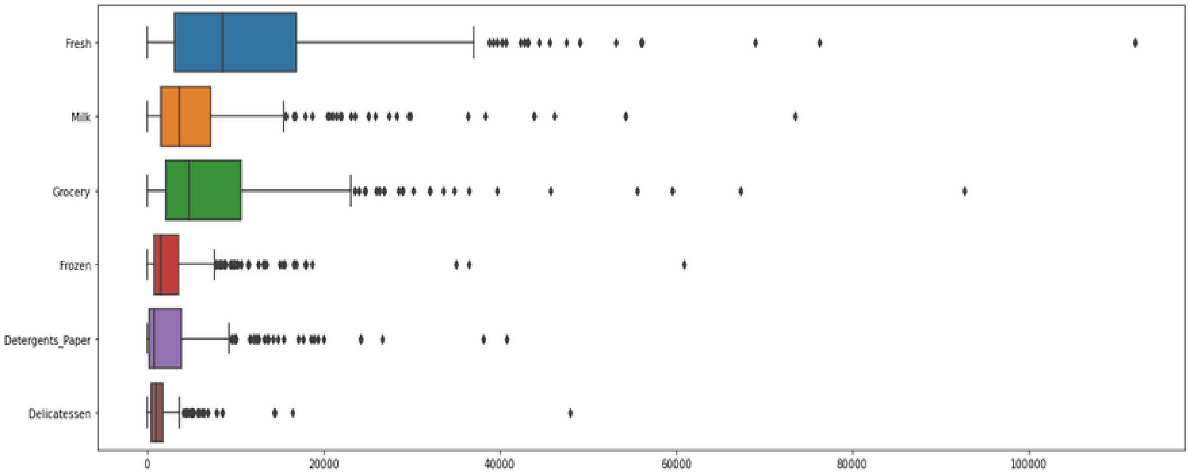


Figure 8 - Boxplot of Annual Spends on different Items.

From the above boxplot it is evident that outliers are present in all 6 items annual spends.

Q 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.

- The spends on different variety of items is almost rightly skewed, it is not evenly distributed.
- Spends are concentrated more on Hotel channel, so there is scope for increasing business on Retail channel.
- Spends are concentrated more on Other region, so there is scope for increasing business in remaining regions.
- Spends on Delicatessen products is less compared to other variety of items, so there is possible scope of business improvement.

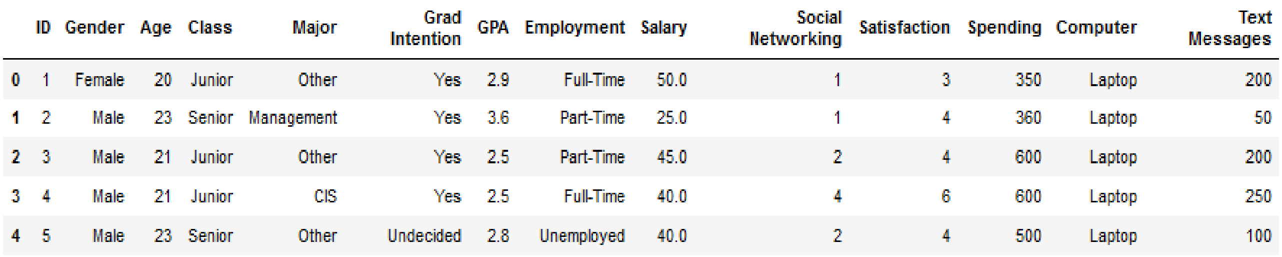
Problem 2

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: ID of a student
2) Gender: Male/Female
3) Age: Age of a student
4) Class: Class to which a student belongs
5) Major: Major area of their studies.
6) Grad Intention: Have intention to undergo Graduation or not
7) GPA: GPA of a student
8) Employment: Employment status of a student
9) Salary: Salary being drawn currently
10) Social Networking: Number of Social Networking Sites students are active
11) Satisfaction: Student's Satisfaction level on a scale of 1 – 5
12) Spending: Spends made by the student
13) Computer: Laptop/Desktop/Tablet
14) Text Messages: Number of text messages used by Student

Sample of the dataset



	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

Table 8 - Dataset Sample

Dataset has 14 columns with 62 rows. Each row in the dataset corresponds to one student. Each column corresponds to different information about the student.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
ID                      int64
Gender                 object
Age                     int64
Class                  object
Major                  object
Grad Intention         object
GPA                   float64
Employment             object
Salary                float64
Social Networking       int64
Satisfaction            int64
Spending                int64
Computer               object
Text Messages           int64
```

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

Check for missing values in the dataset

```
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
ID                   62 non-null     int64
Gender               62 non-null     object
Age                  62 non-null     int64
Class                62 non-null     object
Major                62 non-null     object
Grad Intention       62 non-null     object
GPA                  62 non-null     float64
Employment           62 non-null     object
Salary               62 non-null     float64
Social Networking    62 non-null     int64
Satisfaction         62 non-null     int64
Spending             62 non-null     int64
Computer             62 non-null     object
Text Messages        62 non-null     int64
```

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot

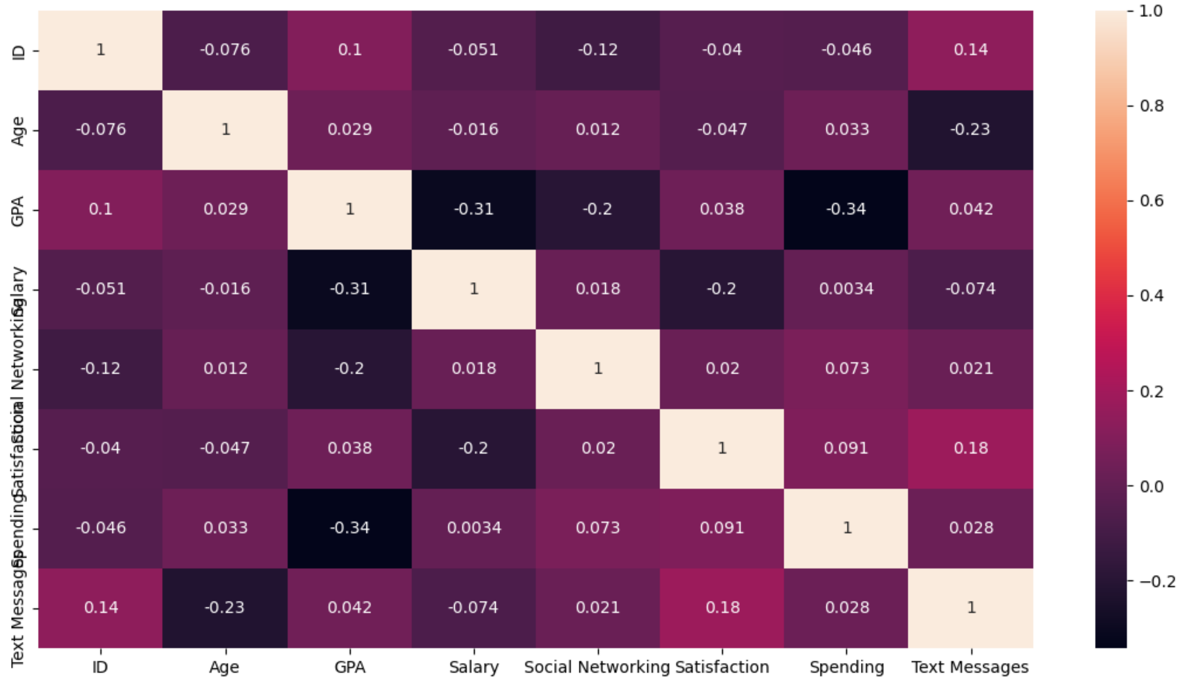
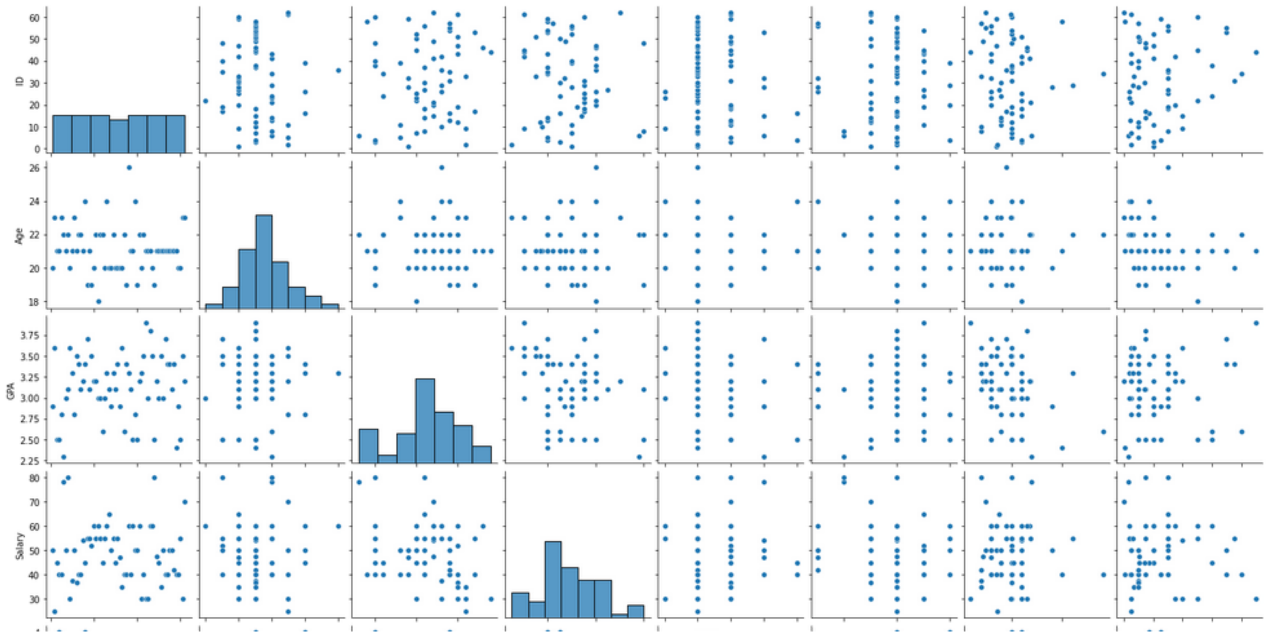


Figure 9 - Correlation Heatmap

From the correlation plot, we can see the correlation among different variables. 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

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



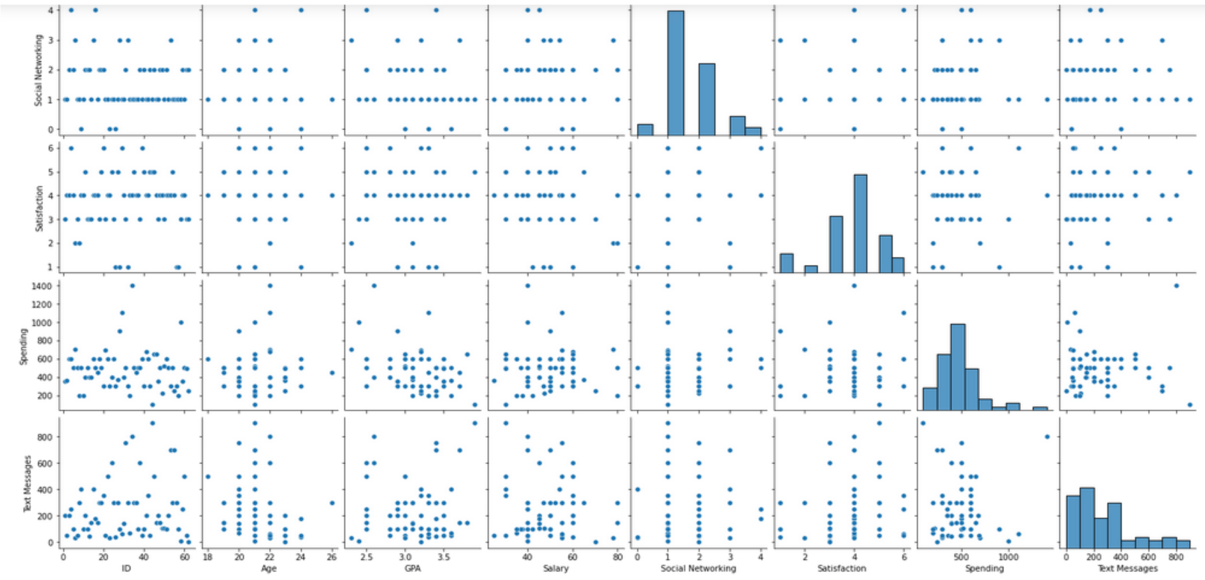
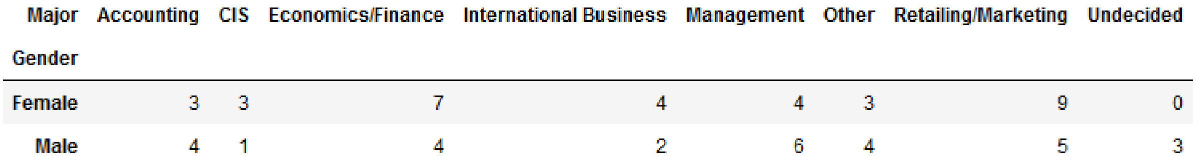


Figure 10 - Pairplot

Q 2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

Q 2.1.1. Gender and Major



Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Table 9 - Cross Tabulation of Gender and Major

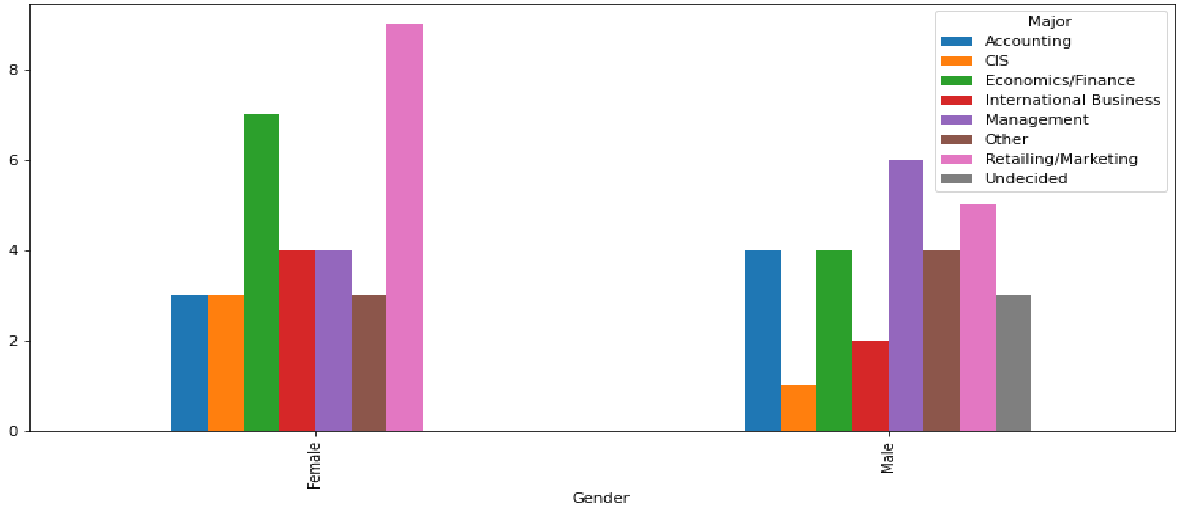
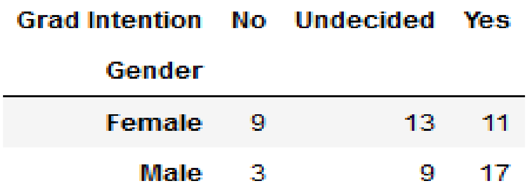


Figure 11 - Crosstabulation Plot of Gender and Major

Q 2.1.2. Gender and Grad Intention



Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Table 10 - Cross Tabulation of Gender and Grad Intention

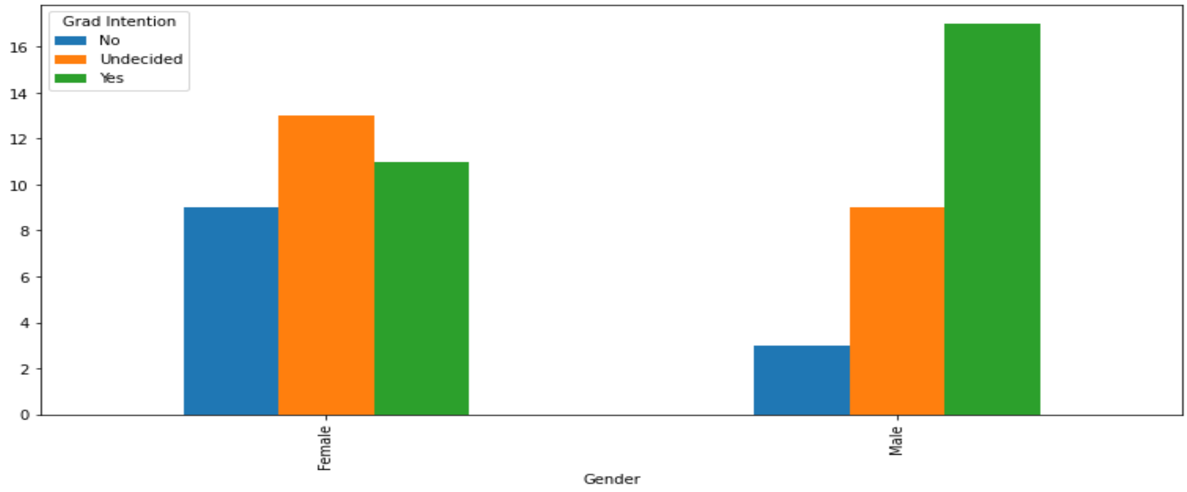
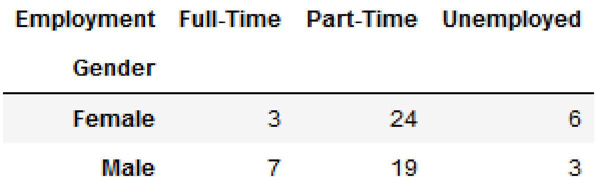


Figure 12 - Crosstabulation Plot of Gender and Grad Intention

Q 2.1.3. Gender and Employment



Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

Table 11 - Cross Tabulation of Gender and Employment

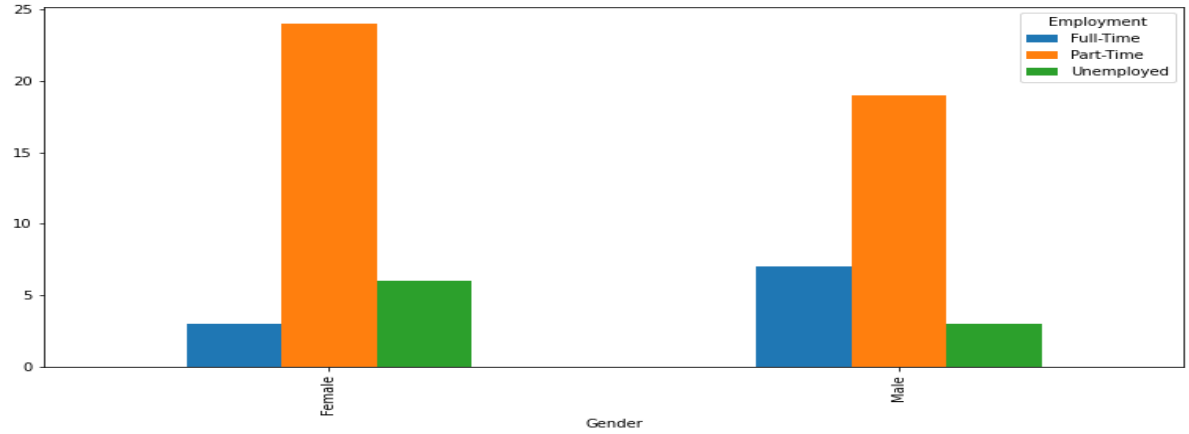
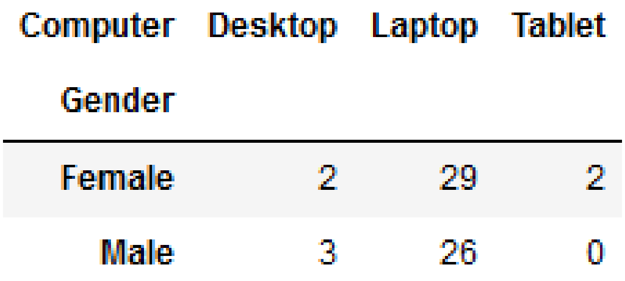


Figure 13 - Crosstabulation Plot of Gender and Employment

Q 2.1.4. Gender and Computer



Computer	Desktop	Laptop	Tablet
Gender			
Female	2	29	2
Male	3	26	0

Table 12 - Cross Tabulation of Gender and Computer

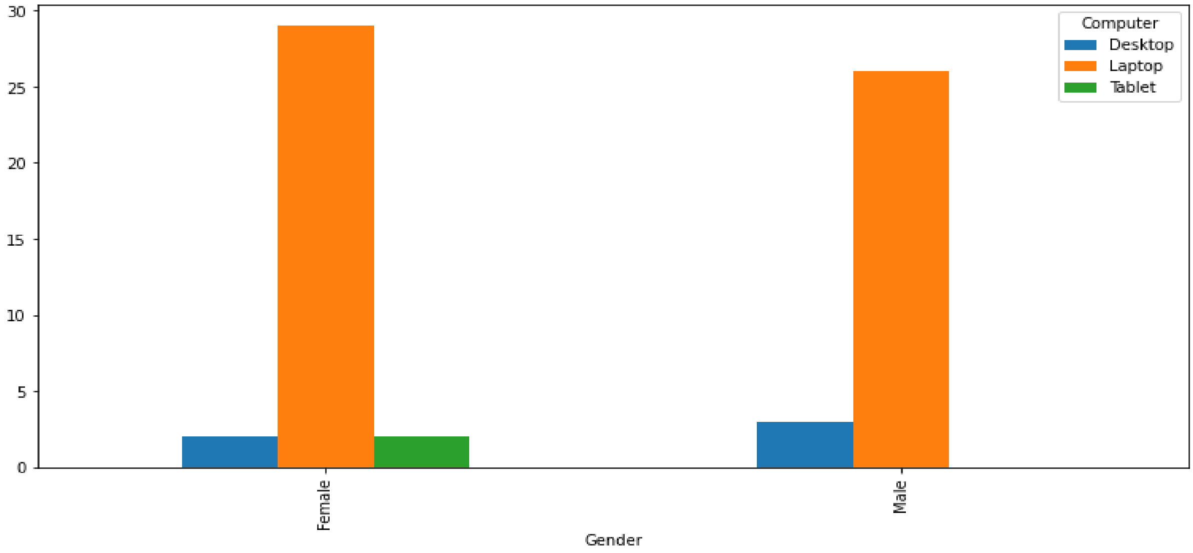


Figure 14 - Crosstabulation Plot of Gender and Computer

Q 2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

Q 2.2.1. What is the probability that a randomly selected CMSU student will be male?
Probability that a randomly selected CMSU student will be male is: 0.47

Q 2.2.2. What is the probability that a randomly selected CMSU student will be female?
Probability that a randomly selected CMSU student will be female is: 0.53

Q 2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

Q 2.3.1. Find the conditional probability of different majors among the male students in CMSU.

1) Conditional Probability that a randomly selected student belonging to Retailing/Marketing will be a male is: 0.17
2) Conditional Probability that a randomly selected student belonging to Economics/Finance will be a male is: 0.14
3) Conditional Probability that a randomly selected student belonging to Management will be a male is: 0.21
4) Conditional Probability that a randomly selected student belonging to Accounting will be a male is: 0.14
5) Conditional Probability that a randomly selected student belonging to Other will be a male is: 0.14
6) Conditional Probability that a randomly selected student belonging to International Business will be a male is: 0.07
7) Conditional Probability that a randomly selected student belonging to CIS will be a male is: 0.03
8) Conditional Probability that a randomly selected student belonging to Undecided will be a male is: 0.1

Q 2.3.2 Find the conditional probability of different majors among the female students of CMSU.

1) Conditional Probability that a randomly selected student belonging to Retailing/Marketing will be a female is: 0.27
2) Conditional Probability that a randomly selected student belonging to Economics/Finance will be a female is: 0.21
3) Conditional Probability that a randomly selected student belonging to Management will be a female is: 0.12
4) Conditional Probability that a randomly selected student belonging to Accounting will be a female is: 0.09
5) Conditional Probability that a randomly selected student belonging to Other will be a female is: 0.09
6) Conditional Probability that a randomly selected student belonging to International Business will be a female is: 0.12
7) Conditional Probability that a randomly selected student belonging to CIS will be a female is: 0.09
8) Conditional Probability that a randomly selected student belonging to Undecided will be a female is: 0.0

Q 2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

Q 2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

Probability that a randomly chosen student is a male and intends to graduate is: 0.27

Q 2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

Probability that a randomly selected student is a female and does NOT have a laptop is: 0.06

Q 2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

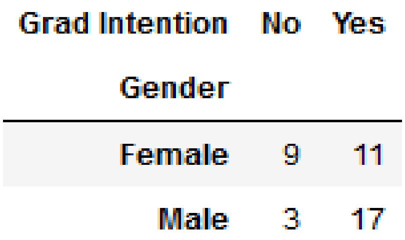
Q 2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?

Probability that a randomly chosen student is a male or has full-time employment: 0.52

Q 2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

Conditional probability that given a female student is randomly chosen, she is majoring in international business or management is: 0.24

Q 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?



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table 13 - Cross Tabulation of Gender and Grad Intention (Yes/No)

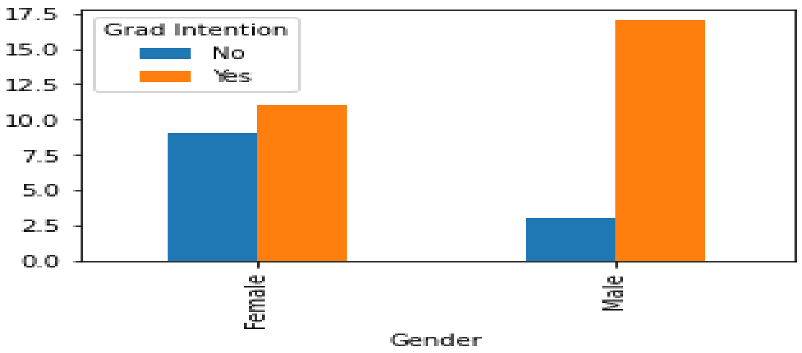


Figure 15 - Crosstabulation Plot of Gender and Grad Intention (Yes/No)

Probability of Graduate Intention as Yes is: 0.45

Probability of Graduate Intention as Yes given Student is Female is: 0.33

Since above two Probabilities are not equal Graduate Intention as Yes and being Female are not independent events.

Q 2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Q 2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

Probability of randomly selected student is having GPA < 3: 0.27

Q 2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

Conditional probability that a randomly selected male earns 50 or more is: 0.48
Conditional probability that a randomly selected female earns 50 or more is: 0.55

Q 2.8. 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. Write a note summarizing your conclusions.

- The mean, median & mode of 'GPA' variable is 3.13, 3.15 and 3.0 respectively.
- The mean, median & mode of 'Salary' variable is 48.54, 50.0 and 40.0 respectively.
- The mean, median & mode of 'Spending' variable is 482, 500 and 500 respectively.

- The mean, median & mode of 'Text Messages' variable is 246, 200 and 300 respectively.

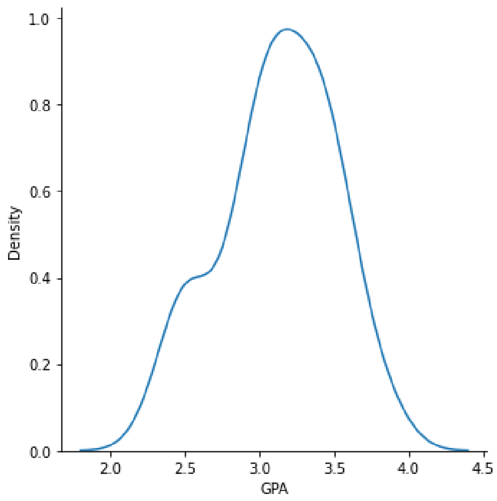


Figure 16 - GPA Kernel Density Plot

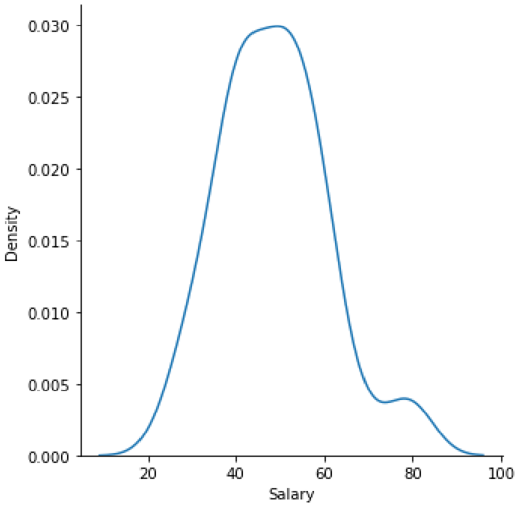


Figure 17 - Salary Kernel Density Plot

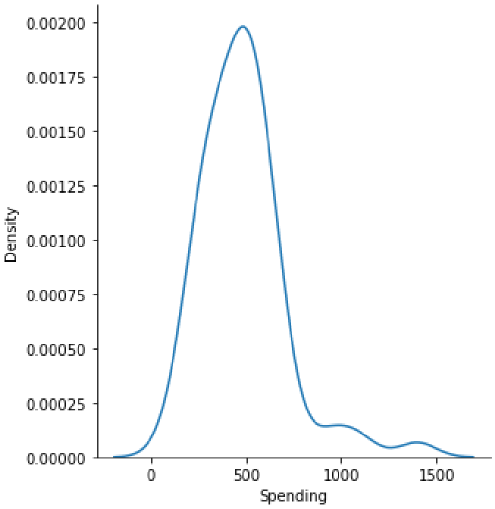


Figure 18 - Spending Kernel Density Plot

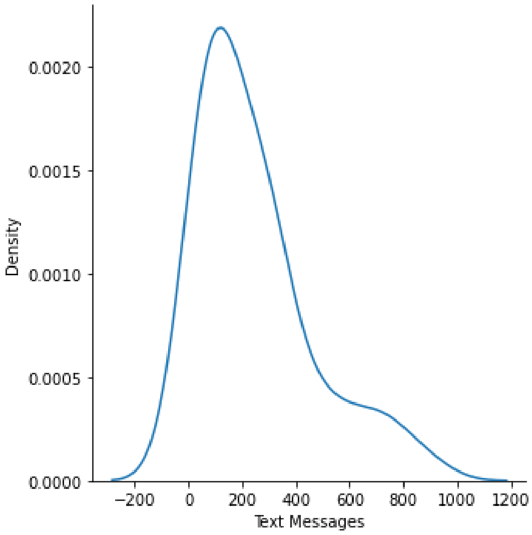


Figure 19 - Text Messages Kernel Density Plot

From the above data and plots we infer following:

- GPA is having almost equal mean, median and mode so data is normally distributed.
- Salary is not having equal mean, median and mode so data is not normally distributed.
- Spending is having almost equal mean, median and mode so data is normally distributed.
- Text Messages is not having equal mean, median and mode so data is not normally distributed.

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.

Data Description

The dataset includes 36 moisture content measurements in pounds per 100 square feet for A shingles and 31 moisture content measurements in pounds per 100 square feet for B shingles.

Q 3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

- The level of significance (Alpha) = 0.05.
- Since the population standard deviation (Sigma) is unknown, we have to use a T-test.
- We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.
- Degree of Freedom:
    - For A shingles we have 36 samples, so N-1 degrees of freedom: 35
    - For B shingles we have 31 samples, so N-1 degrees of freedom: 30
- Since the sole purpose of the test is to check whether the mean moisture content is less than 0.35 pounds per 100 square feet, we would prefer a One-tailed one sample T-test.
- Let us formulate the hypothesis:
    - H0 (null hypothesis): μ (mean moisture content) >= 0.35
    - H1 (alternate hypothesis): μ (mean moisture content) < 0.35

A Shingles:

From the one sample t-test performed, we got the below results:

t-statistic: -1.474, p-value: 0.075

At the level of 5% significance, p-value = 0.075. Since p-value > 0.05 we don't have sufficient statistic al evidence to reject the null hypothesis, that means mean moisture content is not less than 0.35 pou nds per 100 square feet.

B Shingles:

From the one sample t-test performed, we got the below results:

t-statistic: -3.1, p-value: 0.002

At the level of 5% significance, p-value = 0.002. Since p-value < 0.05 we have sufficient statistical evidence to reject the null hypothesis, that means mean moisture content is less than 0.35 pounds per 100 square feet.

Q 3.2 Do you think that the population mean 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?

- The level of significance (Alpha) = 0.05.
- Since the population standard deviation (Sigma) is unknown, we have to use a T-test.
- We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.
- Degree of Freedom:
  - For A shingles we have 36 samples, so N-1 degrees of freedom: 35
  - For B shingles we have 31 samples, so N-1 degrees of freedom: 30
- Since the sole purpose of the test is to check whether both the mean moisture content is equal or not, we would prefer a Two-tailed two sample independent T-test.
- Let us formulate the hypothesis:
  - H0 (null hypothesis): µA == µB (mean moisture content for shingles A and B are equal)
  - H1 (alternate hypothesis): µA != µB (mean moisture content for shingles A and B are not equal)

From the two-sample t-test performed, we got the below results:
t-statistic: 1.29, p-value: 0.202

At the level of 5% significance, p-value = 0.202. Since p-value > 0.05 we don’t have sufficient statistical evidence to reject the null hypothesis, that means mean moisture content is for shingles A and B are equal.

THE END!