SMDM PROJECT

Submitted by,

JIYA JACOB

PGP-DSBA ONLINE

JULY 2021

DATE:12/09/2021

WHOLESALE CUSTOMER DATA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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).

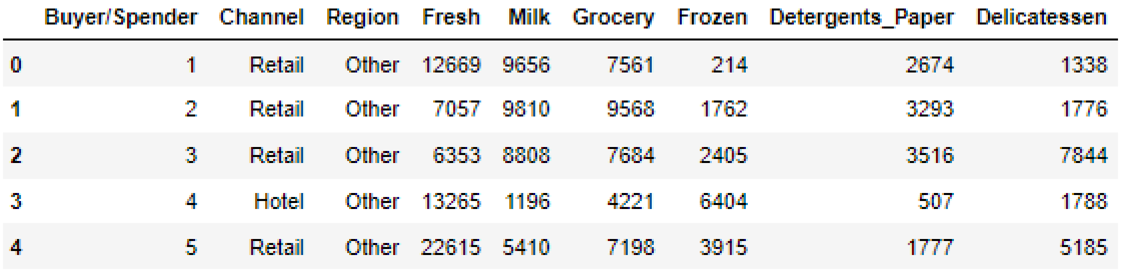
INTRODUCTION

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. Analyze the different attributes of the car make which can help in analyzing the sales of different items through various channels in 3 different regions. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

DATA DESCRIPTION

1. Buyer/Spender- serial number
2. Channel- 2 types; hotel and retail
3. 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:



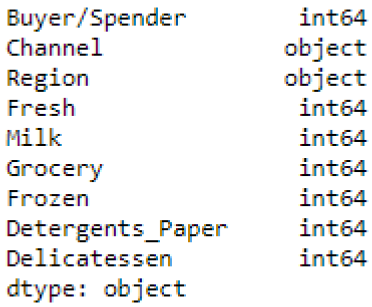
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198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

Table 1: Dataset Sample

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).

EXPLORATORY DATA ANALYSIS

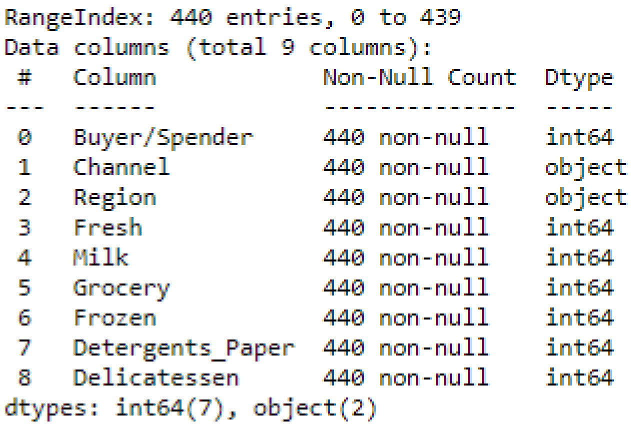
Let us check the types of variables in the data frame.



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

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.

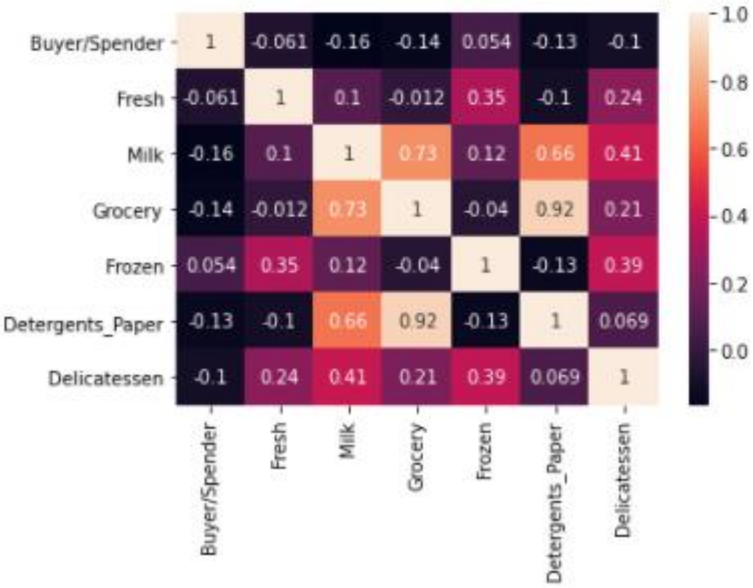
Check for missing values in the dataset:



```
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Buyer/Spender     440 non-null    int64 
 1   Channel           440 non-null    object
 2   Region            440 non-null    object
 3   Fresh             440 non-null    int64 
 4   Milk              440 non-null    int64 
 5   Grocery           440 non-null    int64 
 6   Frozen            440 non-null    int64 
 7   Detergents_Paper  440 non-null    int64 
 8   Delicatessen      440 non-null    int64 
dtypes: int64(7), object(2)
```

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

Correlation Plot



	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Buyer/Spender	1	-0.061	-0.16	-0.14	0.054	-0.13	-0.1
Fresh	-0.061	1	0.1	-0.012	0.35	-0.1	0.24
Milk	-0.16	0.1	1	0.73	0.12	0.66	0.41
Grocery	-0.14	-0.012	0.73	1	-0.04	0.92	0.21
Frozen	0.054	0.35	0.12	-0.04	1	-0.13	0.39
Detergents_Paper	-0.13	-0.1	0.66	0.92	-0.13	1	0.069
Delicatessen	-0.1	0.24	0.41	0.21	0.39	0.069	1

FIG 1: CORRELATION HEATMAP

From the correlation plot, we can see that sales of various items are highly correlated to each other. 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.

Pair plot

Pair plot 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 Detergents_paper. From the histogram we can see that the price of the whole dataset is left skewed.

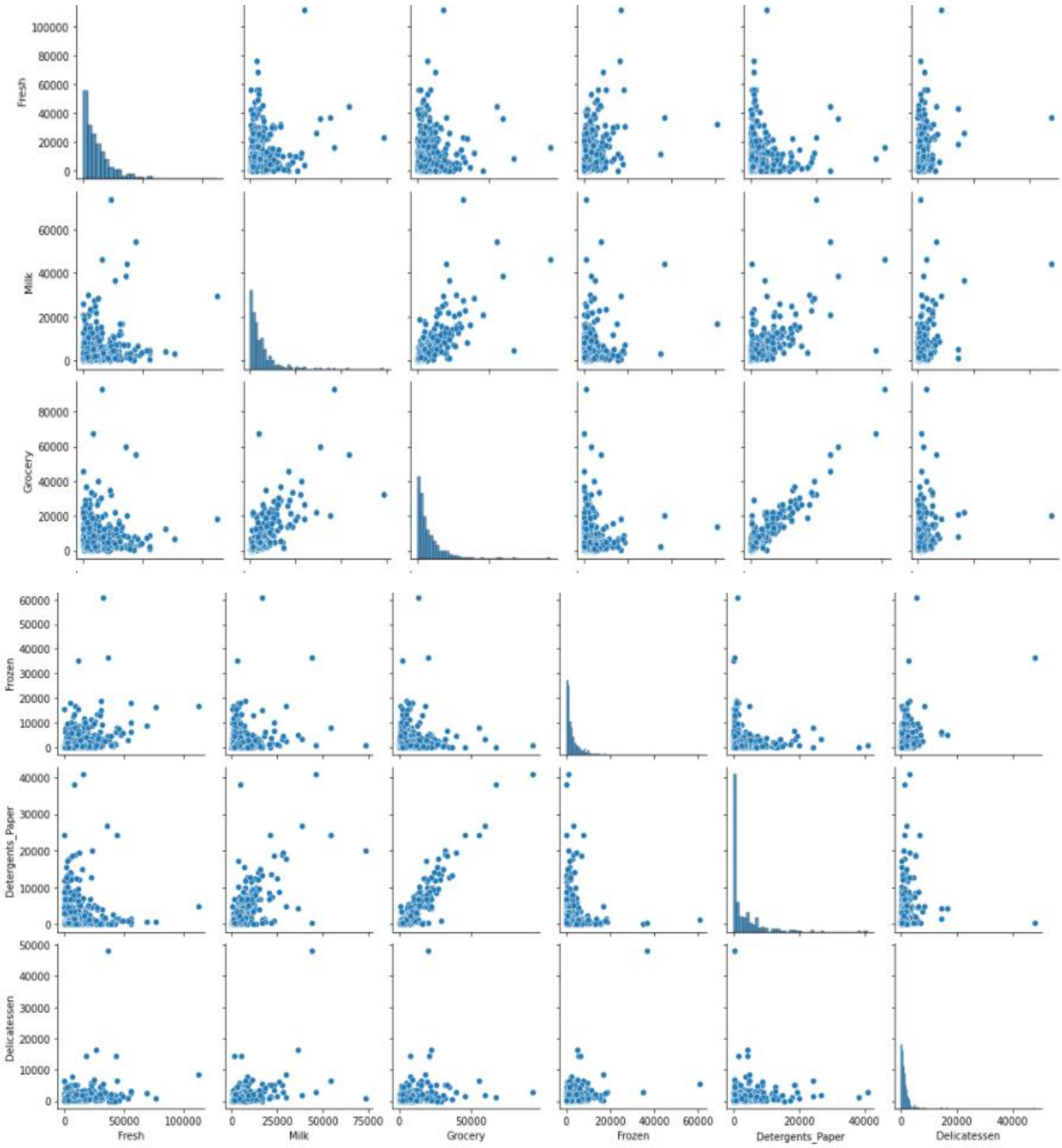
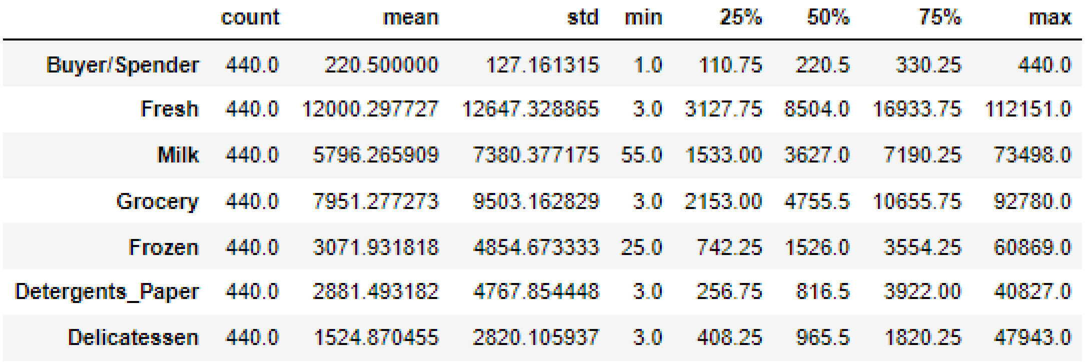


FIG 2: PAIR PLOT

Q1.1: **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 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 centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

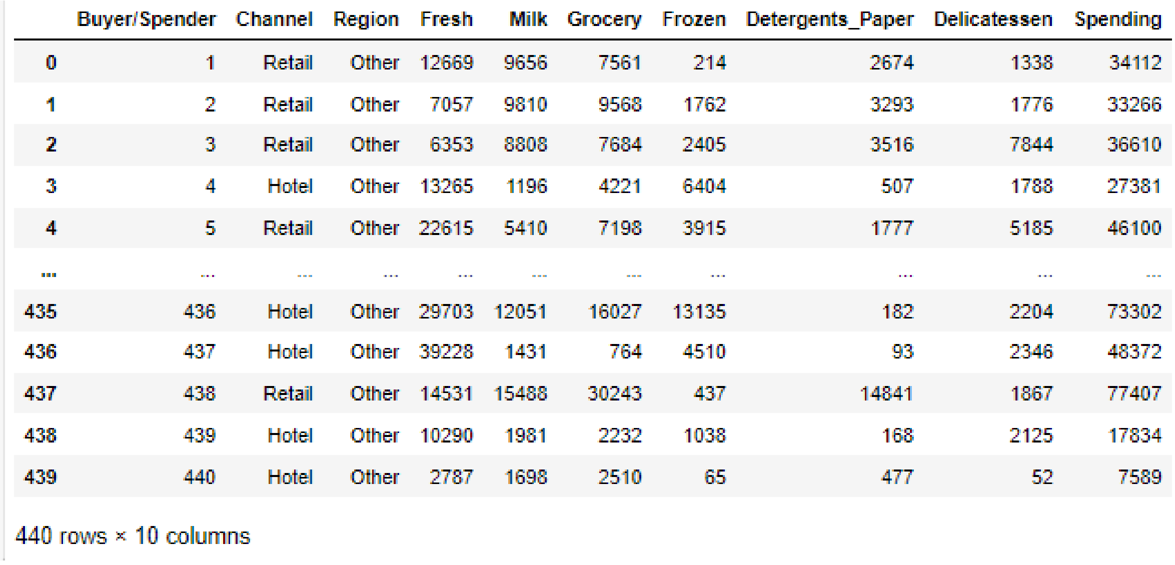


	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

TABLE 2: SUMMARY OF THE DATA

From the descriptive statistics, we can see that there are 6 different types od products that are being sold through the 2 sales channels namely, hotel and retail in 3 different regions i.e., Lisbon, Oporto and other. From the above table, we can see that the annual spending on the item Fresh was the maximum which grosses up to 12000.297727 whereas the annual spending on the item Delicatessen was the least which just sums up to 1524.870455.

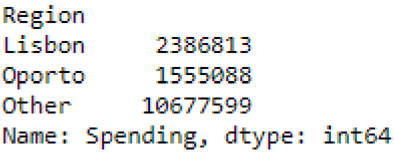
Calculating the total spending of each channel and region



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Spending
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100
...	...	...	...	...	...	...	...	...	...	...
435	436	Hotel	Other	29703	12051	16027	13135	182	2204	73302
436	437	Hotel	Other	39228	1431	764	4510	93	2346	48372
437	438	Retail	Other	14531	15488	30243	437	14841	1867	77407
438	439	Hotel	Other	10290	1981	2232	1038	168	2125	17834
439	440	Hotel	Other	2787	1698	2510	65	477	52	7589

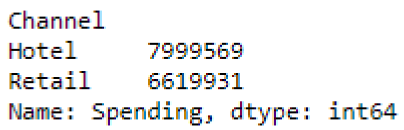
440 rows × 10 columns

Table 3: The total spending of each channel and region.



```
Region
Lisbon     2386813
Oporto     1555088
Other     10677599
Name: Spending, dtype: int64
```

From the above given table, we can conclude that in the region wise, Other spend the most with 10677599 while Oporto spends the least with 1555088.



```
Channel
Hotel     7999569
Retail    6619931
Name: Spending, dtype: int64
```

From the above given table, we can conclude that in the channel wise, Hotel spend the most with 7999569 while Oporto spends the least with 6619931.

Q1.2: **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.**

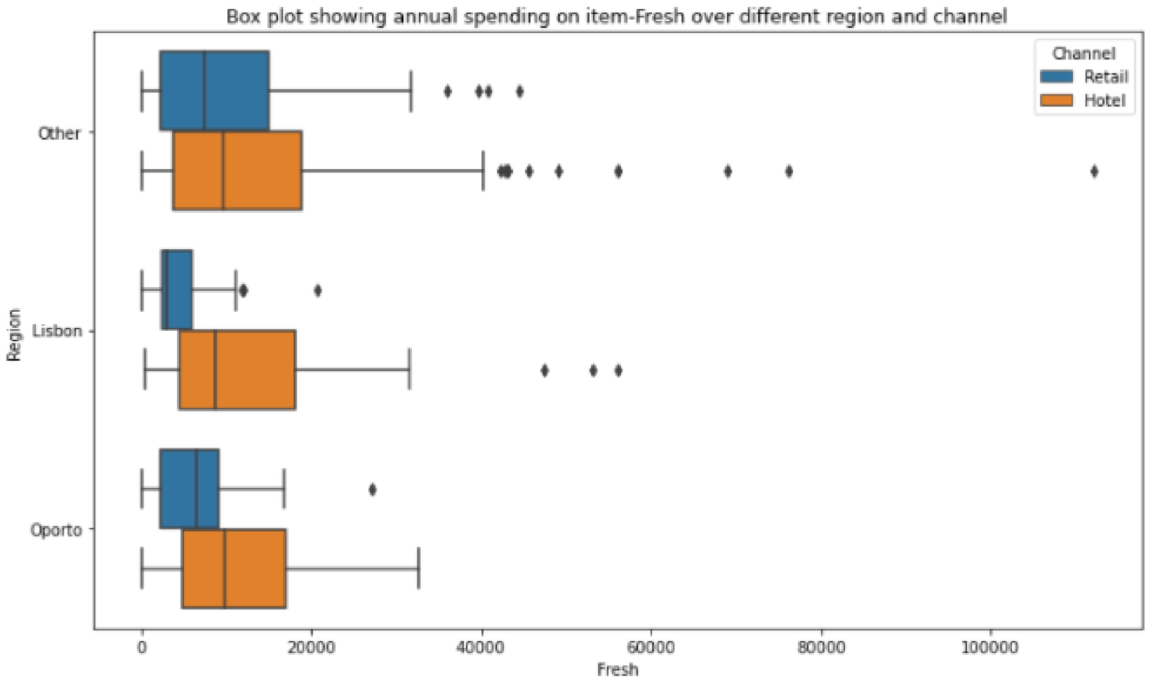


Fig 3.1. Box plot showing the annual spending on item -Fresh over different regions and channels

In the case of item Fresh, the annual spending is more through the hotel channel in all the three regions. The annual spending for item Fresh is more in the region “Other” and the least in “Lisbon”. Since for all the regions, the values are much plotted to the left side, it is a left skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the retail channel in all the three regions are lower than the median of the annual spending through the hotel channel.

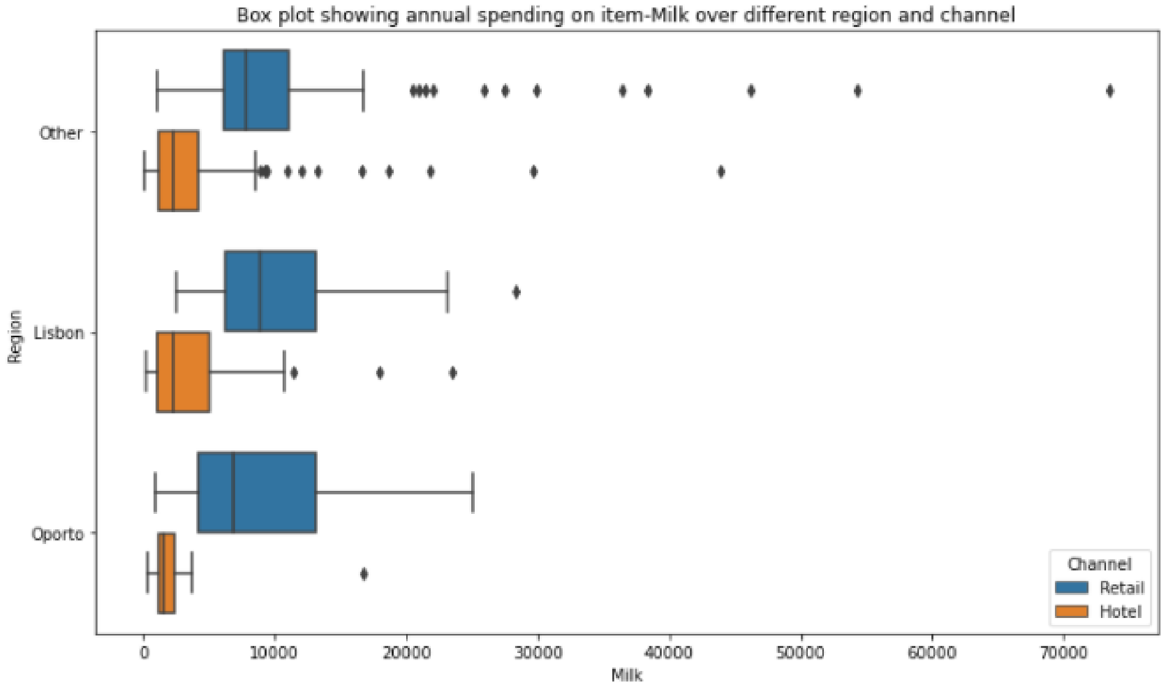


Fig 3.2. Box plot showing the annual spending on item -Milk over different regions and channels

In the case of item Milk, the annual spending is more through the retail channel in all the three regions. The annual spending for item Milk is more in the region “Other” and the least in “Lisbon”. Since for all the regions, the values are much plotted to the right side, it is a right skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the retail channel in all the three regions are higher than the median of the annual spending through the hotel channel.

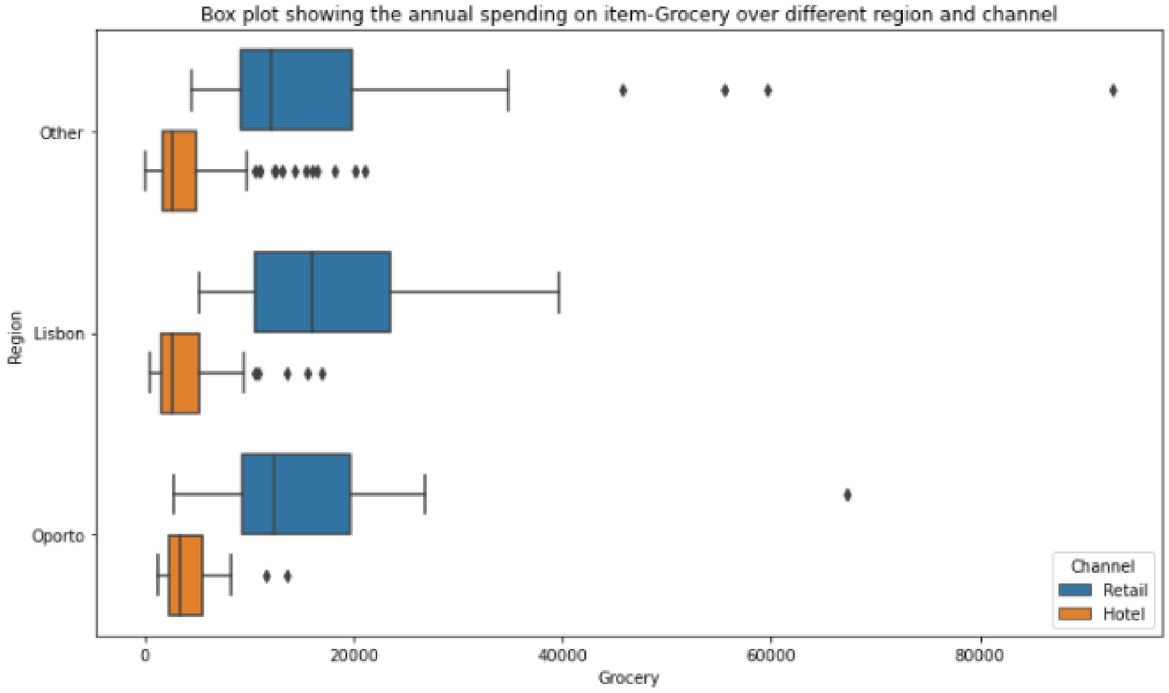


Fig 3.3. Box plot showing the annual spending on item -Grocery over different regions and channels

In the case of item Grocery, the annual spending is more through the Retail channel in all the three regions. The annual spending for item Retail is more in the region “Other” and the least in “Oporto”. Since for all the regions, the values are much plotted to the left side, it is a left skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the retail channel in all the three regions are higher than the median of the annual spending through the hotel channel.

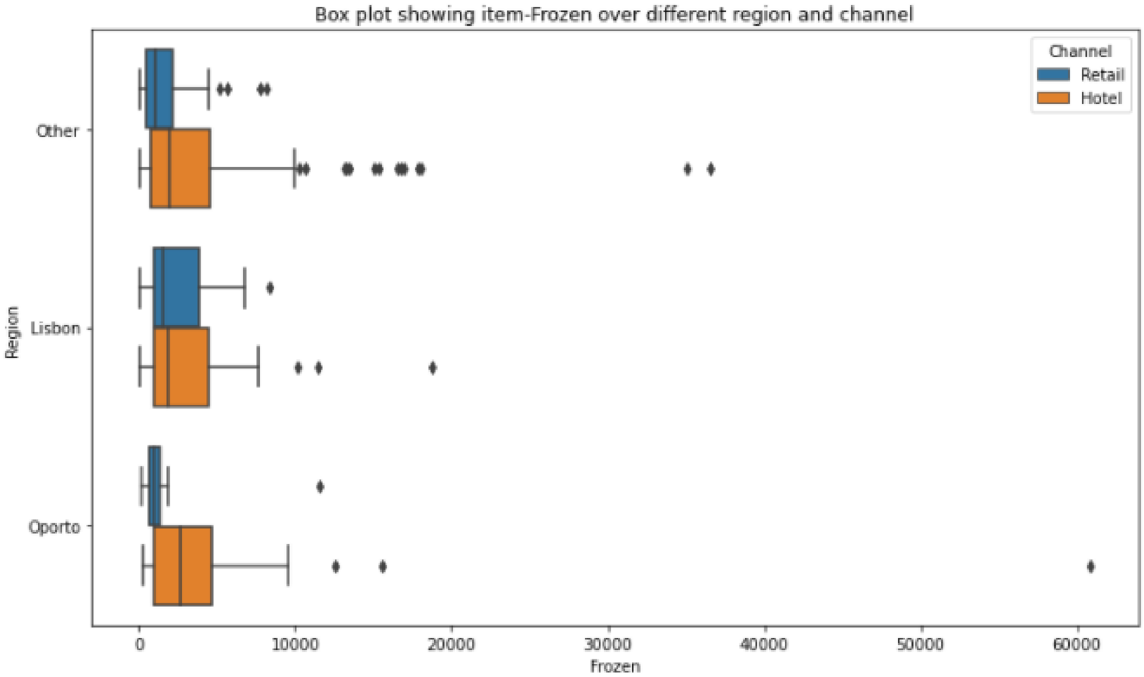


Fig 3.4. Box plot showing the annual spending on item -Frozen over different regions and channels

In the case of item Frozen, the annual spending is more through the hotel channel in all the three regions. The annual spending for item Frozen is more in the region “Other” and the least in “Oporto”. Since for all the regions, the values are much plotted to the right side, it is a right skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the hotel channel in two regions are higher than the median of the annual spending through the retail channel.

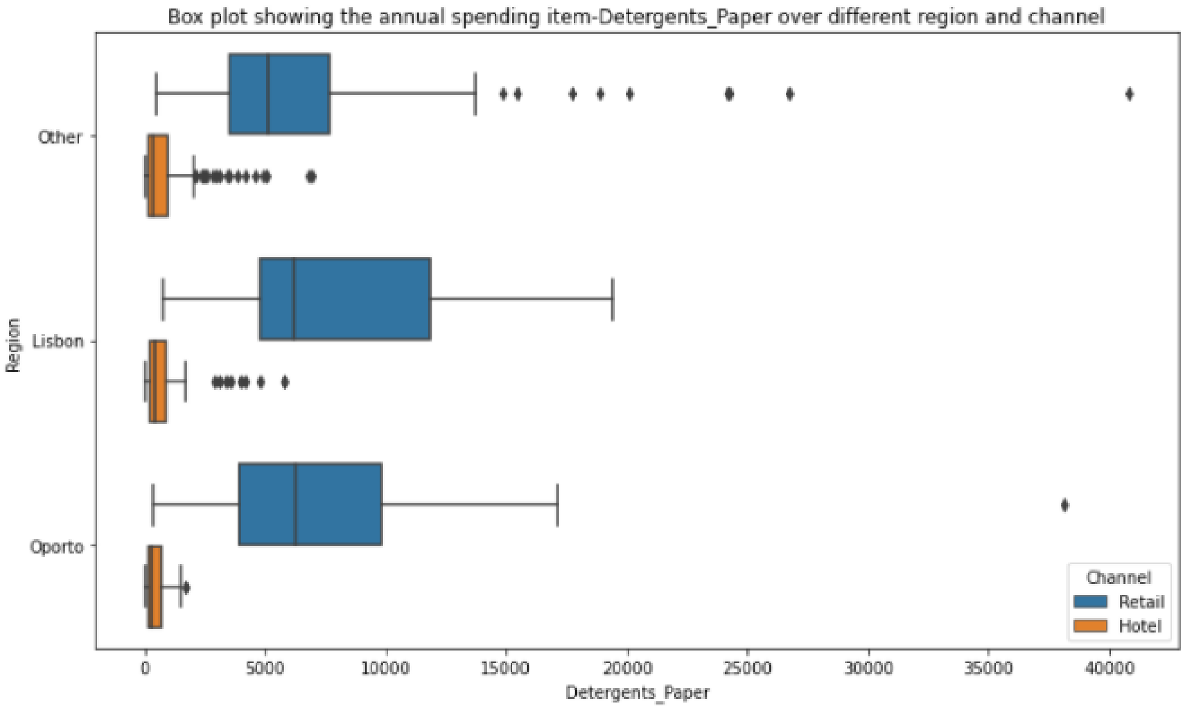


Fig 3.5. Box plot showing the annual spending on item -Detergents_Paper over different regions and channels

In the case of item Detergents_Paper, the annual spending is more through the retail channel in all the three regions. The annual spending for item Detergetns_Paper is more in the region “Other” and the least in “Oporto”. Since for all the regions, the values are much plotted to the right side, it is a right skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the retail channel in all the three regions are higher than the median of the annual spending through the hotel channel.

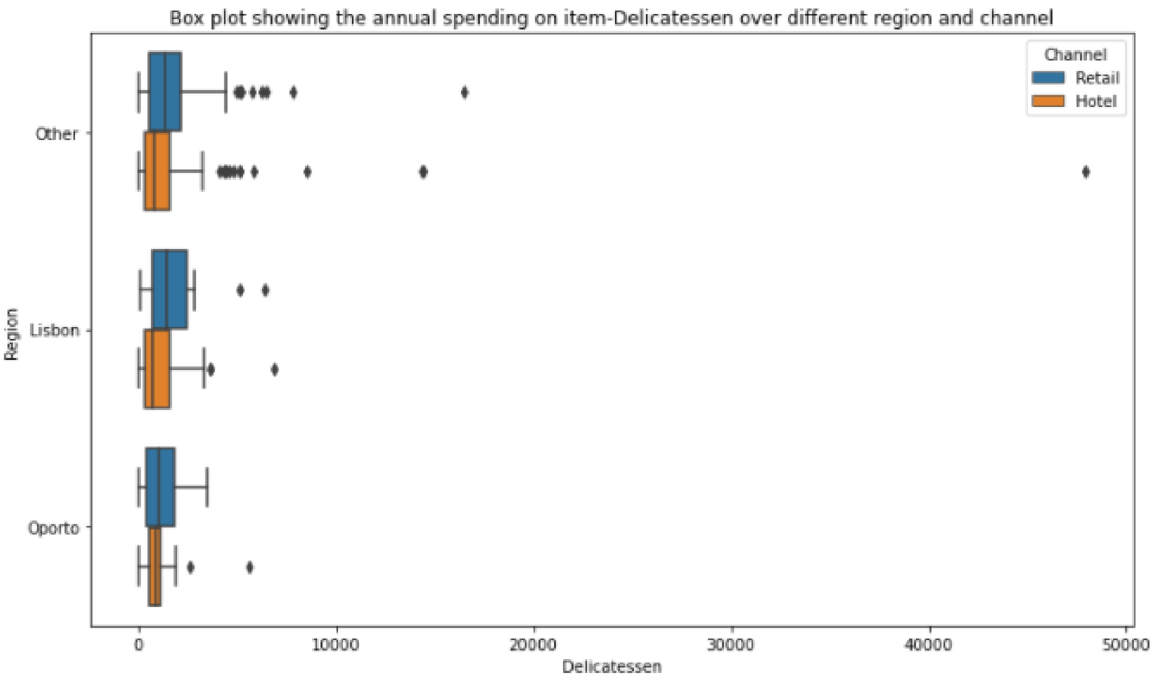
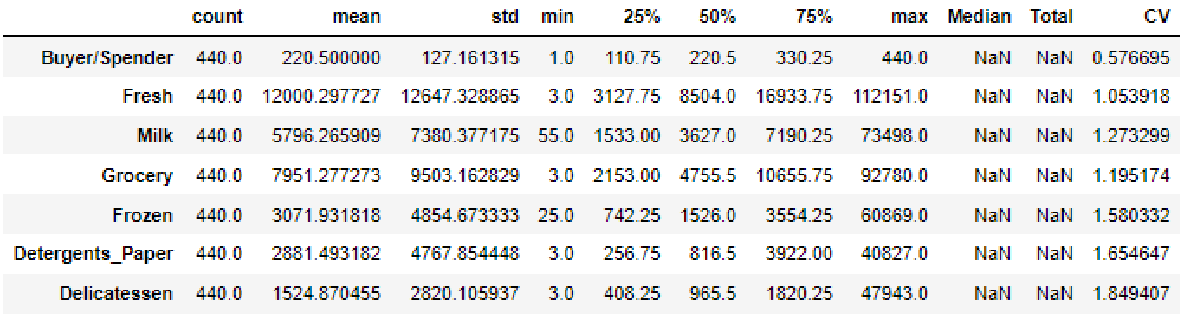


Fig 3.6. Box plot showing the annual spending on item -Delicatessen over different regions and channels

In the case of item Delicatessen, the annual spending is more through the retail channel in all the three regions. The annual spending for item Fresh is more in the region “Other” and the least in “Oporto”. Since for all the regions, the values are much plotted to the right side, it is a right skewed distribution. The longer the box, the more dispersed the data. Accordingly, the region “Other” had the most dispersed data and the greatest number of outliers. Therefore, the region “Other “has the highest range. The median of the annual spending through the retail channel in all the three regions are higher than the median of the annual spending through the hotel channel.

Q1.3: **On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?**



	count	mean	std	min	25%	50%	75%	max	Median	Total	CV
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0	NaN	NaN	0.576695
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	NaN	NaN	1.053918
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	NaN	NaN	1.273299
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	NaN	NaN	1.195174
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	NaN	NaN	1.580332
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	NaN	NaN	1.654647
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	NaN	NaN	1.849407

Table 4: Table showing the coefficient of variation of different items.

The coefficient of variation measures how consistent the different values of the set are from the mean of the dataset. The smaller the cv, higher is the consistency. According to the table shown above, item showing the most inconsistent behaviour is Delicatessen whose cv is 1.849407, while the item showing the least inconsistent behaviour is Fresh whose cv is just 1.053918.

Q1.4:**Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

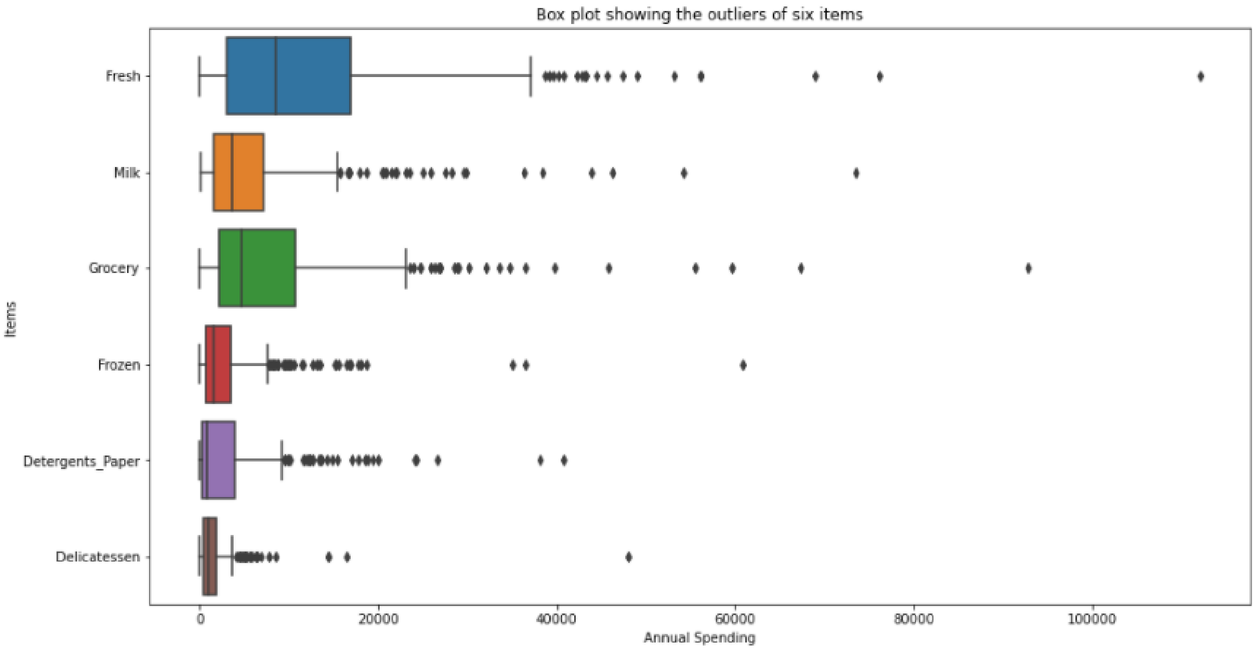


Fig:4 The box plot portraying the outliers of different items

An outlier is an observation that is numerically distant from the rest of the data. While reviewing the boxplot, outliers are the data points that lie outside the whiskers of the box plot. From the above given box plot, we can see that Item fresh has the greatest number of outliers while the item delicatessen has the least number of outliers. Similarly, the greater the distance of the outliers from the median, the greater is the range. Therefore, the item Fresh has the highest range while the item Delicatessen has the smallest range.

Q1.5:**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**

Analysis 1: Out of the 6 items, delicatessen and detergents_paper has comparatively lesser sales than others. While comparing their coefficient of variations we can understand their inconsistency. So wholesale distributor should take measures to take care of the sales of these two items.

Analysis 2: The sales of all the 6 items are highest in the other regions apart from Lisbon and Oporto. So, the wholesale distributor should focus on other regions to increase the sales of the items. Oporto region has the least sales for most for the items. So, the suggestion of eliminating the region “Oporto” is put forward. Among the channels Retail spends the most for the items. So, the sales through retail should be more focussed than through the channel hotel.

UNIVERSITY SURVEY DATA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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).

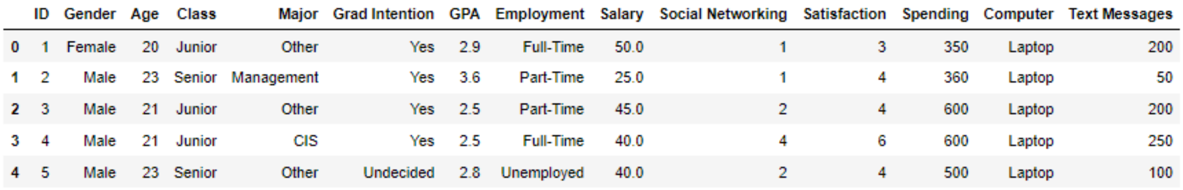
INTRODUCTION

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. Analyze the different attributes of the car make which can help in analyzing the price of the car. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

DATA DESCRIPTION

1. ID- id number of the students from 1 to 62
2. Gender-two types; male or female
3. Age-age of the students; continuous from 18 to 26
4. Class-three types; Junior, senior, sophomore
5. Major- eight types; Management, CIS, Economics/Finance, Retailing/Marketing, Accounting, International Business, Undecided, Other
6. Grad Intention- three types; yes, no, undecided
7. GPA- continuous from 2.3 to 3.9
8. Employment-three times; full-time, part-time, unemployed
9. Salary-continuous from 25.0 to 85.0
10. Social Networking-continuous from 0 to 4
11. Satisfaction-continuous from 1 to 6
12. Spending-continuous from 100 to 1400
13. Computer- two types; Laptops and desktops
14. Text Messages-continuous from 0 to 900

Sample of the dataset:



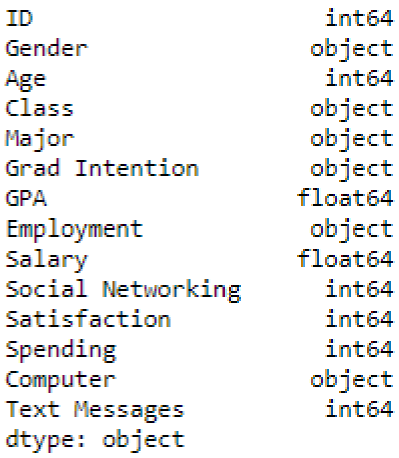
	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	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 1: Dataset Sample

Dataset has 62 entries with 14 variables of 3 datatypes. A survey of 14 questions is asked to the 62 undergraduates and their responses are noted down.

EXPLORATORY DATA ANALYSIS

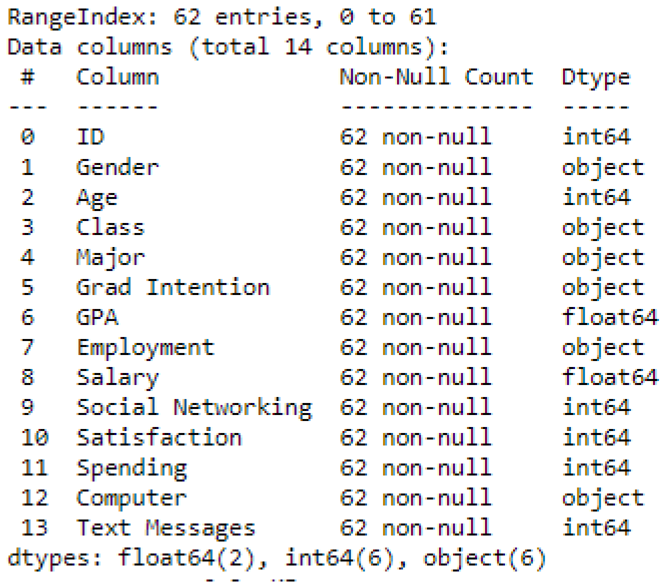
Let us check the types of variables in the data frame.



```
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
dtype: object
```

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

Check for missing values in the dataset:



```
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   ID                 62 non-null     int64  
 1   Gender             62 non-null     object 
 2   Age                62 non-null     int64  
 3   Class              62 non-null     object 
 4   Major              62 non-null     object 
 5   Grad Intention     62 non-null     object 
 6   GPA                62 non-null     float64
 7   Employment         62 non-null     object 
 8   Salary             62 non-null     float64
 9   Social Networking  62 non-null     int64  
 10  Satisfaction       62 non-null     int64  
 11  Spending           62 non-null     int64  
 12  Computer           62 non-null     object 
 13  Text Messages      62 non-null     int64  
dtypes: float64(2), int64(6), object(6)
```

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

Correlation Plot

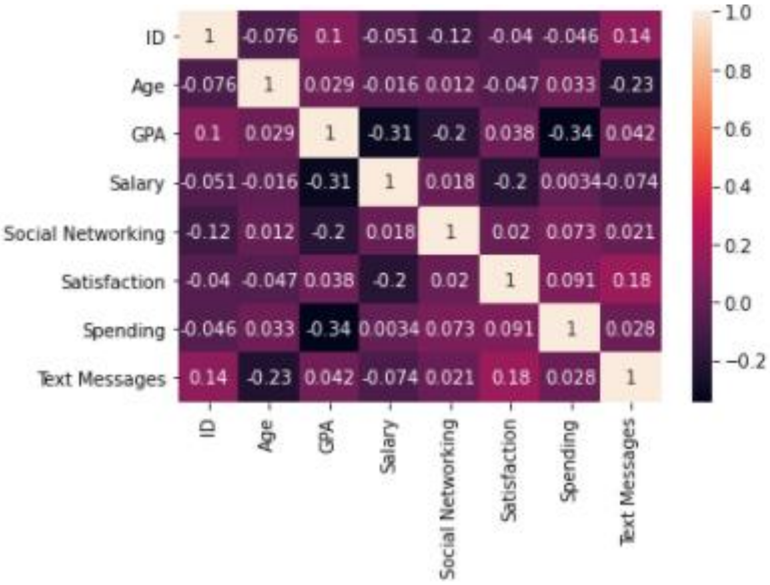


Fig 1: Correlation heatmap

From the correlation plot, we can see that various attributes of the student survey are highly correlated to each other. 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.

Pair plot

Pair plot 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 np kind of linear relationship between any variables in the dataset. But there are many independent variables such as GPA, age, salary and social networking.

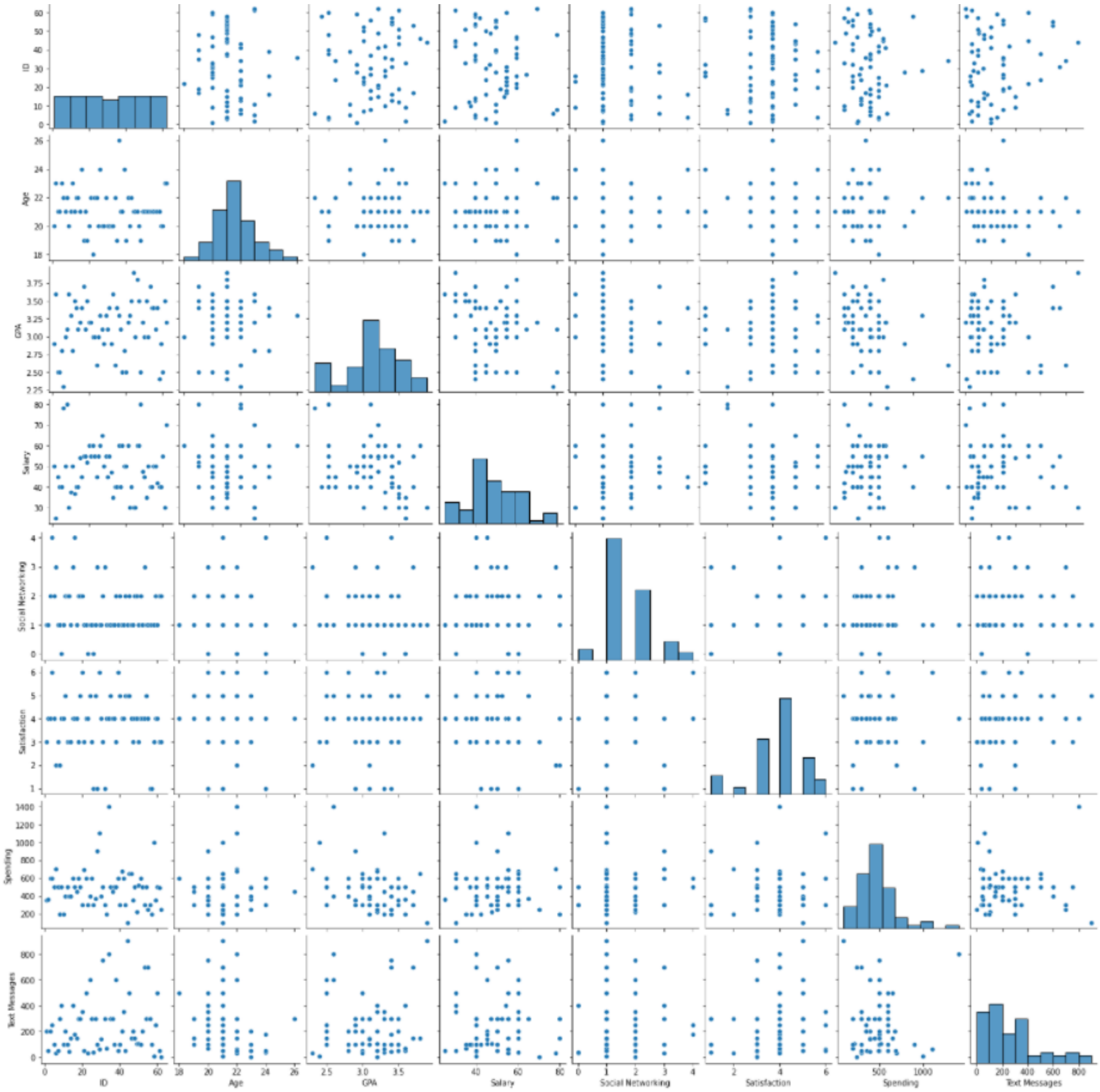
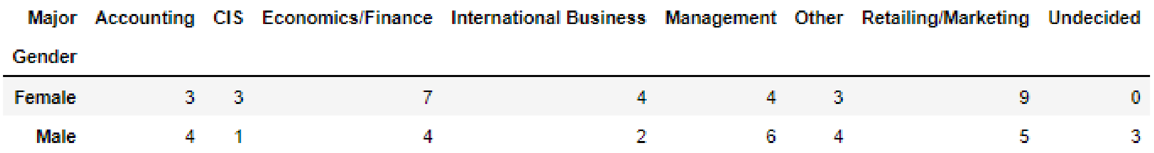


Fig 2: Pair plot of various variables in the student survey.

Q2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

Q2.1.1. Gender and Major

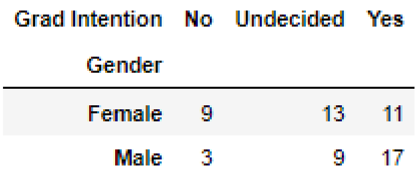


Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Contingency table between gender and Major

The above contingency table is created keeping the gender as row variable and major as column variable.

Q2.1.2. Gender and Grad Intention

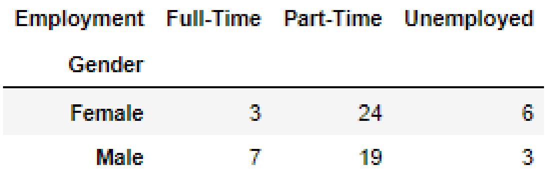


Grad Intention	No	Undecided	Yes
Gender			
Female	9	13	11
Male	3	9	17

Contingency table between Gender and Grad Intention

The above contingency table is created keeping the gender as row variable and major as column variable.

Q2.1.3. Gender and Employment

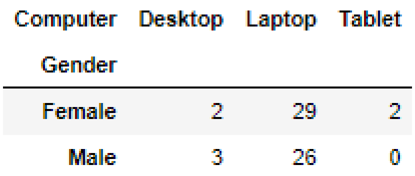


Employment	Full-Time	Part-Time	Unemployed
Gender			
Female	3	24	6
Male	7	19	3

Contingency table between Gender and employment

The above contingency table is created keeping the gender as row variable and major as column variable.

Q2.1.4. Gender and Computer



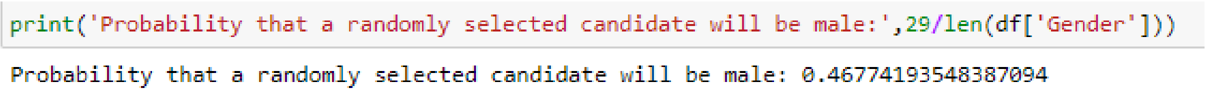
Computer Gender	Desktop	Laptop	Tablet
Female	2	29	2
Male	3	26	0

Contingency table between Gender and Computer

The above contingency table is created keeping the gender as row variable and major as column variable.

Q2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

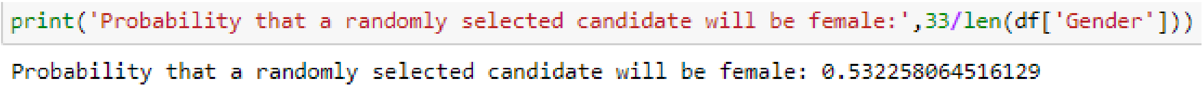
Q2.2.1. What is the probability that a randomly selected CMSU student will be male?



```
print('Probability that a randomly selected candidate will be male:',29/len(df['Gender']))
```

```
Probability that a randomly selected candidate will be male: 0.46774193548387094
```

Q2.2.2. What is the probability that a randomly selected CMSU student will be female?

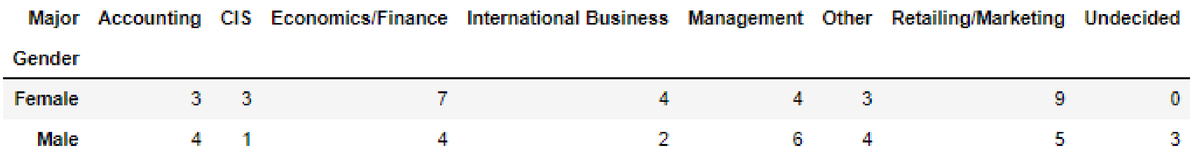


```
print('Probability that a randomly selected candidate will be female:',33/len(df['Gender']))
```

```
Probability that a randomly selected candidate will be female: 0.532258064516129
```

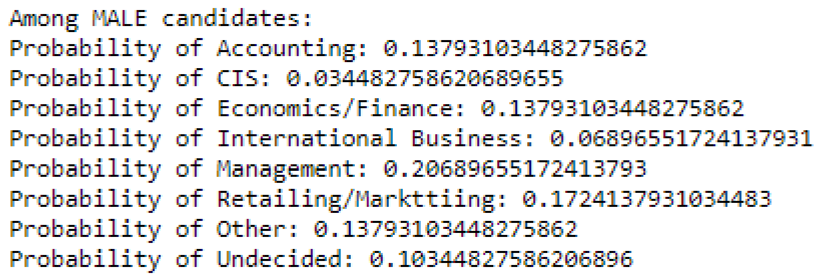
Q2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

Q2.3.1. Find the conditional probability of different majors among the male students in CMSU.



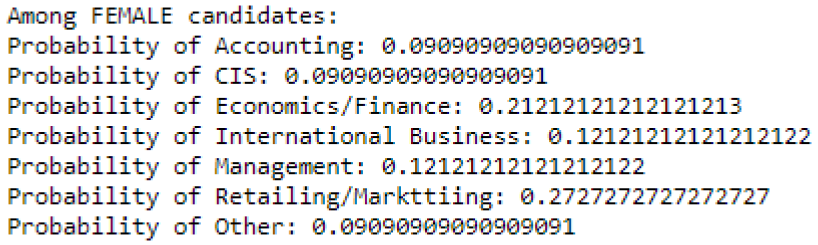
Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

Contingency table of Gender and Major



```
Among MALE candidates:
Probability of Accounting: 0.13793103448275862
Probability of CIS: 0.034482758620689655
Probability of Economics/Finance: 0.13793103448275862
Probability of International Business: 0.06896551724137931
Probability of Management: 0.20689655172413793
Probability of Retailing/Markttiing: 0.1724137931034483
Probability of Other: 0.13793103448275862
Probability of Undecided: 0.10344827586206896
```

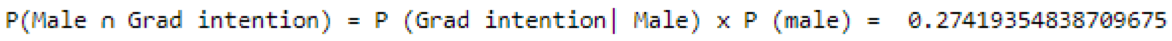
Q2.3.2 Find the conditional probability of different majors among the female students of CMSU.



```
Among FEMALE candidates:
Probability of Accounting: 0.09090909090909091
Probability of CIS: 0.09090909090909091
Probability of Economics/Finance: 0.21212121212121213
Probability of International Business: 0.12121212121212122
Probability of Management: 0.12121212121212122
Probability of Retailing/Markttiing: 0.2727272727272727
Probability of Other: 0.09090909090909091
```

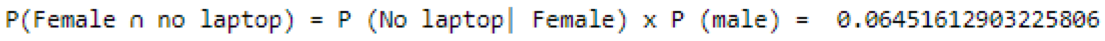
Q2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

Q2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.



```
P(Male ∩ Grad intention) = P (Grad intention| Male) x P (male) =  0.27419354838709675
```

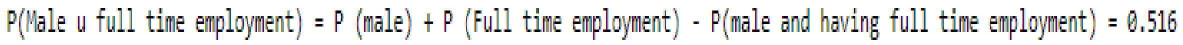
Q2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.



```
P(Female ∩ no laptop) = P (No laptop| Female) x P (male) =  0.06451612903225806
```

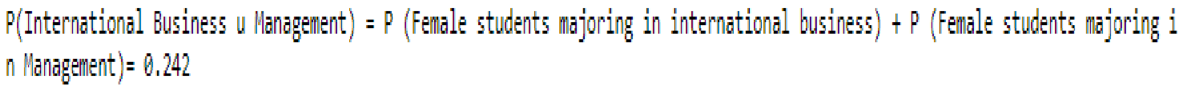
Q2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

Q2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment?



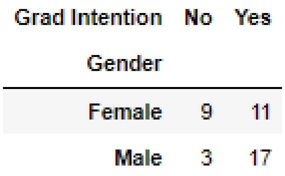
```
P(Male u full time employment) = P (male) + P (Full time employment) - P(male and having full time employment) = 0.516
```

Q2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.



```
P(International Business u Management) = P (Female students majoring in international business) + P (Female students majoring i
n Management)= 0.242
```

Q2.6. 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?



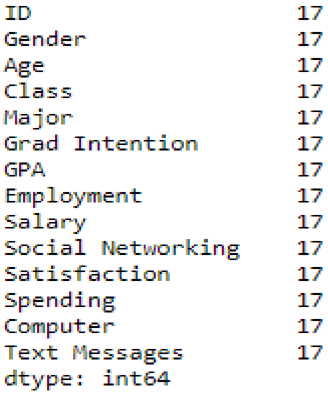
Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

A 2*2 matrix of the female students who intends to graduate or not

Q2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

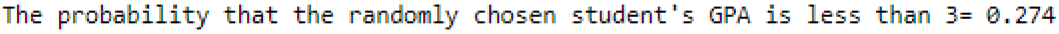
Answer the following questions based on the data

Q2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?



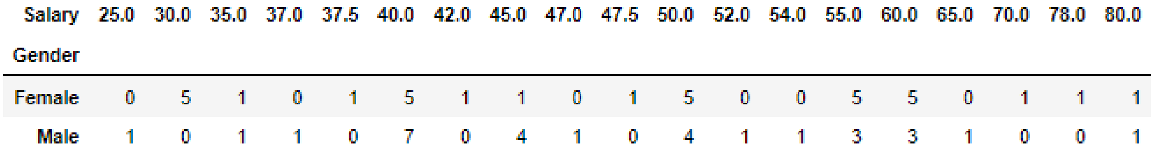
```
ID                     17
Gender                 17
Age                    17
Class                  17
Major                  17
Grad Intention         17
GPA                    17
Employment             17
Salary                 17
Social Networking      17
Satisfaction           17
Spending               17
Computer               17
Text Messages          17
dtype: int64
```

Count of the students whose GPA is less than 3



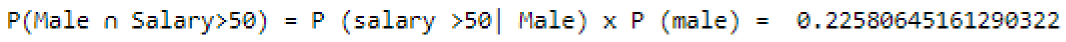
```
The probability that the randomly chosen student's GPA is less than 3= 0.274
```

Q2.7.2. 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.



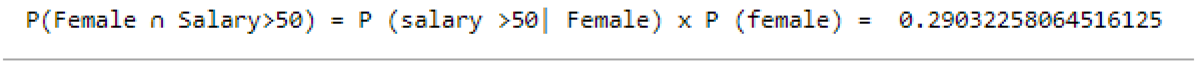
Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0
Gender																			
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1

Contingency table taking gender as row variable and salary as column variable



```
P(Male ∩ Salary>50) = P (salary >50| Male) x P (male) =  0.22580645161290322
```

This shows the conditional probability that a randomly selected male earns 50 or more and it is 0.22580645161290322



```
P(Female ∩ Salary>50) = P (salary >50| Female) x P (female) =  0.29032258064516125
```

This shows the conditional probability that a randomly selected female earns 50 or more and it is 0.29032258064516125

Q2.8. 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.

Hsitogram is a univariate plot and it depicts the variation of a numerical column or continuous variables. The most obvious way to tell if a distribution is a normal distribution is to look at the histogram itself. If the graph is approximately bell-shaped and symmetric about the mean, you can usually assume it is a normal distribution.

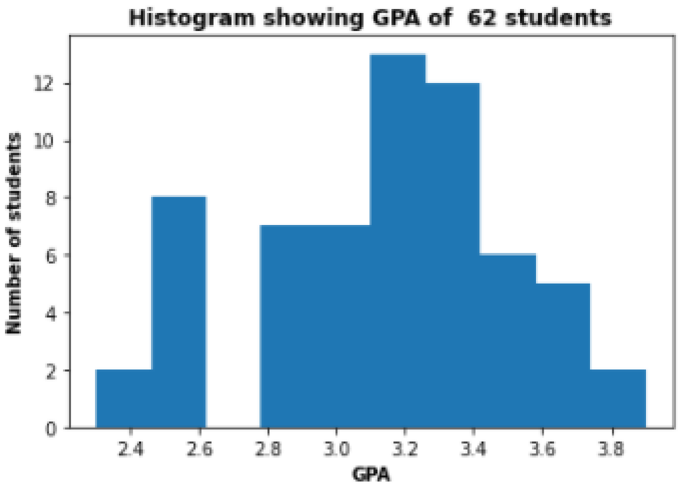
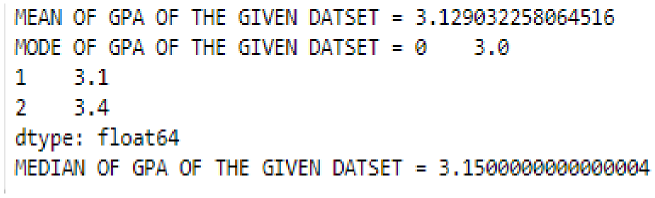


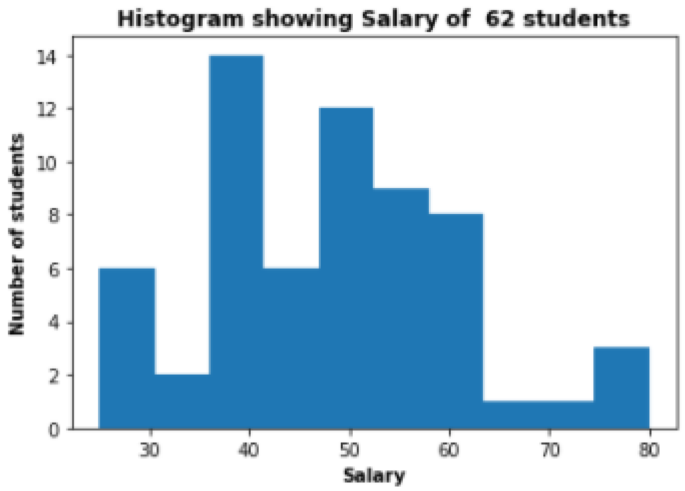
Fig 2: Histogram showing GPA of 62 students

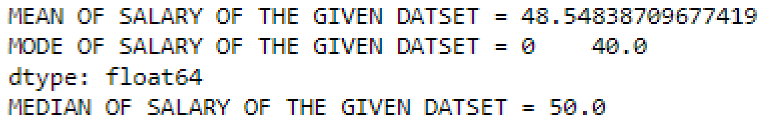


```
MEAN OF GPA OF THE GIVEN DATSET = 3.129032258064516
MODE OF GPA OF THE GIVEN DATSET = 0    3.0
1    3.1
2    3.4
dtype: float64
MEDIAN OF GPA OF THE GIVEN DATSET = 3.1500000000000004
```

This shows the mean, median and mode of the GPA of the students in the given dataset.

From the histogram we can conclude that GPA of the students doesn’t follow a normal distribution. If most of the data are on the right, with a few smaller values showing up on the left side of the histogram, the data are skewed to the left. Hence, this data is skewed to left. When data are skewed left, the mean is smaller than the median as shown above.

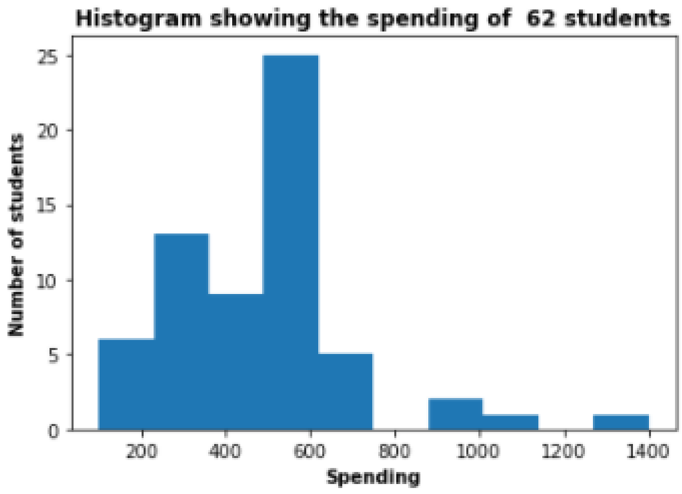


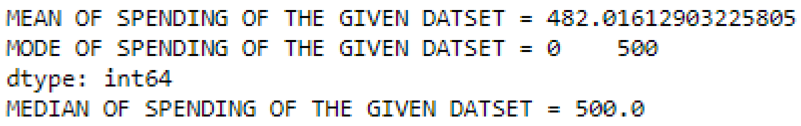


```
MEAN OF SALARY OF THE GIVEN DATSET = 48.54838709677419
MODE OF SALARY OF THE GIVEN DATSET = 0    40.0
dtype: float64
MEDIAN OF SALARY OF THE GIVEN DATSET = 50.0
```

This shows the mean, median and mode of the salary of the students in the given dataset.

From the histogram we can conclude that Salary of the students doesn’t follow a normal distribution. If most of the data are on the right, with a few smaller values showing up on the left side of the histogram, the data are skewed to the left. Hence, this data is skewed to left. When data are skewed left, the mean is smaller than the median as shown above.

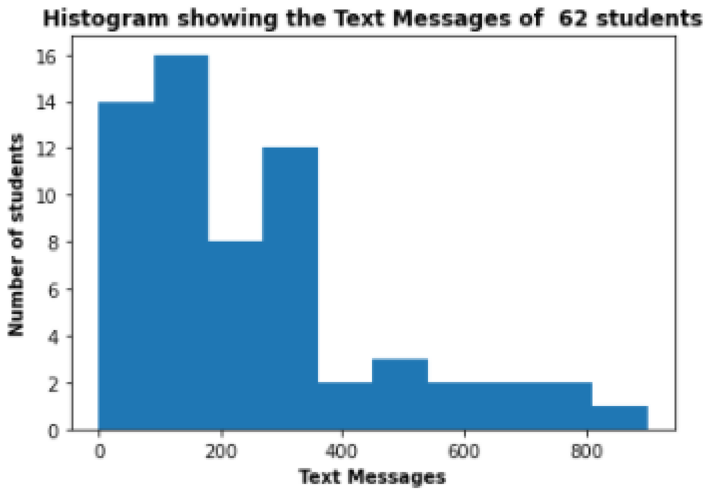


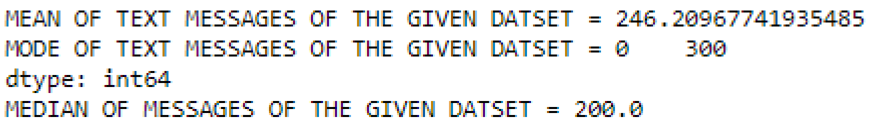


```
MEAN OF SPENDING OF THE GIVEN DATSET = 482.01612903225805
MODE OF SPENDING OF THE GIVEN DATSET = 0    500
dtype: int64
MEDIAN OF SPENDING OF THE GIVEN DATSET = 500.0
```

This shows the mean, median and mode of the Spending of the students in the given dataset.

From the histogram we can conclude that Spending of the students doesn't follow a normal distribution. If most of the data are on the right, with a few smaller values showing up on the left side of the histogram, the data are skewed to the left. Hence, this data is skewed to left. When data are skewed left, the mean is smaller than the median as shown above.





```
MEAN OF TEXT MESSAGES OF THE GIVEN DATSET = 246.20967741935485
MODE OF TEXT MESSAGES OF THE GIVEN DATSET = 0    300
dtype: int64
MEDIAN OF MESSAGES OF THE GIVEN DATSET = 200.0
```

This shows the mean, median and mode of the text messages of the students in the given dataset.

From the histogram we can conclude that Spending of the students doesn't follow a normal distribution. If most of the data are on the left side of the histogram but a few larger values are on the right, the data are said to be skewed to the right. when data are skewed right, the mean is larger than the median as shown above.

CONCLUSION

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). Numerous contingency tables keeping the gender variable as row and several other variables as column is created and corresponding relations is observed and studied.

MANUFACTURING SHINGLES DATA

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

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 colouring 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.

INTRODUCTION

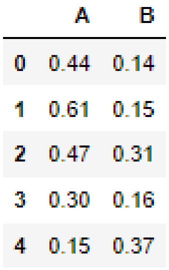
The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. Analyze the different attributes of the car make which can help in analyzing the price of the car. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

DATA DESCRIPTION

Sample A: continuous value from 0.13 to 0.72

Sample B: continuous value from 0.1 to 0.58

Sample of the dataset:



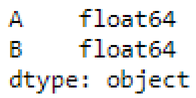
	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

Table 1: Sample dataset

Dataset has 2 variables namely A and B of float data type

EXPLORATORY DATA ANALYSIS

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

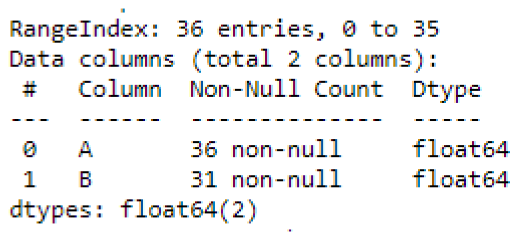


```
A      float64
B      float64
dtype: object
```

Table 2.1. Checking the types of variables in the dataset

There are total 37 rows and 2 columns in the dataset out of which both the columns are of float data type.

Check for missing values in the dataset:



```
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   A       36 non-null     float64
 1   B       31 non-null     float64
dtypes: float64(2)
```

Table 2.2. Checking the missing values in the dataset

From the above results we can see that there are 5 missing values present in the column B of the dataset.

Correlation Plot

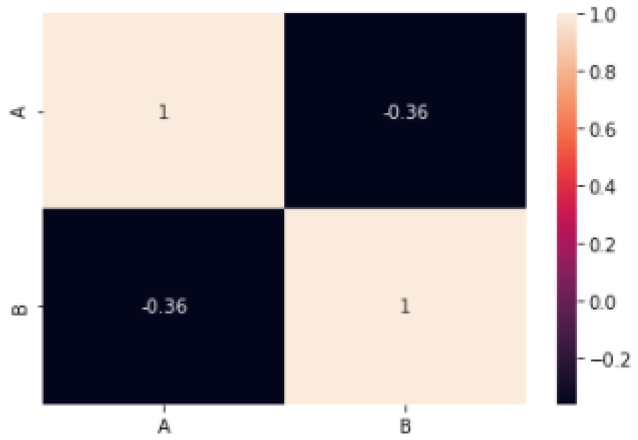


Fig 1: Correlation heatmap

From the correlation plot, we can see that samples A and B are highly correlated to each other. 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.

Q3.1 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.

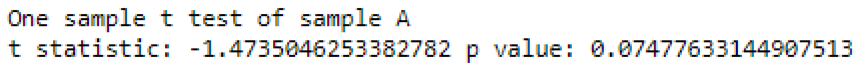
TYPE A SHINGLE

H0: The mean moisture contents in type A shingle is not within the permissible limits

H1: The mean moisture contents in type A shingle is within the permissible limits

α= 0.05, since the alpha is not given, we take the default alpha value=0.05

Even though the sample size is more than 30, the population standard deviation is not known, therefore we conduct one sample t-test.



```
One sample t test of sample A
t statistic: -1.4735046253382782 p value: 0.07477633144907513
```

Since p value > 0.05, we fail to reject the null hypothesis, ie. the mean moisture contents in type A shingle is not within the permissible limits of 0.35 pounds per 100 square feet. p-value = 0.0748 means that the probability of observing a sample of 36 shingles having the sample mean moisture content of 0.3167 pounds per 100 square feet or less is 0.074776

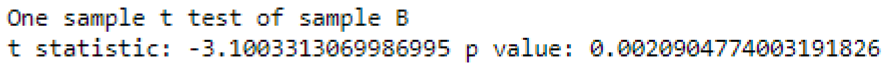
TYPE B SHINGLE

H0: The mean moisture contents in type B shingle is not within the permissible limits

H1: The mean moisture contents in type B shingle is within the permissible limits

α= 0.05, since the alpha is not given, we take the default alpha value=0.05

Even though the sample size is more than 30, the population standard deviation is not known, therefore we conduct one sample t-test.



```
One sample t test of sample B
t statistic: -3.1003313069986995 p value: 0.0020904774003191826
```

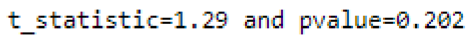
Since p value < 0.05, we reject the null hypothesis, ie. the mean moisture contents in type A shingle is within the permissible limits of 0.35 pounds per 100 square feet. p-value = 0.0021 means that the probability of observing a sample of 31 shingles having the sample mean moisture content of 0.2735 pounds per 100 square feet or less is 0.0020904.

Q3.2 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?

H0: µ(A)= µ(B); population mean of the shingles A and B are equal

Ha: µ(A)! = µ(B); population mean of shingles A and B are not equal α = 0.05; since the alpha is not given, we take the default alpha value=0.05

Since the given two samples are independent samples, we conduct 2 sample t-test



```
t_statistic=1.29 and pvalue=0.202
```

As the p value > α, we cannot reject the null hypothesis; and we can say that population mean for shingles A and B are equal.

- Assumption made is that the distribution if both the samples are normal.

CONCLUSION

Two sample t-test and one sample t -test is conducted according to the given conditions and we have arrived at appropriate conclusions.

THE END