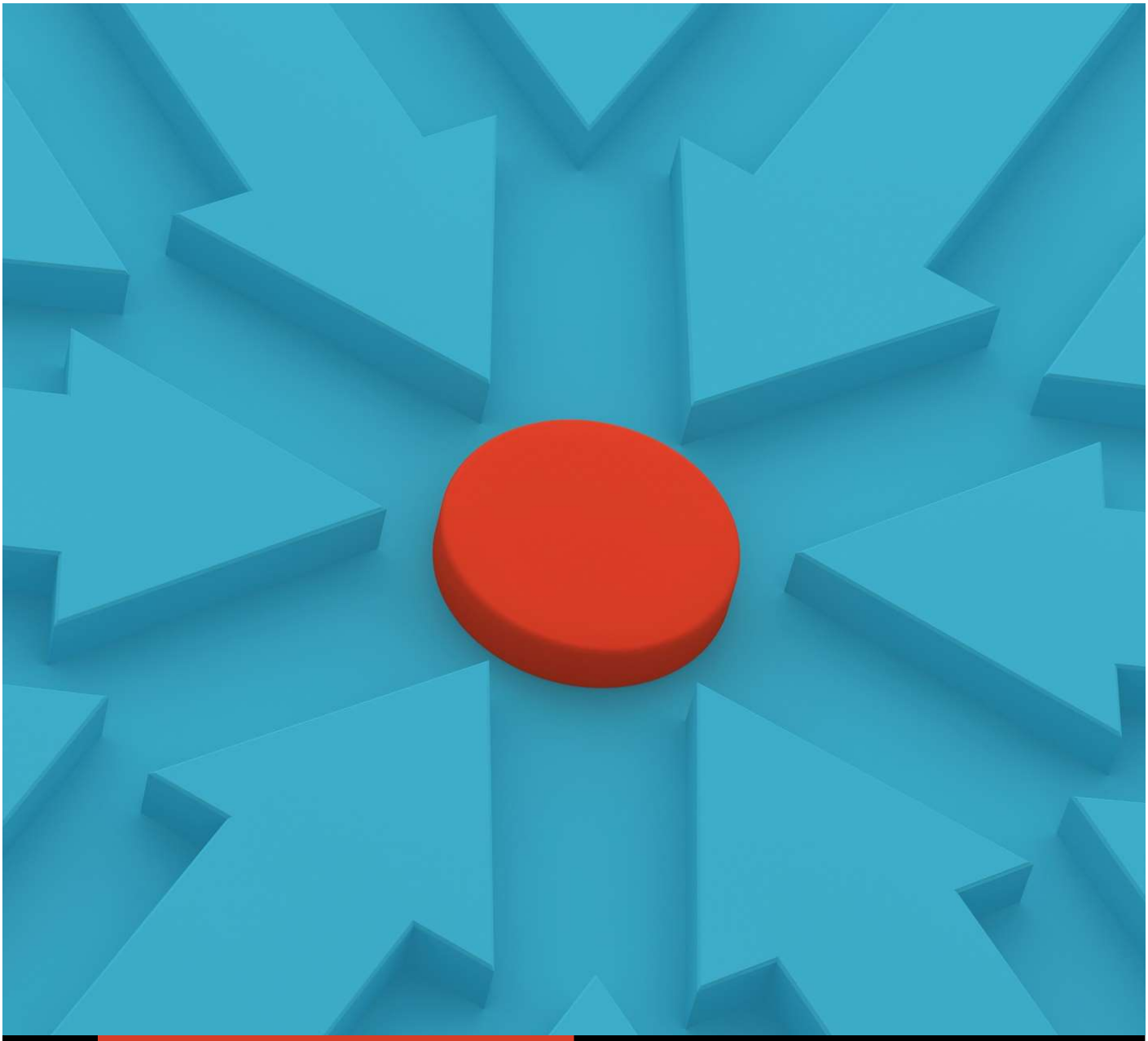


BUSINESS REPORT

Statistical Methods for Decision Making Project
PGP-DSBA Online

Athisya Nadar

9th May 2021

GREAT LEARNING **Athisya@gmail.com**

Table of Contents

Wholesale Customers Analysis 3

Q1.1 4

Q1.2 6

Q1.3 9

Q1.4 13

Q1.5 14

Clear Mountain State University (CMSU) survey 15

Q2.1.1 15

Q2.1.2 16

Q2.1.3 16

Q2.1.4 17

Q2.2.1 17

Q2.2.2 17

Q2.3.1 17

Q2.3.2 18

Q2.4.1 19

Q2.4.2 19

Q2.5.1 19

Q2.5.2 19

Q2.6 20

Q2.7.1 20

Q2.7.2 20

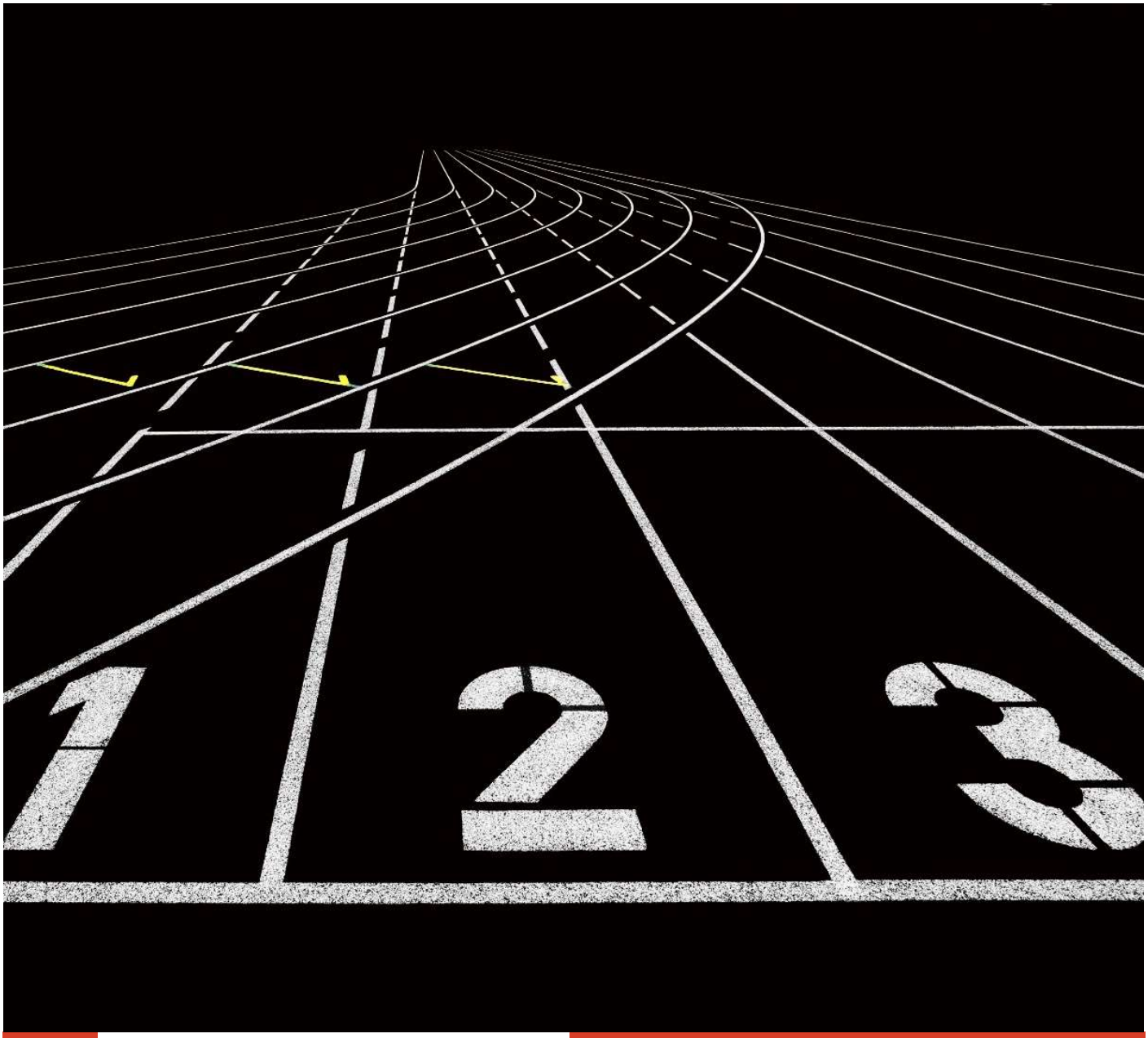
Q2.8.1 21

Q2.8.2 22

ABC asphalt shingles 23

Q3.1 23

Q3.2 24



WHOLESALE CUSTOMERS ANALYSIS

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).

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 is concerned with Data Summarization Graphs/Charts and tables. The methods of descriptive statics include Distribution, which deals with each value's frequency, Measures of Central Tendency and Measures of variability. The most widely used measures of central tendency is Arithmetic Mean, Median, and Mode.

Mean is defined as the arithmetic average of all observations in the data set.
Median is defined as the middle value in the data set arranged in ascending or descending order.
Mode is defined as the most frequently occurring value in the distribution; it has the largest frequency**.**

Measures of Dispersion include Range, IQR, Standard Deviation

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.
Inter-Quartile Range (IQR) is computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. IQR is less affected by outliers.
Standard deviation is the square root of variance in simple words

The table below shows the description of the Wholesale customer dataset:

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

In the table below we can see some sample records which has 2 categorical variable and 6 numerical variables. 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).

	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

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
435	436	Hotel	Other	29703	12051	16027	13135	182	2204
436	437	Hotel	Other	39228	1431	764	4510	93	2346
437	438	Retail	Other	14531	15488	30243	437	14841	1867
438	439	Hotel	Other	10290	1981	2232	1038	168	2125
439	440	Hotel	Other	2787	1698	2510	65	477	52

```
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)
memory usage: 31.1+ KB
```

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	1264414	1521743	2317845	234671	1032270	248988	6619931

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	854833	422454	570037	231026	204136	104327	2386813
Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599

The Region that has spent the most is Other(10677599) and the region that has spent the least is Oporto(1555088).

The Channel that has spent the most is Hotel(7999569) and the channel that has spent the least is Retail(6619931).

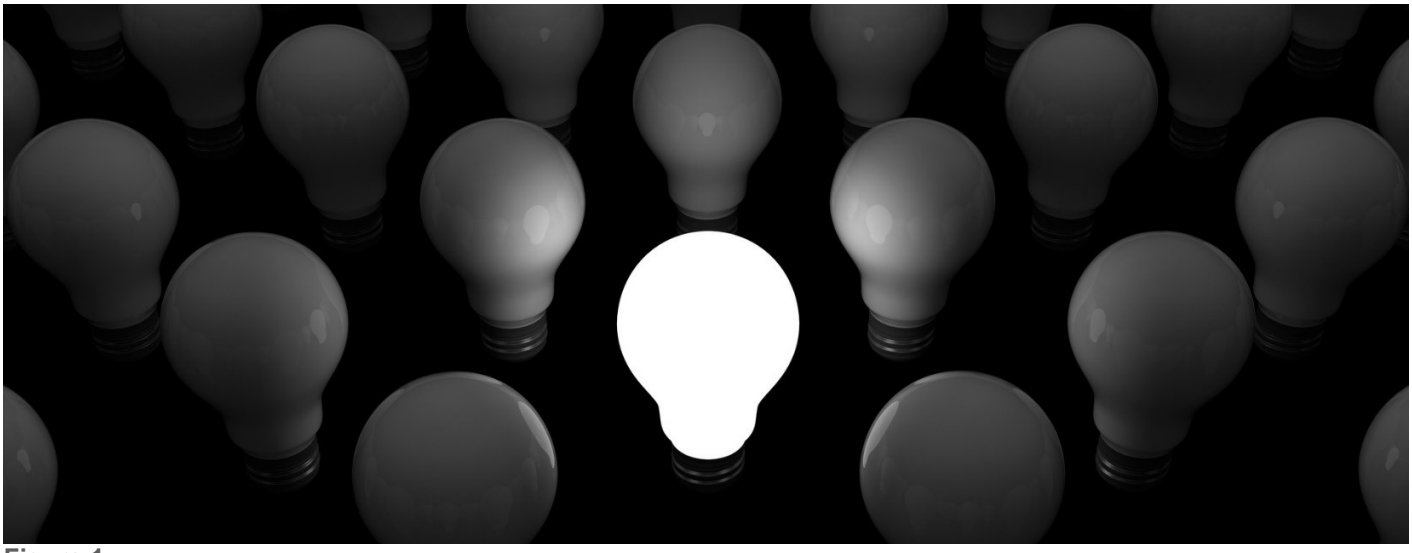


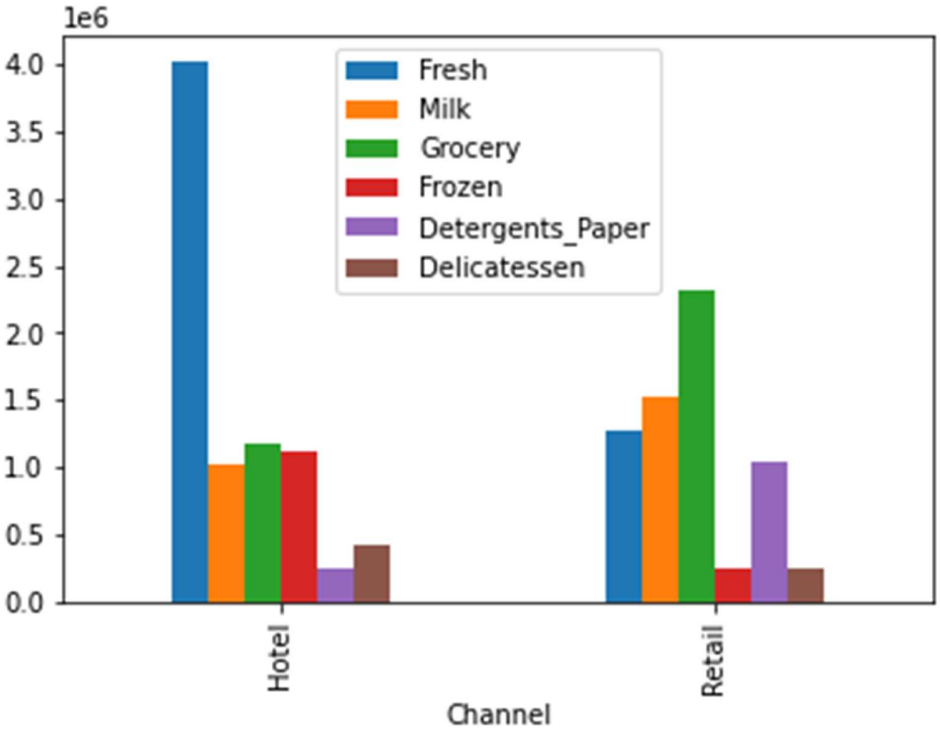
Figure 1

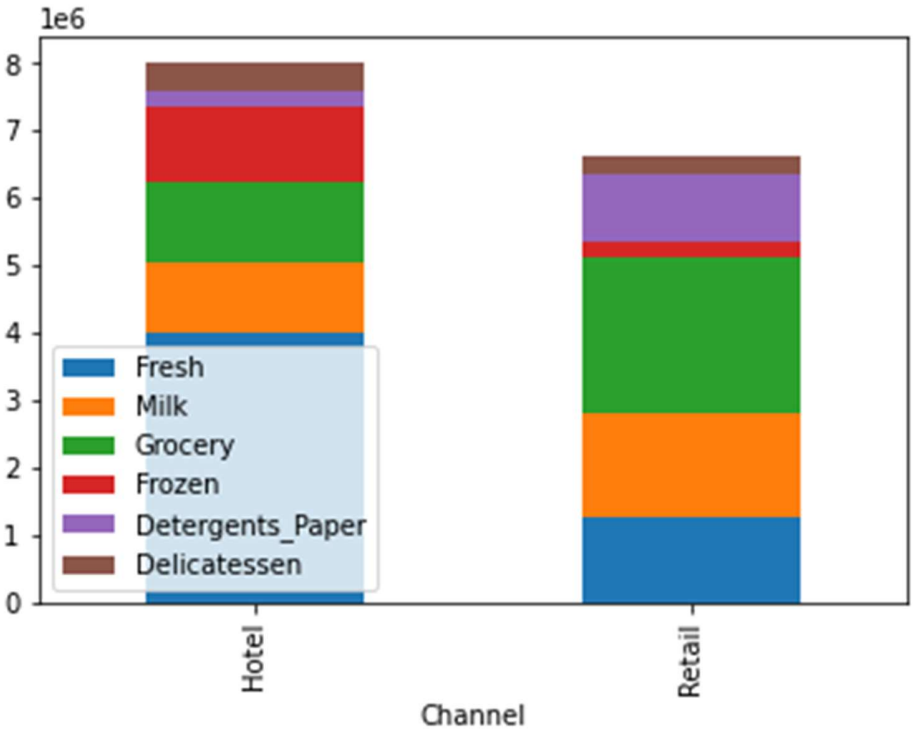
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

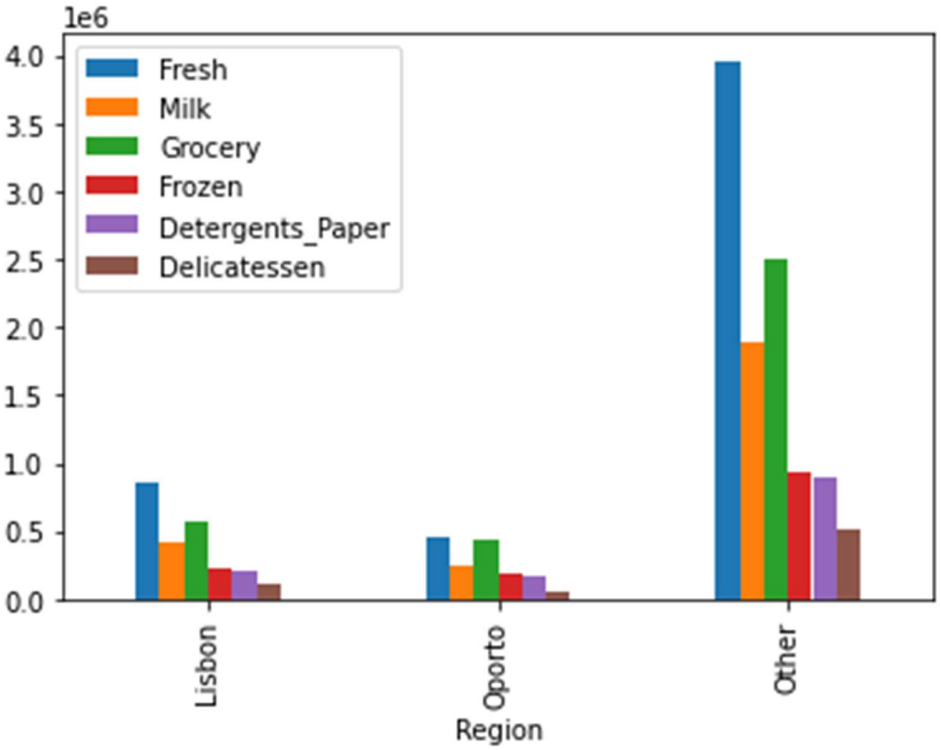
When we sum up the spending across each channel and region, we get the total spending across each channel and region in the following table. the 6 different varieties of items which include Fresh, Milk, grocery, frozen, detergent paper, delicatessen spending can be further summarized in the bar graph

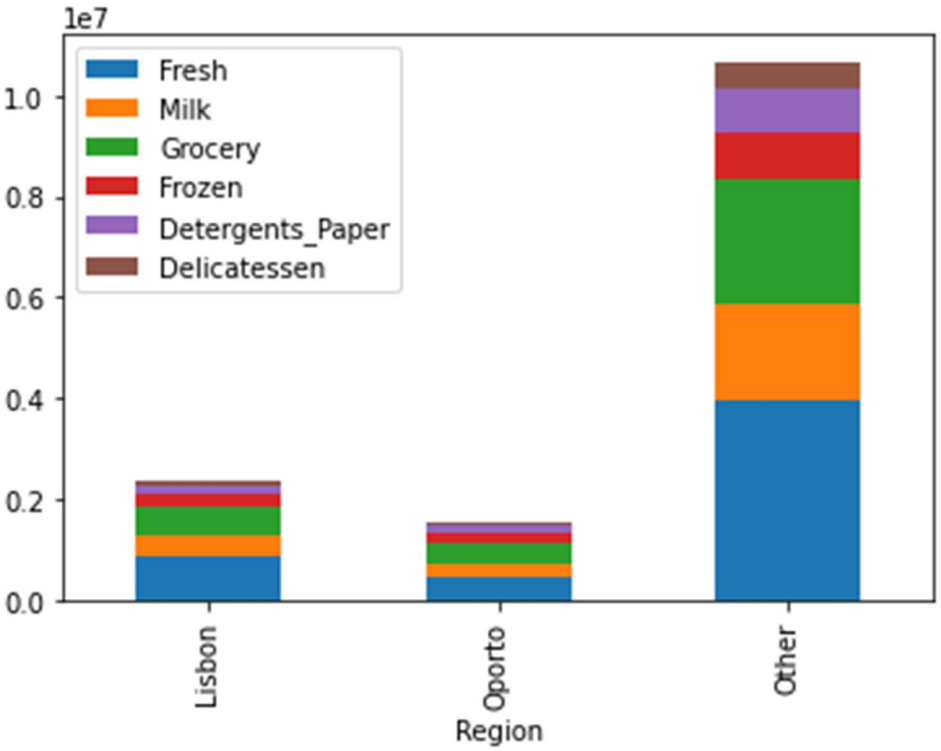
Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	1264414	1521743	2317845	234671	1032270	248988	6619931

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	854833	422454	570037	231026	204136	104327	2386813
Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599


1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen
Hotel
Retail
Channel


1e6
8
7
6
5
4
3
2
1
0
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen
Hotel
Retail
Channel





From the above graph, we can see that at Lisbon most spent product are Fresh products and the least spent product is Delicatessen. At Oporto, the most spent product are Fresh products and least spent products are Delicatessen. In other category, the most spent product are Fresh products and least spent product are Delicatessen

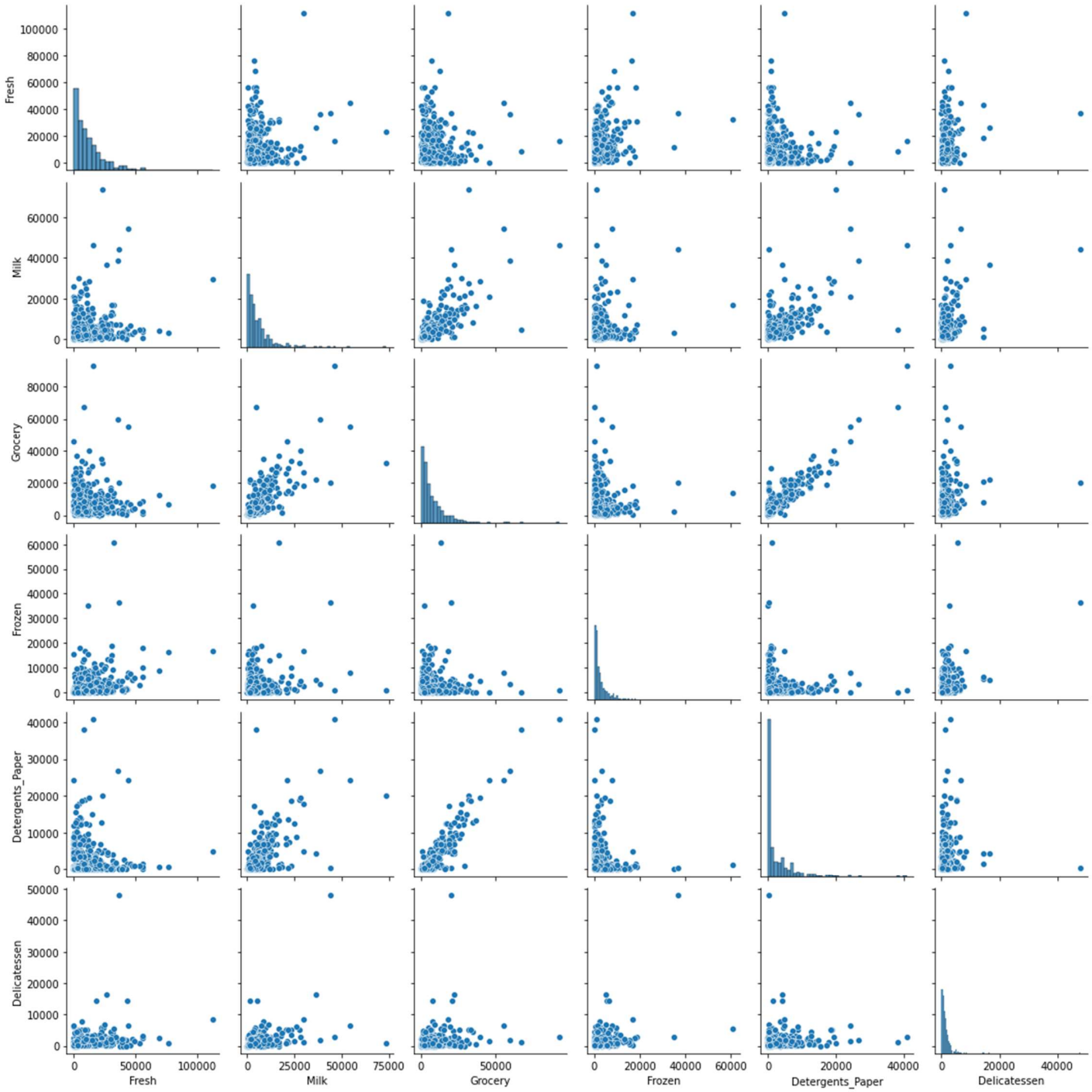
The above graph clearly shows that the most spent product in retail category is Grocery products and least spent product in retail category is the Frozen food products. In Hotel category the most spent product is the Fresh products and least spent product is the Detergents paper

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?

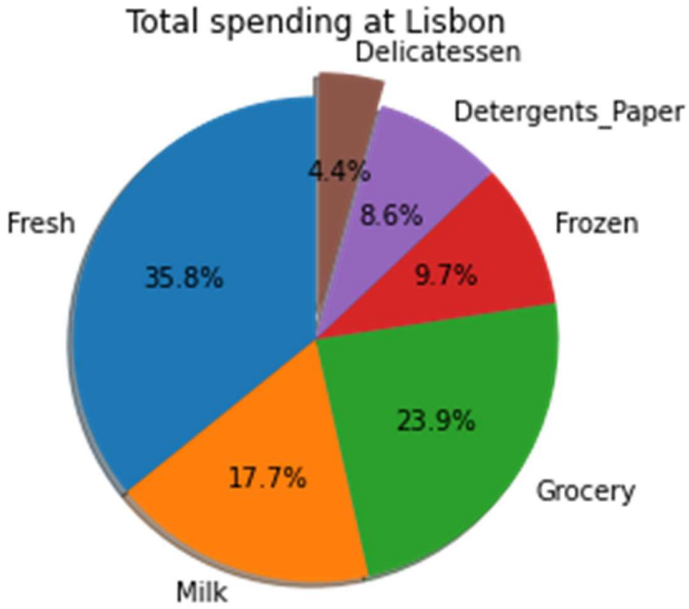
The common descriptive measures of variability are the range, IQR, variance, and standard deviation. To check the inconsistent behavior of an item we can calculate the coefficient of variation of each of the variable. The following pie chart explains how each of the item has performed across the 3 different locations Lisbon, Oporto and other against both retail and hotel category.

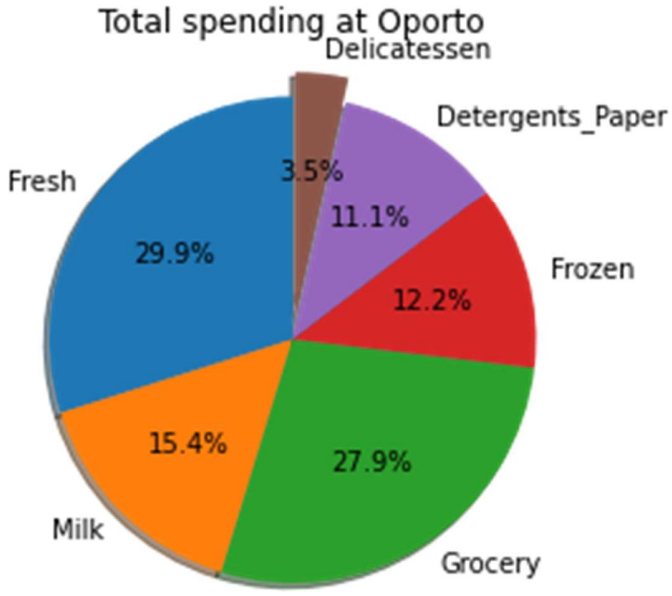
	coeff_var
Fresh	105.259690
Milk	127.113364
Grocery	119.332707
Frozen	157.670810
Detergents_Paper	165.061542
Delicatessen	184.420171

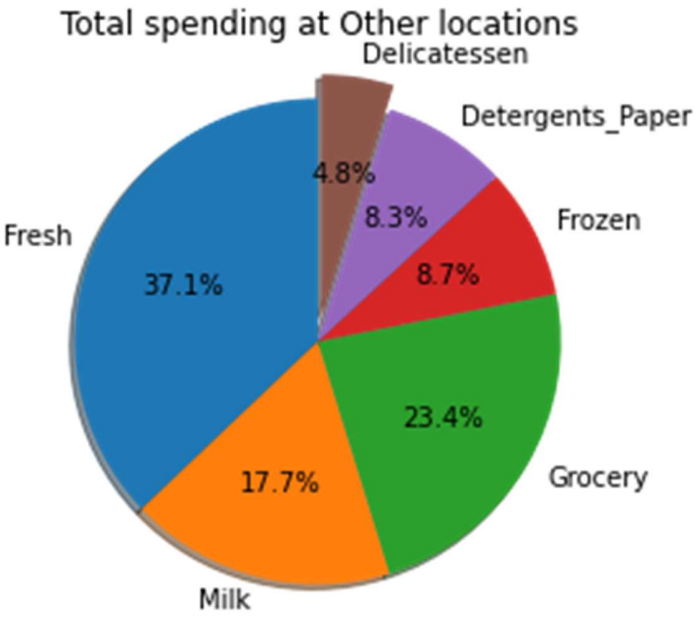
This table shows that coefficient of variance of Fresh products is 105.25% while that of Delicatessen is 184.42%. Therefore, Fresh products show the most inconsistent behavior and Delicatessen shows the least inconsistent behavior

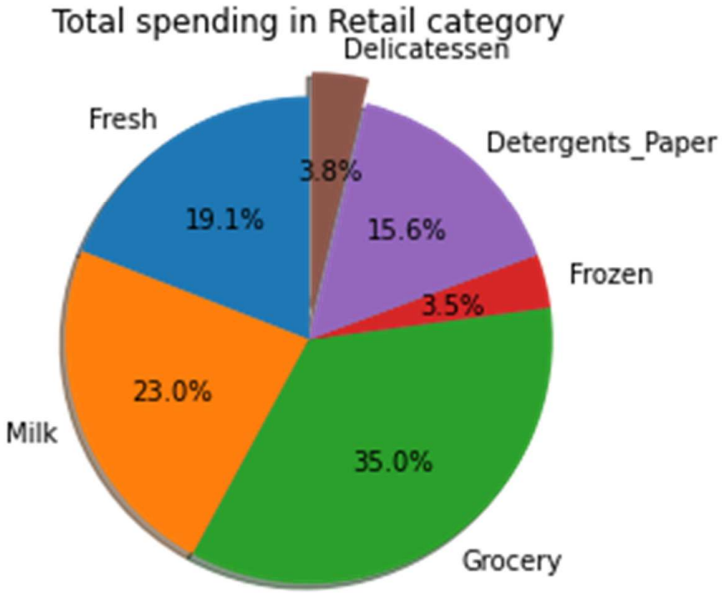


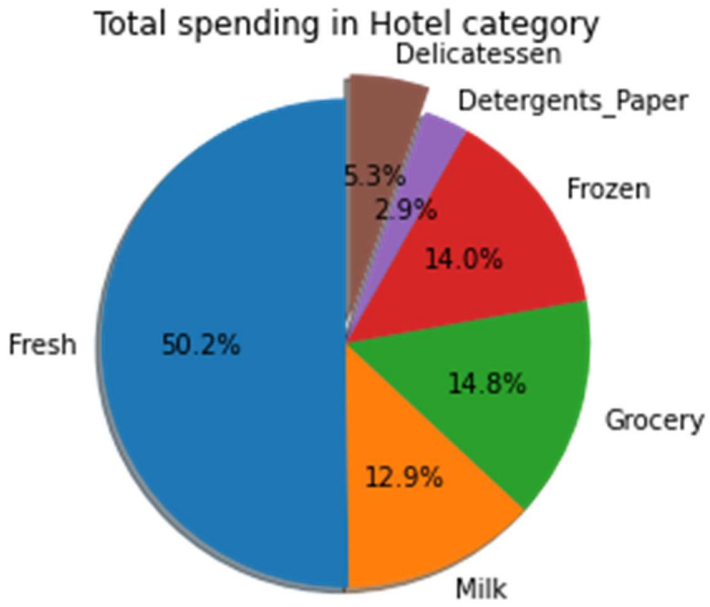
This pair plot helps us to understand the relationship between the 6 food items.


Total spending at Lisbon
Delicatessen
Detergents_Paper
Fresh
Frozen
4.4%
8.6%
35.8%
9.7%
23.9%
17.7%
Grocery
Milk


Total spending at Oporto
Delicatessen
Detergents_Paper
Fresh
3.5%
11.1%
29.9%
Frozen
12.2%
15.4%
27.9%
Milk
Grocery

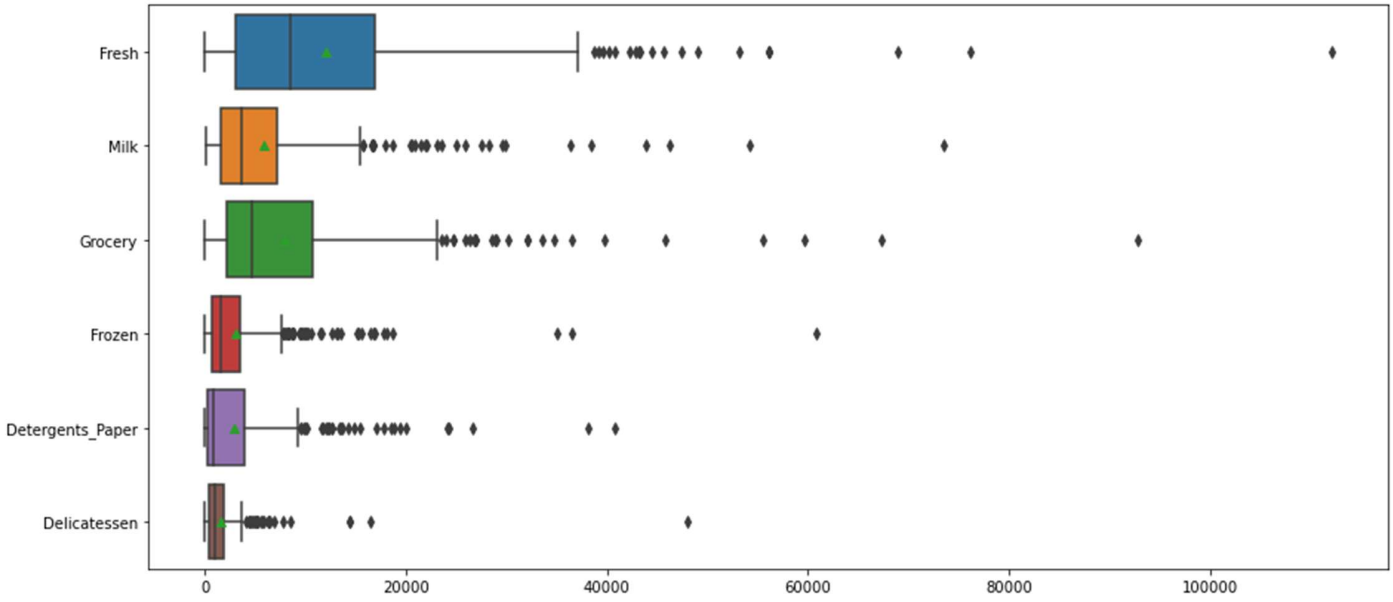

Total spending at Other locations
Delicatessen
Detergents_Paper
Frozen
Fresh
4.8%
8.3%
8.7%
37.1%
23.4%
17.7%
Grocery
Milk


Total spending in Retail category
Delicatessen
Fresh
Detergents_Paper
3.8%
19.1%
15.6%
Frozen
3.5%
23.0%
Milk
35.0%
Grocery



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

From this Boxplot, we can clearly see that all the 6 items have outliers



Outliers are observations in a dataset that don't fit in some way. Perhaps the most common or familiar type of outlier is the observations that are far from the rest of the observations or the center of mass of observations. Outliers can skew statistical measures and data distributions, providing a misleading representation of the underlying data and relationships. Removing outliers from data prior to modeling can result in a better fit of the data and, in turn, more skillful predictions.

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

From this analysis we can conclude that:
(1) When we calculate total, the business spends the most on fresh products across different channels and different regions, so the company needs to ensure that it is driving the most profit from this food item.
(2) Since the Delicatessen show the least inconsistent behavior, the business should invest more in this food item because it is less risky
(3)Fresh products require more spending, to cut cost the wholesale distributor can concentrate more on other food items like Milk, Grocery, Frozen, Detergents paper and Delicatessen

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).

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

From the 62 students total, 33 are female and 29 are male.
The following table shows the number of males and females against each major.

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

	Other	4
	Retailing/Marketing	5
	Undecided	3

2.1.2. Gender and Grad Intention

The following table shows the number of male and female against whether they intent to graduate or no along with some who are undecided

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

2.1.3. Gender and Employment

The following table displays the employment status with the number of males and females for each type of employment.

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

2.1.4. Gender and Computer

The following table show the number of male and female students who use tablet, laptop or desktop.

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

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

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

Probability of male student = number of male student/total number of students=29/62 = 0.46774193548387094

The probability that a randomly selected CMSU student will be male is 0.46774193548387094

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

Probability of female student = number of female student /total number of students=33/62 = 0.532258064516129

The probability that a randomly selected CMSU student will be female is 0.532258064516129

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

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

The following table shows the conditional probability of different majors among the male students in CMSU which is calculated by number of male students in accounting, CIS, economics/finance, international business, management, other, Retailing/Marketing ,undecided/total number of male students

	male_prob	
Gender		
Major		
Accounting	**4**	**0.137931**
CIS	**1**	**0.034483**
Economics/Finance	**4**	**0.137931**
International Business	**2**	**0.068966**
Management	**6**	**0.206897**
Other	**4**	**0.137931**
Retailing/Marketing	**5**	**0.172414**
Undecided	**3**	**0.103448**

2.3.2 Find the conditional probability of different majors among the female students of CMSU.
The following table show the conditional probability of different majors among the female students of CMSU which is calculated by number of females in accounting, CIS, economics/finance, international business, management, other/total number of female students

	Gender	female_prob
Major		
Accounting	**3**	**0.090909**
CIS	**3**	**0.090909**
Economics/Finance	**7**	**0.212121**
International Business	**4**	**0.121212**
Management	**4**	**0.121212**

Other	**3**	**0.090909**
Retailing/Marketing	**9**	**0.272727**

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

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 = number of male students who intends to graduate/total number of students=17/62

Probability that a randomly chosen student is a male and intends to graduate is 0.27419354838709675

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 = number of female students without laptop/total number of students=4/62

Probability that a randomly chosen student is a female and does not have a laptop is 0.06451612903225806

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

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 = number of male students or students who have full time employment/total number of students=32/62

```
Probability that a randomly chosen student is a male or has full time empl
oyment is 0.5161290322580645
```

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= number of female student from international business or management/total number of female students=8/33

Probability that a randomly chosen student is female and has Major in Management or International Business 0.24242424242424243

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?

	yes	no
male	17	3
female	11	9

Two events A and B are said to be independent if the fact that one event has occurred does not affect the probability that the other event will occur. We can see out of 29 male, 17 intent to graduate and out of 33 female only 11 intent to graduate.

Events A and B are independent if the equation $P(A \cap B) = P(A) \cdot P(B)$ holds true.

P(female) = 33/62 = 0.532258064516129
P(intent to graduate) =28/62 = 0.45161290322580644
P(female & intent to graduate) =11/62 = 0.1774193548387097

P(female) * P(intent to graduate) =0.532258064516129*0.45161290322580644
=0.24037460978147762

P(female) * P(intent to graduate) not equal to P(female & intent to graduate)

Hence, the graduate intention and being female are dependent events

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

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

Probability that a randomly chosen student's GPA is less than 3 = number of students with GPA less than 3/total number of students=17/62

Probability that a randomly chosen student's GPA is less than 3 is 0.27419354838709675

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.

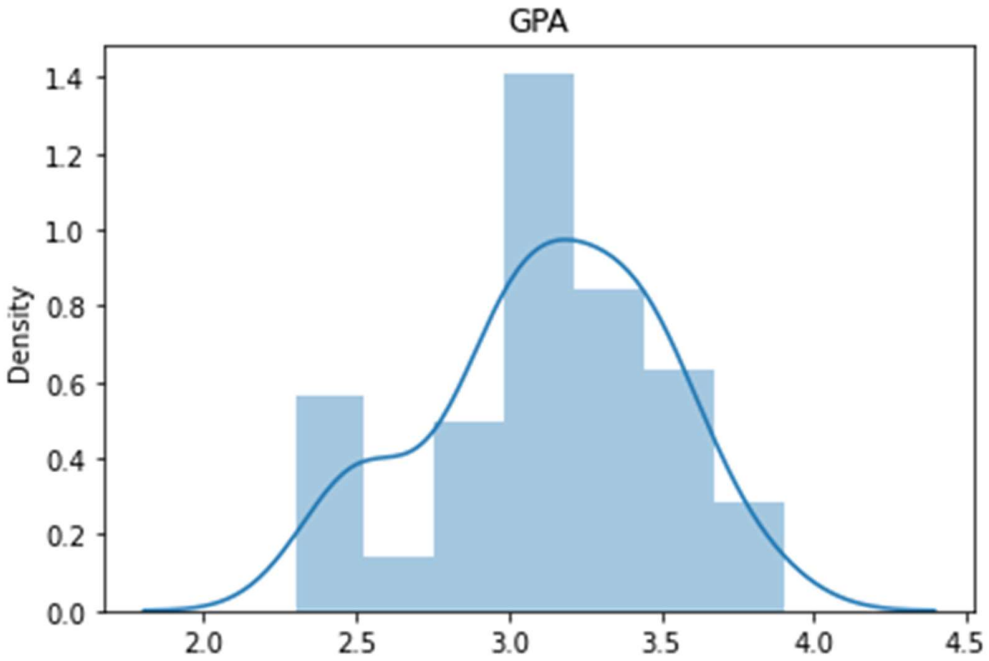
Probability that a randomly selected male earns more than 50 = number male students who earns 50 or more/total number of male=14/29

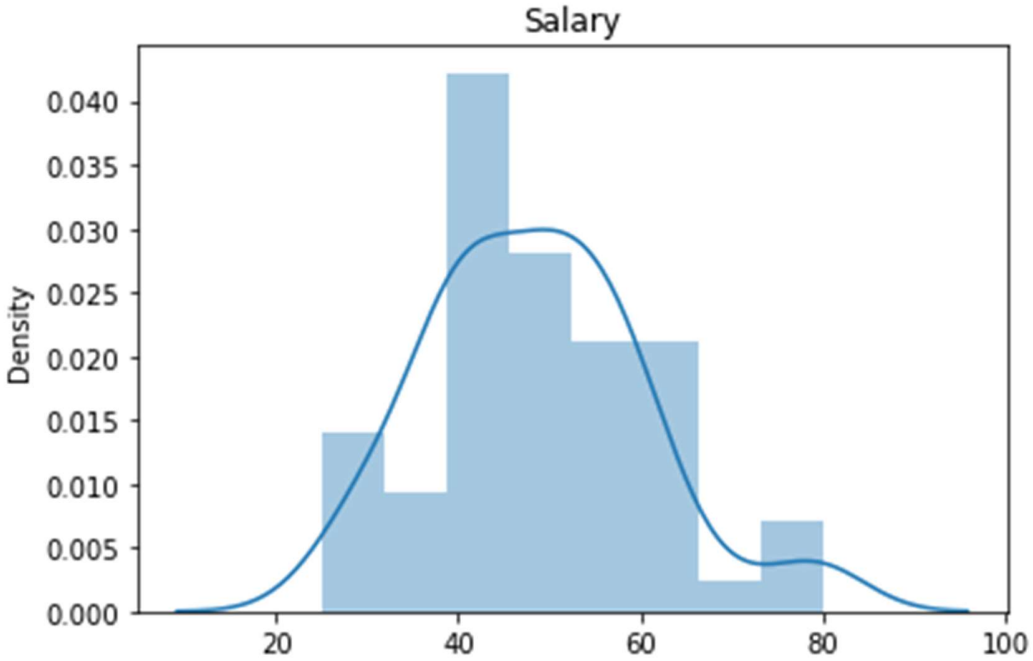
Probability that a randomly selected male earns more than 50 is 0.4827586206896552

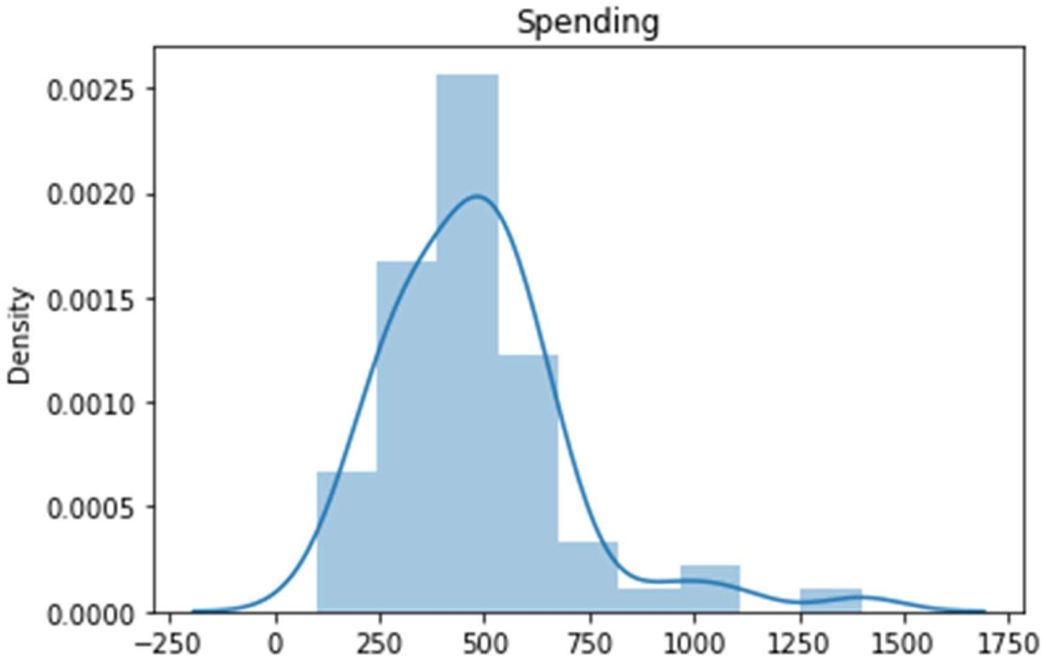
Probability that a randomly selected Female earns more than 50 = number of females earning 50 or more/number of females=18/33

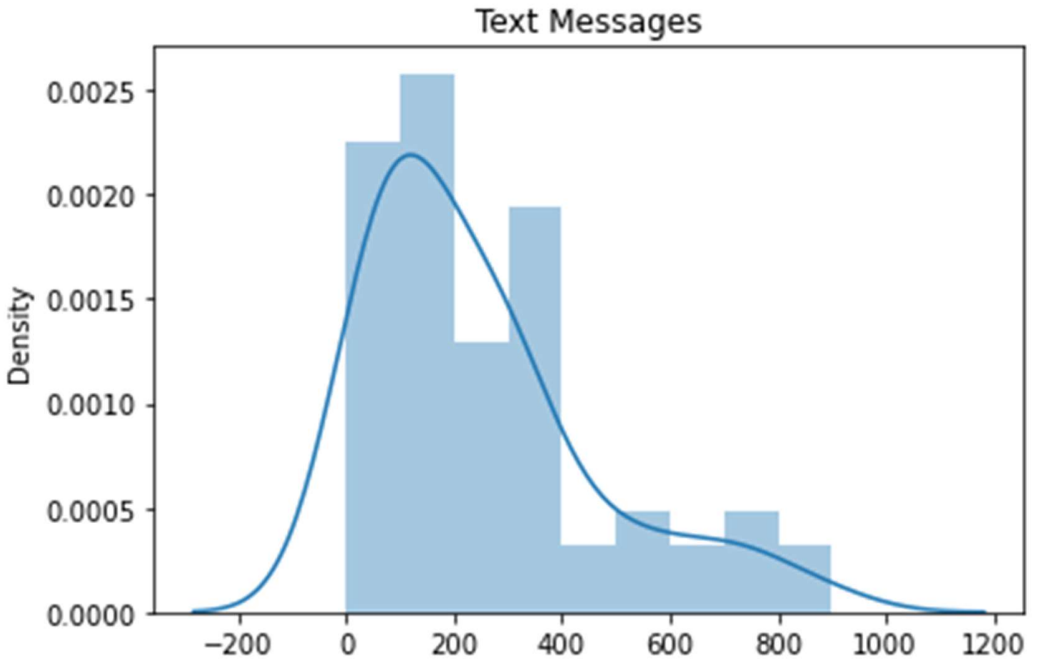
Probability that a randomly selected Female earns more than 50 is 0.5454545454545454

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.









And to confirm whether these four data sets are following normal distribution or not, we done the Shapiro–Wilk test and the output from Python we got –
ShapiroResult for GPA(statistics=0.994, p=0.987)
p-value is more than 0.05
ShapiroResult for Salary(statistics=0.971, p=0.147)
p-value is more than 0.05
ShapiroResult for Spending(statistics=0.984, p=0.589)
p-value is more than 0.05
ShapiroResult for Text Messages(statistics=0.980, p=0.408)
p-value is more than 0.05

By these details we confirm that out of the given four data sets ‘GPA’ ,‘Salary’ , Spending’ and ‘Text Messages’ are following normal distribution.

2.8.2 Write a note summarizing your conclusions for this whole Problem 2.

From this analysis, we can conclude that the sample survey conducted for the students from central Missouri state university shows that there are multiple factors that affect the graduation of a student. The survey conducted by Student News Service at Clear Mountain State University (CMSU) has information about what major the undergrad students are pursuing, whether they intent to graduate, what is their GPA, nature of their employment and their salary, social networking, spending, satisfaction, computer and text messages. Using our analysis, we have constructed contingency tables and calculated probabilities between these variables. We can conclude that in order to help students graduate and find suitable employment the university can work on improving the infrastructure by providing easy access to computers and conducting social networking events. The probabilities of male students graduating is more than that of female students, so female students need more support and choice of major.

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 is calculated. The company would like to show that the mean moisture content is less than 0.35 pound 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.

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.

One sample t testt statistic: -1.4735046253382782

p value: 0.07477633144907513

Since pvalue > 0.05, do not reject H0 .

There is not enough evidence to conclude that the mean moisture content for Sample A shingles is less than 0.35 pounds per 100 square feet. p-value = 0.0748. If the population mean moisture content is in fact no less than 0.35 pounds per 100 square feet, the probability of observing a sample of 36 shingles that will result in a sample mean moisture content of 0.3167 pounds per 100 square feet or less is .0748.

```
t_statistic, p_value = ttest_1samp(df.B, 0.35,nan_policy='omit' )
print('One sample t test \nt statistic: {0} p value: {1} '.format(t_statistic, p_value/2))
```

One sample t testt statistic: -3.1003313069986995 p value: 0.0020904774003191826

Since pvalue < 0.05, reject H0 . There is enough evidence to conclude that the mean moisture content for Sample B shingles is not less than 0.35 pounds per 100 square feet. p-value = 0.0021. If the population mean moisture content is in fact no less than 0.35pounds per 100 square feet, the probability of observing a sample of 31 shingles that will result in a sample mean moisture content of 0.2735 pounds per 100 square feet or less is .0021.

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?

H0 : μ(A)= μ(B)

Ha : μ(A)!= μ(B)

α = 0.05

t_statistic=1.29

pvalue=0.202

As the pvalue > α , do not reject H0;

and we can say that population mean for shingles A and B are equal Test Assumptions When running a two-sample t-test, the basic assumptions are that the distributions of the two populations are normal, and that the variances of the two distributions are the same. If those assumptions are not likely to be met, another testing procedure could be use**.**

Problem 1

Wholesale Customers Analysis

Problem 2

Clear Mountain State University (CMSU)

Problem 3

ABC asphalt shingles

Thank You !