BUSINESS REPORT

Problem 1

1.1.

Ans) Looking at the first 5 entries of the data to get a gist of the dataset using head() function from pandas library in python.

	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

The following information can be inferred from the dataset using info() function from pandas library in python.

```
<class 'pandas.core.frame.DataFrame'>
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
dtypes: int64(7), object(2)
memory usage: 31.0+ KB
```

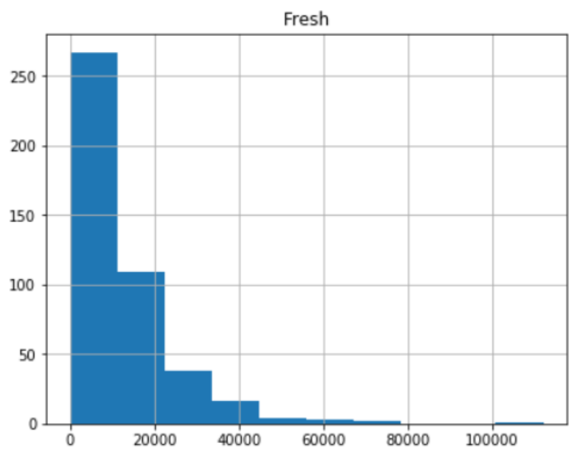
There are **440 entries with 9 columns** of which channel and region are of object data type and others of integer data type.

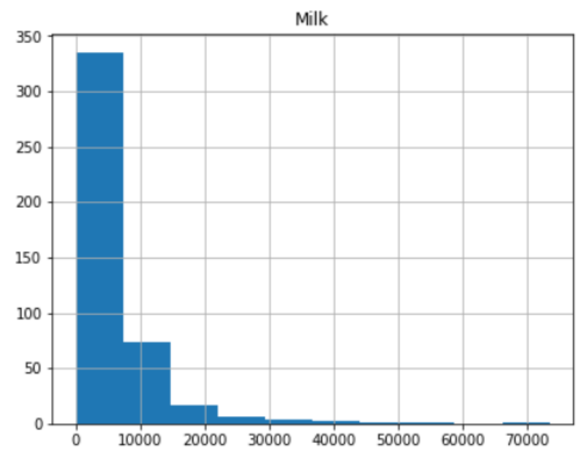
There are **no null values** in the dataset, checked using isnull() function from pandas library in python.

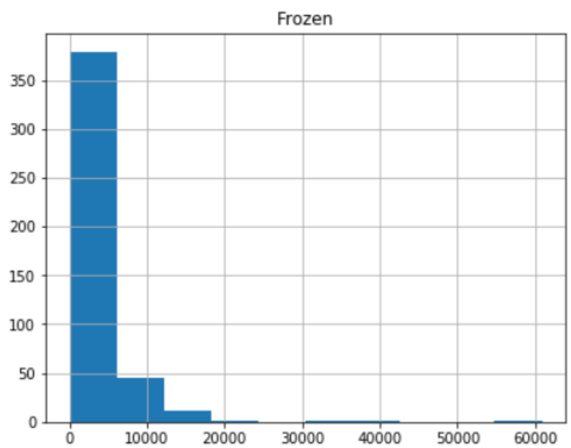
By checking the summary of the data we get the following result, we can get an idea about the 5 point statistics of data, which was checked using describe() function from pandas library in python.

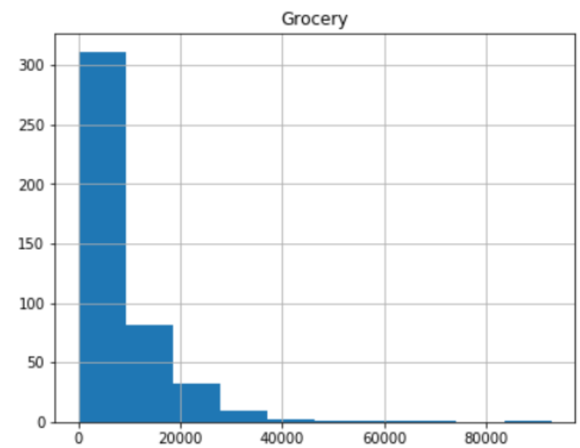
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	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

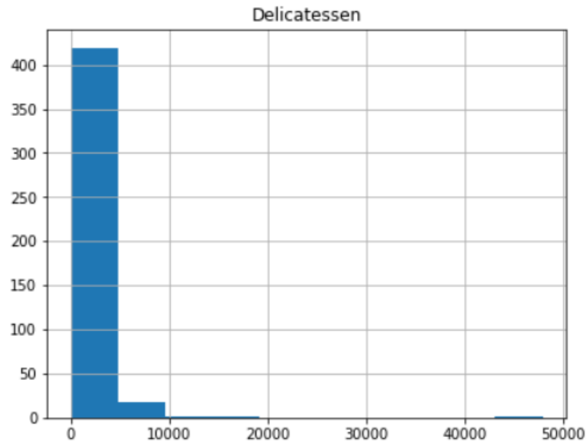
Checking distribution of amount for each variety of products using hist() function from matplotlib library in python.

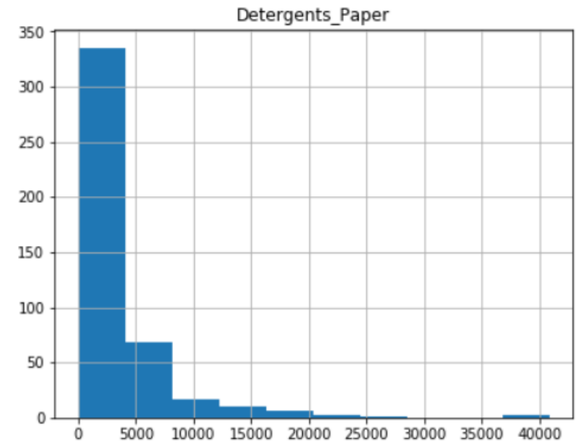






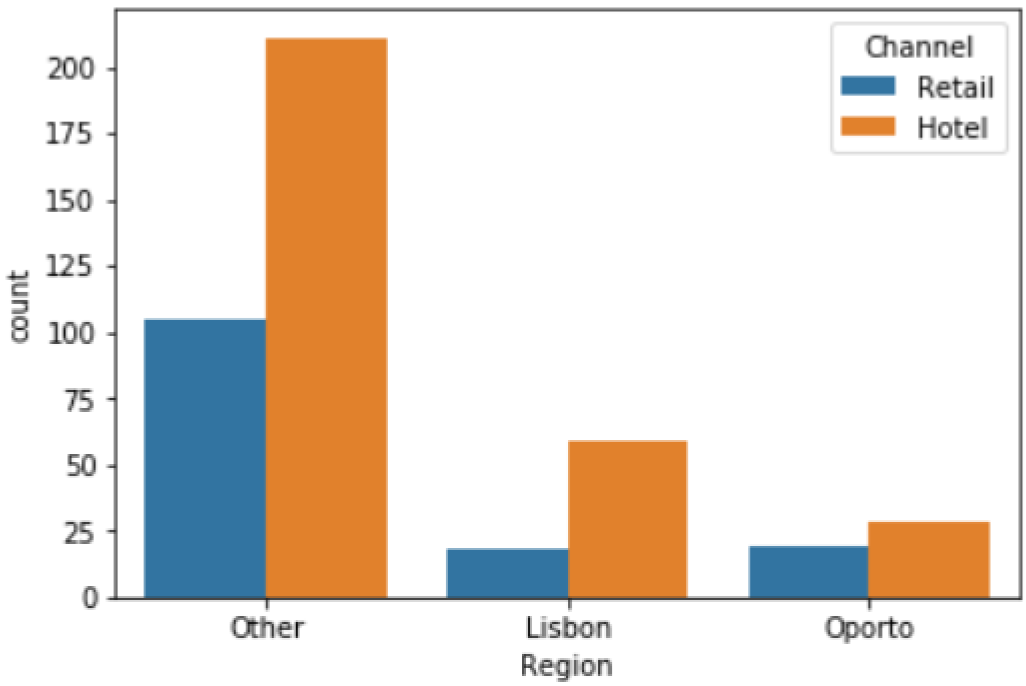






From the above plots we can infer that none of the variables are normally distributed.

The counts of channels for each region using countplot function from seaborn library in python.



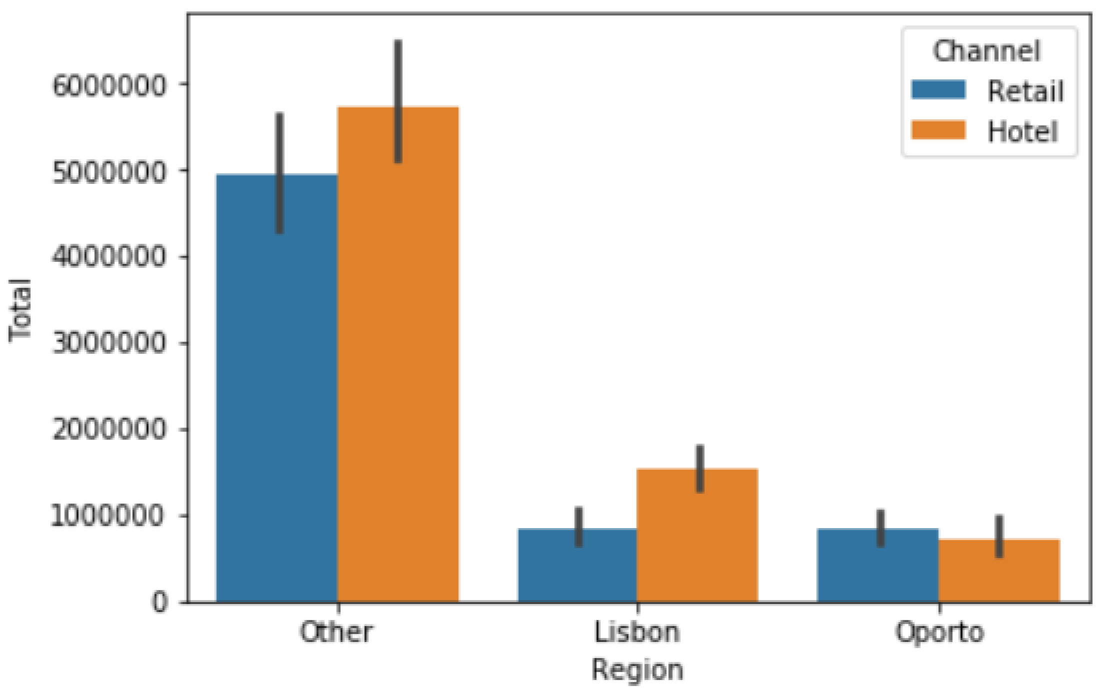
- **Other region and hotel channel** has maximum number of buyer/spender
- **Oporto, Lisbon region and retail channel** seem to have least number of buyer/spender

For ease of calculation, a new column “Total” has been added which will be the sum amount of all 6 varieties of products for a single buyer.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
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

Which Region and which Channel seems to spend more?
Which Region and which Channel seems to spend less?

Ans) Plotting a bar graph of total and regions with channels as hue and estimator as sum, we get



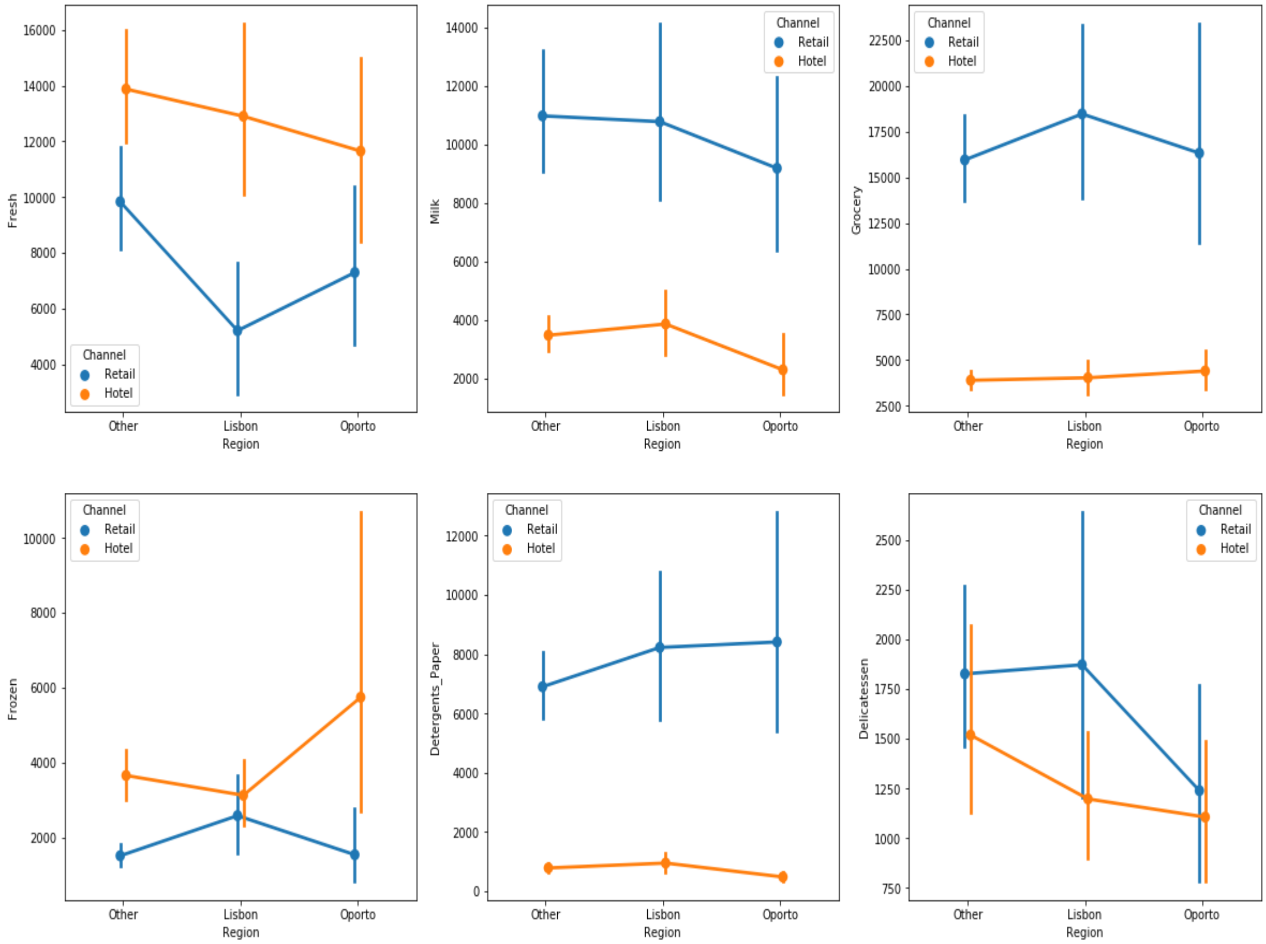
From the above plot, we can make the following observations:

- **Other** region spends the **most**, and **Oporto** the **least**.
- **Hotel** channel spends the **most** and **retail** the **least**.
- **Other region and hotel channel** spends the **most**.
- **Oporto region and hotel channel** spends the **least**.

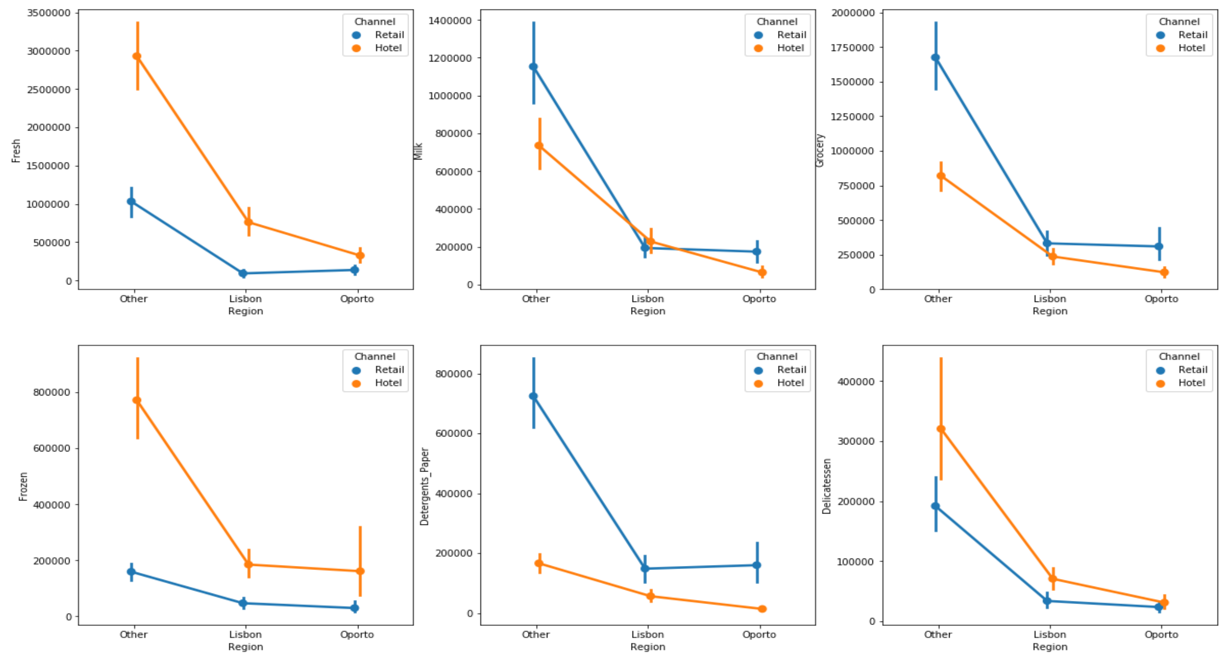
1.2.

Ans)

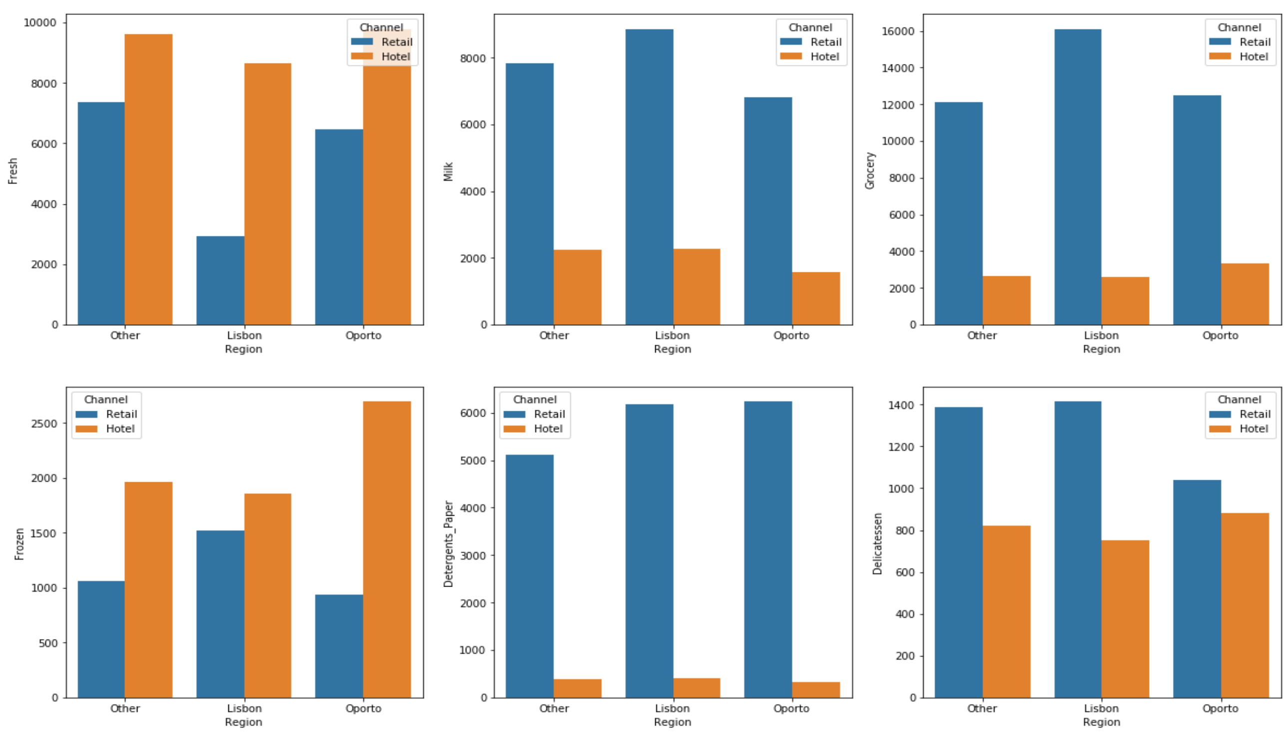
The below plots are point plots with estimator as the **mean** of the total amount spent on the 6 varieties. We can observe that milk, grocery and detergent paper show a similar pattern across all channels with retail channel having more spending.



The next plot is point plots with estimator as the **sum** of the amount spent on the 6 varieties. We can observe that milk, grocery and detergent paper show a similar pattern with retail channel having more spending. Here we also observe that patterns of fresh, frozen and delicatessen look quite similar too with hotel channel having more spending. Other region has the most and Oporto region seems to have the least annual spending across all varieties.



The below plot is of **median** values across region and channel of the varieties spending amount. The values of milk and grocery across the hotel channel seem to appear the closest to each other. Thus we can say that all varieties don't show a similar behavior, but some varieties do show some similar behavior when different factors are considered.



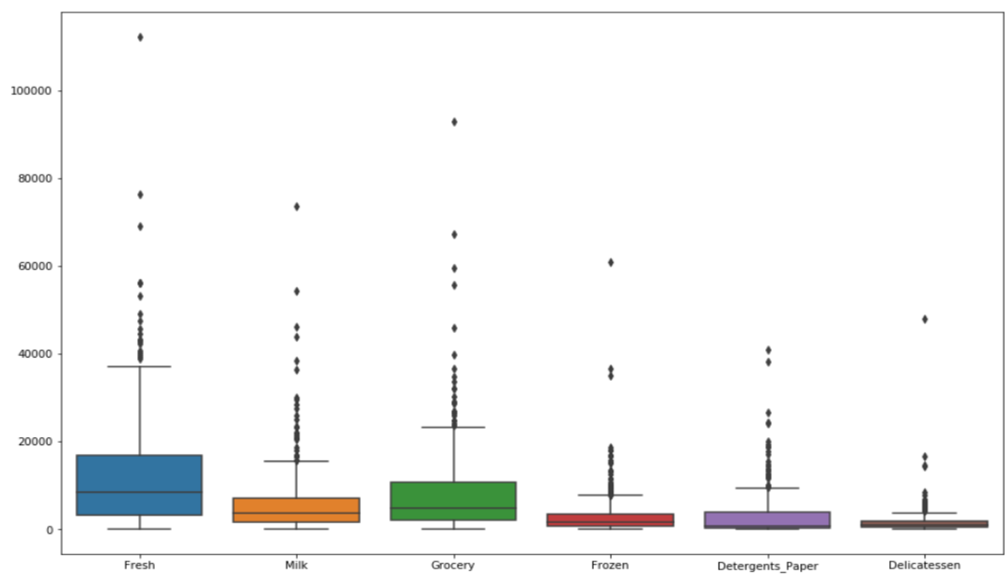
1.3.

	count	mean	std	min	25%	50%	75%	max	cov	IQR	Range
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	1.053918	13806.00	112148.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	1.195174	8502.75	92777.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	1.273299	5657.25	73443.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	1.580332	2812.00	60844.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	1.654647	3665.25	40824.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	1.849407	1412.00	47940.0

When a distribution has lower variability, the values in a dataset are more consistent.

The **range** is the most obvious measure of dispersion and is the difference between the lowest and highest values in a dataset. The **inter-quartile range** is a measure that indicates the extent to which the central 50% of values within the dataset are dispersed. The **standard deviation** is a measure that summarizes the amount by which every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. **If the range, inter quartile range and the standard deviation are higher for a data we can conclude that the spread of the data is higher and hence the data has more variability**. Arranging the 5 point summary of all variables in the descending order as per the measures of variability, we can say that **Fresh** variety has the highest variability and thus has the most inconsistent behavior whereas **Delicatessen** has the least variability and can be considered as the most consistent among all the varieties.

1.4.



From the above plot we can see that each variable has outliers in them.

1.5.

Ans) From the analysis of data, a few recommendations for the wholesale distributor are as follows:

- The distributor should give higher preference to **hotel channel** across all the regions as most spenders/buyers are from this channel
- The distributor should give higher preference to **other region** as most spenders/buyers are from this region and the net spending is also the highest
- **Fresh, milk and grocery** are among the most popular varieties with mean and total spending higher across all regions and channels which implies they are always high in demand and hence they should be stocked well
- **The retail channel in Oporto and Lisbon regions** are performing fairly well across all varieties in general and the distributor can try to target more retail channels from this region
- **Milk, grocery and detergent paper** have the poorest sales among all regions from hotel channel so the distributor can encourage retailers to buy them to increase the net sales.
- **Fresh, frozen and delicatessen** have the poorest sales among all regions from retail channel so the distributor can encourage retailers to buy them to increase the net sales.

Problem 2

Part I

2.1.

Ans) For creating contingency tables, we used the crosstab function from pandas taking x variable as gender to keep it as the row variable

2.1.1. Gender and Major

Major / Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention

Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

2.1.3. Gender and Employment

Employment	Full-Time	Part-Time	Unemployed	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

2.2.

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

Ans)

Total sample size = 62

Total_Male = 29

∴ Probability_male = $\frac{Total\_Male}{Total\ sample\ size}$= $\frac{29}{62}$ = 0.46774193548387094

Probability that a randomly selected CMSU student will be male is 0.4677

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

Ans)

Similarly,

Total_Female = 33

∴ Probbability_Female = $\frac{Total\_Female}{Total\ sample\ size}$= $\frac{33}{62}$ = 0.532258064516129

Probability that a randomly selected CMSU student will be female is 0.5323

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

Ans) Having the below table as reference

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

Applying conditional probability,

P (A/B) = $\frac{P(A\ and\ B)}{P(B)}$

Probability of having **Accounting as major** given the student is **male** =

P $(Accounting/Male)$ = $\frac{P(Accounting\ and\ Male)}{P(Male)} = \frac{4/62}{29/62}$ = 0.1379

Probability of having **CIS as major** given the student is **male** =

P $(CIS/Male)$ = $\frac{P(CIS\ and\ Male)}{P(Male)} = \frac{1/62}{29/62}$ = 0.0345

Probability of having **Economics/Finance as major** given the student is **male** =

P $((Economics/Finance)/Male)$ = $\frac{P(Economics/Finance\ and\ Male)}{P(Male)} = \frac{4/62}{29/62}$ = 0.1379

Probability of having **International Business as major** given the student is **male** =

P $(International\ Business/Male)$ = $\frac{P(International\ Business\ and\ Male)}{P(Male)} = \frac{2/62}{29/62}$ =0.0690

Probability of having **Management as major** given the student is **male** =

P $(Management/Male)$ = $\frac{P(Management\ and\ Male)}{P(Male)} = \frac{6/62}{29/62}$ = 0.2069

Probability of having **Other as major** given the student is **male** =

P $(Other/Male)$ = $\frac{P(Other\ and\ Male)}{P(Male)} = \frac{4/62}{29/62}$ = 0.1379

Probability of having **Retailing/Marketing as major** given the student is **male** =

P $((Retailing/Marketing)/Male)$ = $\frac{P(Retailing/Marketing\ and\ Male)}{P(Male)} = \frac{5/62}{29/62}$ = 0.1724

Probability of having **Undecided as major** given the student is **male** =

P $(Undecided/Male)$ = $\frac{P(Undecided\ and\ Male)}{P(Male)} = \frac{3/62}{29/62}$ = 0.1034

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

Ans) Applying conditional probability and taking the above table as reference,

Probability of having **Accounting as major** given the student is **female** =

P $(Accounting/Female)$ = $\frac{P(Accounting \ and \ Female)}{P(Female)}$ = $\frac{3/62}{33/62}$ = 0.0909

Probability of having **CIS as major** given the student is **female** =

P $(CIS/Female)$ = $\frac{P(CIS \ and \ Female)}{P(Female)}$ = $\frac{3/62}{33/62}$ = 0.0909

Probability of having **Economics/Finance as major** given the student is **female** =

P $((Economics/Finance)/Female)$ = $\frac{P(Economics/Finance \ and \ Female)}{P(Female)}$ = $\frac{7/62}{33/62}$ = 0.2121

Probability of having **International Business as major** given the student is **female** =

P $(International \ Business/Female)$ = $\frac{P(International \ Business \ and \ Female)}{P(Female)}$ = $\frac{4/62}{33/62}$ = 0.1212

Probability of having **Management as major** given the student is **female** =

P $(Management/Female)$ = $\frac{P(Management \ and \ Female)}{P(Female)}$ = $\frac{4/62}{33/62}$ = 0.1212

Probability of having **Other as major** given the student is **female** =

P $(Other/Female)$ = $\frac{P(Other \ and \ Female)}{P(Female)}$ = $\frac{3/62}{33/62}$ = 0.0909

Probability of having **Retailing/Marketing as major** given the student is **female** =

P $((Retailing/Marketing)/Female)$ = $\frac{P(Retailing/Marketing \ and \ Female)}{P(Female)}$ = $\frac{9/62}{33/62}$ = 0.2727

Probability of having **Undecided as major** given the student is **female** =

P $(Undecided/Female)$ = $\frac{P(Undecided\ and\ Female)}{P(Female)}$ = $\frac{0/62}{33/62}$ = 0.0000

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.

Ans) Having the below table as reference

Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

Applying conditional probability,

Probability of having **intent as No** given the student is **male** =

P $(No/Male)$ = $\frac{P(No\ and\ Male)}{P(Male)}$ = $\frac{3/62}{29/62}$ = 0.1034

Probability of having **intent as Undecided** given the student is **male** =

P $(Undecided/Male)$ = $\frac{P(Undecided\ and\ Male)}{P(Male)}$ = $\frac{9/62}{29/62}$ = 0.3103

Probability of having **intent as Yes** given the student is **male** =

P $(Yes/Male)$ = $\frac{P(Yes\ and\ Male)}{P(Male)}$ = $\frac{17/62}{29/62}$ = 0.5862

Find the conditional probability of intent to graduate, given that the student is a female.

Ans) Applying conditional probability and taking the above table as reference,

Probability of having **intent as No** given the student is **female** =

P $(No/Female)$ = $\frac{P(No\ and\ Female)}{P(Female)}$ = $\frac{9/_{62}}{33/_{62}}$ = 0.1034

Probability of having **intent as Undecided** given the student is **female** =

P $(Undecided/Female)$ = $\frac{P(Undecided\ and\ Female)}{P(Female)}$ = $\frac{13/_{62}}{33/_{62}}$ = 0.3939

Probability of having **intent as Yes** given the student is **female** =

P $(Yes/Female)$ = $\frac{P(Yes\ and\ Female)}{P(Female)}$ = $\frac{11/_{62}}{33/_{62}}$ = 0.3333

2.2.4.

Ans) Having below table as reference,

Employment	Full-Time	Part-Time	Unemployed	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

Applying conditional probability,

For **males**,

Probability of having **employment status as Full-Time** given the student is **male** =

P $(FullTime/Male)$ = $\frac{P(FullTime\ and\ Male)}{P(Male)}$ = $\frac{7/_{62}}{29/_{62}}$ = 0.2414

Probability of having **employment status as Part-Time** given the student is **male** =

P $(PartTime/Male) = \dfrac{P(PartTime\ and\ Male)}{P(Male)} = \dfrac{19/62}{29/62}$ = 0.6552

Probability of having **employment status as Unemployed** given the student is **male** =

P $(Unemployed/Male) = \dfrac{P(Unemployed\ and\ Male)}{P(Male)} = \dfrac{3/62}{29/62}$ = 0.1034

For **females**,

Probability of having **employment status as Full-Time** given the student is **female** =

P $(FullTime/Female) = \dfrac{P(FullTime\ and\ Female)}{P(Female)} = \dfrac{3/62}{33/62}$ = 0.0909

Probability of having **employment status as Part-Time** given the student is **female** =

P $(PartTime/Female) = \dfrac{P(PartTime\ and\ Female)}{P(Female)} = \dfrac{24/62}{33/62}$ = 0.7273

Probability of having **employment status as Unemployed** given the student is **female** =

P $(Unemployed/Female) = \dfrac{P(Unemployed\ and\ Female)}{P(Female)} = \dfrac{6/62}{33/62}$ = 0.1818

2.2.5.

Ans) Having the below table as reference,

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

Applying conditional probability,

Probability of having **Laptop as preference** given the student is **male** =

P $(Laptop/Male)$ = $\frac{P(Laptop\ and\ Male)}{P(Male)}$ = $\frac{26/_{62}}{29/_{62}}$ = 0.8966

Probability of having **Laptop as preference** given the student is **female** =

P $(Laptop/Female)$ = $\frac{P(Laptop\ and\ Female)}{P(Female)}$ = $\frac{29/_{62}}{33/_{62}}$ = 0.8788

2.3.

Ans) Two events, A and B, are said to be independent if P (A/B) = P (A) and P (B/A) = P (B). If the probabilities are significantly different, then we conclude the events are not independent.

Considering independence of **major and gender**, by listing out the probabilities together, with and without gender consideration

For **Accounting**:

P(Accounting/male) : P(Accounting/female) : P(Accounting) = 0.1379 : 0.0909 : 0.1129

For **CIS**:

P(CIS/male) : P(CIS/female) : P(CIS) = 0.0345 : 0.0909 : 0.0645

For **Economics/Finance**:

P(EconomicsORFinance/male) : P(EconomicsORFinance/female) : P(EconomicsORFinance) = 0.1379 : 0.2121 : 0.1774

For **International Business:**

P(InternationalBusiness/male) : P(InternationalBusiness/female) : P(InternationalBusiness) = 0.0690 : 0.1212 : 0.0968

For **Management**:

P(Management/male) : P(Management/female) : P(Management) = 0.2069 : 0.1212 : 0.1613

For **Other**:

P(Other/male) : P(Other/female) : P(Other) = 0.1379 : 0.0909 : 0.1129

For **Retailing/Marketing**:

P(RetailingORMarketing/male) : P(RetailingORMarketing/female) : P(RetailingORMarketing) = 0.1724 : 0.2727 : 0.2258

For **Undecided**:

P(Undecided/male) : P(Undecided/female) : P(Undecided) = 0.1034 : 0.0000 : 0.0484

Considering accounting as example, the probability of selecting accounting is different for males and females and also if considered irrespective of gender. The same can be said for other majors when we look at their probabilities. It implies that **gender is not independent from major** selection and selection of major varies with gender.

Considering independence of **graduation intent and gender**, by listing out the probabilities together, with and without gender consideration

for No

P(No/male) : P(No/female) : P(No) = 0.1034 : 0.2727 : 0.1935

for Undecided

P(Undecided/male) : P(Undecided/female) : P(Undecided) = 0.1034 : 0.0000 : 0.3548

for Yes

P(Yes/male) : P(Yes/female) : P(Yes) = 0.5862 : 0.3333 : 0.4516

We see that probability of a decision varies if we take gender into account. Hence we can say that **grad intention is not independent of gender**.

Considering independence of **employment status and gender**, by listing out the probabilities together, with and without gender consideration

for Full-Time

P(Full-Time/male) : P(Full-Time/female) : P(Full-Time) = 0.2414 : 0.0909 : 0.1613 for Part-Time

P(Part-Time/male) : P(Part-Time/female) : P(Part-Time) = 0.6552 : 0.7273 : 0.6935

for Unemployed

P(Unemployed/male) : P(Unemployed/female) : P(Unemployed) = 0.1034 : 0.1818 : 0.1452

We see that probability of employment status varies if we take gender into account. Hence we can say that **employment status is not independent of gender**.

Considering independence of **employment status and gender**, by listing out the probabilities together, with and without gender consideration

for Desktop

P(Desktop/male) : P(Desktop/female) : P(Desktop) = 0.1034 : 0.0606 : 0.0806

for Laptop

P(Laptop/male) : P(Laptop/female) : P(Laptop) = 0.8966 : 0.8788 : 0.8871

for Tablet

P(Tablet/male) : P(Tablet/female) : P(Tablet) = 0.0000 : 0.0606 : 0.0323

We see that probability of choice of computer varies if we take gender into account. For laptop although we see that the values are pretty close, which means for this particular computer they might be independent but overall we can say that **computer is not independent of gender**.

Part II

2.4.

Ans) We perform the Shapiro-Wilk test to test the normality of variables using Shapiro function from scipy library in python. The null hypothesis is that the data was drawn from a normal distribution. We consider the significance level (α) value as 0.05. The p-value for each variable comes out as follows:

Salary = 0.028000956401228905

Spending = 1.6854661225806922e-05

Text Messages = 4.324040673964191e-06

If p-value < α, then we reject the null hypothesis.

We can see that none of the p values are greater than alpha, which means the variables significantly deviate from normal distribution.

It can also be checked by comparing the mean and median (50%) values

	count	mean	std	min	25%	50%	75%	max
Salary	62.0	48.548387	12.080912	25.0	40.0	50.0	55.0	80.0
Spending	62.0	482.016129	221.953805	100.0	312.5	500.0	600.0	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.0	200.0	300.0	900.0

If mean = median, it follows a normal distribution which is not true for any of the given variables. **None of the variables follow normal distribution.**

Problem 3

3.1.

Ans) H_0: μ_A >= 0.35 pound per 100 square feet

H_A: μ_A < 0.35 pound per 100 square feet

3.2.

Ans) H_0: μ_B >= 0.35 pound per 100 square feet

H_A: μ_B < 0.35 pound per 100 square feet

3.3.

Ans) In the given problem, the sample size is fairly large (greater than 30). We will be conducting a 2 sample t-test since there are two independent samples given whose means are to be checked if they are equal or not, and the population variance is unknown. The assumption that we make is

- The data are continuous (not discrete).

- The data variables follow a normal distribution
- The variances of the two populations are equal
- The two samples are independent. There is no relationship between the individuals in one sample as compared to the other
- Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.
- The scale of measurement applied to the data collected follows a continuous or ordinal scale

Our null and alternate hypothesis is given as follows:

H_0: $\mu_A = \mu_B$

H_A: $\mu_A \neq \mu_B$

We set the significance level (α) as **0.05**

We divide Shingles A & Shingles B in two groups; group A and group B respectively, and remove the null values from the shingle B column.

Using python, we carry out the t-test using the **ttest_ind** function which calculates the T-test for the means of *two independent* samples of scores.

t_statistic = 1.289628271966112

p_value = 0.2017496571835328

Since the p_value > α, we **fail to reject** the null hypothesis.

This implies we can say with **95% confidence** that the population means of the two shingles A & B **are equal.**

3.4.

Ans) The assumption that we make is that the population follows a normal distribution.