SMDM

PROJECT REPORT

10-MAY-2020

PREPARED BY
JOTINDER SINGH MATTA
(PGP DSBA MARCH_20A)

Problem 1

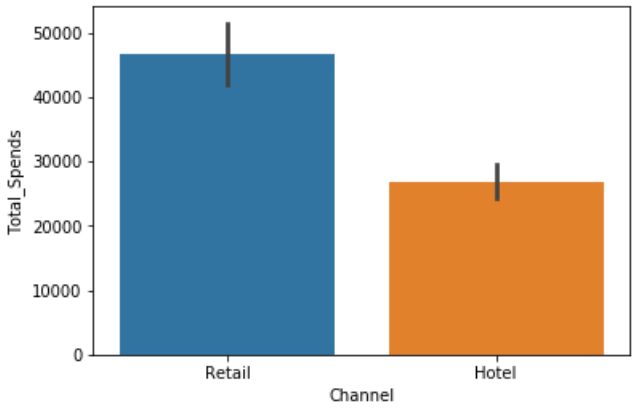
PROBLEM STATEMENT

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 (Wholesale Customer.csv) 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/Restaurant/Café HoReCa, Retail).

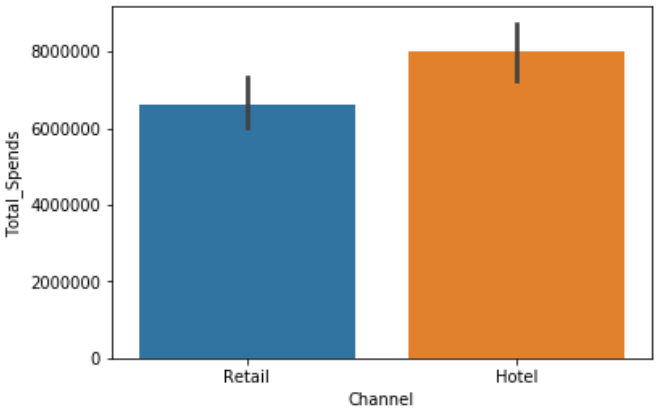
1.1. USE METHODS OF DESCRIPTIVE STATISTICS TO SUMMARIZE DATA.

WHICH REGION AND WHICH CHANNEL SEEMS TO SPEND MORE? WHICH REGION AND WHICH CHANNEL SEEMS TO SPEND LESS?

Although total spend by Hotel channel is more in absolute terms, **however on the average Hotel Channel spends less than the Retail Channel.**

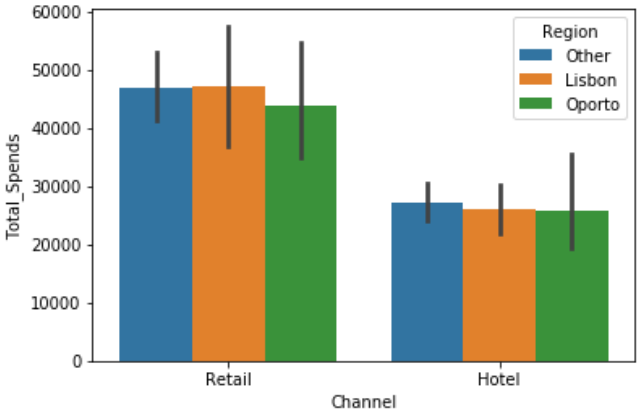


Using Mean



Using Sum

Since there is disparity in the number of records for Hotel and Retail channel, we would choose the Mean for better clarity. **Hence Retail Channel spends more on the average.**



Channel	Region	Total_Spends
Hotel	Lisbon	26073.593220
	Oporto	25683.928571
	Other	27213.635071
Retail	Lisbon	47137.277778
	Oporto	43996.736842
	Other	47004.971429

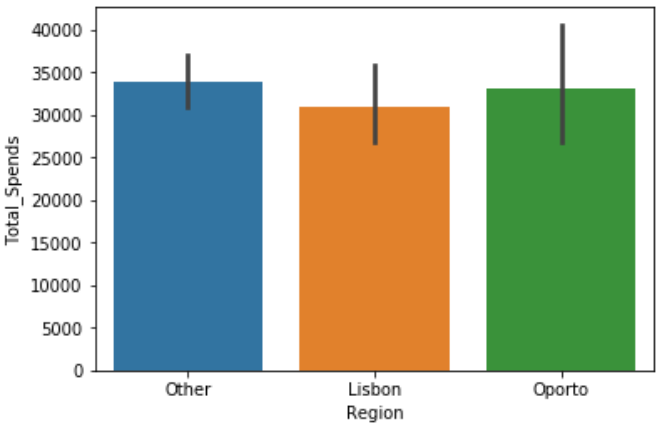
Refer above graph and table to see how regions fare for both the channels i.e. Retail as well as Hotel.

Problem 1

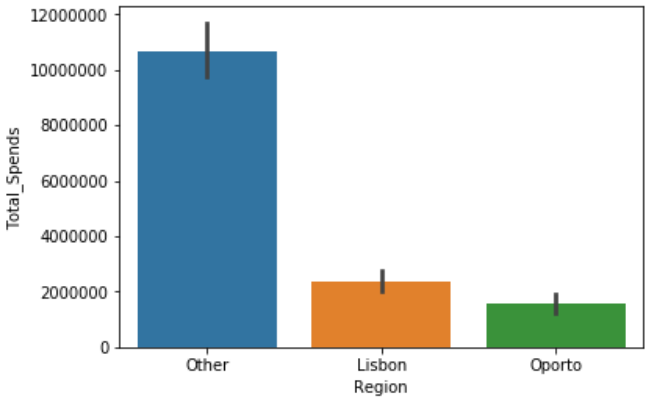
1.1. USE METHODS OF DESCRIPTIVE STATISTICS TO SUMMARIZE DATA.

WHICH REGION AND WHICH CHANNEL SEEMS TO SPEND MORE? WHICH REGION AND WHICH CHANNEL SEEMS TO SPEND LESS? (CONTINUED)..

Although total spend by **Other** region is more in absolute terms **as well as on the average. However on the average Other regions spend very close to Oporto Region.**



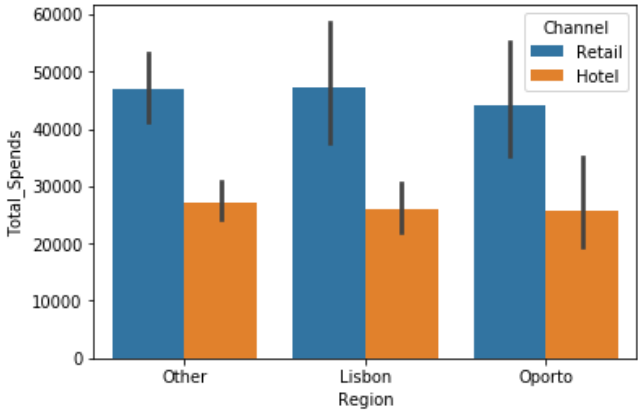
Using Mean



Using Sum

Since there is disparity in the number of records for each region, we would choose the Mean for better clarity. **Hence Other region spends more on the average, followed by Oporto and then Lisbon.**

Region	Total_Spends
Other	33789.870253
Oporto	33086.978723
Lisbon	30997.571429



Region	Channel	Total_Spends
Lisbon	Hotel	26073.593220
	Retail	47137.277778
Oporto	Hotel	25683.928571
	Retail	43996.736842
Other	Hotel	27213.635071
	Retail	47004.971429

Refer above graph and table to see how different Channels fare for different regions.

Problem 1

1.2. THERE ARE 6 DIFFERENT VARIETIES OF ITEMS ARE CONSIDERED. DO ALL VARIETIES SHOW SIMILAR BEHAVIOR ACROSS REGION AND CHANNEL?

Melted the given DataFrame (using pd.melt()) to form two different DataFrame of the below format. The one on the left is then used to plot a factor plot for comparing all the varieties with the different Regions and the one on the right is then used to plot a factor plot for comparing all the varieties with the different Channels.

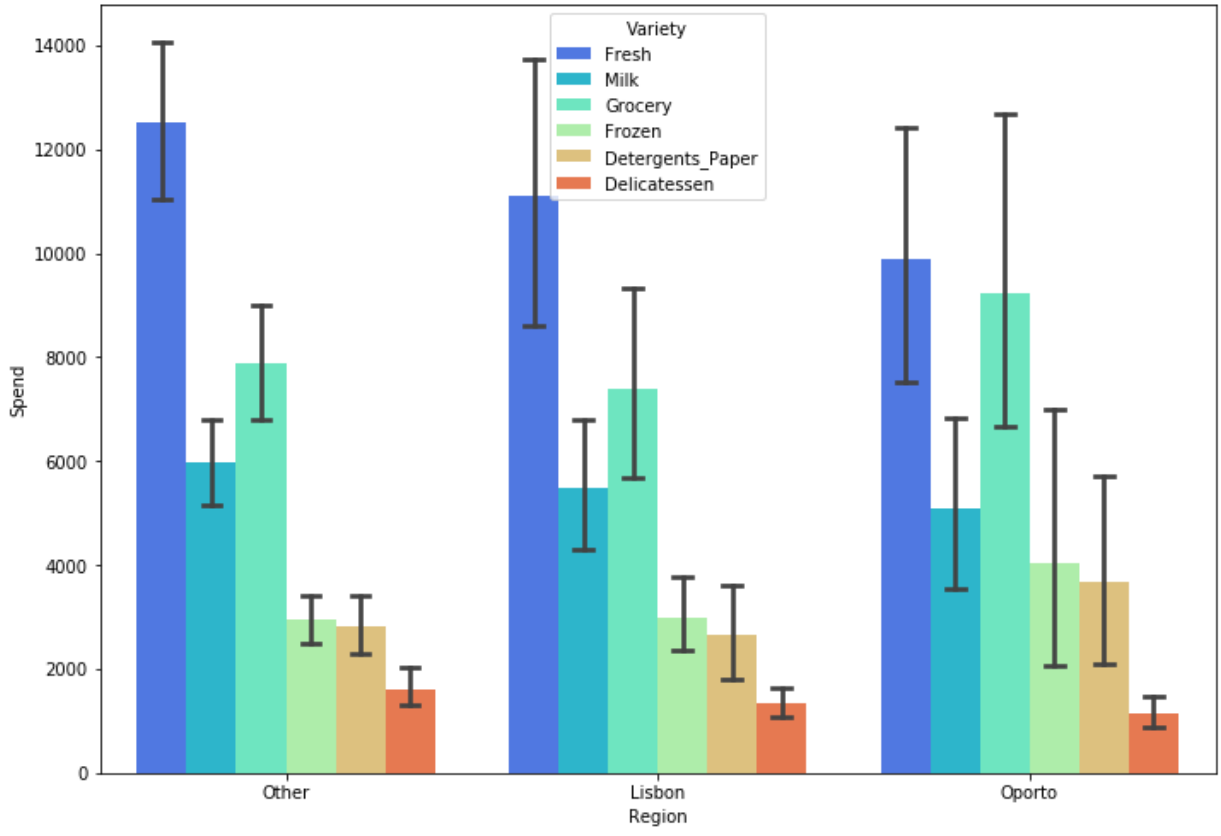
	Region	Variety	Spend
0	Other	Fresh	12669
1	Other	Fresh	7057
2	Other	Fresh	6353
3	Other	Fresh	13265
4	Other	Fresh	22615

Head of DataFrame for Regions

	Channel	Variety	Spend
0	Retail	Fresh	12669
1	Retail	Fresh	7057
2	Retail	Fresh	6353
3	Hotel	Fresh	13265
4	Retail	Fresh	22615

Head of DataFrame for Channels

Factor Plot depecting behaviour of all varieties across regions is as below.

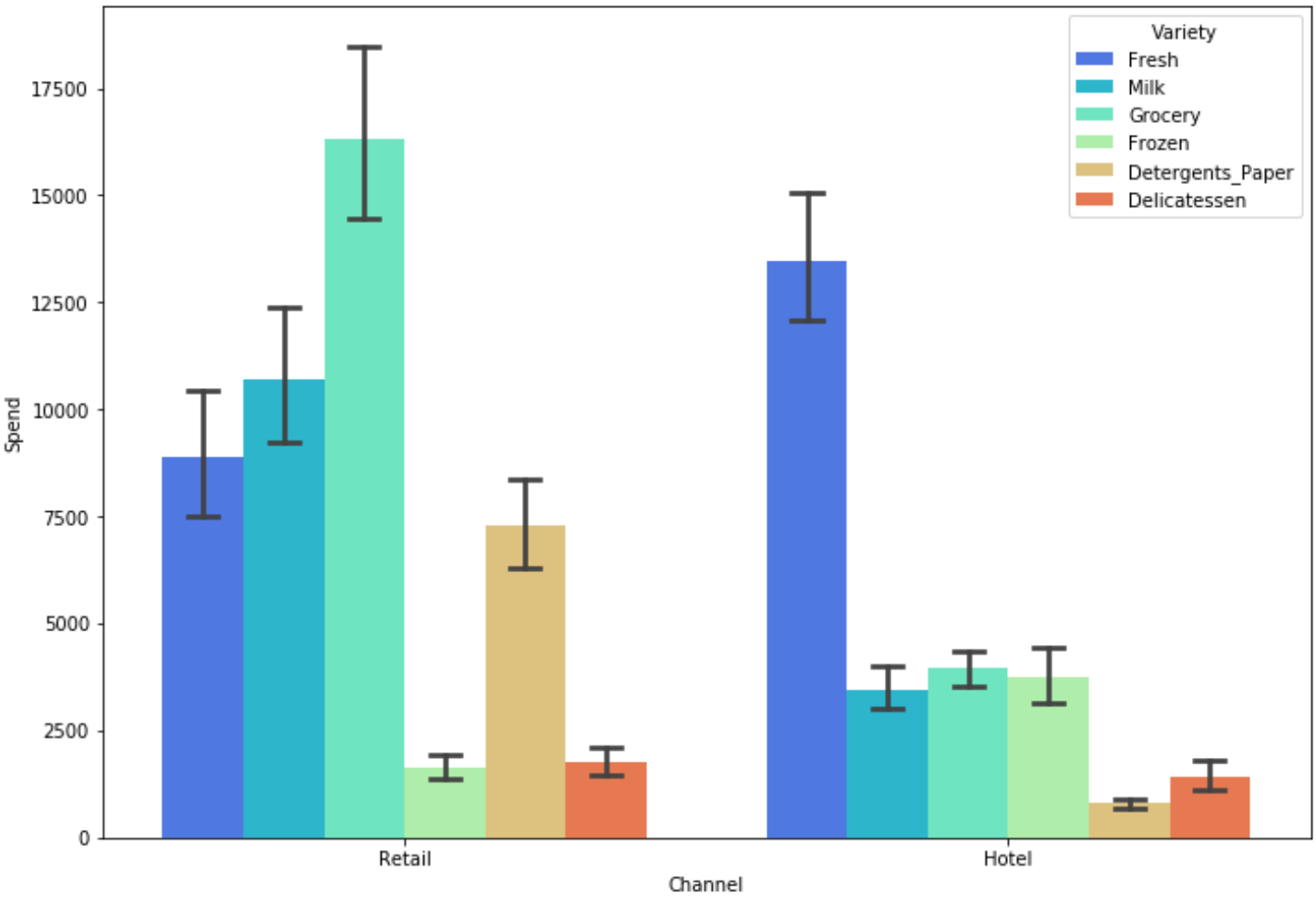


All varieties show almost the same behavior across regions, with an **exception of Grocery which is a little on the higher side for the Oporto region.**

Problem 1

1.2. THERE ARE 6 DIFFERENT VARIETIES OF ITEMS ARE CONSIDERED. DO ALL VARIETIES SHOW SIMILAR BEHAVIOR ACROSS REGION AND CHANNEL? (CONTINUED)..

Factor Plot depecting behaviour of all varieties across Channels is as below.



There is considerable difference in the behaviour of varieties across different Channels.

For instance

- Items from Fresh category sell more in Hotel Channel than Retail Channel.
- There is lot more demand for Grocery items in Retail channel compared to Hotel Channel.
- Similarly Detergents and Paper is more popular in Retail than in Hotel.

With the exception of Fresh and Frozen, all other food verieties sell more in Retail Channel.

Problem 1

1.3. ON THE BASIS OF THE DESCRIPTIVE MEASURE OF VARIABILITY, WHICH ITEM SHOWS THE MOST INCONSISTENT BEHAVIOUR? WHICH ITEMS SHOWS THE LEAST INCONSISTENT BEHAVIOUR?

We looked at various methods to determine the consistency of different variables. Below were the three methods by which we checked for inconsistency of the variables.

1. Standard Deviation
2. Variation
3. Co-efficient of Variation

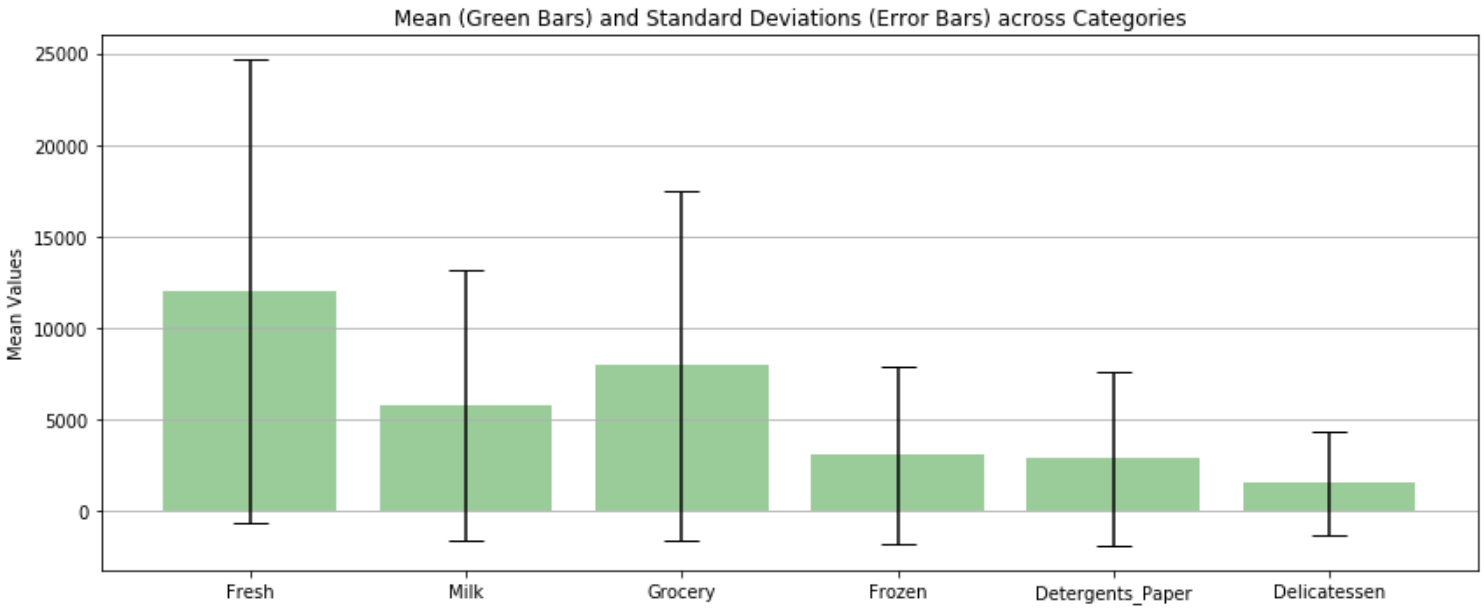
	Standard Deviation
Delicatessen	2820
Detergents_Paper	4768
Frozen	4855
Milk	7380
Grocery	9503
Fresh	12647

	Variance
Delicatessen	7952997
Detergents_Paper	22732436
Frozen	23567853
Milk	54469967
Grocery	90310104
Fresh	159954927

	Co-efficient of Variance
Delicatessen	1.85
Detergents_Paper	1.65
Frozen	1.58
Milk	1.27
Grocery	1.19
Fresh	1.05

We can clearly see from all the above mentioned three approaches that **Delicatessen** shows the the least inconsistent behavior. For Delicatessen we can see least standard deviation, least variation and the maximum co-efficient of variance (since it is inversely propotional).

Also refer to below diagram where the error bars represent the standard deviation for each variety. we can clearly see delicatessen has the smallest error bar of all the varities.



Problem 1

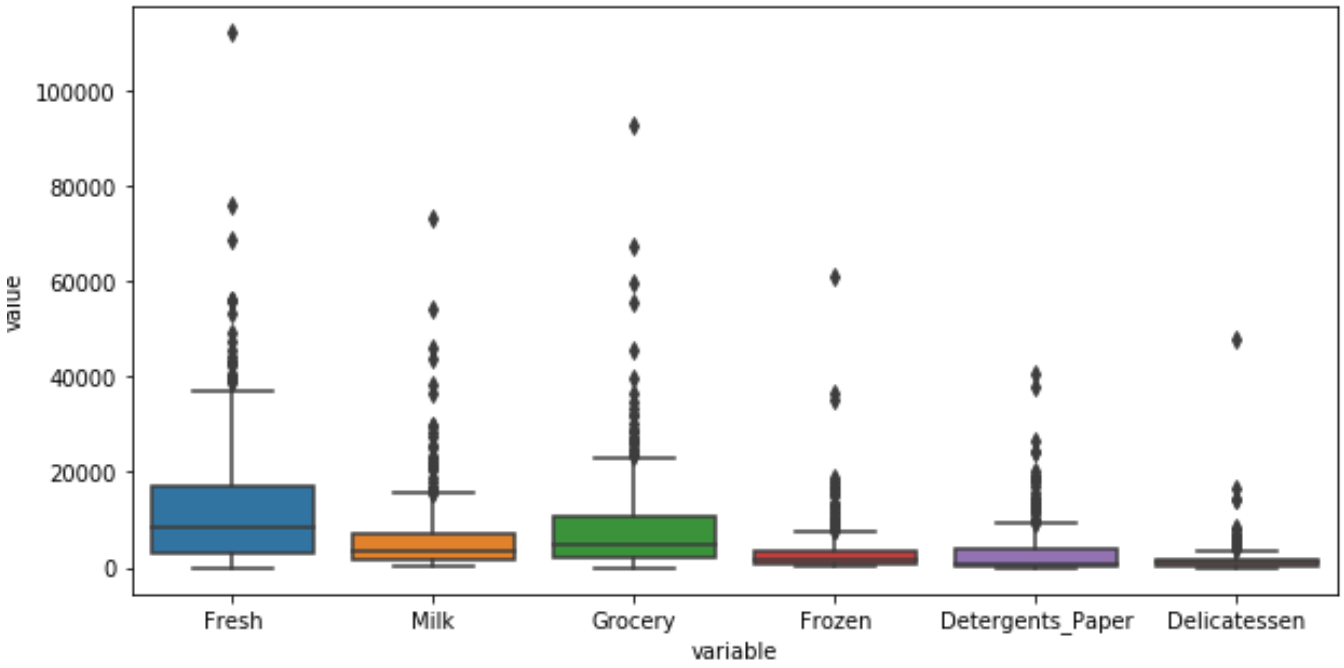
1.4. ARE THERE ANY OUTLIERS IN THE DATA?

- There are 138 outliers in the Hotel Channel compared to 66 in retail channel.
- There are 141, 33, 28 outliers in Other, Lisbon and Oporto Regions respectively.

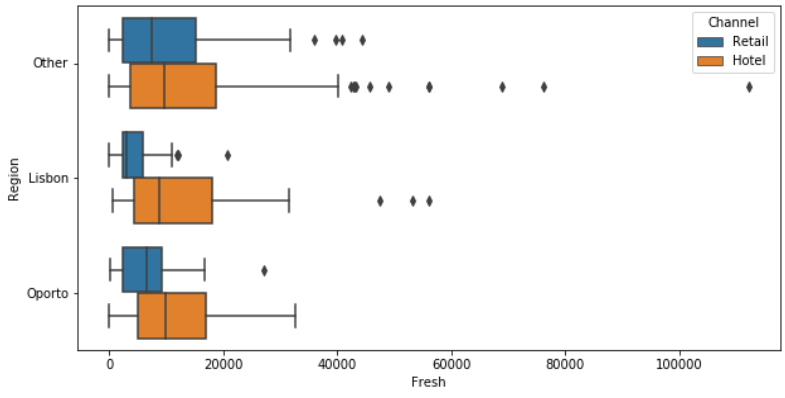
	Channel	Outlier_Count
1	Hotel	138
0	Retail	66

	Region	Outlier_Count
0	Other	141
1	Lisbon	33
2	Oporto	28

Univariate Outlier Data Analysis

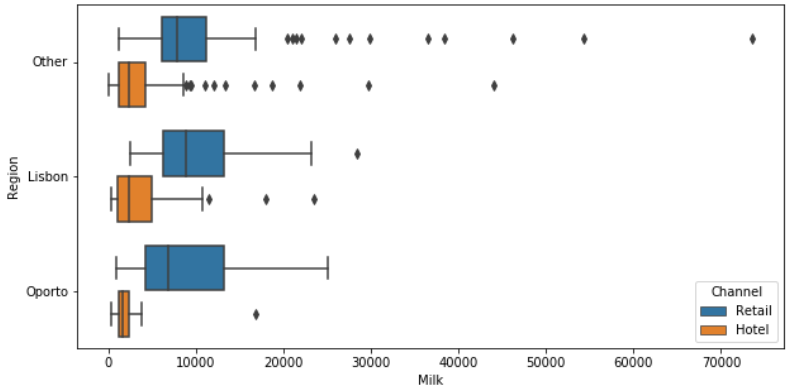


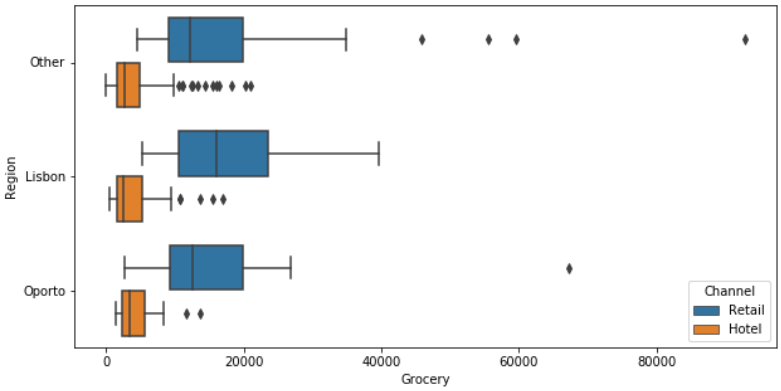
Multivariate Outlier Data Analysis

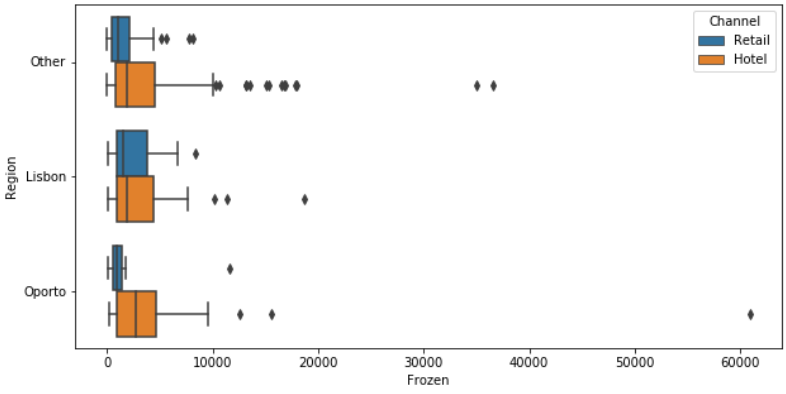


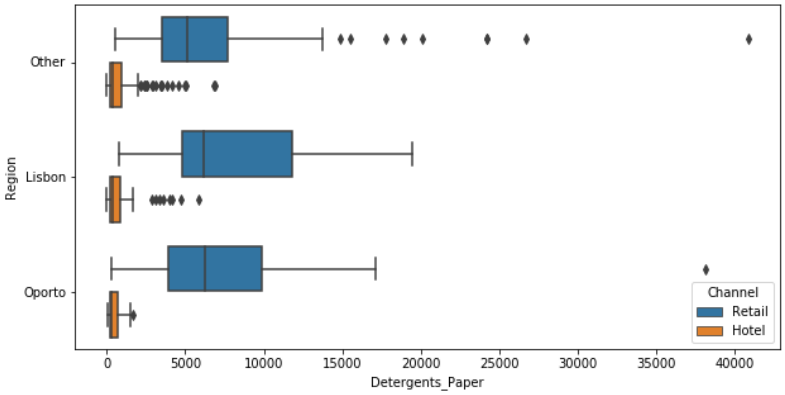
Problem 1

1.4. ARE THERE ANY OUTLIERS IN THE DATA?



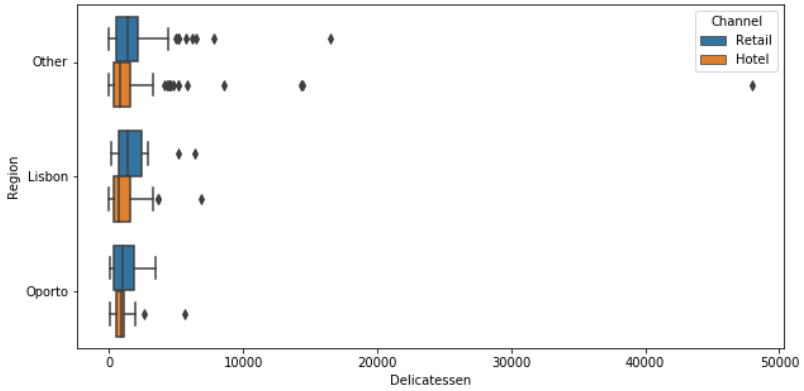






Problem 1

1.4. ARE THERE ANY OUTLIERS IN THE DATA?



1.5. ON THE BASIS OF THIS REPORT, WHAT ARE THE RECOMMENDATIONS?

Based on the above report we can provide below recommendations to the wholesaler.

- Wholesaler needs to sell more **Milk, Grocery & Detergents_Paper** items in **Hotel** Channel.
- Needs to increase sales of **Fresh** items in **Retail** Channel.
- Needs to increase sales of **Milk** in **Oporto** Region via **Hotel** Channel.
- Needs to increase sales of **Delicatessen** in **Oporto** region via **Hotel** channel.
- Needs to increase sales of F**rozen** in **Oporto** region via **Hotel** Channel**.**
- Needs to increase sales of **Detergent** in **Other** region via **Retail** channel.

Overall some improvement is required in Oporto region for Hotel Channel.

Problem 2

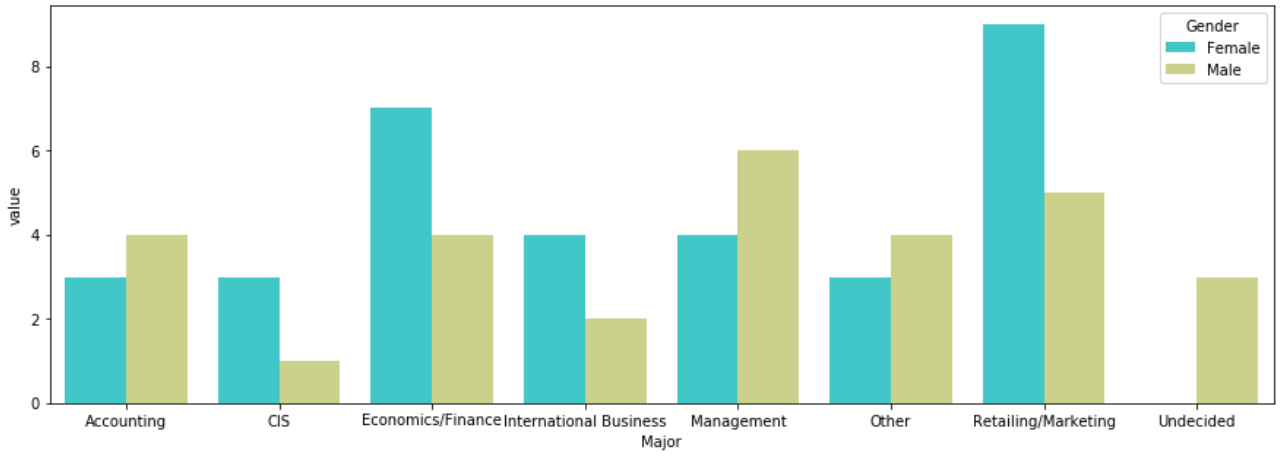
PROBLEM STATEMENT

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.csv file).

2.1. FOR THIS DATA, CONSTRUCT THE FOLLOWING CONTINGENCY TABLES (KEEP GENDER AS ROW VARIABLE)

2.1.1. GENDER AND MAJOR

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

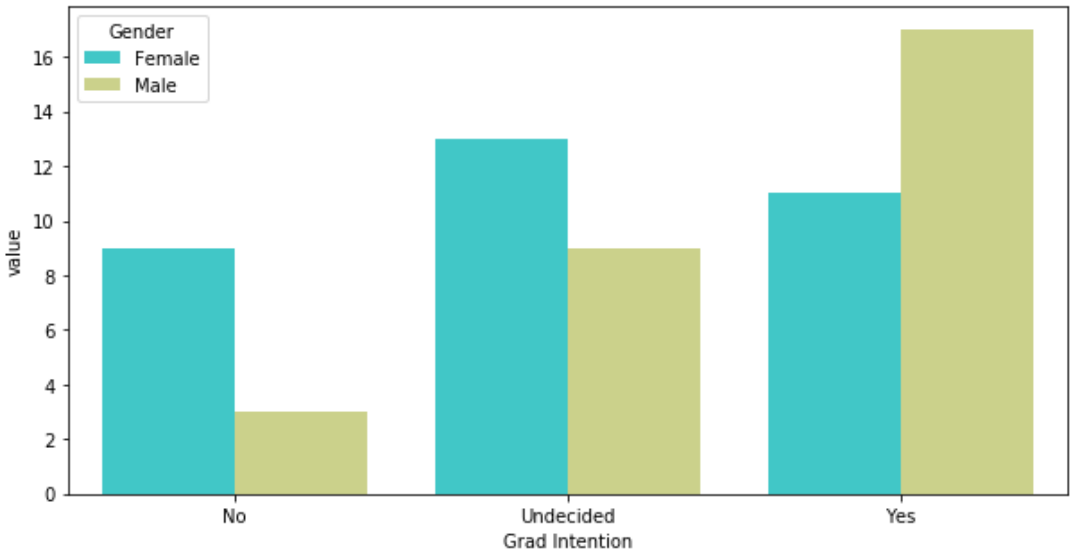


2.1.2. GENDER AND GRAD INTENTION

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

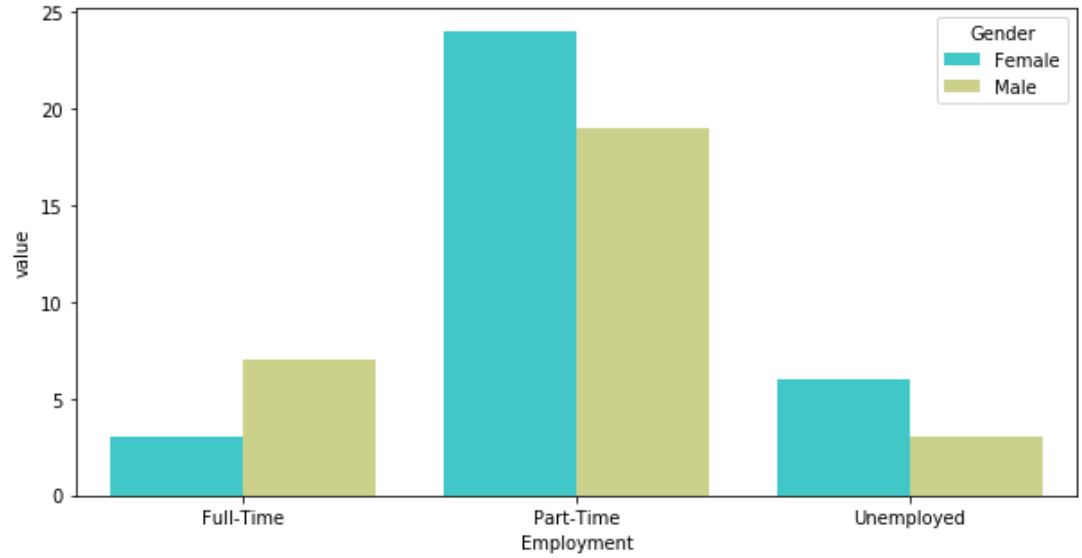
Problem 2

2.1.2. GENDER AND GRAD INTENTION



2.1.3. GENDER AND EMPLOYMENT

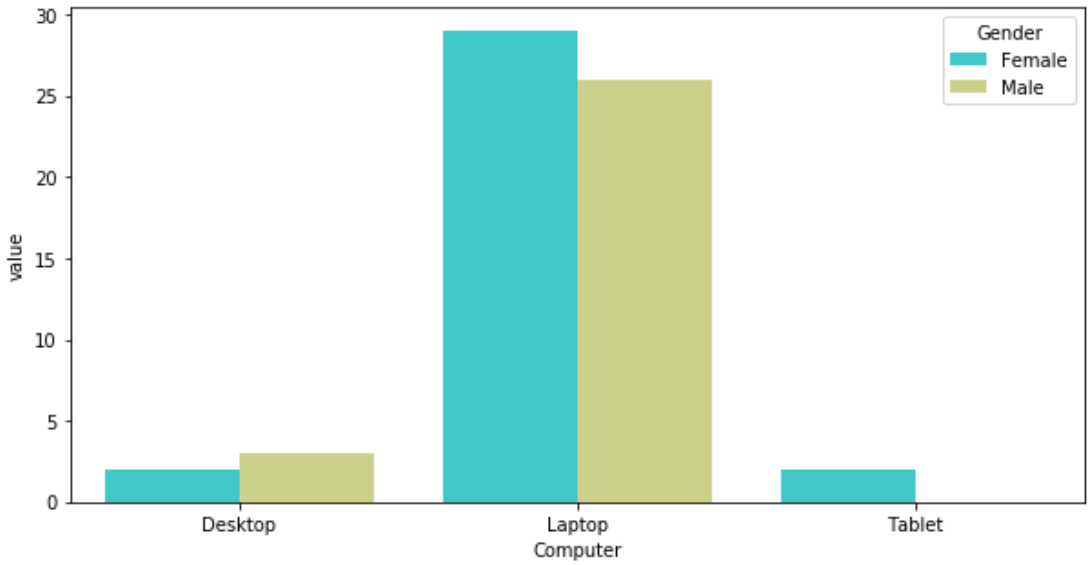
Employment	Full-Time	Part-Time	Unemployed	All
Gender				
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62



Problem 2

2.1.4. GENDER AND COMPUTER

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



2.2. ASSUME THAT THE SAMPLE IS A REPRESENTATIVE OF THE POPULATION OF CMSU. BASED ON THE DATA, ANSWER THE FOLLOWING QUESTIONS:

2.2.1. WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE MALE? WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE FEMALE?

We do this using basic formula of probability.

P(E) = Number of outcomes favourable to E / Total Number of outcomes
Hence,
P(Male) = Number of Males / Total Students
P(Male) = 29 / 62
P(Male) = 0.4677
or
47% Probability that randomly selected CMSU student will be Male.

Problem 2

2.2.1. WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE MALE? WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED CMSU STUDENT WILL BE FEMALE? (CONTINUED)..

Similarly for Females,

P(Female) = Number of Females / Total Students
P(Female) = 33 / 62
P(Female) = 0.5322
or
53% Probability that randomly selected CMSU student will be Female.

2.2.2. FIND THE CONDITIONAL PROBABILITY OF DIFFERENT MAJORS AMONG THE MALE STUDENTS IN CMSU. FIND THE CONDITIONAL PROBABILITY OF DIFFERENT MAJORS AMONG THE FEMALE STUDENTS OF CMSU.

We have used the contingency tables created above to calculate the conditional probabilities.
E.g.

P(Accounting / Male) = Number of Males having Accounting Major / Total Number of Males
P(Accounting / Male) = 4/29
P(Accounting / Male) = 0.1379
or,
13.79 % Probability that a given a male student, he will have Accounting as a Major.

Similarly we have calculated probability for all the other majors and in all other questions using the same formula in Python and have created a contingency table of the probabilities.

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	9.09	9.09	21.21	12.12	12.12	9.09	27.27	0.00	100.0
Male	13.79	3.45	13.79	6.90	20.69	13.79	17.24	10.34	100.0

2.2.3. FIND THE CONDITIONAL PROBABILITY OF INTENT TO GRADUATE, GIVEN THAT THE STUDENT IS A MALE. FIND THE CONDITIONAL PROBABILITY OF INTENT TO GRADUATE, GIVEN THAT THE STUDENT IS A FEMALE.

Grad Intention	No	Undecided	Yes	All
Female	27.27	39.39	33.33	100.0
Male	10.34	31.03	58.62	100.0

Problem 2

2.2.4. FIND THE CONDITIONAL PROBABILITY OF EMPLOYMENT STATUS FOR THE MALE STUDENTS AS WELL AS FOR THE FEMALE STUDENTS.

Employment	Full-Time	Part-Time	Unemployed	All
Female	9.09	72.73	18.18	100.0
Male	24.14	65.52	10.34	100.0

2.2.5. FIND THE CONDITIONAL PROBABILITY OF LAPTOP PREFERENCE AMONG THE MALE STUDENTS AS WELL AS AMONG THE FEMALE STUDENTS.

Computer	Desktop	Laptop	Tablet	All
Female	6.06	87.88	6.06	100.0
Male	10.34	89.66	0.00	100.0

	Laptop
Female	87.88
Male	89.66

Refer to the table at the right, if we are looking probabilities specifically for Laptops.

2.3. BASED ON THE ABOVE PROBABILITIES, DO YOU THINK THAT THE COLUMN VARIABLE IN EACH CASE IS INDEPENDENT OF GENDER?JUSTIFY YOUR COMMENT IN EACH CASE.

We will take a look at all the probabilities calculated above and then, logically looking at the data, we will try to answer the above question.

There were different methods that were tried to answer this

1) Chi-Squared --> This proved all the column variables to be independent of the Gender.

2) Independence Formula -->
i.e. if P(A/B) = P(A) then A is said to be independent of B.
Similarly we tried to calculate say for example
P(Laptop / Male) = P(Male)
26/29 != 29/62
0.89 !=0.46

Using this approach, we could see all the column variables were dependent on Gender, as the probabilities did not match exactly in any of these cases. Refer jupyter notebook for implementation.

However what seemed to make most sense was to look at the data individually and then comment on case by case, whether it is dependent on Gender or not. Which is what we have done below.

Problem 2

2.3. BASED ON THE ABOVE PROBABILITIES, DO YOU THINK THAT THE COLUMN VARIABLE IN EACH CASE IS INDEPENDENT OF GENDER?JUSTIFY YOUR COMMENT IN EACH CASE.

3) Logical Analysis of Data

- **Gender and Major Independence**

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	9.09	9.09	21.21	12.12	12.12	9.09	27.27	0.00	100.0
Male	13.79	3.45	13.79	6.90	20.69	13.79	17.24	10.34	100.0

  - Looking at the data, we can see more females tend to gravitate towards, **Economics/ Finance & Retail/Marketing**. While more males tend to prefer **Management.**
  - Hence we can say Major is **Dependent** on Gender variable.

- **Gender and Graduate Independence**

Grad Intention	No	Undecided	Yes	All
Female	27.27	39.39	33.33	100.0
Male	10.34	31.03	58.62	100.0

  - Looking at the data we can clearly see, females have less intention to graduate with as many as 9 saying, they do not intend to graduate.
  - Males on the other hand were clearly ahead in answering yes for this question.
  - Hence we can say Grad Intention is **Dependent** on Gender variable.

Problem 2

2.3. BASED ON THE ABOVE PROBABILITIES, DO YOU THINK THAT THE COLUMN VARIABLE IN EACH CASE IS INDEPENDENT OF GENDER?JUSTIFY YOUR COMMENT IN EACH CASE.

3) Logical Analysis of Data

- **Gender and Employment Independence**

Employment	Full-Time	Part-Time	Unemployed	All
Female	9.09	72.73	18.18	100.0
Male	24.14	65.52	10.34	100.0

  - Looking at the above data, we can see Females prefer more Part-time jobs and more of them are Unemployed.
  - Also it is evident from the data that males tend to prefer more Full-time job compared to Part-Time.
  - Hence we can say Employment column variable is **Dependent** on Gender variable.

- **Gender and Laptop Independence**

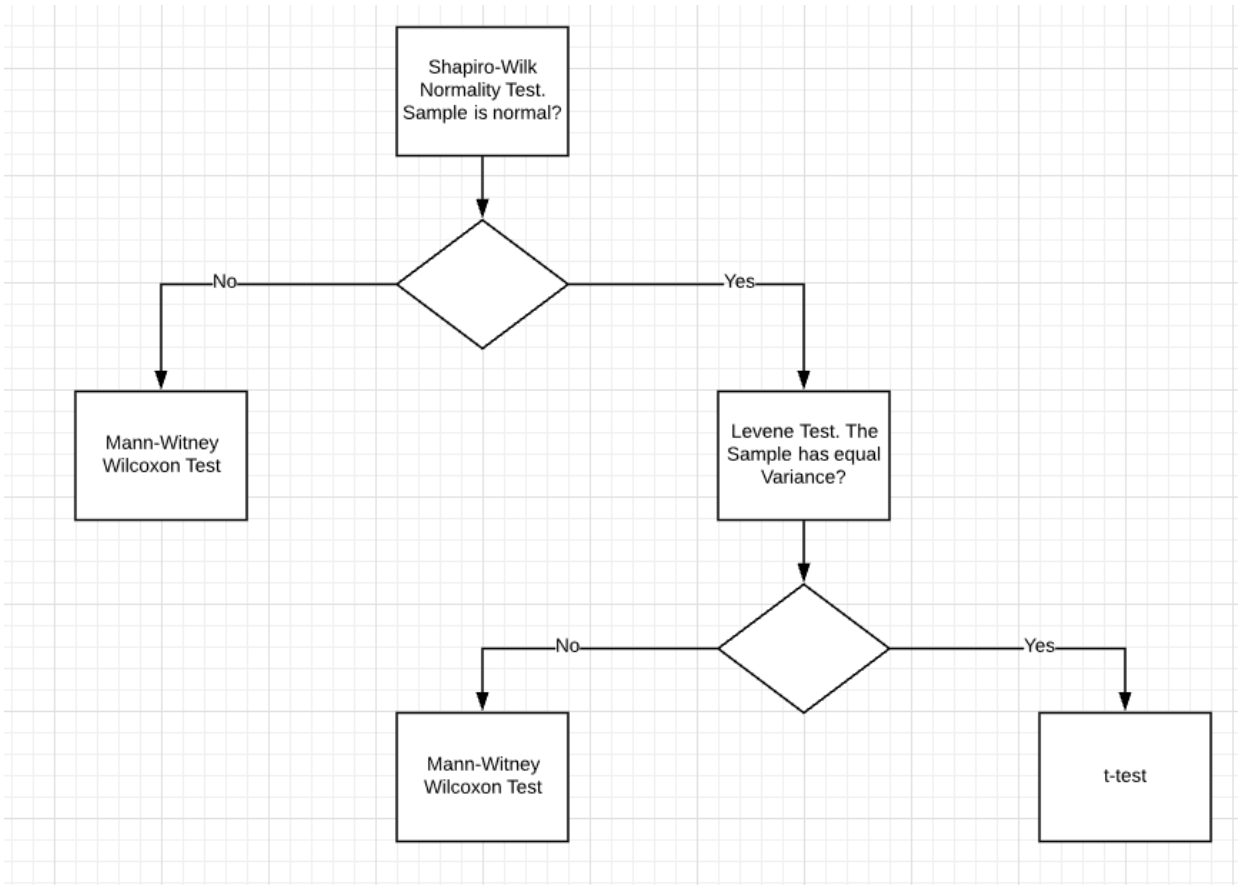
Computer	Desktop	Laptop	Tablet	All
Female	6.06	87.88	6.06	100.0
Male	10.34	89.66	0.00	100.0

  - Looking at above data, we can see although there is some difference in the number of desktops prefered by each gender and also for tablets.
  - However these differences look too minor when compared with the numbers for Laptops.
  - Hence we can see irrespective of the gender, everyone tends to prefer Laptops.
  - Computer column variable is thus **Independent** of the gender variable.

Problem 2

2.4. NOTE THAT THERE ARE THREE NUMERICAL (CONTINUOUS) VARIABLES IN THE DATA SET, SALARY, SPENDING AND TEXT MESSAGES. FOR EACH OF THEM COMMENT WHETHER THEY FOLLOW A NORMAL DISTRIBUTION. WRITE A NOTE SUMMARIZING YOUR CONCLUSIONS. [RECALL THAT SYMMETRIC HISTOGRAM DOES NOT NECESSARILY MEAN THAT THE UNDERLYING DISTRIBUTION IS SYMMETRIC]

We used below approach to check if the three continuous variables are indeed following normal distribution or not.



We used the approach for all the three variables, refer the jupyter notebook for implementation.

Following Hypothesis was formed for these tests.
H0 --> Variable is following Normal Distribution
Ha --> variable is NOT following Normal Distribution

Shapiro test failed for all 3 variables i.e. p-value was less than 0.05, hence we went ahead with Wilcoxon test, this too showed p-value less than 0.05, hence we ended up rejecting the Null hypothesis in all the cases.

Hence **none** of the 3 continuous variables i.e. **Salary, Spending & Text Messages** follow normal distribution.

Problem 2

2.4. NOTE THAT THERE ARE THREE NUMERICAL (CONTINUOUS) VARIABLES IN THE DATA SET, SALARY, SPENDING AND TEXT MESSAGES. FOR EACH OF THEM COMMENT WHETHER THEY FOLLOW A NORMAL DISTRIBUTION. WRITE A NOTE SUMMARIZING YOUR CONCLUSIONS. [RECALL THAT SYMMETRIC HISTOGRAM DOES NOT NECESSARILY MEAN THAT THE UNDERLYING DISTRIBUTION IS SYMMETRIC]

We also applied many other tests like normal test, Anderson darling test and ended with the same result.

Also we checked if the mean,mode and median of the distributions are similar which would indicate towards a normal distribution, however with this approach too, we concluded, none of the 3 continuous variables follow normal distribution.

Distplot and QQplot were also refered.

(P.T.O)

Problem 3

PROBLEM STATEMENT

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. FOR THE A SHINGLES, FORM THE NULL AND ALTERNATIVE HYPOTHESIS TO TEST WHETHER THE POPULATION MEAN MOISTURE CONTENT IS LESS THAN 0.35 POUND PER 100 SQUARE FEET.

The hypothesis that we have formed for this is as below

H0 --> U >=0.35
Ha --> U < 0.35

We have used 1 sample t-test with 0.05 as the significance level here to evaluate, since standard deviation of population is not given.
p-value = 0.14955266289815025

Hence p-value > 0.05
Fail to reject null hypothesis.
U>0.35 for A Shingles

3.2. FOR THE B SHINGLES, FORM THE NULL AND ALTERNATIVE HYPOTHESIS TO TEST WHETHER THE POPULATION MEAN MOISTURE CONTENT IS LESS THAN 0.35 POUND PER 100 SQUARE FEET.

The hypothesis that we have formed for this is as below

H0 --> U >=0.35
Ha --> U < 0.35

We have used 1 sample t-test with 0.05 as the significance level here to evaluate, since standard deviation of population is not given.
p-value = 0.004180954800638365

Hence p-value < 0.05
Reject null hypothesis.
U<0.35 for B Shingles

Problem 3

3.3. DO YOU THINK THAT THE POPULATION MEANS FOR SHINGLES A AND B ARE EQUAL?FORM THE HYPOTHESIS AND CONDUCT THE TEST OF THE HYPOTHESIS.WHAT ASSUMPTION DO YOU NEED TO CHECK BEFORE THE TEST FOR EQUALITY OF MEANS IS PERFORMED?

The hypothesis that we have formed for this is as below

H0 --> Ua = Ub
Ha --> Ua != Ub

We have used ttest_ind, or 2 sample t-test to check for the above hypothesis at 0.05 significance level. We had to use nan_policy='omit' since there were less number of observations for Shingles B compared to Shingles A.

p-value = 0.2017496571835306
Hence p-value > 0.05

Fail to reject Null hypothesis.

Hence we can say that ,
Ua = Ub

Before performing ttest_ind, we had to check for the assumption of equal variances for both the populations.

Hence we checked for variance of Shingles A & Shingles B

p3_df['A'].var() --> 0.018422857142857133
p3_df['B'].var() --> 0.018850322580645163

We can see the variance of Shingles A as well as Shingles B is almost identical, hence ttest_ind can be applied. Post checking this asumption the above mentioned test was applied.

(P.T.O)

Problem 3

3.4. WHAT ASSUMPTION ABOUT THE POPULATION DISTRIBUTION IS NEEDED IN ORDER TO CONDUCT THE HYPOTHESIS TESTS ABOVE?

Below assumptions were made while conducting the hypothesis tests above.

1. The data follows a continuous distribution.
2. Both the samples are selected at random from their populations.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The fourth assumption is a reasonably large sample size is used.
5. The final assumption is homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.