BUSINESS REPORT

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.We will use visualization approach to describe the data. We will also find out outliers.

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?

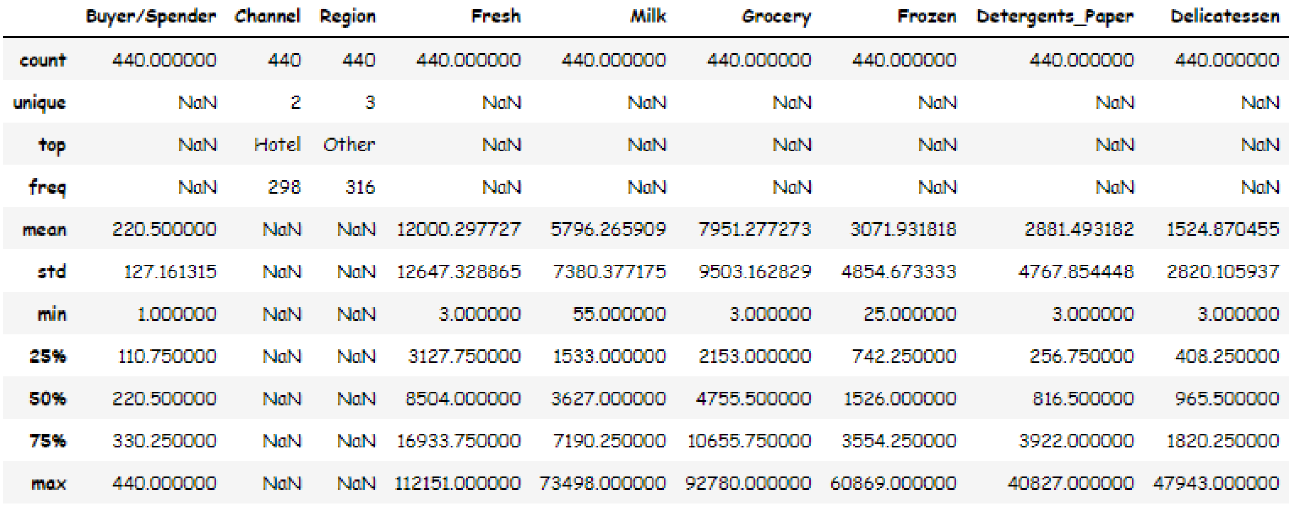
'Region' has 3 unique values with 'other' seems to spend more 316/440=0.718 i.e about 71.8 %

'Channel' has 2 unique values with 'hotel' seems to spend more 298/440=0.677 i.e about 67.7 %

'Region' has 3 unique values with 'Oporto' seems to spend less 47/440=0.106 i.e about 10.6 %

'Channel' has 2 unique values with 'Retail' seems to spend less 142/440=0.322 i.e about 32.2 %

1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

No all varieties do not show similar behaviour across Region and Channel

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour?

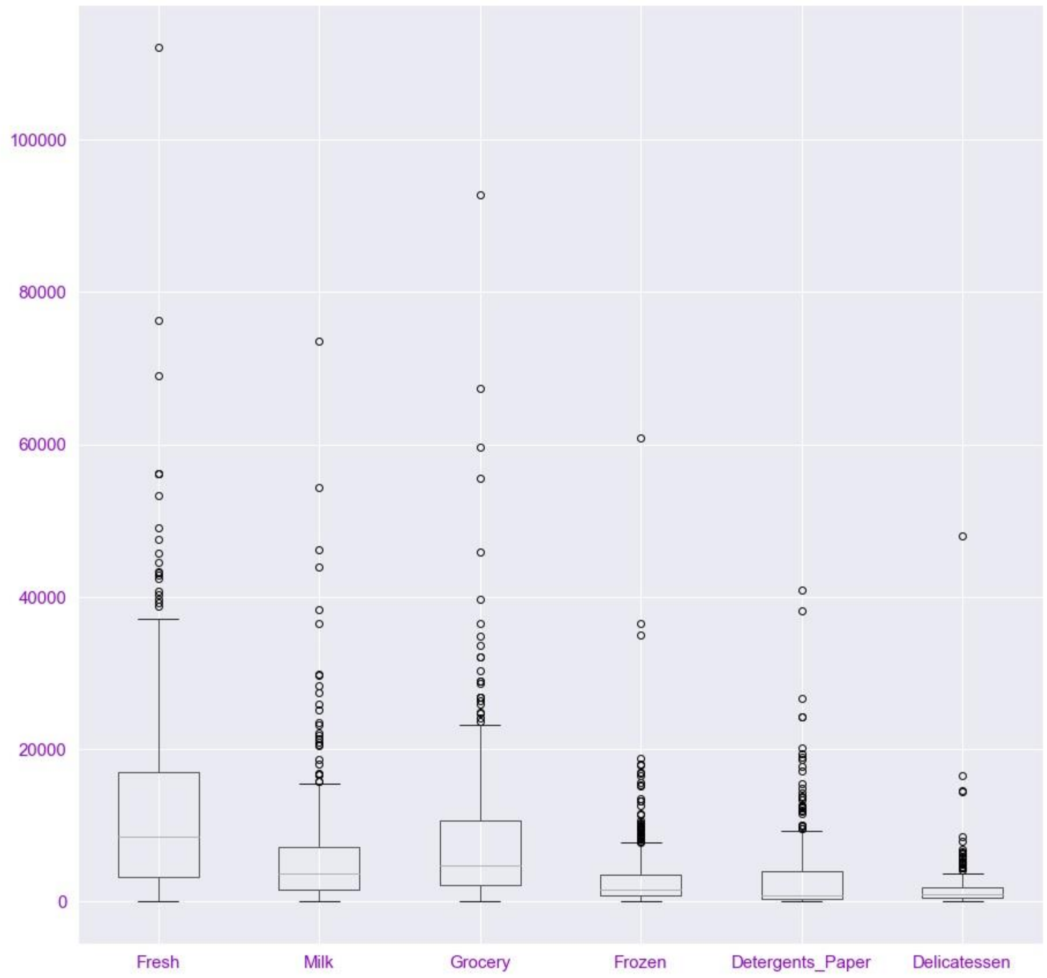
Which items shows the least inconsistent behaviour?

It can be seen that the standard deviation for each of these 6 product categories indicates vast spread of data in each of the product categories. The range is very big.

Fresh item has the Highest std =12647.3 so Fresh item shows the most inconsistent behaviour.

Delicatessen item has the lowest std = 2820.1 so Delicatessen item shows the least inconsistent behaviour.

1.4. Are there any outliers in the data?



Yes there are outliers in the data in all the items across the product range (Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen)

1.5. On the basis of this report, what are the recommendations?

The mean of the dataset indicates that on an average, the annual spending of the large retailers is highest for Fresh and lowest for Delicatessen.

The median of the dataset indicates the middle-most values of the annual spending of the large retailers

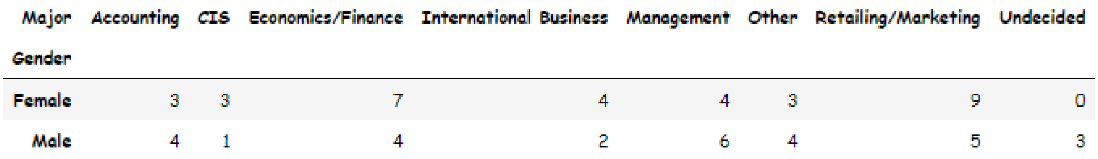
It is recommended that the distributors can increase both Grocery and Fresh items stocks across regions and channels

It is recommended that the distributors can decrease Frozen stock across Regions and channel and decrease Detergents_Paper for Lisbon and others – both in Hotel Channel.

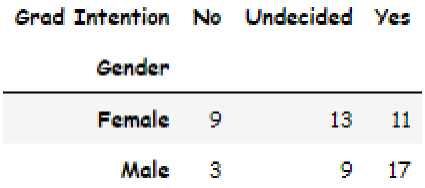
Problem 2

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

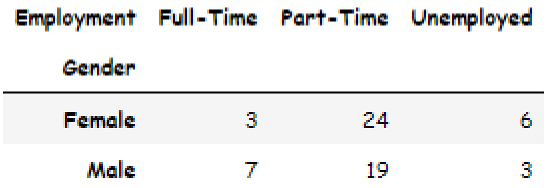
2.1.1. Gender and Major 2.1.2. Gender and Grad Intention 2.1.3. Gender and Employment 2.1.4. Gender and Computer



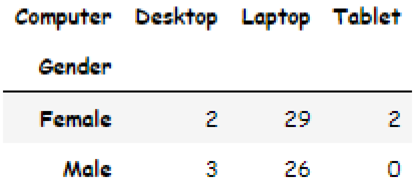
Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Gender								
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3



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



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



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

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?

Probability that a randomly selected CMSU student will be Male P(M)=No. of fav. outcomes/Total no. of outcomes = 29/62=0.468

i.e 46.8% probabaility that selected CMSU student will be male .

Probability that a randomly selected CMSU student will be Female P(F)=No. of fav. outcomes/Total no. of outcomes = 33/62=0.532

i.e 53.2% probabaility that selected CMSU student will be male .

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.

Retailing/Marketing	**14**
Economics/Finance	**11**
Management	**10**
Accounting	**7**
Other	**7**
International Business	**6**
CIS	**4**
Undecided	**3**

FOR MALE:

P(Accounting/Male) = 4/29 =0.137 i.e 13.7%

P(CIS/Male)=1/29 =0.034 i.e 3.4%

P(Economics Finance/Male)=4/29 =0.137 i.e 13.7%

P(International Business/Male)=2/29 =0.068 i.e 6.8%

P(Management/Male/Male/Male/Male)=6/29 =0.206 i.e 20.6%

P(Other/Male/Male/Male)=4/29 =0.137 i.e 13.7%

P(Retailing/Marketing/Male/Male)=5/29 =0.172 i.e 17.2%

P(Undecided/Male)=3/29 =0.103 i.e 10.3%

FOR FEMALE:

P(Accounting/Female) = 3/33 =0.090 i.e 9.0%

P(CIS/Female/Female)=3/33 =0.090 i.e 9.0%

P(Economics Finance/Female)=7/33 =0.212 i.e 21.2%

P(International Business/Female)=4/33=0.121 i.e 12.1%

P(Management/Female)=4/33 =0.121 i.e 12.1%

P(Other/Female)=3/33 =0.090 i.e 9.0%

P(Retailing/Marketing/Female)=9/33 =0.272 i.e 27.2%

P(Undecided/Female)=0/33 =0 i.e 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.

Probability of intent to graduate=G

Given that the student is a Male=M

P of (Gradintention No/Male) =3/29 = 0.103 i.e 10.3%

P of (Gradintention Undecided/Male)=9/29 =0.031 i.e 31.0%

P of (Gradintention Yes/Male) = 17/29 =0.586 i.e 58.6%

Given that the student is a Female=F

P of (Gradintention No/Female)=9/33 = 0.2727 i.e 27.3%

P of (Gradintention Undecided/Female)=13/33 =0.394 i.e 39.4%

P of (Gradintention Yes/Female) = 11/33 =0.333 i.e 33.3%

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

Probability of employment status for the male students =

P of (Full time/Male)=7/29=.241 i.e 24.1%

P of (Part time/Male) =19/29=0.655 i.e 65.5%

P of (Unemployed/Male)=3/29=0.103 i.e 10.3%

Probability of employment status for the Female students =

P of (Full time/Female)=3/33 =0.090 i.e 9.0%

P of (Part time/Female)=24/33=0.727 i.e 72.7%

P of (Unemployed/Female)=6/33=0.182 i.e 18.2%

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

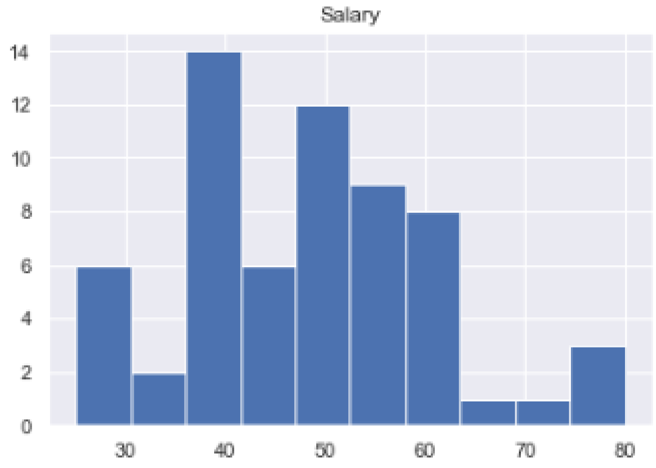
Probability of Laptop Preference among the Male students = P(L/M)=26/62=.419 i.e 41.9%

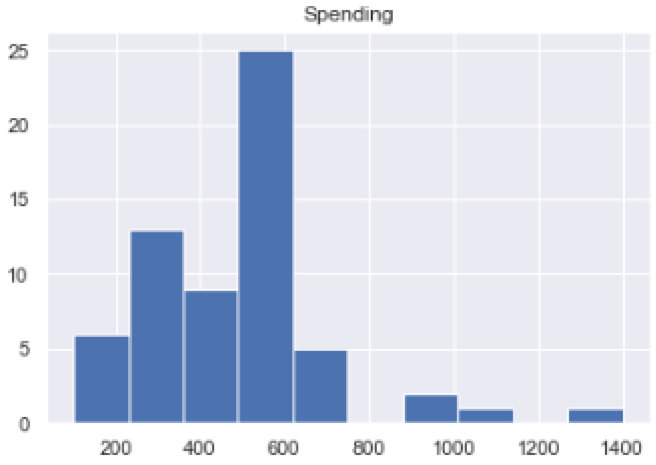
Probability of Laptop Preference among the Female students = P(L/F)=29/62=.468 i.e 46.8%

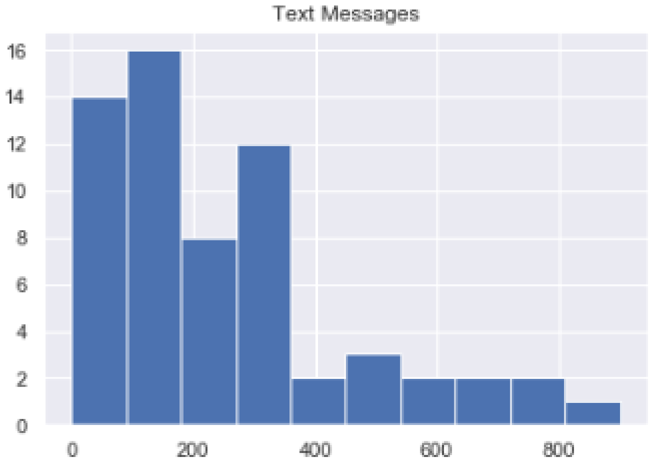
2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? Justify your comment in each case. Part II

No, the column variable in any of the case is not independent of the genderThere are 33 female and 29 male students who responded to the survey. More female students responded to the survey more than the male students and male and female have different probability for all column variables like we have infered bhy calculating probabilities.

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]







None of them follows a normal distribution as per the above histograms. All of the above histograms appears to be right skewed.

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 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 is calculated. The company claims that the mean moisture content cannot be greater 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 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?

We assume that samples are random and both the populations are normally distributed.

Hypothesis

For the two-sample independent tTest,

population means of shingles *A* = u*A*

population means of shingles B = uB

Ho:u*A* = uB

Ha: u*A* /= uB

t=1.28

p value=0.202281

p value>0.05, fail to reject null hypothesis.

At 5% significance level:

The two sample t-test reveals since p-value is greater than 0.05, we fail to reject null hypothesis. At 95% Confidence Interval, there is insufficient evidence to support the claim that population means for shingles *A* and B are equal

3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

Assumptions on population distribution to conduct the above hypothesis tests:

Distribution should be normally distributed before proceeding with the Test

Both *A* & B populations have same variance

Each value is sampled independently for *A* and B

Samples are random