**SMDM PROJECT**

**BUSINESS REPORT**

**Submitted By:**

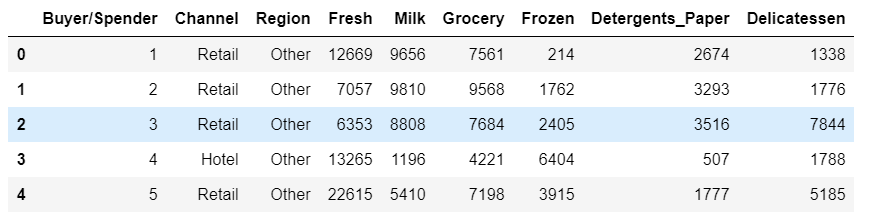
**Prachi Gupta**

**Problem #1**

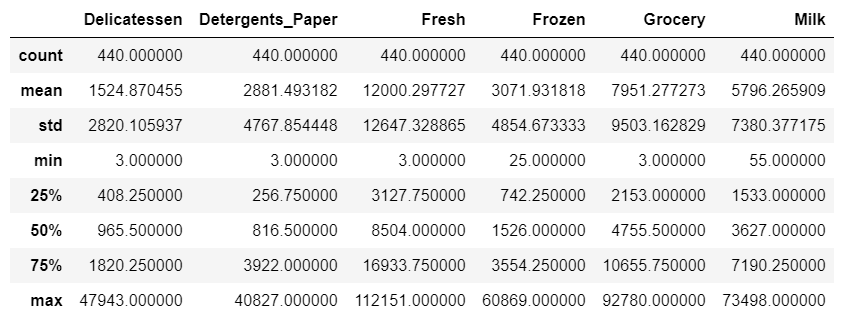
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale Customer.csv) consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

* 1. **a) Use methods of descriptive statistics to summarize data.**

To get insights into data, I have used ‘.**read\_csv()**’ function from pandas library, and to fetch first or last five observations, ‘.**head()**’ or ‘.**tail()**’ functions can be used. I have used ‘.**head()**’ here.

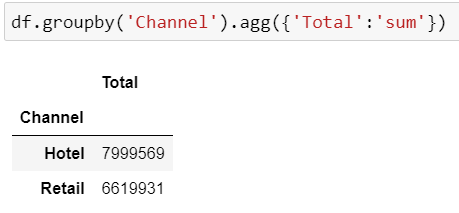
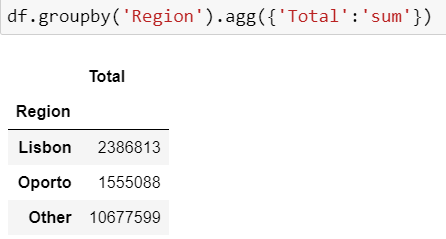


I have used ‘**describe()**’ function in pandas to get descriptive statistics summary. It returns the count, mean, standard deviation, minimum, maximum values and the 3 quantiles of the dataframe provided.



Here, as you can see, mean value is greater than median value for all the columns showing Items, which is shown as 50% (50th percentile) in the index column.

On the basis of above shown calculation using ‘.**Groupby()**’ method, we can conclude that:

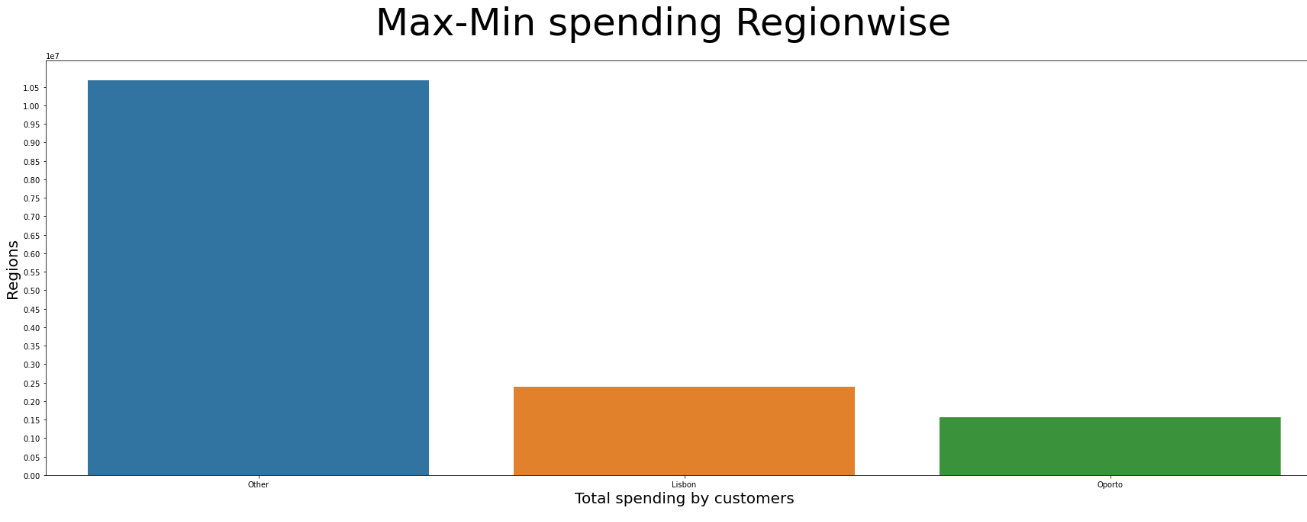
* 1. **b) Which Region and which Channel spend the most:**

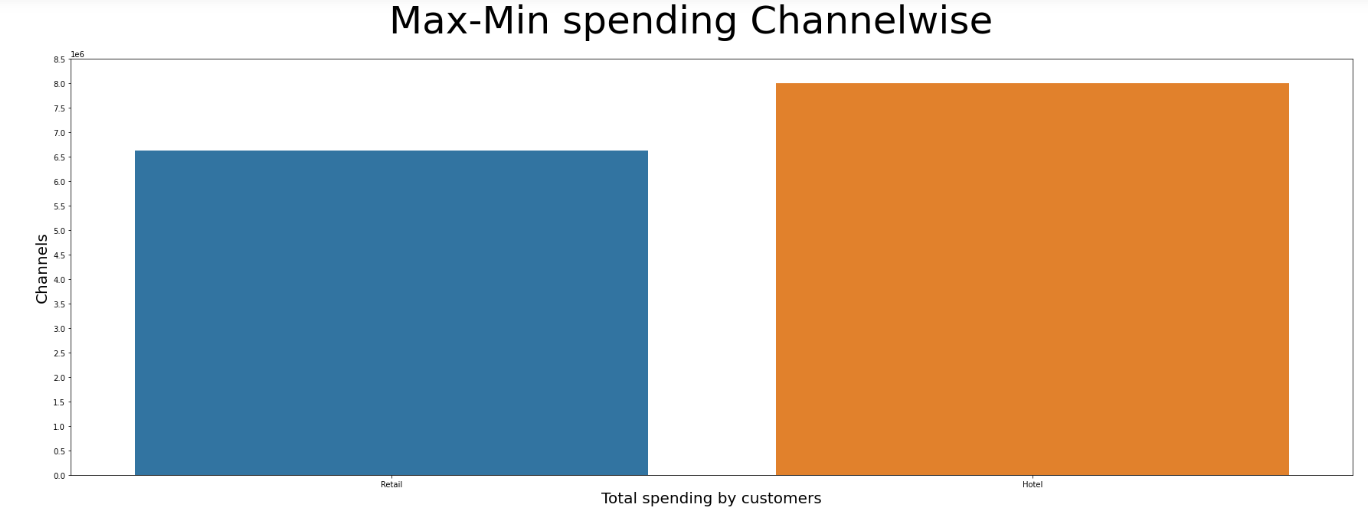
**Region**: Other, **Channel**: Hotel

* 1. **c) Which Region and which Channel spend the least:**

**Region**: Oporto, **Channel**: Retail

Following are the Barplots showing the max-min spending of the customers across Regions & across Channels:





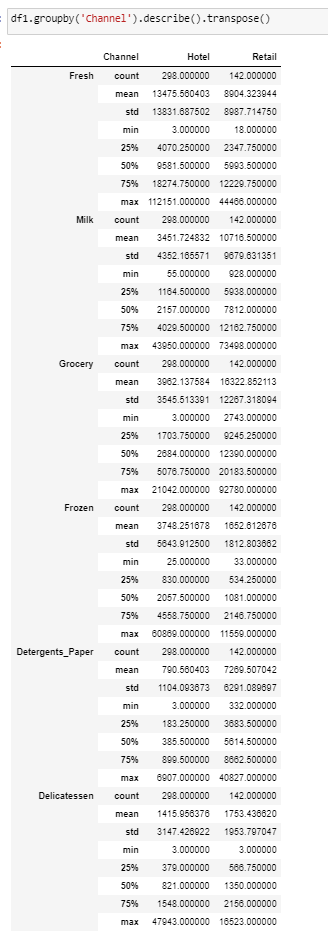
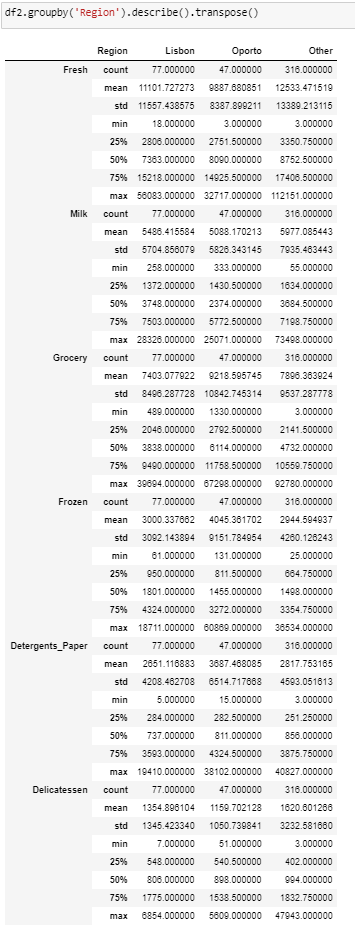
* 1. **There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.**

Performed the descriptive analysis of the 6 varieties, individually

Across 3 ‘**Region’**: Lisbon, Oporto, Other­­­­­

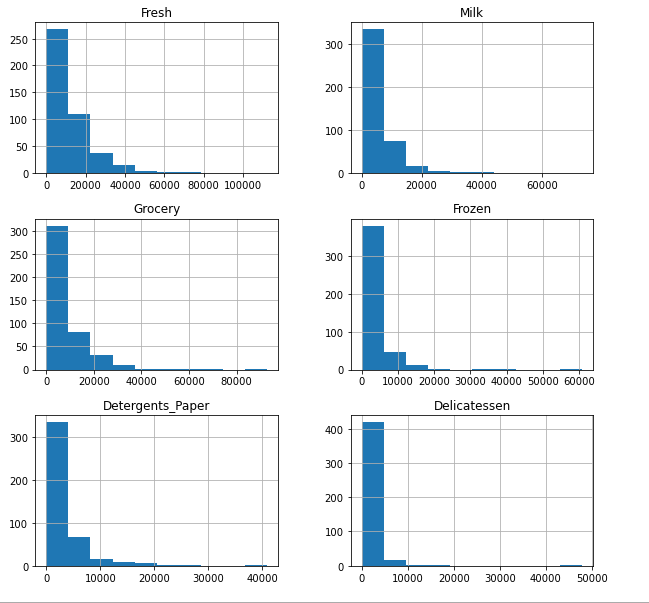
Across 2 ‘**Channel’** : Hotel, Retail using describe() and Histogram plot.

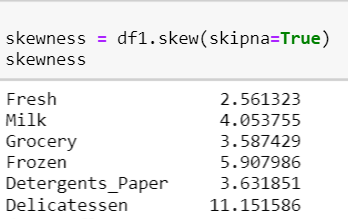
**Following are the descriptive summaries of the 6 varieties across ‘Channel’ and ‘Region’ :**

Following are the observations:

1. There are large differences between 75th %tile values and the max values for all 6 varieties across ‘**Region’** as well as ‘**Channel’** descriptive summaries, which means that there are extreme/unusual spending patterns during some times of the year.
2. The varieties: ‘**Fresh’ , ‘Milk’** and **‘Grocery’** have greater Range (differences between their minimum and maximum values). Hence, the datapoints for these varieties are spread out largely.This implies that the customers buying such items have inconsistent spending patterns annually.
3. Among the 6 varieties of items across ‘**Channel’** and ‘**Region’** , the variety: ‘**Fresh’** has the highest mean as well as median values amidst all other varieties.Hence, this variety has the highest consumption.
4. Since the Inter-Quartile Range (Difference between 75th and 25th percentiles) is relatively very less for the varieties: ‘**Frozen’ , ‘Detergents\_Paper’ , ‘Delicatessen’ ,** the data points for these 3 varieties are more bunched up around the mean, i.e., they are very less spread out.Hence, these varieties show the least inconsistent behaviour.
5. On plotting histogram, it was observed that all 6 varieties have right-skewed/right-tailed distributions i.e., None of the distributions are symmetric about their mean values.Hence, most of the spendings are on the higher side as compared to their mean spending.



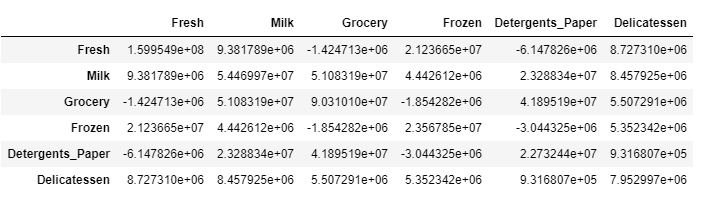


The same can be seen here on calculating skewness across all 6 columns.Positive values indicate right-skewed distributions.The column ‘**Delicatessen’** is highly positively skewed, as compared to other varieties.

1. There are some pairs of varieties which show the maximum co-variance.

For example, ‘**Fresh’** has highest degree of covariance with ‘**Frozen’** ,which Implies that customer generally spending on ‘**Fresh’** also prefers **‘Frozen’.**

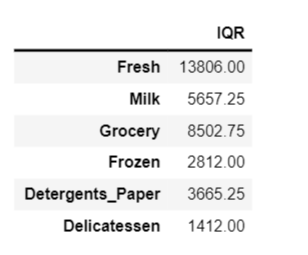
Also, ‘**Fresh’** has lowest degree of covariance with ‘**Detergents\_Paper’** ,which Implies that customer generally spending on ‘**Fresh’** is very less likely to prefer ‘**Detergents\_Paper’.**



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

The 3 feasible descriptive measures of variability are: IQR, Standard deviation, Coefficient of Variance.

In order to find the most/least inconsistent behaviour, we need to determine the nature of dataset first.



The distribution is right skewed, as skewness of all columns comes out as positive.

For a skewed dataset, having larger spread, IQR(=75th percentile -25th percentile) is a better measure to determine behaviour. Higher the IQR, more is the **inconsistent behaviour for that column.**

From the IQR calculation for all 6 item datasets, it can be concluded that:

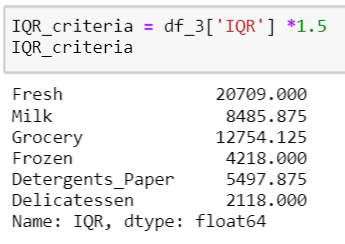
Item having **most** **inconsistent** behaviour **:** ‘**Fresh’** (IQR=13806.00)

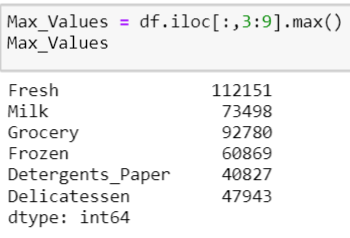
Item having **least** **inconsistent** behaviour **:** ‘**Delicatessen**’ (IQR = 1412.00)

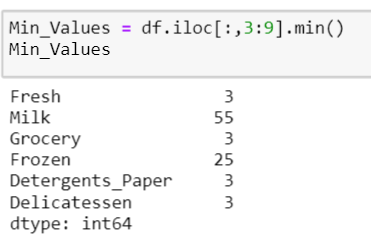
* 1. **Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.**

**As already concluded from descriptive statistics summary,** There are large differences between 75th %tile values and the max values for all 6 varieties across ‘**Region’** as well as ‘**Channel’** descriptive summaries, which means that there are Outliers/Extreme values in our dataset.

**Also, Determined using the formula: Max value > IQR\*1.5 , Min value < IQR\*1.5**



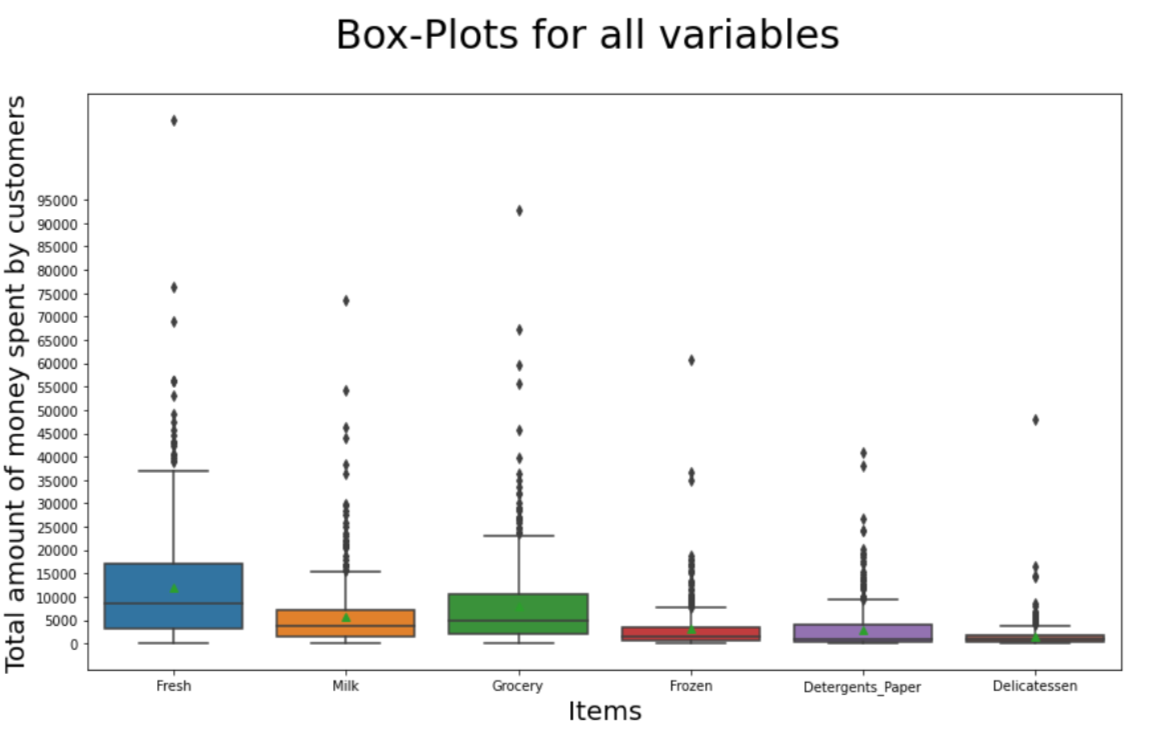




It can be concluded from the calculations that outliers are present in our dataset.

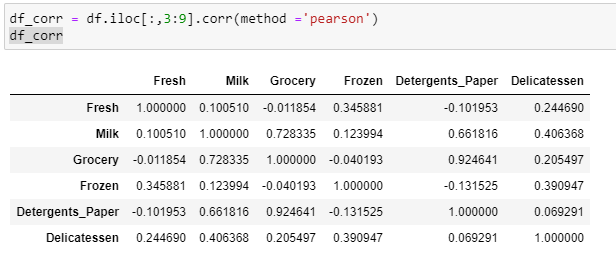
Same is illustrated using Box-Plot below:  
It shows Mean, 25%(Q1), Median(Q2), 75%(Q3) values for each column.

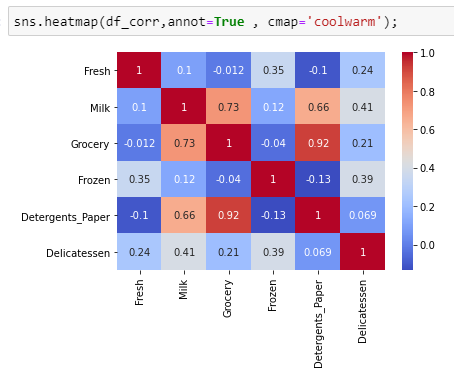
The upper & lower whiskers show the maximum & minimum values for that column respectively.The dotted values above the Upper Whisker represent the ‘Outliers’ or Extreme values.



* 1. **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**

**Calculated the correlation among the 6 varieties first, then plotted the correlation dataframe using Heatmap.**





**Business Insights :**

1. All varieties have extreme values, which clearly shows the unusual spending patterns across the ‘Region’ and ‘Channel’ . Also,Mean values for all the 6 varieties are greater than the median values(50th percentile) across ‘**Region**’ as well as ‘**Channel’** descriptive summaries. This implies that most of the spending amount data is more than the mean values.

**It is recommended that :** The annual data needs to be further drilled down or segregated on the basis of seasons, months or quarter-wise because the outliers could be due to weather changes, festivals or any other factor. For example,

* During festivities , the demand for ‘Grocery’, ‘Delicatessen’ increases.
* In the current COVID-19 times, demand for items under ‘Detergents\_Paper’ like sanitizer,masks have increased drastically.

Hence, the wholesale stores can be more prepared to stock their inventories during such times, by keeping buffer stocks to handle unusual demands.

1. Some varieties have strong/weak correlation between them, as shown in the heatmap plot above , For example,

* ‘Detergents\_Paper’ and ‘Grocery’ have a coefficient variation of 0.92
* ‘Grocery’ and ‘Milk’ have a coefficient variation of 0.73
* ‘Grocery’ and ‘Fresh’ have a coefficient variation of -0.012
* ‘Grocery’ and ‘Frozen’ have a coefficient variation of -0.04

As concluded from the descriptive summary, the spending on items : ‘Grocery’ , ‘Milk’ and ‘Fresh’ is higher as compared to other items.

**It is recommended that :** The sales of items having lesser correlations with the items: ‘Grocery’ , ‘Milk’ or ‘Fresh’ can be increased by giving extra discounts or additional offers . Hence, cross selling can be promoted by providing offers on such combinations, For example,

**On purchase of** ‘Frozen’ or ‘Fresh’ along with ‘Grocery’ should have additional offers.

**Problem #2**

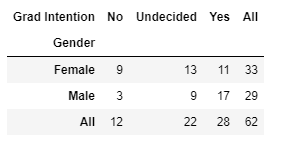
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the *Survey* data set).

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

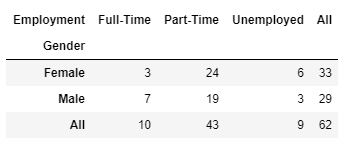
**2.1.1. Gender and Major**



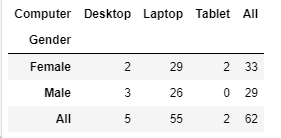
**2.1.2. Gender and Grad Intention**



**2.1.3. Gender and Employment**



**2.1.4. Gender and Computer**



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

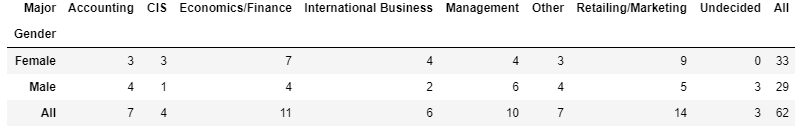
**2.2.1. What is the probability that a randomly selected CMSU student will be male?**

Prob\_RandomStudentIsMale = MaleCount/TotalCount = 29/62

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

Prob\_RandomStudentIsFemale = FemaleCount/TotalCount = 33/62

**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.**

Based on the above shown table, Following are Conditional probabilities for males:

**Prob\_MaleAccounting =** CountOf\_MalesIn\_Accounting / **TotalMaleCount = 4/29**

**Prob\_MaleCIS =** CountOf\_MalesIn\_CIS / **TotalMaleCount = 1/29**

**Prob\_MaleEconomics\_Finance =** CountOf\_MalesIn\_Economics\_Finance / **TotalMaleCount = 4/29**

**Prob\_MaleInternationalBusiness =** CountOf\_MalesIn\_**InternationalBusiness** / **TotalMaleCount = 2/29**

**Prob\_MaleManagement =** CountOf\_**MalesIn\_Management** / **TotalMaleCount = 6/29**

**Prob\_MaleOther =** CountOf\_**MalesIn\_Other**/ **TotalMaleCount = 4/29**

**Prob\_MaleRetailing\_Marketing =** CountOf\_**MalesIn\_Retailing\_Marketing** / **TotalMaleCount = 5/29**

**Prob\_MaleUndecided =** CountOf\_**Males\_Undecided**/ **TotalMaleCount = 3/29**

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

Based on the above shown table, Following are Conditional probabilities for males:

**Prob\_****FemaleAccounting =** CountOf\_FemalesIn\_Accounting / **Total**Fem**aleCount ­= 3/33**

**Prob\_FemaleCIS =** CountOf\_FemalesIn\_CIS / **Total**Fem**aleCount = 3/33**

**Prob\_FemaleEconomics\_Finance =** CountOf\_FemalesIn\_Economics\_Finance / **Total**Fem**aleCount = 7/33**

**Prob\_FemaleInternationalBusiness =** CountOf\_FemalesIn\_**InternationalBusiness** / **Total**Fem**aleCount = 4/33**

**Prob\_FemaleManagement =** CountOf\_Fem**alesIn\_Management** / **Total**Fem**aleCount = 4/33**

**Prob\_FemaleOther =** CountOf\_Fem**alesIn\_Other**/ **Total**Fem**aleCount = 3/33**

**Prob\_FemaleRetailing\_Marketing =** CountOf\_Fem**alesIn\_Retailing\_Marketing** / **Total**Fem**aleCount = 9/33**

**Prob\_FemaleUndecided =** CountOf\_Fem**ales\_Undecided**/ **Total**Fem**aleCount = 0/33**

**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.**

**P(A) :** **Probability That a randomly chosen student is a male**

**P(B) : Probability That a randomly chosen student intends to graduate**

**= 1 - (Probability That a randomly chosen student is undecided)**

**= 1 – 3/29 = 26/29**

As per the addition rule, when the events are mutually exclusive,

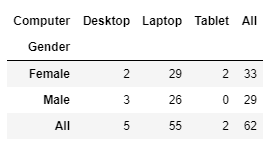
**P(AUB) = P(A)+P(B)**

Applying the formula, we get :

**P(AUB) =** 29/62 \* (26/29) = 26/62

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

**Following is the Created contingency table of Gender and Computer :**



**P(A) :** **Probability That a randomly chosen student is a Female**

**P(B) : Probability That a female doesn’t have a laptop**

**P(A∩B) : Probability that a randomly selected student is a female and does NOT have a laptop.**

From the contingency table shown above, it’s clear that there are 4 females out of 33 who don’t have a laptop

Hence, **P(A∩B)** = 4/62

**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 either a male or has full-time employment?**

**P(A)** :Probability that a Randomly chosen Student Is Male = 29/62

**P(B)** : Probability that a Randomly chosen Student **has full-time employment** = 10/62

**P(A∩B)** :Probability **that a randomly chosen student is a male AND has full-time employment** = 7/62

**P(AUB)** : Probability **that a randomly chosen student is a male OR has full-time employment**

Applying the Addition Rule, we get :

As per the addition rule, when the events aren’t mutually exclusive,

**P(AUB) = P(A)+P(B)–P(A∩B)**

Applying the formula, we get :

**P(AUB) =** 29/62 + 10/62 – 7/62 = 32/62

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

**P(A)** :Probability that a Randomly chosen Student Is Female = 33/62

**P(B)** : Probability of Majoring In International Business Or Management

= Probability of Majoring In International Business + Probability of Majoring In Management

= 4/62 + 4/62 = 8/62

**P(B/A)** : Conditional Probability of Majoring In International Business Or Management Given the Randomly chosen student is Female

When events are not independent,

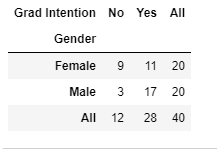
**P(B/A) = P(A∩B) / P(A)**

Applying the formula, we get :

**P(B/A) =** (8/62 ) / (33/62) = 8/33

**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?**

**Following is the Created contingency table of Gender and Intent to Graduate :**



**In order to prove that graduate intention and being female are independent events,**

**Let us assume that:**

**The graduate intention and being female are independent events.Then the Multiplication Rule for Independent events must also hold true.Hence.**

**Prob\_**RandomStudentIs**FemaleANDProb\_RandomStudentHavingGradIntention = Prob\_**RandomStudentIs**Female \* Prob\_RandomStudentHavingGradIntention**

**R.H.S = 20/40 \* 28/40 = 7/20 = 0.35**

**Now, calculating L.H.S using contingency table ,**

**Prob\_**RandomStudentIs**FemaleANDProb\_RandomStudentHavingGradIntention = 11/40 = 0.275**

Since L.H.S  ≠ R.H.S , **We can conclude that graduate intention and being female are not independent 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?**

**CountOfGpaLessThan3 has been calculated using python filter condition on the dataframe:**

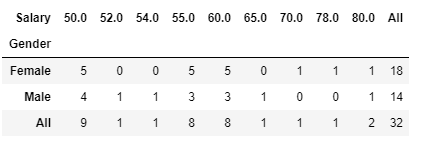
**Length of Dataframe subset where GPA is less than 3 = 17**

**Now,**

**Prob\_ChosenStudentHasGPAlessThan3 = CountOfGpaLessThan3/TotalCount = 17/62**

**2.7.2. a) Find the conditional probability that a randomly selected male earns 50 or more.**

**Entries where male earns 50 or more & where a female earns 50 or more have been calculated using python filter condition and the contingency table has been created:**



**Using the output of contingency table, I have calculated marginal probabilities using another simplified table :**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Salary>=50** | **Salary<50** | **All** |
| **Female** | 18 | 15 | 33 |
| **Male** | 14 | 15 | 29 |
| **All** | 32 | 30 | 62 |

P(A) :Probability that the randomly selected person earns 50 or more

P(B) : Probability that a randomly selected person is male

P(A/B) : Probability that the selected person earns 50 or more Given the person is male.

**P(A/B) = P(A∩B) / P(B) ,**

P(A∩B) : Probability that the selected person earns 50 or more AND the person is male

**From the above shown table, it is clear that there are total 14 males Earning 50 Or more , among a total of 62 people.**

Hence, P(A∩B) = 14/62

P(B) = Number of males / Total Number of undergraduate students attending CMSU

Applying the formula, we get :

**P(A/B)**= (14/62) / (29/62) = 14/29

**2.7.2 b) Find the conditional probability that a randomly selected female earns 50 or more.**

Similarly, for

P(A/B) : Probability that the selected person earns 50 or more Given the person is female.

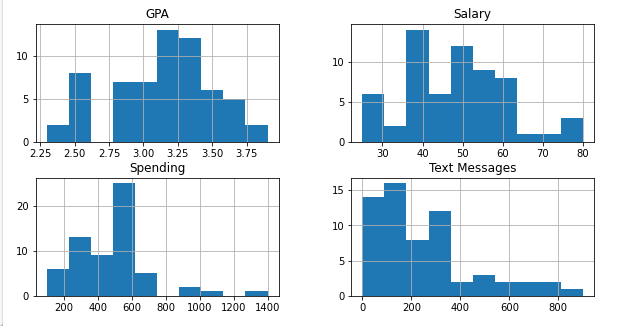
**P(A/B) =** (18/62) / (33/62) = 18/33

**2.8.a) 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.**

**The properties of a random sample having normal distribution are:**

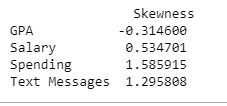
1. **The data distribution has to be symmetric, It cannot be left or right skewed.**
2. **The mean, mode, median are all equal.**
3. **Half of the distribution count is less than the mean & the other half is greater than the mean.**

**Having said that, following are the histograms plotted for all four continuous variables:**



**OBSERVATIONS FROM HISTOGRAM:**

1. **‘GPA’ follows close to normal kind of distribution , appears a bit left-tailed/negatively skewed. The peak is around 3.10 , but the distribution extends backwards to lower values of left side, as compared to the higher values of right side.**
2. **‘Salary’ follows close to normal kind of distribution , appears a bit right-tailed/positively skewed. The peak is around 36 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.**
3. **‘Spending’ follows a positively skewed distribution , is right-tailed/positively skewed. The peak is around 440 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.**
4. **‘Text Messages’ follows a positively skewed distribution , is right-tailed/positively skewed. The peak is around 190 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.**

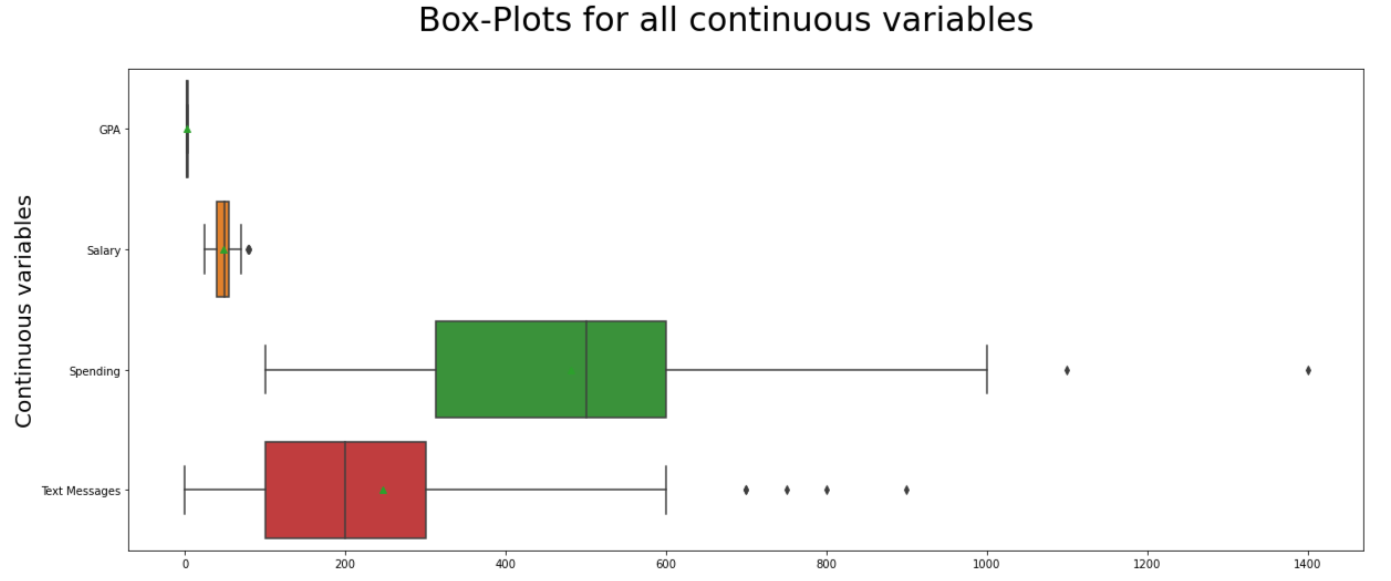


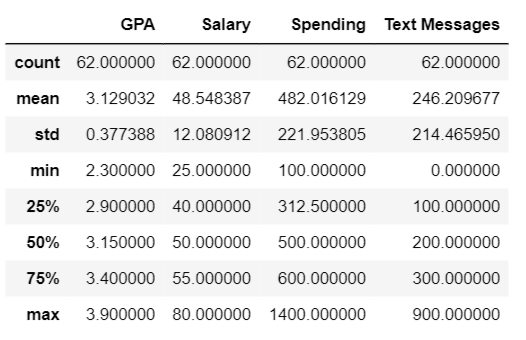
**Same can be concluded upon calculation of Skewness parameter values for all 4 columns.**

**2.8.b)Write a note summarizing your conclusions for this whole Problem 2**.

1. Negative skewness in ‘**GPA’** shows that very few undergraduates are scoring GPA below average.This implies that most of the students are performing well.

2. There is huge difference between **‘Salary’** and ‘**Spending’** of students.Also, the salary obtained by the students is almost same in numbers, but their spending pattern is quite different comparatively.





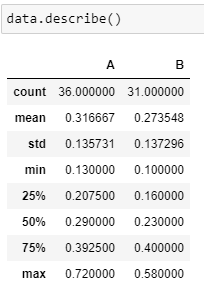
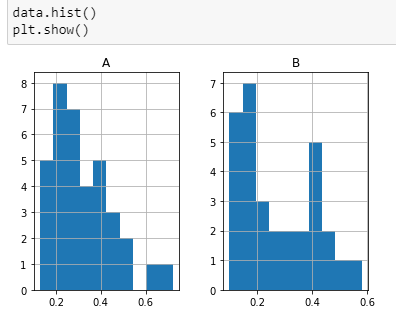
**Problem #3**

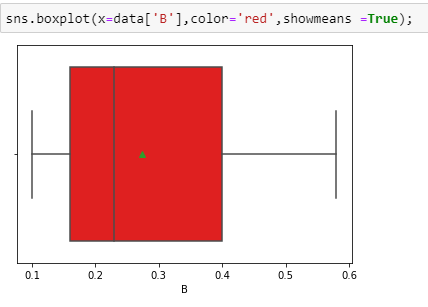
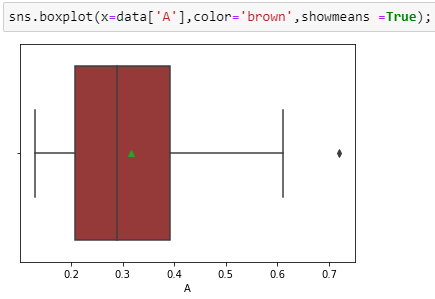
An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging.   In some cases, excessive moisture can cause the granules attached to the shingles for texture and 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 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](https://olympus.greatlearning.in/courses/36669/files/2044162/download?verifier=5qNd7kr9QL6O6SDHPsrzTxD7XD3yIMvHbVRnyGTg&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

* 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.**

**Analyzing the dataset using describe() and by plotting Histogram & box-plot, we get the following :**



* Mean and Median Values of each sample are not much different.
* The 'A' sample data looks more symmetrically distributed, whereas the 'B' sample data looks right skewed.
* But since the population standard deviation (Sigma) is unknown, we have to use a Tstat test.
* Since we need to check if mean moisture contents in both types of shingles are within the permissible limits individually, we'll perform individual **1 sample T-Tests** using **ttest\_1samp** method from scipy library for both "A" & "B".

**FOR ‘A’:**

**Step 1: Define null and alternative hypotheses**

* Null hypothesis states that mean moisture content 𝜇 =< 0.35, Ho => 𝜇 =< 0.35
* Alternative hypothesis states that the mean moisture content 𝜇 > 0.35, 𝐻a => 𝜇 > 0.35

**Step 2: Decide the significance level**

* The level of significance (Alpha ) = 0.05.

**Step 3: Identify the test statistic**

* tstat: -1.4735046253382782
* p-value for one-tail = p-value/2 : 0.07477633144907513

Hence, We have no evidence to reject the null hypothesis since p value > Level of significance

**FOR ‘B’:**

**Step 1: Define null and alternative hypotheses**

* Null hypothesis states that mean moisture content 𝜇 =< 0.35, Ho => 𝜇 =< 0.35
* Alternative hypothesis states that the mean moisture content 𝜇 > 0.35, 𝐻a => 𝜇 > 0.35

**Step 2: Decide the significance level**

* The level of significance (Alpha ) = 0.05.

**Step 3: Identify the test statistic**

* tstat : -3.1003313069986995
* p-value for one-tail = p-value/2 : 0.0020904774003191826

Hence, We have evidence to reject the null hypothesis since p value < Level of significance

**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?**

* We use the **scipy.stats.ttest\_ind** method from scipy library to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

**Step 1: Define null and alternative hypotheses**

* Null hypothesis states that the population mean moisture content for A & B are equal
* Alternative hypothesis states that the mean moisture content for A & B are not equal
* 𝐻0: 𝜇𝐴 - 𝜇𝐵 = 0 i.e.,  𝜇𝐴 = 𝜇𝐵
* 𝐻𝐴: 𝜇𝐴 - 𝜇𝐵 ≠ 0 i.e.,  𝜇𝐴 ≠ 𝜇𝐵

**Step 2: Decide the significance level**

* The level of significance (Alpha\_level) = 0.05 , as this is the default value of Confidence-level when not provided. The population standard deviation is also not known.

**Step 3: Identify the test statistic**

Tstat: 1.2896282719661123

p-value: 0.2017496571835306

Since p-value> Alpha\_level,

We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis.

Hence, We conclude that the population mean for shingles A and B are equal.

**ASSUMPTIONS MADE:**

* This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.
* Before the test for equality of means is performed, we are going to assume that the variance is equal and then compute the necessary statistical values.
* The samples are assumed to be following a normal distribution.