Statistical Methods for Decision Making (SMDM)

Group Assignment

GREAT LAKES INSTITUTE OF MANAGEMENT

greatlearning
Learning for Life

Team Name: SMDM Group6

Members:

1. Nikhil Panchal
2. Madhusudan Shiva
3. Rajesh Chandran
4. Narendra Rai
5. Kunal Sharma

Submission Dt:

Sep 28th, 2020

About Us

This section will cover brief introduction about us. We are seasoned professionals with diverse experience.

Nikhil

Overall **13+ years of cross functional experience** in FMCG and Alcobev industry. Working with **Diageo PLC** since 2017, looking after **Data Analytics CoE for Global Audit & Risk department**. Before to that worked with **Coca-Cola for 10 years** and completed stint in operations, supply chain (direct & indirect), Master data maintenance (SAP cross functional role) and Data analytics (Internal Audit).

Madhu

An experienced **business analytics professional** helping organisation to analyse KPIs, IT data, OT data, risk and control data, etc. and provide insights through reports and dashboards to take informed and timely decisions. Helps leadership to build client and top-level power point presentations. Also develop and maintains share point sites for management of information.

Rajesh

A **Product and Technology leader** with significant Product Management, Professional Services & Technology Consulting experience in **Fortune 500 Organizations**. He has a proven track record of developing roadmaps & building high performance teams. Rajesh comes with strong Product Management background launching SaaS/on-prem software products across DC/Cloud, Mobility & Video domain(s).

Kunal

Around **4 years** of experience in the field of data analysis. he works in capacity of a **subject matter expert**. Help clients collect, process, and to some extent interpret data on key macroeconomic variables for world economies and for some projects key banking and financial indicators.

Narendar

Business Analyst in Cisco having **13 years of total experience** in analysis like Gross margin, manufacturing cost, booking and revenue etc. and help to build the dashboard (in MicroStrategy) and tools like Hyperion & Business objects which are used for SEC reporting, Earnings calls and planning and strategy.

Table of Contents:

Problem1:

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).

Remarks: All the questions of problem1 which explained here, are also performed in python as well. Please refer python notebook "SMDM_Group_Assignment_GRP6_Problem1".

1.1 Use methods of descriptive statistics to summarize data

Background: Descriptive statistics, as the name suggests, is used to describe basic features of a data set. In layman's terms 'descriptive statistics' gives an introduction to the data set by providing a small summary. Measures of central tendencies like mean, median, mode, standard deviation etc. are most commonly used measures in descriptive statistics to provide quantitative summary to a large data set.

EDA (Exploratory Data Analysis): We used 'Exploratory Data Analysis,' before summarizing the data set using descriptive statistics, to learn about the types of variables we are dealing with and whether or not there are any missing values in the data set. We learnt that the data set has 440 entries, 9 columns and no null values. *We found a categorical variable been represented as an integer and changed its field to 'object.'*

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
Buyer               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: 31.1+ KB
```

Figure ③ (Data Types)

```
df.Buyer = df.Buyer.astype('object') ##
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
Buyer               440 non-null object
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(6), object(3)
memory usage: 31.1+ KB
```

Figure ③ converting Buyer from int to obj.

```
df.isnull().sum() ## c
```

```
Buyer               0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

Figure ③ No null data in the dataset

Data Summary:

	count	mean	std	min	25%	50%	75%	max
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

We know from the problem statement that this dataset gives information on annual spend of a wholesale distributor on **6 items in three regions - Lisbon, Oporto, Other - through two channels of distribution – Hotel and Retail.**

We can infer the following looking at the data summary above:

- There ***are 440 total rows*** in this dataset.
- There are **6 numerical variables in the dataset - Fresh, Milk, Grocery, Frozen, Detergent_Paper, and Delicatessen** which helps us summarize the spending trend.
- We learn about the spending behavior on each item by looking into the average, the minimum and the maximum amount spent on an item.
- For all items, **mean is greater than 50% (i.e. median)** which means we have skewness in the data.

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

Background: We looked at the total expenditure and the maximum and the minimum amount spent on each item in different regions. (i.e. Lisbon, Oporto, Other)

We made the same observations for the channel of distribution. (i.e. Hotel and Retail)

We then looked into the spending behavior on these items among regions and channels together as well. (i.e. region and channel)

To get the total spend, we added up all 6 items (i.e. fresh, milk, grocery, frozen, detergent paper and delicatessen) spends together.

	Buyer	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend	Region_Channel
0	1	Retail	Other	12669	9656	7561	214	2674	1338	34112	Other - Retail
1	2	Retail	Other	7057	9810	9568	1762	3293	1776	33266	Other - Retail
2	3	Retail	Other	6353	8808	7684	2405	3516	7844	36610	Other - Retail
3	4	Hotel	Other	13265	1196	4221	6404	507	1788	27381	Other - Hotel
4	5	Retail	Other	22615	5410	7198	3915	1777	5185	46100	Other - Retail

Figure ④ two new columns added "Total Spend & Region_Channel"

Our observations, accompanied by supporting tables and graph, are as follows:

Region- wise summary:

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Lisbon	854,833	422,454	570,037	231,026	204,136	104,327	2,386,813
Oporto	464,721	239,144	433,274	190,132	173,311	54,506	1,555,088
Other	3,960,577	1,888,759	2,495,251	930,492	890,410	512,110	10,677,599

Figure ⑤ Region wise total spend summary

The below table shows the region with the highest total spend across items. We can see that **Region: 'Other' has the highest total spend** (*i.e. 10.7 million*) on items across all regions.

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Other	3,960,577	1,888,759	2,495,251	930,492	890,410	512,110	10,677,599

Figure ⑥ Region with highest spend

The below table shows the region with the lowest total spend across items. We can see that **Region: 'Oporto' has the lowest total spend** (*i.e. 1.6 million*) on items across all regions.

Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Oporto	464,721	239,144	433,274	190,132	173,311	54,506	1,555,088

Figure ⑦ Region with lowest spend

Channel- wise summary:

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Hotel	4,015,717	1,028,614	1,180,717	1,116,979	235,587	421,955	7,999,569
Retail	1,264,414	1,521,743	2,317,845	234,671	1,032,270	248,988	6,619,931

Figure ⑧ Channel wise total spend summary

The below table shows the channel with the highest total spend on items. We can see that **Channel: 'Hotel' has the highest total spend** (*i.e. 8 million*)on items.

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Hotel	4,015,717	1,028,614	1,180,717	1,116,979	235,587	421,955	7,999,569

Figure ⑨ Channel with highest spend

The below table shows the channel with the lowest total spend on items. We can see that **Channel: 'Retail' has the lowest total spend** (*i.e. 6.6 million*)on items.

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Retail	1,264,414	1,521,743	2,317,845	234,671	1,032,270	248,988	6,619,931

Figure ⑩ Channel with lowest spend

Summary on combination of Region and Channel:

Region_Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Lisbon - Hotel	761,233	228,342	237,542	184,512	56,081	70,632	1,538,342
Lisbon - Retail	93,600	194,112	332,495	46,514	148,055	33,695	848,471
Oporto - Hotel	326,215	64,519	123,074	160,861	13,516	30,965	719,150
Oporto - Retail	138,506	174,625	310,200	29,271	159,795	23,541	835,938
Other - Hotel	2,928,269	735,753	820,101	771,606	165,990	320,358	5,742,077
Other - Retail	1,032,308	1,153,006	1,675,150	158,886	724,420	191,752	4,935,522

Figure ⑪ Region & Channel wise total spend summary

The below table shows the region + channel combination with the highest total spend on items. We can see that the maximum amount is spent on these items in **region 'Other' through channel 'Hotel.'** (*i.e. 5.7 million*)

Region_Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Other - Hotel	2,928,269	735,753	820,101	771,606	165,990	320,358	5,742,077

Figure ⑫ Region & Channel with highest spend

The below table shows the region + channel combination with the lowest total spend on items. We can see that the minimum amount is spent on these items in **region 'Oporto' through channel 'Hotel.'** (*i.e. 0.7 million*)

Region_Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_spend
Oporto - Hotel	326,215	64,519	123,074	160,861	13,516	30,965	719,150

Figure ⑬ Region & Channel with lowest spend

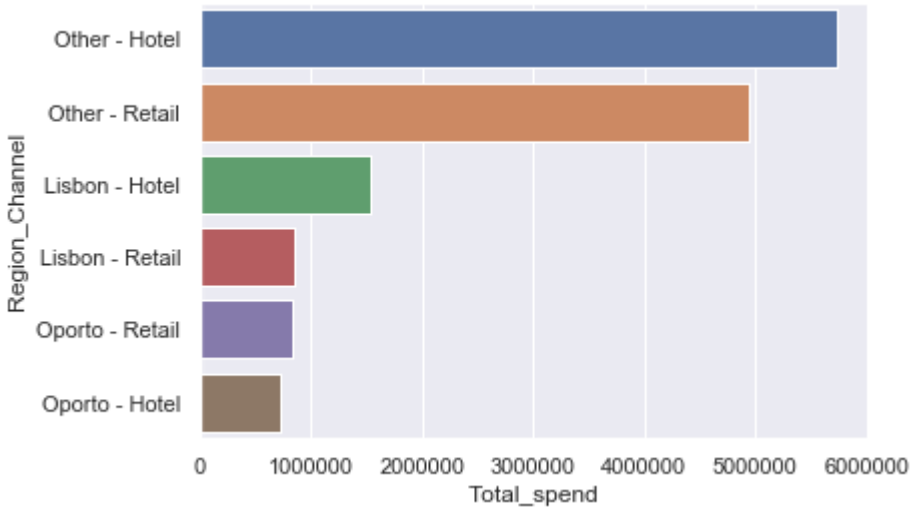


Figure ⑭ Visual Summary around Region & Channel wise total spend

Conclusion: Mentioned below is the summary around region wise, channel wise and Region + Channel wise highest and lowest spend based on the given dataset.

Highest Spend: *[assumed to be in $ currency for this problem as default currency isn't stated]:*

- **Region: Other** has the highest spend with amount of $ 10.7 million.
- **Channel: Hotel** has the highest spend with amount of $ 8 million.
- **Region & Channel:** When we combine region and channel **OTHER & HOTEL** has the highest spend with amount of $ 5.7 million.

Lowest Spend: *[assumed to be in $ currency for this problem as default currency isn't stated]:*

- **Region: Oporto** has the lowest spend with amount of $ 1.6 million.
- **Channel: Retail** has the lowest spend with amount of $ 6.6 million.
- **Region & Channel:** When we combine region and channel **OPORTO & HOTEL** has the lowest spend with amount of $ 0.7 million.

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

Background: In the dataset, there are 6 varieties of items given (i.e. fresh, milk, grocery, frozen, detergent paper and delicatessen. To observe each items behavior across region and channel, we will use diagonal pair plot.

- Region wise check the spending behavior for all items
- Channel wise check the spending behavior for all items

Region wise check the spending behavior:

- **Item: Fresh**

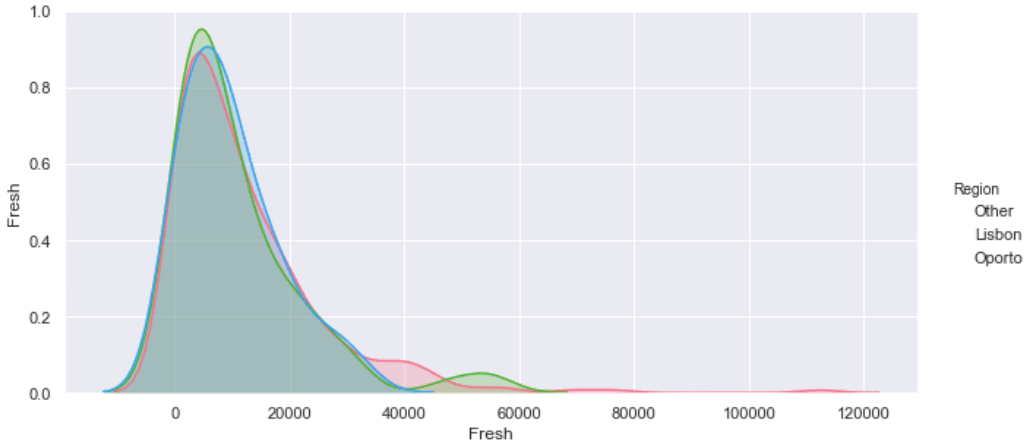


Figure ⑮ Region wise pair plot for item : Fresh

Spend behavior for item 'Fresh' across region: As we can see in the graph that most buyers are spending approx. 10k on this item. **Hence, it can be concluded that item fresh displays similar behavior across regions.**

- **Item: Milk**

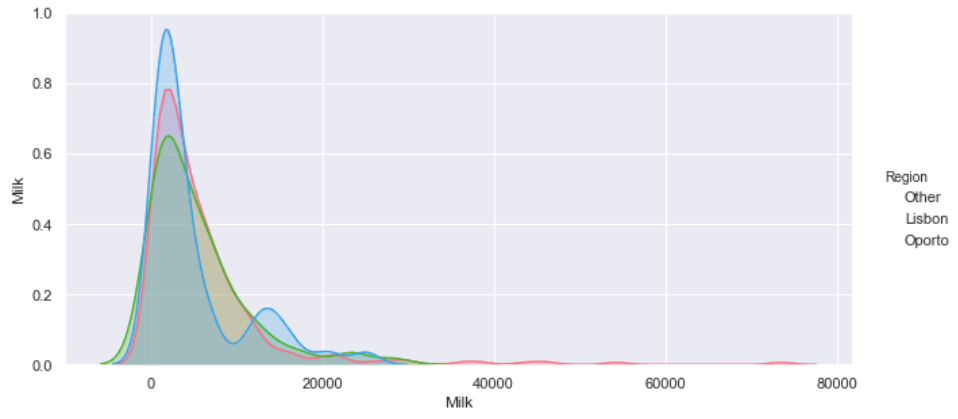


Figure ⑯ Region wise pair plot for item : Milk

Spend behavior for item 'Milk' across region: Buyers spend somewhere between 6k and 10k on milk across region. **Hence, it can be concluded that the spending behavior for item milk is not similar across regions.**

- **Item: Grocery**

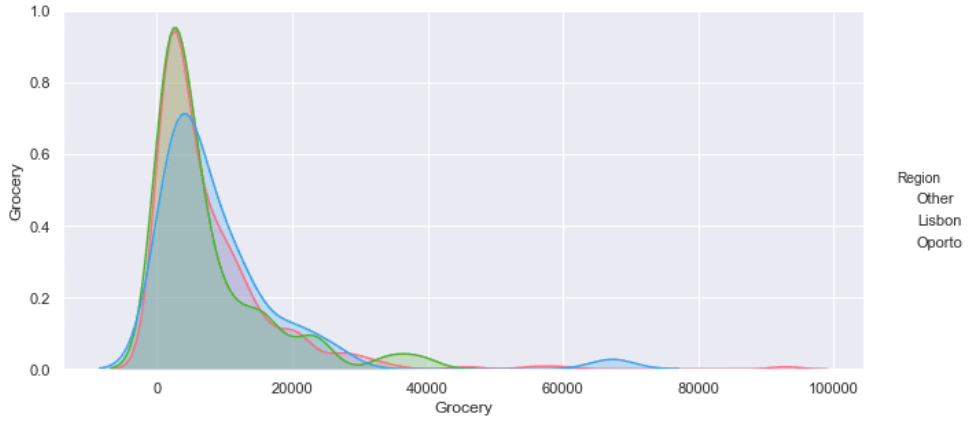


Figure ⑰ Region wise pair plot for item : Grocery

Spend behavior for item 'Grocery' across region: We can see in the graph that most buyers are spending in the range of 7k to 10k on grocery. **Hence, it can be concluded that the spending behavior for item grocery is not similar across regions.**

- **Item: Frozen**

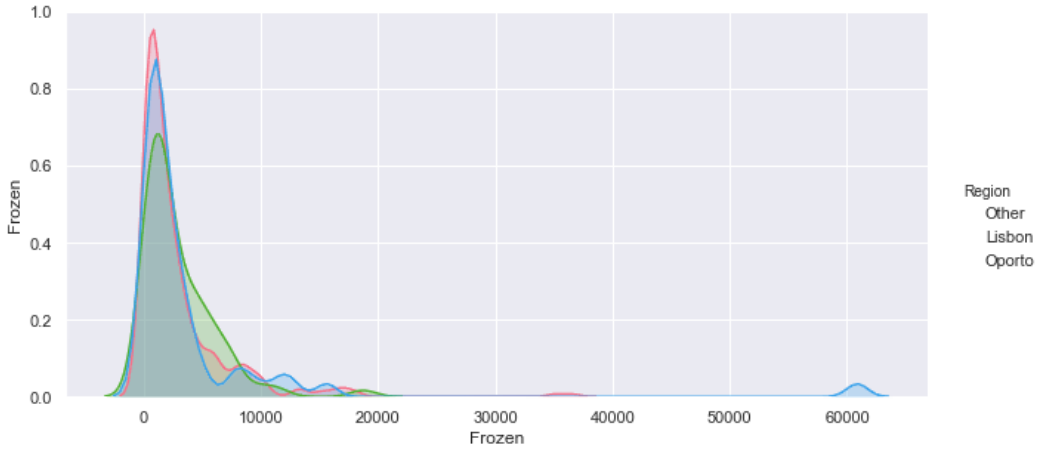


Figure ⑱ Region wise pair plot for item : Frozen

Spend behavior for item 'Frozen' across region: The graph above shows that majority of buyers are spending between 7k and 9k on this item across regions. **Hence, it can be concluded that the spending behavior for item frozen is not similar across regions.**

- **Item: Detergent Paper**

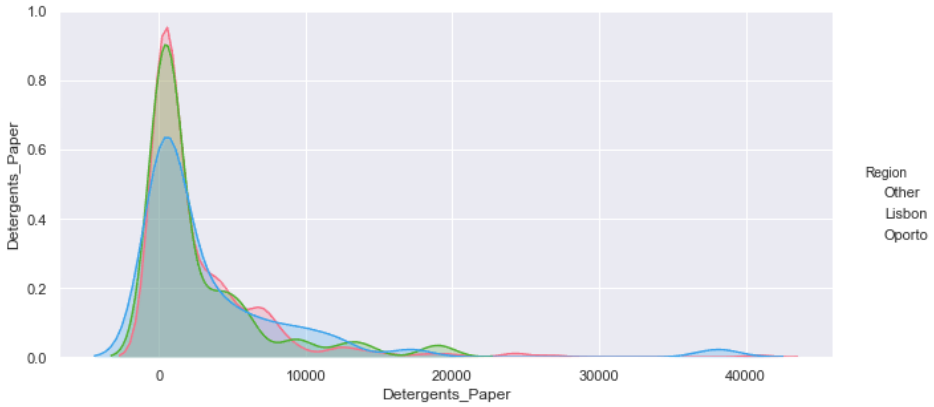


Figure ⑲ Region wise pair plot for item : Detergent Paper

Spend behavior for item 'Detergents Paper' across region: As we can see in the graph that most buyers are spending 6k to 10k on this item across regions. **Hence, it can be concluded that the spending behavior for item Detergents Paper is not similar across regions**.

- **Item: Delicatessen**

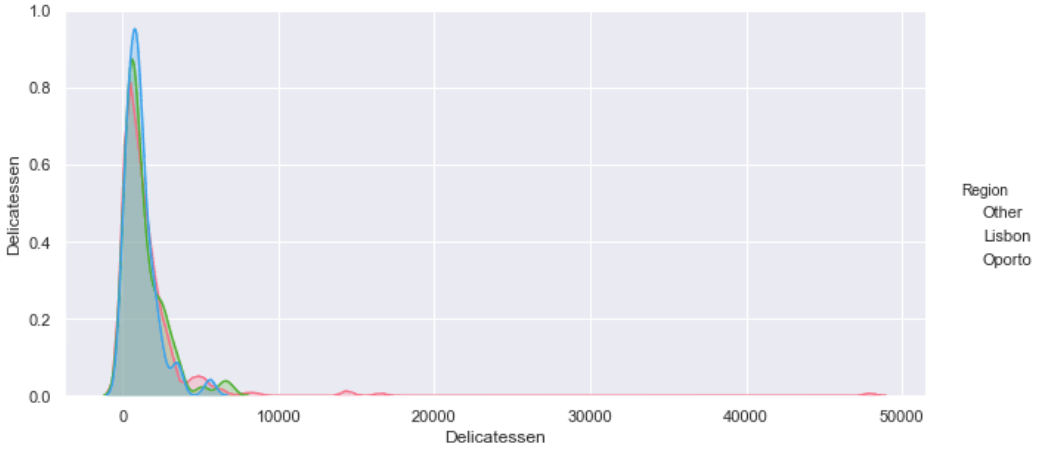


Figure ⑳ Region wise pair plot for item : Delicatessen

Spend behavior for item 'Delicatessen' across region: As we can see expenditure of close to 10k in all three regions for the item. **Hence, the graph above shows that spending behavior for item Delicatessen is almost similar across regions.**

Conclusion: Based on the pair plots it can be concluded that items **'Fresh' and 'Delicatessen' show somewhat similar behavior across the three regions**, whereas there is **considerable difference in the spending behavior on items 'Detergents Paper,' 'Frozen, 'Grocery,' and 'Milk.'**

Channel wise check the spending behavior:

- **Item: Fresh**

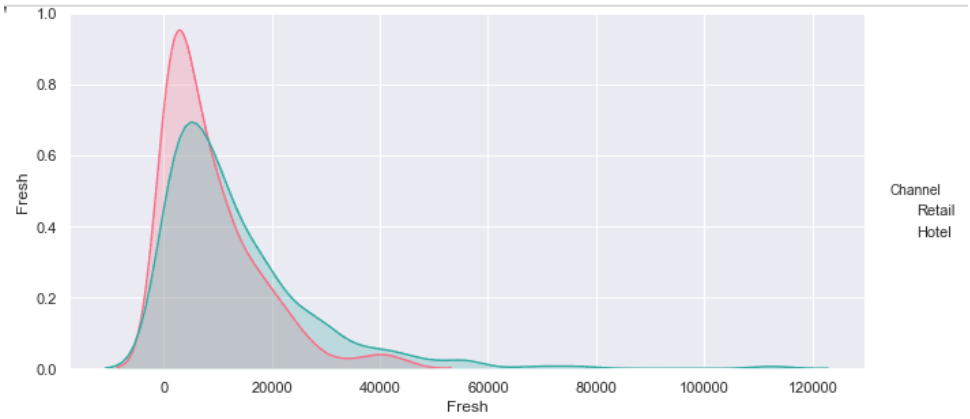


Figure 21 Channel wise pair plot for item : Fresh

Spend behavior for item 'Fresh' across channel: As we can see in the graph that most buyers are spending approximately 6k to 10k on this item. **Hence, it can be concluded that the spending behavior for item fresh is not similar across channels.**

- **Item: Milk**

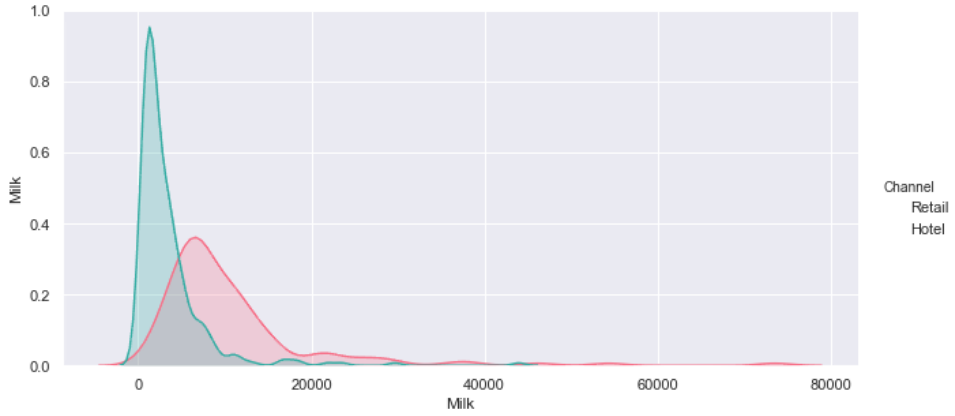


Figure 22Channel wise pair plot for item : Milk

Spend behavior for item 'Milk' channel: The graph above clearly shows considerable difference in the amount spent on milk across channels. We can see that expenditure on the item through channel hotel is approximately 9k, whereas the spending is only around 3.8k through retail channel. **Hence, it can be concluded that the spending behavior for item milk is not similar across channels.**

- **Item: Grocery**

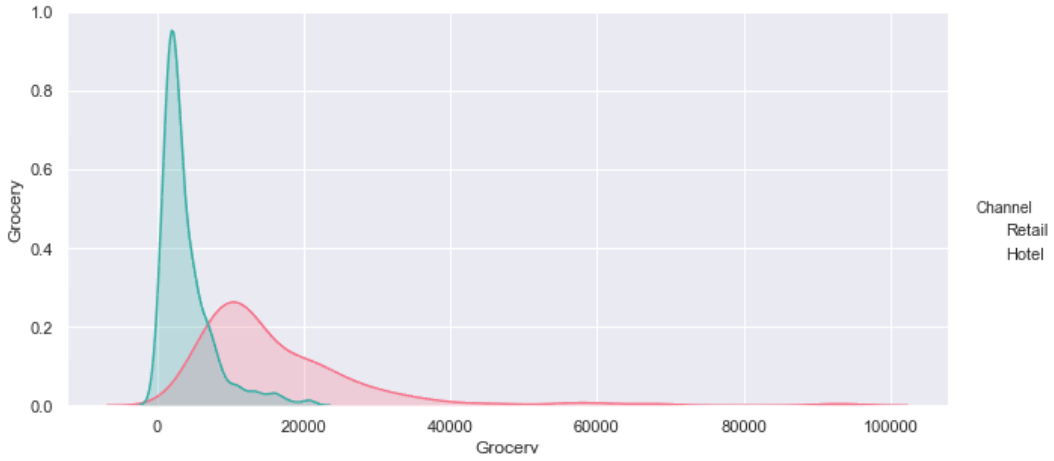


Figure 23 Channel wise pair plot for item : Grocery

Spend behavior for item 'Grocery' across channel: **The graph shows considerable difference in spending behavior on groceries across channels.** The trend is similar to that of the spending behavior on item milk – considerably higher amount spent on the item through channel hotel in comparison to the amount spend through channel retail.

- **Item: Frozen**

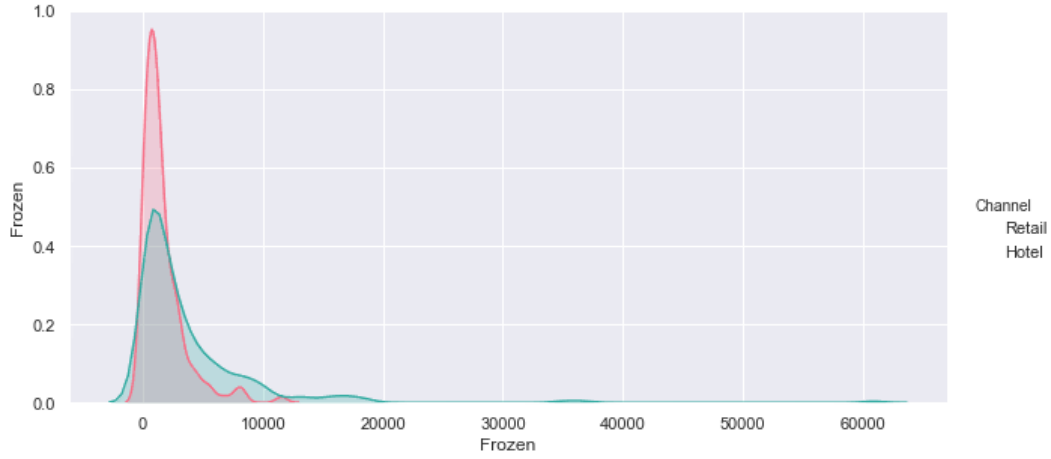


Figure 24 Channel wise pair plot for item : Frozen

Spend behavior for item 'Frozen' across channel: There is a difference of approximately 5k in the amount spent on the item across channels. **It can therefore be concluded that the spending behavior for item frozen is not similar across channels.**

- **Item: Detergent Paper**

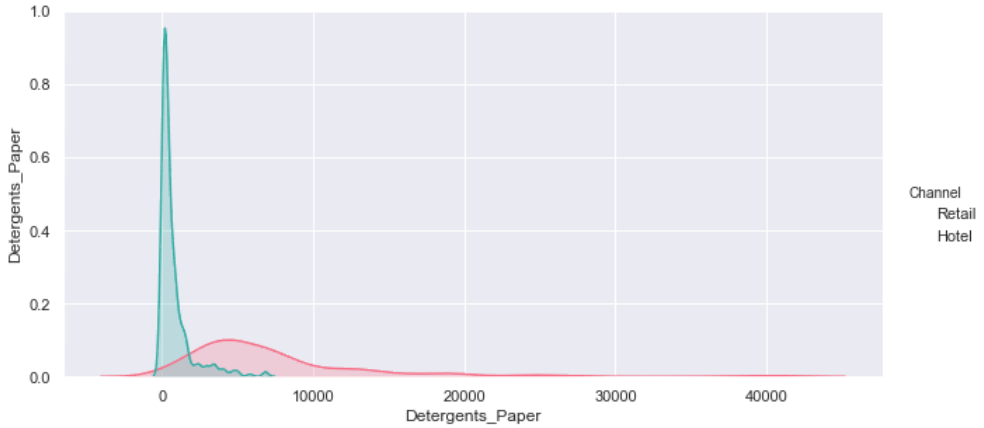


Figure 25 Channel wise pair plot for item : Detergent Paper

Spend behavior for item 'Detergents Paper' across channel: There is a huge difference in the amount spent on this item through channel hotel in comparison to channel retail. **The graph clearly shows that there is no similarity in the spending behavior for item Detergents paper across channels.**

- **Item: Delicatessen**

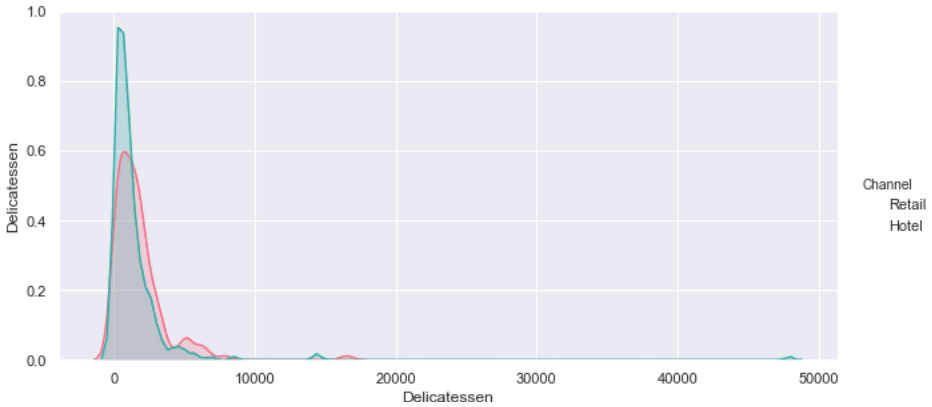


Figure 26 Channel wise pair plot for item : Delicatessen

Spend behavior for item 'Delicatessen' across channel: The graph shows that approximately 6k is spent on the item through 'Retail' and approximately 9k is spent on it through channel 'Hotel.' **It clearly shows lack of any spending similarity for 'Delicatessen' across the two channels.**

Conclusion: Looking at the amount spent by buyers on each item across the given channels **it can be concluded that the spending behavior for none of the 6 items is similar across the two channels.** The data also shows that spending on each item through 'Hotel' is comparatively more than spending on each item through 'Retail.'

1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behavior? Which items shows the least inconsistent behavior?

Background: A ***measure of variability (spread or dispersion)*** is a summary statistic that represents ***the amount of dispersion in a dataset***. measures of variability define how far away the data points tend to fall from the center (i.e. mean). A *low dispersion* indicates that the data points tend to be *clustered tightly around the center*. *High dispersion* signifies that *they tend to fall further away*.

Below are the types of measure of variability:

- Range (R)
- Interquartile Range (IQR)
- Variance
- Standard deviation (std)
- Coefficient of variation (CV)

For our calculation, we will be using range, standard deviation and coefficient of variation to find out behavior of the dataset.

Let's calculate item wise range, standard deviation and coefficient of variance and find out the inconsistency in the behavior.

- **Item: Fresh**

```
##Fresh Item Measure of variability

Mean_Fresh = df['Fresh'].mean()
Range_Fresh = df['Fresh'].max() - df['Fresh'].min()
Std_Fresh = df['Fresh'].std()
cv_fresh = df['Fresh'].std()/df['Fresh'].mean()

print('Measure of variability for fresh item are:')
print('1. Mean:' ,Mean_Fresh)
print('2. Range:' ,Range_Fresh)
print('3. Standard Deviation:' ,Std_Fresh)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_fresh))
```

```
Measure of variability for fresh item are:
1. Mean: 12000.297727272728
2. Range: 112148
3. Standard Deviation: 12647.328865076894
4. Coefficient of variation: 105.39%
```

Figure 27 Measure of variability: Fresh

The same if we calculate the mean, range, standard deviation & coefficient of variation for other items then, the below table shows the coefficient of variance for each item:

```
##Milk Item Measure of variability

Mean_Milk = df['Milk'].mean()
Range_Milk = df['Milk'].max() - df['Milk'].min()
Std_Milk = df['Milk'].std()
cv_Milk = df['Milk'].std()/df['Milk'].mean()

print('Measure of variability for Milk item are:')
print('1. Mean:' ,Mean_Milk)
print('2. Range:' ,Range_Milk)
print('3. Standard Deviation:' ,Std_Milk)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_Milk))
```

```
Measure of variability for Milk item are:
1. Mean: 5796.265909090909
2. Range: 73443
3. Standard Deviation: 7380.377174570843
4. Coefficient of variation: 127.33%
```

Figure 32 Measure of variability: Milk

```
##Grocery Item Measure of variability

Mean_Grocery = df['Grocery'].mean()
Range_Grocery = df['Grocery'].max() - df['Grocery'].min()
Std_Grocery = df['Grocery'].std()
cv_Grocery = df['Grocery'].std()/df['Grocery'].mean()

print('Measure of variability for Grocery item are:')
print('1. Mean:' ,Mean_Grocery)
print('2. Range:' ,Range_Grocery)
print('3. Standard Deviation:' ,Std_Grocery)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_Grocery))
```

```
Measure of variability for Grocery item are:
1. Mean: 7951.277272727273
2. Range: 92777
3. Standard Deviation: 9503.162828994346
4. Coefficient of variation: 119.52%
```

Figure 32 Measure of variability: Grocery

```
##Frozen Item Measure of variability

Mean_Frozen = df['Frozen'].mean()
Range_Frozen = df['Frozen'].max() - df['Frozen'].min()
Std_Frozen = df['Frozen'].std()
cv_Frozen = df['Frozen'].std()/df['Frozen'].mean()

print('Measure of variability for Frozen item are:')
print('1. Mean:' ,Mean_Frozen)
print('2. Range:' ,Range_Frozen)
print('3. Standard Deviation:' ,Std_Frozen)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_Frozen))
```

```
Measure of variability for Frozen item are:
1. Mean: 3071.931818181818
2. Range: 60844
3. Standard Deviation: 4854.673332592367
4. Coefficient of variation: 158.03%
```

Figure 32 Measure of variability: Frozen

```
##Detergents_Paper Item Measure of variability

Mean_DP = df['Detergents_Paper'].mean()
Range_DP = df['Detergents_Paper'].max() - df['Detergents_Paper'].min()
Std_DP = df['Detergents_Paper'].std()
cv_DP = df['Detergents_Paper'].std()/df['Detergents_Paper'].mean()

print('Measure of variability for Detergents Paper item are:')
print('1. Mean:' ,Mean_DP)
print('2. Range:' ,Range_DP)
print('3. Standard Deviation:' ,Std_DP)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_DP))
```

```
Measure of variability for Detergents Paper item are:
1. Mean: 2881.4931818181817
2. Range: 40824
3. Standard Deviation: 4767.8544479042
4. Coefficient of variation: 165.46%
```

Figure 32 Measure of variability: Detergent Paper

```
Mean_Delicatessen = df['Delicatessen'].mean()
Range_Delicatessen = df['Delicatessen'].max() - df['Delicatessen'].min()
Std_Delicatessen = df['Delicatessen'].std()
cv_Delicatessen = df['Delicatessen'].std()/df['Delicatessen'].mean()

print('Measure of variability for Delicatessen item are:')
print('1. Mean:' ,Mean_Delicatessen)
print('2. Range:' ,Range_Delicatessen)
print('3. Standard Deviation:' ,Std_Delicatessen)
print('4. Coefficient of variation:' ,"{:.2%}".format(cv_Delicatessen))
```

```
Measure of variability for Delicatessen item are:
1. Mean: 1524.8704545454545
2. Range: 47940
3. Standard Deviation: 2820.1059373693975
4. Coefficient of variation: 184.94%
```

Figure 32 Measure of variability: Delicatessen

Coefficient of variance is a measure of dispersion of data points around mean. Higher the coefficient of variance, greater the level of dispersion around mean.

Conclusion: Below table shows that **all the items in the data set have dispersion**.

However, the item '**Delicatessen' has highest dispersion and therefore is the item with most inconsistent behavior**.

On the other hand, 'Fresh' is the item with least amount of dispersion around mean. In other words, **'Fresh' is the item with least inconsistent behavior.**

	Items	Coeffficient of variation
0	Fresh	105.39%
1	Milk	127.33%
2	Grocery	119.52%
3	Frozen	158.03%
4	Detergents Paper	165.46%
5	Delicatessen	184.94%

Figure 33 Item wise coefficient of variation

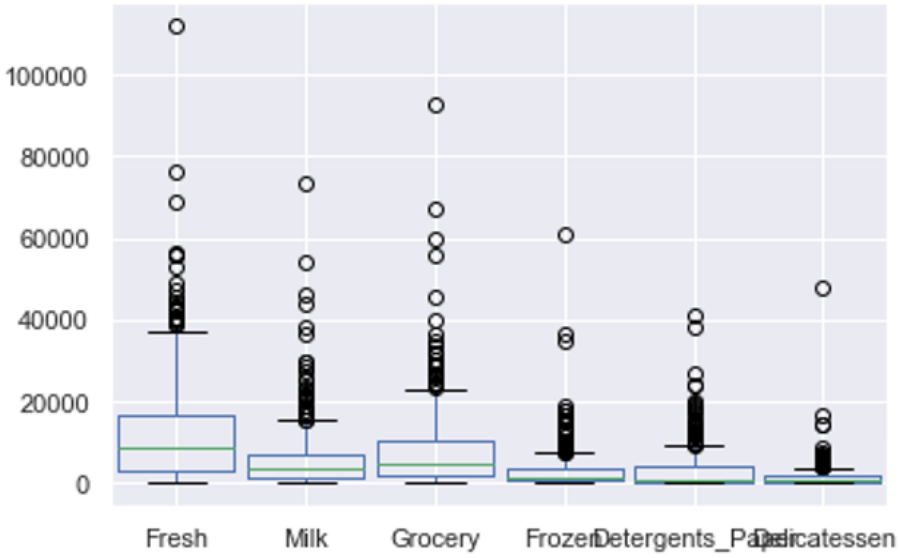
1.4 Are there any outliers in the data?

Background: There are 2 common ways to find out outliers in numerical variables:

- **Z Score:** To identify associated values [above defined threshold] *(if the details of z score are outside of the range of -3 to 3 than we can conclude that given numeric variable has the outlier.)*
- **Box Plot:** Visual representation of the outliers.

We have calculated *z score for all items and have plotted 'box-plot' for each item to see whether or not there are outliers in the data*. *(refer python notebook for detail working)*

Below is a summary boxplot with all 6 items together to support our concluding statement.



Z score calculation summary: Below table shows the values of outliers in the dataset for each item.

Item	Outliers
Fresh	[56159, 56082, 76237, 112151, 56083, 53205, 68951]
Milk	[36423, 54259, 29892, 38369, 46197, 73498, 29627, 43950, 28326]
Grocery	[55571, 59598, 45828, 92780, 39694, 36486, 67298]
Frozen	[35009, 18028, 36534, 18711, 60869, 17866]
Detergent Paper	[24171, 17740, 26701, 24231, 40827, 20070, 18906, 19410, 18594, 38102]
Delicatessen	[16523, 14472, 14351, 47943]

Conclusion: Box plot for each item and their z scores show that **all numerical variables (i.e. fresh, milk, grocery, frozen, detergent paper, delicatessen) has the outliers.**

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

The data set shows clear difference in spending over the six items across regions and across the given channels. For instance, the region 'other' sees the maximum amount spent on all items and in that region 'hotel' is the channel that accounts for majority amount spent on all items. It is safe to infer that out of the three regions, Lisbon, Oporto, and Other, the region 'other' spends the most on the items. On the other hand, out of the two channels in place, spending on items is the most through 'Hotel.'

To conclude our recommendation, we have bifurcated into 2 categories:

- Business oriented recommendations
- Data facts oriented recommendations

Business oriented recommendations:

Assuming the main goal/objective of the wholesale distributor is to grow Sales by increasing retailer spend across the board, we recommend:

- A deeper analysis by the wholesaler to study the regions showing a comparatively lower spending behavior, i.e Oporto and Lisbon and determine the root-cause. There can be multiple reasons for the difference in spending behavior [for which we don't have adequate number of data points]. For e.g. the reasons can be the difference in the ratio of working population, disposable income, spending behavior, average age of the residents etc.
- The wholesale retailer will have to draw these and many such inferences and come up with a plan to boost sales/expenditure on items in a region which currently shows less spending and through channel that is performing badly.

The wholesaler can provide discounts on items, different offers to make them look more affordable, if spending is an issue in a particular region, or add in a reward offer on the purchase of the item.

The wholesaler can continue to direct more quantity towards the channel and the region that brings in more revenue at the same time look at ways to attract more buyers in regions and through channels that are currently performing poorly.

Data facts oriented recommendations:

Sometimes it's best to keep outliers in the dataset as they might capture valuable insights. Retaining these points can be hard, particularly when it reduces statistical significance! However, excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the dataset. When considering whether to remove an outlier, we need to evaluate if it appropriately reflects our target population, buyers, items, region, channel and methodology around the same.

Did anything unusual happen while measuring these observations, such as focusing on specific region or channel or buyer or items.

If the outlier in question is:

- A measurement error or data entry error correct the error if possible. If we can't fix it, remove that observation because we know it's incorrect.
- When we decide to remove outliers, we need to document the excluded data points and explain your reasoning.
- We must be able to attribute a specific cause for removing outliers.
- Another approach is to perform the analysis with and without these observations and discuss the differences. Comparing results in this manner is particularly useful when we are unsure about removing an outlier and when there is substantial disagreement within the dataset.

Problem2:

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).

Remarks: All the questions of problem2 which explained here, are also performed in python as well. Please refer python notebook "SMDM_Group_Assignment_GRP6_Problem2".

Part I

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

Background: A **contingency table** cross-tabulates or tallies the values of two or more categorical variables. It also helps to study the pattern that may exist between the variables. Tallies can be shown as either of the following based on the contingency table we want to construct.

a. **a frequency OR No. of occurrence**
b. **a percentage of the overall total**
c. a percentage of the row total
d. a percentage of the column total

Example: In the simplest contingency table, one that contains only two categorical variables, the joint responses appear in the table such that the tallies of the one variable are located in the rows and the tallies of the other variable are located in the columns.

In this dataset, we have been asked to construct contingency table for various categorical variables with considering gender as row variable. As per above brief introduction, we will use option (a) and option (b) to build the contingency tables.

2.1.1. Gender and Major

Contingency Table : Gender and Major

	MAJOR								
GENDER	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	**Total**
Female	3	3	7	4	4	3	**9**		33
Male	4	1	4	2	6	4	5	3	29
Total	**7**	**4**	**11**	**6**	**10**	**7**	**14**	**3**	**62**

Contingency Table : Gender and Major (based on percentage of overall total)

	MAJOR								
GENDER	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	**Total**
Female	5%	5%	11%	6%	6%	5%	**15%**	0%	53%
Male	6%	2%	6%	3%	10%	6%	8%	5%	47%
Total	**11%**	**6%**	**18%**	**10%**	**16%**	**11%**	**23%**	**5%**	**100%**

As per above table, we have completed the contingency table after all **62 undergraduates** responses have been tallied. It shows that there are 9 responses that have gender as female and major as retailing/marketing. **In summarising all 16 joint responses(2 gender X 8 major)** , the table reveals that Gender(Female) and Major(Retailing/Marketing) has the highest joint response (i.e. no. of count as 9 & % of overall total as 15%).

In other words, **highest responses** received (i.e. 9 out of 62 (i.e. 15%)) from *female gender and they are holding retailing/marketing as a major*.

2.1.2. Gender and Grad Intention

Contingency Table : Gender and Grad Intention				
	GRAD INTENTION			
GENDER	Yes	No	Undecided	Total
Female	11	9	13	33
Male	17	3	9	29
Total	**28**	**12**	**22**	**62**

Contingency Table : Gender and Grad Intention (based on percentage of overall total)				
	GRAD INTENTION			
GENDER	Yes	No	Undecided	Total
Female	18%	15%	21%	53%
Male	27%	5%	15%	47%
Total	**45%**	**19%**	**35%**	**100%**

This table shows that there are 17 responses that have gender as male and grade intention as yes. **In summarising all 6 (2 gender X 3 grad intention) joint responses**, the table reveals that gender(male) and grad intention (yes) has the highest joint response (i.e. no. of count as 17 & % of overall total as 27%).

In other words, **highest respons**es received (i.e. 17 out of 62 (i.e. 27%)) *from male gender and with grad intention as Yes*.

2.1.3. Gender and Employment

Contingency Table : Gender and Employment				
	EMPLOYMENT			
GENDER	Full-Time	Part-Time	Unemployed	Total
Female	3	24	6	33
Male	7	19	3	29
Total	**10**	**43**	**9**	**62**

Contingency Table : Gender and Employment (based on percentage of overall total)				
	EMPLOYMENT			
GENDER	Full-Time	Part-Time	Unemployed	Total
Female	5%	39%	10%	53%
Male	11%	31%	5%	47%
Total	**16%**	**69%**	**15%**	**100%**

This table shows that there are 24 responses that have gender as female and employment as part-time. **In summarising all 6 joint responses (2 gender X 3 employment)**, the table reveals that gender(female) and employment (part-time) has the highest joint response (i.e. no. of count as 24 & % of overall total as 39%).

In other words, **highest responses** received (i.e. 24 out of 62 (i.e. 39%)) from *female gender and with employment as part-time.*

2.1.4. Gender and Computer

Contingency Table : Gender and Computer				
	COMPUTER			
GENDER	Laptop	Desktop	Tablet	Total
Female	29	2	2	33
Male	26	3		29
Total	**55**	**5**	**2**	**62**

Contingency Table : Gender and Computer (based on percentage of overall total)				
	COMPUTER			
GENDER	Laptop	Desktop	Tablet	Total
Female	47%	3%	3%	53%
Male	42%	5%	0%	47%
Total	**89%**	**8%**	**3%**	**100%**

This table shows that there are 29 responses that have gender as female and computer as laptop. **In summarising all 6 joint responses (2 gender X 3 computer)**, the table reveals that gender(female) and computer(laptop) has the highest joint response (i.e. no. of count as 29 & % of overall total as 47%).

In other words, **highest responses** received (i.e. 29 out of 62 (i.e. 47%)) *from female gender and they are having laptop as well*.

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?

Background: A probability is the **numerical value representing the chance, likelihood or possibility** that a particular event will occur. The probability involved is a proportion or fraction whose *value ranges between 0 and 1 including 0 and 1*.

Probability of occurrence:

$$Probability\ of\ occurrence = \frac{X}{T}$$

Where

X = Number of ways in which the event occurs

T = Total number of possible outcomes

In this dataset, total responses we have received are 62 and amongst that 33 were female and 29 were male.

GENDER	Nos.
Male	29
Female	33
Total	**62**

Probability that a randomly selected CMSU student will be male

$$P(Male) = \frac{Male\ respondents}{Total\ respondents}$$

$$= \frac{29}{62}$$

$$= 0.468$$

There is a **46.8%** chance that a randomly selected student will be **male.**

Probability that a randomly selected CMSU student will be female

$$P(Female) = \frac{Female\ respondents}{Total\ respondents}$$

$$= \frac{33}{62}$$

$$= 0.532$$

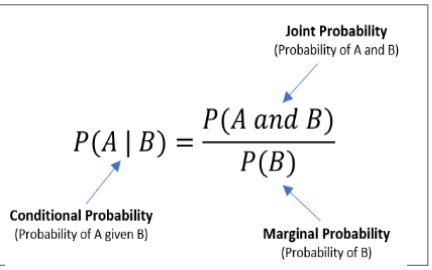
There is a **53.2%** chance that a randomly selected student will be **female.**

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.

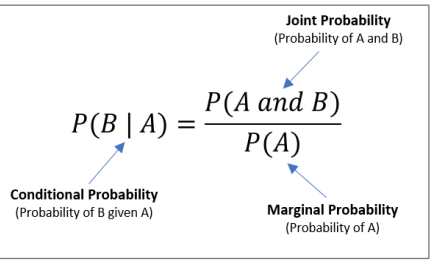
Background: Conditional probability refers to the probability of event A, given the information about the occurrence of another event, B.

Conditional probability:

The probability of ***A given B*** is equal to the probability of ***A and B*** divided by the probability of ***B***.



The probability of ***B given A*** is equal to the probability of ***A and B*** divided by the probability of ***A***.

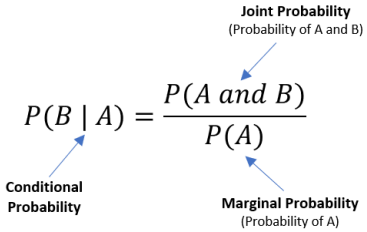


Now, let's find out the conditional probability of different majors among the male students.

1. **Major = Accounting given Student = Male**

Let's consider,

A = Student is Male
B = Major is Accounting



Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{4/62}{29/62} = \frac{4}{29} = \mathbf{0.138}$$

$$P(Major\ is\ Accounting \mid Student\ is\ Male) = 0.138$$

Therefore, given that the selected student is male, there is a **13.8%** chance that the male student also has major as accounting.

2. **Major = CIS given Student = Male**

Consider: A = Student is Male , B = Major is CIS

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{1/62}{29/62} = \frac{1}{29} = \mathbf{0.034}$$

$$P(Major\ is\ CIS \mid Student\ is\ Male) = 0.034$$

Therefore, given that the selected student is male, there is a **3.4%** chance that the male student also has major as CIS.

Likewise, the same way other majors also used to calculate conditional probability.

Major = Economics/Finance given Student = Male

Consider: A = Student is Male , B = Major is Economics/Finance

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{4/62}{29/62} = \frac{4}{29} = \mathbf{0.138}$$

$$P(Major\ is\ Economics\_Finance \mid Student\ is\ Male) = 0.138$$

Therefore, given that the selected student is male, there is a **13.8%** chance that the male student also has major as economics/finance.

Major = International Business given Student = Male

Consider: A = Student is Male , B = Major is International Business

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{2/62}{29/62} = \frac{2}{29} = \mathbf{0.069}$$

$$P(Major\ is\ International\ Business \mid Student\ is\ Male) = 0.069$$

Therefore, given that the selected student is male, there is a **6.9%** chance that the male student also has major as international business.

Major = Management given Student = Male

C Consider: A = Student is Male , B = Major is Management

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{6/62}{29/62} = \frac{6}{29} = \mathbf{0.207}$$

$$P(Major\ is\ Management \mid Student\ is\ Male) = 0.207$$

Therefore, given that the selected student is male, there is a **20.7%** chance that the male student also has major as management.

Major = Other given Student = Male

Consider: A = Student is Male , B = Major is Other

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{4/62}{29/62} = \frac{4}{29} = \mathbf{0.138}$$

$$P(Major\ is\ other \mid Student\ is\ Male) = 0.138$$

Therefore, given that the selected student is male, there is a **13.8%** chance that the male student also has major as other.

Major = Retailing/Marketing given Student = Male

Consider: A = Student is Male , B = Major is Retailing/Marketing

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{5/62}{29/62} = \frac{5}{29} = \mathbf{0.172}$$

$$P(Major\ is\ Retailing\_Marketing \mid Student\ is\ Male) = 0.172$$

Therefore, given that the selected student is male, there is a **17.2%** chance that the male student also has major as retailing/marketing.

Major = Undecided given Student = Male

Consider: A = Student is Male , B = Major is Undecided

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{3/62}{29/62} = \frac{3}{29} = \mathbf{0.103}$$

$$P(Major\ is\ Undecided \mid Student\ is\ Male) = 0.103$$

Therefore, given that the selected student is male, there is a **10.3%** chance that the male student also has major as undecided.

In summary, below table gives the conditional probability of different majors among the male students:

Contingency Table : Gender and Major									
	MAJOR (as B)								
GENDER (as A)	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	**Total**
Female	3	3	7	4	4	3	9		33
Male	4	1	4	2	6	4	5	3	29
Total	**7**	**4**	**11**	**6**	**10**	**7**	**14**	**3**	**62**
P (A and B) *(Joint Probability)*	*4/62*	*1/62*	*4/62*	*2/62*	*6/62*	*4/62*	*5/62*	*3/62*	
P (A) = ***Male*** *(Marginal Probability)*	*29/62*	*29/62*	*29/62*	*29/62*	*29/62*	*29/62*	*29/62*	*29/62*	
P (B | A) *(Conditional Probability)*	0.138	0.034	0.138	0.069	0.207	0.138	0.172	0.103	

Now, let's find out the conditional probability of different majors among the female students.

If we will follow the same approach as above then below is the summary table gives the conditional probability of different majors among the female students

Contingency Table : Gender and Major								
	MAJOR (as Y)							
GENDER (as X)	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	**Total**
Female	3	3	7	4	4	3	9		33
Male	4	1	4	2	6	4	5	3	29
Total	**7**	**4**	**11**	**6**	**10**	**7**	**14**	**3**	**62**
P (X and Y) (Joint Probability)	*3/62*	*3/62*	*7/62*	*4/62*	*4/62*	*3/62*	*9/62*	*0/62*
*P (X) = **Female** (Marginal Probability)*	*33/62*	*33/62*	*33/62*	*33/62*	*33/62*	*33/62*	*33/62*	*33/62*
P (Y | X) (Conditional Probability)	0.091	0.091	0.212	0.121	0.121	0.091	0.273	0.000

Major = Accounting given Student = Female

Consider: X = Student is Female , Y = Major is accounting

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{3/62}{33/62} = \frac{3}{33} = \mathbf{0.091}$$

$$P(\text{Major is Accounting} \mid \text{Student is Female}) = 0.091$$

Therefore, given that the selected student is female, there is a **9.1%** chance that the female student also has major as accounting.

Major = CIS given Student = Female

Consider: X = Student is Female , Y = Major is CIS

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{3/62}{33/62} = \frac{3}{33} = \mathbf{0.091}$$

$$P(\text{Major is CIS} \mid \text{Student is Female}) = 0.091$$

Therefore, given that the selected student is female, there is a **9.1%** chance that the female student also has major as CIS.

Major = Economics/Finance given Student = Female

Consider: X = Student is Female , Y = Major is Economics/Finance

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{7/62}{33/62} = \frac{7}{33} = \mathbf{0.212}$$

$$P(\text{Major is Economics\_Finance} \mid \text{Student is Female}) = 0.212$$

Therefore, given that the selected student is female, there is a **21.2%** chance that the female student also has major as economics/finance.

Major = International Business given Student = Female

Consider: X = Student is Female , Y = Major is International Business

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{4/62}{33/62} = \frac{4}{33} = \mathbf{0.121}$$

$$P(\text{Major is International Business} \mid \text{Student is Female}) = 0.121$$

Therefore, given that the selected student is female, there is a **12.1%** chance that the female student also has major as International business.

Major = Management given Student = Female

Consider: X = Student is Female , Y = Major is Management

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{4/62}{33/62} = \frac{4}{33} = \mathbf{0.121}$$

$$P(\text{Major is Management} \mid \text{Student is Female}) = 0.121$$

Therefore, given that the selected student is female, there is a **12.1%** chance that the female student also has major as management.

Major = Other given Student = Female

Consider: X = Student is Female , Y = Major is Other

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{3/62}{33/62} = \frac{3}{33} = \mathbf{0.091}$$

$$P(\text{Major is other} \mid \text{Student is Female}) = 0.091$$

Therefore, given that the selected student is female, there is a **9.1%** chance that the female student also has major as other.

Major = Retailing/Marketing given Student = Female

Consider: X = Student is Female , Y = Major is Retailing/Marketing

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{9/62}{33/62} = \frac{9}{33} = \mathbf{0.273}$$

$$P(\text{Major is Retailing\_Marketing} \mid \text{Student is Female}) = 0.273$$

Therefore, given that the selected student is female, there is a **27.3%** chance that the female student also has major as retailing/marketing.

Major = Undecided given Student = Female

Consider: X = Student is Female , Y = Major is Undecided

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{0/62}{33/62} = \frac{0}{33} = \mathbf{0}$$

$$P(\text{Major is Undecided} \mid \text{Student is Female}) = 0$$

Therefore, given that the selected student is female, there is a **0%** chance that the female student also has major as undecided.

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

We will follow the same approach as 2.2.2. Below is the summary table gives the conditional probability of intent to graduate given that the student is male.

Contingency Table : Gender and Grad Intention				
	GRAD INTENTION (as B)			
GENDER (as A)	Yes	No	Undecided	**Total**
Female	11	9	13	33
Male	17	3	9	29
Total	**28**	**12**	**22**	**62**
P (A and B) *(Joint Probability)*	*17/62*	*3/62*	*9/62*	
P (A) = ***Male*** *(Marginal Probability)*	*29/62*	*29/62*	*29/62*	
P (B | A) *(Conditional Probability)*	0.586	0.103	0.310	

Grade intention = Yes given Student = Male

Consider: A = Student is Male , B = Grade intention as Yes

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{17/62}{29/62} = \frac{17}{29} = \mathbf{0.586}$$

$$P(\text{Grade intention as Yes} \mid \text{Student is Male}) = 0.586$$

Therefore, given that the selected student is male, there is a **58.6%** chance that the male student also inclined towards grade intention.

Grade intention = No given Student = Male

Consider: A = Student is Male , B = Grade intention as No

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{3/62}{29/62} = \frac{3}{29} = \mathbf{0.103}$$

$$P(\text{Grade intention as No} \mid \text{Student is Male}) = 0.103$$

Therefore, given that the selected student is male, there is a **10.3%** chance that the male student also declined towards grade intention.

Grade intention = Undecided given Student = Male

Consider: A = Student is Male , B = Grade intention as Undecided

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{9/62}{29/62} = \frac{9}{29} = \mathbf{0.310}$$

$$P(\text{Grade intention as Undecided} \mid \text{Student is Male}) = 0.310$$

Therefore, given that the selected student is male, there is a **31%** chance that the male student also not sure (undecided) towards grade intention.

below is the summary table gives the conditional probability of intent to graduate given that the student is female.

Contingency Table : Gender and Grad Intention				
	GRAD INTENTION (as Y)			
GENDER (as X)	Yes	No	Undecided	**Total**
Female	11	9	13	33
Male	17	3	9	29
Total	**28**	**12**	**22**	**62**
P (X and Y) *(Joint Probability)*	*11/62*	*9/62*	*13/62*	
P (X) = ***Female*** *(Marginal Probability)*	*33/62*	*33/62*	*33/62*	
P (Y | X) *(Conditional Probability)*	0.333	0.273	0.394	

Grade intention = Yes given Student = Female

Consider: A = Student is Female , B = Grade intention as Yes

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{11/62}{33/62} = \frac{11}{33} = \mathbf{0.333}$$

$$P(\text{Grade intention as Yes} \mid \text{Student is Female}) = 0.333$$

Therefore, given that the selected student is female, there is a **33.3%** chance that the female student also inclined towards grade intention.

Grade intention = No given Student = Female

Consider: A = Student is Female , B = Grade intention as No

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{9/62}{33/62} = \frac{9}{33} = \mathbf{0.273}$$

$$P(\text{Grade intention as No} \mid \text{Student is Female}) = 0.273$$

Therefore, given that the selected student is female, there is a **27.3%** chance that the female student also declined towards grade intention.

Grade intention = Undecided given Student = Female

Consider: A = Student is Female , B = Grade intention as Undecided

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{13/62}{33/62} = \frac{13}{33} = \mathbf{0.394}$$

$$P(\text{Grade intention as Undecided} \mid \text{Student is Female}) = 0.394$$

Therefore, given that the selected student is female, there is a **39.4%** chance that the female student also not sure (undecided) towards grade intention.

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

Contingency Table : Gender and Employment

GENDER (as A)	EMPLOYMENT (as B) Full-Time	Part-Time	Unemployed	Total
Female	3	24	6	33
Male	7	19	3	29
Total	10	43	9	62

P (A and B) (Joint Probability)	*7/62*	*19/62*	*3/62*
*P (A) = **Male** (Marginal Probability)*	*29/62*	*29/62*	*29/62*
P (B | A) (Conditional Probability)	0.241	0.655	0.103

We will follow the same approach as 2.2.2. Summary & Detail table gives the conditional probability of employment status for male students.

Employment = Full-time given Student = Male

Consider: A = Student is Male , B = Employment as Full-time

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{17/62}{29/62} = \frac{7}{29} = \mathbf{0.241}$$

$$P(\text{Full} - \text{time} \mid \text{Student is Male}) = 0.241$$

Therefore, given that the selected student is male, there is a **24.1%** chance that the male student also having full-time employment.

Employment = Part-time given Student = Male

Consider: A = Student is Male , B = Employment as Part-time

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{19/62}{29/62} = \frac{19}{29} = \mathbf{0.655}$$

$$P(\text{Part} - \text{time} \mid \text{Student is Male}) = 0.655$$

Therefore, given that the selected student is male, there is a **65.5%** chance that the male student also having part-time employment.

Employment = Unemployed given Student = Male

Consider: A = Student is Male , B = Employment as Unemployed

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{3/62}{29/62} = \frac{3}{29} = \mathbf{0.103}$$

$$P(\text{Unemployed} \mid \text{Student is Male}) = 0.103$$

Therefore, given that the selected student is male, there is a **10.3%** chance that the male student also unemployed.

Contingency Table : Gender and Employment				
	EMPLOYMENT (as Y)			
GENDER (as X)	Full-Time	Part-Time	Unemployed	**Total**
Female	3	24	6	33
Male	7	19	3	29
Total	**10**	**43**	**9**	**62**
P (X and Y) *(Joint Probability)*	*3/62*	*24/62*	*6/62*	
P (X) = ***Female*** *(Marginal Probability)*	*33/62*	*33/62*	*33/62*	
P (Y | X) *(Conditional Probability)*	0.091	0.727	0.182	

Summary & Detail table gives the conditional probability of employment status for female students.

Employment = Full-time given Student = Female

Consider: X = Student is Female , Y = Employment as Full-time

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{3/62}{33/62} = \frac{3}{33} = \mathbf{0.091}$$

$$P(Full-time \mid Student\ is\ Female) = 0.091$$

Therefore, given that the selected student is female, there is a **9.1%** chance that the female student also having full-time employment.

Employment = Part-time given Student = Female

Consider: A = Student is Female , B = Employment as Part-time

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{24/62}{33/62} = \frac{24}{33} = \mathbf{0.727}$$

$$P(Part-time \mid Student\ is\ Female) = 0.727$$

Therefore, given that the selected student is female, there is a **72.7%** chance that the female student also having part-time employment.

Employment = Unemployed given Student = Female

Consider: A = Student is Female , B = Employment as Unemployed

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{6/62}{33/62} = \frac{6}{33} = \mathbf{0.182}$$

$$P(Unemployed \mid Student\ is\ Female) = 0.182$$

Therefore, given that the selected student is female, there is a **18.2%** chance that the female student also unemployed.

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

We will follow the same approach as 2.2.2. Summary & Detail table gives the conditional probability of employment status for male students.

Contingency Table : Gender and Computer				
	COMPUTER (as B)			
GENDER (as A)	Laptop	Desktop	Tablet	**Total**
Female	29	2	2	33
Male	26	3		29
Total	**55**	**5**	**2**	**62**
P (A and B) *(Joint Probability)*	*26/62*	*3/62*	*0/62*	
P (A) = ***Male*** *(Marginal Probability)*	*29/62*	*29/62*	*29/62*	
P (B | A) *(Conditional Probability)*	0.897	0.103	0.000	

Computer = Laptop given Student = Male

Consider: A = Student is Male , B = Computer as Laptop

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{26/62}{29/62} = \frac{26}{29} = \mathbf{0.897}$$

$$P(Laptop \mid Student\ is\ Male) = 0.897$$

Therefore, given that the selected student is male, there is a **89.7%** chance that the male student also having laptop.

Computer = Desktop given Student = Male

Consider: A = Student is Male , B = Computer as Desktop

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{3/62}{29/62} = \frac{3}{29} = \mathbf{0.103}$$

$$P(Desktop \mid Student\ is\ Male) = 0.103$$

Therefore, given that the selected student is male, there is a **10.3%** chance that the male student also having desktop.

Computer = Tablet given Student = Male

Consider: A = Student is Male , B = Computer as Tablet

Hence, conditional probability:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0/62}{29/62} = \frac{0}{29} = \mathbf{0}$$

$$P(Tablet \mid Student\ is\ Male) = 0$$

Therefore, given that the selected student is male, there is a **0%** chance that the male student also having tablet.

Contingency Table : Gender and Computer

GENDER (as X)	COMPUTER (as Y) Laptop	Desktop	Tablet	Total
Female	29	2	2	33
Male	26	3		29
Total	**55**	**5**	**2**	**62**

P (X and Y) *(Joint Probability)*	*29/62*	*2/62*	*2/62*
P (X) = ***Female*** *(Marginal Probability)*	*33/62*	*33/62*	*33/62*
P (Y | X) *(Conditional Probability)*	0.879	0.061	0.061

Summary & Detail table gives the conditional probability of employment status for female students.

Computer = Laptop given Student = Female

Consider: X = Student is Female , Y= Computer as Laptop

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{29/62}{33/62} = \frac{29}{33} = \mathbf{0.879}$$

$$P(Laptop \mid Student\ is\ Female) = 0.879$$

Therefore, given that the selected student is female, there is a **87.9%** chance that the female student also having laptop.

Computer = Desktop given Student = Female

Consider: X = Student is Female , Y = Computer as Desktop

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{2/62}{33/62} = \frac{2}{33} = \mathbf{0.061}$$

$$P(Desktop \mid Student\ is\ Female) = 0.061$$

Therefore, given that the selected student is female, there is a **6.1%** chance that the female student also having desktop.

Computer = Tablet given Student = Female

Consider: X = Student is Female , Y = Computer as Tablet

Hence, conditional probability:

$$P(Y \mid X) = \frac{P(X \text{ and } Y)}{P(X)} = \frac{2/62}{33/62} = \frac{2}{33} = \mathbf{0.061}$$

$$P(Tablet \mid Student\ is\ Female) = 0.061$$

Therefore, given that the selected student is female, there is a **6.1%** chance that the female student also having tablet.

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

Background: How to check Independence?, when the outcome of one event does not affect with the probability of occurrence of another event, the events are said to be independent. Below equation is used to determine the independence.

Example: Two events, A and B are independent if and only if

$$P(B \mid A) = P(B)$$

Conditional Probability (Probability of B given A)

Marginal Probability (Probability of B)

Let's check independence for each column variable with Gender.

1. Gender & Major:

As per 2.2.2, we already know probability of major given gender. now, to determine independence we need to populate probability major.

$$P(Major = Accounting) = \frac{Accounting\ respondents}{Total\ respondents} = \frac{7}{62} = 0.113$$

$$P(Major = Accounting) = 0.113$$

$$P(Major = CIS) = \frac{CIS\ respondents}{Total\ respondents} = \frac{4}{62} = 0.065$$

$$P(Major = CIS) = 0.065$$

Likewise, calculate the other major's probability as well (refer below table for calculation) and then compare whether ***P(Major | Gender) = P(Major)*** *or not. If, both the probabilities are same then we can conclude that they are independent*.

Contingency Table : Gender and Major

GENDER	MAJOR: Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	Total
Female	3	3	7	4	4	3	9		33
Male	4	1	4	2	6	4	5	3	29
Total	**7**	**4**	**11**	**6**	**10**	**7**	**14**	**3**	**62**
*P (**Majorwise**)* *(Marginal Probability)*	*7/62*	*4/62*	*11/62*	*6/62*	*10/62*	*7/62*	*16/62*	*3/62*
	0.113	0.065	0.177	0.097	0.161	0.113	0.226	0.048
*P (Majorwise | **Male**)* *(Conditional Probability)*	0.138	0.034	0.138	0.069	0.207	0.138	0.172	0.103
Independence *P (Majorwise | **Male**) = P (Majorwise)*	0.138 ≠ 0.113 **Dependent**	0.034 ≠ 0.065 **Dependent**	0.138 ≠ 0.177 **Dependent**	0.069 ≠ 0.097 **Dependent**	0.207 ≠ 0.161 **Dependent**	0.138 ≠ 0.113 **Dependent**	0.172 ≠ 0.226 **Dependent**	0.103 ≠ 0.048 **Dependent**
*P (Majorwise | **Female**)* *(Conditional Probability)*	0.091	0.091	0.212	0.121	0.121	0.091	0.273	0.000
Independence *P (Majorwise | **Female**) = P (Majorwise)*	0.091 ≠ 0.113 **Dependent**	0.091 ≠ 0.113 **Dependent**	0.212 ≠ 0.177 **Dependent**	0.121 ≠ 0.097 **Dependent**	0.121 ≠ 0.097 **Dependent**	0.091 ≠ 0.113 **Dependent**	0.273 ≠ 0.226 **Dependent**	0.000 ≠ 0.048 **Dependent**

So, we can say that row variables (i.e. gender) **does affect (dependent)** with the probability of column variables (i.e. major).

2. Gender & Grad intention:

As per 2.2.3 we already know probability of Grad intention given gender. now, to determine independence we need to populate probability grad intention.

$$P(Grade\ Intention = Yes) = \frac{Yes\ respondents}{Total\ respondents} = \frac{28}{62} = 0.452$$

$$P(Grade\ Intention = Yes) = 0.452$$

$$P(Grade\ Intention = No) = \frac{No\ respondents}{Total\ respondents} = \frac{12}{62} = 0.194$$

$$P(Grade\ Intention = No) = 0.194$$

$$P(Grade\ Intention = Undecided) = \frac{Undecided\ respondents}{Total\ respondents} = \frac{22}{62} = 0.355$$

$$P(Grade\ Intention = Undecided) = 0.355$$

Now, compare whether **P(Grade intention | Gender) = P(Grade intention)** or not.

Contingency Table : Gender and Grad Intention				
	GRAD INTENTION			
GENDER	Yes	No	Undecided	**Total**
Female	11	9	13	33
Male	17	3	9	29
Total	**28**	**12**	**22**	**62**
*P (**Grad Intention**)* *(Marginal Probability)*	*28/62*	*12/62*	*22/62*	
	0.452	0.194	0.355	
*P (Grad Intention | **Male**)* *(Conditional Probability)*	0.586	0.103	0.310	
Independence *P(Grad Intention | **Male**) = P(Grad Intention)*	0.586 ≠ 0.452 **Dependent**	0.103 ≠ 0.194 **Dependent**	0.310 ≠ 0.355 **Dependent**	
*P (Grad Intention | **Female**)* *(Conditional Probability)*	0.333	0.273	0.394	
Independence *P(Grad Intention | **Female**) = P(Grad Intention)*	0.333 ≠ 0.452 **Dependent**	0.273 ≠ 0.194 **Dependent**	0.394 ≠ 0.355 **Dependent**	

So, we can say that row variables (i.e. gender) **does affect (dependent)** with the probability of column variables (i.e. grad intention).

3. Gender & Employment:

As per 2.2.4 we already know probability of employment given gender. now, to determine independence we need to populate probability employment.

$$P(Employment = Full-time) = \frac{Full-time\ respondents}{Total\ respondents} = \frac{10}{62} = 0.161$$

$$P(Employment = Full-time) = 0.161$$

$$P(Employment = Part-time) = \frac{Part-time\ respondents}{Total\ respondents} = \frac{43}{62} = 0.694$$

$$P(Employment = Part-time) = 0.694$$

$$P(Employment = Unemployed) = \frac{Unemployed\ respondents}{Total\ respondents} = \frac{9}{62} = 0.145$$

$$P(Employment = Unemployed) = 0.145$$

Now, compare whether **P(Employment | Gender) = P(Employment)** or not.

Contingency Table : Gender and Empolyment				
	EMPLOYMENT			
GENDER	Full-Time	Part-Time	Unemployed	**Total**
Female	3	24	6	33
Male	7	19	3	29
Total	**10**	**43**	**9**	**62**
*P (**Employment**) =* *(Marginal Probability)*	*10/62* 0.161	*43/62* 0.694	*9/62* 0.145	
*P (Employment | **Male**)* *(Conditional Probability)*	0.241	0.655	0.103	
Independence *P(Employment | **Male**) = P(Employment)*	0.241 ≠ 0.161 **Dependent**	0.655 ≠ 0.694 **Dependent**	0.103 ≠ 0.145 **Dependent**	
*P (Employment | **Female**)* *(Conditional Probability)*	0.091	0.727	0.182	
Independence *P(Employment | **Female**) = P(Employment)*	0.091 ≠ 0.161 **Dependent**	0.727 ≠ 0.694 **Dependent**	0.182 ≠ 0.145 **Dependent**	

So, we can say that row variables (i.e. gender) **does affect (dependent)** with the probability of column variables (i.e. employment).

4. Gender & Computer:

As per 2.2.5 we already know probability of computer given gender. now, to determine independence we need to populate probability computer.

$$P(Computer = Laptop) = \frac{Laptop\ respondents}{Total\ respondents} = \frac{53}{62} = 0.887$$

$$P(Computer = Laptop) = 0.887$$

$$P(Computer = Desktop) = \frac{Desktop\ respondents}{Total\ respondents} = \frac{5}{62} = 0.081$$

$$P(Computer = Desktop) = 0.081$$

$$P(Computer = Tablet) = \frac{Tablet\ respondents}{Total\ respondents} = \frac{2}{62} = 0.032$$

$$P(Computer = Tablet) = 0.032$$

Now, compare whether **P(Computer | Gender) = P(Computer)** or not.

Contingency Table : Gender and Computer				
	COMPUTER			
GENDER	Laptop	Desktop	Tablet	**Total**
Female	29	2	2	33
Male	26	3		29
Total	**55**	**5**	**2**	**62**
*P (**Computer**) = (Marginal Probability)*	*53/62*	*5/62*	*2/62*
	0.887	0.081	0.032
*P (Computer | **Male**) (Conditional Probability)*	0.897	0.103	0.000
*Independence P(Computer | **Male**) = P(Computer)*	0.897 ≠ 0.887 **Dependent**	0.103 ≠ 0.081 **Dependent**	0.00 ≠ 0.032 **Dependent**
*P (Computer | **Female**) (Conditional Probability)*	0.879	0.061	0.061
*Independence P(Computer | **Female**) = P(Computer)*	0.879 ≠ 0.887 **Dependent**	0.061 ≠ 0.081 **Dependent**	0.061 ≠ 0.032 **Dependent**

So, we can say that row variables (i.e. gender) **does affect (dependent)** with the probability of column variables (i.e. computer).

Hence, it is proved that all the column variables (i.e. Major, Grad intention, Employment and Computer) are dependent with the probability of row variables (i.e. gender).

Part II

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]

Background: In the dataset, we **have 3 numerical variables (i.e. Salary, Spending & Text messages)** and we want to check whether it follows the normal distribution or not?

To evaluate normality, the **normal distribution has several important theoretical/visual characteristics** that needs to be checked as mentioned below:

a. Skewness:
   - mean < median (negative OR left skewed distribution)
   - mean = median (symmetrical OR normal distribution)
   - mean > median (positive OR right skewed distribution)
b. The IQR (interquartile range) equals to 1.33 standard deviations
c. The range is approximately equal to 6 standard deviations.
d. Box plot
e. Shapiro's test
f. QQ plot (quantile-quantile plot)

Numerical Variable: Salary

In the given dataset we have 62 entries for Salary. To check the type of the distribution let's perform above activities. Below are the descriptive statistics and five number summary for salary variable.

Salary			
Descriptive Statistics		*Five Number Summary*	
Mean	48.55	Minimum	25
Standard Error	1.53	First Quartile (Q1)	40
Median	50.00	Median	50
Mode	40.00	Third Quartile (Q3)	55
Standard Deviation	12.08	Maximum	80
Sample Variance	145.95		
Kurtosis	0.42	IQR	15
Skewness	0.53	(Q3 less Q1)	
Range	55.00		
Minimum	25.00		
Maximum	80.00		
Sum	3010.00		
Count	62.00		

a. *The mean (48.55) is less than the median (50)*. (In a normal distribution, the mean and median are equal) Hence, **it is not normally distributed.**

b. *The interquartile range (IQR) (15) is approximately 1.24 standard deviation.* (In a normal distribution, the interquartile range is 1.33 standard deviation) Hence, **it is not normally distributed.**

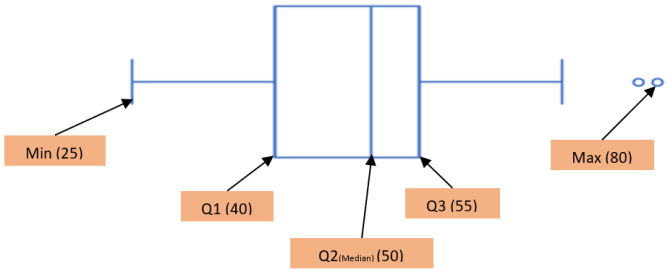
c. *The range (55) is approximately 4.55 standard deviations*. (In a normal distribution, range is 6 standard deviation) Hence, **it is not normally distributed.**

d. Box Plot: Let's compare (Q3 (3rd quartile) – Q2 (median)) and (Q2 (median) – Q1 (1st quartile)). If both are equal then check the whisker (line) both the side. If that also is same then we can say it is bell shaped. In this data set,

*(Q3 (3rd quartile) – Q2 (median)) (55 – 50 = 5) is **lesser than***

(Q2 (median) – Q1 (1st quartile)) (50 – 40 = 10).

Hence, **it is not normally distributed.**



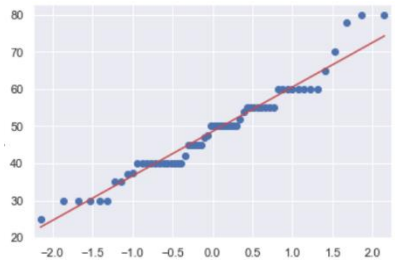
e. Shapiro's test: In shapiro's test, if p value is less than alpha (i.e. 0.05), we can say that data is not normally distributed. Here, Null hypothesis would be data is normally distributed and alternate hypothesis would be data is not normally distributed.

```
## Shapiro test
stat, p = shapiro(df1)
print('Statistics=%.3f, p=%.3f' % (stat, p))
alpha = 0.05
Salary_Shapiro = np.where(p > alpha, "normally distributed", "not normally distributed")
print('Shapiro test for salary variable stats that data is', Salary_Shapiro)
```

```
Statistics=0.957, p=0.028
Shapiro test for salary variable stats that data is not normally distributed
```

In salary variable, p value is 0.028 which is lesser then alpha value(p low -> null go).
Hence, **it is not normally distributed.**

f. QQ plot: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



In Salary variable, *QQ plot showing the scatter plot of points in a diagonal line, which is not fitting the expected diagonal pattern*. Hence, it is **not normally distributed.**

Conclusion: Based on above justifications, we have seen that salary variable has mean less than median which means skewness is negative side. The IQR and range are also lesser than what would be expected in normal distribution. In Box plot we can see whisker is normally distributed, but box is skewed left side. Shapiro & QQ plot also said the same. Thus, we can conclude that the data characteristics of salary variable **differs from theoretical/visual properties of normal distribution.**

Numerical Variable: Spending

In the given dataset we have 62 entries for Spending. To check the type of the distribution let's perform above activities. Besides are the descriptive statistics and five number summary for spending variable.

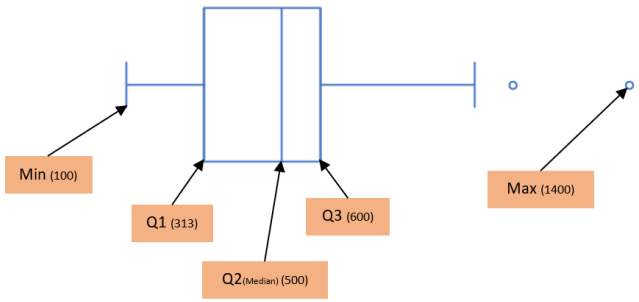
Spending			
Descriptive Statistics		*Five Number Summary*	
Mean	482.02	Minimum	100
Standard Error	28.19	First Quartile (Q1)	313
Median	500.00	Median	500
Mode	500.00	Third Quartile (Q3)	600
Standard Deviation	221.95	Maximum	1400
Sample Variance	49263.49		
Kurtosis	4.56	IQR	288
Skewness	1.59	(Q3 less Q1)	
Range	1300.00		
Minimum	100.00		
Maximum	1400.00		
Sum	29885.00		
Count	62.00		

a. *The mean (482.02) is less than the median (500).* (In a normal distribution, the mean and median are equal) Hence, **it is not normally distributed.**

b. *The interquartile range (IQR) (288) is approximately 1.29 standard deviation.* (In a normal distribution, the interquartile range is 1.33 standard deviation) Hence, **it is not normally distributed.**

c. *The range (1300) is approximately 5.86 standard deviations.* (In a normal distribution, range is 6 standard deviation) Hence, **it is not normally distributed.**

d. **Box Plot:** Let's compare (Q3 (3rd quartile) – Q2 (median)) and (Q2 (median) – Q1 (1st quartile)). If both are equal then check the whisker (line) both the side. If that also is same then we can say it is bell shaped. In this data set,



*(Q3 (3rd quartile) – Q2 (median)) (600 – 500 = 100) is **lesser than** (Q2 (median) – Q1 (1st quartile)) (500 – 313 = 187).* Hence, **it is not normally distributed.**

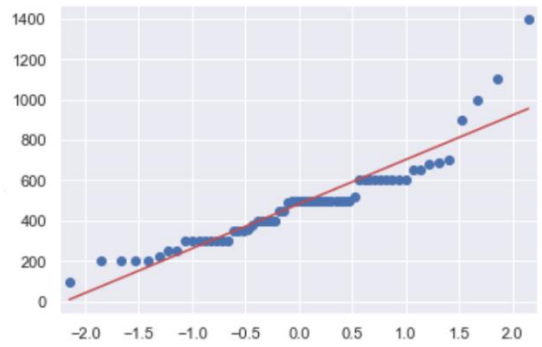
e. **Shapiro's test:** In shapiro's test, if p value is less than alpha (i.e. 0.05), we can say that data is not normally distributed. Here, Null hypothesis would be data is normally distributed and alternate hypothesis would be data is not normally distributed.

```
## Shapiro test
stat, p = shapiro(df1)
print('Statistics=%.3f, p=%.6f' % (stat, p))
alpha = 0.05
Spending_Shapiro = np.where(p > alpha, "normally distributed", "not normally distributed")
print('Shapiro test for spending variable stats that data is', Spending_Shapiro)
```

```
Statistics=0.878, p=0.000017
Shapiro test for spending variable stats that data is not normally distributed
```

In spending variable, p value is 0.000017 which is lesser then alpha value(p low -> null go). Hence, **it is not normally distributed.**

f. **QQ plot:** A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



In Spending variable, *QQ plot showing the scatter plot of points in a diagonal line, which is not fitting the expected diagonal pattern*. Hence, it is **not normally distributed.**

Conclusion: Based on above justifications, we have seen that spending variable has mean less than median which means skewness is negative side. The IQR and range are also lesser than what would be expected in normal distribution. In Box plot we can see box is skewed left side. Shapiro & QQ plot also said the same. Thus, we can conclude that the data characteristics of spending variable **differs from theoretical/visual properties of normal distribution**.

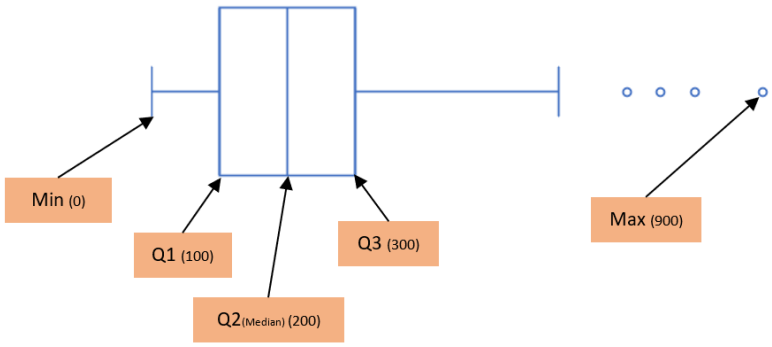
Numerical Variable: Text Messages

In the given dataset we have 62 entries for Text messages. To check the type of the distribution let's perform above activities. Below are the descriptive statistics and five number summary for text messages variable.

Text Messages			
Descriptive Statistics		*Five Number Summary*	
Mean	246.21	Minimum	0
Standard Error	27.24	First Quartile (Q1)	100
Median	200.00	Median	200
Mode	300.00	Third Quartile (Q3)	300
Standard Deviation	214.47	Maximum	900
Sample Variance	45995.64		
Kurtosis	1.14	IQR	200
Skewness	1.30	(Q3 less Q1)	
Range	900.00		
Minimum	0.00		
Maximum	900.00		
Sum	15265.00		
Count	62.00		

a. *The mean (246.21) is greater than the median (200)*. (In a normal distribution, the mean and median are equal) Hence, **it is not normally distributed.**
b. *The interquartile range (IQR) (200) is approximately 0.93 standard deviation.* (In a normal distribution, the interquartile range is 1.33 standard deviation) Hence, **it is not normally distributed.**
c. *The range (900) is approximately 4.20 standard deviations.* (In a normal distribution, range is 6 standard deviation) Hence, **it is not normally distributed.**
d. **Box Plot:** Let's compare (Q3 (3rd quartile) – Q2 (median)) and (Q2 (median) – Q1 (1st quartile)). If both are equal then check the whisker (line) both the side. If that also is same then we can say it is bell shaped.

In this data set,(Q3 (3rd quartile) – Q2 (median)) (300 – 200 = 100) is ***same as*** *(Q2 (median) – Q1 (1st quartile)) (200 – 100 = 100) but* ***whisker (line) is longer in right side then the left side***. Hence, **it is not normally distributed.**



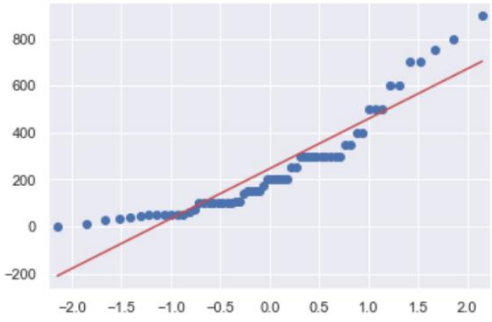
e. Shapiro's test: In shapiro's test, if p value is less than alpha (i.e. 0.05), we can say that data is not normally distributed. Here, Null hypothesis would be data is normally distributed and alternate hypothesis would be data is not normally distributed.

```
## Shapiro test
stat, p = shapiro(df1)
print('Statistics=%.3f, p=%.6f' % (stat, p))
alpha = 0.05
TM_Shapiro = np.where(p > alpha, "normally distributed", "not normally distributed")
print('Shapiro test for text messages variable stats that data is', TM_Shapiro)
```

```
Statistics=0.859, p=0.000004
Shapiro test for text messages variable stats that data is not normally distributed
```

In text messages variable, p value is 0.000004 which is lesser then alpha value (p low -> null go). Hence, **it is not normally distributed.**

f. QQ plot: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



In Text messages variable, *QQ plot showing the scatter plot of points in a diagonal line, which is not fitting the expected diagonal pattern*. Hence, it is **not normally distributed.**

Conclusion: Based on above justifications, we have seen that spending variable has mean greater than median which means skewness is positive side. The IQR and range are also lesser than what would be expected in normal distribution. In Box plot we can see box are same, but whisker is skewed longer in right side then the left side. Thus, we can conclude that the data characteristics of text messages variable **differs from theoretical/visual properties of normal distribution.**

Problem3:

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 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.

For the A shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

H0<=0.35

HA>0.35

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

H0<=0.35

HA>0.35

Remarks: All the questions of problem3 which explained here, are also performed in python as well. Please refer python notebook "SMDM_Group_Assignment_GRP6_Problem3".

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?

Background: Company is manufacturing asphalt shingles. Recently they got feedback from customer that shingles contains excessive moisture (i.e. quality issue). Company wants perform hypothesis testing to two samples (i.e. A and B) that compare statistic from samples selected from two populations.

Assumption: If we assume that the random samples are **independently selected from two populations** and that the **populations are normally distributed and have equal variances** then we can use **pooled variance t test** (i.e. Means of two independence populations) to determine whether there is a significant difference between the means. We are also assuming **level of significance (α) = 0.05 (5%)**

Approach: we will follow **6 steps** approach to perform pooled variance t test (i.e. Means of two independence populations)

Step1: Identify Data Given

#	Data points	A Shingles	B Shingles
1	Sample measurement count	$n_a = 36$	$n_b = 31$
2	Sample mean	$\bar{X}_a = 0.317$	$\bar{X}_b = 0.274$
3	Sample standard deviation	$S_a = 0.136$	$S_b = 0.137$
4	Level of significance	$\alpha = 0.05$	

Step2: Define Hypothesis (H0 and H1)

Using subscripts to distinguish between the population means of the first population (i.e. A shingles) μ_a, and the population mean of the second population (i.e. B shingles) μ_b.

The null hypothesis of no difference in the means of two independent populations can be stated as

H_0 (null hypothesis): $\mu_a = \mu_b$ OR $\mu_a - \mu_b = 0$

and the alternative hypothesis, that means are different can be stated as

H_1 (alternative hypothesis): $\mu_a \neq \mu_b$ OR $\mu_a - \mu_b \neq 0$

Step3: Calculate tSTAT

Pooled variance t test for the difference between two means

$$t_{STAT} = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{\sqrt{S_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}}$$

WHERE

$$S_p^2 = \frac{(n_a - 1)S_a^2 + (n_b - 1)S_b^2}{(n_a - 1) + (n_b - 1)}$$

S_p^2 = Pooled variance
$\bar{X}_a$ = sample mean of A shingles
$\bar{X}_b$ = sample mean of B shingles
n_a = sample count of A shingles
n_b = sample count of B shingles
S_a = sample variance of A shingles

AND S_b = sample variance of B shingles

To calculate t_{STAT}, first we need to calculate S_p^2

$$S_p^2 = \frac{(36-1)(0.136)^2 + (31-1)(0.137)^2}{(36-1)+(31-1)}$$

$$S_p^2 = \frac{(0.647) + (0.563)}{(65)}$$

$$S_p^2 = 0.0186$$

Now let's calculate t_{STAT}

$$t_{STAT} = \frac{(0.317 - 0.274) - (0)}{\sqrt{(0.0186)^2\left(\frac{1}{36} + \frac{1}{31}\right)}}$$

$$= \frac{(0.317 - 0.274) - (0)}{\sqrt{(0.0186)\left(\frac{1}{36} + \frac{1}{31}\right)}}$$

$$= \frac{(0.043)}{\sqrt{(0.001117)}}$$

$$= \frac{(0.043)}{(0.03341)}$$

$$t_{STAT} = 1.289$$

Step4: Identify the Reject Zone

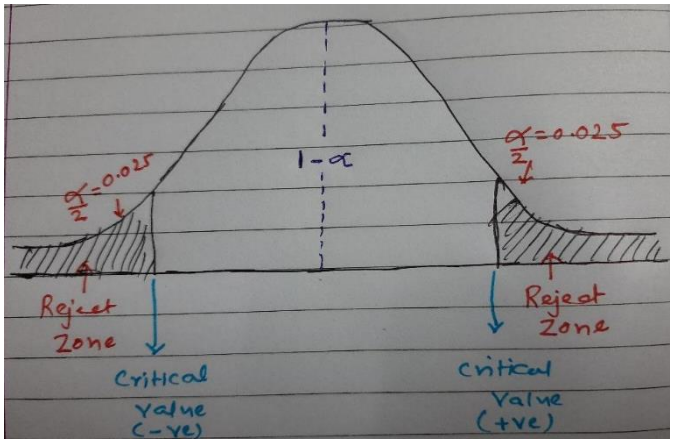


Figure 34 - Rejection Zone

For a given level of significance, α = 0.05, in a two-tail test, we should reject the null hypothesis if the computed t_{STAT} is **greater than the upper-tail critical value** from the t distribution or if the computed t_{STAT} is **less than the lower-tail critical value** from the t distribution.

Step5(a): Find out tCRIT

To find out t_{CRIT}, we need degrees of freedom (dof) and (α/2):

dof = $(n_a + n_b - 2)$ = (36 + 31 -2) = **65**

α/2 = (0.05/2) = **0.025**

Now refer below table to see t_{CRIT} based on (dof) and (α/2)

Degrees of Freedom	Upper-Tail Areas			
	0.25	0.10	0.05	0.025
51	0.6793	1.2984	1.6753	2.0076
52	0.6792	1.2980	1.6747	2.0066
53	0.6791	1.2977	1.6741	2.0057
54	0.6791	1.2974	1.6736	2.0049
55	0.6790	1.2971	1.6730	2.0040
56	0.6789	1.2969	1.6725	2.0032
57	0.6788	1.2966	1.6720	2.0025
58	0.6787	1.2963	1.6716	2.0017
59	0.6787	1.2961	1.6711	2.0010
60	0.6786	1.2958	1.6706	2.0003
61	0.6785	1.2956	1.6702	1.9996
62	0.6785	1.2954	1.6698	1.9990
63	0.6784	1.2951	1.6694	1.9983
64	0.6783	1.2949	1.6690	1.9977
65	0.6783	1.2947	1.6686	1.9971

Figure 35 - t table to check t CRIT value

$$t_{CRIT} = \pm 1.997$$

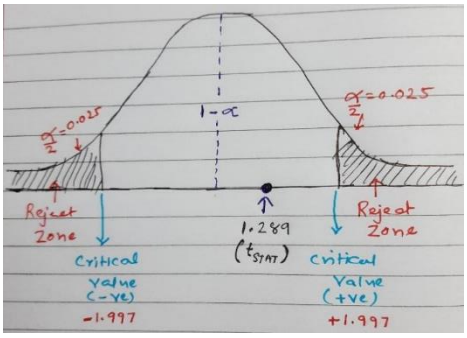
Step5(b): Calculate pvalue

To calculate p_{value}, we need to use excel function as below:

p_{value} = T.DIST.2T(t_{STAT}, dof) = T.DIST.2T(1.289,65) = **0.201**

$$p_{value} = 0.201$$

Step6(a): Compare tSTAT with tCRIT



We **fail to reject the null hypothesis** because **t_{STAT} (1.289)** value is less than **t_{CRIT} (1.997)** zone.

Step6(b): Compare pvalue with α

$$p_{value} = 0.201 > \alpha = 0.05$$

This p_{value} indicates that if the population means are equal, the probability of the observing a difference in the two-sample means is 0.201. Because the p_{value} (0.201) is greater than α (0.05), there is **no sufficient evidence to reject the null hypothesis**.

So, we can conclude that the mean of A shingles and B shingles are same.

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

Assumption: The random samples in the **populations are normally distributed and have equal variances** with **level of significance (α) = 0.05 (5%).**