SMDM PROJECT REPORT

PGP-DSBA

SRIKANTH PR 8/9/20 SMDM Project

Contents:

1 Problem 1: Descriptive Statistics:

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).

Attribute Information:

Region: [Nominal Data] [Categorical Variable]

- Lisbon
- Oporto
- Other

Channel: [Nominal Data] [Categorical Variable]

- Hotel
- Retail

Products: [Continuous Data] [Numerical Variable]

- Fresh Products.
- Milk Products.
- Grocery Products.
- Frozen Products.
- Detergent_paper Products.
- Delicatessen Products.

Exploratory Data Analysis:

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

Dataset has 9 variables; Buyer/Spender has unique row number for every transaction detail. There are 2 types of channel (Hotel and Retail). There are 3 Regions (Lisbon, Oporto and Other) and rest are the 6 varieties for which the spending has been provided.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    object
 1   Region            440 non-null    object
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicatessen      440 non-null    int64
dtypes: int64(6), object(2)
memory usage: 27.6+ KB
```

Info of the Dataset

All the variables are in numerical format except Region and channel are in object format.
There are total 440 observations and 8 columns in the dataset.
There are no missing values found in the dataset.

1.1 Use methods of descriptive statistics to summarize data:
Which Region and which Channel seems to spend more?
Which Region and which Channel seems to spend less?

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

From the above descriptive statistics, average spending on Fresh is 1200, Milk is 5796, Grocery is 7951, Frozen is 3071, Detergent_paper is 2881 and Delicatessen is 1525. From this result we can say that highest spending amount is on grocery.

Calculate median for all the variables:

```
The Median of Fresh is 8504.0
The Median of Milk is 3627.0
The Median of Grocery is 4755.5
The Median of Frozen is 1526.0
The Median of Detergents_Paper is 816.5
The Median of Delicatessen is 965.5
```

From the above result, we could show that the median values is same as 50th percentile of the dataset. Since the mean and median of all the six variables are not the same, we can say that the variables are highly skewed.

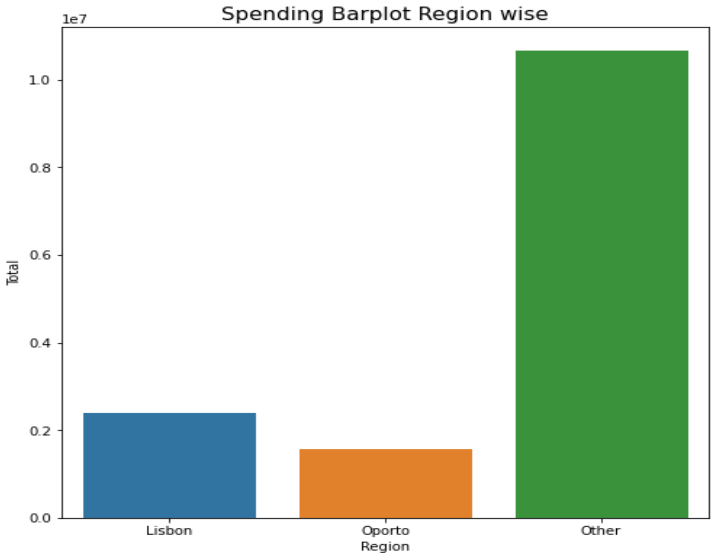
Mode Calculation:

The "mode" is the value that occurs most often

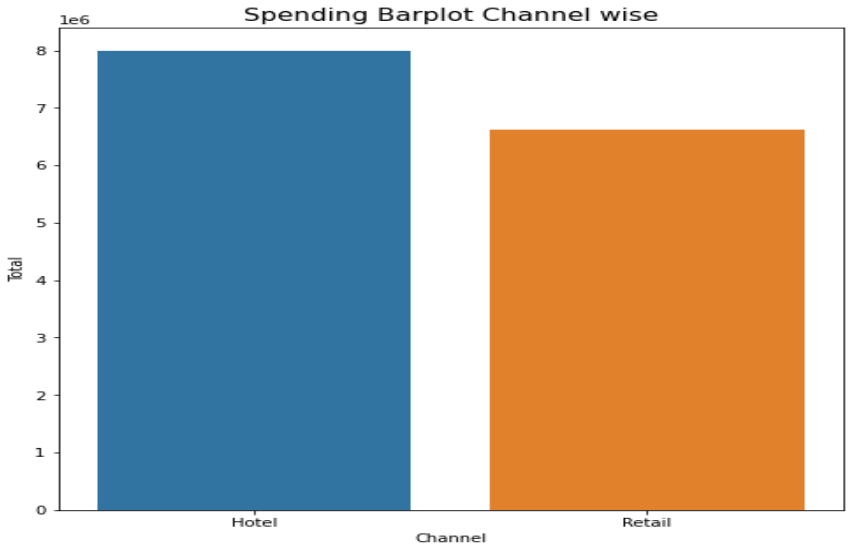
Since all the variables except Region and channel is unique numerical values, there will be no mode. We can find the mode of Region and channel.

The most often occur value (mode) of Region is **Other**.
The most often occur value (mode) of Channel is **Hotel**.

That means that the most frequent value present in the Region column is **other** and the most frequent value in the channel is **Hotel**.

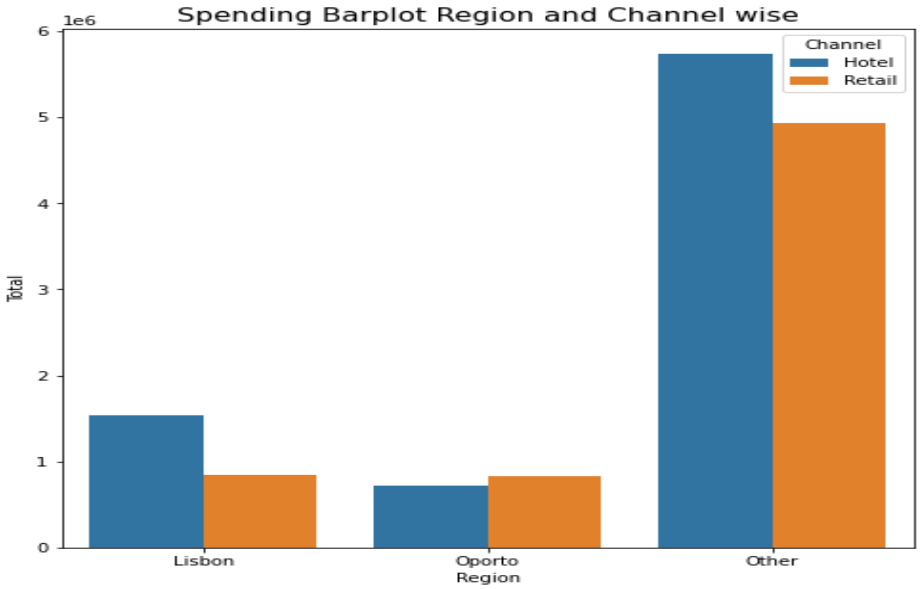


From the above bar plot, we can see that other region is spending the highest and Oporto Region is spending the least.



From the above Results, We find out:

1. The 'Hotel' channel spent more amount: 7,999,569 Euro.
2. The 'Retail' channel spent less amount: 6,619,931 Euro.



From the above bar plot, Other Region which has the highest spending's, hotel channel has more spending's than retail channel. Same pattern can be seen for Lisbon. However, in Oporto Region, Retail has the more spending as compare to the Hotel channel.

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

To check behavior of 6 different varieties, we will subset the dataset with respect to region and channel and analyze the descriptive statistics.

Analysis of varieties in different channel:

- **Retail.describe()**

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	142.000000	142.000000	142.000000	142.000000	142.000000	142.000000
mean	8904.323944	10716.500000	16322.852113	1652.612676	7269.507042	1753.436620
std	8987.714750	9679.631351	12267.318094	1812.803662	6291.089697	1953.797047
min	18.000000	928.000000	2743.000000	33.000000	332.000000	3.000000
25%	2347.750000	5938.000000	9245.250000	534.250000	3683.500000	566.750000
50%	5993.500000	7812.000000	12390.000000	1081.000000	5614.500000	1350.000000
75%	12229.750000	12162.750000	20183.500000	2146.750000	8662.500000	2156.000000
max	44466.000000	73498.000000	92780.000000	11559.000000	40827.000000	16523.000000

- Total Count of spending's done by Retail is 142. From varying standard deviation ranging from (1812 to 12267) with high range. We found that all the variables don't show similar behavior.
- The minimum amount spent on Grocery is the highest and Delicatessen is the lowest.

- **Hotel.decsribe()**

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	298.000000	298.000000	298.000000	298.000000	298.000000	298.000000
mean	13475.560403	3451.724832	3962.137584	3748.251678	790.560403	1415.956376
std	13831.687502	4352.165571	3545.513391	5643.912500	1104.093673	3147.426922
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	4070.250000	1164.500000	1703.750000	830.000000	183.250000	379.000000
50%	9581.500000	2157.000000	2684.000000	2057.500000	385.500000	821.000000
75%	18274.750000	4029.500000	5076.750000	4558.750000	899.500000	1548.000000
max	112151.000000	43950.000000	21042.000000	60869.000000	6907.000000	47943.000000

- Total Count of spending's done by Hotel is 298. From varying standard deviation ranging from (1104 to 13832) with high range. We found that all the variables don't show similar behavior.
- The minimum amount spend on Milk is the highest. There are 4 variables on which hotel has spent the same amount of minimum amount of 3.

Analysis of varieties in different Region:

- **Lisbon.describe()**

- Total Count of spending's done by Lisbon is 77. From varying standard deviation ranging from (1345 to 11557) with high range. We found that all the variables don't show similar behavior for Lisbon Region.
- The minimum amount spend on Grocery is the highest and Detergents Paper is the lowest.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	11101.727273	5486.415584	7403.077922	3000.337662	2651.116883	1354.896104
std	11557.438575	5704.856079	8496.287728	3092.143894	4208.462708	1345.423340
min	18.000000	258.000000	489.000000	61.000000	5.000000	7.000000
25%	2806.000000	1372.000000	2046.000000	950.000000	284.000000	548.000000
50%	7363.000000	3748.000000	3838.000000	1801.000000	737.000000	806.000000
75%	15218.000000	7503.000000	9490.000000	4324.000000	3593.000000	1775.000000
max	56083.000000	28326.000000	39694.000000	18711.000000	19410.000000	6854.000000

- **Oporto.describe()**

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.000000
mean	9887.680851	5088.170213	9218.595745	4045.361702	3687.468085	1159.702128
std	8387.899211	5826.343145	10842.745314	9151.784954	6514.717668	1050.739841
min	3.000000	333.000000	1330.000000	131.000000	15.000000	51.000000
25%	2751.500000	1430.500000	2792.500000	811.500000	282.500000	540.500000
50%	8090.000000	2374.000000	6114.000000	1455.000000	811.000000	898.000000
75%	14925.500000	5772.500000	11758.500000	3272.000000	4324.500000	1538.500000
max	32717.000000	25071.000000	67298.000000	60869.000000	38102.000000	5609.000000

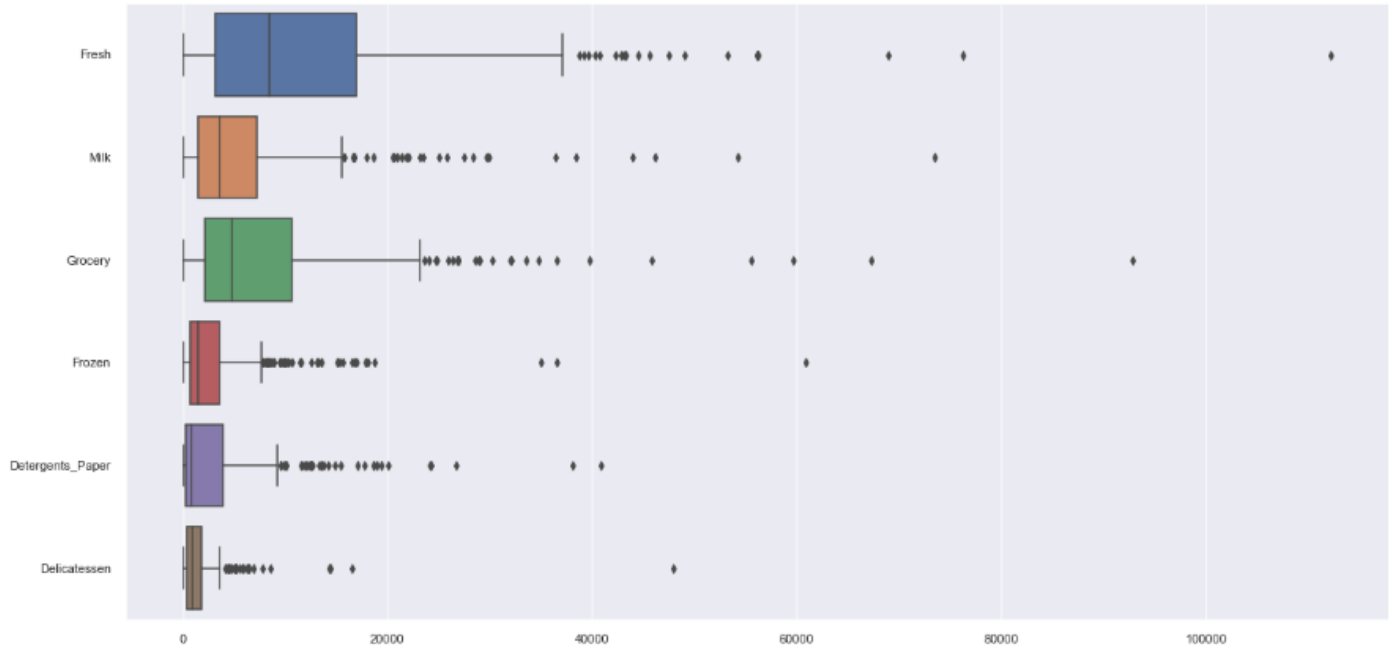
- Total Count of spending's done by Oporto is 47. From varying standard deviation ranging from (1050 to 10843) with high range. We found that all the variables don't show similar behavior for Oporto Region.
- The minimum amount spend on Grocery is the highest and Fresh is the lowest.

- **Other.describe()**

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	316.000000	316.000000	316.000000	316.000000	316.000000	316.000000
mean	12533.471519	5977.085443	7896.363924	2944.594937	2817.753165	1620.601266
std	13389.213115	7935.463443	9537.287778	4260.126243	4593.051613	3232.581660
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3350.750000	1634.000000	2141.500000	664.750000	251.250000	402.000000
50%	8752.500000	3684.500000	4732.000000	1498.000000	856.000000	994.000000
75%	17406.500000	7198.750000	10559.750000	3354.750000	3875.750000	1832.750000
max	112151.000000	73498.000000	92780.000000	36534.000000	40827.000000	47943.000000

- Total Count of spending done by other is 316. From varying standard deviation ranging from (3232 to 13389) with high range. We found that all the variables don't show similar behavior for Other Region.
- The minimum amount spend on Milk is the highest and There are 4 variables on which other region has spent the same amount of minimum amount of 3.

1.3 Are there any outliers in the data?



- From the above plot: We can say that all the 6 items in the wholesale dataset have extreme Values.
- The List of Outliers are as follows:
  ['Fresh', 'Milk, 'Grocery', 'Frozen', 'Detergent Paper', 'Delicatessen']

1.4 On the basis of the descriptive measure of variability:
Which item shows the most inconsistent behaviour?
Which item shows the least inconsistent behaviour?

Descriptive measures of variability is used to describe the amount of spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, standard deviation. We will use coefficient of variation here.

The Coefficient of variation (CV) is a statistical measure of the dispersion of data points in a series around the mean. It is a useful statistic for comparing the degree of variation from one data series to another, even if the means are very different from one another.

$$CV = \sigma/\mu$$

Where σ = standard deviation and μ = mean.

```
The coefficient of variation for Fresh is 1.053
The coefficient of variation for Milk is 1.272
The coefficient of variation for Grocery is 1.194
The coefficient of variation for Frozen is 1.579
The coefficient of variation for Detergents_Paper is 1.653
The coefficient of variation for Delicatessen is 1.847
```

From the above result of CV, we found that Delicatessen shows the most inconsistent behavior and Fresh shows the least inconsistent behavior.

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

From the all the analysis done, below are the Observations and recommendations:

1. Out of all the regions, Other region is spending the highest and Oporto is spending the lowest.
2. Hotel is spending more than Retail.
3. Out of all the 6 varities, the highest spending was done on Fresh followed by Grocery, Milk, Frozen, Detergent_papers, Delicatessen.
4. There are outliers present in the dataset.

2 Problem 2: Probability and it's Distribution:

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).

Exploratory Data Analysis:

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100

- Dataset has 14 variables, which has the different values for the particular response. ID is the variable which has the unique row number for each response.

Data-types of the Variables:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 62 non-null     int64
 1   Gender             62 non-null     object
 2   Age                62 non-null     int64
 3   Class              62 non-null     object
 4   Major              62 non-null     object
 5   Grad Intention     62 non-null     object
 6   GPA                62 non-null     float64
 7   Employment         62 non-null     object
 8   Salary             62 non-null     float64
 9   Social Networking  62 non-null     int64
 10  Satisfaction       62 non-null     int64
 11  Spending           62 non-null     int64
 12  Computer           62 non-null     object
 13  Text Messages      62 non-null     int64
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

2.1 For this data, construct the following contingency tables:

2.1.1 Gender and Major

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

2.1.2 Gender and Grad Intention

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

2.1.3 Gender and Employment

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

2.1.4 Gender and Computer

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

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

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

Male Probability = (Total number of male students)/ (Total number of students at the university)

Prob_male = 29/62 =0.4677

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

Female Probability = (Total number of Female students)/ (Total number of students at the university)

Prob_Female = 33/62 =0.532

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.?

- By looking into the contingency table of 'Gender and Majors' and by using Conditional probability formula,We can calculate the conditional probability of different Majors given the probability of Gender:

- Conditional Probability Formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Where:
P(A|B) = Probability of Event A given Event B has occured.
P(A and B) = Probability of Event A occured and Event B occured.
P(B) = Probability of Event B.

- Calculating Conditional Probability for different Majors among the Male are as below:

Formula to calculate Different Majors among the Male:

- $P(Accounting|Male) = \frac{P(Accounting \cap Male)}{P(Male)}$
- $P(CIS|Male) = \frac{P(CIS \cap Male)}{P(Male)}$
- $P(Economics - Finance|Male) = \frac{P(Economics - Finance \cap Male)}{P(Male)}$
- $P(International\,Business|Male) = \frac{P(International\,Business \cap Male)}{P(Male)}$
- $P(Management|Male) = \frac{P(Management \cap Male)}{P(Male)}$
- $P(Retailing - Marketing|Male) = \frac{P(Retailing - Marketing \cap Male)}{P(Male)}$
- $P(Others|Male) = \frac{P(Others \cap Male)}{P(Male)}$
- $P(Undecided|Male) = \frac{P(Undecided \cap Male)}{P(Male)}$

From the Gender and Major contingency table,we get:

- $P(Male) = \frac{29}{62}$
- $P(Accounting \cap Male) = \frac{4}{62}$
- $P(CIS \cap Male) = \frac{1}{62}$
- $P(Economics - Finance \cap Male) = \frac{4}{62}$
- $P(International\,Business \cap Male) = \frac{2}{62}$
- $P(Management \cap Male) = \frac{6}{62}$
- $P(Retailing - Marketing \cap Male) = \frac{5}{62}$
- $P(Others \cap Male) = \frac{4}{62}$
- $P(Undecided \cap Male) = \frac{3}{62}$

Final Output of different Majors among the Male students:

- $P(Accounting|Male) = \frac{4}{29}$
- $P(CIS|Male) = \frac{1}{29}$
- $P(Economics - Finance|Male) = \frac{4}{29}$
- $P(International Business|Male) = \frac{2}{29}$
- $P(Management|Male) = \frac{6}{29}$
- $P(Retailing - Marketing|Male) = \frac{5}{29}$
- $P(Others|Male) = \frac{4}{29}$
- $P(Undecided|Male) = \frac{3}{29}$

➢ [Insights Drawn from the above Caculation]:

The probabilitiy of CMSU Male students selecting :
- Accounting has a major is 13.80%
- CIS has a major is 3.45%
- Economics-Finance has a major is 13.80%
- International Business has a major is 6.90%
- Management has a major is 20.68%
- Retailing-Marketing has a major is 17.25%

➢ From the above results we can say selecting Management has a Major among the Male students is very high and selecting CIS has a major among the Male students is very less.
➢ Also, we can see, 10.35 % among the Male students have still not decided that which major has to be selected.
➢ 13.80 % among the male students opted Others Major apart from the one listed.

- Calculating Conditional Probability for different Majors among the Female are as below:

Formula to calculate Different Majors among the Female:

- $P(Accounting|Female) = \frac{P(Accounting \cap Female)}{P(Female)}$
- $P(CIS|Female) = \frac{P(CIS \cap Female)}{P(Female)}$
- $P(Economics - Finance|Female) = \frac{P(Economics - Finance \cap Female)}{P(Female)}$
- $P(International Business|Female) = \frac{P(International Business \cap Female)}{P(Female)}$
- $P(Management|Female) = \frac{P(Management \cap Female)}{P(Female)}$
- $P(Retailing - Marketing|Female) = \frac{P(Retailing - Marketing \cap Female)}{P(Female)}$
- $P(Others|Female) = \frac{P(Others \cap Female)}{P(Female)}$
- $P(Undecided|Female) = \frac{P(Undecided \cap Female)}{P(Female)}$

From the Gender and Major contingency table,we get:

- $P(Female) = \frac{33}{62}$
- $P(Accounting \cap Female) = \frac{3}{62}$
- $P(CIS \cap Female) = \frac{3}{62}$
- $P(Economics - Finance \cap Female) = \frac{7}{62}$
- $P(International Business \cap Female) = \frac{4}{62}$
- $P(Management \cap Female) = \frac{4}{62}$
- $P(Retailing - Marketing \cap Female) = \frac{9}{62}$
- $P(Others \cap Female) = \frac{3}{62}$
- $P(Undecided \cap Female) = \frac{0}{62}$

Final Output of different Majors among the Female students:

- $P(Accounting|Female) = \frac{3}{33}$
- $P(CIS|Female) = \frac{3}{33}$
- $P(Economics - Finance|Female) = \frac{7}{33}$
- $P(International Business|Female) = \frac{4}{33}$
- $P(Management|Female) = \frac{4}{33}$
- $P(Retailing - Marketing|Female) = \frac{9}{33}$
- $P(Others|Female) = \frac{3}{33}$
- $P(Undecided|Female) = 0$

➢ [Insights Drawn from the above Caculation]:

The probabilitiy of CMSU Female students selecting :

- Accounting has a major is 9%
- CIS has a major is 9%
- Economics-Finance has a major is 21.21%
- International Business has a major is 12.12%
- Management has a major is 12.12%
- Retailing-Marketing has a major is 27.27%

➢ From the above results we can say selecting 'Economics-Finance' and 'Retailing-Marketing' has a Major among the Female students is very high.
➢ Selecting 'Accounting' and 'CIS' has a major is of same percentage and also with 'Management' and 'International Business' is been distributed equally among the female students.

- Also, it's happy to see that all the female students have opted one or the other major when compared to Male students where 10.35 % Male students have still not decided that which major has to be selected.

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.?

- Calculating Conditional Probability for Grad Intention among the Male are as below:

Formula to calculate conditional probability of intent to graduate among the Male:

- $P(YES|Male) = \frac{P(YES \cap Male)}{P(Male)}$
- $P(NO|Male) = \frac{P(NO \cap Male)}{P(Male)}$
- $P(UNDECIDED|Male) = \frac{P(UNDECIDED \cap Male)}{P(Male)}$

From the Gender and Grad_Intention contingency table,we get:

- $P(Male) = \frac{29}{62}$
- $P(Yes \cap Male) = \frac{17}{62}$
- $P(NO \cap Male) = \frac{3}{62}$
- $P(Undecided \cap Male) = \frac{9}{62}$

Final Output of different Grad_Intention among the Male students:

- $P(Yes|Male) = \frac{17}{29}$
- $P(NO|Male) = \frac{3}{29}$
- $P(UNDECIDED|Male) = \frac{9}{29}$

[Insights Drawn from the above Caculation]:

- The probability of intent to graduate among the male students is 58.62 %
- The probability of the male students who do not wish to or intent to graduate is 10.34%.
- The probability of the male students who have still not decided to graduate is of 31.04%.

- Calculating Conditional Probability for Grad Intention among the Female are as below:

Formula to calculate conditional probability of intent to graduate among the Female:

- $P(YES|Female) = \frac{P(YES \cap Female)}{P(Female)}$
- $P(NO|Female) = \frac{P(NO \cap Female)}{P(Female)}$
- $P(UNDECIDED|Female) = \frac{P(UNDECIDED \cap Female)}{P(Female)}$

From the Gender and Grad_Intention contingency table,we get:

- $P(Female) = \frac{33}{62}$
- $P(Yes \cap Female) = \frac{11}{62}$
- $P(NO \cap Female) = \frac{9}{62}$
- $P(Undecided \cap Female) = \frac{13}{62}$

Final Output of different Grad_Intention among the Female students:

- $P(Yes|Female) = \frac{11}{33}$
- $P(NO|Female) = \frac{9}{33}$
- $P(UNDECIDED|Female) = \frac{13}{33}$

[Insights Drawn from the above Caculation]:

- The probability of intent to graduate among the Female students is of 33.33 %
- The probability of the female students who do not wish to or intent to graduate is 27.27%.
- The probability of the female students who have still not decided to graduate is 39.40%.

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

- Calculating Conditional Probability of Employement status among the Male are as below:

Formula to calculate conditional probability of Employment status among the Male:

- $P(Full - Time|Male) = \frac{P(Full-Time \cap Male)}{P(Male)}$
- $P(Part - Time|Male) = \frac{P(Part-Time \cap Male)}{P(Male)}$
- $P(Unemployed|Male) = \frac{P(Unemployed \cap Male)}{P(Male)}$

From the Gender and Employment contingency table,we get:

- $P(Male) = \frac{29}{62}$
- $P(Full - Time \cap Male) = \frac{7}{62}$
- $P(Part - Time \cap Male) = \frac{19}{62}$
- $P(Unemployed \cap Male) = \frac{3}{62}$

Final Output of different Employement status among the Male students:

- $P(Full-Time|Male) = \frac{7}{29}$
- $P(Part-Time|Male) = \frac{19}{29}$
- $P(Unemployed|Male) = \frac{3}{29}$

[Insights Drawn from the above Caculation]:

- The probability of Full-Time Employee among the male students is of 24.14 %
- The probability of Part-Time Employee among the male students is of 65.52 %
- The probability of UnEmployed among the male students is of 10.34 %

- Calculating Conditional Probability of Employement status among the Male are as below:

Formula to calculate conditional probability of Employment status among the Female:

- $P(Full-Time|Female) = \frac{P(Full-Time \cap Female)}{P(Female)}$
- $P(Part-Time|Female) = \frac{P(Part-Time \cap Female)}{P(Female)}$
- $P(Unemployed|Female) = \frac{P(Unemployed \cap Female)}{P(Female)}$

From the Gender and Employment contingency table,we get:

- $P(Female) = \frac{33}{62}$
- $P(Full-Time \cap Female) = \frac{3}{62}$
- $P(Part-Time \cap Female) = \frac{24}{62}$
- $P(Unemployed \cap Female) = \frac{6}{62}$

Final Output of different Employment status among the Female students:

- $P(Full-Time|Female) = \frac{3}{33}$
- $P(Part-Time|Female) = \frac{24}{33}$
- $P(Unemployed|Female) = \frac{6}{33}$

[Insights Drawn from the above Caculation]:

- The probability of Full-Time Employee among the female students is of 9 %
- The probability of Part-Time Employee among the female students is of 72.72 %
- The probability of UnEmployed among the female students is of 18.18 %

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

Formula to calculate conditional probability of Laptop preference among the Male:

- $P(Laptop|Male) = \frac{P(Laptop \cap Male)}{P(Male)}$

From the Gender and Computer contingency table,we get:

- $P(Male) = \frac{29}{62}$
- $P(Laptop \cap Male) = \frac{26}{62}$

Final Output of Laptop preference among the Male students:

- $P(Laptop|Male) = \frac{26}{29}$

[Insights Drawn from the above Caculation]:

- The probability of prefering laptop among the male students is of 89.65 %

Formula to calculate conditional probability of Laptop preference among the Female:

- $P(Laptop|Female) = \frac{P(Laptop \cap Female)}{P(Male)}$

From the Gender and Computer contingency table,we get:

- $P(Female) = \frac{33}{62}$
- $P(Laptop \cap Female) = \frac{29}{62}$

Final Output of Laptop preference among the Female students:

- $P(Laptop|Female) = \frac{29}{33}$

[Insights Drawn from the above Caculation]:

- The probability of prefering laptop among the female students is of 87.87 %

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

- Two events A and B are independent if and only if the below condition is satisfied:

$$P(A|B) = P(A)$$

Where:

P(A|B) : Conditional probability of A given B
P(A) : Marginal probability of A

- The Conditional probability of different major among the male or the female students is not equal to the marginal probability of Major, So the two events "Major" and "Gender" are not Independents events.

- The Conditional probability of Grad-Intention among the male or the female students is not equal to the marginal probability ofGrad-Intention, So the two events "Intent to Graduate" and "Gender" are not Independents events.

- The Conditional probability of Employement status among the male or the female students is not equal to the marginal probability of Employement status, So the two events "Employement status" and "Gender" are not Independents events.

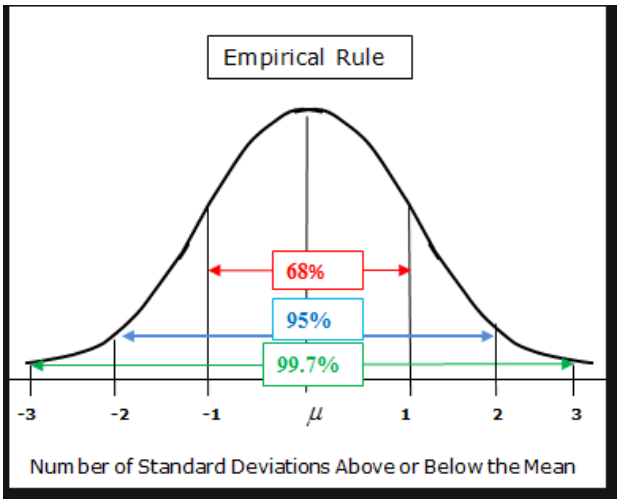
- The Conditional probability of Laptop preference among the male or the female students is not equal to the marginal probability of Laptop preference, So the two events "Laptop preference" and "Gender" are not Independents events.

- Overall in each case, the column variable is not Independent of Gender.

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 and Write a note summarizing your conclusions.

[Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric].

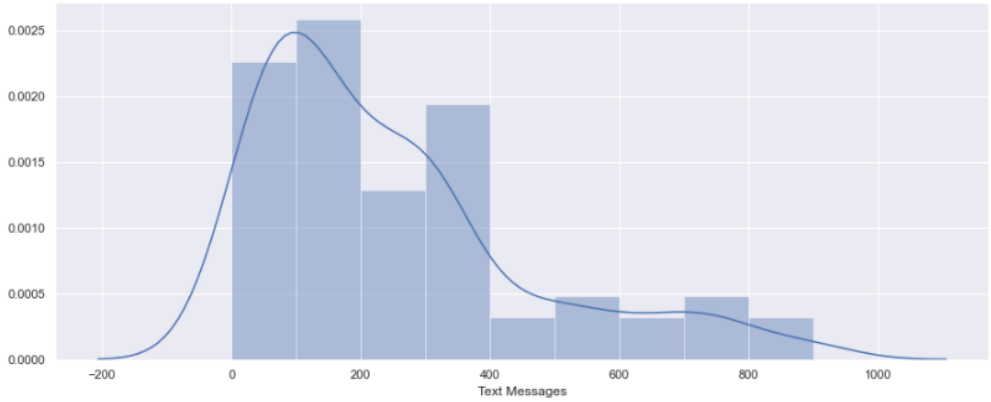
- For a continuous data to follow a normal distribution, it must satisfy emperical rule condition.
- Emperical rule states that:
  - ✓ 68 % of the data in a perfect bell shaped distribution lies within One standard deviation : μ ± σ.
  - ✓ 95 % of the data in a perfect bell shaped distribution lies within two standard deviation : μ ± 2σ.
  - ✓ 99.7 % of the data in a perfect bell shaped distribution lies within Three standard deviation : μ ± 3σ.
- Perfect Bell shape or Normal Distribution look like the below graph:



- For a Normal Distribution, the mean, median and mode all the three values are equal.
- The curve is symmetric at the centre.

1. Check for Text Message Numerical variable whether it follows normal Distribution or not?

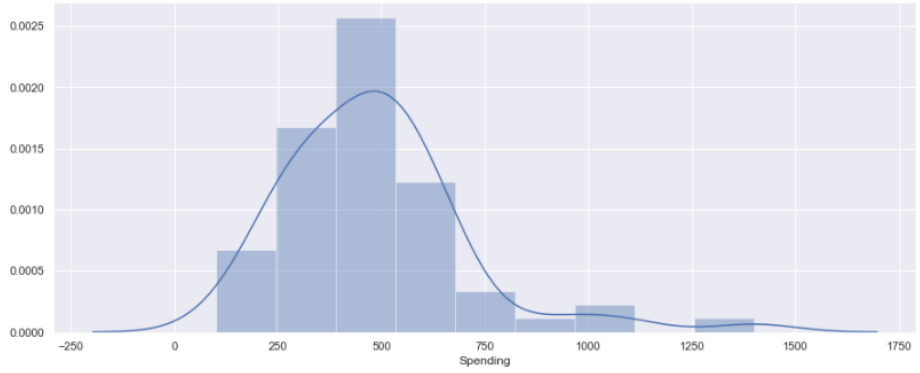
[Output]: Using histogram and KDE plot trying to visualize whether the data is distributed normally or not.



- From the above plot we can clearly see that the data is not symmetry around the mean and the data does not follow normal distribution.
- We can still confirm and get accurate ,by caculating the cumulative density function of normal distribution.
- For "Text Messages" Numerical Variable, the data lies within the values for :
  - ✓ One standard Deviation is : 42.53 %
  - ✓ Two standard Deviation is : 80.20 %
  - ✓ Three standard Deviation is : 96.79%
- So the above results failed to follow the emperical rule 68-95-99.7 % to follow a perfect bell shaped or a Normal distribution.
- Hence, Text Message Numerical variable do not follow a Normal Distribution.

2. Check for "Spending'' Numerical variable whether it follows normal Distribution or not?

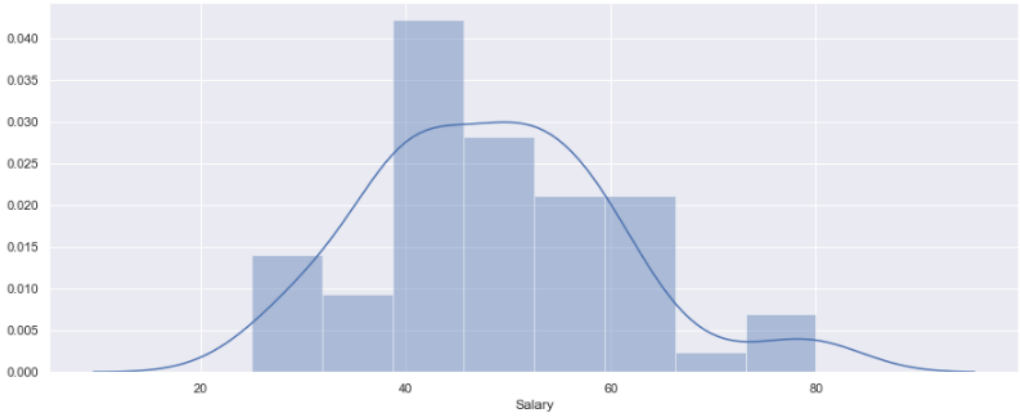
[Output]: Using histogram and KDE plot trying to visualize whether the data is distributed normally or not.



- From the above plot we can clearly see that the data is not symmetry around the mean and the data does not follow normal distribution.
- We can still confirm and get accurate results,by caculating the cumulative density function of normal distribution.
- For "Spending" Numerical variable the data lies within the values for :
  - ✓ One standard Deviation is : 11.98 %
  - ✓ Two standard Deviation is : 43.18 %
  - ✓ Three standard Deviation is : 79.62%
- So the above results failed to follow the emperical rule 68-95-99.7 % to call it a perfect bell shaped or a Normal distribution.
- Hence, Spending Numerical variable do not follow a Normal Distribution but tends to approach Normal Distribution by adding more number of sample size.

3. Check for "Salary" Numerical variable whether it follows normal Distribution or not?

[Output]: Using histogram and KDE plot trying to visualize whether the data is distributed normally or not.



- From the above plot we can clearly see that the data is not symmetry around the mean and the data does not follow normal distribution.
- To get accurate results, caculate the cumulative density function of normal distribution.
- For "Salary" Numerical variable the data lies within the values for :
  - ✓ One standard Deviation is : 0.12 %
  - ✓ Two standard Deviation is : 2.17 %
  - ✓ Three standard Deviation is :15.40 %
- So the above results failed to follow the emperical rule 68-95-99.7 % to call it a perfect bell shaped or a Normal distribution.
- Hence, Salary Numerical variable do not follow a Normal Distribution.
- As per the Central Limit theorem, with growing sample size, the data tends to follow normal distribution, So with increasing sample size we can attain the Normal distribution for all 3 numerical variables.

3 Problem 3: Hypothesis Testing

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company claims that the mean moisture content cannot be greater than 0.35 pound per 100 square feet.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

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:

H_0<=0.35

H_A>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:

H_0<=0.35

H_A>0.35

Exploratory Data Analysis:

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

- Dataset has 2 variables A and B, which has the measurement of moisture present per 100 sq.ft
- Both variables are float in data types.

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

The company claims that moisture content is less than 0.35 lbs/100 sqft.

Null hypothesis is the claim or the status quo. Only under strong evidence, the null hypothesis is to be rejected.

Null hypothesis states that mean moisture content is: μ<= 0.35 pound per 100 square feet
Alternate hypothesis states that: μ > 0.35 pound per 100 square feet

H_0: μ <= 0.35
H_A: μ > 0.35

```
The test statistic is -1.4735046253382782
```

```
The p-value is 0.9252236685509249 which is greater than the level of significance, hence we fail to reject the Null hypothesis
```

Conclusion: Null hypothesis that mean moisture content is less than 0.35 lbs/100 sqft cannot be rejected.

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

The company claims that moisture content is less than 0.35 lbs/100 sqft.

Null hypothesis is the claim or the status quo. Only under strong evidence, the null hypothesis is to be rejected.

Null hypothesis states that mean moisture content is: μ<= 0.35 pound per 100 square feet
Alternate hypothesis states that: μ > 0.35 pound per 100 square feet

H_0: μ <= 0.35
H_A: μ > 0.35

```
The test statistic is -3.1003313069986995
```

```
The p-value is 0.9979095225996808 which is greater than the level of significance, hence we fail to reject the Null hypothesis
```

Conclusion: Null hypothesis that mean moisture content is less than 0.35 lbs/100 sqft cannot be rejected.

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?

- Formulate Null and Alternate hypothesis:

  - In testing, whether the population mean weight of moisture content are same in both the shingles of ABC asphalt company, the null hypothesis states that the mean weight of moisture content of shingles A and B are the same, μA equals μB. The alternative hypothesis states that the mean weight of moisture content are different, μA is not equal to μB.

  - H_0 : μA = μB
  - H_1 : μA ≠ μB

- The assumptions for t-test:

  1. The first assumptions made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale.
  2. The second assumption made is that of a simple random sample, that the data is collected from a randomly selected portion of the total population.
  3. The distribution of the moisture content in both populations follows a normal distribution.
  4. The fourth assumption is a reasonably large sample size is used.
  5. The final assumption is homogeneity of variance. Homogeneous or equal, variance exists when the standard deviations of samples are approximately equal.

```
One sample t-test
t statistics: 1.2896282719661123 p value: 0.2017496571835306

Level of significance: 0.05

Our one-sample t-test p-value 0.2017496571835306

We have no evidence to reject the null hypothesis since p-value > level of significance
```

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

- Assumptions about the population distribution made are:

  - Need to check whether the 2 samples taken from a population follows Normal Distribution or not by using shapiro test.
  - Since the given sample size is greater than 30, as per the central limit theorem, it states that with larger sample size the distribution approaches Normal distribution.