Report by

Souravi Sinha

SMDM PROJECT REPORT

STATISTICAL ANALYSIS

Sunday, 12 September 2021

TABLE OF CONTENTS

INTRODUCTION

This report consist of three problems.

Wholesale Customers Analysis

Students Survey Analysis

AB shingles Moisture Analysis

PROBLEM - 1 SUMMARY

WHOLESALE CUSTOMERS ANALYSIS

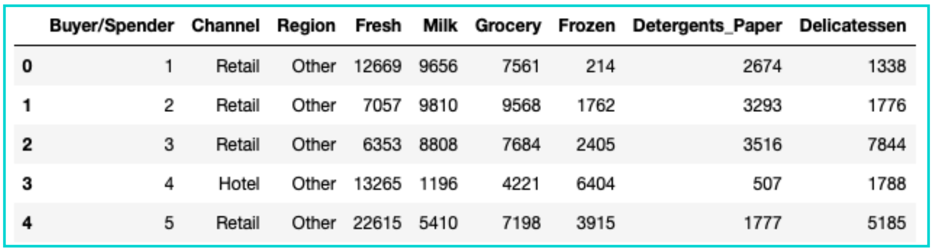
A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon,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

Exploratory Data Analysis

Data Description

1. Buyer/Spender : Serial numbers corresponding to either buyer or spender.
2. Channel : Type of sales channel it is i.e. Retail, Hotel.
3. Region : Retailer's region, It would be from Lisbon, Oporto or Other.
4. Fresh : Annual spending of fresh orders by the customer.
5. Milk : Annual spending of milk orders by the customer.
6. Grocery : Annual spending of Grocery ordered by the customer.
7. Frozen : Annual spending of Frozen products ordered by the customer.
8. Detergents_Paper : Annual spending of Detergents_Paper ordered by the customer.
9. Delicatessen : Annual spending of Delicatessen products ordered by the customer.

Sample of the dataset



	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

Figure 1. Dataset Sample

Data types of different variable

Column	Datatype
Buyer/Spender	int64
Channel	object
Region	object
Fresh	int64
Milk	int64
Grocery	int64
Frozen	int64
Detergents_Paper	int64
Delicatessen	int64

Table 1 : Data Information

Details

- Dataset has a total of 440 rows and 9 columns.
- Out of that 2 are objects and rest are int type.

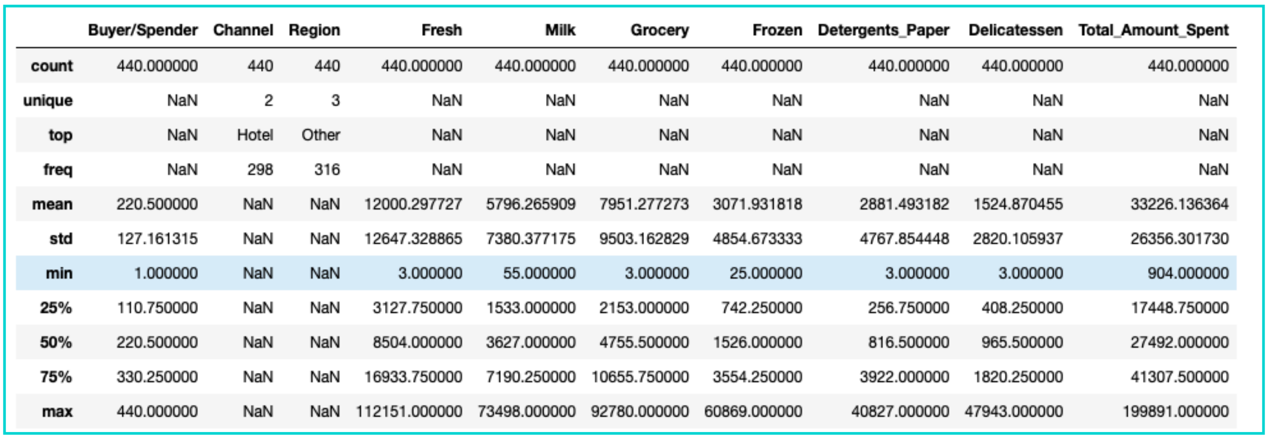
Missing data analysis for the variables

Column	Is Missing Data Present
Buyer/Spender	FALSE
Channel	FALSE
Region	FALSE
Fresh	FALSE
Milk	FALSE
Grocery	FALSE
Frozen	FALSE
Detergents_Paper	FALSE
Delicatessen	FALSE

Table 2 : Missing data analysis

PROBLEMS

I. Use methods of descriptive statistics to summarise data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?



	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total_Amount_Spent
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455	33226.136364
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000

Figure 2. Descriptive Statistics

Descriptive Statistics Inference

- Dataset has 3 different Regions and 2 different Channels.
- The average amount spent by the customers combining all the products is ***33226.13***
- The maximum total amount spent by the customer is ***199891.00***
- NaN shows that the values cannot be calculated for that particular variables.

Calculating the Total Amount Spent per Regions and Channels

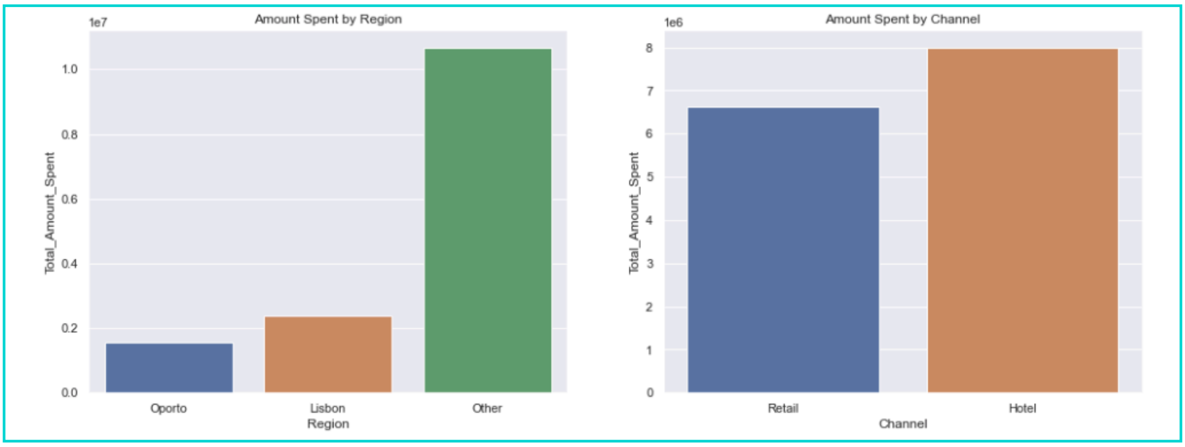


Figure 3. Total Amount Spent per Regions and Channels

Inference

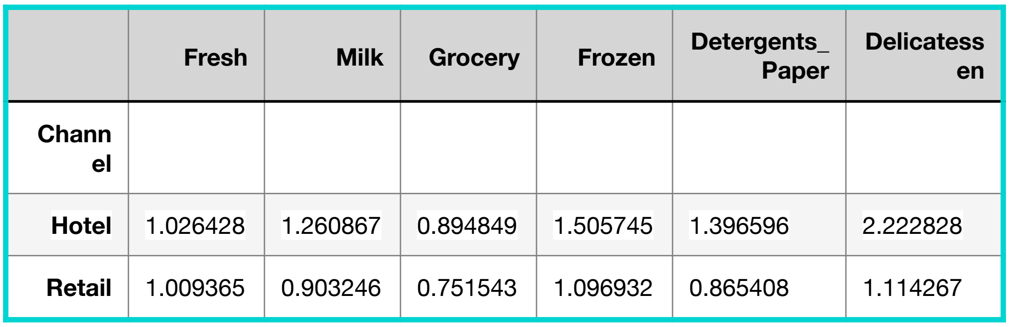
- After calculating the Total Amount Spent per Channels, we can conclude that ***least amount*** is spent by ***Retail*** and ***most*** by ***Hotel*** items.
- After calculating the ***Total Amount Spent per Regions***, we can conclude that ***least amount*** is spent on ***Oporto*** and ***max*** on ***Other*** items

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

Descriptive behaviour across Regions and Channels:

Channels:

Below is the coefficient of variation of the 6 varieties of the items across Channels.



	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicatess en
Chann el						
Hotel	1.026428	1.260867	0.894849	1.505745	1.396596	2.222828
Retail	1.009365	0.903246	0.751543	1.096932	0.865408	1.114267

Table 3 : CV across Channels

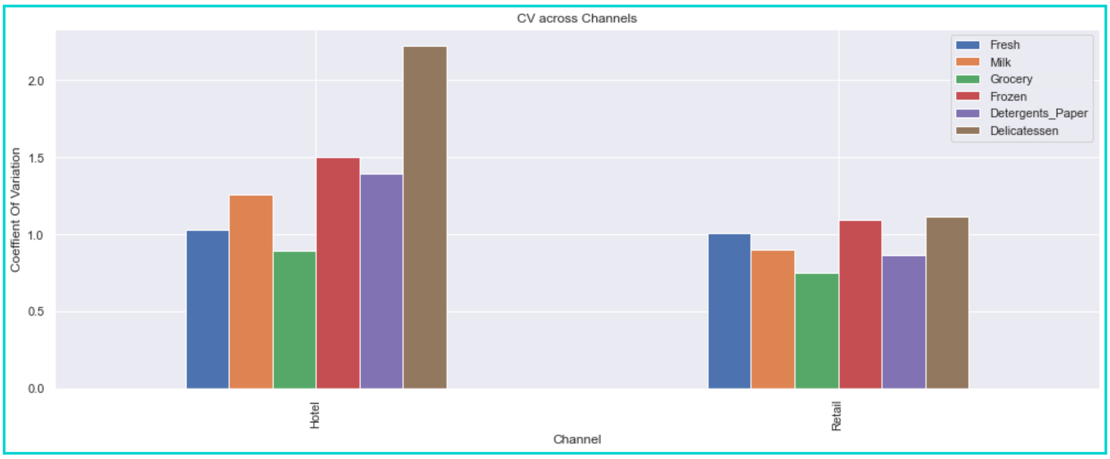


Figure 4: CV across all Channels

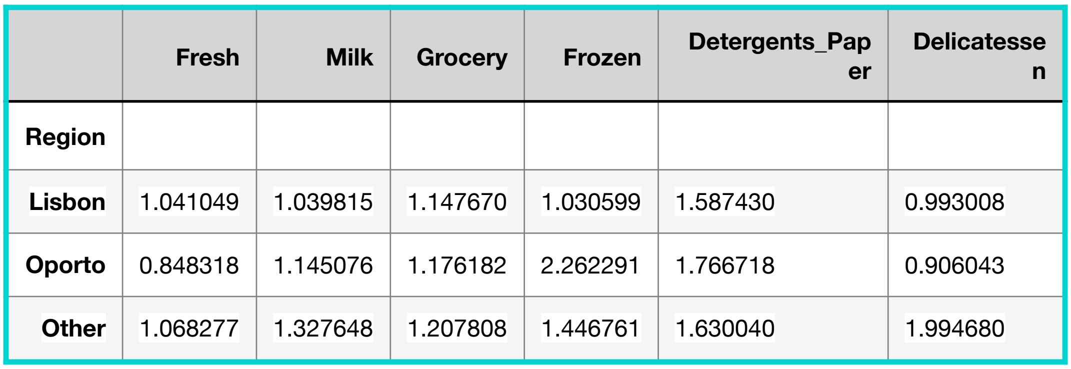
Inference

We can observe the following from the above analysis :

- ***Grocery (Retail Channel)*** has the lowest coefficient of variation and ***Delicatessen(Hotel Channel)*** has the highest cv.
- For Retail, ***Grocery*** has the lowest and ***Delicatessen*** has the highest cv.
- For Hotel, ***Grocery*** has the lowest and ***Delicatessen*** has the highest cv.

Regions:

Below is the coefficient of variation of the 6 varieties of the items across Regions.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Region						
Lisbon	1.041049	1.039815	1.147670	1.030599	1.587430	0.993008
Oporto	0.848318	1.145076	1.176182	2.262291	1.766718	0.906043
Other	1.068277	1.327648	1.207808	1.446761	1.630040	1.994680

Table 4 : CV across Regions

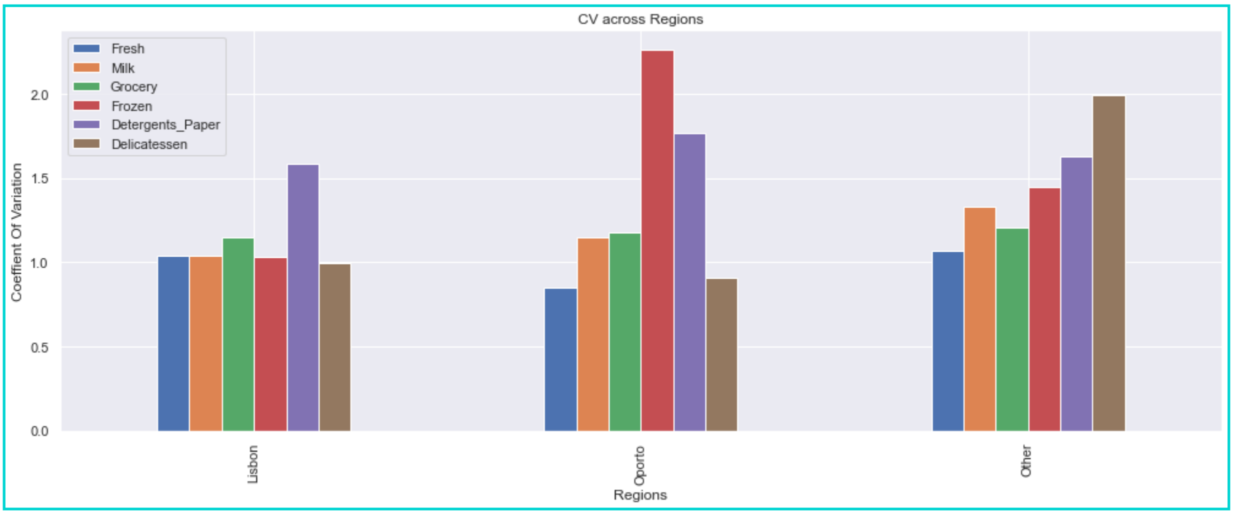


Figure 5: CV across all Regions

Inference

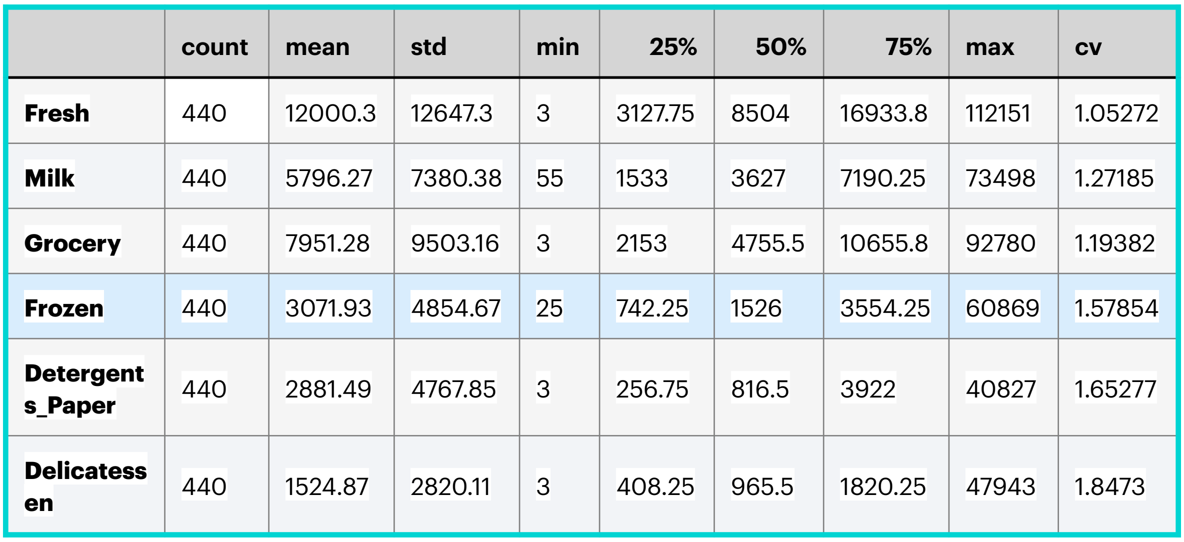
We can observe the following from the above analysis :

- ***Fresh (Oporto Region)*** has the lowest coefficient of variation and ***Frozen (Oporto Region)*** has the highest cv.
- For Lisbon Region, ***Delicatessen*** has the lowest and ***Detergents_Paper*** has the highest cv.
- For Oporto Region, ***Fresh*** has the lowest and ***Frozen*** has the highest cv.
- For Other Region, ***Fresh*** has the lowest and ***Delicatessen*** has the highest cv.

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

Descriptive behaviour across all the items:

Below is the combined description about the 6 varieties of the products overall. As we are comparing across various products , hence we will use coefficient of variation (cv).



	count	mean	std	min	25%	50%	75%	max	cv
Fresh	440	12000.3	12647.3	3	3127.75	8504	16933.8	112151	1.05272
Milk	440	5796.27	7380.38	55	1533	3627	7190.25	73498	1.27185
Grocery	440	7951.28	9503.16	3	2153	4755.5	10655.8	92780	1.19382
Frozen	440	3071.93	4854.67	25	742.25	1526	3554.25	60869	1.57854
Detergents_Paper	440	2881.49	4767.85	3	256.75	816.5	3922	40827	1.65277
Delicatessen	440	1524.87	2820.11	3	408.25	965.5	1820.25	47943	1.8473

Table 5 : Behavioural data analysis

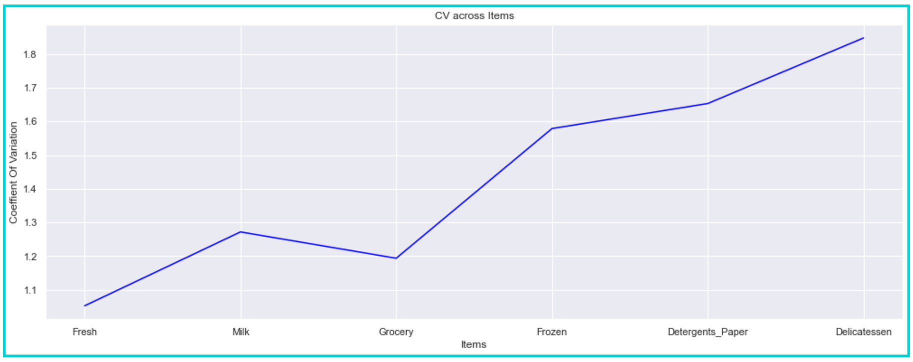


Figure 6. CV across all items

As per the above data, we can conclude the ***Fresh products have the least variation*** and ***Delicatessen products have the highest variation***.

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

Observations :

To find outliers, I have plotted Box plot. From the below plot, we can observe the following:

- All the items have outliers.
- Max outlier lies with ***Fresh*** items.
- Min outlier lies with ***Detergents_Paper***

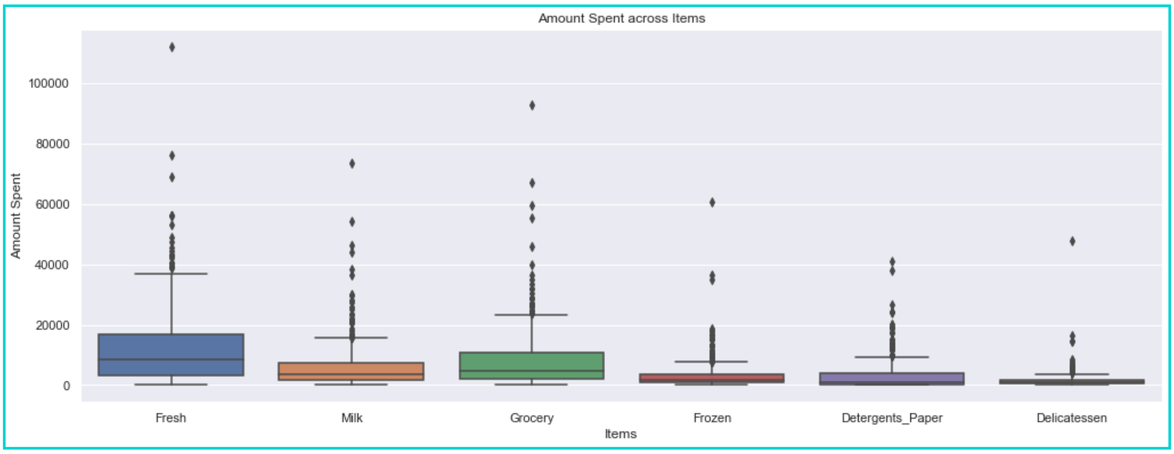


Figure 6. CV across all items

V. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

In order to suggest the measures to improve business or eradicate problems, we need to find the correlation among different variables.

Inference

Based on the overall statistical analysis as well as the correlation, following are the observations:

- Hotel buyers spend more amount than Retailers.
- Most of the amount is spent on Fresh and Groceries, across all the Regions.

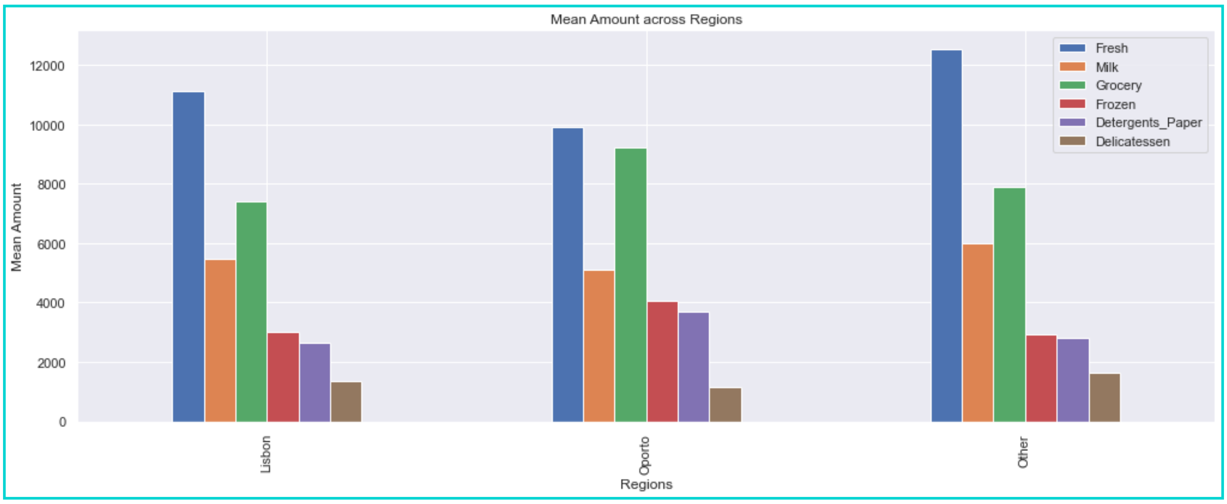


Figure 7. Mean Amount across all Regions

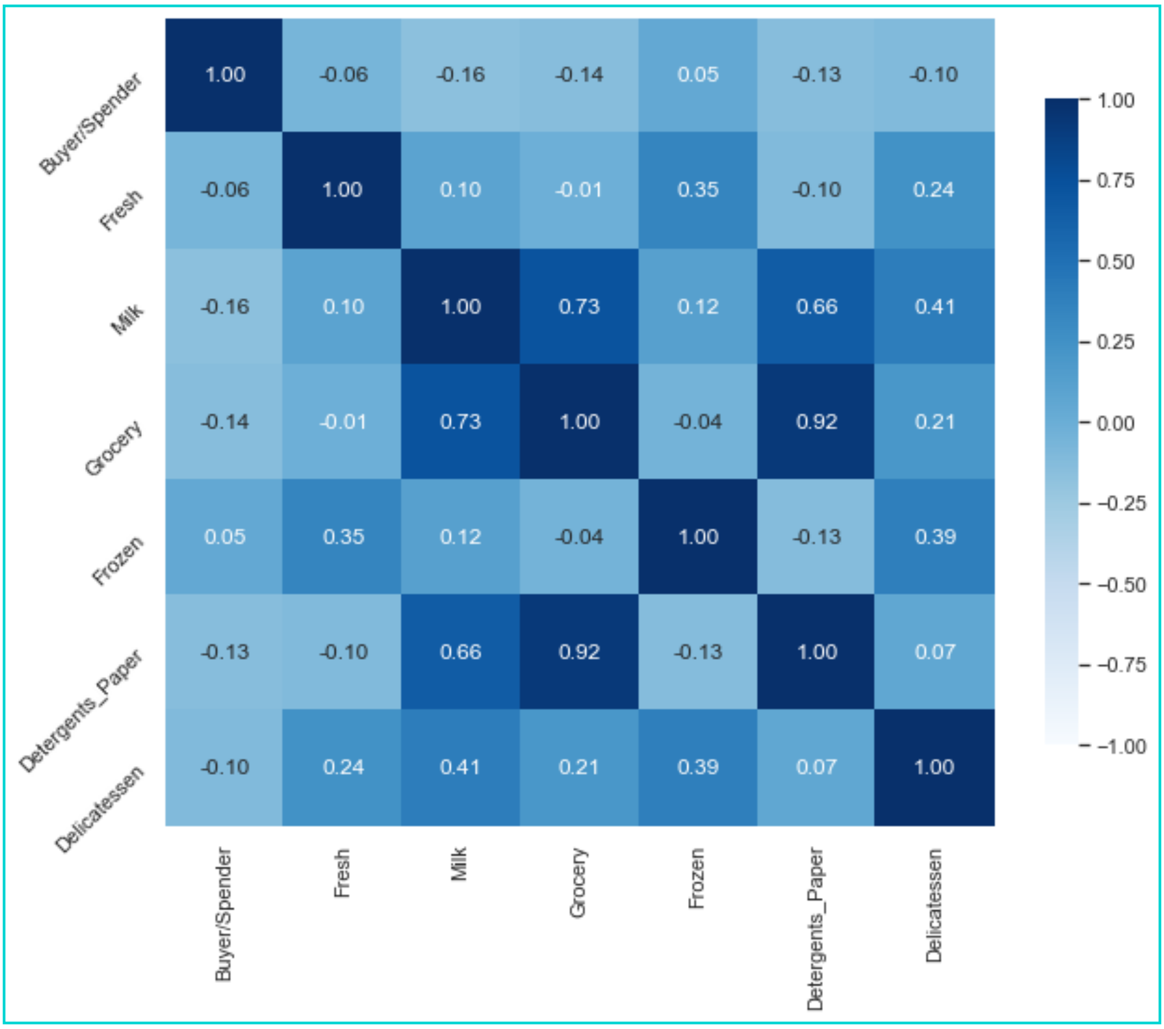


Figure 8. Correlation across all items

- Least of the amount is spent on Delicatessen, across all the Regions.
- Detergents_Paper and Grocery is highly correlated.
- Milk and Grocery is highly correlated.
- Detergents_Paper and Milk is highly correlated.
- Grocery has the lowest coefficient of variation and Delicatessen has the highest cv.
- All the items have outliers.
- Max outlier lies with Fresh items.
- Min outlier lies with Detergents_Paper

Recommendations:

- More stocks of Fresh and Grocery can be sold compared to other items.
- Hotel buyers can be concentrated on more.
- Less stock of Delicatessen can be maintained as it contributes to least amount.
- Detergents_Paper, Milk and Grocery are correlated the most, hence they can be sold together the most.
- Outliers from the data has to be removed for furthermore analysis.

PROBLEM - 2 SUMMARY

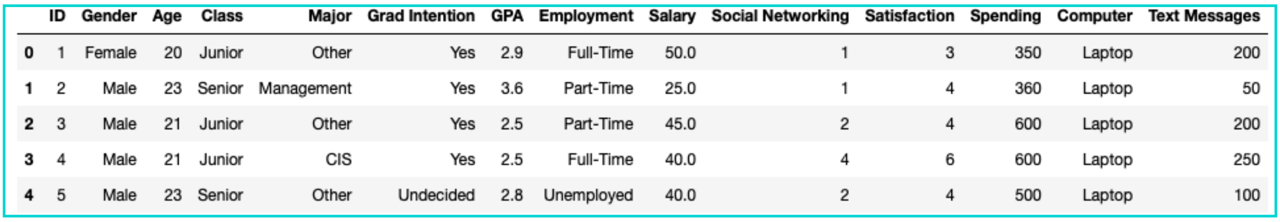
STUDENTS SURVEY ANALYSIS

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

Data Description

1. ID: Serial numbers corresponding to the students.
2. Gender: Gender of students.
3. Age: Student's age.
4. Class: Student's class
5. Major: Major subject of student.
6. Grad Intention: Graduate intention of the student.
7. GPA: Student's Grade Point Average
8. Employment: Type of employment e.g. - Part-Time, Full-Time and Unemployed
9. Salary: Salary of the student
10. Social Networking: Student is social networking sites.
11. Satisfaction:
12. Spending: Spending amount by the student.
13. Computer: Types of computer used e.g. Laptop, Desktop and Tablet.
14. Text Messages: Number of text messages per student.

Sample of the dataset

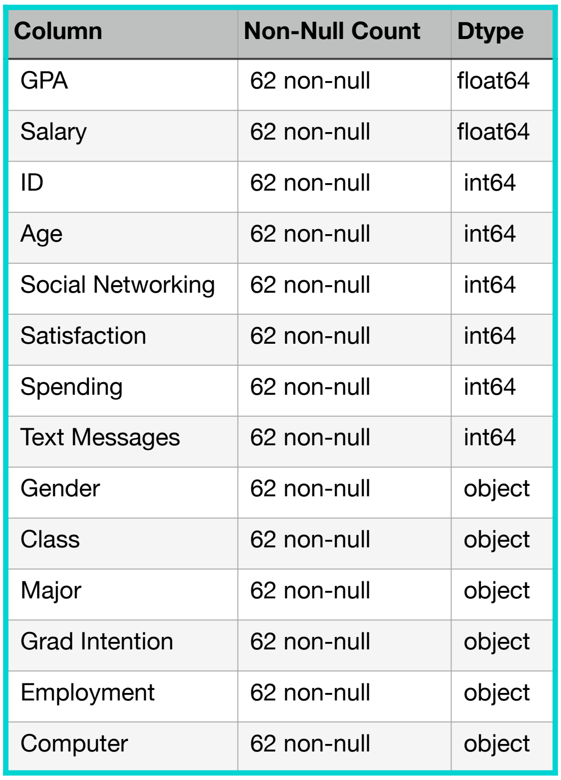


	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop	50
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	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

Figure 9. Dataset Sample 2

Exploratory Data Analysis

Data types of different variable



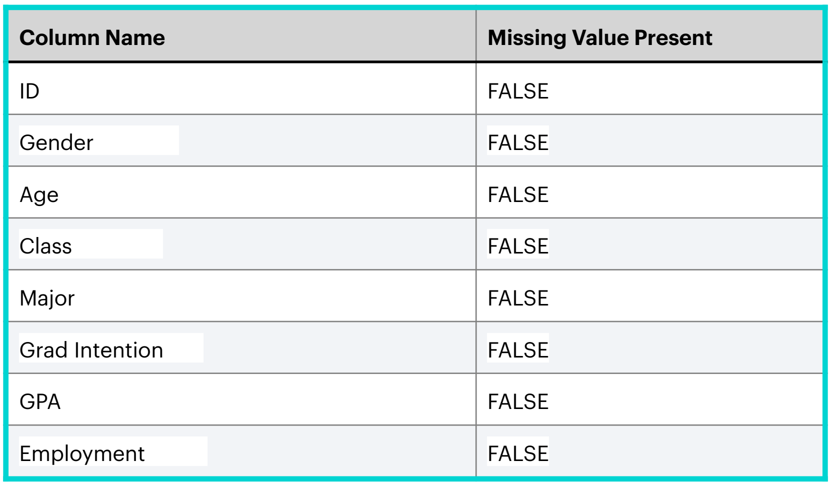
Column	Non-Null Count	Dtype
GPA	62 non-null	float64
Salary	62 non-null	float64
ID	62 non-null	int64
Age	62 non-null	int64
Social Networking	62 non-null	int64
Satisfaction	62 non-null	int64
Spending	62 non-null	int64
Text Messages	62 non-null	int64
Gender	62 non-null	object
Class	62 non-null	object
Major	62 non-null	object
Grad Intention	62 non-null	object
Employment	62 non-null	object
Computer	62 non-null	object

Table 6 : Problem 2 Datatypes

Details

- Dataset has a total of 62 rows and 14 columns.
- Out of that 2 are float types, 6 Int types and 6 are object type.

Missing data analysis for the variables



Column Name	Missing Value Present
ID	FALSE
Gender	FALSE
Age	FALSE
Class	FALSE
Major	FALSE
Grad Intention	FALSE
GPA	FALSE
Employment	FALSE

Table 7 : Missing data present

Salary	FALSE
Social Networking	FALSE
Satisfaction	FALSE
Spending	FALSE
Computer	FALSE
Text Messages	FALSE

Table 7 : Missing data present

This shows that there is no missing data.

PROBLEMS

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

1. Gender and Major

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

Table 8 : Gender and Major

2. Gender and Grad Intention

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

Table 9 : Gender and Grad Intention

3. Gender and Employment

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

Table 10: Gender and Employment

4. Gender and Computer

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

Table 11 : Gender and Computer

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

Male_Prob	Female_Prob
0.467742	0.532258

Table 11 : Probabilities of the Genders

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

The probability that a randomly selected CMSU student will be male is 0.46774193548387094

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

The probability that a randomly selected CMSU student will be female is 0.5322580645161

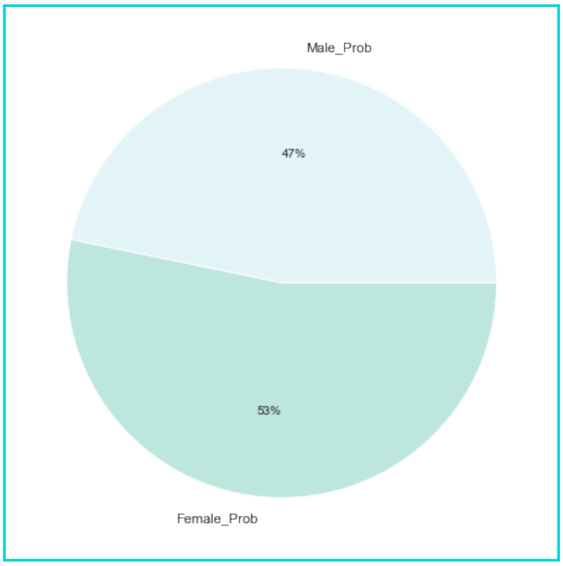


Figure 10. Male vs Female probability

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

1. Find the conditional probability of different majors among the male students in CMSU.

Gender	Major	Probability
Male	Management	0.206897
Male	Retailing/Marketing	0.172414
Male	Accounting	0.137931
Male	Economics/Finance	0.137931
Male	Other	0.137931
Male	Undecided	0.103448
Male	International Business	0.068966
Male	CIS	0.034483

Among MALE candidates:

- Probability of Management Major: 0.20689655172413793
- Probability of Retailing/Marketing Major: 0.1724137931034483
- Probability of Accounting Major: 0.13793103448275862
- Probability of Economics/Finance Major: 0.13793103448275862
- Probability of Other Major: 0.13793103448275862
- Probability of Undecided Major: 0.10344827586206896
- Probability of International Business Major: 0.06896551724137931
- Probability of CIS Major: 0.034482758620689655

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

Gender	Major	Probabilities
Female	Retailing/Marketing	0.272727
Female	Economics/Finance	0.212121
Female	International Business	0.121212
Female	Management	0.121212
Female	Accounting	0.090909
Female	CIS	0.090909
Female	Other	0.090909

Table 12 : Probabilities of the females in Majors

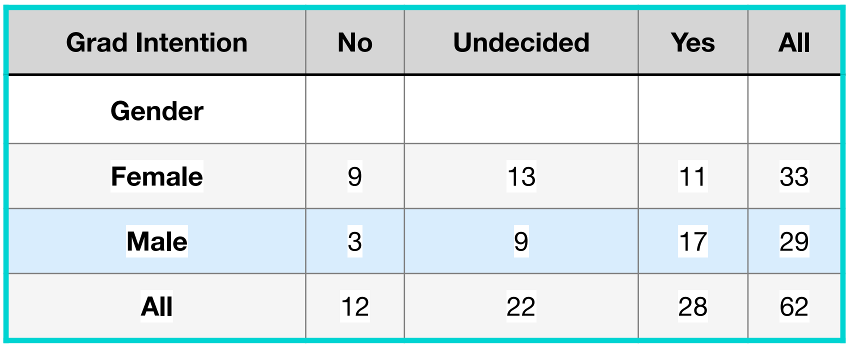
Among FEMALE candidates:

- Probability of Retailing/Marketing Major: 0.2727272727272727
- Probability of Economics/Finance Major: 0.21212121212121213
- Probability of International Business Major: 0.12121212121212122
- Probability of Management Major: 0.12121212121212122
- Probability of Accounting Major: 0.09090909090909091
- Probability of CIS Major: 0.09090909090909091
- Probability of Other Major: 0.09090909090909091
- Probability of Undecided Major: 0.0

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

1. Find the probability That a randomly chosen student is a male and intends to graduate.

Probability that a randomly chosen student is a male and intends to graduate is 0.21123829344432882

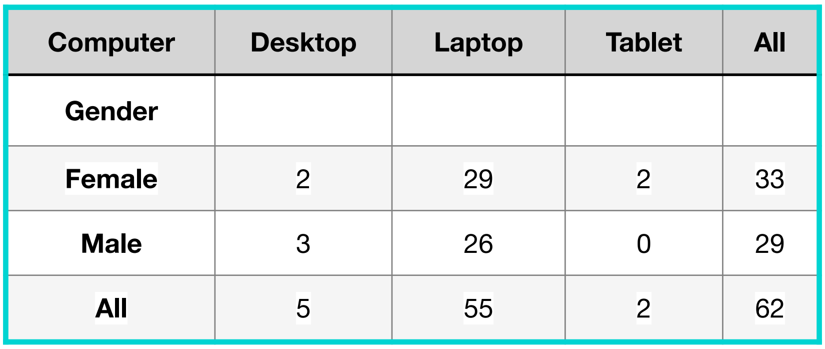


Grad Intention	No	Undecided	Yes	All
Gender				
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

Table 13 : Gender vs Grad Intention

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

Probability that a randomly selected student is a female and does NOT have a laptop is 0.060093652445369405



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

Table 14 : Gender and Computer

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

1. Find the probability that a randomly chosen student is a male or has full-time employment?

Probability that a randomly chosen student is a male or has full-time employment is 0.6290322580645161

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

Table 15 : Gender and Employment

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

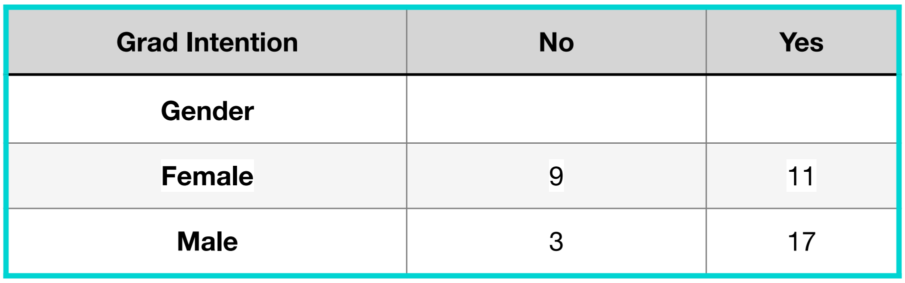
Conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 0.24242424242424243

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

Table 16 : Gender and Major

VI. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Events A and B are independent if the equation P(A∩B) = P(A) · P(B) holds true. In this case,



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table 17 : Gender and Grad Intention

$$P(FemaleAndGraduateIntention) <> P(Female) * P(GraduateIntention)$$

Hence, the graduate intention and being female are not independent events.

VII. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.Answer the following questions based on the data

1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

If a student is chosen randomly, the probability that his/her GPA is less than 3 is 0.27419354838709675

2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.

- Probability that a randomly selected male earns 50 or more is 0.4827586206896552
- Probability that a randomly selected female earns 50 or more is 0.5454545454545454

VIII. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarising your conclusions.

1. GPA

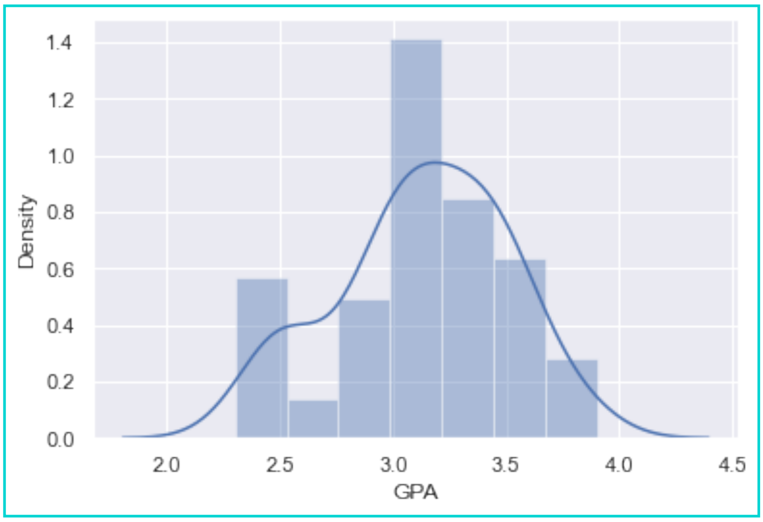


Figure 13: Students GPA as normal distribution
GPA Mean : 3.129032258064516
GPA Standard Deviation : 0.3773883926969118

2. Salary

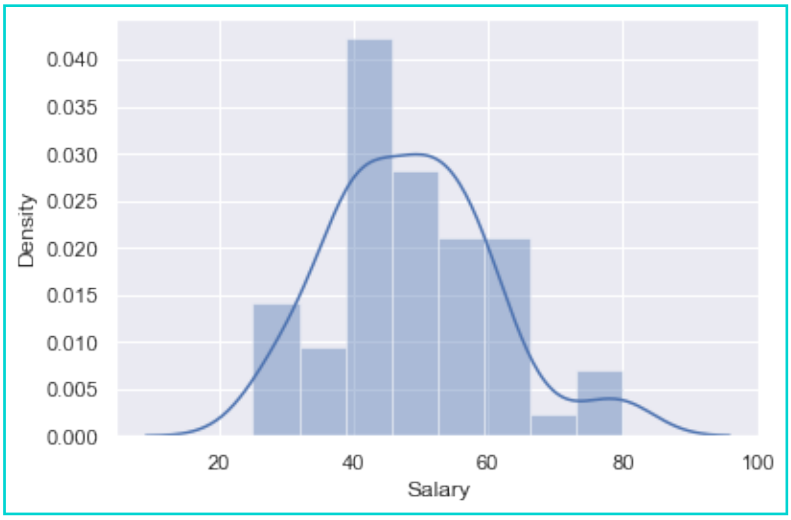


Figure 14: Students Salary as normal distribution
Salary Mean : 48.54838709677419
Salary Standard Deviation : 12.080912216337277

3. Spending

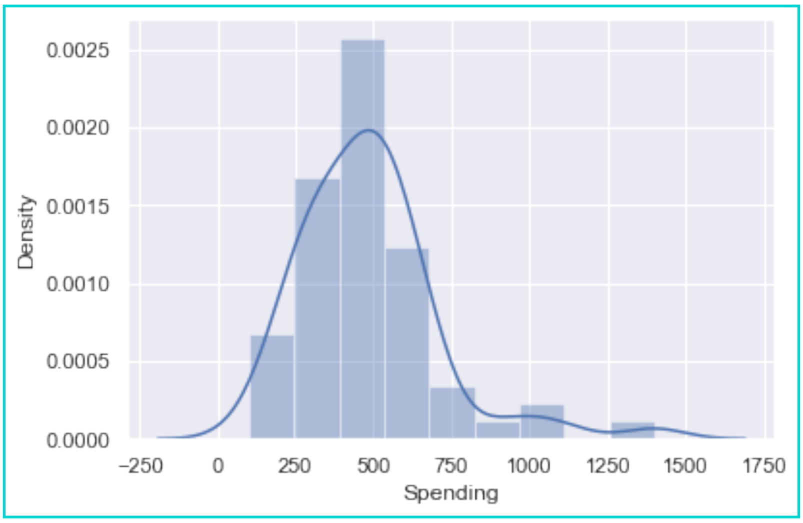


Figure 15: Students Spendings as normal distribution
Spending Mean : 482.01612903225805
Spending Standard Deviation : 221.95380496596204

4. Text Messages

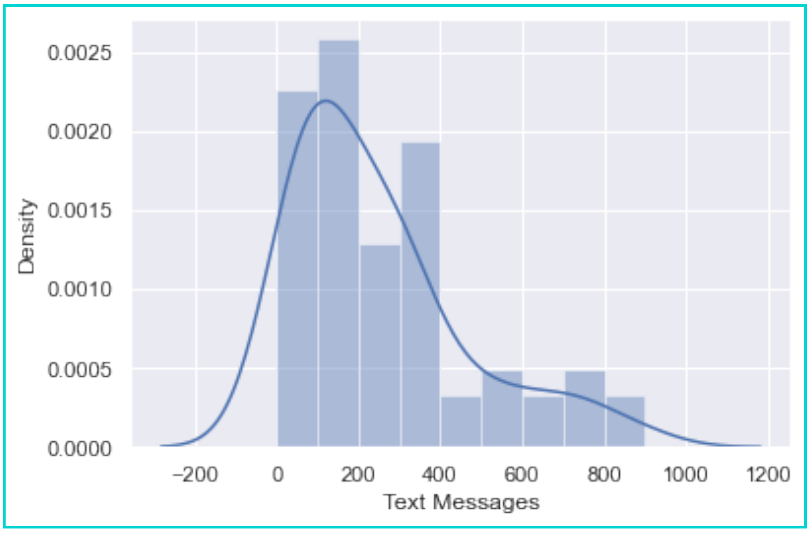


Figure 16: Students Text Messages as normal distribution
Text Messages Mean : 246.20967741935485
Text Messages Standard Deviation :
214.4659503026961

Mean and Standard Deviation per Column

Column Name	Mean	Std
GPA	3.129	0.377
Salary	48.548	12.080
Spending	482.016	221.953
Text Messages	246.209	214.465

Mean and Standard Deviation per Column

Empirical Rule Verification Per Column

Column Name	% within 1 SD	% within 2 SD	% within 3 SD	Empirical Rule Verified
GPA	1.535E-11	1.574E-08	6.073E-06	No
Salary	0.126	2.176	15.419	No
Spending	11.990	43.182	79.625	No
Text Messages	42.530	80.206	96.796	No

Empirical Rule Verification Per Column

PROBLEM - 3 SUMMARY

AB SHINGLES MOISTURE ANALYSIS

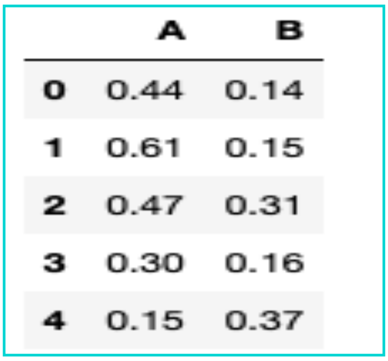
An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging.   In some 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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds 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.

Data Description

1. A: It Includes 36 measurements (in pounds per 100 square feet) for A shingles.
2. B: It Includes 31 measurements (in pounds per 100 square feet) for B shingles.

Sample of the dataset



	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

Figure 17: Dataset Sample

Exploratory Data Analysis

Data types of different variable

Column	Non-Null Count	Dtype
A	36 non-null	float64
B	31 non-null	float64

Table 17 : Problem 3 Datatypes

Details

- Dataset has a total of 36 rows and 2 columns.
- Both the columns are of float datatypes.

Missing data analysis for the variables

Column Name	Missing Value Present
A	FALSE
B	TRUE

Table 18 : Missing Data Present

This shows that there is missing data present in column B.

Problems

I. Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

In this question, the conclusion will be drawn with following Steps:

- **Step 1: State the Null and Alternative Hypothesis**
  - ➡ Alternative hypothesis (HA) : The mean moisture content > 0.35
  - ➡ Null hypothesis (H0) : The mean moisture content = 0.35

  To perform Hypothesis Testing, the following assumptions must hold,

  - ➡ The variables must follow continuous distribution
  - ➡ The sample must be randomly collected from the population
  - ➡ The underlying distribution must be normal. Alternatively, if the data is continuous, but may not be assumed to follow a normal distribution, a reasonably large sample size is required. CLT asserts that sample mean follows a normal distribution, even if the population distribution is not normal, when sample size is at-least 30.

- **Step 2: Decide the significance level**
  - ➡ Here we select Alpha = 0.05.
  - ➡ The sample size for A Shingles is 36
  - ➡ The sample size for B Shingles is 31

- **Step 3: Identify the test Statistic**
  - ➡ We are not aware of the Population standard deviation, hence we use ttest for 1 Sample here, for A, B separately

- **Step 4: Calculate the p_value for A Shingles**
  - ➡ Here we will use ttest_1samp to find the P-value. This function returns t statistics and 2-tailed P-value.
  - ➡ The result came as : 0.07477

- **Step 5: Calculate the P-value for B Shingles**
  - ➡ Here we will use ttest_1samp to find the P-value. This function returns t statistics and 2-tailed P-value.
  - ➡ The result came as : 0.00209

- **Step 6: Conclusion**
  - ➡ **A Shingles Conclusion :**

    Since P-value is 0.07477633144907513, which is more than the significance of a test(i.e. 0.05), Hence we do not have enough evidence to reject the Null Hypothesis that the mean moisture content of A Shingles is less than or equal to 0.35.
  - ➡ **B Shingles Conclusion :**

    Since P-value is 0.0020904774003191826, which is less then the significance of a test(i.e. 0.05), Hence we have enough evidence to reject the Null Hypothesis that mean moisture content of B Shingles less than or equal to 0.35.

II. Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

In this question, the conclusion will be drawn with following Steps:

- **Step 1: State the Null and Alternative Hypothesis**
  - ➡ Alternative hypothesis (HA) : The mean population of A & B are not equal.
  - ➡ Null hypothesis (H0) : The mean population of A & B are equal.

To perform Hypothesis Testing, the following assumptions must hold,

- ➡ The variables must follow continuous distribution
- ➡ The sample must be randomly collected from the population
- ➡ The underlying distribution must be normal. Alternatively, if the data is continuous, but may not be assumed to follow a normal distribution, a reasonably large sample size is required. CLT asserts that sample mean follows a normal distribution, even if the

population distribution is not normal, when sample size is at-least 30.

- **Step 2: Decide the significance level**
  - ➡ Here we select Alpha = 0.05.
  - ➡ The sample size for A Shingles is 36
  - ➡ The sample size for B Shingles is 31

- **Step 3: Identify the test Statistic**
  - ➡ We are dealing with 2 independent samples, hence we use ttest for 2 Sample here, for both A & B.

- **Step 4: Calculate the P-value for A and B Shingles**
  - ➡ Here we will use ttest_ind to find the P-value. This function returns t statistics and 2-tailed P-value.
  - ➡ The result came as : 0.201749

- **Step 5: Conclusion**
  - ➡ Since P-value is 0.2017496571835306, which is more then the significance of a test(i.e. 0.05), Hence we do not have enough evidence to reject the Null Hypothesis that mean population of A & B are equal.

.... *Thank you* ....