PGP - DSBA

STATISTICAL METHODS FOR DECISION MAKING

Project Report – July 2022

Shruti Jha
7-10-2022

Great Learning
POWER AHEAD

Contents

Problem 1 - Wholesale Customers Analysis 3

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least? 3

1.2 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. 5

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour? 7

1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 8

1.5 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. 9

Problem 2 – CMSU Student Analysis 10

2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 10

2.1.1. Gender and Major 10

2.1.2. Gender and Grad Intention 10

2.1.3. Gender and Employment 10

2.1.4. Gender and Computer 10

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 11

2.2.1. What is the probability that a randomly selected CMSU student will be male? 11

2.2.2. What is the probability that a randomly selected CMSU student will be female? 11

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 12

2.3.1. Find the conditional probability of different majors among the male students in CMSU. 12

2.3.2 Find the conditional probability of different majors among the female students of CMSU. 12

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 13

2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate. 13

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop. 13

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 14

2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment? 14

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 14

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events? 15

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. 16

2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 16

2.7.2. 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. 16

2.8. 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 summarizing your conclusions. 17

Problem 3 - ABC Asphalt Shingles Analysis 19

3.1 Do you think there is evidence that mean moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps. 19

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed? 21

Problem 1 - 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).

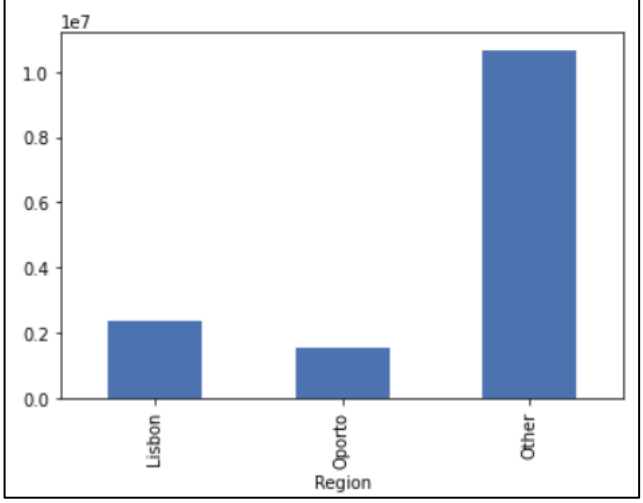
1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel spent the most? Which Region and which Channel spent the least?

Here is the summary of data, showing the total data count or rows, i.e. 440.

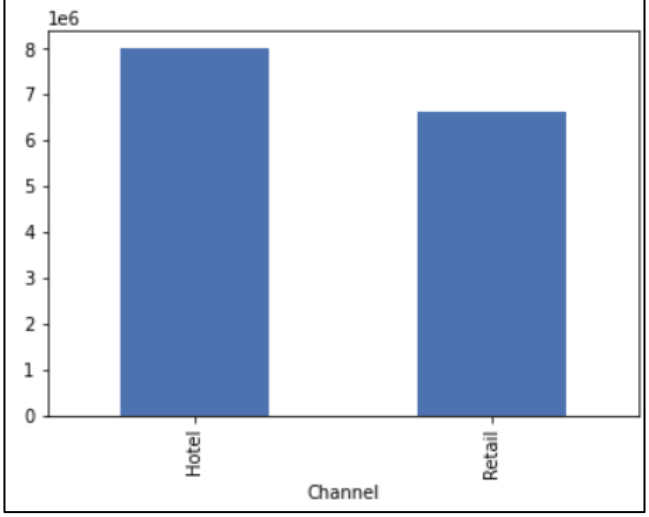
- Unique, top and most frequent data is presented for categorical datatype.
- It can also be inferred that there are 440, 2 and 3 unique customers, channels and regions, respectively.
- Highest number of purchases was done by **Hotel** channel and **Other** region by making payments **298** and **316** times, respectively.
- The highest amount of **$112,151** was spent on Fresh items.
- The average amount spent on Fresh item is highest among all items, this tells that this item is either expensive or consumed more among all the items.
- The amount spent for each item is highly skewed. Given that the items listed would be used on daily basis and based on their perishability they mostly get purchased in small lots, costing less amount per purchase.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	440.0	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	NaN	NaN	NaN	5796.265909	7380.377175	55.0	1533.0	3627.0	7190.25	73498.0
Grocery	440.0	NaN	NaN	NaN	7951.277273	9503.162829	3.0	2153.0	4755.5	10655.75	92780.0
Frozen	440.0	NaN	NaN	NaN	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	NaN	NaN	NaN	2881.493182	4767.854448	3.0	256.75	816.5	3922.0	40827.0
Delicatessen	440.0	NaN	NaN	NaN	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0
Total Spent	440.0	NaN	NaN	NaN	33226.136364	26356.30173	904.0	17448.75	27492.0	41307.5	199891.0

```
Region
Lisbon      2386813
Oporto      1555088
Other      10677599
Name: Total Spent, dtype: int64
```

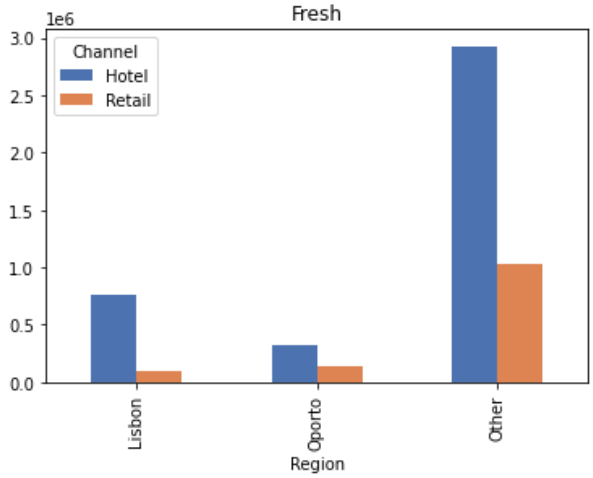


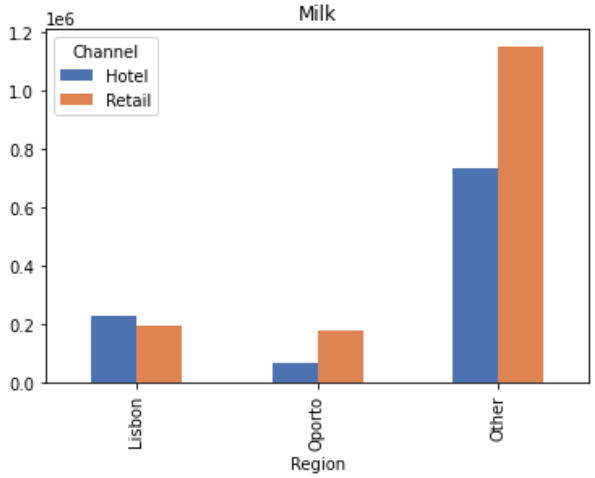

```
Channel
Hotel     7999569
Retail    6619931
Name: Total Spent, dtype: int64
```

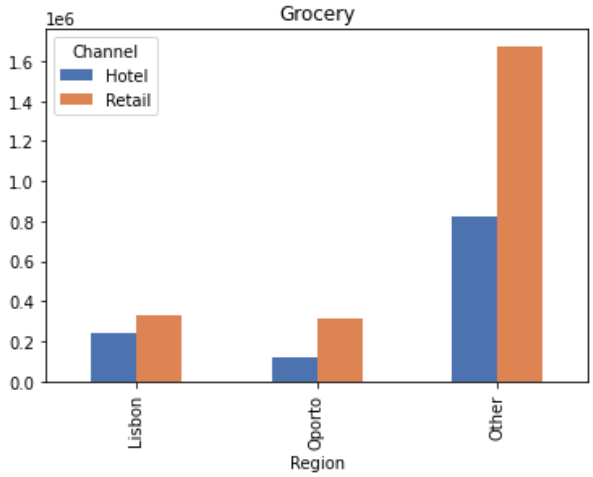


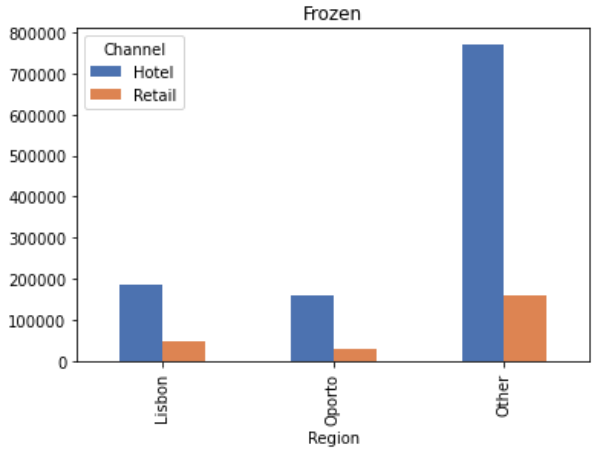
- As per the data calculated above, we can say that '**Other**' regions and '**Hotel'** channel have spent the most, and
- The '**Oporto'** region and '**Retail'** channel spent the least amount on the items.

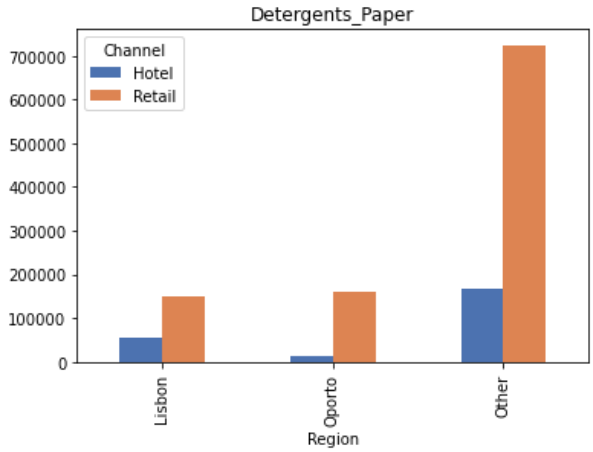
1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

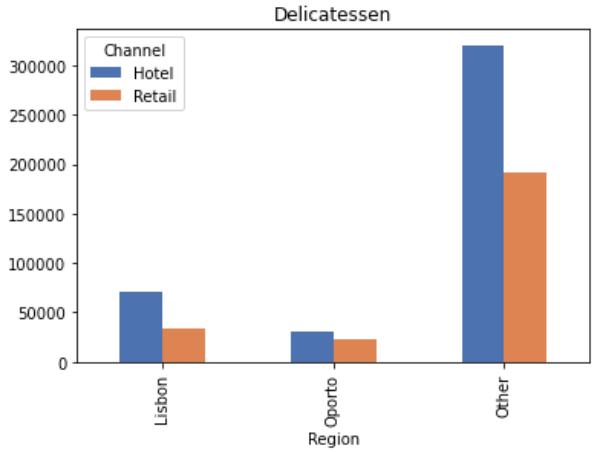




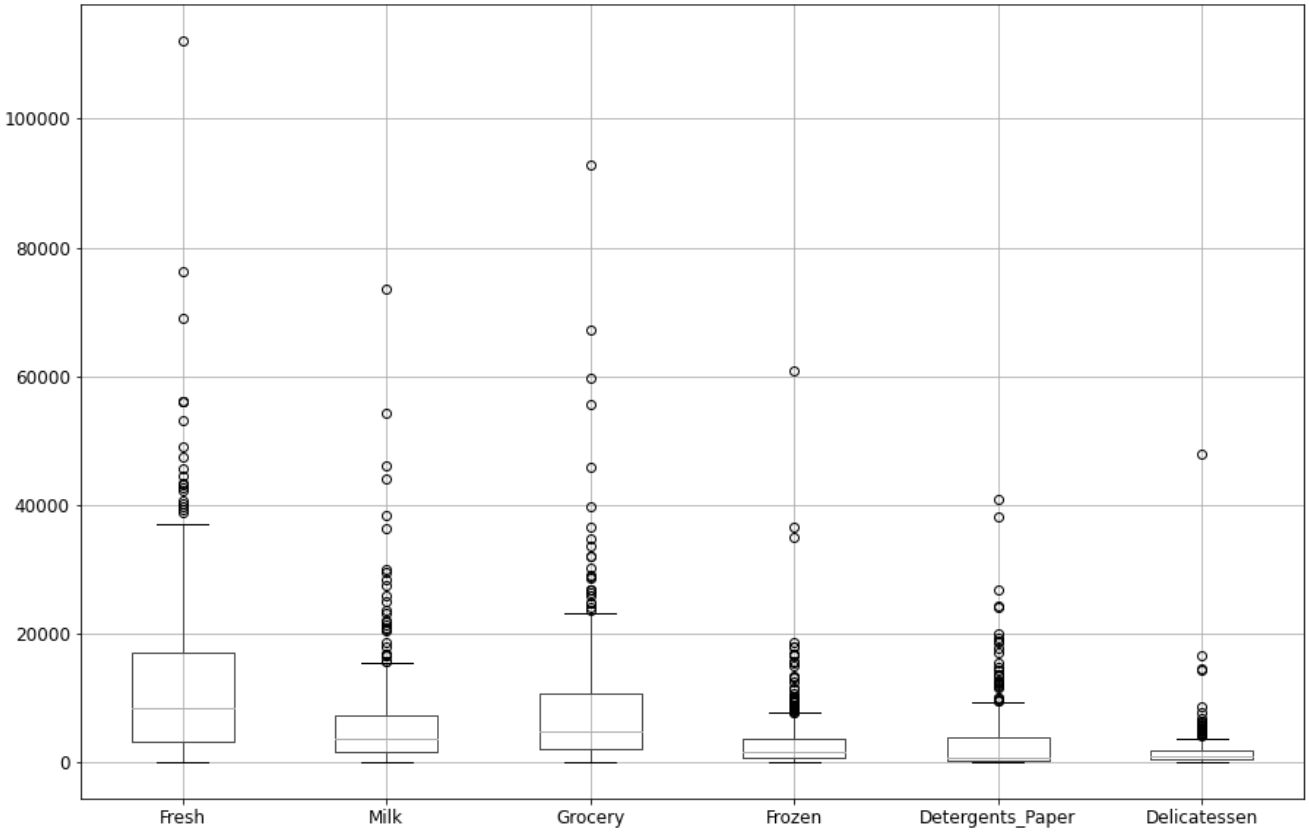








- Milk, grocery and detergent papers had been spent on more by the retail channel,
- And the Hotel channel had spent more on fresh, frozen and delicatessen items, in all the regions.



- Among all the items, highest amount has been spent on Fresh produce and Delicatessen item was spent on the least amount.

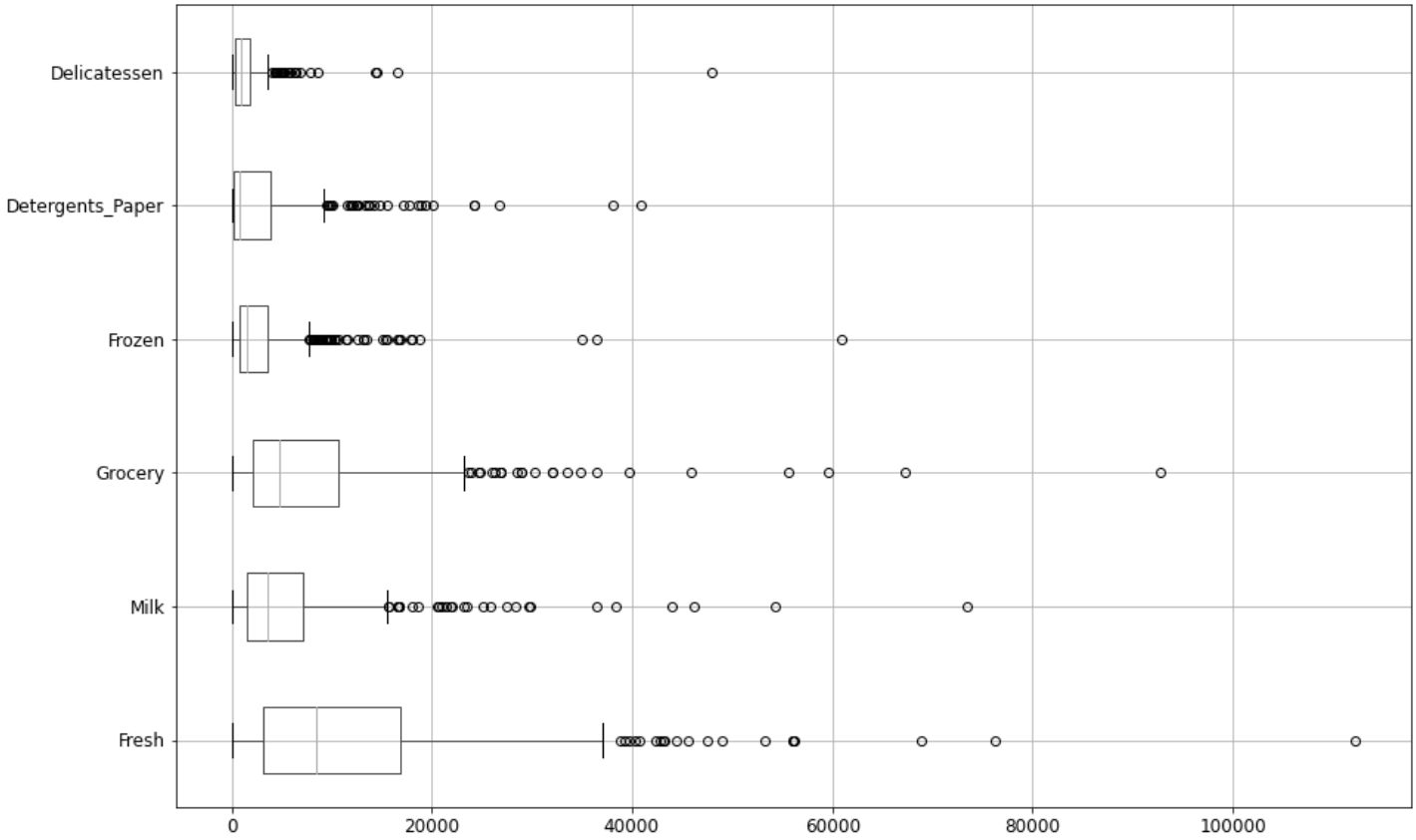
1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

	Skewness	Coefficient of Variation
Fresh	2.552583	1.052720
Milk	4.039922	1.271851
Grocery	3.575187	1.193815
Frozen	5.887826	1.578536
Detergents_Paper	3.619458	1.652766
Delicatessen	11.113534	1.847304

As per above calculations:

- Purchases of Delicatessen item shows the most inconsistent behaviour, with highest levels of skewness and coefficient of variation.
- Purchases of Fresh item shows the least inconsistent behaviour, with lowest levels of skewness and coefficient of variation.

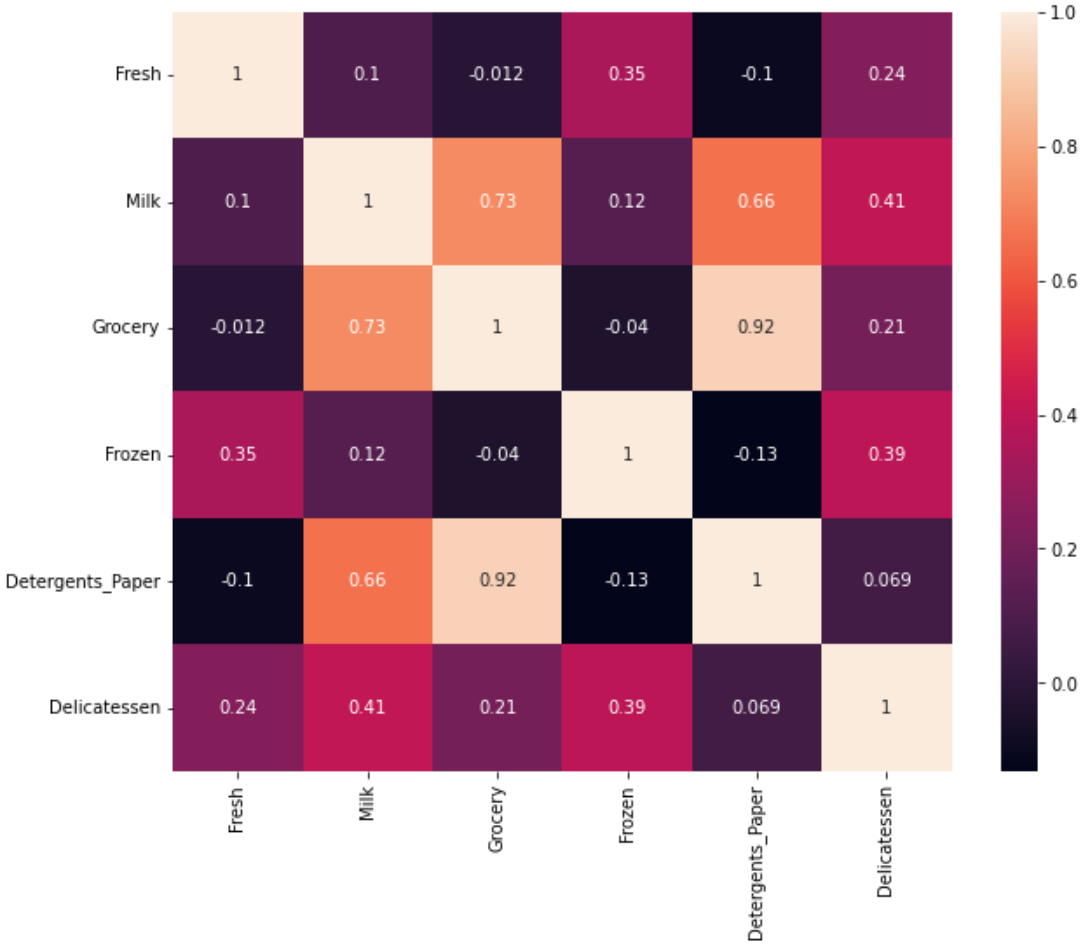
1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.



- There are quite a lot of extreme values present for all the items.
- The distribution of annual spendings on each item is extremely right skewed.
- Most of the annual spendings are on the lower side.
- Through this we can infer that, based on the perishability of all the items, customers would buy them in small lots, costing less amount per purchase.
- Hence, the outliers represent the less often time when these items were purchased in big lots of higher amounts.

1.5 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.

1. Based on the analysis, we can say that wholesaler can improve their business by:
    - Tackling the inconsistencies present in the purchase of each product. Mainly the delicatessen items, which showed highly inconsistent behaviour.



    - As per the above figure, delicatessen products show moderately high correlation with Milk. Hence, we can say that buyers who came to buy delicatessens also ended up buying Milk, most of the times.
    - Similarly, correlation between frozen and delicatessen items is also good enough to say that customers reached out to buy these products together more often.
    - Given the fact that frozen, delicatessen items and milk need proper refrigeration, all of these items can be placed together to increase the chances of customers to buy them together.

2. As we had already seen in section 1.2, that for fresh, frozen and delicatessen items hotels have made a significantly higher purchase than retail channel. Wholesaler can target retail channel to increase the sales of these products among them.

3. Similarly, Milk, grocery and detergent papers had been spent on more by the retail channel. From the above heatmap, these 3 products show the highest correlation. We can infer that these products are purchased together more often by retail customers. Wholesaler can target hotels to increase the sales of these products.

This analysis can help the business in improve and make their sales consistent for all the items across all the regions and channels.

Problem 2 – CMSU Student 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).

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

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 representative of the population of CMSU. Based on the data, answer the following question:

col_0	Count
Gender	
Female	33
Male	29
All	62

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

As per the above contingency table, we can say that the probability that a randomly selected CMSU student will be male is 0.467742 or 46.77%.

2.2.2. 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 0.532258 or 53.23%.

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

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

Using the normalized contingency table, we have calculated the conditional probability of different majors chosen by male students, as shown below.

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

- Through this we can infer that majority (20.69%) of male students opt for Management and the least (3.45%) number of male students prefer CIS as their major.
- Out of the total students, 10.34% are indecisive as to which major to choose and all of them are male.

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

Using the normalized contingency table, we have calculated the conditional probability of different majors chosen by female students, as shown below.

Major	
Accounting	0.090909
CIS	0.090909
Economics/Finance	0.212121
International Business	0.121212
Management	0.121212
Other	0.090909
Retailing/Marketing	0.272727
Undecided	0.000000

- Through this we can infer that majority (27.27%) of female students opt for Retailing/Marketing as their major.

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

2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate.

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

Using the probability multiplication rule, the probability that a randomly chosen student is a male and intends to graduate is 0.274194 or 27.42%.

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

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

The probability that a randomly selected student is a female and does NOT have a laptop is 0.064516 or 6.45%.

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

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

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

Probability that a randomly chosen student is a male or has full-time employment is 0.612903 or 61.29%.

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

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

Probability that given a female student is randomly chosen, she is majoring in international business or management is 0.242424 or 24.24 %.

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

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

For 2 events to be independent, this condition is to be satisfied:

$P(A \cap B) = P(A) * P(B)$

$P(\text{Grad Intention} \cap \text{Female}) = P(\text{Grad Intention}) * P(\text{Female})$

$P(\text{Grad Intention}) = 28 / 40 = 0.7$

$P(\text{Female}) = 20 / 40 = 0.5$

$P(\text{Grad Intention}) * P(\text{Female}) = 0.7 * 0.5 = 0.35$

$P(\text{Grad Intention} \cap \text{Female}) = 11 / 40 = 0.275$

Since the probability multiplication of both events is not equal to combined event, so the graduate intention and being female are not independent events.

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data

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

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	All
Gender																	
Female	0.016129	0.016129	0.032258	0.000000	0.016129	0.048387	0.080645	0.032258	0.064516	0.048387	0.032258	0.064516	0.016129	0.032258	0.016129	0.016129	0.532258
Male	0.000000	0.000000	0.064516	0.032258	0.032258	0.016129	0.032258	0.080645	0.032258	0.032258	0.080645	0.032258	0.032258	0.000000	0.000000	0.000000	0.467742
All	0.016129	0.016129	0.096774	0.032258	0.048387	0.064516	0.112903	0.112903	0.096774	0.080645	0.112903	0.096774	0.048387	0.032258	0.016129	0.016129	1.000000

If a student is chosen randomly, the probability that his/her GPA is less than 3 is 0.274194 or 27.42%.

2.7.2. 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.

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0
Gender																			
Female	0.000000	0.151515	0.030303	0.000000	0.030303	0.151515	0.030303	0.030303	0.000000	0.030303	0.151515	0.000000	0.000000	0.151515	0.151515	0.000000	0.030303	0.030303	0.030303
Male	0.034483	0.000000	0.034483	0.034483	0.000000	0.241379	0.000000	0.137931	0.034483	0.000000	0.137931	0.034483	0.034483	0.103448	0.103448	0.034483	0.000000	0.000000	0.034483
All	0.016129	0.080645	0.032258	0.016129	0.016129	0.193548	0.016129	0.080645	0.016129	0.016129	0.145161	0.016129	0.016129	0.129032	0.129032	0.016129	0.016129	0.016129	0.032258

The conditional probability that a randomly selected male earns 50 or more is 0.482759 or 48.28%.

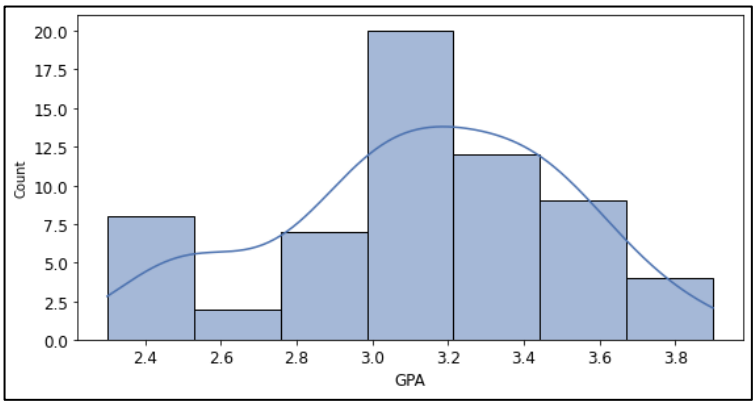
The conditional probability that a randomly selected female earns 50 or more is 0.545455 or 54.55%.

2.8. 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 summarizing your conclusions.

To figure out whether or not each variable is normally distributed, employed 3 methods mentioned below:

1. **Histograms**
2. **Mean and median** (normally distributed data has mean ≃ median)
3. **Mean and standard deviation** (~95% values fall under ±2σ from the mean for normally distributed data, as per the industry standard)

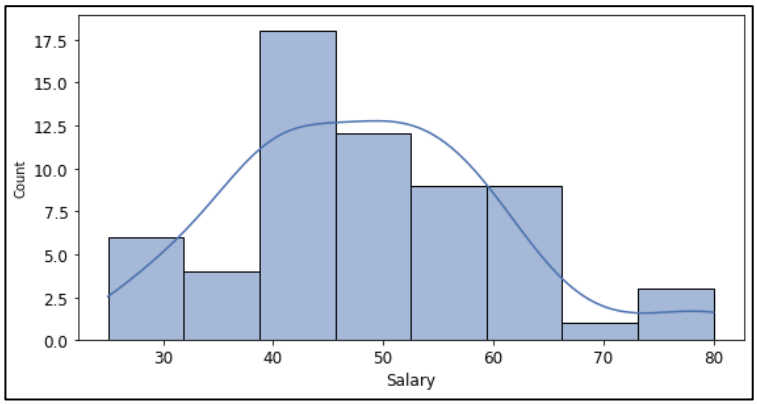
- **GPA**:



Mean = 3.129032
Median = 3.15
Standard deviation = 0.377388

- As per the histogram, the data seems to be normally distributed.
- The difference between mean and median (0.02) is nominal.
- 3.129032 ± (2 * 0.377388) = 96% of the values fall under 2σ, which is ~95%.
- Hence, we can conclude that GPA follows a normal distribution.

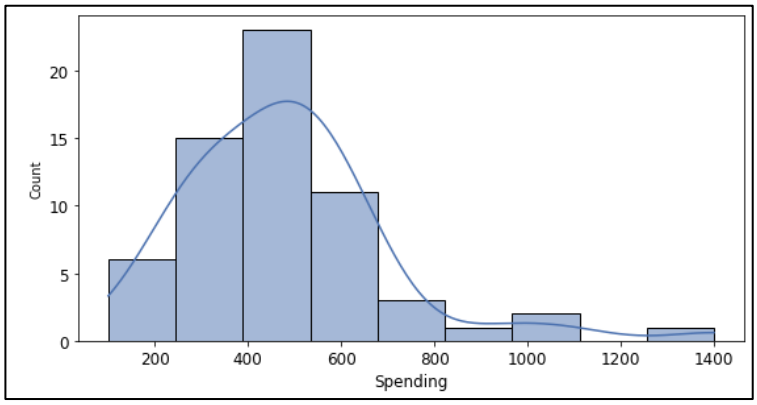
- **Salary**:



Mean = 48.548387
Median = 50
Standard deviation = 12.080912

- As per the histogram, the data seems to be normally distributed.
- The difference between mean and median (1.45) is very less.
- 48.548387 ±(2 * 12.080912) = 95.16% of the values fall under 2σ, which is ~95%.
- Hence, we can conclude that Salary also follows a normal distribution.

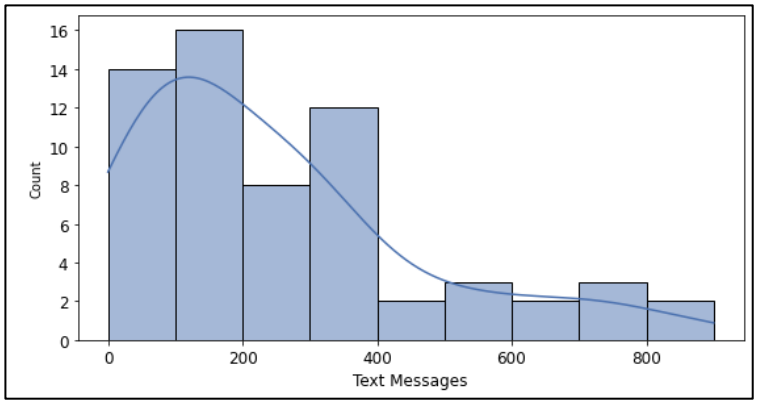
- **Spending**:



Mean = 482.016129
Median = 500
Standard deviation = 221.953805

  - As per the histogram, the data seems to be right skewed.
  - The difference between mean and median (17.98) is high.
  - 482.016129 ±(2 * 221.953805), i.e. only 90.32% of the values fall under 2σ.
  - Hence, we can conclude that Spending doesn't follow a normal distribution.

- **Text Messages**:



Mean = 246.209677
Median = 200
Standard deviation = 214.465950

  - As per the histogram, the data seems to be right skewed.
  - The difference between mean and median (46.21) is high.
  - 246.209677 ±(2 * 214.465950), i.e. only 43.54% of the values fall under 2σ.
  - Hence, we can conclude that Text Messages doesn't follow a normal distribution.

Problem 3 - ABC Asphalt Shingles 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 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 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.

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

- Alpha = 0.05
- Left tailed test
- 1 sample T-test

- 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:

  H1 = $\mu_{\text{moisture content}}$ > 0.35 pounds
  H0 = $\mu_{\text{moisture content}}$ <= 0.35 pounds

P-value = 0.07477633144907513

As we can see, the pvalue is > 0.05. So, there are insufficient evidence that mean moisture content per 100 square feet in A shingles is more than permissible limits (0.35 pounds).

So, we failed to reject the null hypothesis. Hence, we can conclude that at 95% confidence level mean moisture content in A shingles is within the permissible limits.

- 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:

  H1 = $\mu_{\text{moisture content}}$ > 0.35 pounds
  H0 = $\mu_{\text{moisture content}}$ <= 0.35 pounds

P-value = 0.0020904774003191826

As we can see for B shingles, the pvalue is < 0.05. We can say that at 95% confidence level there is sufficient evidence that mean moisture content in B shingles is more than the permissible limits.

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

H1 = μ(A) = μ(B)
H0 = μ(A) != μ(B)
Alpha = 0.05
2-tailed test
Independent T-test

Output:
P-value = 0.2017496571835306

As we can see that the pvalue is > 0.05. So, at 95% confidence level, we failed to reject the null hypothesis. Hence, the population mean for shingles A and B are not 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. Central Limit Theorem asserts that sample mean follows a normal distribution, even if the population distribution is not normal, when sample size is at least 30.
- For 2 sample t-test, the population variances of 2 distributions must be equal.