SMDM Project

Nandha Keshore Utti

PG-DSBA Online

March’ 22

Date: 21/05/2022

Contents

List of figures

List of tables

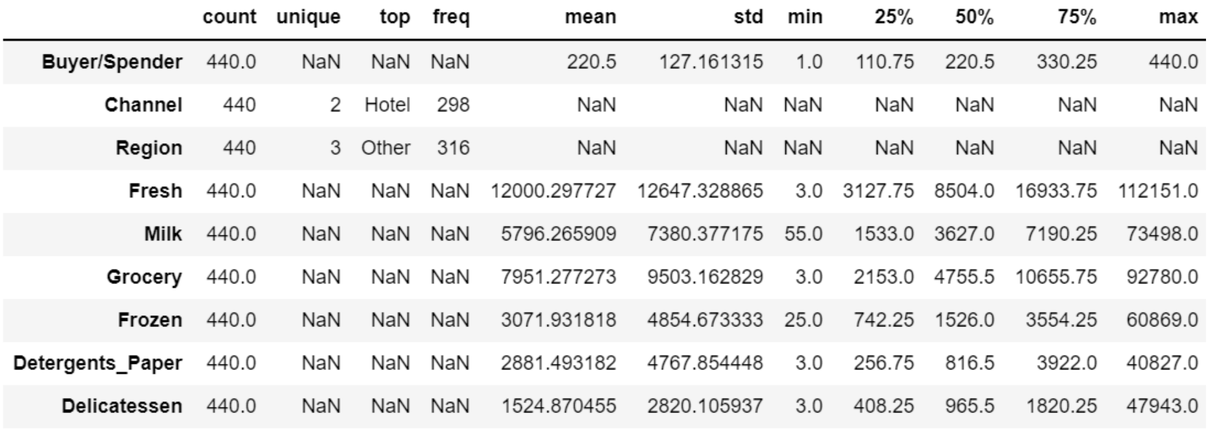
Problem 1 (Wholesale customers analysis)

Problem Statement:

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?

Summarization of the data:



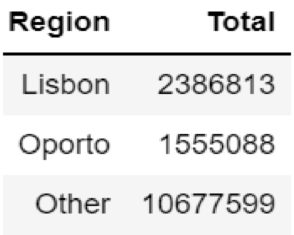
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	NaN	NaN	NaN	220.5	127.161315	1.0	110.75	220.5	330.25	440.0
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440.0	NaN	NaN	NaN	12000.297727	12647.328865	3.0	3127.75	8504.0	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

Table.1

Interpretation:

- There are 3 different regions and 2 different channels with total of 440 no. of buyer/spender.
- Distributor is spending maximum on 'Fresh' item and minimum on 'Delicatessen' item on for 440 buyer/spenders.

Description across given 3 different regions:



Region	Total
Lisbon	2386813
Oporto	1555088
Other	10677599

Table.2

- Distributor has spent the most on 'Other' region and least on 'Oporto' region
- Graphical representation is as shown below:

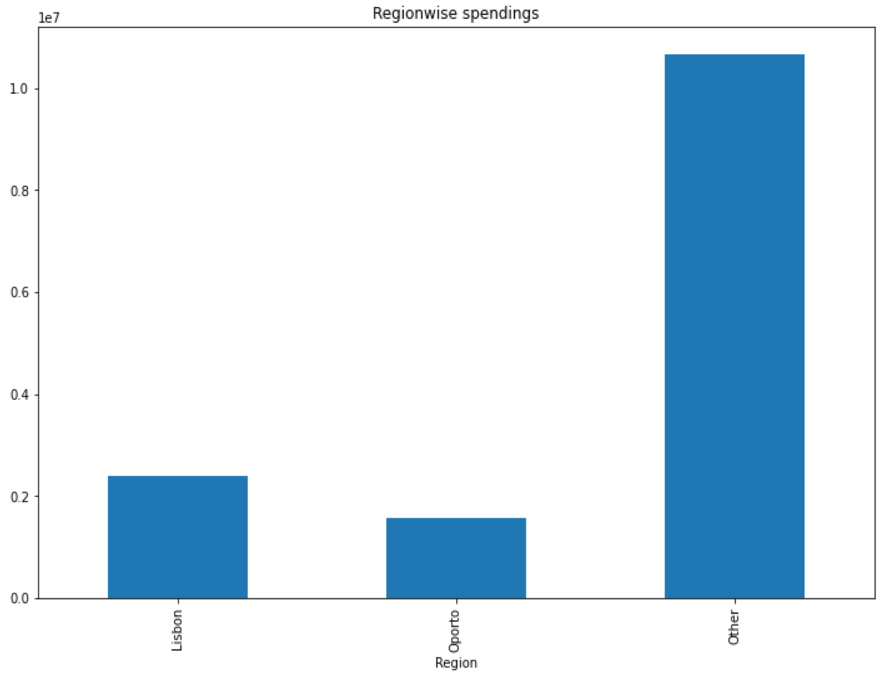
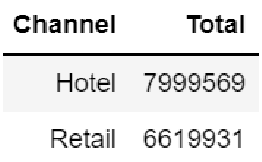


Fig.1

Description across given 2 different channels:



Channel	Total
Hotel	7999569
Retail	6619931

Table.3

- Distributor has spent the most on ‘Hotel’ channel and least on ‘Retail’ channel
- Graphical representation is as shown below:

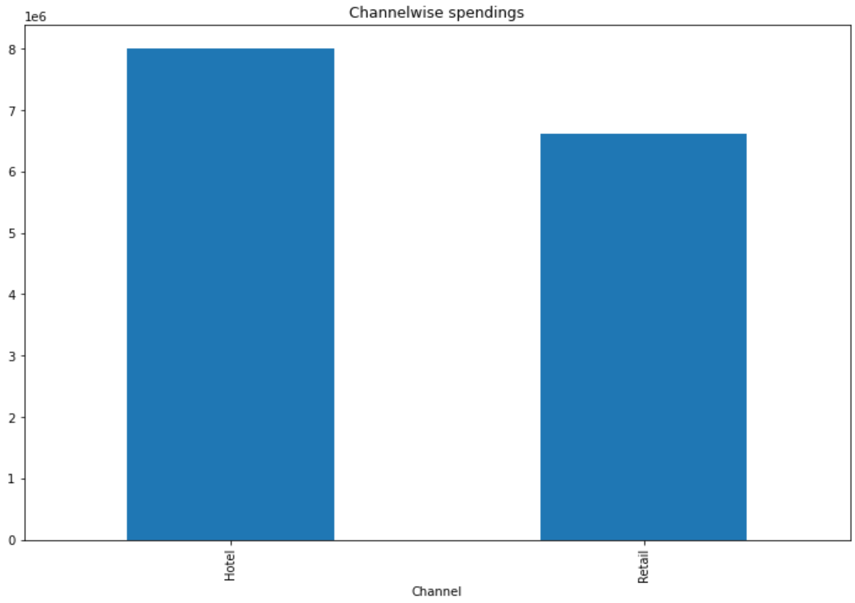
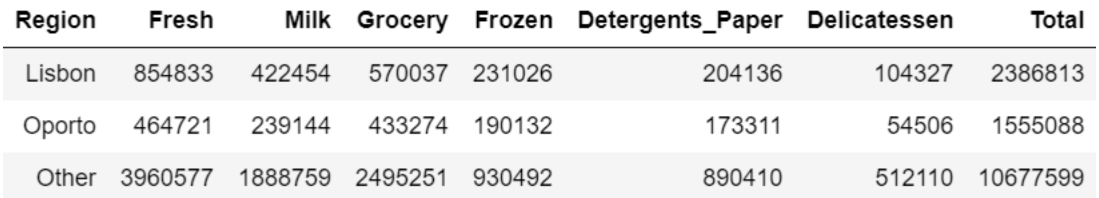


Fig.2

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.

Region wise spending on 6 varieties of items:

- Data representation:



Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	854833	422454	570037	231026	204136	104327	2386813
Oporto	464721	239144	433274	190132	173311	54506	1555088
Other	3960577	1888759	2495251	930492	890410	512110	10677599

Table.4

- Graphical representation:

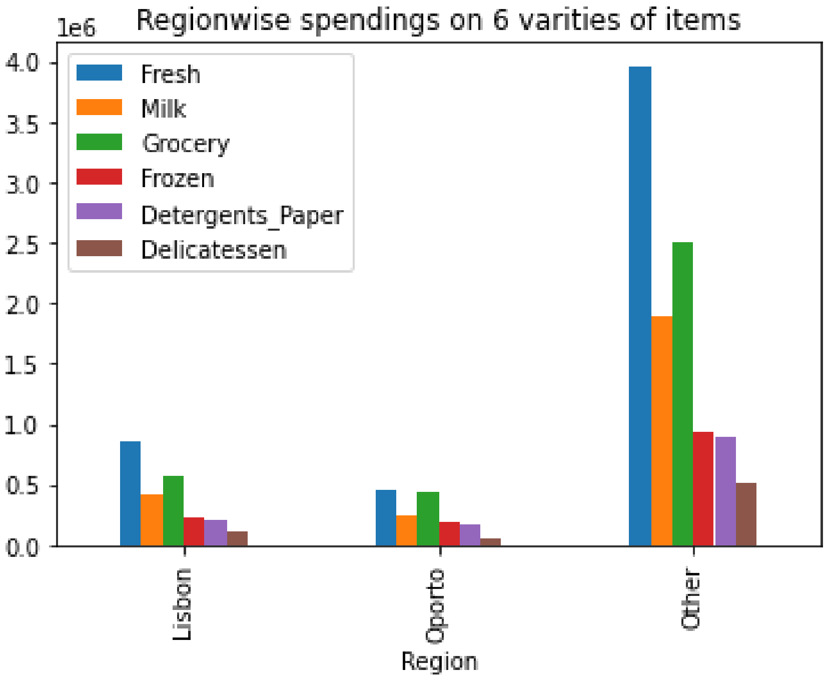
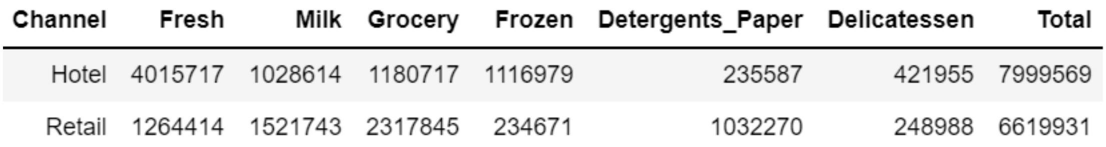


Fig.3

- Other regions have spent the most on every variety of item.
- Oporto region has spent the least compared to other two regions.
- But among all varieties of items, Fresh, Milk, Grocery are the items on which all three different regions have spent the most and they have spent the least on Frozen, Detergents paper, Delicatessen.

Channel wise spending on 6 varieties of items:

- Data representation:



Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	1264414	1521743	2317845	234671	1032270	248988	6619931

Table.5

- Graphical representation:

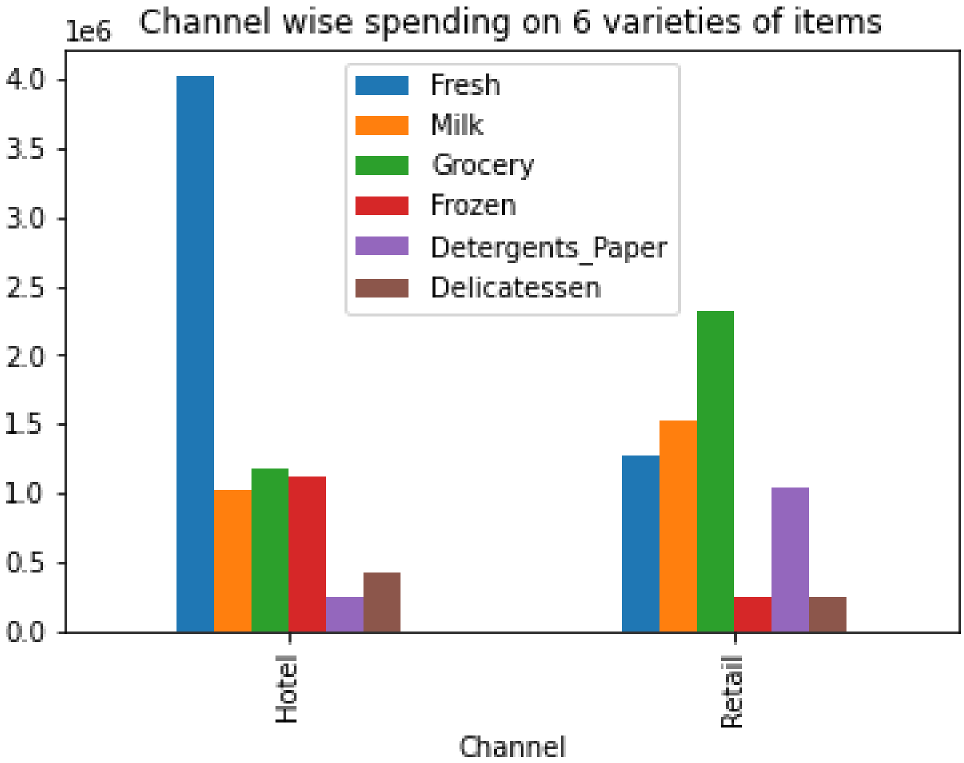


Fig.4

- Spendings on 6 items behaving differently in two channels
- Hotel channels have spent the maximum amount on Fresh item compared to all other items nearly by more than double amount. Nearly same amount spent on items such as milk, grocery and frozen. And least spent on detergents paper and delicatessen
- Retail channel has spent its most amount on Grocery. Fresh, milk, detergents paper is the spent by them in medium range. And least spent on frozen, delicatessen
- There is inconsistency in spendings on Fresh, grocery, frozen and detergent paper across both the channels

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?

- ‘Delicatessen’ has shown most inconsistent behaviour and ‘Fresh’ item has shown the least inconsistent behaviour

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

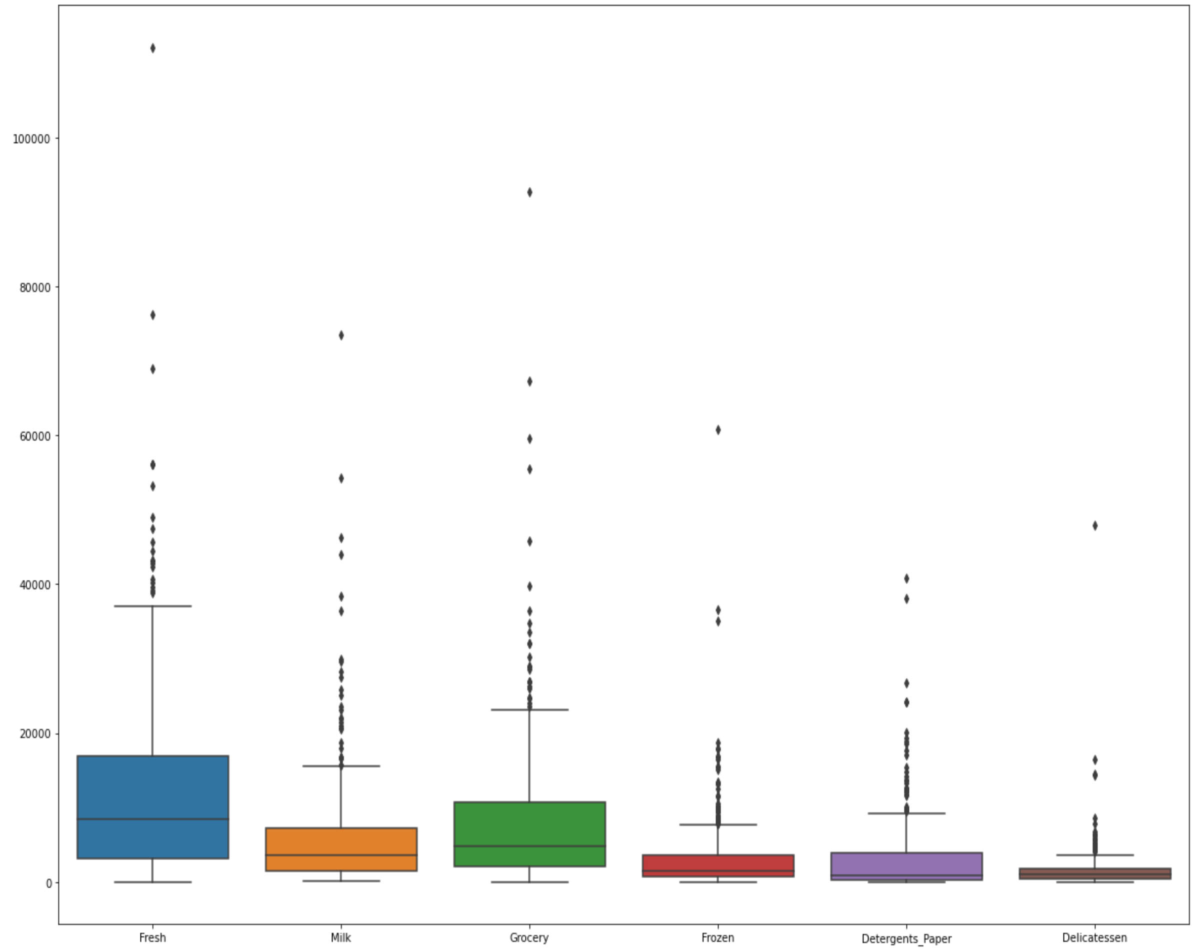


Fig.5

- Boxplot of all items show that every item has outliers in their spendings

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.

- Distributor should highly concentrate on ‘Oporto’ region on every item as there is huge difference in spendings compared to ‘Other’ region. Same for ‘Lisbon’ also
- Even though ‘Fresh’ item is highly spent overall, but inconsistency among two channels, distributor need to focus in this area and rectify it, same kind of inconsistency for ‘Frozen’ and ‘Detergents paper’ also
- ‘Delicatessen’ is having most inconsistency overall, need to focus more on this item compared other 5 items

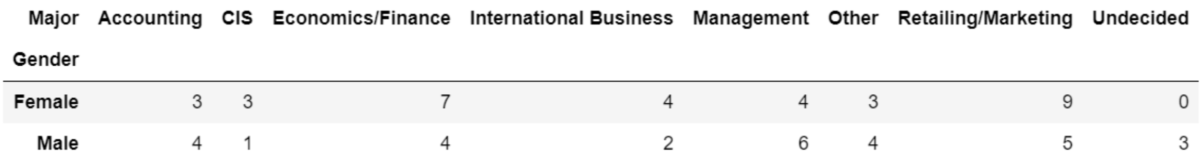
Problem 2 (Undergraduate students)

Problem Statement:

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)

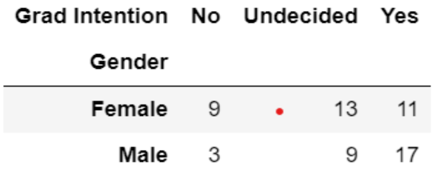
2.1.1. Gender and Major



Major	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.6

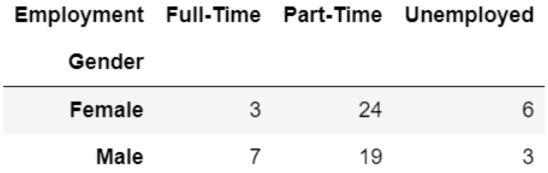
2.1.2. Gender and Grad Intention



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

Table.7

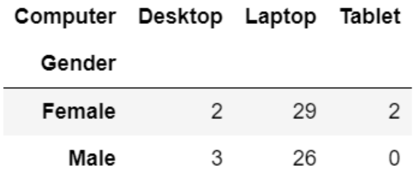
2.1.3. Gender and Employment



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

Table.8

2.1.4. Gender and Computer



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

Table.9

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

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

- Probability that a randomly selected CMSU student will be male is 0.47

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

- Probability that a randomly selected CMSU student will be female is 0.53

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.

- Probability of different majors among the male students in CMSU are as below:
- Probability of "Accounting" is 0.14
- Probability of "CIS" is 0.03
- Probability of "Economics/Finance" is 0.14
- Probability of "International Business" is 0.04
- Probability of "Management" is 0.21
- Probability of "Other" is 0.14
- Probability of "Retailing/Marketing" is 0.17
- Probability of "Undecided" is 0.1

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

- Probability of different majors among the female students in CMSU are as below:
- Probability of "Accounting" is 0.09
- Probability of "CIS" is 0.09
- Probability of "Economics/Finance" is 0.21
- Probability of "International Business" is 0.12
- Probability of "Management" is 0.12
- Probability of "Other" is 0.09
- Probability of "Retailing/Marketing" is 0.27
- Probability of "Undecided" is 0.0

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.

- Probability that a randomly chosen student is a male and intends to graduate is 0.27

2.4.2 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 0.06

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?

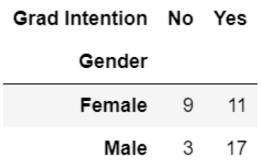
- Probability that a randomly chosen student is a male or has full time employment is 0.52

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

- Probability that given randomly chosen person is female student and she is majoring in international business or management is 0.24

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?

Contingency table of Gender and Indent to Graduate at 2 levels (Yes/No):



Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table.10

- Probability of grad intention [p(A)] is 0.7
- Probability of being female [p(B)] is 0.5
- Probability of grad intention and being female [p(A∩B)] is 0.275
- p(A∩B) ≠ p(A)*p(B)
- Both 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?

- Probability that his/her GPA is less than 3 is 0.37

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.

- Probability that a randomly selected male earns 50 or more is 0.48
- Probability that a randomly selected female earns 50 or more is 0.55

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.

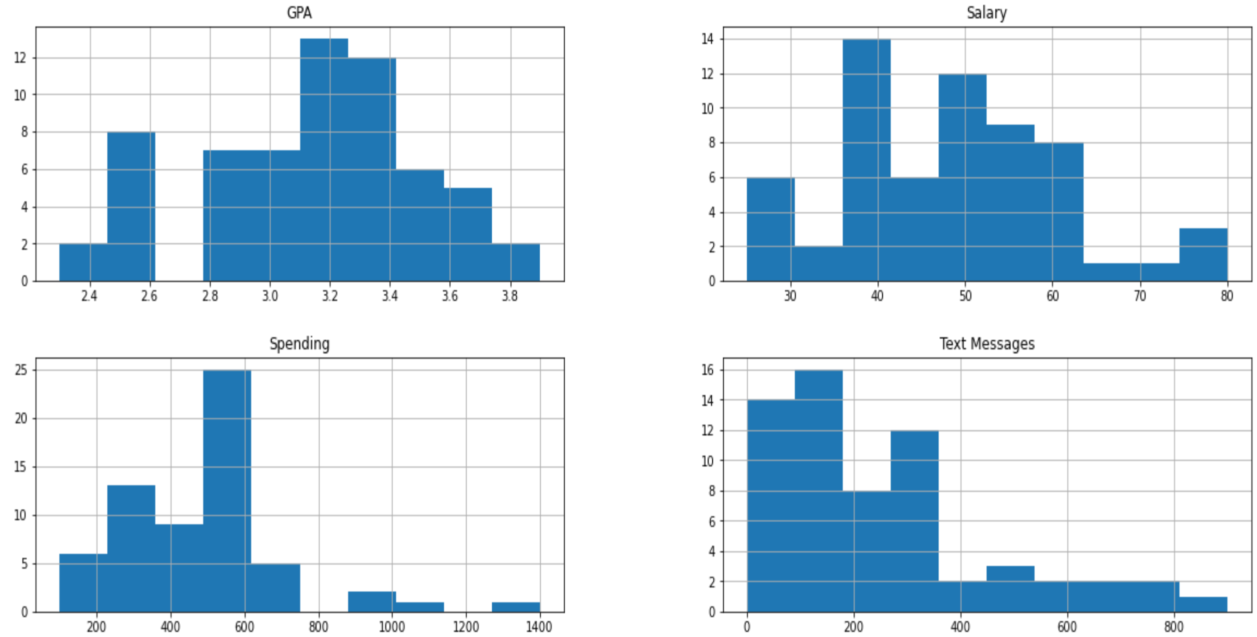


Fig.6

- Spending and text messages are highly right skewed
- GPA and salary distribution is approximately symmetric

Problem 3 (Shingles moisture analysis)

Problem statement:

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.

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

We will do hypothesis testing for both types of shingles (A, B) to check moisture contents are within the permissible limits or not.

For A type shingles:

- Step-1: Define null and alternate hypothesis

H_0: Moisture contents of A type shingles is less than or equal to 0.35

H_1: Moisture contents of A type shingles is more than 0.35

- Step-2: Define the significance level

α=0.05

- Step-3: Identify the test statistic

We do not know the standard deviation and n=36, so we use the t distribution and t-stat for **1 sample T-test**

- Step-4: Calculate p-value and test statistic

By python programming, it is calculated that t-value is -1.47 and p-value is 0.07

- Step-5: Decide to reject or accept null hypothesis

p-value is greater than confidence level, so there is no enough evidence to reject the null hypothesis

For B type shingles:

- Step-1: Define null and alternate hypothesis

H_0: Moisture contents of B type shingles is less than or equal to 0.35

H_1: Moisture contents of B type shingles is more to 0.35

- Step-2: Define the significance level

α=0.05

- Step-3: Identify the test statistic

We do know the standard deviation and n=31, so we use the t distribution and t-stat for **1 sample T-test**

- Step-4: Calculate p-value and test statistic

By python programming, it is calculated that t-value is -3.10, p-value is 0.0020

- Step-5: Decide to reject or accept null hypothesis

p-value is less than confidence level, so we reject null hypothesis i.e., there is enough evidence to reject the null hypothesis

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

We perform hypothesis testing to check for evidence whether mean values of moisture contents of shingles A and B are equal or not

- Step-1: Define null and alternate hypothesis

H_0: Mean of A = Mean of B

H_1: Mean of A ≠ Mean of B

- Step-2: Define the significance level

α=0.05

- Step-3: Identify the test statistic

We have two samples and we do not know the population standard deviation. So, we use the t distribution and the t-stat for **two sample unpaired test.**

- Step-4: Calculate p-value and test statistic

By python programming, it is calculated that t-value is 1.29 and p-value is 0.20

- Step-5: Decide to reject or accept null hypothesis

p-value is greater than 0.05, so there is no enough evidence to reject null hypothesis

THE END