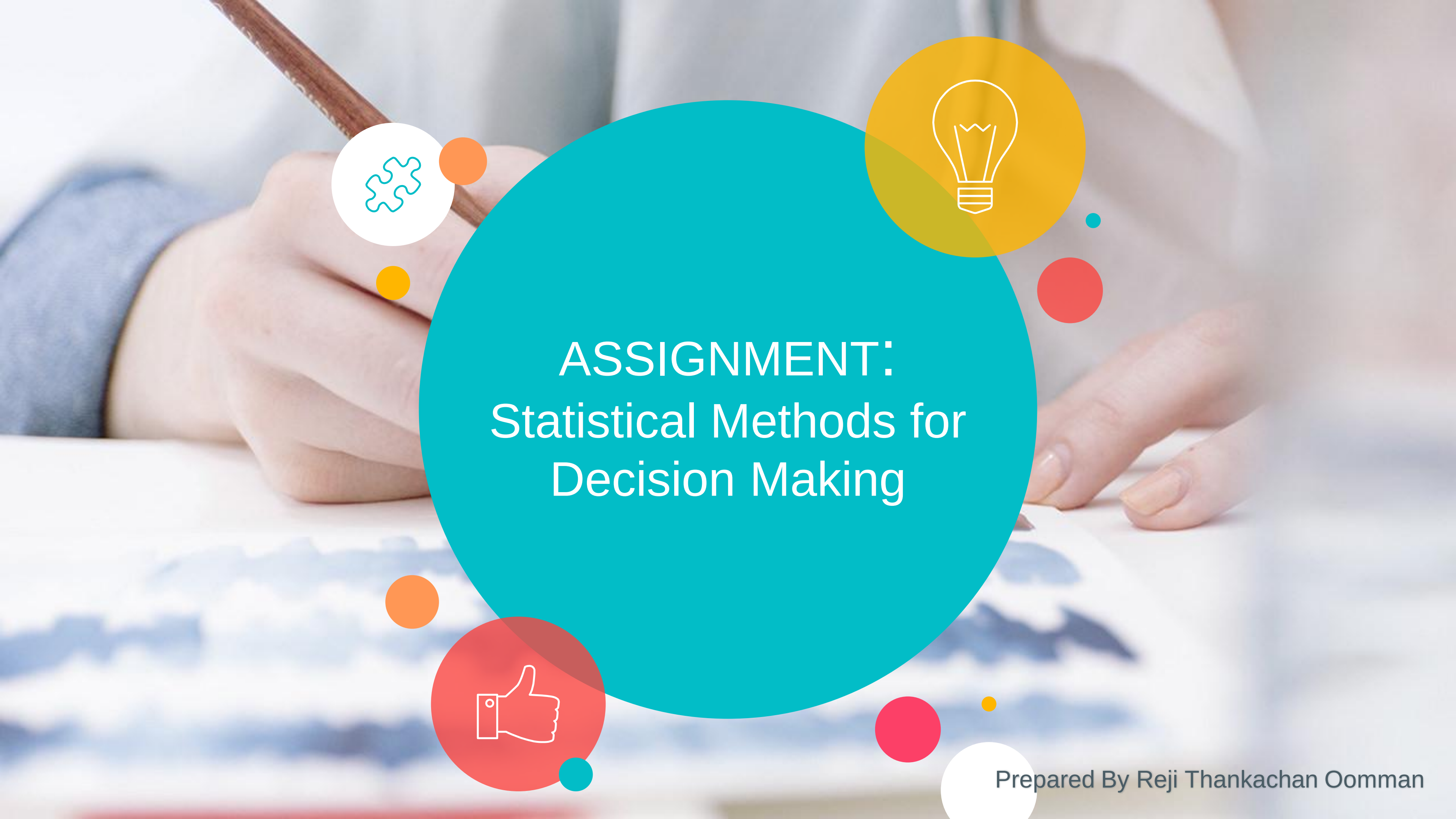


ASSIGNMENT: Statistical Methods for Decision Making

Prepared By Reji Thankachan Oomman

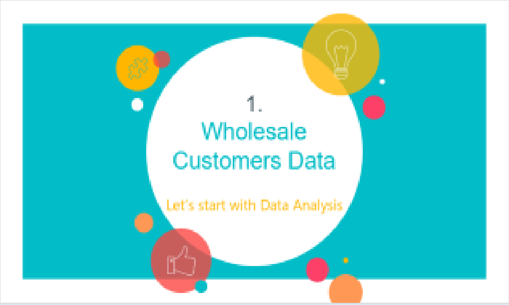
Preface:

This assignment involves drawing inferences from 3 case studies, namely - Wholesale Customer Data (Store Sales), University Survey Data & Manufacturing Shingles Data. The concepts of various measures of Descriptive Statistics, Probability and Probability Distributions and various measures of Estimation & Hypothesis Testing are used to analyse these case studies

Skills and Tools

Descriptive Statistics, Probability & Probability Distributions, Estimation, Hypothesis Testing

CASE STUDIES:



1. Wholesale Customers Data

Let's start with Data Analysis

2. Clear Mountain State University: Student survey

Let's start with Data Analysis

3. ABC asphalt shingles

Let's start with A & B sample analysis

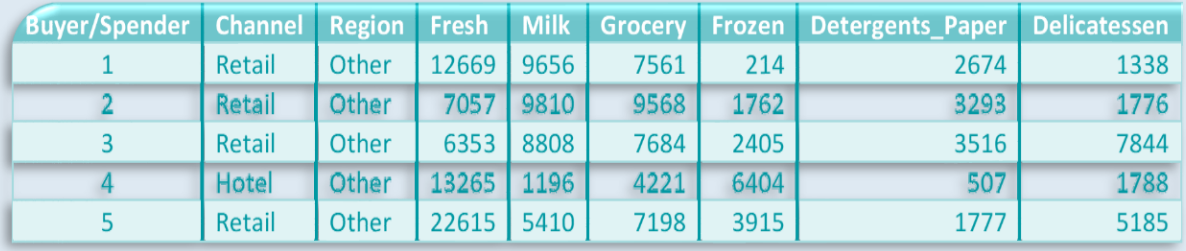
1.
Wholesale Customers Data

Let’s start with Data 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).

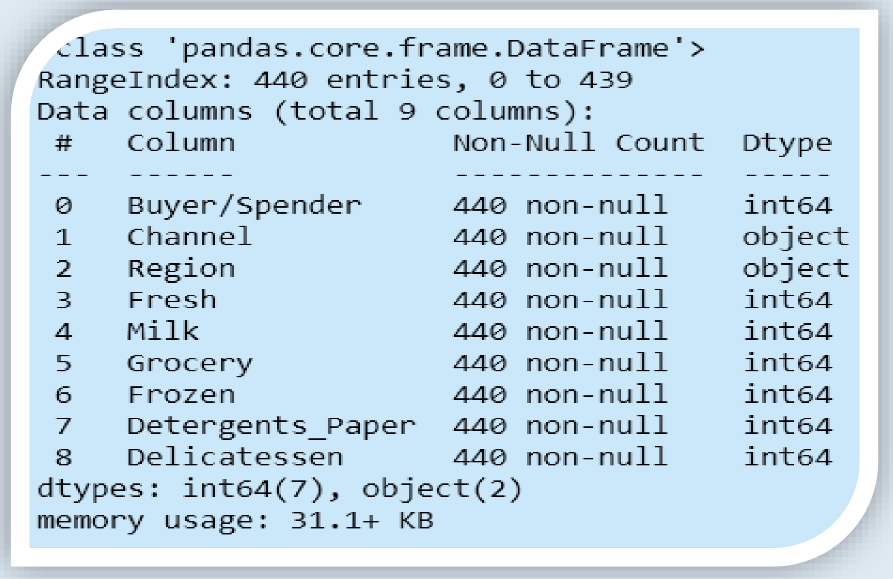
INTRODUCTION



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

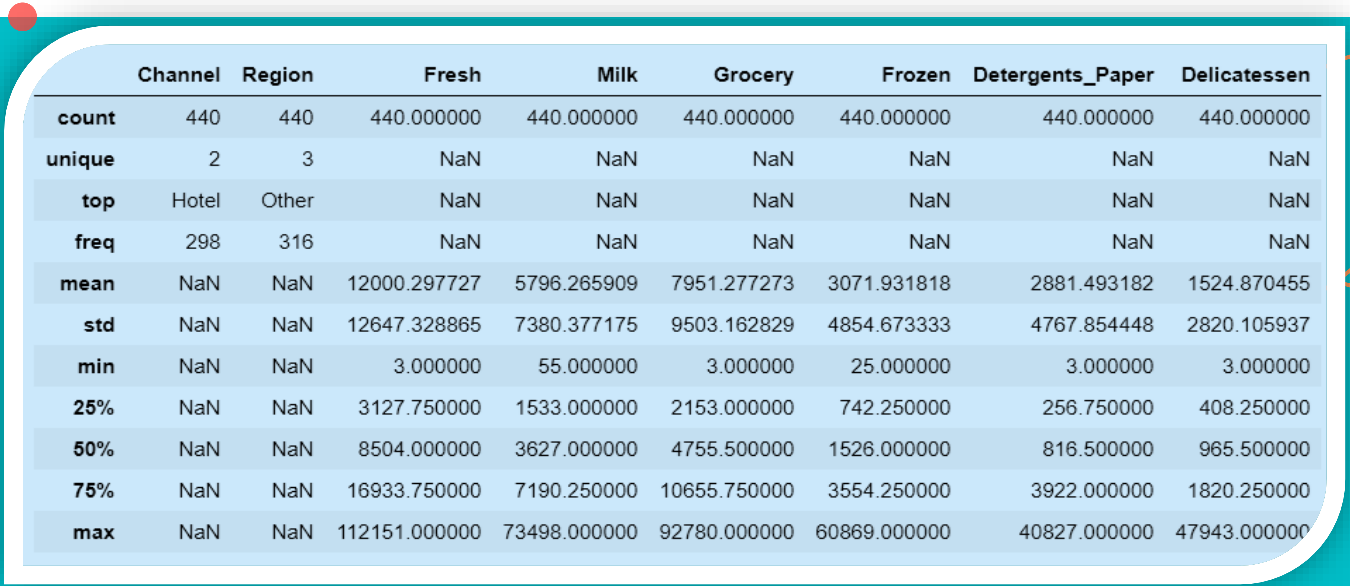
Sample Dataset has total of 440 entries and 9 columns. Observing the first five entries where ‘Buyer/Spender’ is a unique identifier integer type column, ‘Channel’ and ‘Region’ both are object type categorical columns while 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen' is integer type continuous columns.

Checking for null values



```
class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Buyer/Spender      440 non-null    int64
 1   Channel            440 non-null    object
 2   Region             440 non-null    object
 3   Fresh              440 non-null    int64
 4   Milk               440 non-null    int64
 5   Grocery            440 non-null    int64
 6   Frozen             440 non-null    int64
 7   Detergents_Paper   440 non-null    int64
 8   Delicatessen       440 non-null    int64
dtypes: int64(7), object(2)
memory usage: 31.1+ KB
```

From the above results, it is evident that there are no null values present in the dataset.



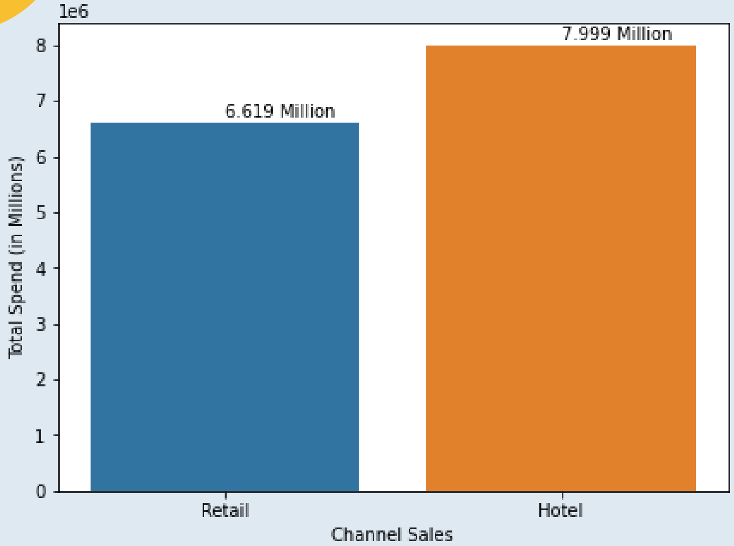
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	2	3	NaN	NaN	NaN	NaN	NaN	NaN
top	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN
freq	298	316	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	12000.297727	5796.265909	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

Exploratory Data Analysis

Following observations can be made from above descriptive table:

- ✓ "Hotel" has the top frequency among two Channels
- ✓ "Other" has the top frequency among three Regions
- ✓ "Fresh" items has highest range of spending with 112,148 monetary units.
- ✓ "Detergents_Paper" items has lowest range of spending with 40,824 monetary units.

1.1 Channel wise analysis in terms of spending

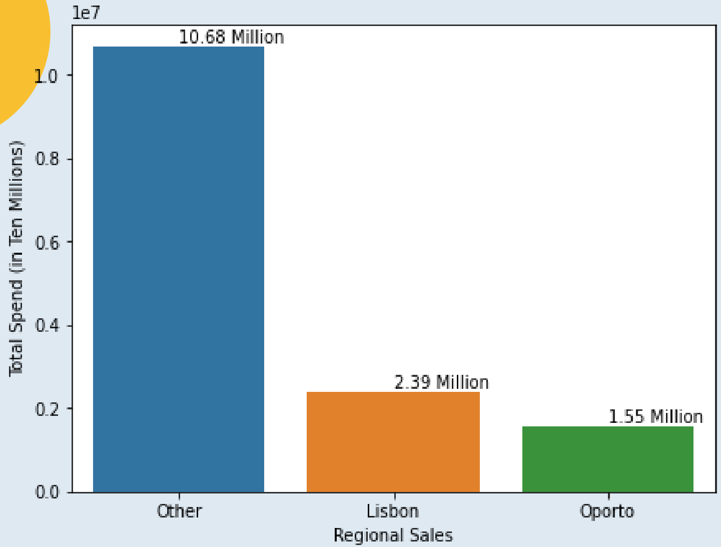


Channel	Total Spend
Hotel	7,999,569
Retail	6,619,931

In terms of channel wise observation from the above plot shows there are two channels – Hotel and Retail:

- Channel which seems to spend more is **Hotel with total spending of 7.99 million monetary units.**
- Channel which seems to spend less is **Retail with total spending of 6.62 million monetary units.**

1.1 Region wise analysis in terms of spending

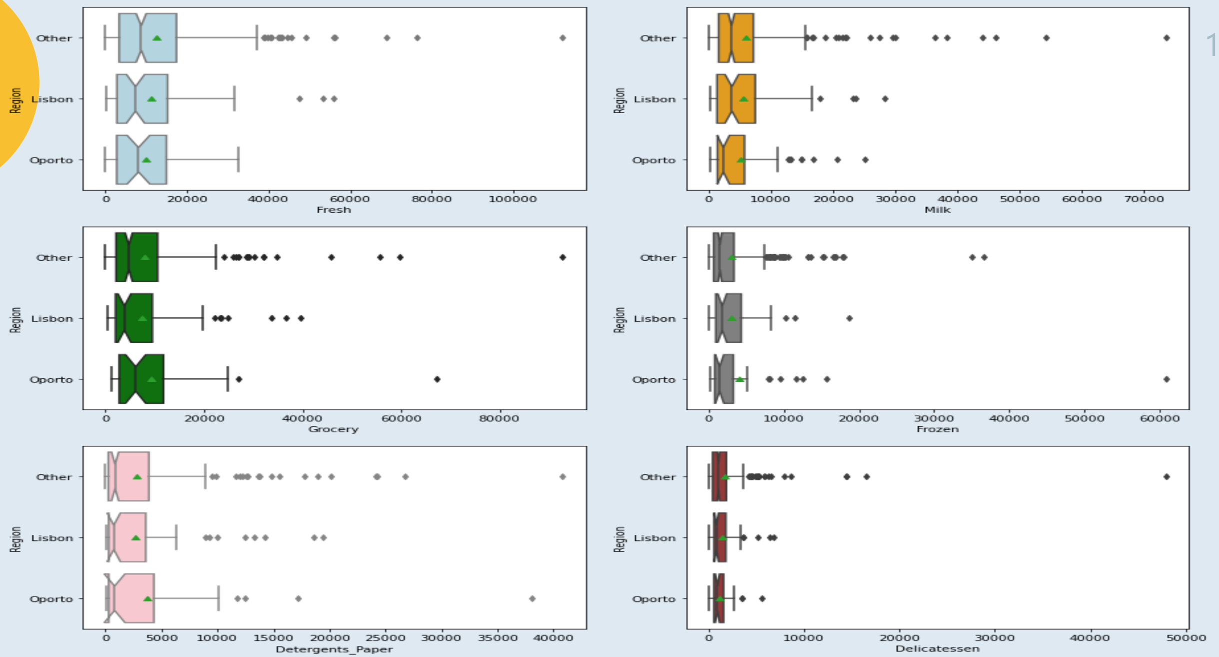


Region	Total Spend
Other	10,677,599
Lisbon	2,386,813
Oporto	1,555,088

In terms of region wise observation from the above plot shows that there are three regions – **Lisbon, Oporto and Other**:

- **Region which seems to spend more** is **Other region** with total spending of 10.677 million monetary units.
- **Region which seems to spend less** is **Oporto region** with total spending of 1.555 million monetary units.

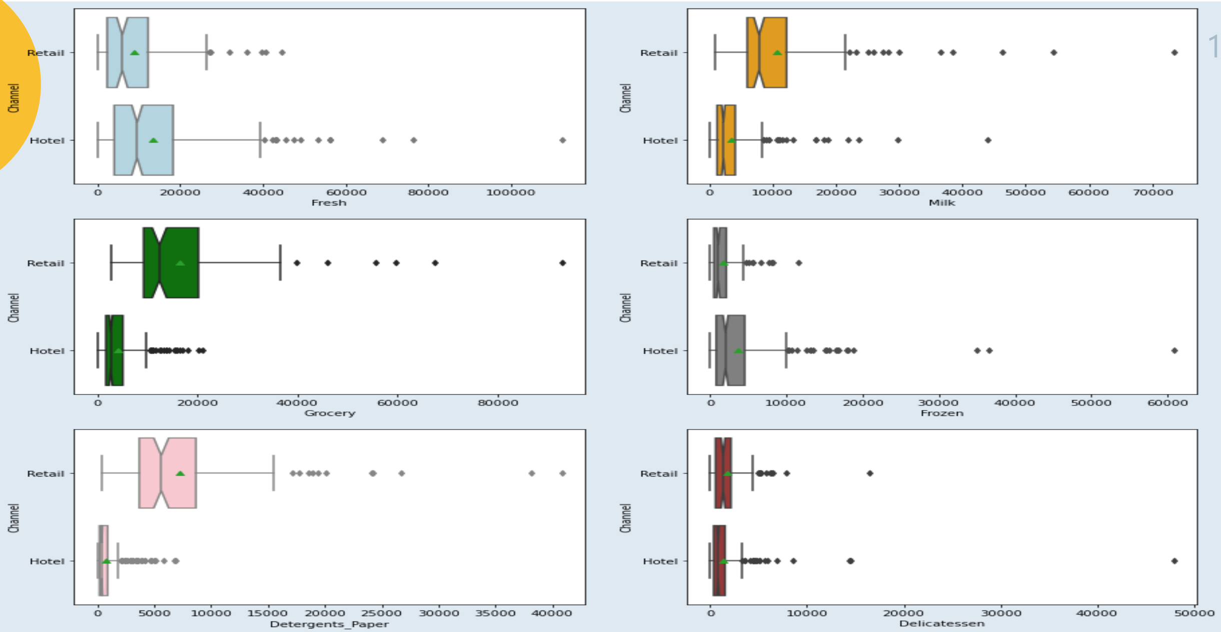
1.2 Region wise Behavior analysis in 6 different items



The above plot shows the behavior of all six items individually across three regions:

- All six items has **mean which is more than the median** across all three regions.
- All six items are **right skewed (positively skewed) and asymmetric** across all three regions.
- **Other region** has **maximum number of outlier data values** in all six items when compared to other two regions, out of which one of them is common in all six items (will be discovered in further analysis).

1.2
Channel wise Behavior analysis in 6 different items



The above plot shows the behavior of all six items individually across two Channels:

- All six items has **mean which is more than the median** across both channels
- All six items are **right skewed (positively skewed) and asymmetric** across both channels
- **Both Hotel and Retail Channels** has **outlier data values** in all six items (will be discovered in further analysis).

1.3 Behavioural Consistency of all items: most and least inconsistent

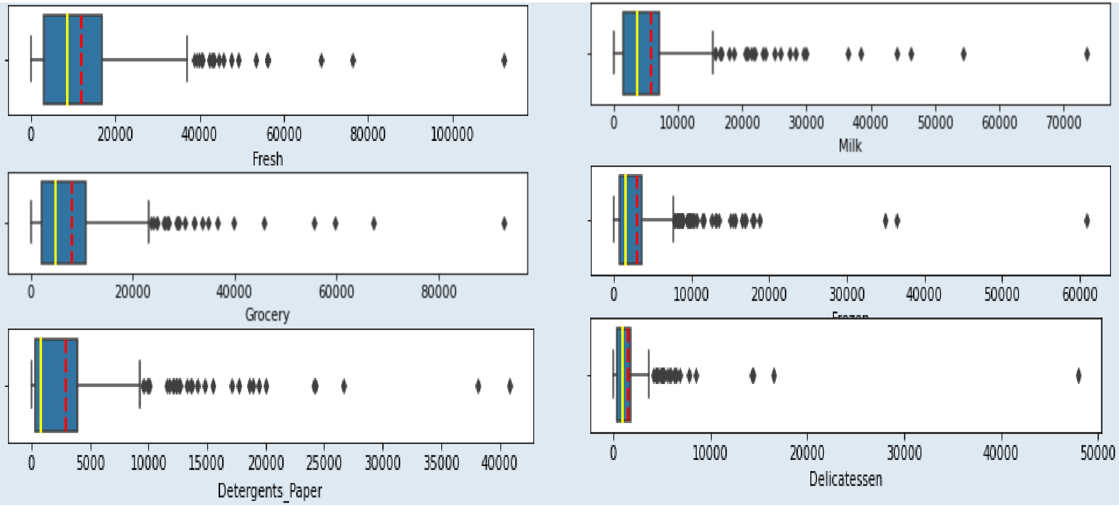
ITEMS	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Mean	12000	5796	7951	3071	2881	1524
Standard Deviation	12647	7380	9503	4854	4767	2820
Coefficient of Variance %	105%	127%	120%	158%	165%	185%

By observing coefficient of variation % we can check the relative measure of variability in each item:

❖ Delicatessen items shows the most inconsistent behavior with coefficient of variation percentage of 185 % which means more is the spread of the data around its mean making it most inconsistent when compared to other items.

❖ Whereas fresh items show the least inconsistent behavior with coefficient of variation percentage of 105 % which means less is the spread of the data around its mean making it least inconsistent when compared to other items.

1.4
Outliers in Data



Both visualizing and quantitative approach into each item column was performed to confirm the presence of outliers and following outlier rows were found:

- ❑ There are 67 Outliers in at least 1 column, or Outliers in at least 1 column are: [129, 258, 259, 2, 4, 393, 143, 401, 17, 277, 406, 24, 409, 282, 283, 155, 285, 29, 287, 413, 289, 411, 36, 38, 425, 173, 431, 176, 303, 435, 52, 436, 309, 310, 312, 195, 196, 70, 71, 72, 73, 328, 200, 202, 205, 334, 337, 338, 339, 209, 88, 218, 91, 93, 349, 351, 230, 239, 112, 240, 370, 371, 372, 377, 126, 381, 254]
- ❑ There are 41 outliers in at least 2 columns, or Outliers in more than 1 columns are: [384, 265, 145, 22, 23, 28, 284, 163, 39, 40, 171, 427, 45, 43, 47, 304, 49, 181, 437, 183, 56, 61, 319, 65, 325, 201, 331, 77, 333, 211, 85, 86, 87, 216, 343, 92, 358, 103, 109, 251, 125]
- ❑ There are 17 outliers in at least 3 columns, or Outliers in more than 2 columns are: [65, 163, 325, 47, 92, 49, 145, 211, 181, 85, 86, 56, 183, 251, 28, 61, 216]
- ❑ There are 4 outliers in at least 4 columns, or Outliers in more than 3 columns are: [251, 92, 181, 47]
- ❑ There is 1 outlier in at least 5 columns, or Outliers in more than 4 columns are: [47]
- ❑ There is 1 outlier in all 6 columns, or Outliers in more than 5 columns are: [47]

1.5 Recommendations for Business growth

Following recommendations are suggested from business perspective:

1. Region

As the other region has more spending, the business in Lisbon and Oporto should be focused for business growth opportunity.

2. Channel

As the Hotel channel has more spending, the business in Retail channel should be focused for business growth opportunity.

3. Perishable Items

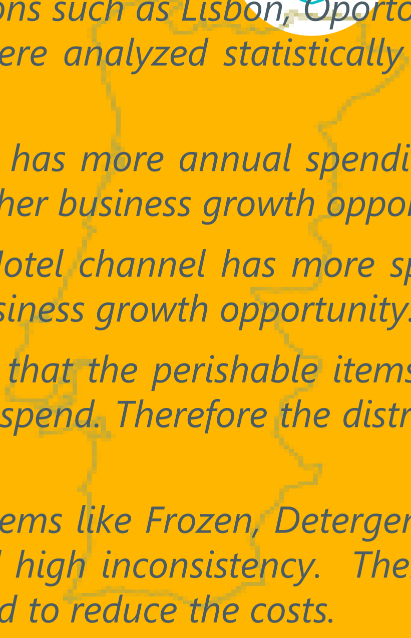
The data clearly shows the perishable items like Fresh , Milk and Grocery items are having higher average spend. Therefore the distribution of these items needs to be in focus to increase the sales.

4. Non Perishable Items

The data clearly shows the non perishable items like Frozen, Detergents_Paper and Delicatessen are having lower average spend and high inconsistency. Therefore the distribution of these items needs to be further analyzed to reduce the costs.

Summary:

A Portuguese wholesale distributor data of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions such as Lisbon, Oporto, Other through 2 distinct sales channel which are Hotel and retail were analyzed statistically and following conclusions are made:



- *In terms of regions Other region has more annual spending, the business in Lisbon and Oporto should be focused for further business growth opportunity.*
- *With respect to sales channels Hotel channel has more spending, the business in Retail channel should be focused for business growth opportunity.*
- *It is quite evident from the data that the perishable items like Fresh , Milk and Grocery items are having higher average spend. Therefore the distribution of these items needs to be in focus to increase the sales.*
- *Apparently, the non perishable items like Frozen, Detergents_Paper and Delicatessen are having lower average spend and high inconsistency. Therefore the distribution of these items needs to be further analyzed to reduce the costs.*

2.

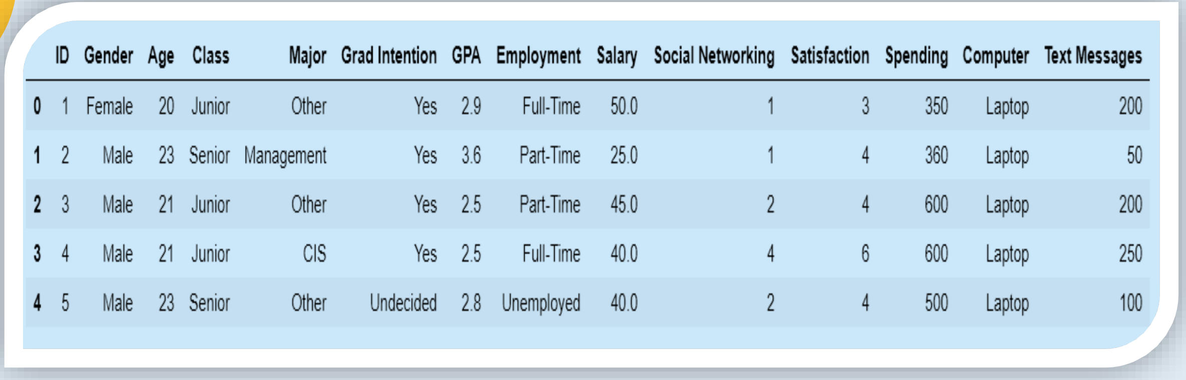
Clear Mountain State University: Student survey

Let’s start with Data Analysis

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

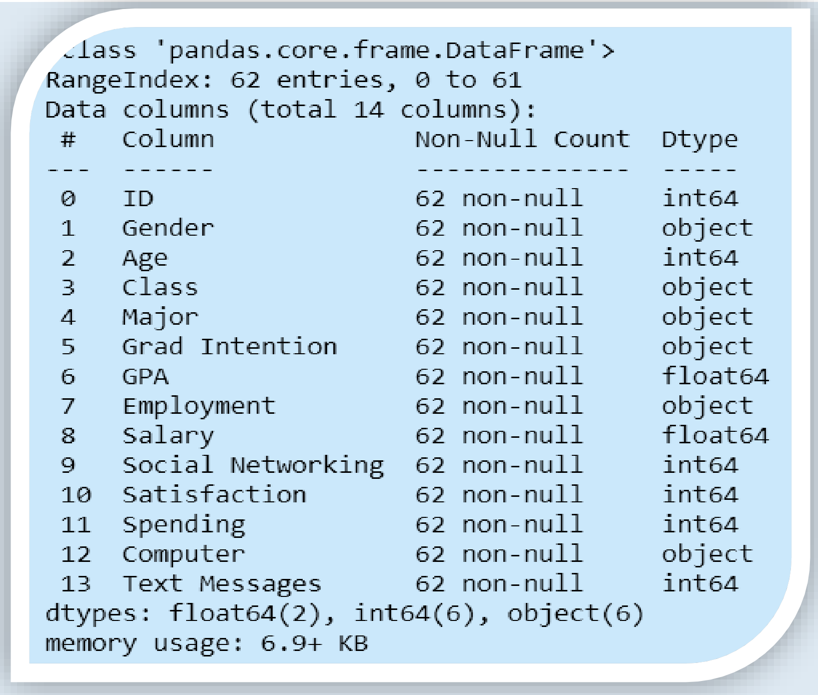
INTRODUCTION



	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

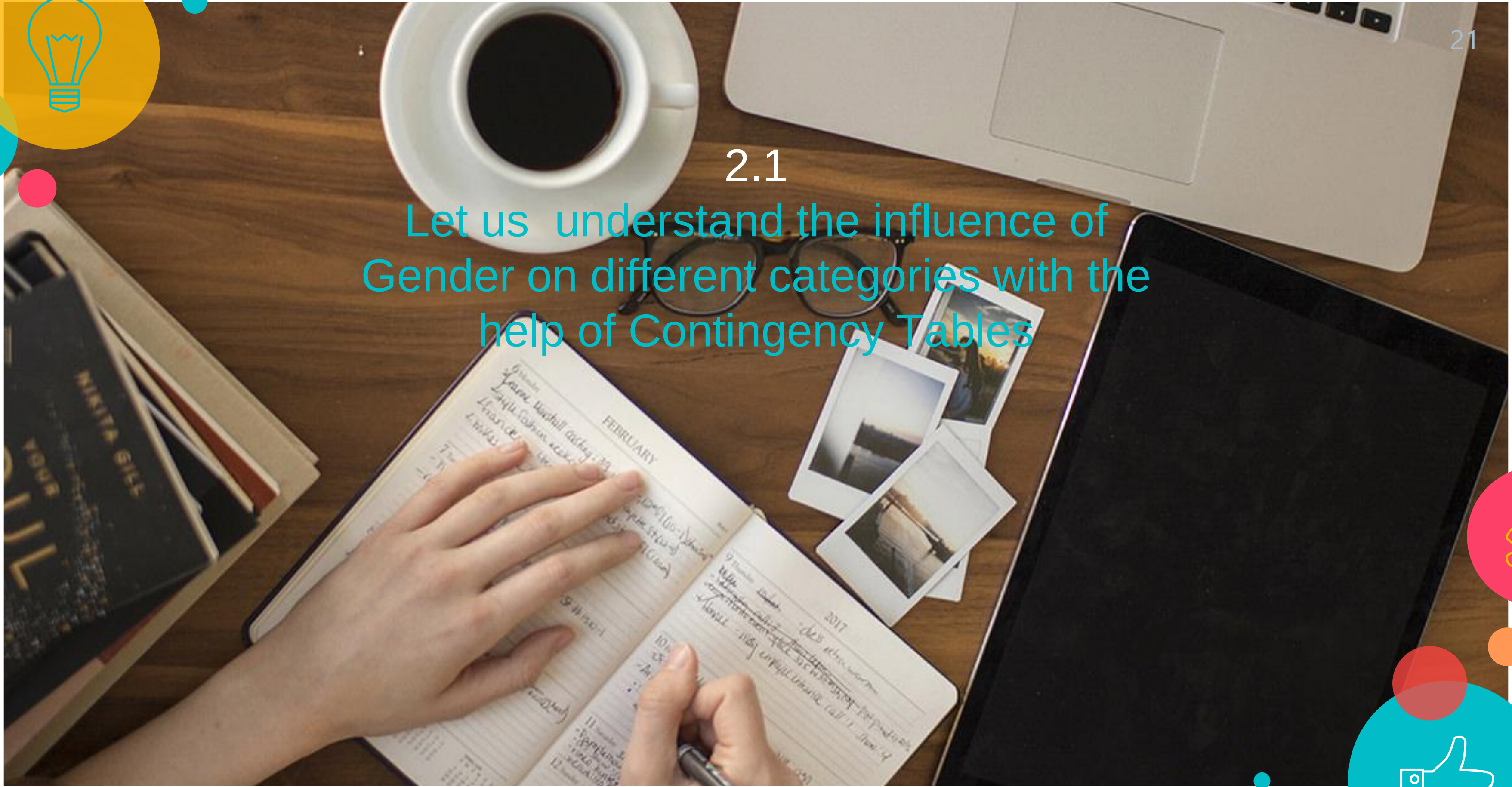
Sample Dataset has total of 62 entries and 14 columns. Observing the first five entries where ‘ID’ is a unique identifier integer type column, ‘Gender’, ‘Class’, ‘Major’, ‘Grad intention’, ‘Employment’, and ‘Computer’ are object type categorical columns while ‘Age', ‘Social Networking', ‘Satisfaction', ‘Spending’ and ‘Text Messages’ are integer type continuous columns and ‘GPA', ‘Salary’ are float type continuous columns.

Checking for null values



```
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
```

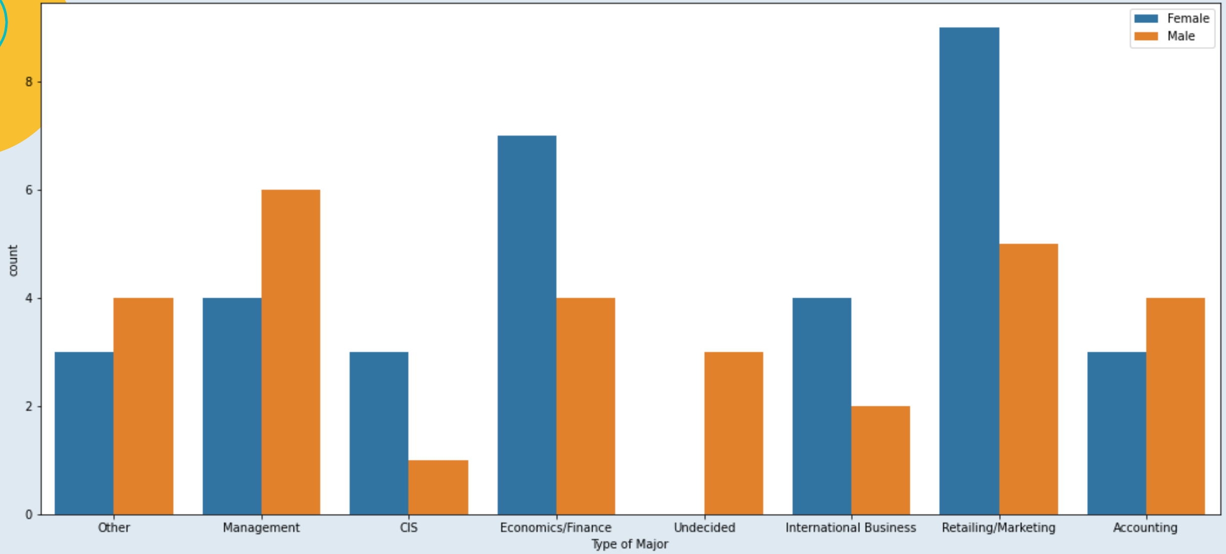
From the above results, it is evident that there are no null values present in the dataset.



2.1

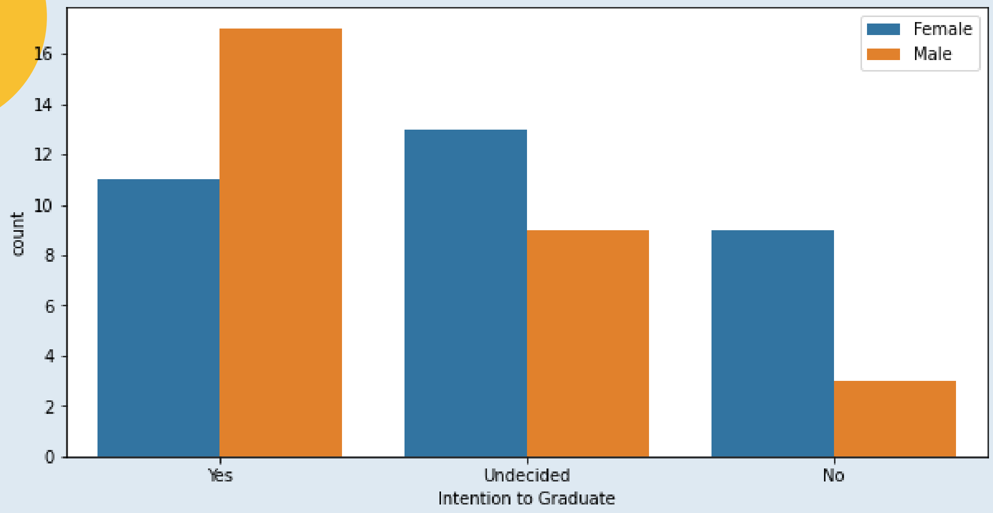
Let us understand the influence of Gender on different categories with the help of Contingency Tables

2.1.1 Contingency Table : Gender and Major



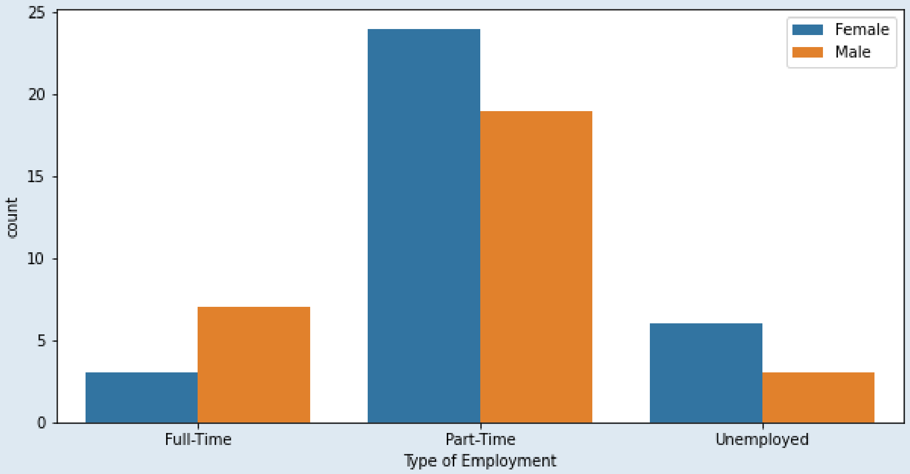
Major Gender	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/ Marketing	Undecided	Marginal
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Marginal	7	4	11	6	10	7	14	3	62

2.1.2
Contingency Table :
Gender and Grad Intention



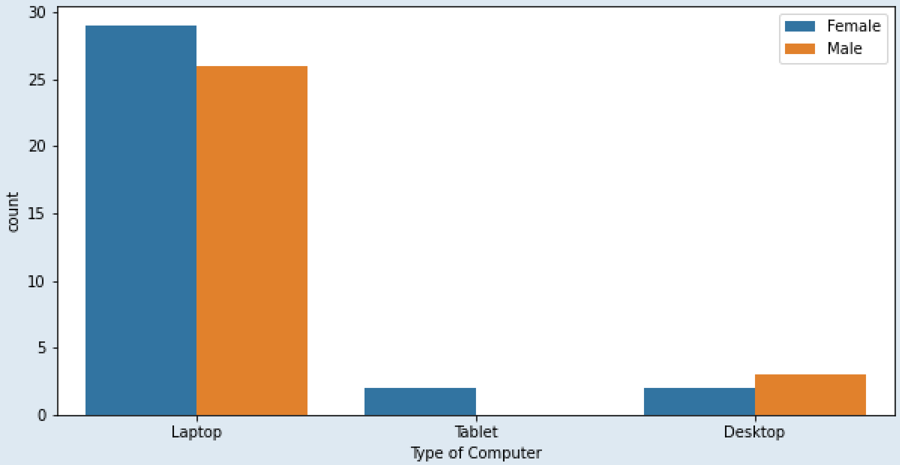
Grad Intention Gender	No	Undecided	Yes	Marginal
Female	9	13	11	33
Male	3	9	17	29
Marginal	12	22	28	62

2.1.3 Contingency Table : Gender and Employment

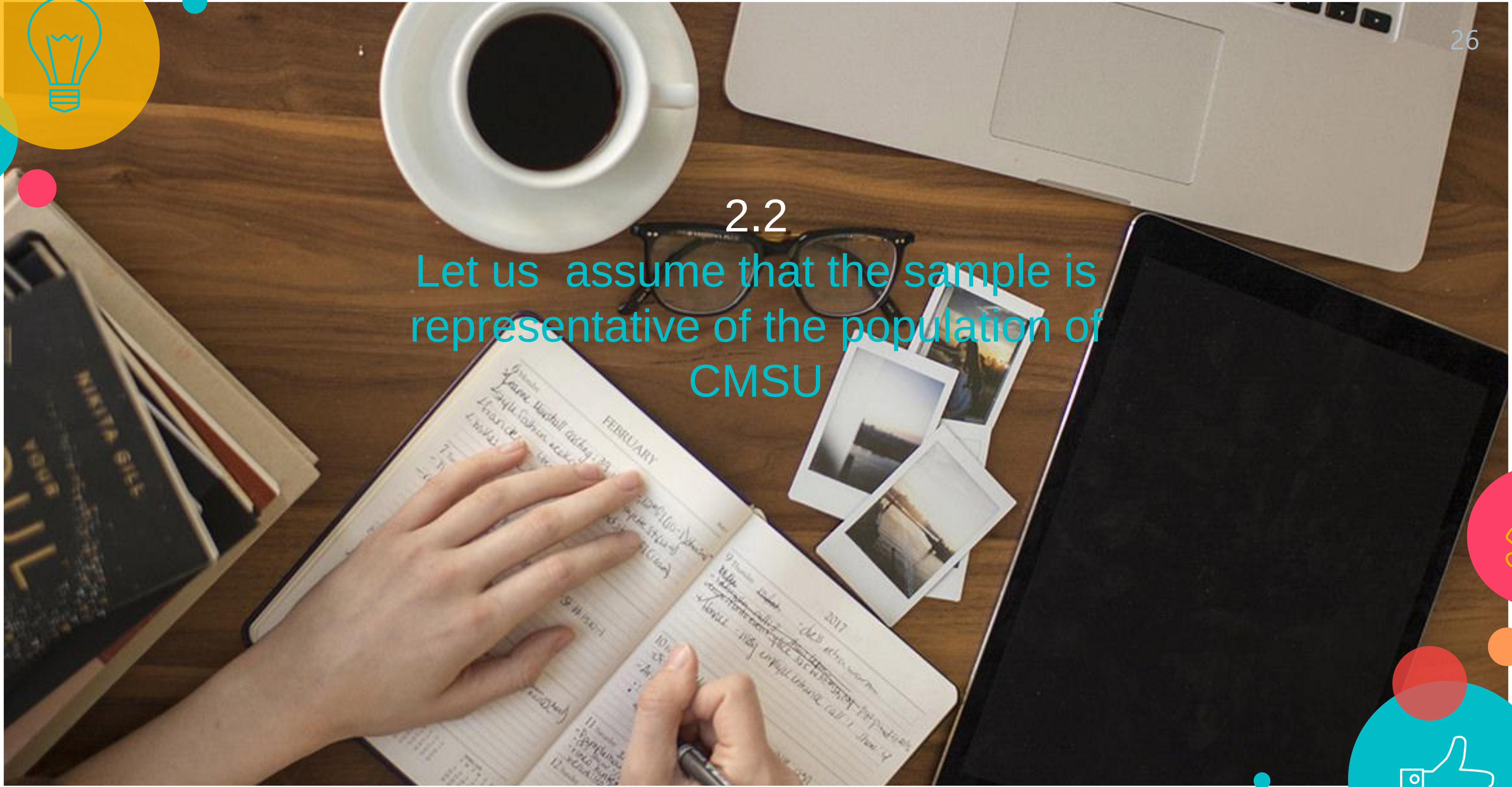


Employment Gender	Full-time	Part-time	Unemployed	Marginal
Female	3	24	6	33
Male	7	19	3	29
Marginal	10	43	9	62

2.1.4 Contingency Table : Gender and Computer



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



2.2

Let us assume that the sample is representative of the population of CMSU

2.2.1. The probability that a randomly selected CMSU student will be male

Total students in sample = 62

29 Male students	33 Female Students

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

The probability of randomly selected CMSU student will be male is 29/62 which is 0.4677
Or The chances of random selection of male student is 46.77%

2.2.2. The probability that a randomly selected CMSU student will be female

Total students in sample = 62

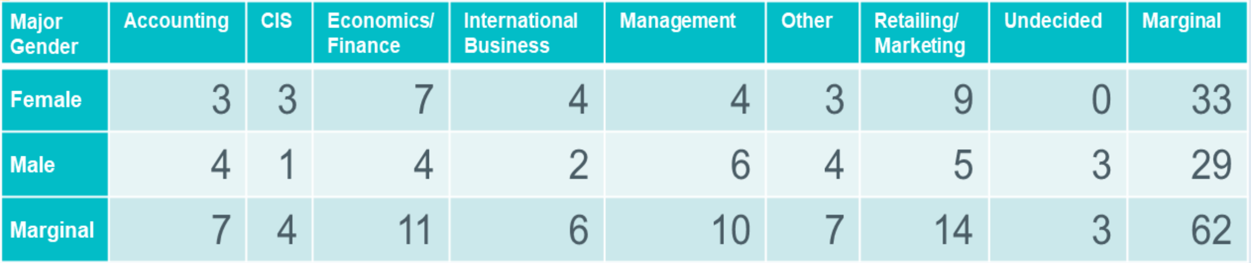
29 Male students	33 Female Students

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

The probability of randomly selected CMSU student will be female is 33/62 which is 0.5322
Or The chances of random selection of male student is 53.22%

2.3.1.

The conditional probability of different majors among the male students in CMSU



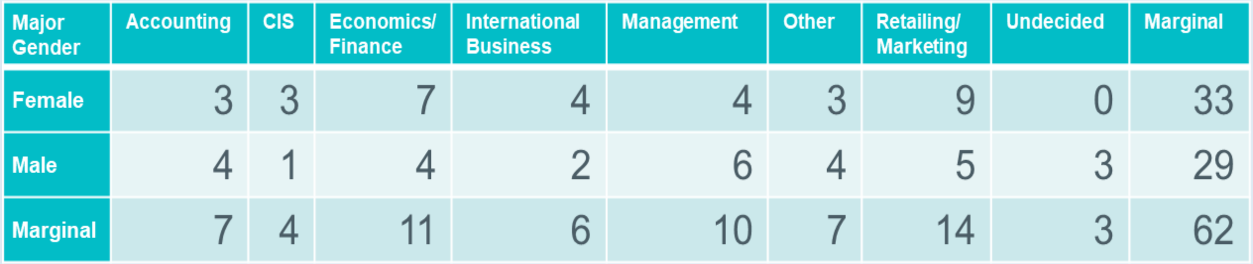
Major Gender	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/ Marketing	Undecided	Marginal
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Marginal	7	4	11	6	10	7	14	3	62

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- The conditional probability of Accounting majors among the male students in CMSU is 0.138 or chances are 13.79 %
- The conditional probability of CIS majors among the male students in CMSU is 0.034 or chances are 3.45 %
- The conditional probability of Economics/Finance majors among the male students in CMSU is 0.138 or chances are 13.79 %
- The conditional probability of International Business majors among the male students in CMSU is 0.069 or chances are 6.90 %
- The conditional probability of Management majors among the male students in CMSU is 0.207 or chances are 20.69 %
- The conditional probability of Other majors among the male students in CMSU is 0.138 or chances are 13.79 %
- The conditional probability of Retailing/Marketing majors among the male students in CMSU is 0.172 or chances are 17.24 %
- The conditional probability of Undecided majors among the male students in CMSU is 0.103 or chances are 10.34 %

2.3.2.

The conditional probability of different majors among the female students in CMSU

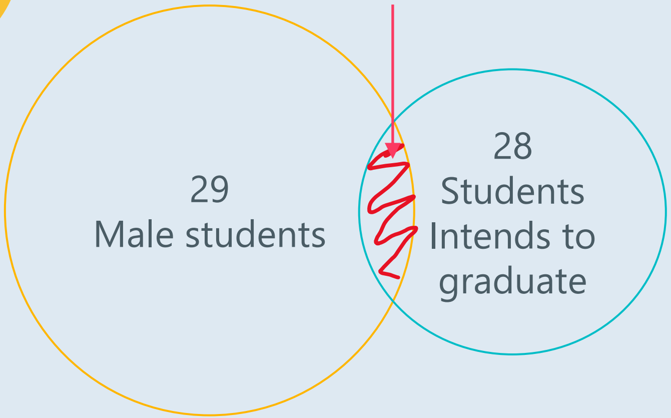


Major Gender	Accounting	CIS	Economics/ Finance	International Business	Management	Other	Retailing/ Marketing	Undecided	Marginal
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
Marginal	7	4	11	6	10	7	14	3	62

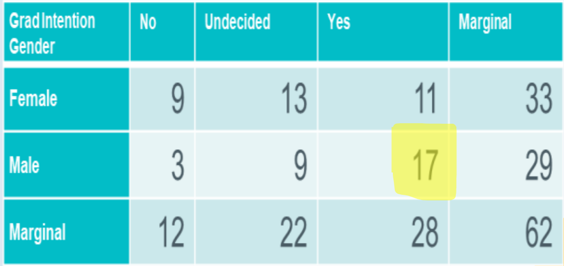
Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- ❖ The conditional probability of Accounting majors among the female students in CMSU is 0.09091 or chances are 9.09 %
- ❖ The conditional probability of CIS majors among the female students in CMSU is 0.09091 or chances are 9.09 %
- ❖ The conditional probability of Economics/Finance majors among the female students in CMSU is 0.21212 or chances are 21.21 %
- ❖ The conditional probability of International Business majors among the female students in CMSU is 0.12121 or chances are 12.12 %
- ❖ The conditional probability of Management majors among the female students in CMSU is 0.12121 or chances are 12.12 %
- ❖ The conditional probability of Other majors among the female students in CMSU is 0.09091 or chances are 9.09 %
- ❖ The conditional probability of Retailing/Marketing majors among the female students in CMSU is 0.27273 or chances are 27.27 %
- ❖ The conditional probability of Undecided majors among the female students in CMSU is 0.00 or chances are 0 %

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



17 male students who intend to graduate

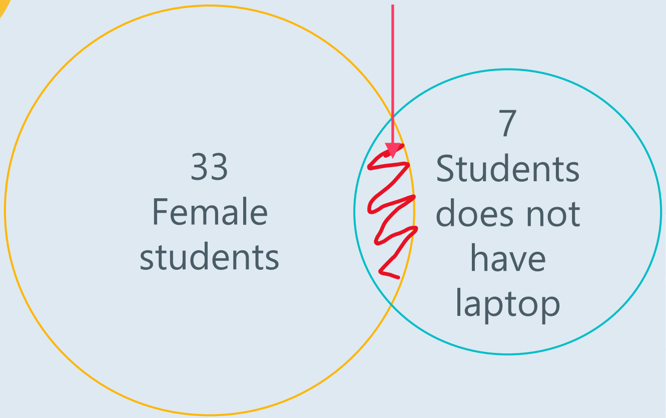


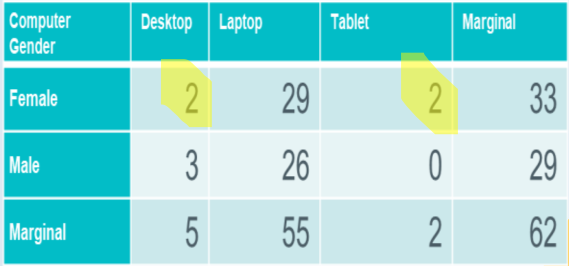
Grad Intention Gender	No	Undecided	Yes	Marginal
Female	9	13	11	33
Male	3	9	17	29
Marginal	12	22	28	62

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- ❖ Number of Male students who intend to graduate is : 17
- ❖ Number of male students is : 29
- ❖ The probability that a randomly chosen student is a male and intends to graduate is 0.5862 or chances are 58.62 %

2.4.2. The probability that a randomly chosen student is a female and does NOT have a Laptop.


4 female students who does not have laptop

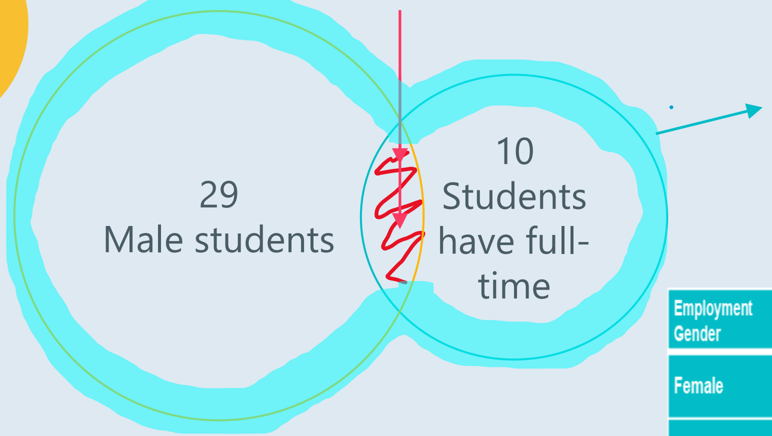


Computer Gender	Desktop	Laptop	Tablet	Marginal
Female	2	29	2	33
Male	3	26	0	29
Marginal	5	55	2	62

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- Number of Female students who does not have laptop : 4
- Number of Female students is : 33
- The probability that a randomly chosen student is a female and does not have a laptop is 0.1212 or chances are 12.12 %

2.5.1. The probability that a randomly chosen student is either a male or has full-time employment.



7 male students who have full-time employment

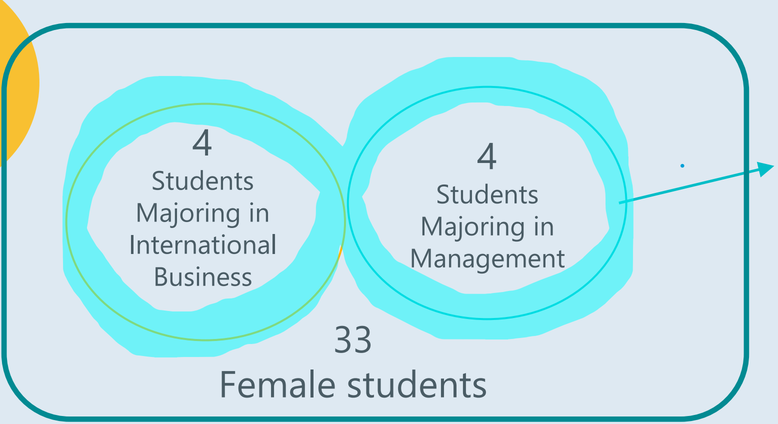
Highlighted Area corresponding to either Male or Full-time students

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

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- Probability of male students is : 0.468
- Probability of students with full-time employment is : 0.161
- Probability of male students with full-time employment: 0.113
- Since the Probability of male students with full time is included in both probability of male and probability of full-time employment individually. Therefore we will reduce the same from the sum of the both to get the probability that a randomly chosen student is either a male or has full-time employment, which is 0.5161 or chances are 51.61 %.

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


4
Students Majoring in International Business
4
Students Majoring in Management
33
Female students

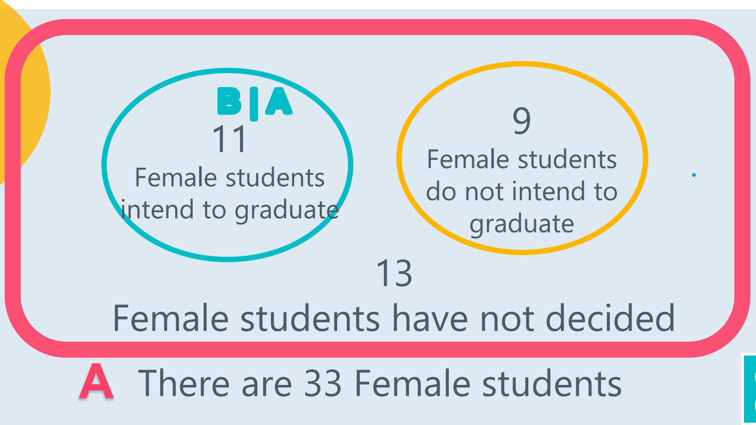
Highlighted Area corresponding to Female students either majoring in International business or Management

Major Gender	International Business	Management
Female	4	4

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- Number of female students who are majoring in international business is : 4
- Probability of female students who are majoring in international business is : 0.1212
- Number of female students who are majoring in Management is : 4
- Probability of female students who are majoring in Management is : 0.1212
- The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is 0.2424 or chances are 24.24 %

2.6.
The graduate intention and being female are independent events.



B|A

11

Female students intend to graduate

9

Female students do not intend to graduate

13

Female students have not decided

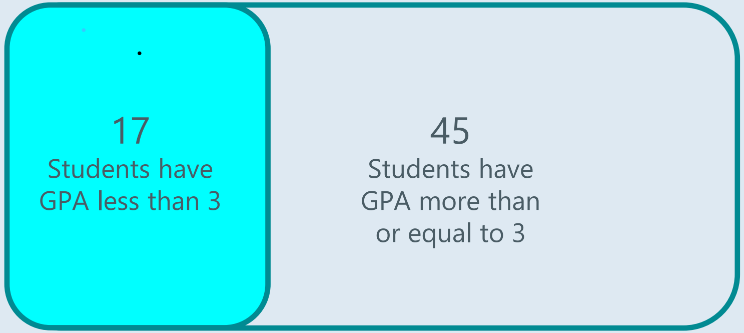
A There are 33 Female students

Gender Grad Intention	Female	Male
No	9	3
Yes	11	17

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- Let A be the event of student being female then P(A) is 0.5322 (i.e.. 33 /62) or 53.22% chances.
- Let B be the event of student intending to graduate then P(B) is 0.4516 (i.e.. 28 /62) or 45.16% chances.
- The conditional probability of female students given intention to graduate is P(A|B) is 0.3929 (i.e.. 11/28) or 39.29% chances.
- The conditional probability of students who intend to graduate given female is P(B|A) is 0.3333 (i.e.. 11/33) or 33.33 % chances.
- Since P(A|B) ≠ P(A) or P(B|A) ≠ P(B), Therefore graduate intention and being female are not independent events.

2.7.1. The probability of student having GPA less than 3



17 Students have GPA less than 3

45 Students have GPA more than or equal to 3

There are total 62 students

GPA Value	No. of Students
Less than 3	17
More than or equal to 3	45
Total	62

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- No. of students with GPA values less than 3 is : 17
- If a student is chosen randomly, the probability that his/her GPA is less than 3 is 0.2742 or have 27.42 % chances

2.7.2. The conditional probability that a randomly selected male earns 50 or more.

Gender	Salary status	No. of students
Male	Salary 50 or more	14
	Salary Less than 50	15
Female	Salary 50 or more	18
	Salary Less than 50	15

14 Male students with salary less than 50	15 Male students with salary 50 or more	18 Female students with salary 50 or more	15 Female students with salary less than 50

There are total 62 students

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- No. of Male students with Salary 50 or more is : 14
- No. of Male students in the sample is : 29
- The conditional probability that a randomly selected male earns 50 or more is 0.4827 or have 48.27 % chances

2.7.2. The conditional probability that a randomly selected female earns 50 or more.

Gender	Salary status	No. of students
Male	Salary 50 or more	14
	Salary Less than 50	15
Female	Salary 50 or more	18
	Salary Less than 50	15

14 Male students with salary 50 or more	15 Male students with salary than 50	18 Female students with salary 50 or more	15 Female students with salary less than 50

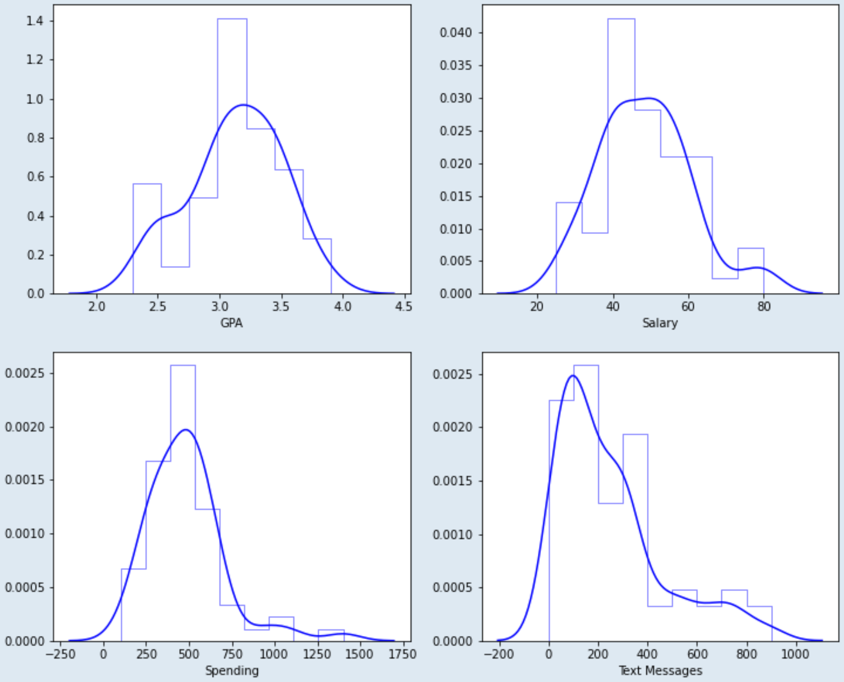
There are total 62 students

Assuming the sample is representative of the population of CMSU and out of total 62 students sample survey collected.

- No. of Female students with Salary 50 or more is : 18
- No. of Female students in the sample is : 33
- The conditional probability that a randomly selected female earns 50 or more is 0.5454 or have 54.55 % chances

2.8

Checking normality of distribution of each of the four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.



Let us analyze the shape of distribution :

- GPA follows almost normal distribution as the mean, median and mode looks closely in line with each other

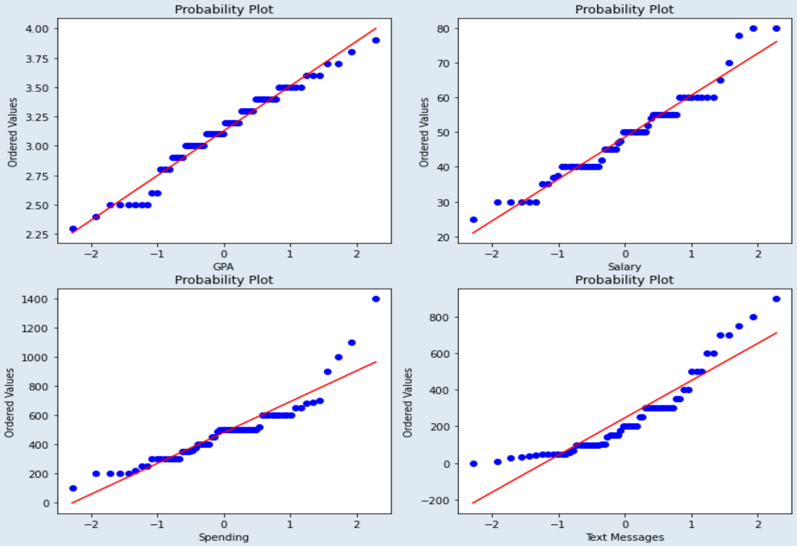
- Salary, Spending and Text messages do not follow normal distribution and are right skewed as the mean looks more towards the right than the median.

Further analysis of normality in distribution can be done with the help of Shapiro-wilk test.

2.8

Shapiro-wilk test:

The Shapiro-Wilk test evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution. The tests assume that that the sample was drawn from a Gaussian distribution. Technically this is called the null hypothesis, or H0. A threshold level is chosen called alpha, typically 5% (or 0.05), that is used to interpret the p-value.



Shapiro-wilk test results:

❑ GPA :
- Statistics = 0.969, p value = 0.112
- GPA distribution looks Normal (fail to reject H0)

❑ Salary :
- Statistics=0.957, p=0.028
- Salary distribution does not look Normal (reject H0)

❑ Spending :
- Statistics=0.878, p=0.000
- Spending distribution does not look Normal (reject H0)

❑ Text Messages :
- Statistics=0.859, p=0.000
- Text Messages distribution does not look Normal (reject H0)

Summary:

The survey data of 62 undergraduate students collected by the Student News Service at Clear Mountain State University (CMSU). Assuming the sample is representative of the population of CMSU following conclusions were made :

- *The strength of Female students (ie. 53%) are slightly more than the Male students (ie. 46%).*
- *Majoring in Management was top preference of Male students while Economics/Finance was of keen interest among Female students*
- *Intention to graduate was more masculine than feminine.*
- *In terms of employment only 14.5% were unemployed, rest 84.5% of students had either part-time or full-time jobs.*
- *Laptops are most commonly used (by approx. 88% students) compared to desktop and tablets.*
- *Less than one-third of students (ie. 27.42%) have GPA score less than 3.*
- *Almost 52% of students have the salary of 50 or more.*

3.

ABC asphalt shingles

Let’s start with A & B sample 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 coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

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

3.1. Testing whether type A sample shingles mean moisture content are within permissible limits.

Step 1 : Define null and alternative hypothesis

In testing the mean moisture content in type A is less than 0.35 pound per 100 square feet.

Null hypothesis states that mean moisture content in type A, μ is equal and greater than 0.35 pounds per 100 sq feet .

Alternative hypothesis states that the mean moisture content in type A, μ is less than 0.35 pounds per 100 sq feet .

- $H0$: $\mu >= 0.35$
- HA: $\mu < 0.35$

Step 2 : Decide the significance level

Here we select $\alpha = 0.05$

Step 3: Identify the test statistic

In the given problem statement, we do not know the population standard deviation and n = 36. So we use the t distribution and the $tSTAT$ test statistic.

Step 4: Calculate the p - value and test statistic

One sample t test calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

One sample t test t statistic:	**-1.4735046253382782**
p value:	**0.07477633144907513**

Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

We have no evidence to reject the null hypothesis since one tailed p value > Level of significance

Our one-sample t-test one tailed p-value= **0.07477633144907513**

The test concludes that type A sample shingles have mean moisture content equal or greater than 0.35 pounds per 100 sq. feet.

3.1. Testing whether type B sample shingles mean moisture content are within permissible limits.

Step 1 : Define null and alternative hypothesis

In testing the mean moisture content in type B is less than 0.35 pound per 100 square feet.

Null hypothesis states that mean moisture content in type B, μ is equal and greater than 0.35 pounds per 100 sq feet .

Alternative hypothesis states that the mean moisture content in type B, μ is less than 0.35 pounds per 100 sq feet .

- $H0$: μ >= 0.35
- HA: μ < 0.35

Step 2 : Decide the significance level

Here we select $\alpha = 0.05$

Step 3: Identify the test statistic

In the given problem statement, we do not know the population standard deviation and n = 36. So we use the t distribution and the $tSTAT$ test statistic.

Step 4: Calculate the p - value and test statistic

One sample t test calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and the two-tailed p value.

One sample t test t statistic: -3.1003313069986995 p value: 0.0020904774003191826

Step 5: Decide to reject or accept null hypothesis

Level of significance: 0.05

We have evidence to reject the null hypothesis since one tailed p value < Level of significance

Our one-sample t-test one tailed p-value= 0.0020904774003191826

The test concludes that type B sample shingles have mean moisture content less than 0.35 pounds per 100 sq. feet.

3.2. Hypothesis Testing whether type A and type B shingles population mean moisture content are equal.

The test concludes that the type A and type B shingles have equal population mean moisture content.

Step 1: Define null and alternate hypothesis

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

- $H0$: $\mu A - \mu B = 0$ i.e $\mu A = \mu B$
- HA: $\mu A - \mu B \neq\neq 0$ i.e $\mu A \neq\neq \mu B$

Step 2: Decide the significance level

Here we select $\alpha = 0.05$ and the population standard deviation is not known.

Step 3: Identify the test statistic

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are different.
- The sample is a large sample, n > 30. So you use the t distribution and the $tSTAT$ test statistic for two sample unpaired test.

Step 4: Calculate the p-value and test statistic

This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values.

This test assumes that the populations have identical variances.

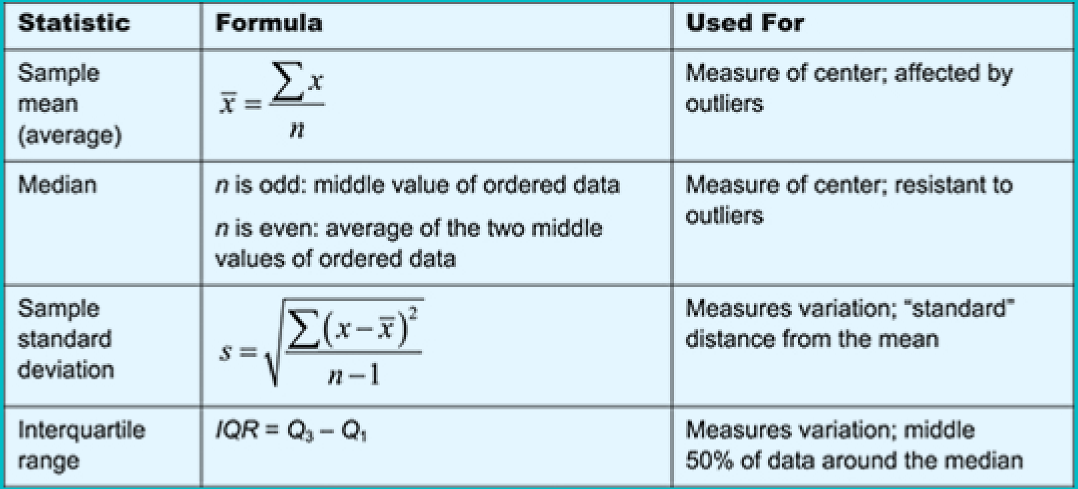
T statistic	**1.2896282719661123**
P Value	**0.2017496571835306**

Step 5: Decide to reject or accept null hypothesis

two-sample t-test p-value= 0.2017496571835306

We do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis

Basic statistical formulas used:



Statistic	Formula	Used For
Sample mean (average)	$\bar{x} = \frac{\sum x}{n}$	Measure of center; affected by outliers
Median	*n* is odd: middle value of ordered data *n* is even: average of the two middle values of ordered data	Measure of center; resistant to outliers
Sample standard deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	Measures variation; "standard" distance from the mean
Interquartile range	$IQR = Q_3 - Q_1$	Measures variation; middle 50% of data around the median

Thanks!!

Submitted by Reji Thankachan Oomman