CHETAN DUDHANE

Registered Mail Id - chetan.dudhane@gmail.com

Submitted on - 13-Spe-2020

SMDM PROJECT REPORT

Introduction

- This report consists of Analysis of 3 Problem Statements
    - Problem 1 - Wholesale Customer Analysis
    - Problem 2 - Clear Mountain State University (CMSU) Survey
    - Problem 3 - A & B Shingles Moisture Content
- Please find the Jupyter Code Notebook in the Google Drive link below. Analysis code is in Python. Datasets used are in the same directory. - https://bit.ly/3idodQc

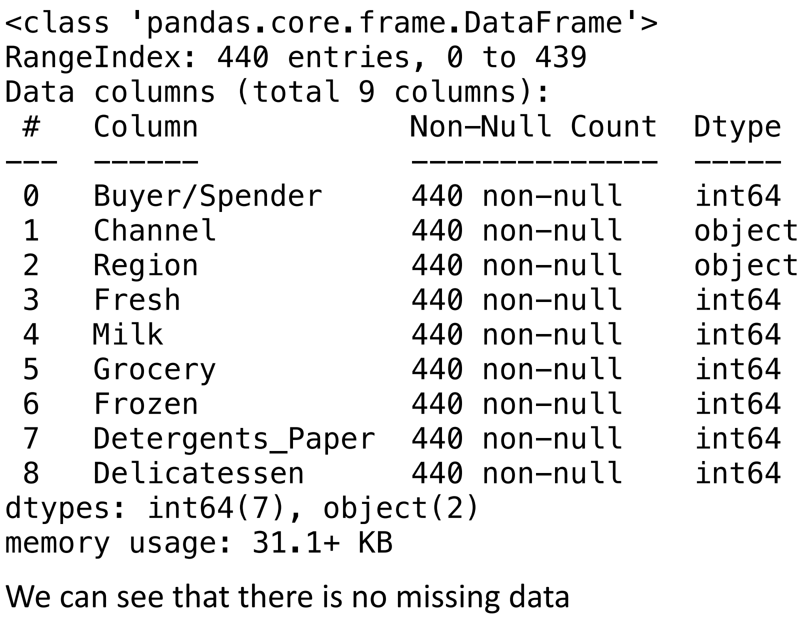
Problem 1 - Wholesale Customer Analysis

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.A Exploratory Analysis

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
6	Retail	Other	9413	8259	5126	666	1795	1451
7	Retail	Other	12126	3199	6975	480	3140	545
8	Retail	Other	7579	4956	9426	1669	3321	2566
9	Hotel	Other	5963	3648	6192	425	1716	750
10	Retail	Other	6006	11093	18881	1159	7425	2098

Table 1.1 : Wholesale Customer Analysis - First 10 rows



```
<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
```

We can see that there is no missing data

Table 1.2 : Summary Info of the whole Data

1.B Basic Understanding of Data Exploration

1. There is NO missing data in the dataset
2. Total number of Large Retailers (Total Records) = 440
3. There are 2 Channels of Distribution - RETAIL AND HOTEL
4. Regions under consideration are - LISBON, OPORTO and OTHER
5. Product categories are - FRESH, MILK, FROZEN, GROCERY,

   DETERGENTS_PAPER and DELICATESSEN

1.C Descriptive Data Analysis

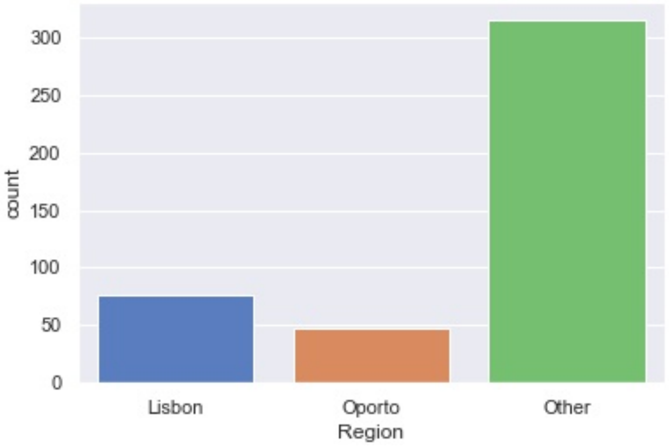
1. The distribution of retailers

   according to Region is-

   Lisbon - 77

   Oporto - 47

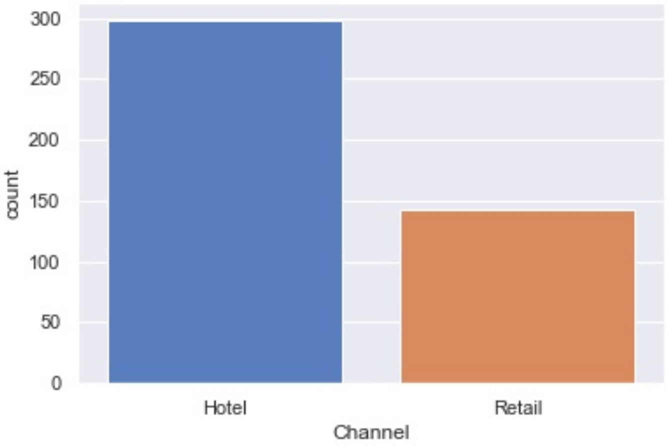
   Other - 316



2. The distribution of retailers according to Channel is-

Hotel - 298

Retail - 142



	count	unique	top	freq	mean	std	min	0.25	0.50	0.75	max
Buyer/ Spender	440				221	127	1	111	221	330	440
Channel	440	2	Hotel	298							
Region	440	3	Other	316							
Fresh	440				12000	12647	3	3128	8504	16934	112151
Milk	440				5796	7380	55	1533	3627	7190	73498
Grocery	440				7951	9503	3	2153	4756	10656	92780
Frozen	440				3072	4855	25	742	1526	3554	60869
Detergents_Paper	440				2881	4768	3	257	817	3922	40827
Delicatessen	440				1525	2820	3	408	966	1820	47943

Table 1.3 : Descriptive Statistics of Wholesale Customer Analysis

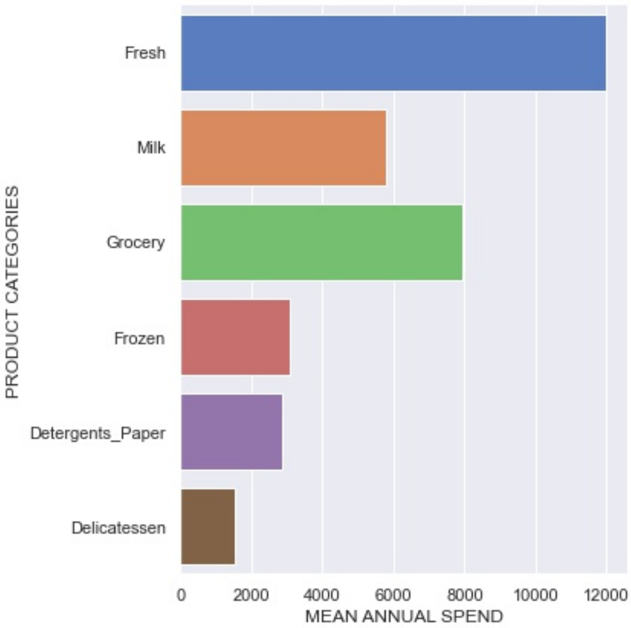
3. Top 3 Product Categories according to their Total Annual Spends (across all regions) is-

Fresh - 5,280,131

Grocery - 3,498,562

Milk - 2,550,357

These top 3 amount to about 77% of Total Annual spends.



4. The Mean and Median Annual Spends (rounded) per product is-

Fresh ---> Mean = 12,000, Median = 8,504 [Right Skewed]

Milk ---> Mean = 5,796, Median = 3,627 [Right Skewed]

Grocery ---> Mean = 7,951, Median = 4,756 [Right Skewed]

Frozen ---> Mean = 3,072, Median = 1,526 [Right Skewed]

Detergents_Paper ---> Mean = 2,881, Median = 816 [Right Skewed]

Delicatessen ---> Mean = 1,525, Median = 966 [Right Skewed]

5. Product Categories - Frozen, Detergents_Paper and Delicatessen have very low contribution to the whole distribution, only about 23%

Q 1.1 Use methods of descriptive statistics to summarise data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

- Table 3 given above (Pg 3) lists the Descriptive Statistics and Summary of the Data
- Region - wise Total Annual Spends is

  Lisbon - 2,386,813
  Oporto - 1,555,088
  Other - 10,677,599

As is evident from the Figure 1.1.a,

Region - **OTHER** spends the **most**

Region - **OPORTO** spends the **least**

- Channel - wise Total Annual Spends is

Hotel - 7,999,569

Retail - 6,619,931

As is evident from the Figure 1.1.b,

Channel - HOTEL spends the most.

Channel - RETAIL spends the least

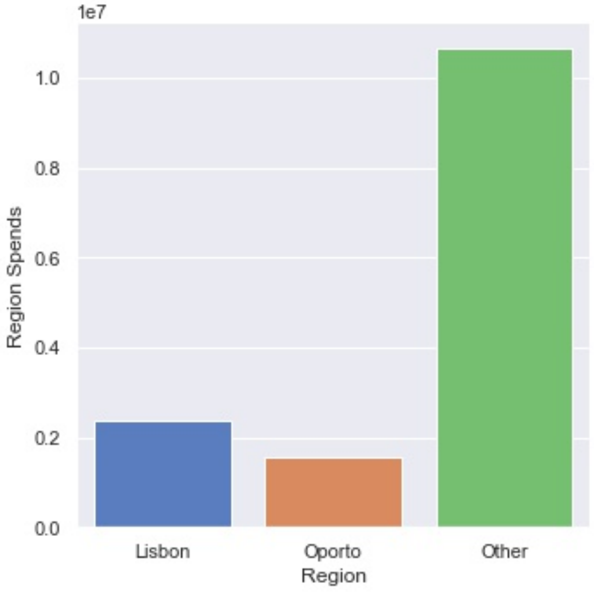


Fig 1.1.a

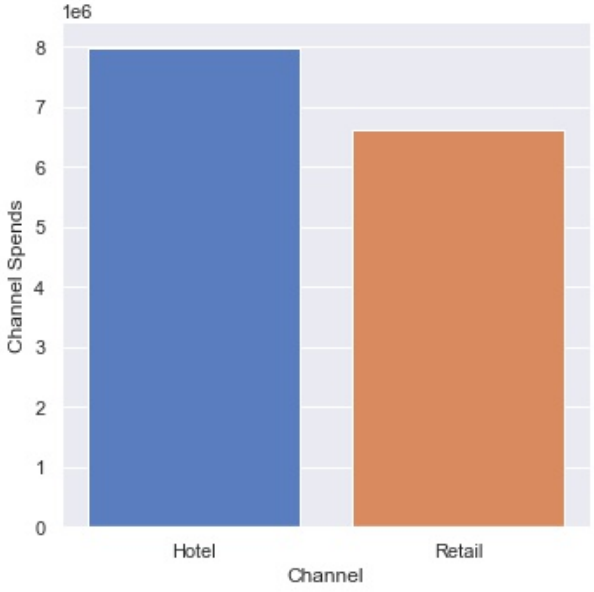


Fig 1.1.b

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

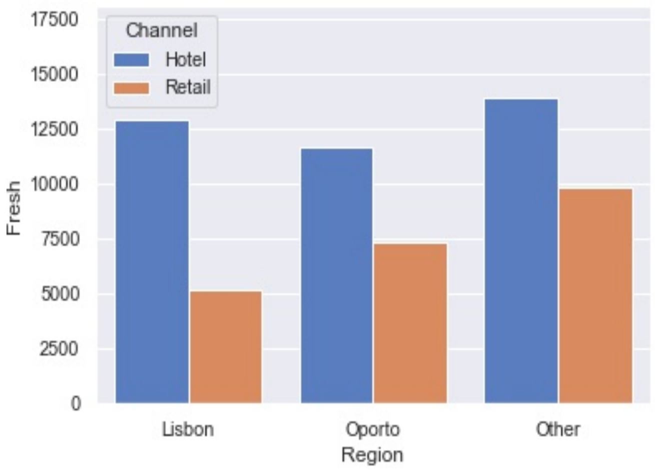


Fig 1.2.a

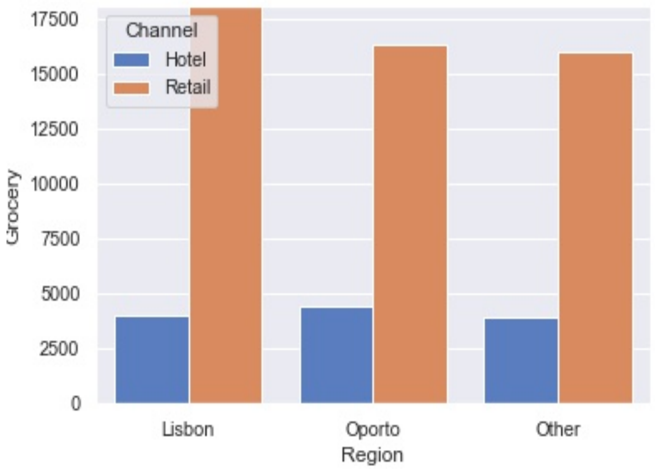


Fig. 1.2.b

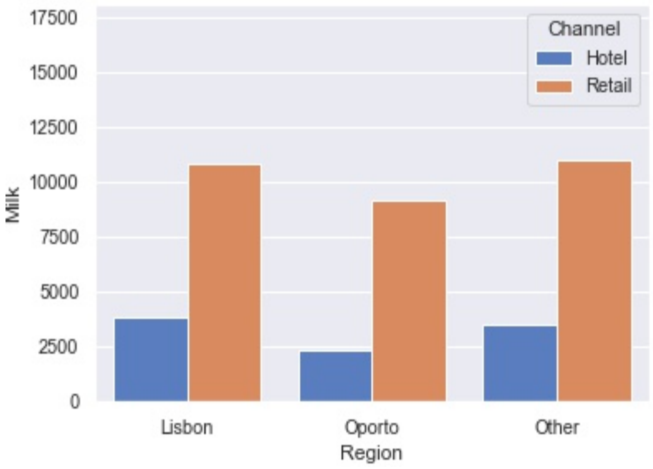


Fig 1.2.c

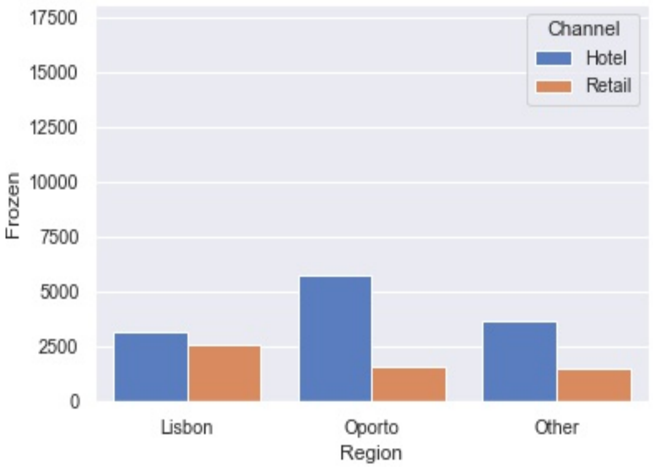


Fig. 1.2.d

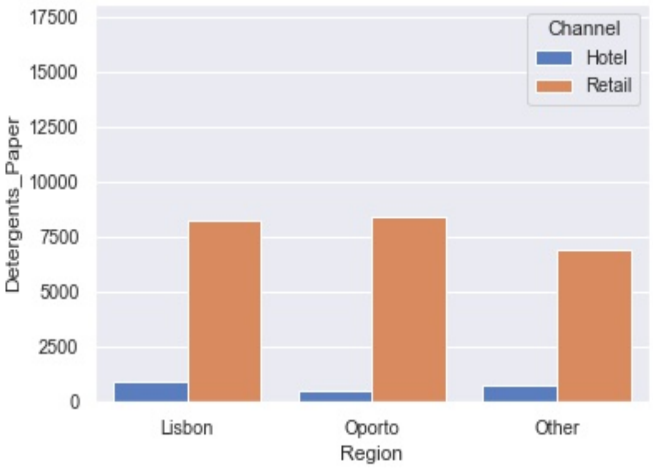


Fig 1.2.e

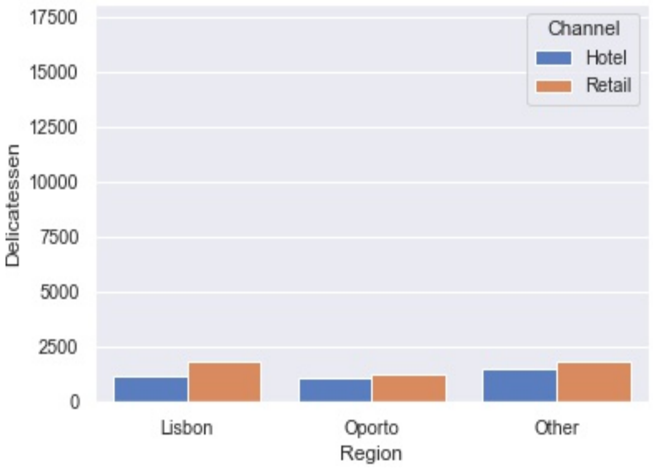


Fig. 1.2.f

- The above figures have standardised y-scale and shows mean spends of all items across Regions and Channels
- Comparing the graphs, it can be seen that almost all items leaving Delicatessen show large variation in their spends across Regions and Channels.
- Lets analyse item-wise :
    - FRESH (Fig. 1.2.a) -
        - Across all regions, **Hotels** spend the most on Fresh
        - Amongst Lisbon wholesalers, Hotels spend about 2.5 times that of Retail, while in Oporto and Other regions it is 1.5 times
    - GROCERY (Fig 1.2.b) -
        - Across all regions, **Retail** spends the most on Grocery

        - Across all regions, Retail spends more than 4 times that of Hotel on this

    - MILK (Fig 1.2.c)
        - Across all regions, **Retail** spends the most on Milk
        - In Oporto, Retail spends about 4 times that of Hotels, while in Lisbon and Other regions, it is about 3 times
    - FROZEN (Fig 1.2.d)
        - Across all regions, **Hotels** spend the most on Frozen
        - Mainly in Oporto, Hotels spend 3.7 times that of Retail on this.
    - DETERGENTS_PAPER (Fig 1.2.e)
        - Across all regions, **Retail** spend the most on Detergents_Paper
        - Across all regions, we see a very large variation in spends here.
        - Oporto Retail spends 10 times than that of Hotels while, while Lisbon and Other region Retail spends 8+ times than that of Hotels
    - DELICATESSEN (Fig 1.2.f)
        - Across all regions, **Retail** spends the most on Delicatessen, but the difference is marginal

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

- Descriptive measure of variability is Coefficient of Variation
- Statistical formulae for Coefficient of Variation is given as follows -

$$C.V = \frac{StandardDeviation}{Mean}$$

- Coefficient of Variation of all items is -
    - FRESH, CV = 1.0539179237473149
    - GROCERY, CV = 1.1951743730016824
    - MILK, CV = 1.2732985840065414
    - FROZEN, CV = 1.5803323836352914

  - DETERGENTS_PAPER, CV = 1.6546471385005155
  - DELICATESSEN, CV = 1.8494068981158382

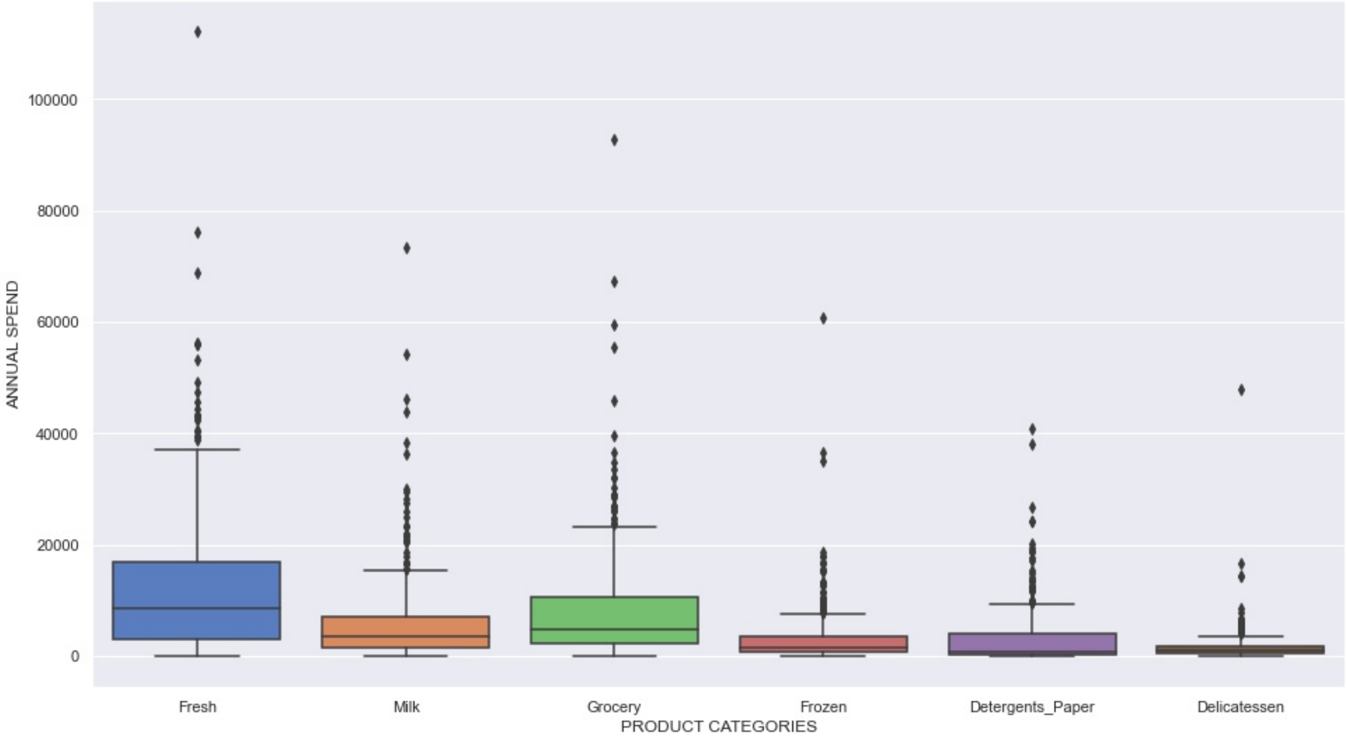
- Max CV = 1.85 Delicatessen

  Min CV = 1.05 Fresh

- **MOST INCONSISTENT** BEHAVIOUR IS OF ITEM **DELICATESSEN**

  LEAST INCONSISTENT BEHAVIOUR IS OF ITEM **FRESH**

Q 1.4 Are there any outliers in the data?



- As we can see, ALL items in the data have Outliers
- All items have Outliers on the MAX side
- Distribution of All items is RIGHT SKEWED

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

- FRESH, GROCERY and MILK - These items account for 77% of annual spends of the Wholesale Distributor Company across Portugal.
- These 3 items should be cross-checked against its Sales and should be checked for wastage to maximise profits

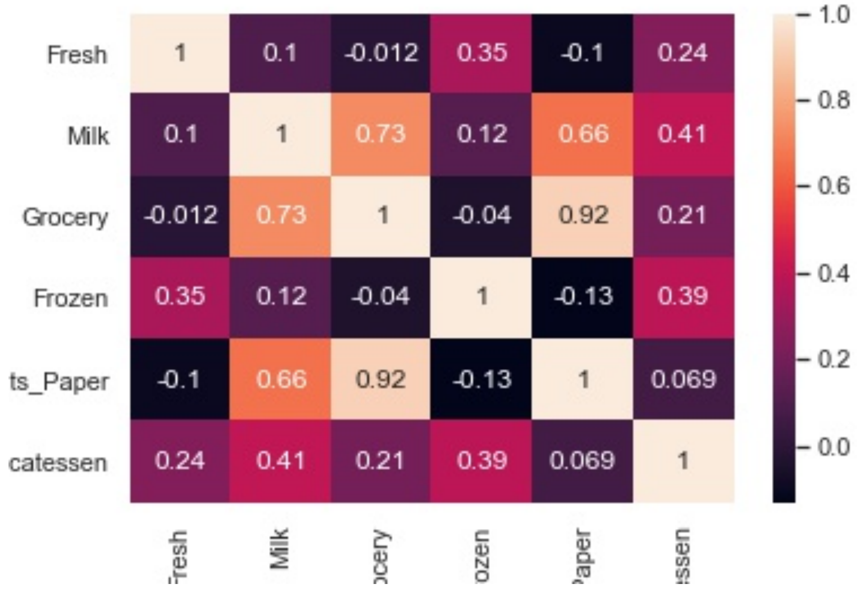
- The following Buyer/Spenders fall in the median range of 27000 to 28000 of Total Annual Spends done.

These are the fence sitters, they should be sufficiently incentivised to spend and procure more. And eventually, sell more

Buyer/Spender	Channel	Region	Total Spend by Buyer/ Spender
215	Retail	Lisbon	27938
165	Retail	Other	27863
84	Hotel	Other	27862
299	Retail	Oporto	27792
17	Retail	Other	27679
388	Hotel	Other	27559
33	Hotel	Other	27425
342	Retail	Other	27408
4	Hotel	Other	27381
107	Retail	Other	27289
183	Hotel	Other	27251

Table 1.4 : Buyer/Spenders falling in the median range of 27000 and 28000 of Total Spends

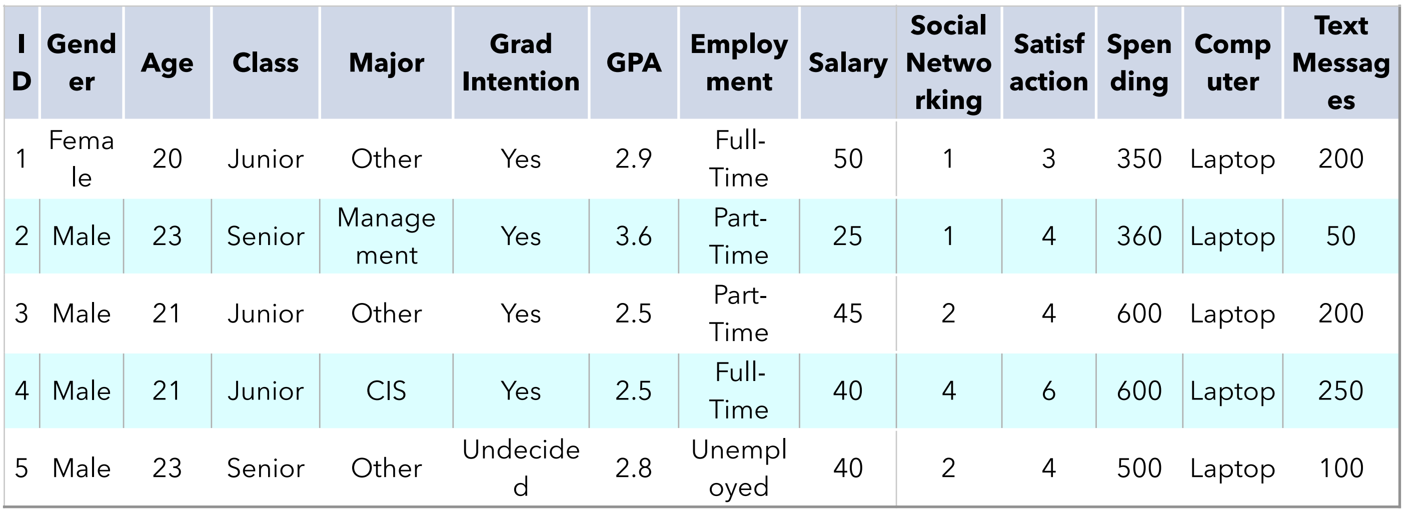
- High Correlation in Spends between Detergents_Paper and Grocery, corr = 0.92
- Also, Good Correlation between Grocery and Milk, corr = 0.73
- These correlation should be considered for the Spends and Stock forecasting for these items together.



Problem 2 : Clear Mountain State University (CMSU) Survey

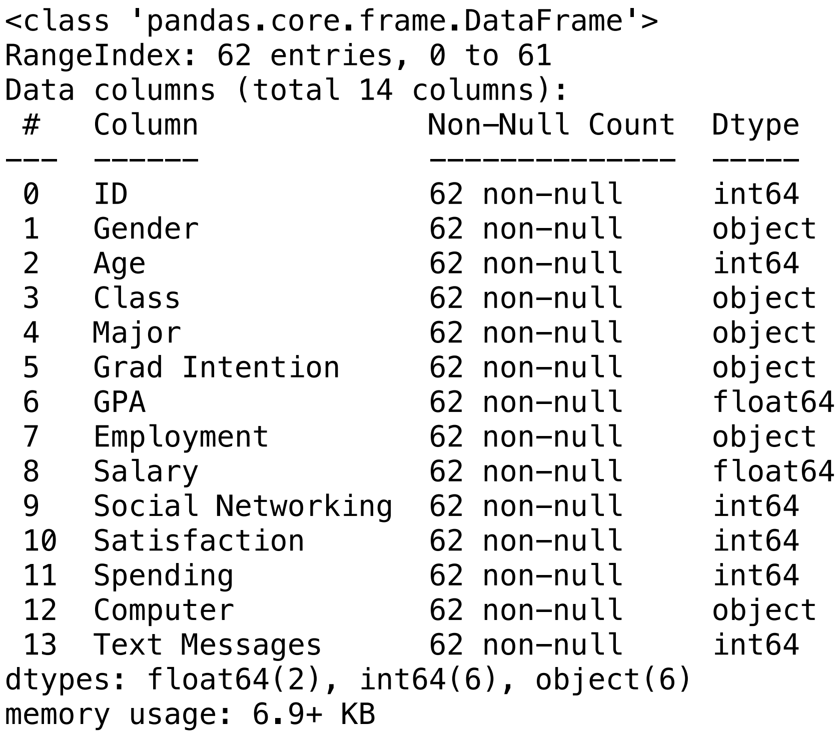
The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the Survey data set).

2.A Exploratory Analysis



ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
1	Female	20	Junior	Other	Yes	2.9	Full-Time	50	1	3	350	Laptop	200
2	Male	23	Senior	Management	Yes	3.6	Part-Time	25	1	4	360	Laptop	50
3	Male	21	Junior	Other	Yes	2.5	Part-Time	45	2	4	600	Laptop	200
4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40	4	6	600	Laptop	250
5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40	2	4	500	Laptop	100

Table 2.1 : First 5 rows of CMSU Survey data



```
<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
```

Table 2.2 : Summary Info of whole data

2.B Basic Understanding of Data Exploration

1. Total number of Undergrads surveyed = 62

2. Total number of Questions (Variables) asked per candidate = 14

   a. Categorical Variables = 6 (Gender, Class, Major, Grad Intention, Employment, Computer)

   b. Continuous Variables = 8

      - Data Type Integer = 6 (ID, Age, Social Networking, Satisfaction, Spending, Text Messages)

      - Data Type Float = 2 (GPA, Salary)

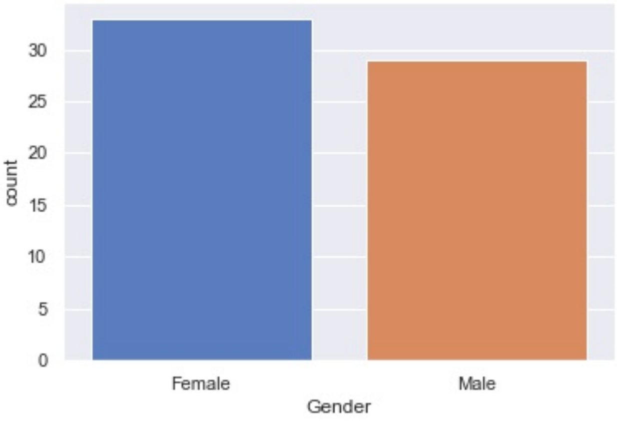
2.C Descriptive Data Analysis

	count	unique	top	freq	mean	std	min	0.25	0.50	0.75	max
ID	62				32	18	1	16	32	47	62
Gender	62	2	Female	33							
Age	62				21	1	18	20	21	22	26
Class	62	3	Senior	31							
Major	62	8	Retailing/Marketing	14							
Grad Intention	62	3	Yes	28							
GPA	62				3	0	2	3	3	3	4
Employment	62	3	Part-Time	43							
Salary	62				49	12	25	40	50	55	80
Social Networking	62				2	1	0	1	1	2	4
Satisfaction	62				4	1	1	3	4	4	6
Spending	62				482	222	100	313	500	600	1400
Computer	62	3	Laptop	55							
Text Messages	62				246	214	0	100	200	300	900

Table 2.3 : Descriptive Statistics of CMSU Survey data

1. Gender wise Distribution -

- Female - 33
- Male - 29



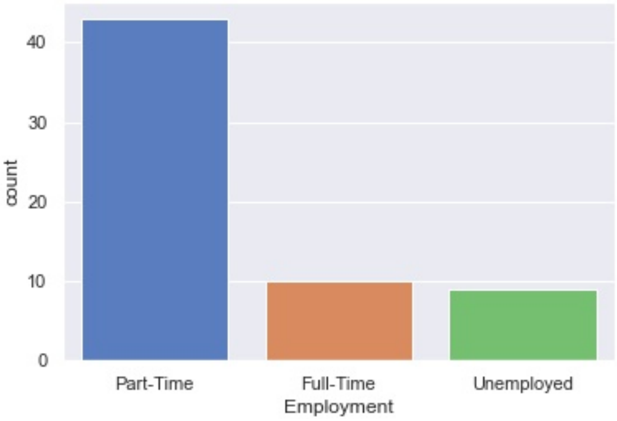
2. Mean Age - 21

3. Maximum are good in studies and have good scores -

- **Mean GPA - 3.13**

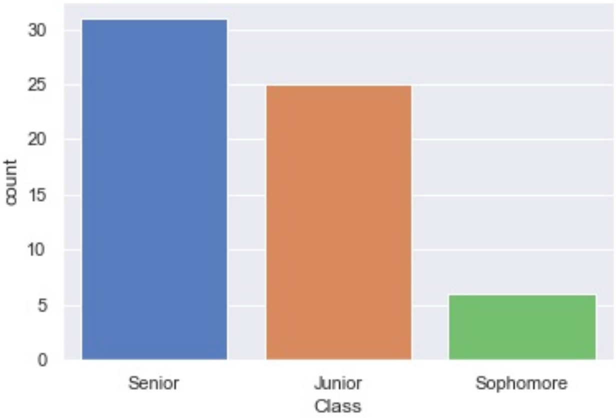
4. Maximum are Part-Time employed -

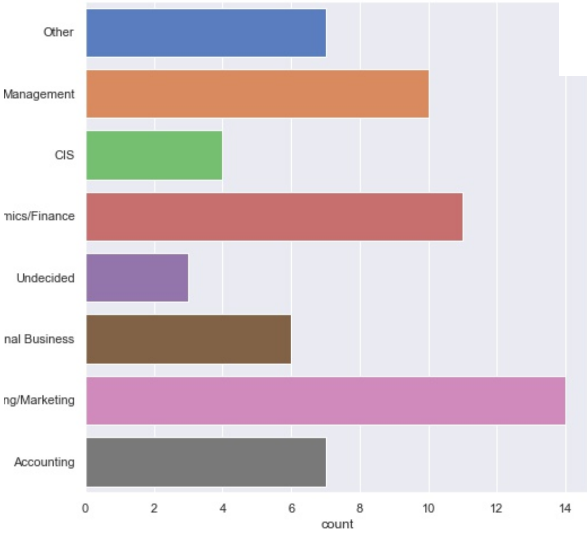
- Part-Time - 43
- Full-Time - 10
- Unemployed - 9



5. Class wise Distribution -

- Senior - 31
- Junior - 25
- Sophomore - 6





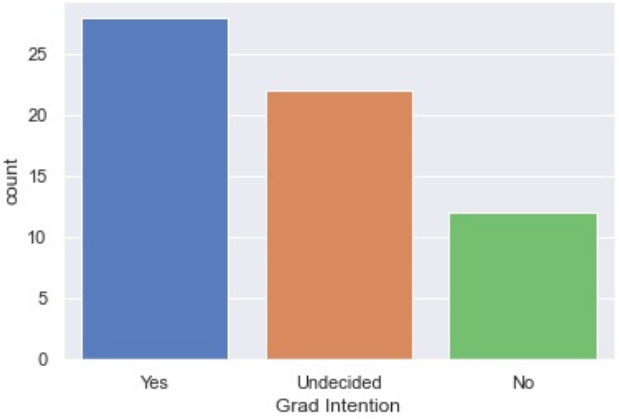
6. Study Major wise Distribution (Top 3) -

- Retailing/Marketing - 14
- Economics.Finance - 11
- Management - 10

8. Majority would like to pursue Graduation -

(Undecided participants should be targeted by CMSU Marketing to convince them to pursue Grad)

- Pursue Grad - 28
- Undecided - 22
- Do not want to pursue Grad - 12



9. Maximum have very low Social Networking skills/presence

(CMSU cannot target communication over social media - Needs to be over mails or messages)

- **Mean Social Networking rating - 1**

10. Very high Satisfaction rating amongst the Undergrad Participants

- **Mean Satisfaction rating - 4**

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

Q 2.1.1. Gender and Major

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

Q 2.1.2. Gender and Grad Intention

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

Q 2.1.3. Gender and Employment

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

Q 2.1.4. Gender and Computer

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

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

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

- Probability that a randomly selected student is a Male :

$$P(Male) = \frac{29}{62} \;=\; \mathbf{0.46774}$$

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

- Probability that a randomly selected student is a Female :

$$P(Female) = 1 - P(Male)$$

$$=. \mathbf{0.53225}$$

$$\text{also,}\quad P(Female) = \frac{33}{62} =. \mathbf{0.53225}$$

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

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

- If the Student is Male, then the Conditional Probability of different Majors is -
  - **Accounting**, $P(Accounting|Male) = \frac{4}{29}$ = **0.13793**
  - **CIS,** $P(CIS|Male) = \frac{1}{29}$ = **0.03448**
  - **Economics/Finance,** $P(Economics - Finance|Male) = \frac{4}{29}$ = **0.13793**
  - **International Business,** $P(InternationalBusiness|Male) = \frac{2}{29}$ = **0.06896**
  - **Management,** $P(Management|Male) = \frac{6}{29}$ = **0.20689**
  - **Other,** $P(Other|Male) = \frac{4}{29}$ = **0.13793**
  - **Retailing/Marketing,** $P(Retailing - Marketing|Male) = \frac{5}{29}$ = **0.17241**
  - **Undecided,** $P(Undecided|Male) = \frac{3}{29}$ = **0.10344**

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

- If the Student is Female, then the Conditional Probability of different Majors is -
  - **Accounting**, $P(Accounting|Female) = \frac{3}{33}$ = **0.0909**
  - **CIS,** $P(CIS|Female) = \frac{3}{33}$ = **0.0909**
  - **Economics/Finance,** $P(Economics - Finance|Female) = \frac{7}{33}$ = **0.21212**
  - **International Business,** $P(InternationalBusiness|Female) = \frac{4}{33}$ = **0.12121**

- **Management,** $P(Management | Female) = \frac{4}{33} =$ **0.12121**
- **Other,** $P(Other | Female) = \frac{3}{33} =$ **0.0909**
- **Retailing/Marketing,** $P(Retailing - Marketing | Female) = \frac{9}{33} =$ **0.27272**
- **Undecided,** $P(Undecided | Female) = \frac{0}{33} =$ **0**

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

Q 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

$= P(Male\ AND\ Intends\ to\ Grad)$

$= P(Male)\ x\ P(Intends\ to\ Grad\ |\ Male|$

$= \frac{29}{62}\ x\ \frac{17}{29}$

= **0.27419**

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

- Probability that a randomly chosen student is a Female and DOES NOT have Laptop

$= P(Female\ AND\ NO\ Laptop)$

$= P(Female)\ x\ P(No\ Laptop\ |\ Female$

$= \frac{33}{62}\ x\ \frac{4}{33}$

= **0.06451**

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

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

- Probability that a randomly chosen student is either a Male or has Full Time Employability

$$= P(Male\ OR\ FullTime\ Employability)$$

$$= P(Male)\ +\ P(FullTime\ Employability)$$

$$-\ P(Male\ AND\ FullTime\ Emplyability)$$

$$= \frac{29}{62} + \frac{10}{62} - \left(\frac{29}{62}\ x\ \frac{7}{29}\right)$$

= **0.51612**

Q 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 a randomly chosen Female student is majoring in International Business(IB) or Management (Mgt)

$$= P(Female\ majors\ in\ IB\ OR\ Mgt)$$

$$= P(IB\ |\ Female)\ +\ P(Mgt\ |\ Female)$$

$$= \frac{4}{33} + \frac{4}{33}$$

= **0.24242**

Q 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 Gender	No	Yes	Total
Female	9	11	**20**
Male	3	17	**20**
Total	**12**	**28**	**40**

- Grad Intention and Being Female is an Independent Event, if

$P(Female \bigcap GradIntent\ Yes)$ =. $P(Female)\ x\ P(GradIntent\ Yes)$

- Now, from the above Contingency table,
  - $P(Female) = \frac{20}{40} = 0.5$
  - $P(GradIntent\ Yes) = \frac{28}{40} = 0.7$
  - $P(Female \bigcap GradIntent\ Yes) =$

$$= P(Female)\ x\ P(GradIntent\ Yes \mid Female)$$

$$= \frac{20}{40}\ x\ \frac{11}{20}$$

= 0.275

- So, it can be seen that,

$$P(Female \bigcap GradIntent\ Yes) \neq P(Female)\ x\ P(GradIntent\ Yes)$$

- So, we conclude that,

Being a Female and Graduate Intentions are NOT INDEPENDENT Events

Q 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

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

Gender **GPA**	**Female**	**Male**	**Total**
Less than 3	8	9	**17**
3 or more	25	20	**45**
Total	**33**	**29**	**62**

- Probability that a randomly chosen student has GPA less than 3

$= \frac{17}{62}$ = **0.27419**

Q 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.

Gender Salary	Female	Male	Total
Less than 50	15	15	**30**
50 or more	18	14	**32**
Total	**33**	**29**	**62**

- Probability that a Male earns 50 or more

$$= P(Salary\ 50\ or\ more \mid Male)$$

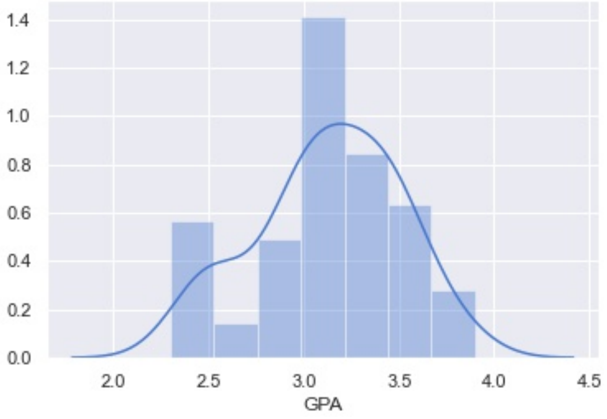
$$= \frac{14}{29} = \mathbf{0.48275}$$

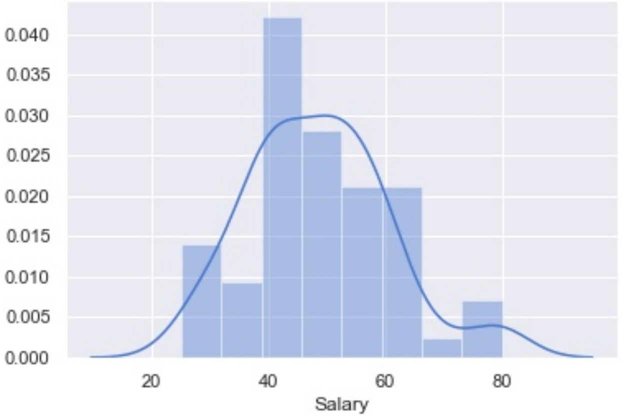
- Probability that a Female earns 50 or more

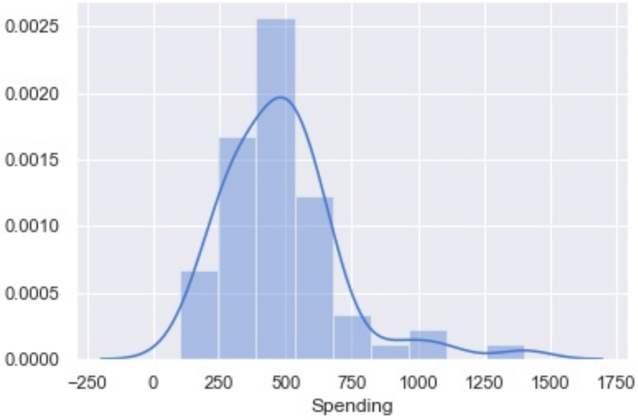
$$= P(Salary\ 50\ or\ more \mid Female)$$

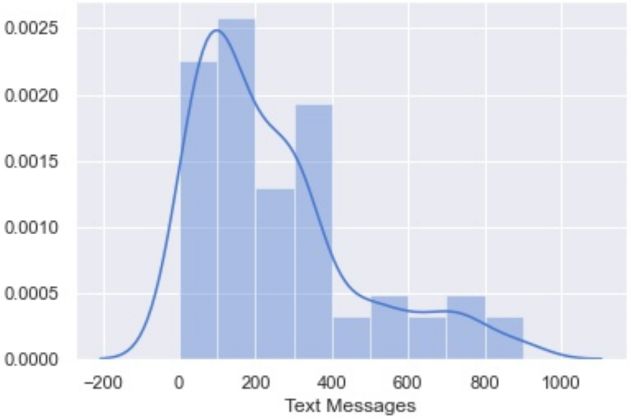
$$= \frac{18}{33} = \mathbf{0.54545}$$

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 summarising your conclusions









- GPA - As seen in the fig, the Kernel Estimation Density (KDE) curve of GPA is fairly symmetric about its mean. The curve is slightly erratic on the left side.

  We can conclude GPA quite nearly follows Normal Distribution.

- SALARY - As seen in the fig, the Kernel Estimation Density (KDE) curve of SALARY is fairly symmetric about its mean. The curve is slightly erratic on the right extreme..

  We can conclude SALARY nearly follows Normal Distribution

- SPENDING - As seen in the fig, the Kernel Estimation Density (KDE) curve of SPENDING is almost perfect symmetric about its mean. The curve is slightly right skewed

  We can conclude SPENDING follows Normal Distribution

- TEXT MESSAGES - As seen in the fig, the Kernel Estimation Density (KDE) curve of TEXT MESSAGES is quite erratic in its behaviour on the right side. Its not peaking at its mean.

  We can conclude SALARY is quite far from Normal Distribution

2.D Final Conclusion

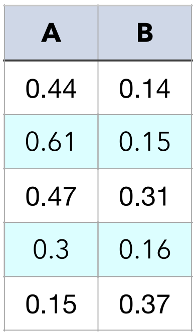
- Majority would like to pursue Graduation. Undecided should be followed up by the Marketing to convince them to pursue Graduation
- Majority of the participants are part-Time employed. It calls for needs to hold classes in the evenings or over the weekends.
- Participants show poor Social Networking Skills/presence, hence messaging by the University should be strictly over mails and phone.
- Participants show high satisfaction ratings.

Problem 3 : A & B Shingles Moisture Content

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 pound per 100 square feet.

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

3.A Exploratory Analysis



A	B
0.44	0.14
0.61	0.15
0.47	0.31
0.3	0.16
0.15	0.37

Table 3.1 : A & B Shingles
First 5 Rows

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   A       36 non-null     float64
 1   B       31 non-null     float64
dtypes: float64(2)
memory usage: 704.0 bytes
```

Table 3.2 : A & B Shingles Data Info

	count	mean	std	min	0.250	0.500	0.750	max
A	36	0.317	0.136	0.130	0.208	0.290	0.393	0.720
B	31	0.274	0.137	0.100	0.160	0.230	0.400	0.580

Table 3.3 : Descriptive Statistics of Shingles A and B

3.A Basic Understanding of Exploratory Analysis

1. There are 2 types of Shingles - A & B

2. Values in the data are the Moisture Content in the Shingles (in pounds per 100 square feet)

3. No. of entries of Shingles A - 36

   No. of entries of Shingles B - 31

4. Shingles A ———> Mean = 0.317, Median = 0.290

   Shingles B ———-> Mean = 0.274, Median = 0.230

Q 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.

- **SHINGLES A**
  - <u>Step 1 : State NULL and ALTERNATE Hypothesis</u>

    Assuming that as status quo, Shingles A do not meet the required level of moisture content and hence we declare Null and Alternate Hypothesis as follows-

    NULL HYPOTHESIS : $H_o : \mu \geq 0.35$

    ALTERNATE HYPOTHESIS : $H_a : \mu < 0.35$ (Left Tailed)

  - <u>Step 2 : Decide Level of Significance</u>

    As, its not given, assuming industry standard of 95% Confidence level. Hence,

    Level of Significance : $\alpha = 0.05$

- Step 3 : Identify the Test
    - As, Population Standard Deviation is not given/known, we cannot use Z-Test.
    - We use t-distribution test for one sample.
- Step 4 : Calculate t statistic and p value
    - $t_{stat} =$ **-1.473504**

      p value (2 Tailed) = 0.149552

      p value (1 Tailed) = **0.074776**
- Step 5 : Conclude based on the result
    - Here, p value = 0.074776

      α = 0.05

      p value > α (Level of Significance)
    - **HENCE, For SHINGLES A, WE HAVE NO EVIDENCE TO REJECT THE NULL HYPOTHESIS**
    - **HENCE, WE CONCLUDE THAT MEAN MOISTURE CONTENT IN SHINGLES A IS NOT WITHIN THE PERMISSIBLE LIMITS**

- **SHINGLES B**
    - Step 1 : State NULL and ALTERNATE Hypothesis

      Assuming that as status quo, Shingles B do not meet the required level of moisture content and hence we declare Null and Alternate Hypothesis as follows-

      NULL HYPOTHESIS : $H_o : \mu \geq 0.35$

      ALTERNATE HYPOTHESIS : $H_a : \mu < 0.35$ (Left Tailed)
    - Step 2 : Decide Level of Significance

      As, its not given, assuming industry standard of 95% Confidence level. Hence,

      Level of Significance : $\alpha = 0.05$

- Step 3 : Identify the Test
    - As, Population Standard Deviation is not given/known, we cannot use Z-Test.
    - We use t-distribution test for one sample.
- Step 4 : Calculate t statistic and p value
    - t_{stat} = **-3.100331**

      p value (2 Tailed) = 0.004180

      p value (1 Tailed) = **0.002090**
- Step 5 : Conclude based on the result
    - Here, p value = 0.002090

      α = 0.05

      p value < α (Level of Significance)
    - **HENCE, For SHINGLES B, WE HAVE EVIDENCE TO REJECT THE NULL HYPOTHESIS**
    - **HENCE, WE CONCLUDE THAT MEAN MOISTURE CONTENT IN SHINGLES B IS WITHIN THE PERMISSIBLE LIMITS**

Q 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?

- **Assumptions for Test of Equality of Means**
    - The data follows Normal Distribution
    - The variances of two samples are equal (If not, then we use Welch's t - Test)
    - For Independent t - Test, data of both samples should have no relation to each other.

- **Step 1 : State NULL and ALTERNATE Hypothesis**
  - Assuming that as status quo, Shingles A and B have the same Mean Moisture Content
  - Hence, we declare Null and Alternate Hypothesis as follows-

    NULL HYPOTHESIS : $\mu_a = \mu_b$

    ALTERNATE HYPOTHESIS : $\mu_a \neq \mu_b$ (Two Tailed)
- **Step 2 : Decide Level of Significance**
  - As, its not given, assuming industry standard of 95% Confidence level. Hence,

    Level of Significance : $\alpha = 0.05$
- **Step 3 : Decide on the Test**
  - Here, Population Standard Deviation is not given/known, we cannot use Z-Test.
  - We use the t distribution and the t_{stat} test statistic for two sample unpaired test. (Independent t-test)
  - Here, Sample sizes are not same, hence the equal variance t-statistic is no longer equal to the unequal variance t-statistic:

    We set equal_var = False
- **Step 4 : Calculate t statistic and p value**
  - $t_{stat} =$ **1.288508**

    p value (2 Tailed) = 0.202258
- **Step 5 : Conclude based on the result**
  - Here, p value > α (Level of Significance)
  - We do not have enough evidence to reject the Null Hypothesis in favour of Alternate Hypothesis
  - **HENCE, WE CONCLUDE THAT MEAN MOISTURE CONTENT IN SHINGLES A AND B IS SAME.**

3.B Final Conclusion

- **MEAN MOISTURE CONTENT IN SHINGLES A IS NOT WITHIN THE PERMISSIBLE LIMITS**
- **MEAN MOISTURE CONTENT IN SHINGLES B IS WITHIN THE PERMISSIBLE LIMITS**
- **MEAN MOISTURE CONTENT IN SHINGLES A AND B IS SAME.**