# SDNM Project Report

Prepared by - Bhanu

Index

WHOLESALE CUSTOMER ANALYSIS

Problem Statement:

A wholesale distributor operating in different regions of Portugal has information on the 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 channels (Hotel, Retail).

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

Problem understanding: We have to find the region and channel which has spent more and less. Before that, the below snippet helps us understand the data in a broader way that is mean we can find the count, unique values, frequency, mean, standard deviation max and, min and also IQR values of each variable in the dataset.

| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Buyer/Spender | 440 | NaN | NaN | NaN | 220.5 | 127.161 | 1 | 110.75 | 220.5 | 330.25 | 440 |
| 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 | NaN | NaN | NaN | 12000.3 | 12647.3 | 3 | 3127.75 | 8504 | 16933.8 | 112151 |
| Milk | 440 | NaN | NaN | NaN | 5796.27 | 7380.38 | 55 | 1533 | 3627 | 7190.25 | 73498 |
| Grocery | 440 | NaN | NaN | NaN | 7951.28 | 9503.16 | 3 | 2153 | 4755.5 | 10655.8 | 92780 |
| Frozen | 440 | NaN | NaN | NaN | 3071.93 | 4854.67 | 25 | 742.25 | 1526 | 3554.25 | 60869 |
| Detergents_Paper | 440 | NaN | NaN | NaN | 2881.49 | 4767.85 | 3 | 256.75 | 816.5 | 3922 | 40827 |
| Delicatessen | 440 | NaN | NaN | NaN | 1524.87 | 2820.11 | 3 | 408.25 | 965.5 | 1820.25 | 47943 |

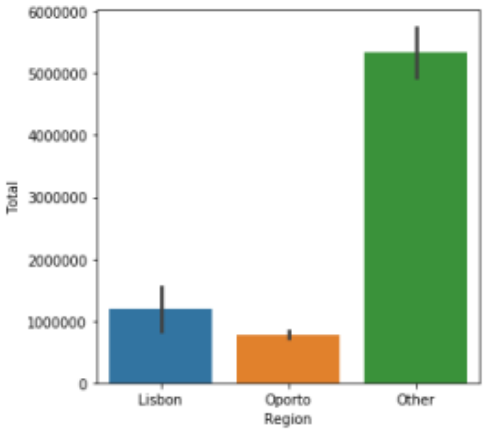
We can conclude that the dataset has

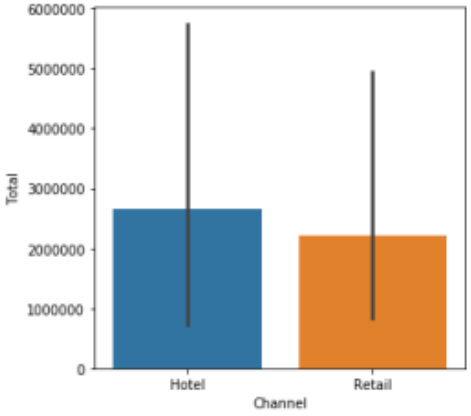
440 counts in all the variables

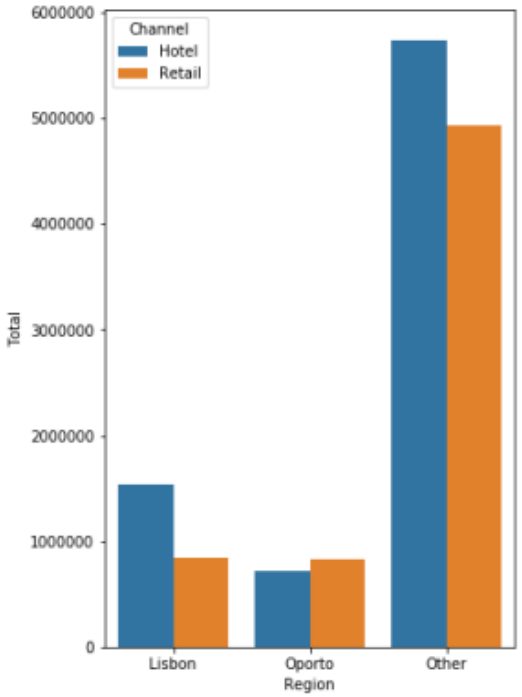
- Two unique values in the channel variable
- Three unique values in the Region variable
- Mean values of the variable are different
- The min value seems to be 3 for Fresh, Grocery, Detergents Paper, and Delicatessen
- From the IQR values we understand the range of data lies in the 25%, 50%, 75%
- The max value seems to be 112151 held by Fresh

| | Region | Channel | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicatessen | Total |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Lisbon | Hotel | 761233 | 228342 | 237542 | 184512 | 56081 | 70632 | 1538342 |
| 1 | Lisbon | Retail | 93600 | 194112 | 332495 | 46514 | 148055 | 33695 | 848471 |
| 2 | Oporto | Hotel | 326215 | 64519 | 123074 | 160861 | 13516 | 30965 | 719150 |
| 3 | Oporto | Retail | 138506 | 174625 | 310200 | 29271 | 159795 | 23541 | 835938 |
| 4 | Other | Hotel | 2928269 | 735753 | 820101 | 771606 | 165990 | 320358 | 5742077 |
| 5 | Other | Retail | 1032308 | 1153006 | 1675150 | 158886 | 724420 | 191752 | 4935522 |

From the below graph we can conclude that the Region-Other spends more Channel-Hotel spends more Region-Oporto spends less Channel –retail spends less



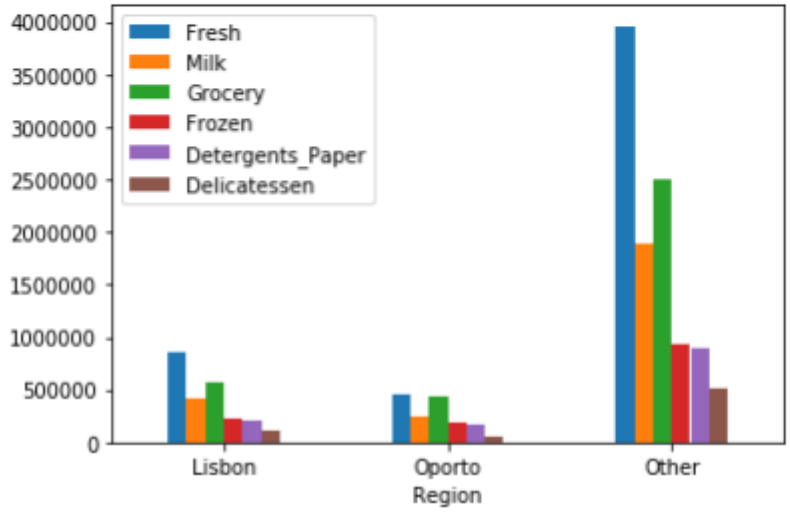

To conclude if we see, we understand in other region hotel spends more and in Oporto region retail spends less.

1.2There are 6 different varieties of items considered. Do all varieties show similar behavior across Region and channels?

Problem understanding:

We have 6 varieties so, if we see across Region we can find insights on each of 6 varieties.

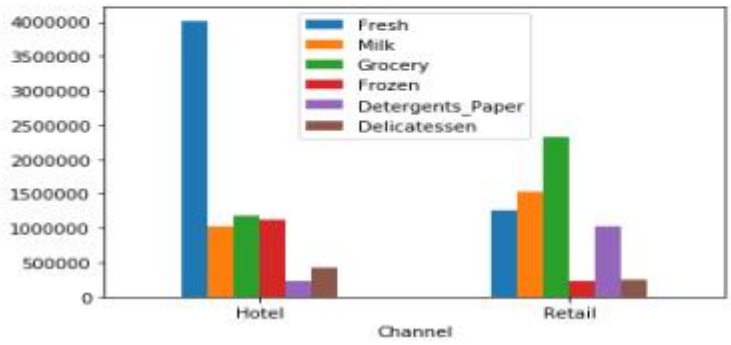
| Region | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicatessen |
|---|---|---|---|---|---|---|
| Lisbon | 854833 | 422454 | 570037 | 231026 | 204136 | 104327 |
| Oporto | 464721 | 239144 | 433274 | 190132 | 173311 | 54506 |
| Other | 3960577 | 1888759 | 2495251 | 930492 | 890410 | 512110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




In OTHER REGION we can the spending is maximum on all varieties, and in the OPORTO REGION we can find the spending is less on all varieties.

We have 6 varieties so, if we see across Channel we can find insights on each of 6 varieties.

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




Across channel if we the spending on varieties is different for each products, we can Fresh variety spends large in Hotel channel where it is less in Retail channel.

1.3based on a descriptive measure of variability, which item shows the most inconsistent behavior? Which items show the least inconsistent behavior?

Problem understanding

We can use IQR method or STD to find the MOST and LEAST inconsistent item.

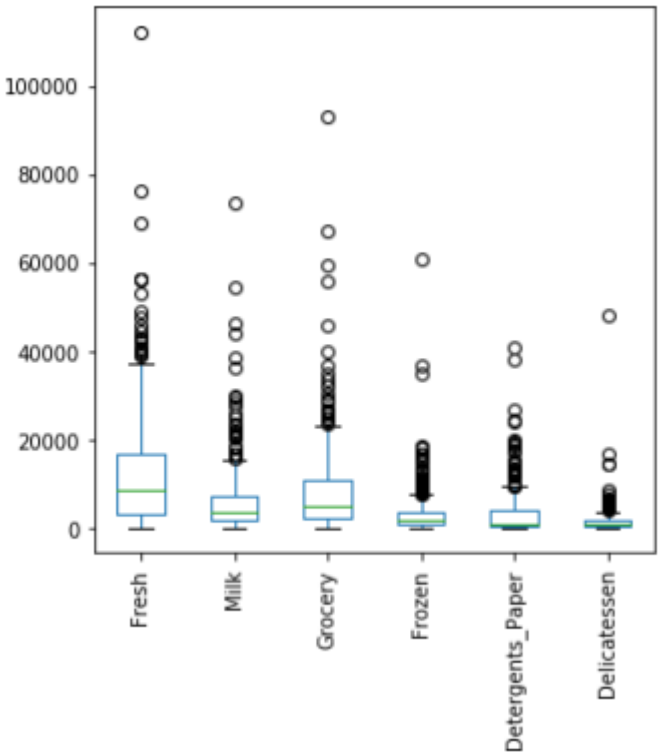
IQR for all 6 varieties

Fresh 13806.00
Milk 5657.25
Grocery 8502.75
Frozen 2812.00
Detergents Paper 3665.25
Delicatessen 1412.00

Fresh varieties seem to be the most inconsistent in terms of spending by the buyer Delicatessen varieties seem to be the least inconsistent in terms of spending by the buyer

1.4Are there any outliers in the data?

YES, all 6 varieties have outliers in the dataset. We can use IQR or box plot to find the outliers in the dataset


100000
80000
60000
40000
20000
0
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicatessen

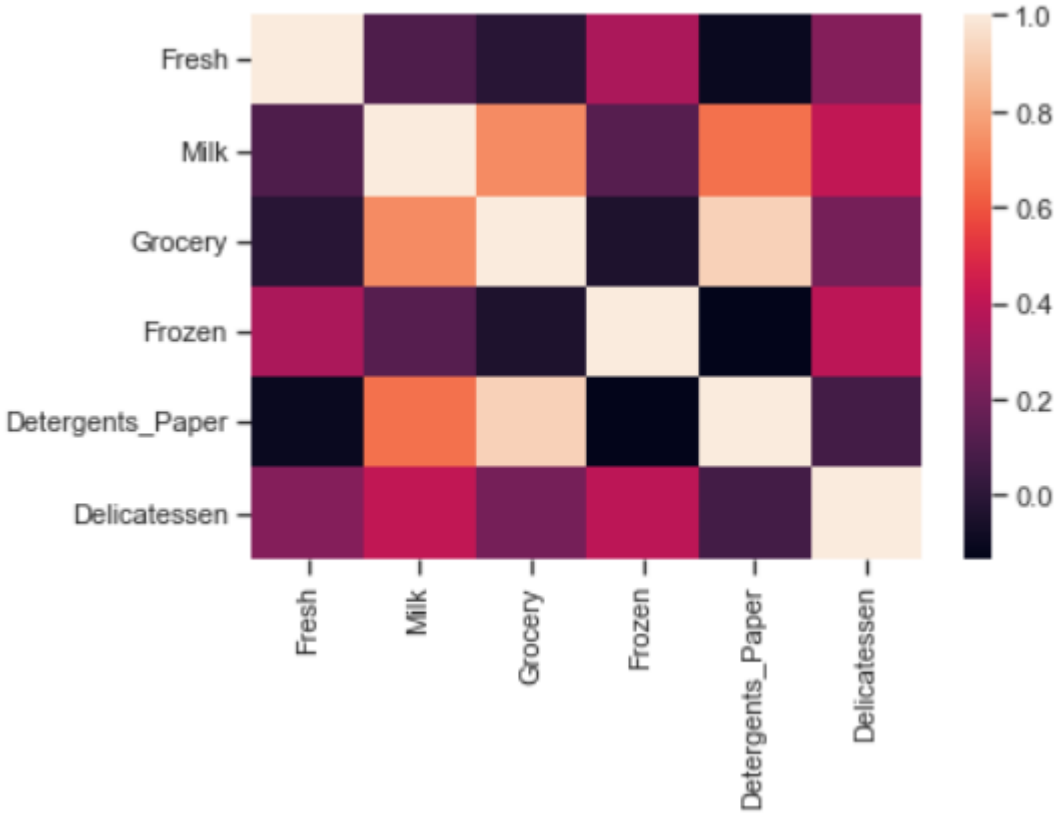
By IQR method,

| | Channel | Delicatessen | Detergents_Paper | Fresh | Frozen | Grocery | Milk | Region |
|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | False |
| 1 | False | False | False | False | False | False | False | False |
| 2 | False | True | False | False | False | False | False | False |
| 3 | False | False | False | False | False | False | False | False |
| 4 | False | True | False | False | False | False | False | False |

1.5Based on this report, what are the recommendations?

Based on the analysis done so far it is evident that there are many buyers in another region so we could improve our retailer base in Lisbon and Oporto region as well We can find evidence Fresh items are considerably used by many retailers among all the region. We find that varieties like Frozen, Detergents paper, and Delicatessen are not Popular among the retailers so we can try to maximize these products where the demand Is more. As of

now,there are only two modes of sales either Retail or Hotel. Considering the base of the region we can use multiple methods had to reach our customers



PROBLEM 2 – 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).

EDA: In basic EDA we could understand more about the data such as

SHAPE-(62, 14)

HEAD-We can use this to understand the first five rows and columns in the dataset

INFO-we have 62 entries and 13 column and we don't have any Null Values

We have, GPA and SALARY as float values ID, AGE, SOCIAL NETWORKING, SATISFACTION, SPENDING, TEXT MESSAGES as integer values

GENDER, CLASS, MAJOR, GRAD INTENTION, EMPLOYMENT AND COMPUTER as object values NULL VALUES = 0

```
ID                0
Gender            0
Age               0
Class             0
Major             0
Grad Intention    0
GPA               0
Employment        0
Salary            0
Social Networking 0
Satisfaction      0
Spending          0
Computer          0
Text Messages     0
dtype: int64
```

DESCRIPTIVE STATICS OF THE DATASET

| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ID** | 62 | NaN | NaN | NaN | 31.5 | 18.0416 | 1 | 16.25 | 31.5 | 46.75 | 62 |
| **Gender** | 62 | 2 | Female | 33 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Age** | 62 | NaN | NaN | NaN | 21.129 | 1.43131 | 18 | 20 | 21 | 22 | 26 |
| **Class** | 62 | 3 | Senior | 31 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Major** | 62 | 8 | Retailing/Marketing | 14 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Grad Intention** | 62 | 3 | Yes | 28 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **GPA** | 62 | NaN | NaN | NaN | 3.12903 | 0.377388 | 2.3 | 2.9 | 3.15 | 3.4 | 3.9 |
| **Employment** | 62 | 3 | Part-Time | 43 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Salary** | 62 | NaN | NaN | NaN | 48.5484 | 12.0809 | 25 | 40 | 50 | 55 | 80 |
| **Social Networking** | 62 | NaN | NaN | NaN | 1.51613 | 0.844305 | 0 | 1 | 1 | 2 | 4 |
| **Satisfaction** | 62 | NaN | NaN | NaN | 3.74194 | 1.21379 | 1 | 3 | 4 | 4 | 6 |
| **Spending** | 62 | NaN | NaN | NaN | 482.016 | 221.954 | 100 | 312.5 | 500 | 600 | 1400 |
| **Computer** | 62 | 3 | Laptop | 55 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Text Messages** | 62 | NaN | NaN | NaN | 246.21 | 214.466 | 0 | 100 | 200 | 300 | 900 |

We have unique values in Gender, class, Major, Grad Intention, Employment, and Computer From descriptive statistics or five-point summary,

- No of the female is 33
- No of the male is 29
- The max-age of students is 26 and the median age is 21
- Class we have 3 unique values senior, junior, sophomore
- In majors we have 8 majors
- Students have scored a 3.9 GPA but the median remains around to be 3.15
- Retailing/marketing is the most preferred Major by students
- 28/62 have grad intent
- Part-time seems to be more when compared to full-time job
- 55 have laptops for the education

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

2.1.1. Gender and Major

2.1.2. Gender and Grad Intention

2.1.3. Gender and Employment

2.1.4. Gender and Computer

2.1.1. Gender and Major

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

2.1.2. Gender and Grad Intention

| Grad Intention<br>Gender | No | Undecided | Yes |
|---|---|---|---|
| Female | 9 | 13 | 11 |
| Male | 3 | 9 | 17 |

2.1.3. Gender and Employment

| Employment<br>Gender | Full-Time | Part-Time | Unemployed |
|---|---|---|---|
| Female | 3 | 24 | 6 |
| Male | 7 | 19 | 3 |

2.1.4. Gender and Computer

| Computer<br>Gender | Desktop | Laptop | Tablet |
|---|---|---|---|
| Female | 2 | 29 | 2 |
| Male | 3 | 26 | 0 |

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?

Number of male (A) = 29

Total Number of students (B) = 62

P (A/B) =29/62

The probability that a randomly selected CMSU student will be male is 46.77419354 8387096 %

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

Number of female (A) = 33

Total Number of students (B) = 62

P (A/B) =33/62

The probability that a randomly selected CMSU student will be female is 53.2258064 516129 %

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.

The conditional probability of different Majors P (Different Majors/ Male)

The snippet shows the probability of male choosing different majors

| Major<br>Gender | Accounting | CIS | Economics/Finance | International Business | Management | Other | Retailing/Marketing | Undecided |
|---|---|---|---|---|---|---|---|---|
| Female | 0.090909 | 0.090909 | 0.212121 | 0.121212 | 0.121212 | 0.090909 | 0.272727 | 0.000000 |
| Male | 0.137931 | 0.034483 | 0.137931 | 0.068966 | 0.206897 | 0.137931 | 0.172414 | 0.103448 |
| All | 0.112903 | 0.064516 | 0.177419 | 0.096774 | 0.161290 | 0.112903 | 0.225806 | 0.048387 |

2.3.2 Find the conditional probability of different majors among the female students of CMSU. P (Conditional Majors/ Female)

The snippet shows the probability of female choosing different majors

| Major<br>Gender | Accounting | CIS | Economics/Finance | International Business | Management | Other | Retailing/Marketing | Undecided |
|---|---|---|---|---|---|---|---|---|
| Female | 0.090909 | 0.090909 | 0.212121 | 0.121212 | 0.121212 | 0.090909 | 0.272727 | 0.000000 |
| Male | 0.137931 | 0.034483 | 0.137931 | 0.068966 | 0.206897 | 0.137931 | 0.172414 | 0.103448 |
| All | 0.112903 | 0.064516 | 0.177419 | 0.096774 | 0.161290 | 0.112903 | 0.225806 | 0.048387 |

2.4. Assume that the sample is 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.

P (Grad Intent Yes/ Male) = 17/29

| Grad Intention | No | Undecided | Yes |
|---|---|---|---|
| Gender | | | |
| Female | 0.272727 | 0.393939 | 0.333333 |
| Male | 0.103448 | 0.310345 | 0.586207 |
| All | 0.193548 | 0.354839 | 0.451613 |

The probability that a randomly chosen student is a male and intends to graduate is 58.62%

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

P (Have a laptop/ female) = 29/33

P (does not have a laptop/ female) = 1- P (Have a laptop/ female) = 1-0.88=12%

| Computer | Desktop | Laptop | Tablet |
|---|---|---|---|
| Gender | | | |
| Female | 0.060606 | 0.878788 | 0.060606 |
| Male | 0.103448 | 0.896552 | 0.000000 |
| All | 0.080645 | 0.887097 | 0.032258 |

The probability that a randomly selected student is a female and does NOT have a laptop is 12%

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 either a male or has full-time employment?

Probability of randomly selected student is male P (A) = 46.77%

Probability of randomly selected student has a fulltime job P (B) = 16.13%

Probability of male having a fulltime job P (A and B) = 11.29%

P = p_of_male_stu+p_of_fulltime_emp-p_of_male_fulltime_emp = 51.61%

The probability that a randomly chosen student is either a male or has full-time employment 51.61290322580645 %

The probability that a randomly selected student is a female and does not have laptop is 1-0.88

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 a female student is randomly chosen, she is majoring in international business or management 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 | All |
|---|---|---|---|
| Gender | | | |
| Female | 9 | 11 | 20 |
| Male | 3 | 17 | 20 |
| All | 12 | 28 | 40 |

| Grad Intention | No | Yes |
|---|---|---|
| Gender | | |
| Female | 0.45 | 0.55 |
| Male | 0.15 | 0.85 |
| All | 0.30 | 0.70 |

CONCLUSION: The probability that a randomly selected Student is Female 50.0 The probability that a randomly selected student is female and intends to graduate 55.0 % They 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

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

The probability that his/her GPA is less than 3 is 27.419354838709676 %

2.6.2. 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.

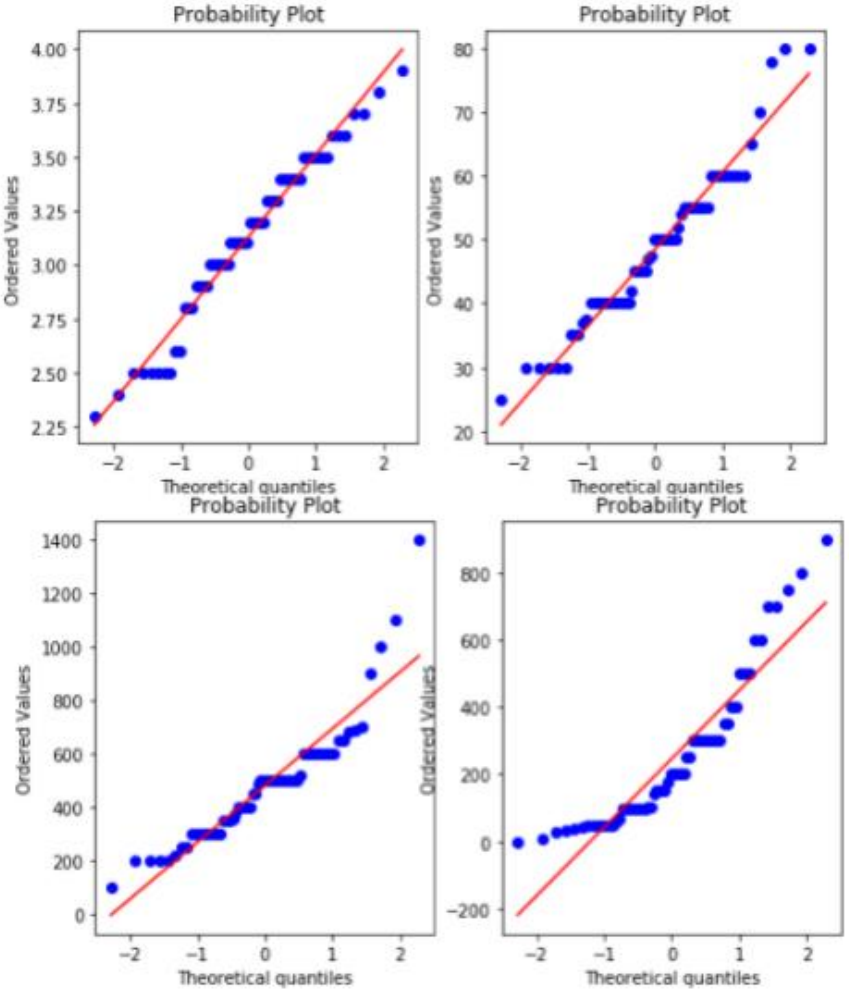
| Salary | False | True |
|---|---|---|
| Gender | | |
| False | 0.454545 | 0.545455 |
| True | 0.517241 | 0.482759 |

The probability that a randomly selected male earns 50 or more is 48%

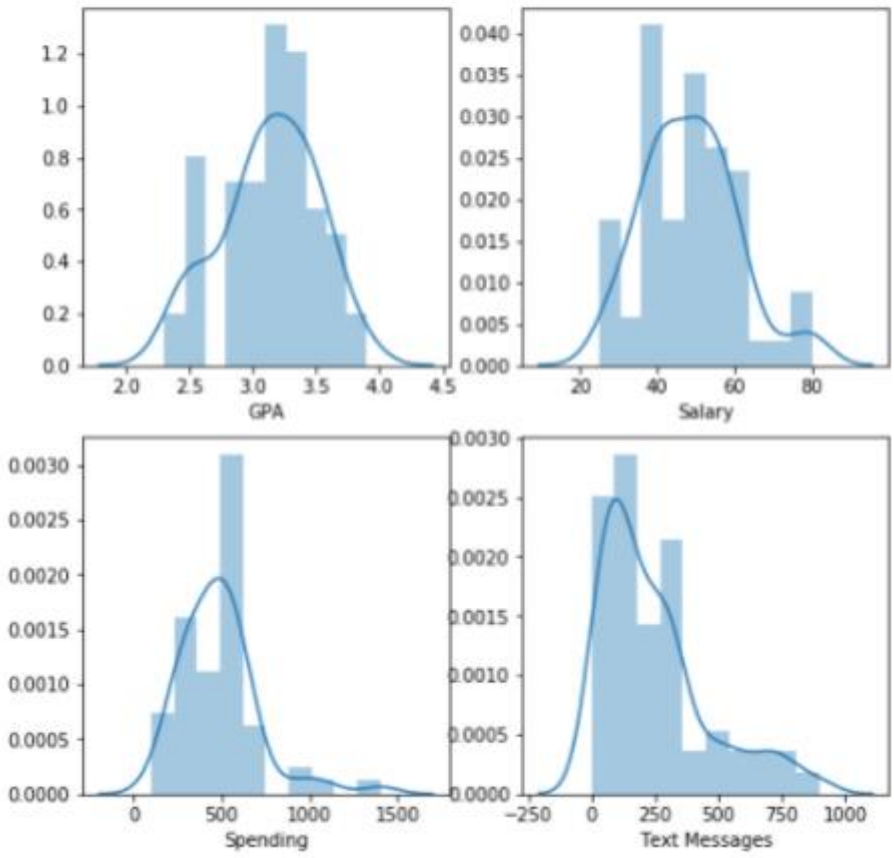
| Salary | False | True |
|---|---|---|
| Gender | | |
| False | 0.517241 | 0.482759 |
| True | 0.454545 | 0.545455 |

The probability that a randomly selected female earns 50 or more is 54%

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.


Probability Plot
Ordered Values
Theoretical quantiles
Probability Plot
Ordered Values
Theoretical quantiles
Probability Plot
Ordered Values
Theoretical quantiles
Probability Plot
Ordered Values
Theoretical quantiles

```
skew value of GPA is -0.3146000894506981
skew value of Salary is 0.5347008436225946
skew value of Spending is 1.5859147414045331
skew value of Text Message is 1.2958079731054333
```




The probability plot can be used to find whether the dataset follows a normal distribution or not, in our dataset we can find the points that follow a straight line and we can say that all the GPA, salary, spending, and text messages follow a normal distribution. Looking at the skew value if the value is zero it is symmetric data if we have a negative value for the skew that indicates that the data are skewed left and the positive value of skew indicates the data are skewed toward right.

## CONCLUSION

We have a dataset of students answering to the survey and we have 62 responses from the students both male and female. We have an almost equal number of male and female students. Many students have the intention of graduating the retailing and marketing seem to have chosen by quite number of students. 2/3 of the students are looking for a part-time job. The mean salary means to be around 50

SHINGLES ANALYSIS A & B

PROBLEM:

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

EDA

We have two variable A and B

NULL VALUES are present in B

```
A    0
B    5
```

FIVE POINT SUMMARY

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| A | 36.0 | 0.316667 | 0.135731 | 0.13 | 0.2075 | 0.29 | 0.3925 | 0.72 |
| B | 31.0 | 0.273548 | 0.137296 | 0.10 | 0.1600 | 0.23 | 0.4000 | 0.58 |

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.

SOLUTION:

In this problem we have provided with two independent samples of shingles A and B population standard deviation is unknown and hence we can't perform z test. So we have to go with t-test.

Since we have to find the mean moisture level is less than the permissible limit for the both samples we have perform one sample t-test for sample A and sample B.

SAMPLE A

STEP 1: DEFINE NULL AND ALTERNATE HYPOTHESIS

The null hypothesis states that the moisture content of sample A is greater or than equal to the permissible limit, $\mu \geq 0.35$

The alternative hypothesis states that the moisture content of sample A is less than permissible limit, $\mu < 0.35$

$H0 : \mu \geq 0.35$

$H : \mu < 0.35$

STEP 2: DECIDE THE SIGNIFICANCE LIMIT

Since alpha value is not given in the question we assume it has alpha = 0.05

STEP 3 IDENTIFY THE TEST STATISTIC

We have sample A and we do not know the population standard deviation. Sample size n=36. We use the t distribution and the $tSTAT$ test statistic for one sample t-test.

STEP 4: CALCULATE THE P - VALUE AND TEST STATISTIC

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Xbar = 0.316667

S = 0.135731

N = 36

Mu = 0.35

Tstat = -1.4735

(P Value/2) = 0.0747

STEP 5: DECIDE TO REJECT OR ACCEPT NULL HYPOTHESIS

Since tstat > p_value, we fail to reject the null hypothesis

We conclude that the moisture content is greater than permissible limit in sample A. SAMPLE B

STEP 1: DEFINE NULL AND ALTERNATE HYPOTHESIS

The null hypothesis states that the moisture content of sample B is greater or than equal to the permissible limit, $\mu \geq 0.35$

The alternative hypothesis states that the moisture content of sample B is less than permissible limit, $\mu < 0.35$

$H0 : \mu \geq 0.35$

$H : \mu < 0.35$

STEP 2: DECIDE THE SIGNIFICANCE LIMIT Since alpha value is not given in the question we assume it has alpha = 0.05

STEP 3 IDENTIFY THE TEST STATISTIC

We have sample A and we do not know the population standard deviation. Sample size n=31. We use the t distribution and the $tSTAT$ test statistic for one sample t-test.

STEP 4: CALCULATE THE P - VALUE AND TEST STATISTIC

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Xbar = 0.2735

S = 0.1372

N = 31

Mu = 0.35

Tstat = -3.1003

P Value = 0.0020

STEP 5: DECIDE TO REJECT OR ACCEPT NULL HYPOTHESIS

Since tstat < p_value, we reject the null hypothesis We conclude that the moisture content is less than permissible limit in sample B

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?

STEP 1 DEFINE NULL AND ALTERNATIVE HYPOTHESIS

In testing whether the mean for shingles A and Shingles B are the same, the null hypothesis states that the mean of shingle A to mean of shingle B are the same, $\mu{A}$ equals $\mu{B}$.

The alternative hypothesis states that the mean are different, $\mu {A}$ is not equal to $\mu{B}$ STEP 2: DECIDE THE SIGNIFICANCE LIMIT

Since alpha value is not given in the question we assume it has alpha = 0.05

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 not the same. The sample size is, n > 30. So we use the t distribution and the $tSTAT$ test statistic for two sample test.

Two tail test

STEP 4: CALCULATE THE P - VALUE AND TEST STATISTIC

$$t = (M_1 - M_2)/\sqrt{(s^2_{M_1} + s^2_{M_2})}$$

CALCULATION: N1= 36

N2 =31

M1= 0.32

M2 = 0.27

S 2 1 = 0.02

S2 1 = 0.02

DF1 = 35 DF2 = 30

Tstat 1.2896282719661123

P Value 0.2017496571835306

STEP 5 DECIDE TO REJECT OR ACCEPT THE NULL HYPOTHESIS Since tstat > p_value, we fail to reject the null hypothesis

We conclude that mean for shingles A and singles B are not the same.