April 2021 | PG-DSBA-Online

SMDM PROJECT REPORT

Submitted By: Dev Tripathi

Contents

List of Figures

List of Tables

Problem 1:

Wholesale Customers Analysis

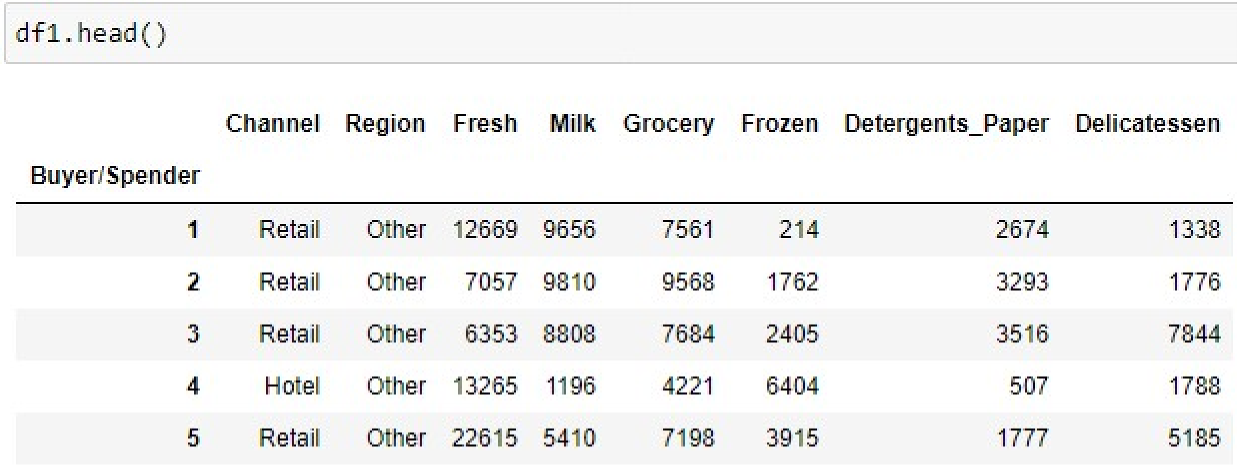
Information about dataset-

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. This 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).

Now, starting with exploratory data analysis-

EDA

Dataset-



```
df1.head()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Buyer/Spender								
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

FIGURE 1: WHOLESALE CUSTOMERS DATASET (HEAD)

Variables information:

#	Column	Non-Null	Count	Data type
0	Channel	440	non-null	object
1	Region	440	non-null	object

2	Fresh	440	non-null	int64
3	Milk	440	non-null	int64
4	Grocery	440	non-null	int64
5	Frozen	440	non-null	int64
6	Detergents_Paper	440	non-null	int64
7	Delicatessen	440	non-null	int64

Table 1: Variables Information of Wholesale Customers Dataset

‘Channel’ and ‘Region’ are categorical variables as mentioned earlier. ‘Fresh’, ’Milk’, ‘Grocery’, ‘Frozen’, ‘Detergent_Paper’ and ‘Delicatessen’ are numerical variable which contains information about spending corresponding to their item type. There are no null values present in this dataset.

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

Solution: Summarization of Data using Descriptive Statistics methods:

	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicatessen**	**Total**
count	440	440	440	440	440	440	440
mean	12000.3	5796.266	7951.277	3071.932	2881.493	1524.87	15429.57
std	12647.33	7380.377	9503.163	4854.673	4767.854	2820.106	15560.82
min	3	55	3	25	3	3	227
25%	3127.75	1533	2153	742.25	256.75	408.25	5715.75
50%	8504	3627	4755.5	1526	816.5	965.5	11180
75%	16933.75	7190.25	10655.75	3554.25	3922	1820.25	18898.25
max	112151	73498	92780	60869	40827	47943	137577

Table 2: Summarization of Data using methods of Descriptive Statistics

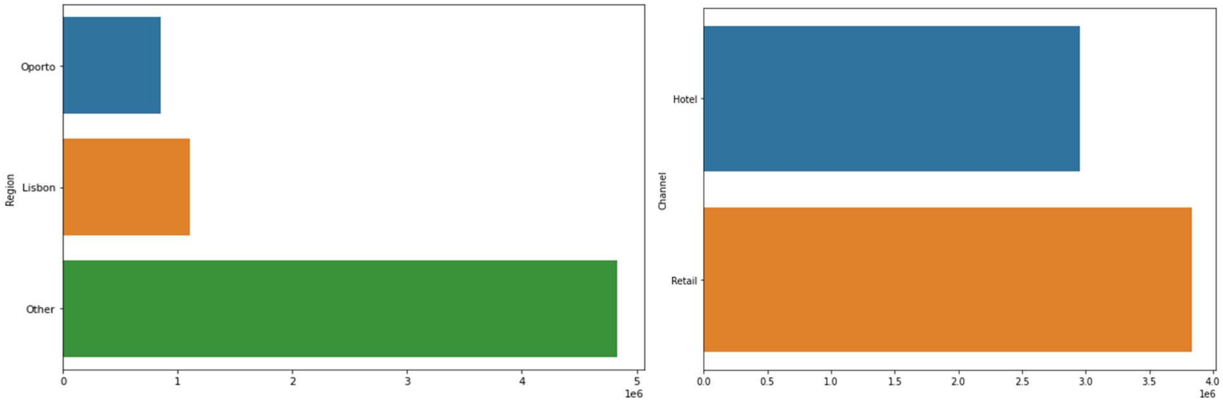
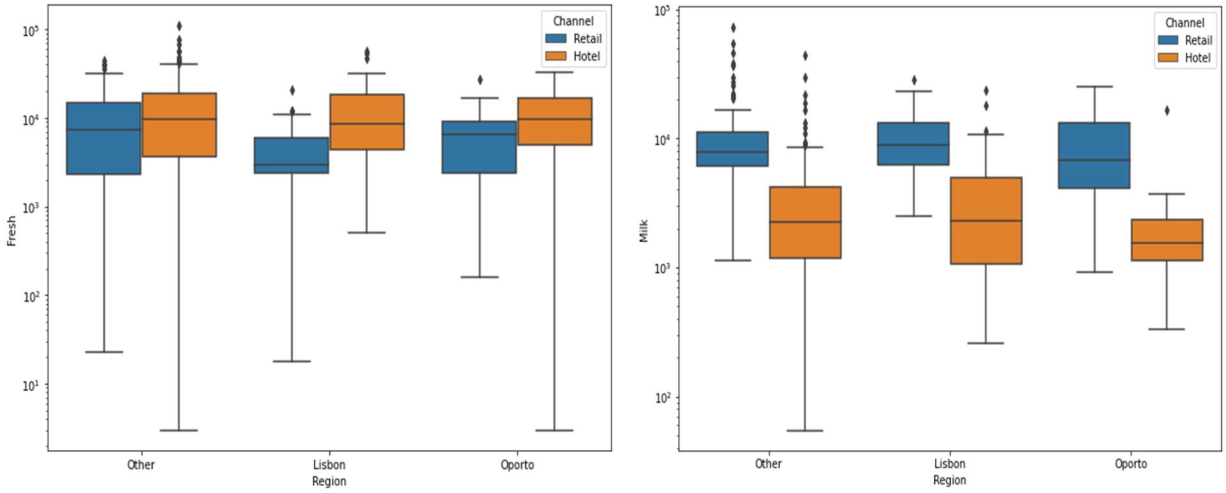


Figure 2: Bar plots of Spending In Different Regions (A) And Channels (B)

As the above figure depicts, the Region and Channel which spent the most are **'Other'** and **'Retail'** respectively. Similarly, we can observe, the Region and Channel which spent the least are **'Oporto'** and **'Hotel'** respectively.

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

Solution: For describing the different varieties statistically, we can use the boxplot function in seaborn library. Since, the range is huge for all these variables, the normal scale will be unable to show the details due to overlapping lines. Hence, the log scale for y axis is used.



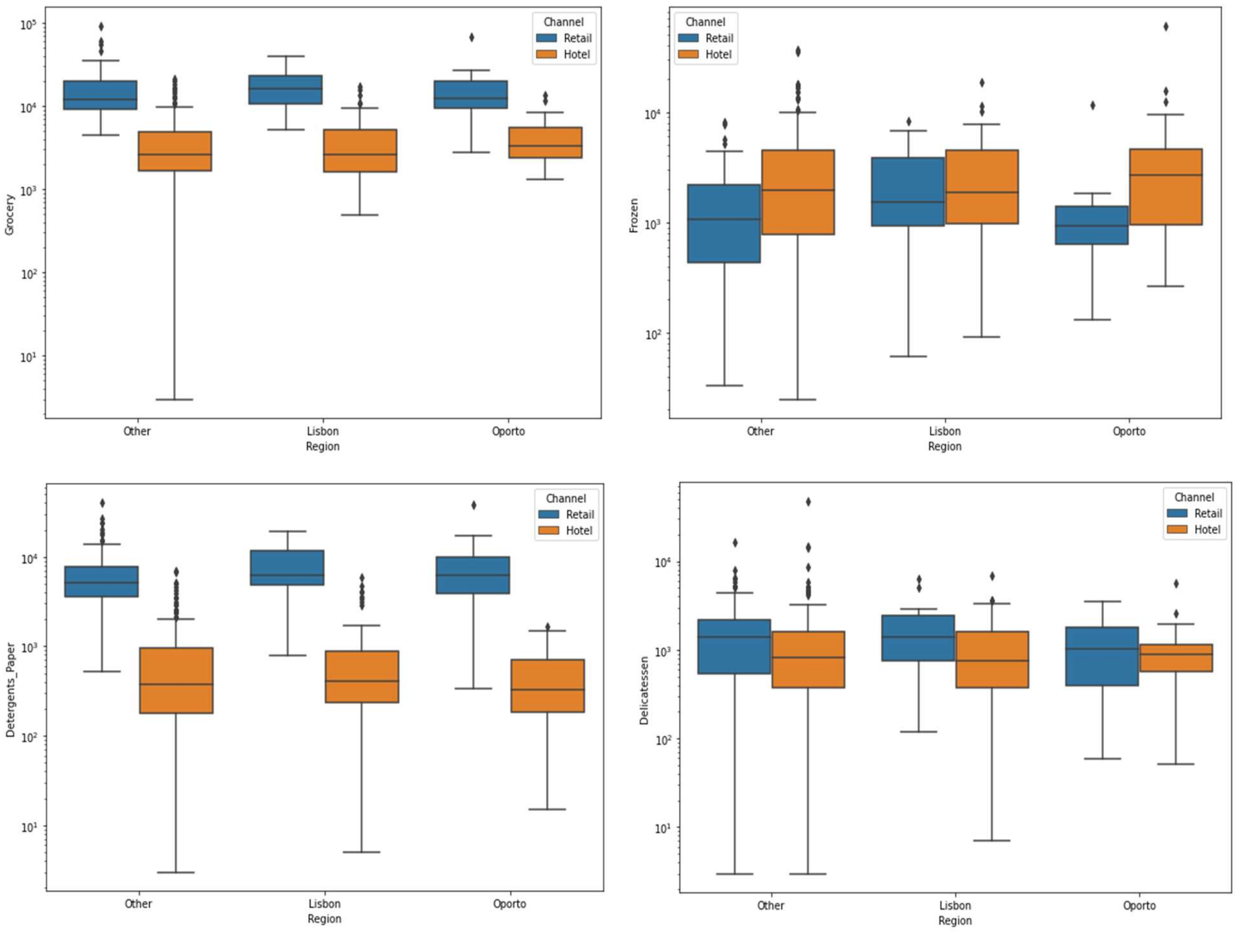


FIGURE 3: BOX PLOTS FOR DIFFERENT ITEMS

Now, using the coefficient of variation for depicting the differences/ similarities for different Regions and Channels for these items, since coefficient of variation is better measure of the variation when it comes to comparing two variables. The following table shows the same:

		Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
Region	**Channel**						
Lisbon	**Hotel**	1.010366	1.362187	0.948436	1.038772	0.893848	1.101167
	Retail	0.844395	0.651707	1.012104	0.911902	0.547926	0.595605
Oporto	**Hotel**	0.93837	0.865205	0.755994	1.957877	0.681008	1.265113
	Retail	0.836982	0.959034	0.917003	1.562595	0.836754	0.70016

Other	**Hotel**	2.406988	1.394922	1.060061	1.352192	0.922363	1.289886
	Retail	1.154817	0.868697	0.975375	0.989504	0.767229	0.958414

Table 3: CV Values for Different Items

Following observations can be made from the above table and boxplots:

1. Behavior of Detergents_Paper, Grocery and Milk is quite different for different Channels. Interestingly, the behavior is similar across different regions (median is almost same) if we exclude the extreme values.
2. For the Detergents_Paper and Milk item, the dispersion of data for Lisbon region is quite different for different channels (please refer to CV table).
3. For Fresh, Frozen and Delicatessen items, the behavior is very similar for different regions and channels, if we exclude the extreme values.
4. For the Delicatessen items in the Other region, the data dispersion for different channels i.e. for Hotel and Retail is very different, which can observed from the CV values for each case.
5. For Grocery items in the Other region, the minimum values and the range is very different for different channels.

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

Solution: For this, using the boxplot we can check the consistency of data i.e. the smaller the box is, the more consistent is the data

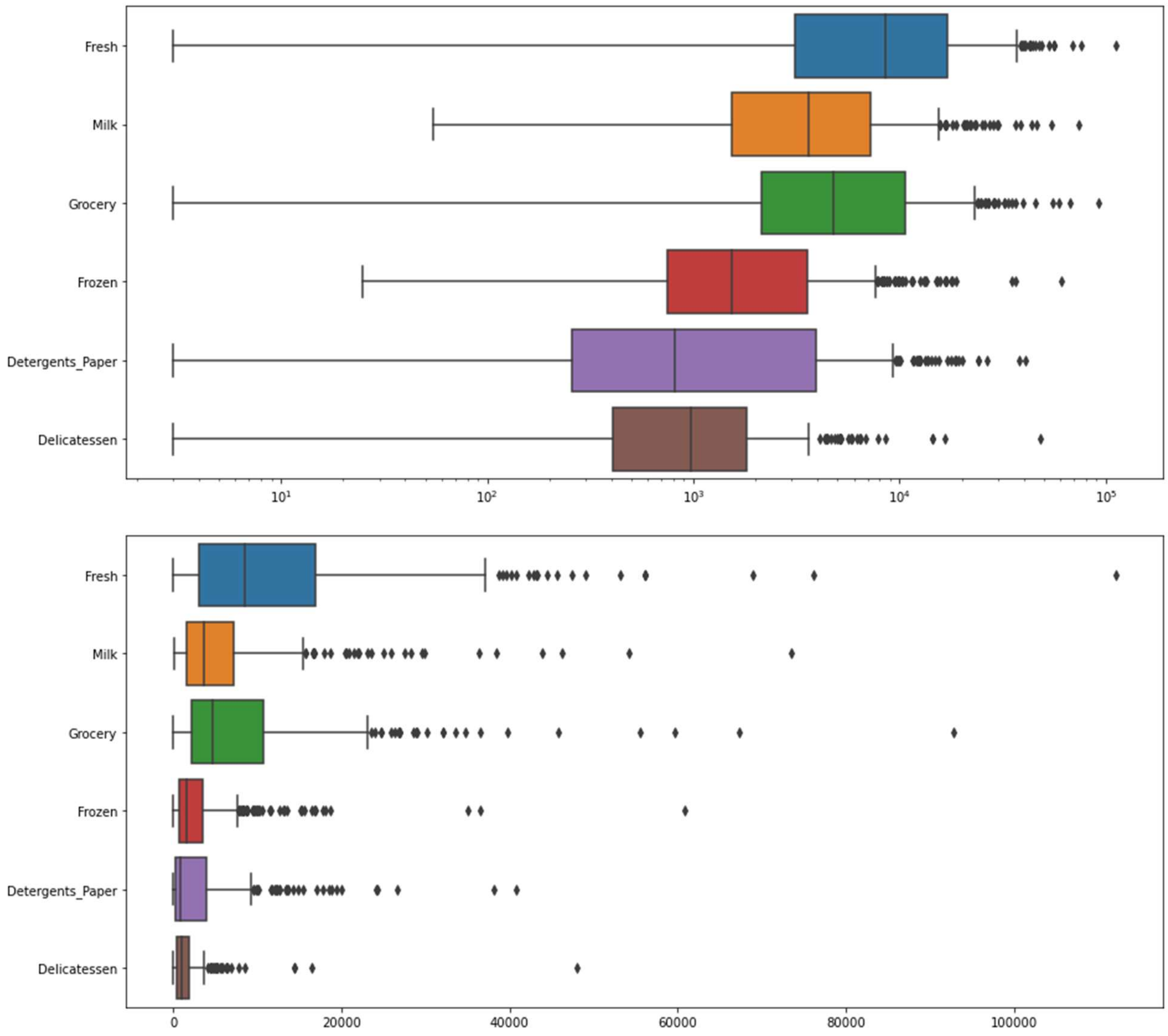


Figure 4: Box Plots for Different Items (A) log scale (B) Normal Scale

Following conclusions can be made from this plot,

1. For item Fresh, Grocery and Detergent_Paper the data seems to be more inconsistent than the other items.
2. The outliers are extremely high compared to median for each item in the data.

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

Solution:

As, we can observe from the figure 3, all the items have got outliers in the data.

1.5 On the basis of this report, what are the recommendations? How can your analysis help the business to solve its problem? Answer from the business perspective

Solution:

1. Since, the data variations across different regions and channels, are quite high for some items, for the wholesale distributor can look into the problems or challenges the retailers face.

2. It can also help the distributor for making strategic expansion plans.

Problem 2:

CMSU Students Survey Dataset Analysis

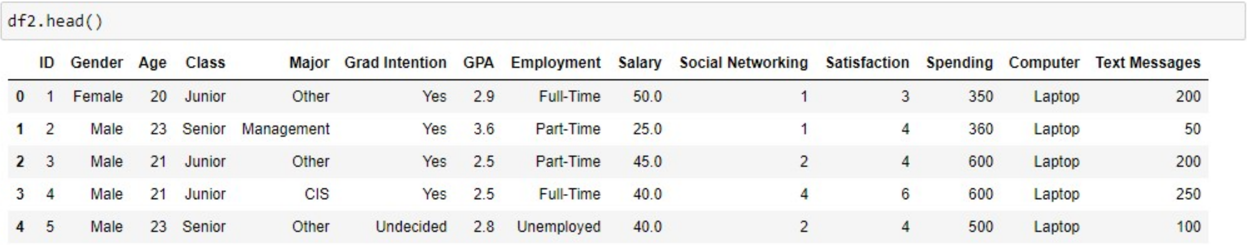
Information about dataset-

The Student News Service at Clear Mountain State University (CMSU) has gathered the data about the undergraduate students that attend CMSU. CMSU created and distributed a survey of 14 questions and receives responses from 62 undergraduates.

Now, let's start with exploratory data analysis -

EDA

Dataset-



```
df2.head()
```

```
   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
```

FIGURE 5: CMSU UNDERGRAD STUDENTS SURVEY DATASET (HEAD)

Variables Information-

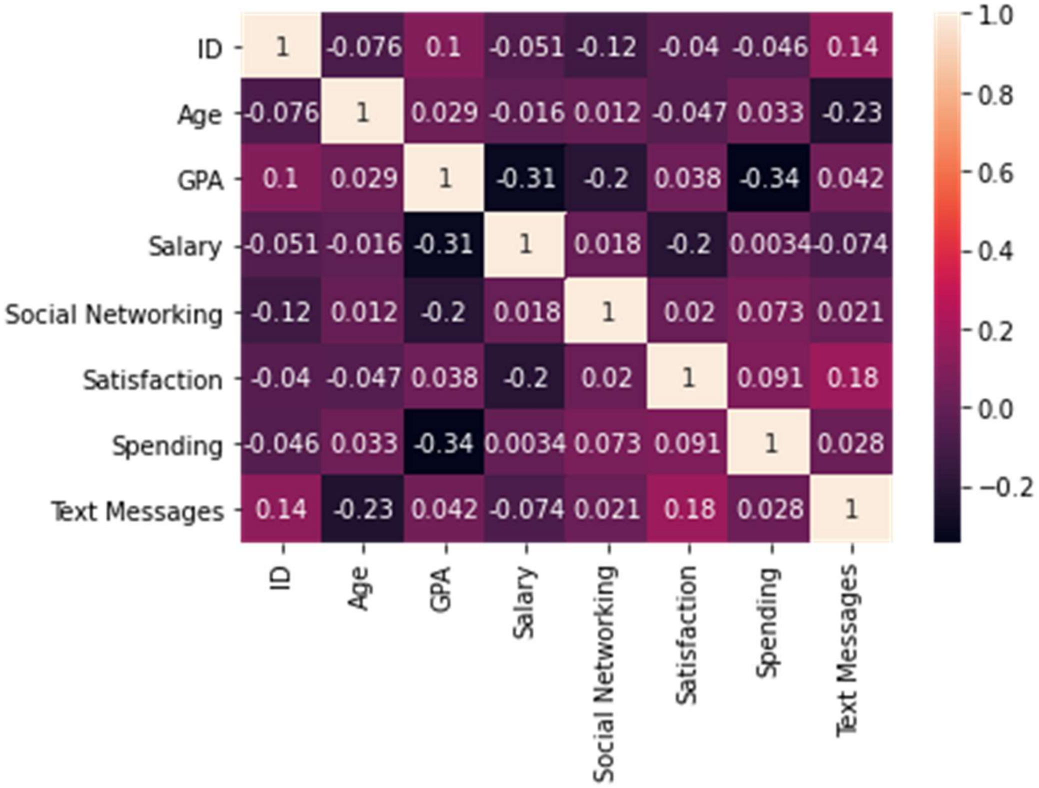
The variables are nothing but the questions asked in the survey from students:

Index	Column	Non-Null	Count	Data type
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

TABLE 4: VARIABLE INFORMATION OF SURVEY DATASET

Heat Map-



	ID	Age	GPA	Salary	Social Networking	Satisfaction	Spending	Text Messages
ID	1	-0.076	0.1	-0.051	-0.12	-0.04	-0.046	0.14
Age	-0.076	1	0.029	-0.016	0.012	-0.047	0.033	-0.23
GPA	0.1	0.029	1	-0.31	-0.2	0.038	-0.34	0.042
Salary	-0.051	-0.016	-0.31	1	0.018	-0.2	0.0034	-0.074
Social Networking	-0.12	0.012	-0.2	0.018	1	0.02	0.073	0.021
Satisfaction	-0.04	-0.047	0.038	-0.2	0.02	1	0.091	0.18
Spending	-0.046	0.033	-0.34	0.0034	0.073	0.091	1	0.028
Text Messages	0.14	-0.23	0.042	-0.074	0.021	0.18	0.028	1

FIGURE 6: CORRELATION HEAT MAP OF SURVEY DATASET VARIABLES

As we can clearly notice, the correlation values are coming out to be very low. Hence, there is no significant correlation exists between any two variables.

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

Solution:

Contingency Table: A contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. It is mainly used for studying the correlation between two variables.

2.1.1 Gender and Major:

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

Table 5: Contingency Table Between Gender and Major

2.1.2. Gender and Grad Intention

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

Table 6: Contingency Table Between Gender and Grad Intention

2.1.3. Gender and Employment

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

Table 7: Contingency Table Between Gender and Employment

2.1.4. Gender and Computer

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

Table 8: Contingency Table Between Gender and Computer

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?

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

Solution:

2.2.1 The probability that a randomly selected CMSU will be male:

$$\text{P (Male)} = \frac{No.\ of\ Males}{Total\ no.\ of\ students} = \frac{29}{62} = 0.46774193548$$

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

$$\text{P (Female)} = \frac{No.\ of\ Females}{Total\ no.\ of\ students} = \frac{33}{62} = \ 0.53225806451$$

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.

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

Solution:

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

For calculation of the conditional probability of different majors among the male students we can use the Gender vs. Major contingency table we created earlier

Say, for example, we want to know P (Accounting | Male) so it can be calculated as:

$$P\ (Accounting \mid Male) = \frac{P\ (Accounting \cap Male)}{P\ (Male)} = \frac{4/62}{29/62} = \frac{4}{29} = \ 0.137931$$

Similarly, we can obtain values for other 'Majors'.

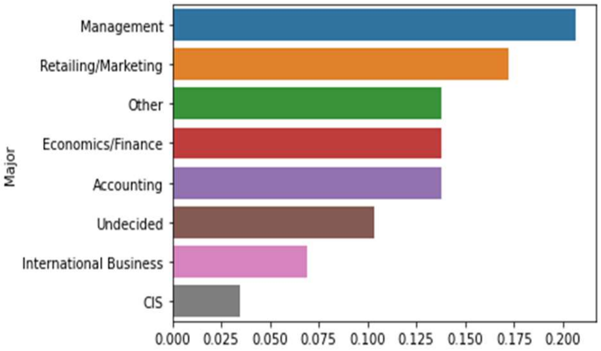


Figure 7: Bar plot for Probabilities of Different Majors Among Males

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

The same procedure, which we used in 2.3.1 can be used for Females and following plot can be obtained:

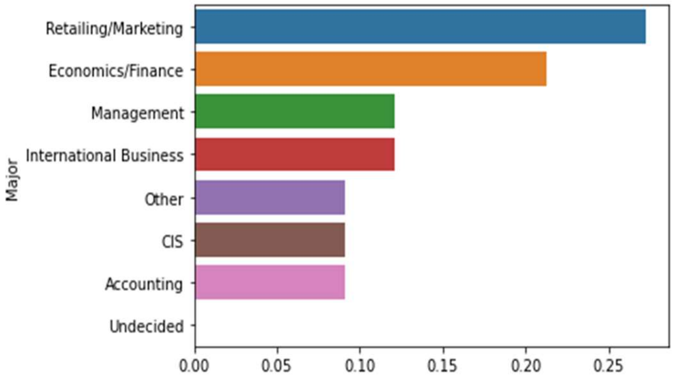


Figure 8: Bar plot for Probabilities of Different Majors Among Females

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

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

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

Solution:

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

Similar procedure which we did in the previous problem can be used for finding the required probability:

For calculation of the conditional probability of different intentions among the male students towards graduation we can use the Gender and Grad Intention contingency table we created earlier

Say, for example, we want to know P (Yes | Male) so it can be calculated as:

P (Yes | Male) $= \frac{P\ (Yes \cap Male)}{P\ (Male)} = \frac{\frac{17}{62}}{\frac{29}{62}} = \frac{17}{29} = 0.586207$

Similarly, we can obtain values for other ‘Grad intentions’.

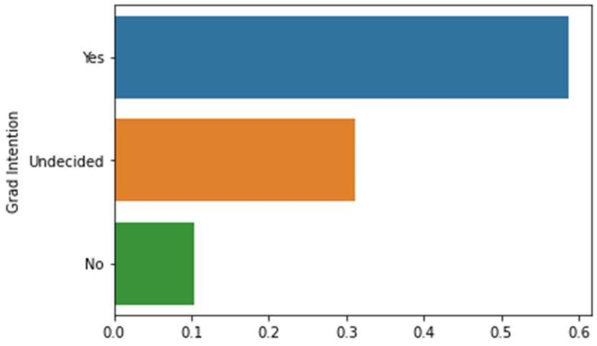


FIGURE 9: BAR PLOT FOR PROBABILITY OF GRAD INTENTION AMONG MALES

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

$$P\ (NO\ Laptop\ |\ Female)\ = \frac{P\ (NO\ Laptop\ \cap\ Female)}{P\ (Female)} = \frac{\frac{4}{62}}{\frac{33}{62}} = \frac{4}{33} = 0.121212$$

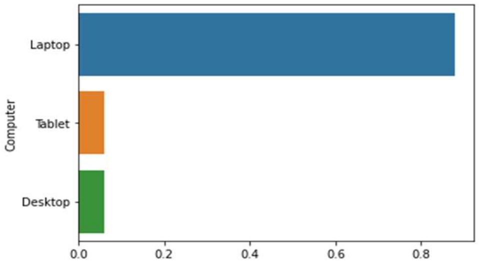


FIGURE 10: BAR PLOT FOR PROBABILITY OF COMPUTER TYPE AMONG FEMALES

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?

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

Solution:

2.5.1. The probability that a randomly chosen student is either a male **(M)** or has full-time employment **(FTE)**

$$P\ (M\ \cup\ FTE)\ =\ P\ (M)\ +\ P\ (FTE) - P\ (M\ \cap\ FTE)$$

Now, since $P\,(M) = 0.46774;\ P\,(FTE) = \frac{10}{62} = 0.161290;\ P\,(M \cap FTE) = \frac{7}{62} = 0.112903$ we get;

$$\boldsymbol{P\,(M \cup FTE) = 0.51612}$$

Hence, the probability that a randomly chosen student is either a male (M) or has full-time employment (FTE) is **0.51612.**

2.5.2. The conditional probability that given a female **(F)** student is randomly chosen, she is majoring in international business **(IB)** or management **(M).**

Since, any student can choose only one Major, the event P (IB) and P (M) are mutually exclusive. Which implies,

$$\boldsymbol{P\,(IB \cap M) = 0}$$

Hence, probability that given a female student is randomly chosen, she is majoring in IB or M is:

$$P\,((IB\ \cup\ M)|F) = P\,(IB\,|F)\ +\ P\,(M|F) - 0$$

$$= 4/33\ +\ 4/33\ - 0$$

$$= 0.242424$$

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?

Solution:

Contingency Table:

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Table 9: Contingency Table of Gender and Intent to Graduate at 2 levels

Now, if we assume that this sample is representative of student population, grad intention is not independent of gender, because the percentage of ‘No’ grad intention is quite high among the female than Males, which we can clearly observe from the table 9.

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

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

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.

Solution:

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

For this we are going to need number of students having GPA less than 3. From our analysis we obtained this value to be 24. Hence,

$$\boldsymbol{P\ (GPA < 3)} = \frac{24}{62} = \ \mathbf{0.3871}$$

2.7.2. a.) The conditional probability that a randomly selected male (M) earns (E) 50 or more.

$$\boldsymbol{P\ (E \geq 50\ |\ M)}\ = \frac{P\ (E>50\cap\ M)}{P\ (M)} = \frac{10/62}{29/62} = \frac{10}{29} = \ \mathbf{0.34482}$$

b.) The conditional probability that a randomly selected female (F) earns (E) 50 or more.

$$\boldsymbol{P\ (E \geq 50\ |\ F)}\ = \frac{P\ (E>50\cap\ F)}{P\ (F)} = \frac{10/62}{33/62} = \frac{10}{33} = \ \mathbf{0.30303}$$

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 for this whole Problem 2.

Solution:

For checking whether the given numerical variables follow normal distribution or not there are many tests which can be performed, like Shapiro Test, K-S test etc. In this analysis, we are using Shapiro Wilk test.

The Shapiro-Wilk test tests the **null hypothesis that the data was drawn from a normal distribution.**

From our, analysis we found that only for the GPA variable, the p-value was coming out to be greater than significance level which we have assumed to be equal to 0.05.

That means, we do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis. Hence, only GPA value follows a normal distribution at significance level of 0.05.Rest of the variables, which are Salary, Spending and text Messages do not follow the normal distribution.

Summary

Following conclusions can be made from this analysis,

1. Among male students, the Marketing and Management majors are popular. On the other hand, among females, Marketing and Finance are very popular majors.

2. Mean and median for student satisfaction is 3.4 and 4 respectively. Hence, we can conclude that the students are satisfied with CMSU. (Assuming 6 is the maximum rating one can give)

3. Most of the students independent of their gender, have laptops. A very few of them use other computers like tablets or desktops.

4. If we look at the distribution of Text Messages for different classes, the sophomore class has got highest mean and median values compared to other two classes (i.e. Junior, Senior)

5. Interestingly, the GPA variable follows normal distribution for the students of CMSU.

Problem- 3:

A & B Shingles Dataset Analysis

Information about dataset-

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

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

Solution:

H_0: **The null hypothesis:** It is a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0. In other words, alternative hypothesis is nothing but the claim one wants to test.

For the given problem,

- **Null Hypothesis (*H_0*):** $\mu_A = 0.35$
- **Alternate Hypothesis (*H_A*):** $\mu_A < 0.35$

Basic Information on the Test performed for testing null hypothesis:

The difference in means between two Normal distributions with unknown variance follows a Student's T-distribution. The t-test is any statistical hypothesis test in which the test statistic follows a Student's T-distribution under the null hypothesis.

Significance Level (α): Significance level assumed for this test is 0.05. Hence, the confidence level is 95%.

T-test Results-

One sided T-test is performed and the test statistic, p-value obtained are following:

t-statistic → -1.4735

p-value → 0.1496

Since p-value is 0.1495; thus we do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis at the significance level of 0.05.

Hence, for sample A, moisture content is not less than 0.35 pound per 100 square feet.

Similarly, for sample B:

- **Null Hypothesis (*H₀*):** $\mu_B = 0.35$
- **Alternate Hypothesis (*H_A*):** $\mu_B < 0.35$

T-test Results-

One sided T-test is performed and the test statistic, p-value obtained are following:

t-statistic → -3.1003

p-value → 0.0041

Since p-value is 0.0041; thus we have enough evidence to reject the null hypothesis in favor of alternative hypothesis at the significance level of 0.05.

Hence, for sample B, moisture content is less than 0.35 pound per 100 square feet.

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?

Solution:

For testing equality of population mean two sample t-test can be performed.

Two Sample T-test:

Before performing the two sample t-test we must check whether **the variances of the two sample are equal or not**. Because if this doesn't hold we can't perform the two sample t-test.

In our analysis, we found that the difference between variance of sample A and B is very small. Hence, we are assuming that **variances for the two samples are equal.**

- Null Hypothesis (***H₀***): $\mu_A = \mu_B$
- Alternate Hypothesis (***HA***): $\mu_A \neq \mu_B$

<u>T-test Results-</u>

One sided T-test is performed and the test statistic, p-value obtained are following:

t-statistic →1.2896

p-value → 0.10087

Since p-value is 0.10087; thus we do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis at the significance level of 0.05.

Hence, population mean for shingles A and B are equal.