Table of Contents

Contents

Executive Summary

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

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and using central tendency and other parameter. The data consist of 6 varieties products and 3 different region 440 time sales. Total size of data 3960. This assignment should help the student in exploring the summary statistics, contingency tables, conditional probabilities & hypothesis testing.

Data Description

1. Buyer/Spender : Costumer count.
2. Channel : Retail & hotel.
3. Region : Other, Lisbon& Oporto.
4. Fresh : Verity of preached amount.
5. Milk : Verity of preached amount.
6. Grocery : Verity of preached amount.
7. Frozen : Verity of preached amount.
8. Detergents Paper : Verity of preached amount.
9. Delicatessen : Verity of preached amount.

Sample of the dataset

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185
5	6	Retail	Other	9413	8259	5126	666	1795	1451
6	7	Retail	Other	12126	3199	6975	480	3140	545
7	8	Retail	Other	7579	4956	9426	1669	3321	2566
8	9	Hotel	Other	5963	3648	6192	425	1716	750
9	10	Retail	Other	6006	11093	18881	1159	7425	2098

Table.1-Dataset Sample

Dataset has 9 variables with 6 different types of Product. Each product has different sets of attributes. Based on the characteristic price of the product is defined.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

#	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

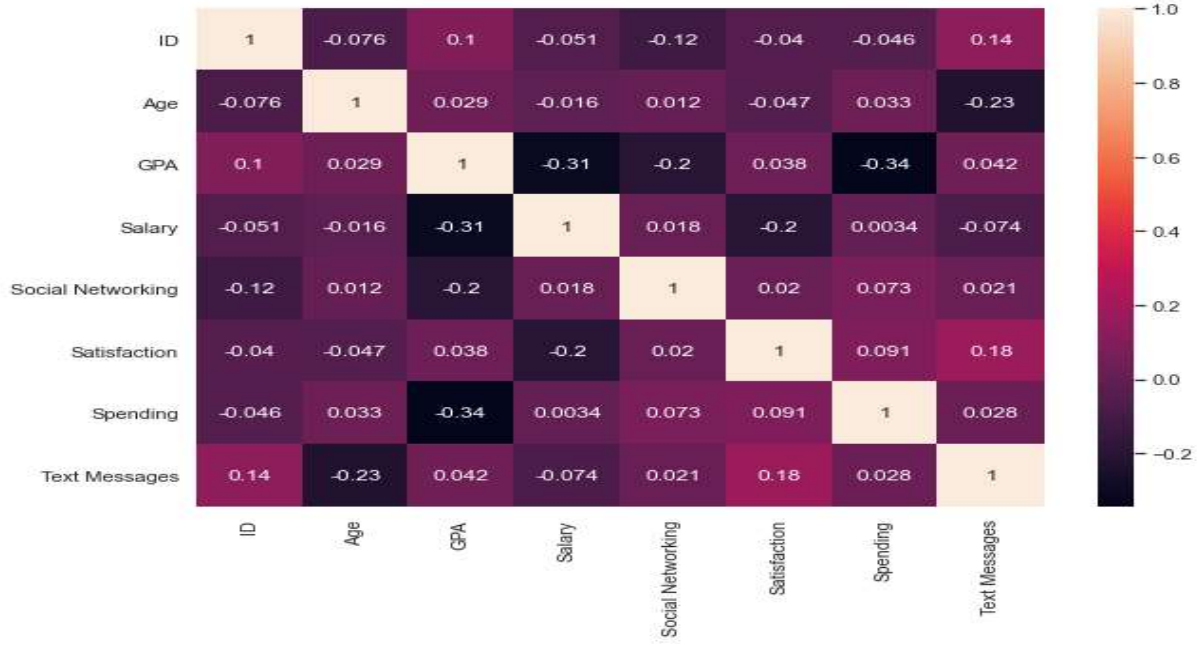
There are total 440 rows and 9 columns in the dataset. 2object and 7 integer dataset.

Check for missing values in the dataset

```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

From the above results we can see that there is no missing value present in the dataset.

Correlation Plot

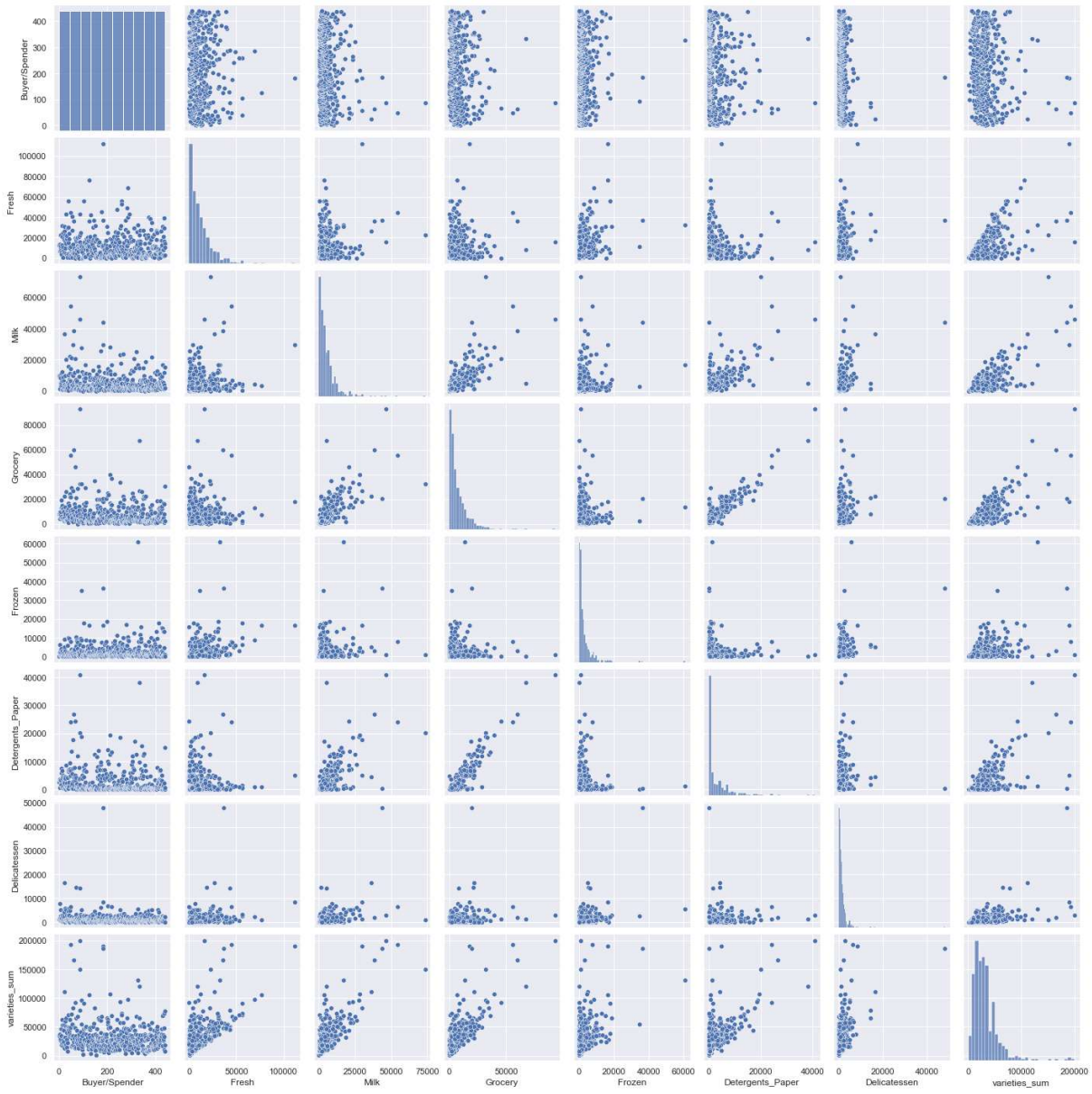


From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 0.2 or 1 are highly positively correlated. Correlation values near to 0 are not correlated to each other.

Pair plot

Pair-plot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

As per plot shown below, there is a linear relationship between Ferozen and Detergents Paper.



Q1.1.1 Use methods of descriptive statistics to summarize data.

Descriptive statistics helps to know the various aspects like max, min values of all the columns, their mean and standard deviation. Their values at 25%, 50% and 75% can also be determined from here. The Descriptive summary is shown below.

	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Fig.-1.0

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185
5	6	Retail	Other	9413	8259	5126	666	1795	1451
6	7	Retail	Other	12126	3199	6975	480	3140	545
7	8	Retail	Other	7579	4956	9426	1669	3321	2566
8	9	Hotel	Other	5963	3648	6192	425	1716	750
9	10	Retail	Other	6006	11093	18881	1159	7425	2098

Fig.1.0.1

Dataset has 9 variables Buyer/ Spender, Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen. Channel and Region both are categorical columns while Buyer/ Spender, Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen are integer.

```
Other      316
Lisbon      77
Oporto      47
Name: Region, dtype: int64
```

```
whsale.Channel.value_counts()
```

```
Hotel      298
Retail     142
Name: Channel, dtype: int64
```

Fig.1.0.2

Region and channel both rows wise data customer visit and region of counts.

Q1.1.2 /Q1.1.3 Which Region and which Channel spent the most? .Which Region and which Channel spent the least?

Channel	varieties_sum
Hotel	7999569
Retail	6619931

Region	varieties_sum
Lisbon	2386813
Oporto	1555088
Other	10677599

` Fig.1.1.2/1.1.3

From the above result:-

1. Ans-1.1.2 as per observation report in channel, **hotel** are total spending **most** spent as camper retail.
2. Ans-1.1.3 as per observation report in region, **Oporto** are total spending **lest** spent as camper rest Lisbon & Other.

Q1.2. 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.

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955
Retail	25986	1264414	1521743	2317845	234671	1032270	248988

Fig.1.2.0

From the above result:-

1. As per observation report in Channel 'Hotel': Average Highest Spending in Fresh items and Lowest Spending in Detergents Paper items.
2. As per observation report in Channel 'Retail': Average Highest Spending in Grocery items and Lowest Spending in Frozen items

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Lisbon	18095	854833	422454	570037	231026	204136	104327
Oporto	14899	464721	239144	433274	190132	173311	54506
Other	64026	3960577	1888759	2495251	930492	890410	512110

Fig.1.2.1

From the above result:-

1. As per observation report in Region 'Lisbon': Average Highest Spending in Fresh items and Lowest Spending in Delicatessen items.
2. As per observation report in Region 'Oporto': Average Highest Spending in Fresh items and Lowest Spending in Detergents_paper items.
3. As per observation report in Region 'Other': Average Highest Spending in Milk items and Lowest Spending in Delicatessen items.

Q1.3 On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?

- **Below is the output from Python :-**

Coefficient of variation (CV)	inconsistent behaviour
CV_for_Fresh	1.0539179237648593
CV_for_Milk	1.2732985841005522
CV_for_Grocery	1.1951743729613995
CV_for_Frozen	1.5803323838615222
CV_for_Detergents_Paper	1.6546471384293562
CV_for_Delicatessen	1.849406897322304

Fig-1.3 Descriptive table

From the above result:-

After calculate the data we show that CV (coefficient of variation) value as per formula CV=Std/Mean, As per descriptive table Fig-1.3 Fresh item behaviour are least inconsistent b ehaviour compare to rest of item. As we can see that Item 'Delicatessen' has highest CV. And - **Fresh" item has lowest CV,** so it is showing least inconsistent behaviour.

Q1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

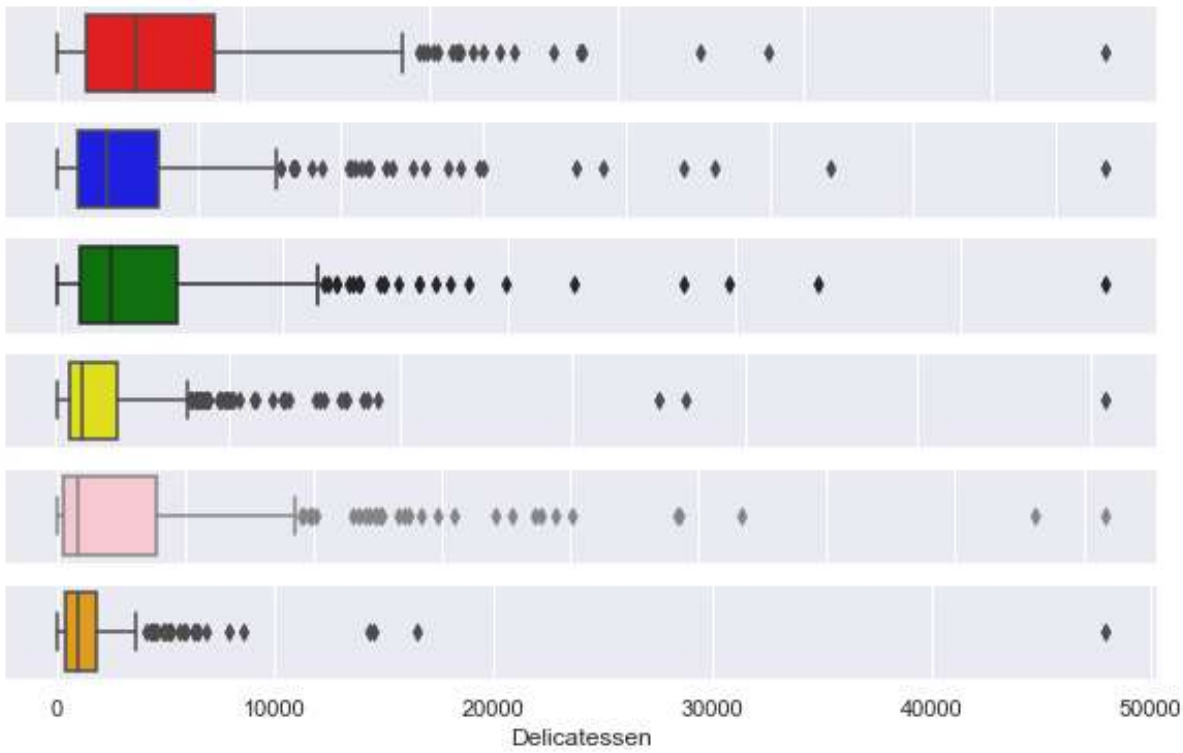


Fig.1.4 Boxplot

From the above result:-

As per boxplot we see that all 6 item have outliers...

Q1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective.

Ans:-On the basis of my analysis, I find out that there are inconsistencies in spending of diff erent items by calculating CV, which should be minimized. The spending of Hotel and Retail channel are different which should be more or less equal. And also spent should equal for di fferent regions. Need to focus on other items also than “Fresh” and “Grocery” item. We nee d to keep more stocks of items depending on different regions and Channel to avoid any sho rtage.

PROBLEM – 2

Clear Mountain State University (CMSU) Data Analysis

Table of Contents

Contents

Executive Summary

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).

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and using central tendency and other parameter. The data consist of 14 columns in this data 2 float, 6 integer, & 6 object total 62 row. Total size of data 868. This assignment should help the analysis of graduated intention, employment, Gender& salary verity define, contingency tables, conditional probabilities & hypothesis testing.

Data Description

#	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

Sample of the dataset

1. ID: Numbering of verity.
2. Gender: Male, Female.
3. Age: Male female age mention.
4. Class: (Junior, senior, sophomore).
5. Major: (Accounting, CIS, Economics/Finance, International Business, Management, Other, Retailing/Marketing, Undecided).
6. GPA: ID Gender wise mark.
7. Employment: full-time, part-time, unemployment.
8. Social Networking: verity of data.
9. Salary: Employs salary.
10. Graduation intention: (yes, no, Undecided).
11. Satisfaction: verity of data
12. Spending: verity of data
13. Computer: (Laptop, Desktop, tablet).
14. Text massages: verity of data.

Sample of the dataset

	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

In this sample 14 columns and 62 no of verity rows.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
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
```

Check for missing values in the dataset

```
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
```

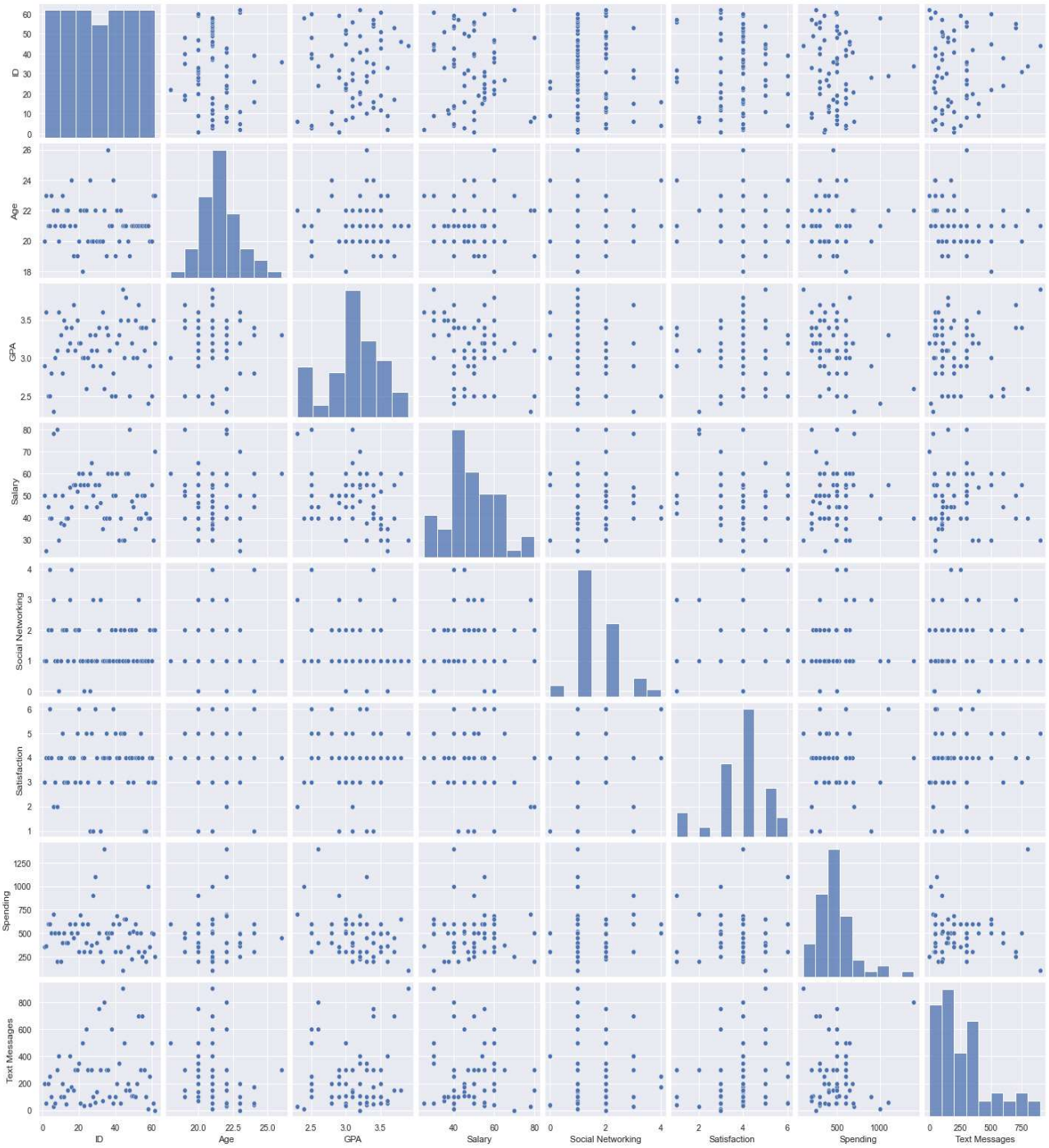
From the above results we can see that there is no missing value present in the dataset.

Correlation Plot

From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to- 0.2 or 1.0 are highly positively correlated. Correlation values near to 0.0 are not correlated to each other.

Pair Plot

There is no linear relationship between any 2 plots which shows that correlation is very weak .All the plots are scattered and showing no much significance.



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

Q2.1.1. Gender and Major-

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

Q 2.1.2. Gender and Grad Intention

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

Q2.1.3. Gender and Employment

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

Q2.1.4. Gender and Computer

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

Q2.2. Assume that the sample is a 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?

Solution -Probability of Male student randomly selected ,using formula= P(A)=m/n. I have two popution taking M is "Male" and B is Major. m=no of ways of accurrence of Male . n= no of total outcome .

Probability of randomly selected CMSU male student 0.46774193548387094

Q2.2.2 what is the probability that a randomly selected CMSU student will be female?

Solution -Probability of Male student randomly selected ,using formula=$P(A) = m/n$. I have two popution taking B is "Female" and C is Major. m=no of ways of accurrence of Female . n= no of total outcome .

Probability of randomly selected CMSU female student 0.532258064516129

Q2.3. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

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

Q2.3.1 Find the conditional probability of different majors among the male students in CMSU.

Solution - conditional Probability of Male different major ,using formula = $P(A|B) = P(A \text{ Int } B)/P(B)$. I have two popution taking A is "Major" and B is Male. P(A *Int* B)=intersection A and B . P(B)= m/n .

1. conditional probability of male for Management 0.4423305588585018
2. conditional probability of male for Retailing/Marketing 0.36860879904875155
3. conditional probability of male for Accounting 0.2948870392390012
4. conditional probability of male for Economics/Finance 0.2948870392390012
5. conditional probability of male for Other 0.2948870392390012
6. conditional probability of male for Undecided 0.2211652794292509
7. conditional probability of male for International Business 0.1474435196195006
8. conditional probability of male for CPI 0.0737217598097503

Q2.3.2 Find the conditional probability of different majors among the female students of CMSU.

Solution - conditional Probability of Female different major ,using formula = $P(A|B) = P(A \text{ Int } B)/P(B)$. I have two popution taking A is "Major" and B is Female. P(A *Int* B)=intersection A and B . P(B)= m/n .

1. conditional probability of female for Retailing/Marketing 0.5830721003134797
2. conditional probability of female for Economics/Finance 0.4535005224660397
3. conditional probability of female for International Business 0.25914315569487983
4. conditional probability of female for Management 0.25914315569487983
5. conditional probability of female for Accounting 0.1943573667711599
6. conditional probability of female for CIS 0.1943573667711599
7. conditional probability of female for Other 0.1943573667711599

Q2.4. Assume that the sample is a representative of the population of CMSU. Based on the dat a, answer the following question:

Q2.4.1 Find the probability that a randomly chosen student is a male and intends to graduate.

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

Solution -Probability of Male student intends to graduate ,using formula= P(A)=m/n. I have two popution taking "M" is "Male" and "B" is Yes(Graduate Intention). m=no of ways of accurrence of Male . n= no of total outcome .

P (Male & Intends to Graduate) = 17/62 = 0.27419354838709675

Q2.4.2 Find the probability that a randomly selected student is a female and does NOT have a l aptop.

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

Solution -Probability of Female student does NOT have a laptop,using formula= P(A)=m/n. I have two popution taking "F" is "Female" and "B" is NOT have a laptop (have=Desktop+Tablet). m=no of ways of accurrence of Female(have=Desktop+Tablet) . n= no of total outcome .

Probability (Female & Doesn't have laptop) = 4/62 = `0.06451612903225806`

Q2.5. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

Q2.5.1 Find the probability that a randomly chosen student is either a male or has a full-time employment

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

Solution -finding probability of all condition fast either a male , secound codition is male & female and thard codition is only male & full time. using formula $P(A \text{ UNI } B) = P(A) + P(B) - P(A \text{ INT } B)$

P (Male or Full-time Employment) = (29+10-7)/62 = 0.5161290322580645

Q2.6 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 graduate intention and being female are independent events?

Ans:-

Grad Intention ⟹ / Gender ⬇	No	Yes	All
Female	9	11	20
Male	3	17	20
All	12	28	40

Contingency table

After Removing the Undecided column,

P (Female) = 20/40 =0.5 #Unconditional Probability

P(Female | grad Intention) = 11/28 = `0.39285714285714285 # Conditional Probability`

Ratio of both probabilities = 0.5/0.39 = **1.282051282051282**

As we can see from the Ratio, that unconditional probability is 28% larger than Conditional probability Means there is noticeable difference between two probabilities. So we can say that these 2 events are not
Independent. Concluding these 2 events are dependent.

Q2.7 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	Gender	
2.3	Female	1
2.4	Female	1
2.5	Male	4
	Female	2
2.6	Male	2
2.8	Male	2
	Female	1
2.9	Female	3
	Male	1
3.0	Female	5
	Male	2

Solution -Probability of GPA<3.0 student in graduae ,using formula=$P(A) = m/n$. I have two popution taking A is GPA<3.0 and B is total student. m=no of ways of accurrence of GPA<3.0 student . n= no of total outcome student.

Probability of GPA<3.0 student less GPA =17/62=0.27419354838709675

Since there are 17 students whose GPA is less than 3 out of 62 students, so probability that his/her GPA is less than 3 is 27.4%.

Q2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
All	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Solution -Probability of salary>=50 ,using formula=$P(A) = m/n$. I have two popution taking A is salary>=50 and B is only Female & Male.

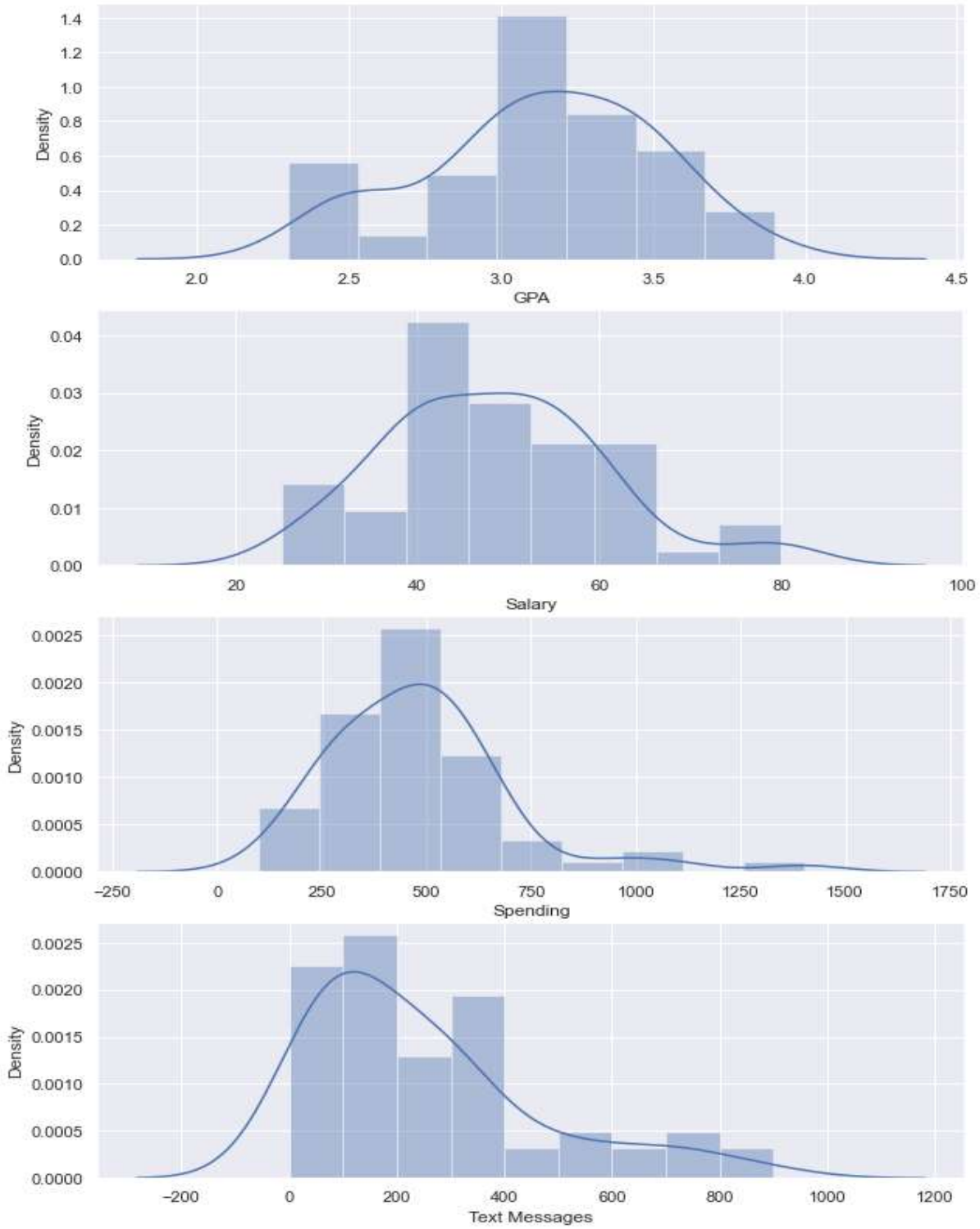
P (salary 50 or more | male) = 14/29 = 0.4827586206896552

1. There are 14 males who earns 50 or more out of 29 males. So required probability t hat a randomly selected male earns 50 or more is 48.2%.

P (salary 50 or more | Female) = 18/33 = 0.5454545454545454

2. Similarly there are 18 females who earns 50 or more out of 33, so required probabili ty that a randomly selected female earns 50 or more is 54.54%.

Q2.8.1 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.



As we see from the Graphs plotted above for all 4 continuous variables, none of them are having normal distribution.

Q2.8.2 Write a note summarizing your conclusions for this whole Problem 2.

Ans-
As per data analysis, there is a very weak correlation between all the parameters, which sho w that one parameter is not majorly affecting the other parameters. Also, the most famous major for females are Retailing/Marketing, and another likely famous major is Economics/Fi nance for Females. Whereas males are mostly interested in both management and retailing/ marketing. I also found that highest GPA 3.9 is attained by a female and also the lowest GPA is attained by a Female which is 2.3. Another interesting thing is that maximum spending is done by male majoring in Retailing/Marketing which is 1400. There are 4 continuous variabl es and rest are categorical variables. Salary is Male verses Female total of 50 to more emplo yer take.

PROBLEM – 3

Manufacturers of ABC asphalt shingles Data Analysis

Table of Contents

Contents

Executive Summary

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and using test of the hypothesis & ttest other parameter value.in this data seat two type of variable mention A & B.

Descriptive summary of the data

	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

Sample of the dataset

```
<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)
```

Missing values in the dataset

```
A    0
B    5
dtype: int64
```