SMDM PROJECT REPORT

DSBA

By – Pavan Kumar R Naik

Purpose

- This document is a business case study report for my project in the subject ‘Statistical Methods of Decision Making’
- Document gives a detailed explanation of various approaches used, their insights and inferences
- Tools used for analysis : Python and Jupiter Notebook
- Packages used: NumPy, Pandas, Seaborn, Matplotlib and SciPy

Contents

Problem 1 6

Business scenario 6

- 1.1 Use methods of descriptive statistics to summarize data 6
    - Which Region and which Channel spent the most? 8
    - Which Region and which Channel spent the least? 8
    - Visualization 8
    - Insights 8
- 1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? 9
    - Summary of all products grouped by channel and region 9
    - Box plot to find the trend of the products between channel and region 10
    - Insights 10
- 1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour? 11
    - Finding standard deviation and mean of the products 11
    - Plotting the standard deviation and mean of the products 11
    - Insights 11
- 1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 12
- 1.5 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 12

Contents

Problem 2 13

Business scenario 13

- 2.1. For this data, construct the following contingency tables 13
  - 2.1.1. Gender and Major 13
  - 2.1.2. Gender and Grad Intention 13
  - 2.1.3. Gender and Employment 13
  - 2.1.4. Gender and Computer 14
- 2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 14
  - 2.2.1. What is the probability that a randomly selected CMSU student will be male? 14
  - 2.2.2. What is the probability that a randomly selected CMSU student will be female? 14
- 2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 14
  - 2.3.1. Find the conditional probability of different majors among the male students in CMSU 14
  - 2.3.2 Find the conditional probability of different majors among the female students of CMSU 15
- 2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 15
  - 2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate 15
  - 2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop 16
- 2.5. Assume that the sample is representative of the population of CMSU 16
  - 2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment? 16
  - 2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management 16

Contents

- 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?..... 17
- 2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.... 17
  - 2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 17
  - 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..... 17
- 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..... 18
  - GPA..... 18
  - Salary..... 18
  - Spending..... 19
  - Text Messages..... 19
  - Insights.....19
- Problem 3..... 20

  Business scenario 20
- 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..... 20
- 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?..... 21

Problem 1 :

Business Scenario –

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).

1.1. Use methods of descriptive statistics to summarize data.

a). Dataset Head

	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

b). Dataset has any null values –

```
Do dataset has any na values?  False
```

c). Type of variables in Dataset

```
<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
```

d). Summary of a Dataset

	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	220.500000	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	127.161315	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	110.750000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	220.500000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	330.250000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	440.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

e).Preprocessing of Data

Removed the buyer and spender column as its required for the further analysis.(head after removing buyer and spender column)

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

Grouping the data by region and channel.

		Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Region	Channel						
Lisbon	Hotel	761233	228342	237542	184512	56081	70632
	Retail	93600	194112	332495	46514	148055	33695
Oporto	Hotel	326215	64519	123074	160861	13516	30965
	Retail	138506	174625	310200	29271	159795	23541
Other	Hotel	2928269	735753	820101	771606	165990	320358
	Retail	1032308	1153006	1675150	158886	724420	191752

Creating a total expenditure column by summing the values of all the products.

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

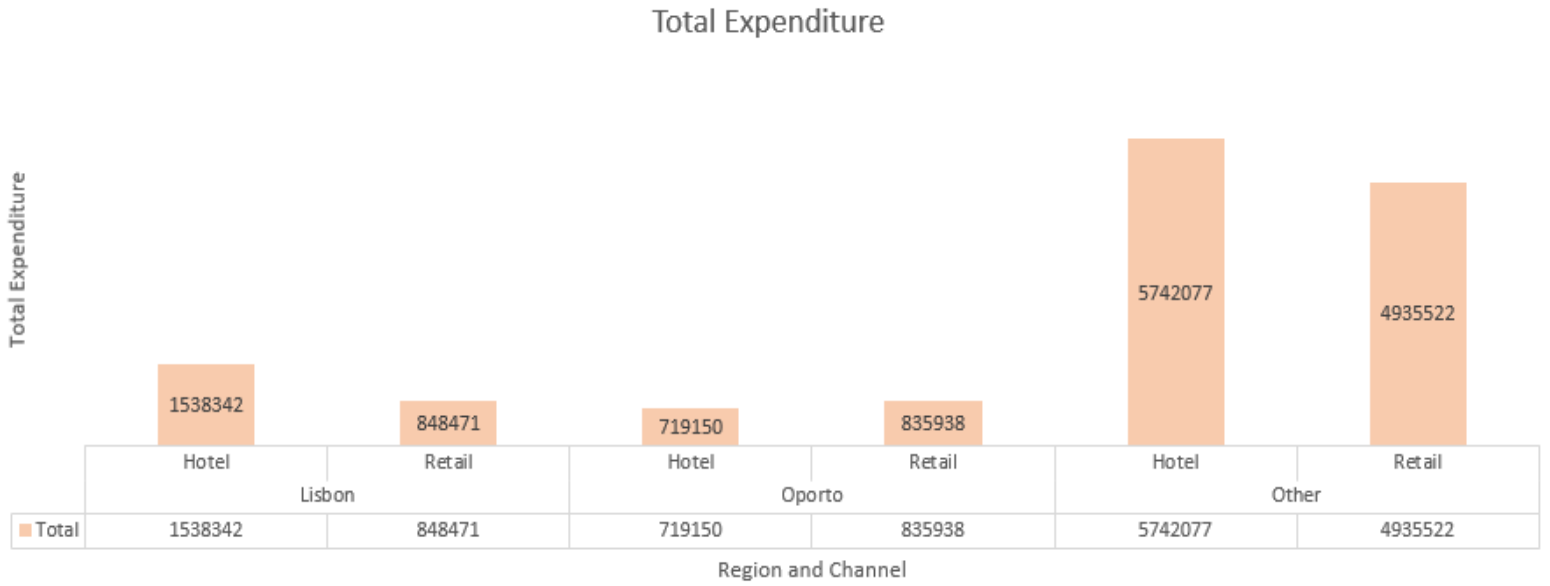
f). Which Region and which Channel spent the most?

```
Region  Channel
Other   Hotel      5742077
Name: Total_Expenditure, dtype: int64
```

g). Which Region and which Channel spent the least?

```
Region  Channel
Oporto  Hotel      719150
Name: Total_Expenditure, dtype: int64
```

- Visualization



h). Insights

From the above analysis we can identify that maximum amount spent/maximum revenue generated for selling the products is from the 'Hotel' channel which belongs to 'Other' region.

The Hotel from Opoto region spent the least amount.

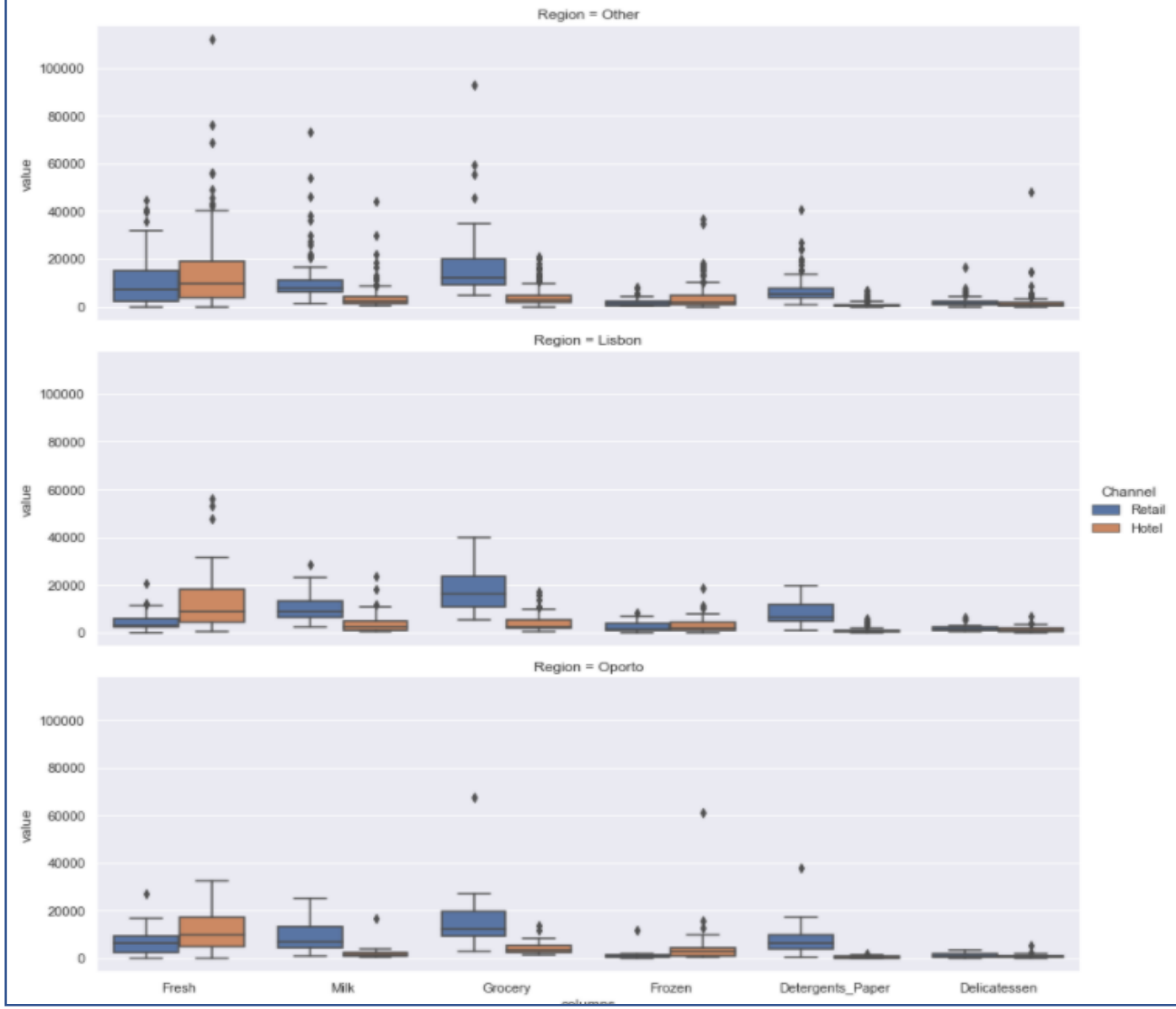
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel?

a)Summary of all products grouped by regions and channels

	Region	Lisbon		Oporto		Other	
	Channel	Hotel	Retail	Hotel	Retail	Hotel	Retail
Fresh	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	12902.254237	5200.000000	11650.535714	7289.789474	13878.052133	9831.504762
	std	12342.008901	5415.521495	8969.362752	6867.934548	14746.572913	9635.394129
	min	514.000000	18.000000	3.000000	161.000000	3.000000	23.000000
	25%	4437.500000	2378.250000	4938.250000	2368.000000	3702.500000	2343.000000
	50%	8656.000000	2926.000000	9787.000000	6468.000000	9612.000000	7362.000000
	75%	18135.000000	5988.000000	17031.500000	9162.000000	18821.000000	15076.000000
	max	56083.000000	20782.000000	32717.000000	27082.000000	112151.000000	44466.000000
Milk	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	3870.203390	10784.000000	2304.250000	9190.789474	3486.981043	10981.009524
	std	4298.321195	6609.221463	2968.628697	6611.354136	4508.505269	10574.827178
	min	258.000000	2527.000000	333.000000	928.000000	55.000000	1124.000000
	25%	1071.000000	6253.250000	1146.000000	4148.500000	1188.500000	6128.000000
	50%	2280.000000	8866.000000	1560.500000	6817.000000	2247.000000	7845.000000
	75%	4995.500000	13112.250000	2344.750000	13127.500000	4205.000000	11114.000000
	max	23527.000000	28326.000000	16784.000000	25071.000000	43950.000000	73498.000000
Grocery	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	4026.135593	18471.944444	4395.500000	16326.315789	3886.734597	15953.809524
	std	3629.644143	10414.687844	3048.298815	14035.453775	3593.506056	12298.935356
	min	489.000000	5265.000000	1330.000000	2743.000000	3.000000	4523.000000
	25%	1620.000000	10634.250000	2373.750000	9318.500000	1666.000000	9170.000000
	50%	2576.000000	16106.000000	3352.000000	12469.000000	2642.000000	12121.000000
	75%	5172.500000	23478.750000	5527.500000	19785.500000	4927.500000	19805.000000
	max	16966.000000	39694.000000	13626.000000	67298.000000	21042.000000	92780.000000

Frozen	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	3127.322034	2584.111111	5745.035714	1540.578947	3656.900474	1513.200000
	std	3276.460124	2424.774577	11454.478518	2473.266471	4956.590848	1504.498737
	min	91.000000	61.000000	264.000000	131.000000	25.000000	33.000000
	25%	966.000000	923.500000	962.250000	639.500000	779.000000	437.000000
	50%	1859.000000	1522.000000	2696.500000	934.000000	1960.000000	1059.000000
	75%	4479.000000	3843.000000	4617.000000	1410.000000	4542.500000	2194.000000
	max	18711.000000	8321.000000	60869.000000	11559.000000	36534.000000	8132.000000
Detergents_Paper	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	950.525424	8225.277778	482.714286	8410.263158	786.682464	6899.238095
	std	1305.907616	5515.878798	425.310506	8286.748255	1099.970640	6022.091110
	min	5.000000	788.000000	15.000000	332.000000	3.000000	523.000000
	25%	237.000000	4818.250000	182.750000	3900.000000	176.500000	3537.000000
	50%	412.000000	6177.000000	325.000000	6236.000000	375.000000	5121.000000
	75%	874.000000	11804.750000	707.000000	9837.500000	948.500000	7677.000000
	max	5828.000000	19410.000000	1679.000000	38102.000000	6907.000000	40827.000000
Delicatessen	count	59.000000	18.000000	28.000000	19.000000	211.000000	105.000000
	mean	1197.152542	1871.944444	1105.892857	1239.000000	1518.284360	1826.209524
	std	1219.945304	1626.486667	1056.778800	1065.438042	3663.183304	2119.052222
	min	7.000000	120.000000	51.000000	59.000000	3.000000	3.000000
	25%	374.000000	746.000000	567.250000	392.500000	378.500000	545.000000
	50%	749.000000	1414.000000	883.000000	1037.000000	823.000000	1386.000000
	75%	1621.500000	2456.500000	1146.000000	1815.000000	1582.000000	2158.000000
	max	6854.000000	6372.000000	5609.000000	3508.000000	47943.000000	16523.000000

b). CAT plot (type box) of all the products group by Region and Channel to find the trend of the products between Regions and Channels.



c). Insights

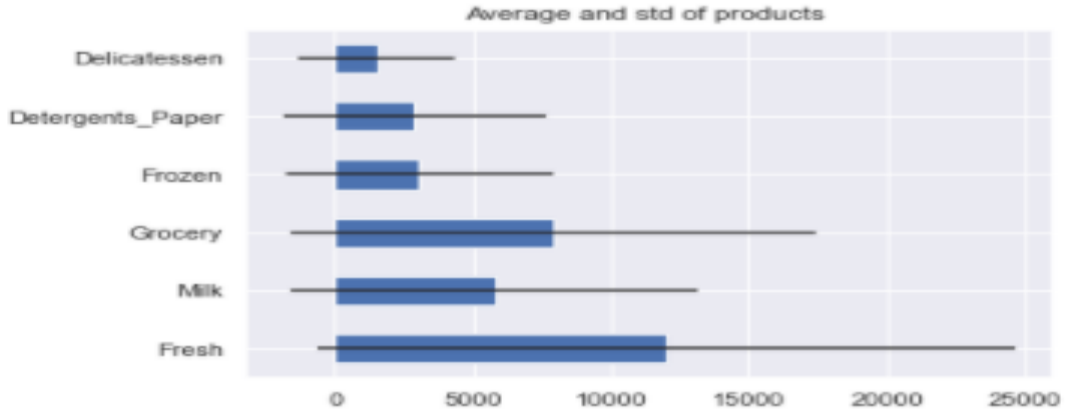
From the box plot we could find that the product varieties shows almost similar behavior across different Regions and Channels.

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

a). Finding standard deviation and mean of the products

	standard_devation	mean
Fresh	12647.328865	12000.297727
Milk	7380.377175	5796.265909
Grocery	9503.162829	7951.277273
Frozen	4854.673333	3071.931818
Detergents_Paper	4767.854448	2881.493182
Delicatessen	2820.105937	1524.870455

b). Plotting the standard deviation and mean of the products



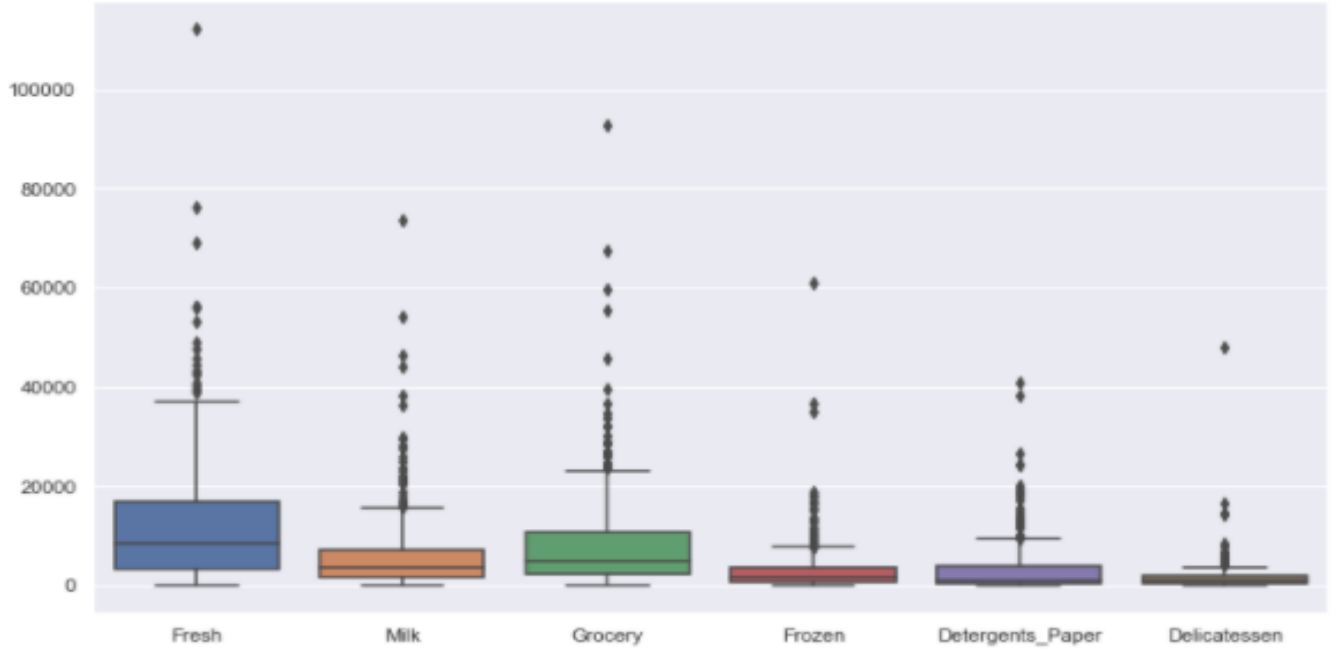
c). Insights

A high standard deviation shows that the data is widely spread and a low standard deviation shows that the data are clustered closely around the mean,

Based on the above analysis Fresh variety has most inconsistent behaviour and Delicatessen has least inconsistent behaviour.

1.4. Are there any outliers in the data?

a). Boxplot to visualise the outliers



b). Insights – Yes, we can find all the 6 products have outliers.

1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem?

- Based on the channel and region analysis, my recommendation to the wholesale distributor is to improve the sales in Oporto and Lisbon region.
- Based on the product analysis, we found that Fresh and Frozen products are sold more in hotels , grocery and milk are sold more in retails channel. So the distributor may use this info to optimize their marketing to boost the sales.
- Based on the regional behaviour analysis, this show similar behaviour in all the regions. Identified that outliers are more in Other region than in Oporto and Lisbon region.
- Based on the outlier analysis on the products, all the products have outliers and fresh items has an extreme outlier.

Problem 2:

Business Scenario -

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

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

2.1.1. 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

2.1.2. Gender and Grad Intention

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

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

2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following questions

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

The probability that a randomly selected CMSU student will be male = 29/62 = 0.46774193548387094 ie 47 %

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

The probability that a randomly selected CMSU student will be female =33/62 = 0.532258064516129 ie 53 %

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.

Gender vs 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

probability of of having Accounting as a major among the male students in CMSU is 4/29 = 0.13793103448275862 ie 14 %
probability of of having CIS as a major among the male students in CMSU is 1/29 = 0.034482758620689655 ie 3 %
probability of of having Economics/Finance as a major among the male students in CMSU is 4/29 = 0.13793103448275862 ie 14 %
probability of of having International Business as a major among the male students in CMSU is 2/29 = 0.06896551724137931 ie 7 %
probability of of having Management as a major among the male students in CMSU is 6/29 = 0.20689655172413793 ie 21 %
probability of of having Other as a major among the male students in CMSU is 4/29 = 0.13793103448275862 ie 14 %
probability of of having Retail/Marketing as a major among the male students in CMSU is 5/29= 0.1724137931034483 ie 17 %
probability of of having Undecided major among the male students in CMSU is 3/29 = 0.10344827586206896 ie 10 %

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

Gender vs 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

```
probability of of having Accounting as a  major among the female students in CMSU is 3/33=                0.09090909090909091 ie
9 %
probability of of having CIS as a  major among the female students in CMSU is 3/33 =                      0.09090909090909091 ie
9 %
probability of of having Economics/Finance as a  major among the female students in CMSU is 7/33=         0.21212121212121213 ie
21 %
probability of of having International Business as a  major among the female students in CMSU is 4/33 =   0.12121212121212122 ie
12 %
probability of of having Management as a  major among the female students in CMSU is  4/33 =              0.12121212121212122 ie
12 %
probability of of having Other as a  major among the female students in CMSU is  3/33 =                   0.09090909090909091 ie
9 %
probability of of having Retail/Marketing  as a  major among the female students in CMSU is 9/33 =        0.2727272727272727 ie
27 %
probability of of having Undecided major among the female students in CMSU is 0/33=                       0.0 ie 0 %
```

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.

Gender vs intent to graduate

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

```
probability of intent to graduate, given that the student is a male is 17/33 = 0.5151515151515151 ie 52 %
```

2.4.1.Find the conditional probability of intent to graduate, given that the student is a female.

probability of intent to graduate, given that the student is a female is 11/33 = 0.3333333333333333 ie 33 %

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

Gender vs Laptop

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

probability of students are female and does not prefer a laptop is (1-29)/33 0.12121212121212122 ie 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?

Gender vs Employment

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

probability that a randomly chosen student is either a male or has full-time employment is 29/62+3/62 = 0.5161290322580645 ie 52 %

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

Gender vs Major

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
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

probability that given a female student is randomly chosen, she is majoring in international business or management is 4/33 + 4/33 = 0.24242424242424243 ie 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?

Gender vs Intent

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

```
gender vs grad Intention
dof=1
[[ 6. 14.]
 [ 6. 14.]]
critical= 3.841458820694124
Stat= 2.9761904761904767
Independent (fail to reject H0)
```

- Based on the Chi Square test analysis graduate intention and being female are two independent events.

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

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

Gender vs GPA

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
Gender																
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0

probability that a randomly chosen student is having GPA less than 3 is 1-(45/62) = 0.27419354838709675 ie 27 %

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.

probability of a randomly selected male earns 50 or more is 14/29 = 0.4827586206896552 ie 48 %

probability of a randomly selected female earns 50 or more is 18/33 = 0.5454545454545454 ie 55 %

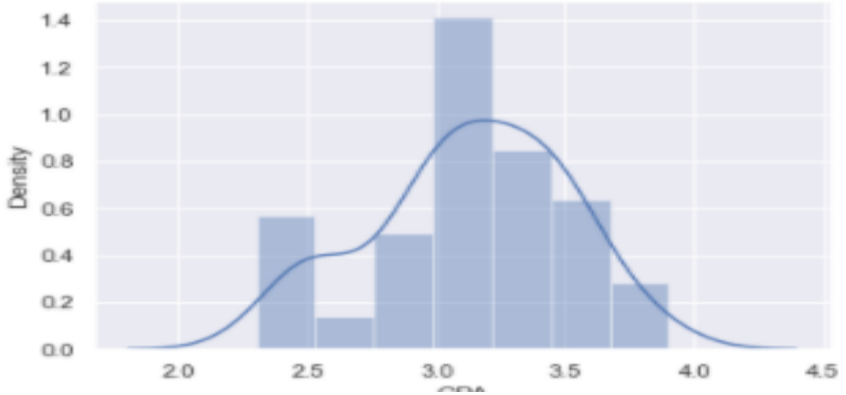
2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions for this whole Problem 2.

- **H0 = Distribution is normal**
- **Ha = Distribution is not normal**
- **p <= alpha: reject H0,**
- **p > alpha: fail to reject H0**
- **Alpha = 0.05**

GPA

```
stat, p = normaltest(df2['GPA'])
print('Statistics=%.3f, p=%.3f' % (stat, p))

if p > alpha:
      print('We have enough evidence that distribution is normal (fail to reject H0)')
else:
    print('We have enough evidence that distribution is not normal (p< alpha reject H0)')
```

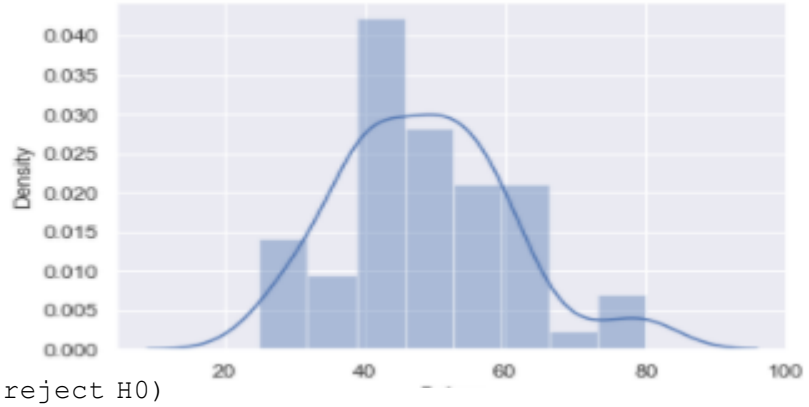



```
Statistics=1.953, p=0.377 We have enough evidence that distribution is normal (fail to reject H0)
```

Salary

```
stat, p = normaltest(df2['Salary'])

print('Statistics=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('We have enough evidence that distribution is normal (fail to reject H0)')
else:
    print('We have enough evidence that distribution is not normal (p< alpha reject H0)')
```



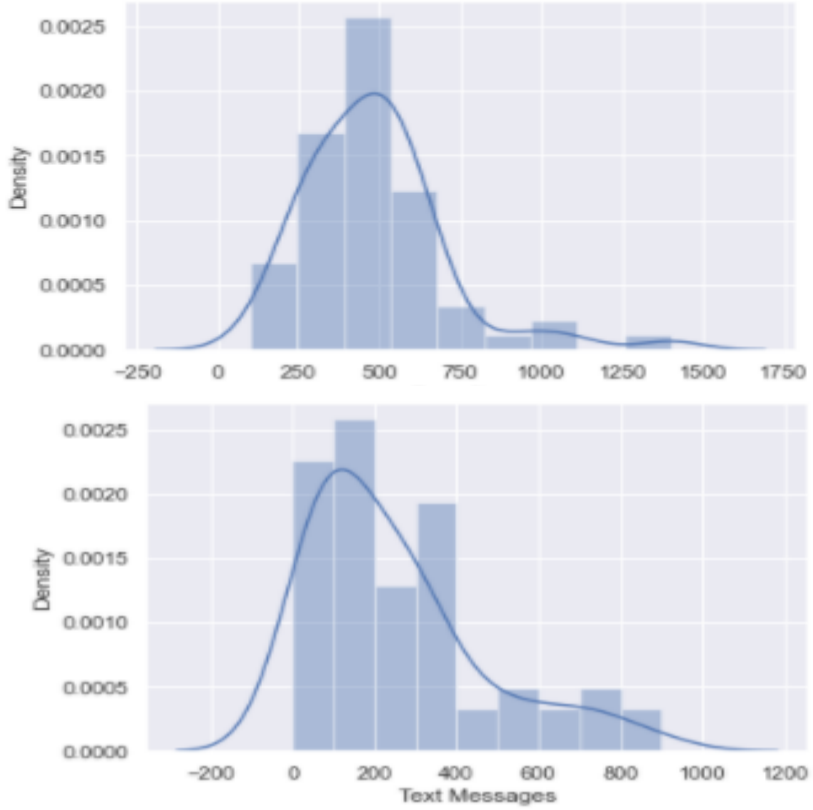

```
Statistics=3.846, p=0.146 We have enough evidence that distribution is normal (fail to reject H0)
```

Spending

```
stat, p = normaltest(df2['Spending'])

print('Statistics=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('We have enough evidence that distribution is normal (fail to reject H0)')
else:
    print('We have enough evidence that distribution is not normal (p< alpha reject H0)')
```

```
Statistics=30.496, p=0.000
We have enough evidence that distribution is not normal (p< alpha reject H0)
```



Text Messages

```
stat, p = normaltest(df2['Text Messages'])

print('Statistics=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('We have enough evidence that distribution is normal (fail to reject H0)')
else:
    print('We have enough evidence that distribution is not normal (p< alpha reject H0)')
```

```
Statistics=16.348, p=0.000
We have enough evidence that distribution is not normal (p< alpha reject H0)
```

Insights:

From the above analysis we identified that Text messages and Spending don't follow normal distribution

Where as GPA and Salary follow normal distribution.

Problem 3:

Business Scenario –

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.

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.

Ho = Mean moisture contents of A and B are not with in limits

Ha = Mean moisture contents of A and B are with in limits

Alpha = 0.35

```
t_statistic,p_value = ttest_ind(df3['A'],df3.dropna()['B'])
t_statistic,p_value
```

```
T Statistic = 1.289628271966112
P Value =  0.2017496571835328
```

```
if p_value < alpha:
    print('We have enough evidence to reject the null hypothesis in favour of alternative hypothesis')
    print('We conclude that the  mean moisture content of A and B are with in limits')
else:
    print('We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis (p > alpha)')
    print('We conclude that the  mean moisture content of A and B are not with in limits.')
```

We have enough evidence to reject the null hypothesis in favor of alternative hypothesis We conclude that the mean moisture content of A and B are with in limits.

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?

Hypothesis;

H0 :S1 = S2

Ha: S1 is not equal to S2

Alpha =0.05

Assumption: We assume that both the populations are normally distributed and samples are random and have unequal variance.

```
tstat 1.289628271966112 P Value 0.2017496571835328
We do not have enough evidence to reject the null hypothesis in favor of alternative hypothesis (p > alpha)
We conclude that the means for shingles A and B are equal.
```

For the above T test, samples must be random and normally distributed and Variance of the population is unknown.

Thank you,

Pavan Kumar R Naik