SMDM PROJECT REPORT

By Karthik Sreeram R

AUGUST 2, 2020

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose

This document is the business report for my final project in the subject “Statistical Methods for decision making “.

This document gives us a detailed explanation of various approaches used, their insight and inferences.

Tools used analysis: Python and Jupiter notebook.
Packages used: NumPy, pandas, seaborn, os, matplotlib, SciPy and sweetviz

Contents

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 (Wholesale Customer.csv) 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.

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

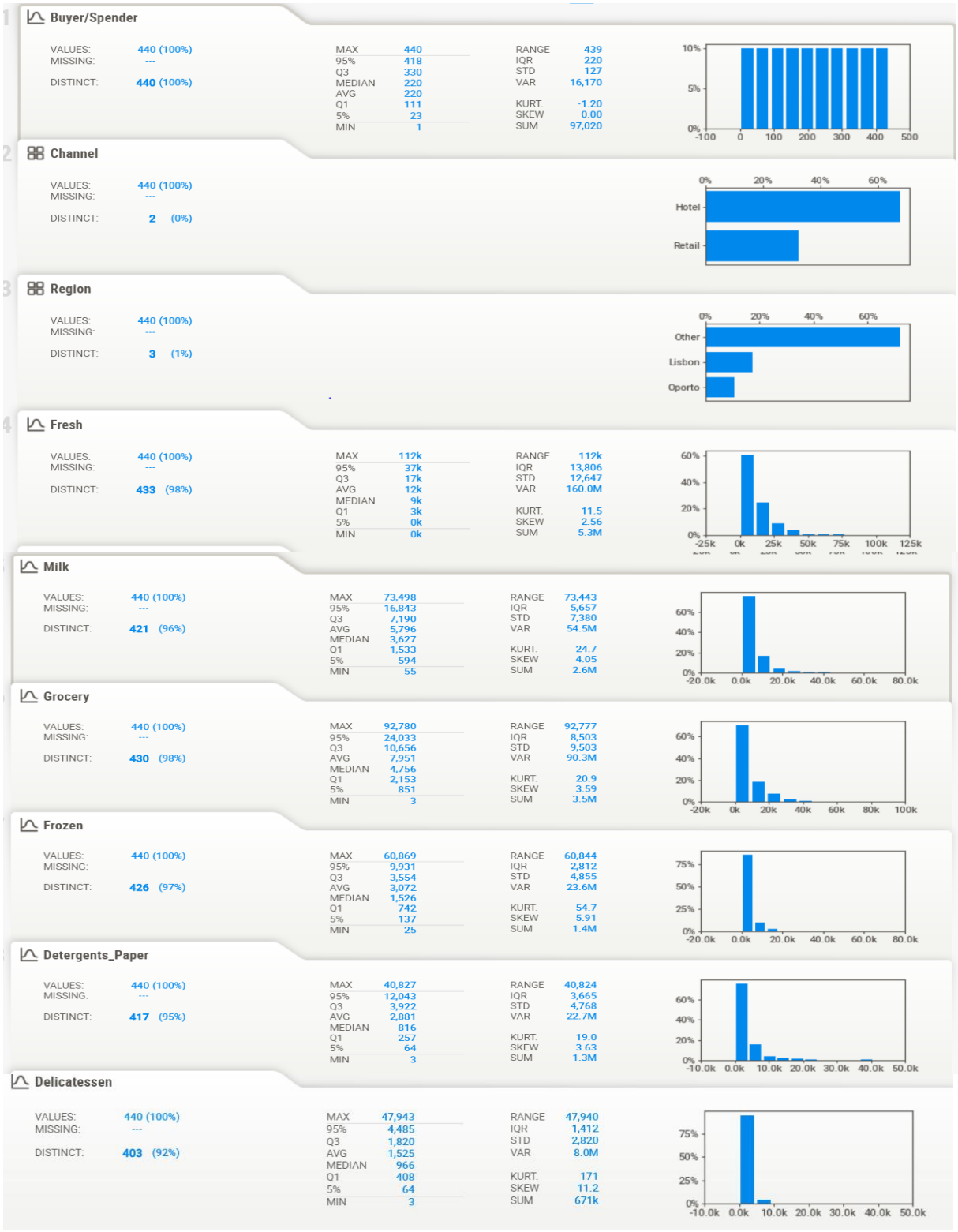
c.) Type of the variables in dataset

```
<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 the 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.) EDA using sweet viz to visualize the summary for each variable as well to underrated the data



f.) Preprocessing of the data

Removed the buyer/spender column as it is not valuable for the further analysis . (printed the head after removing the buyer column

[8]:

	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 (as it is required for this question)

6]:

		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.

[7]:

		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

1.1.1. Which Region and which Channel seems to spend more?

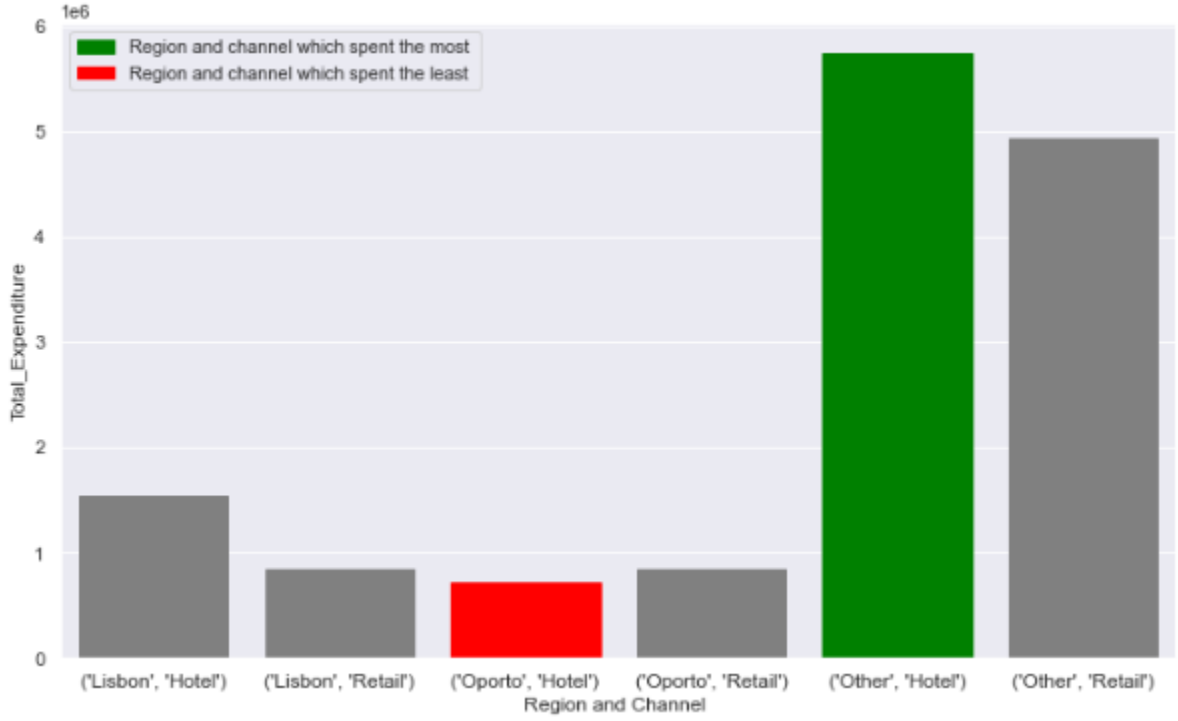
a.) Which Region and which Channel seems to spend more?

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

b.) Which Region and which Channel seems to spend less?

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

c.) Visualization



d.) Insights:

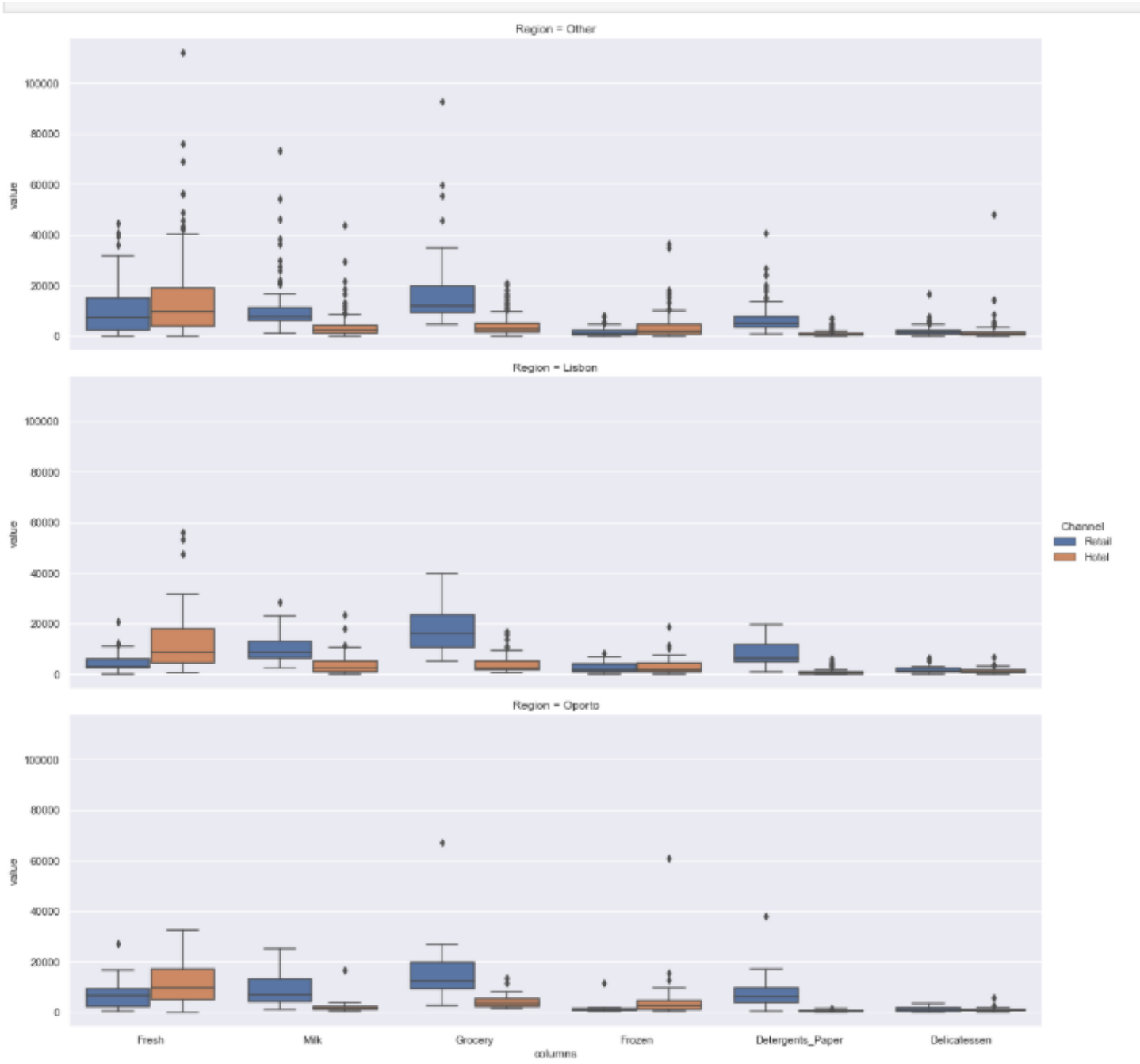
From the above analysis we find out that the maximum amount of revenue for selling the products is from the "Hotels "channel which belong to "Other" region i.e. they spent the most.
The Hotels from Opoto region spent the least amount of money.

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

a.) Summary of all products grouped by Regions and channels

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
gents_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 grouped by region and Channel to find the trend of the products between regions and channels .



c.) Insights

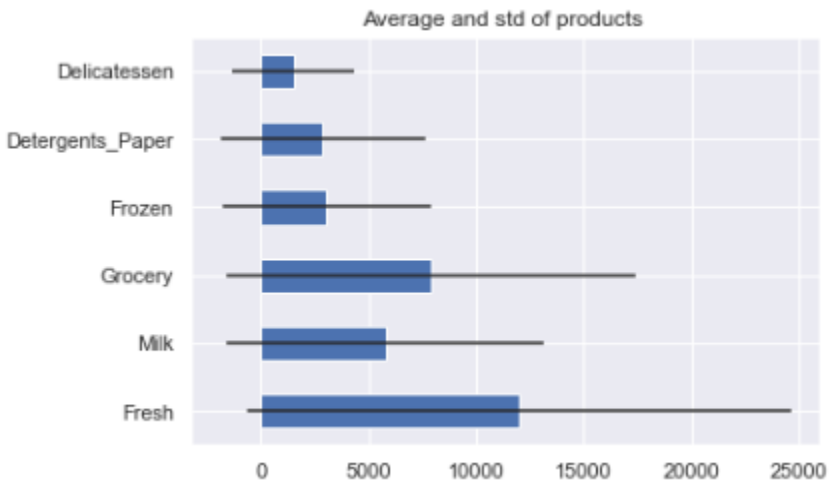
From The above graph we can find out the product varieties shows almost similar behavior across different Regions and Channels.

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

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

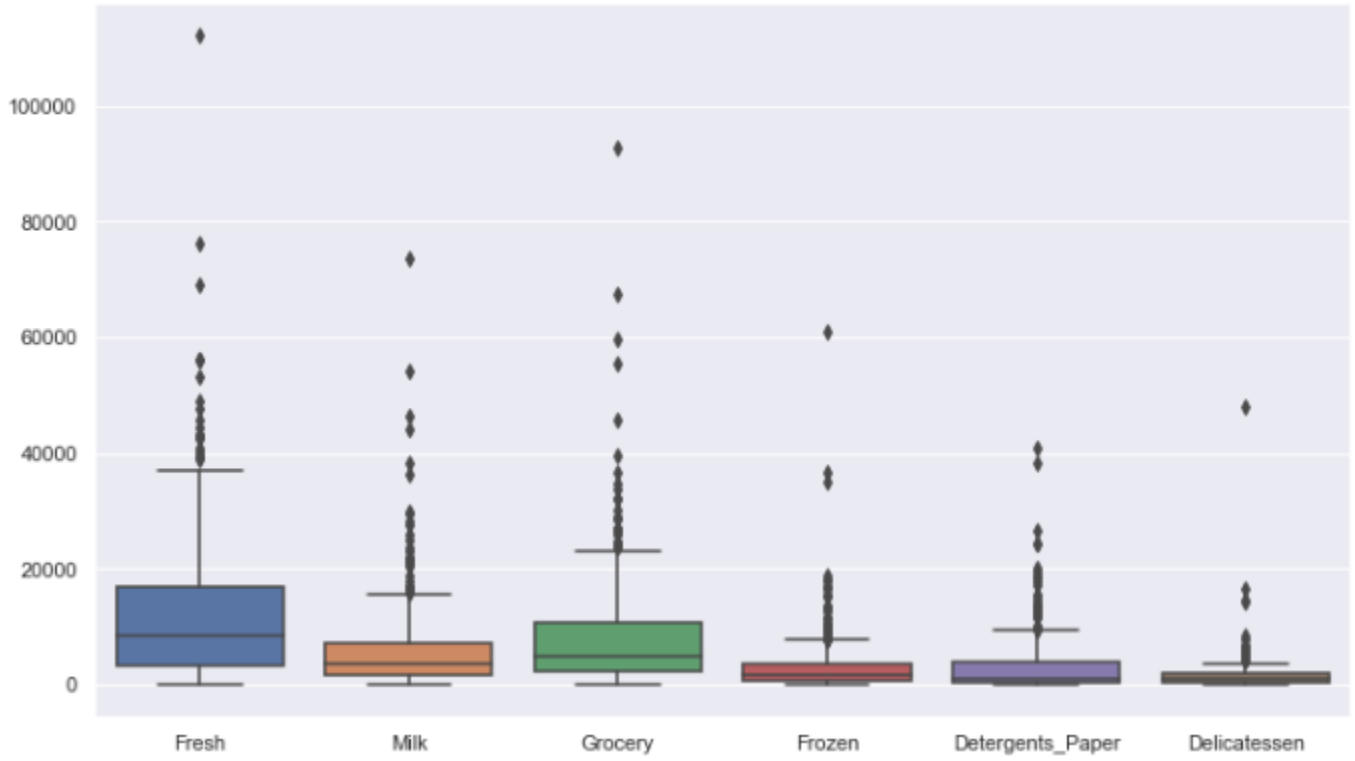


c.) Insights:

A high standard deviation shows that the data is widely spread (less reliable) 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 behavior and Delicateseen has least inconsistent behavior.

1.4. Are there any outliers in the data?

a.) Box plot to visualize the outliers



b.) Insights

yes, there are outliners in the data. From the above analysis we can find that all the 6 products have outliers

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

Based on a 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 can find out that Fresh and Frozen products are sold more to the Hotels and Grocery and milk are sold more to the retail channel. So, the distributor may use this information to optimize their marketing techniques and it may boost the sales.

Based on the reginal behavior analysis, they show similar behavior on the all the region. But There is more outlier in the Other region than in the Oporto and Lisbon region. Require further data collection and analysis on these outliers to find the reason.

Based on the outlier analysis on the products, all products have outliers. But 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 (stored in the Survey.csv file)

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

2.1.1. Gender and 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

2.1.2. Gender and Grad Intention

[49]:

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

2.1.3. Gender and Employment

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

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet	All
Gender				
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?

a.) Will be Male.

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

b.) will be female?

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

2.2.2. Find the conditional probability of different majors among the students in CMSU.

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

a.) Among Male Students

```
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 %
```

b.) Among the Female students

```
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 %
```

2.2.3. Find the conditional probability of intent to graduate

Gender vs Intent to graduate

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

a.) Given that the student is a male.

```
probability of intent to graduate, given that the student is a male is 17/29 =   0.5862068965517241 ie 59 %
```

b.) 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 %
```

2.2.4. Find the conditional probability of employment status

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

a.) For the male students

```
The probability of employment status for the male students
 probability of Fulltime  employment status for the male students  is 7/29 =                                                0.2413793103448276 ie 24 %
 probability of part-time  employment status for the male students  is  19/29 =                                          0.6551724137931034 ie 66 %
 probability of Unemployed status for the male students  is 3/29=                                                        0.10344827586206896 ie 10 %
 probability of some type of (either Fulltime or part time ) employment status for the male students is  (19+7)/29 =  0.896551724137931 ie 90 %
```

b.) For the female students

```
The probability of employment status for the female students
 probability of Fulltime  employment status for the female students  is 3/33 =                                          0.09090909090909091 ie 9 %
 probability of part-time  employment status for the female students  is 24/33 =                                         0.7272727272727273 ie 73 %
 probability of Unemployed status for the female students  is 6/33 =                                                     0.18181818181818182 ie 18 %
 probability of some type of (either Fulltime or part time ) employment status for the female students  is (24+3)/33 =  0.8181818181818182 ie 93 %
```

2.2.5. Find the conditional probability of laptop preference

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

a.) Among the male students

```
probability of laptop preference among the male students  is 26/29 =   0.896551724137931 ie 90 %
```

b.) Among the female students

```
probability of laptop preference among the female students  is 29/33  0.8787878787878788 ie 88 %
```

2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender? Justify your comment in each case.

If Statistic >= Critical Value: significant result, reject null hypothesis (accept Ha), dependent.

If Statistic < Critical Value: not significant result, fail to reject null hypothesis (aceept H0), independent.

2.3.1 Cases:

a.) Case 1: Gender and Major

Observed

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

Expected

```
 gender vs major
dof=7
[[3.72580645 2.12903226 5.85483871 3.19354839 5.32258065 3.72580645
  7.4516129  1.59677419]
 [3.27419355 1.87096774 5.14516129 2.80645161 4.67741935 3.27419355
  6.5483871  1.40322581]]
critical= 14.067140449340169
Stat= 7.084844866036089
Independent (fail to reject H0)
```

b.) Case 2: Gender and Grad Intention

Observed

gender and Grad grad Intention

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

Expected

```
 gender vs grad Intention
dof=2
[[ 6.38709677 11.70967742 14.90322581]
 [ 5.61290323 10.29032258 13.09677419]]
critical= 5.991464547107979
Stat= 4.774796781066374
Independent (fail to reject H0)
```

c.) Case 3: Gender and Employment status

Observed

gender vs Employement status

]:

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

```
 gender vs Employement status
dof=2
Expected [[ 5.32258065 22.88709677  4.79032258]
 [ 4.67741935 20.11290323  4.20967742]]
critical= 5.991464547107979
Stat= 2.9355495613715337
Independent (fail to reject H0)
```

d.) Case 4: Gender and computer

Observed :

gender vs Computer

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

```
 gender vs computer
dof=2
Expected [[ 2.66129032 29.27419355  1.06451613]
 [ 2.33870968 25.72580645  0.93548387]]
critical= 5.991464547107979
Stat= 2.114372565783224
Independent (fail to reject H0)
```

2.3.2 Insights:

The critical value is calculated and interpreted for all the cases , finding that indeed the column variable in each case is independent of Gender (fail to reject H0) .

2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions. [Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

p <= alpha: reject H0, not normal.

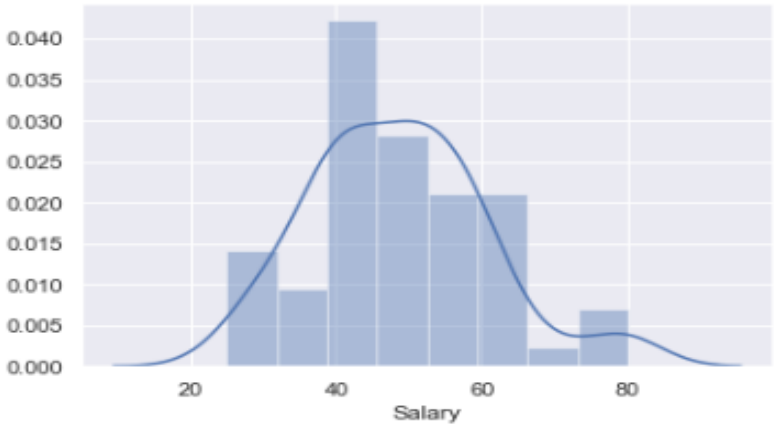
p > alpha: fail to reject H0, normal.

```
alpha = 0.05
```

a.) Salary

```
shapiro test for the variable Salary
Statistics=0.957, p=0.028
Sample does not look Gaussian (p< alpha reject H0)
```

```
Histogram (distrubution plot for varaible salary  )
<matplotlib.axes._subplots.AxesSubplot at 0x137f8b6b670>
```



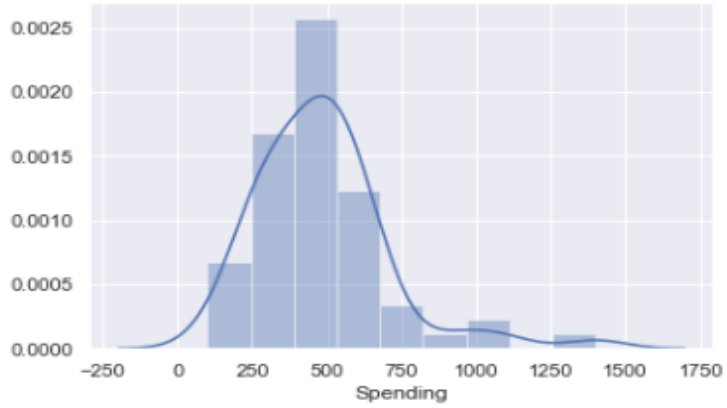
It looks like normal distribution (gaussian) so to confirm I will do D'Agostino's K^2 test.

```
D'Agostino's K^2 Test for variable salary
Statistics=3.846, p=0.146
Sample looks Gaussian (fail to reject H0)
```

b.) Spending

```
shapiro test for the variable Spending
Statistics=0.878, p=0.000
Sample does not look Gaussian (reject H0)
```

```
Histogram (distrubution plot for varaible spending )
<matplotlib.axes._subplots.AxesSubplot at 0x137f92470d0>
```

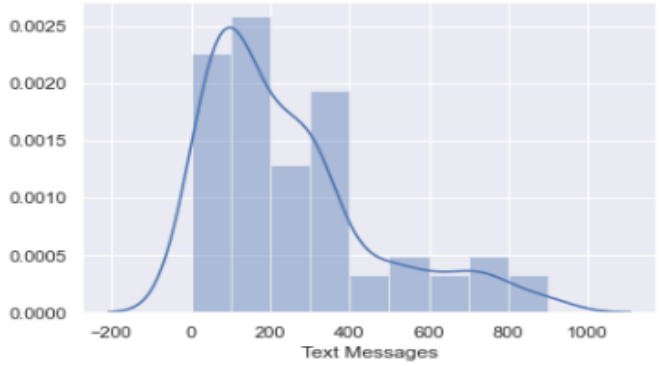


```
D'Agostino's K^2 Test for variable spending
Statistics=30.496, p=0.000
Sample does not look Gaussian (reject H0)
```

c.) Text Messages

```
shapiro test for the variable Text Messages
Statistics=0.859, p=0.000
Sample does not look Gaussian (reject H0)
```

```
Histogram (distrubution plot for varaible Text Messages )
<matplotlib.axes._subplots.AxesSubplot at 0x137f4ac7730>
```




```
D'Agostino's K^2 Test for variable Text Messages
Statistics=16.348, p=0.000
Sample does not look Gaussian (reject H0)
```

2.4.1 Insights:

From the above analysis, we can find that text messages and spending do not follow normal distribution. Salary even though it does not look like gaussian in the shaprio test, In accordance with D'Agostino k squared test and distribution plot, it follows 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 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 claims that the mean moisture content cannot be greater 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.

For the A shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:
$H0 \leq 0.35$
$H1 > 0.35$

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:
$H0 \leq 0.35$
$H1 > 0.35$

3.1 Do you think that the population means 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
H1: S1 is not equal to S2

Assumption: we assume that both the populations are normally distributed and samples are random (as both the sample A and B are above 30 picked randomly)also we assumed unequal variances of the populations. (population variance unknown

```
alpha = 0.05


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

3.2 What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

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