SMDM
PROJECT REPORT

By

M.P.KARTHIKEYAN

Contents Problem

List of Figures

List of Tables

Problem 1

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).

Executive summary

A wholesale distributor has data about 440 large retailers operating in 3 different regions(Lisbon, Oporto, Other) across two different sales channels(Hotel, Retail) and their spending on 6 different varieties of products. With this dataset we must analyse spending pattern of different retailers on different products.

Introduction

The given dataset contains details about 440 retailers spending on 6 different items across 2 different channels.Exploratory Data analysis is done.The most & least preferred product variety,the consistency of the product varieties,the preferred sales channel are to be found.The distribution of product varieties across regions and channels are to be calculated.Descriptive statistics and measure of central tendency are calculated

Sample Dataset

Table-1 Sample 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

Exploratory Data Analysis

Let us check the type of variables

Buyer/Spender	int64
Channel	object
Region	object
Fresh	int64
Milk	int64
Grocery	int64
Frozen	int64
Detergents_Paper	int64
Delicatessen	int64

The dataset contains 440 rows and 9 columns.Out of 9 columns 2 columns are Object type and 7 columns are integer type.

Check for missing values in dataset

Buyer/Spender	440 non null int64
Channel	440 non null object

Region	440 non null object
Fresh	440 non null int64
Milk	440 non null int64
Grocery	440 non null int64
Frozen	440 non null int64
Detergents_Paper	440 non null int64
Delicatessen	440 non null int64

From the above values it is clear that there are no missing values in dataset.

1.1Use methods of descriptive statistics to summarize data.

Descriptive statistics are used to describe about the variables in dataset by giving short summaries about the sample and the measures of data.

The most recognized types of descriptive statistics are measures of centre**: the mean, median, and mode**, which are used at almost all levels of math and statistics.The Total Spent of all the retailers are calculated respectively from the dataset.

Table1.1-Summary of data

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total spent
count	440.000000	440	440	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
unique	NaN	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Hotel	Other	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	298	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	220.500000	NaN	NaN	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455	33226.136364
std	127.161315	NaN	NaN	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937	26356.301730
min	1.000000	NaN	NaN	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	904.000000
25%	110.750000	NaN	NaN	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000	17448.750000
50%	220.500000	NaN	NaN	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000	27492.000000
75%	330.250000	NaN	NaN	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000	41307.500000
max	440.000000	NaN	NaN	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000	199891.000000

From the table we can see that most no. of purchases(316 out of 400) were made in ‘Other’ region .Also we can see that 298 out of 400 purchases are made through ‘Hotel’ making it the most preferred channel.On an average a retailer spend 12000.297 on Fresh , 5796.265 on Milk , 7951.277 on Grocery, 3071.932 on Frozen ,2881.493 on Detergent Paper,1524.870 on Delicatessen.

NaN Values are present in some variables as the measures of centre can’t be calculated.

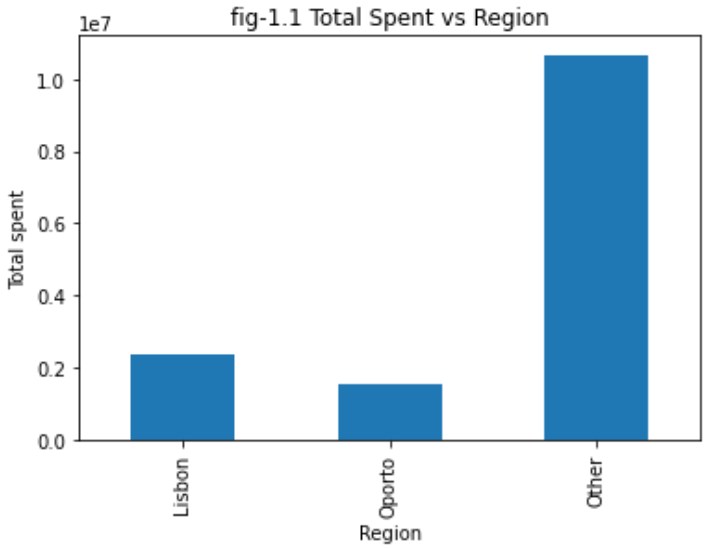
Calculating Total spent by different Regions & Channels

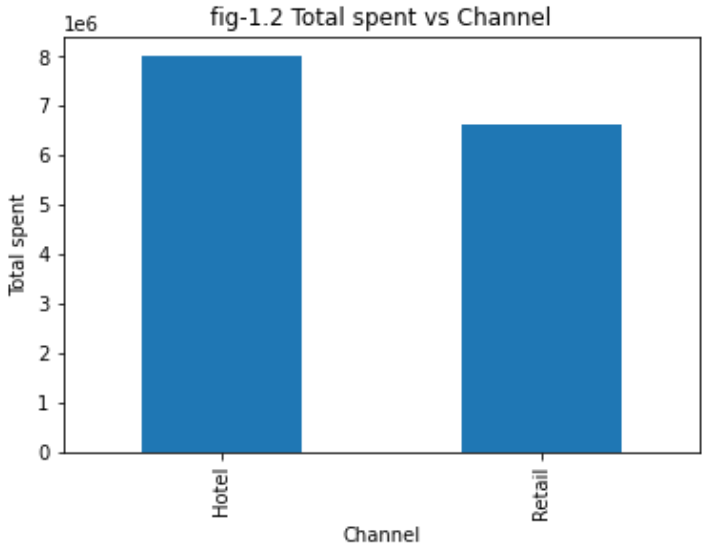
Region	**Total spent**
1.Lisbon	= 2386813

2.Oporto = 1555088
3.Other = 10677599

Channel **Total spent**

1.Hotel = 7999569
2.Retail = 6619931





1.1.1 Which Region and which Channel spent the most?

Thus it is evident from the plots (fig 1.1 & 1.2) that '**Other**' region spent the most and the most spent sales channel is '**Hotel**'

1.1.2 Which Region and which Channel spent the least?

'**Oporto**' region is the least spent region and the least spent channel is '**Retail'.**

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.

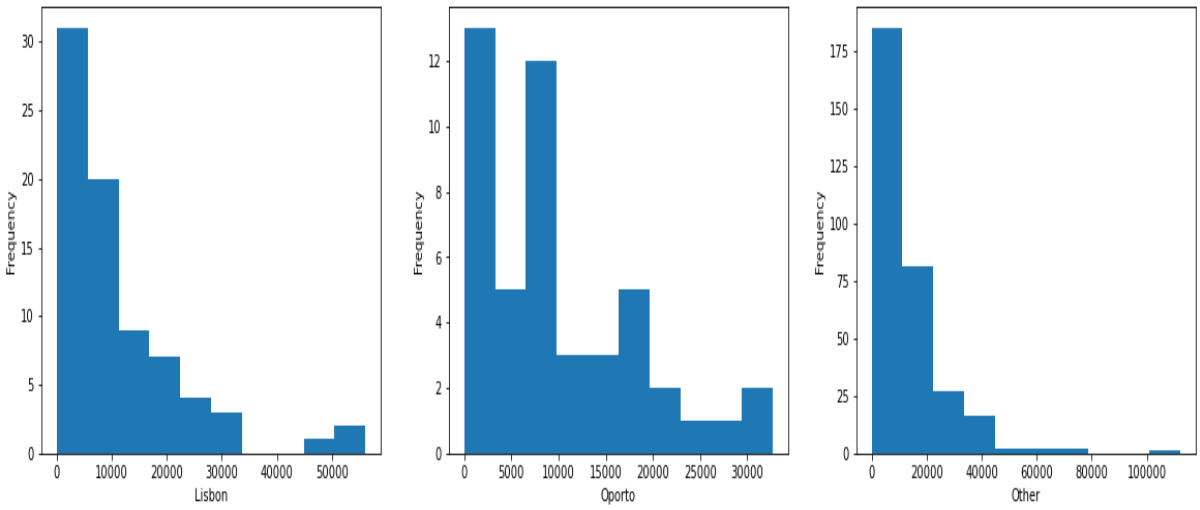
Let's calculate the spread of six different varieties across all the regions and channels

1.2.1Calculating the distribution of Fresh variety across different Regions and Channels.

Table1.2-Summary of Fresh across Regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	3.160000e+02
mean	11101.727273	9887.680851	1.253347e+04
std	11557.438575	8387.899211	1.338921e+04
min	18.000000	3.000000	3.000000e+00
25%	2806.000000	2751.500000	3.350750e+03
50%	7363.000000	8090.000000	8.752500e+03
75%	15218.000000	14925.500000	1.740650e+04
max	56083.000000	32717.000000	1.121510e+05
Total	854833.000000	464721.000000	3.960577e+06
CV	1.041049	0.848318	1.068277e+00

fig-1.3-Distribution of Fresh across regions



From the descriptive statistics we can calculate the spread of Fresh across different regions.On an average a retailer from Lisbon spends 11101.727 on fresh variety.Average

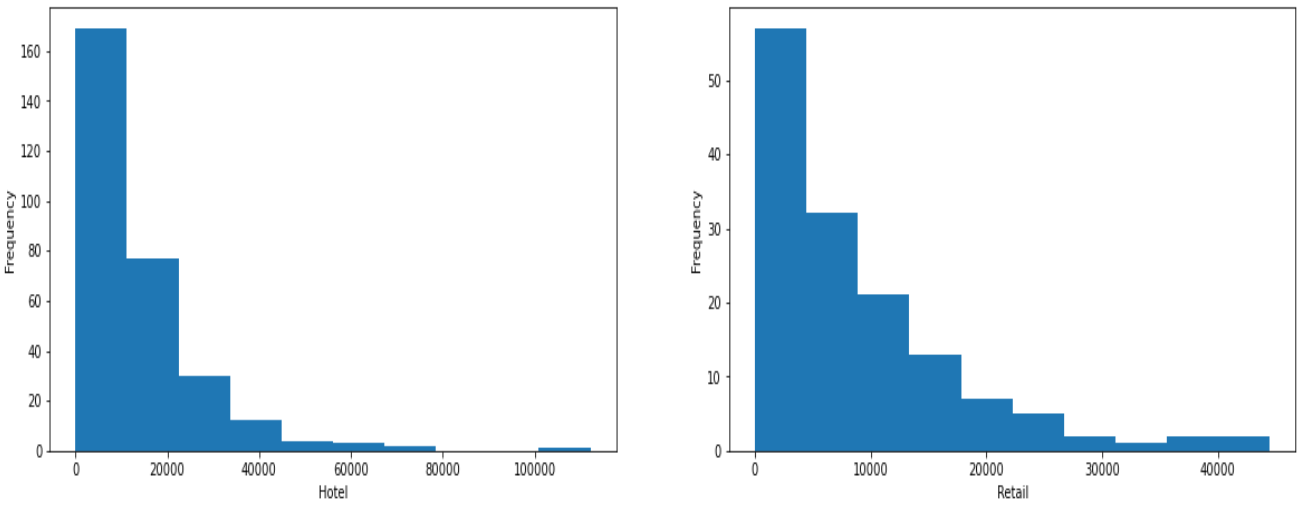
spend of retailer on fresh variety from Oporto is 9887.680 and that of retailer from Other region is 12533.47.

From the plot we can state that the distribution of Fresh across regions is right skewed.

Table1.3-Summary of Fresh across Channels

Channel	Hotel	Retail
count	2.980000e+02	1.420000e+02
mean	1.347556e+04	8.904324e+03
std	1.383169e+04	8.987715e+03
min	3.000000e+00	1.800000e+01
25%	4.070250e+03	2.347750e+03
50%	9.581500e+03	5.993500e+03
75%	1.827475e+04	1.222975e+04
max	1.121510e+05	4.446600e+04
Total	4.015717e+06	1.264414e+06
CV	1.026428e+00	1.009365e+00

fig -1.4 Distribution of Fresh across Channels



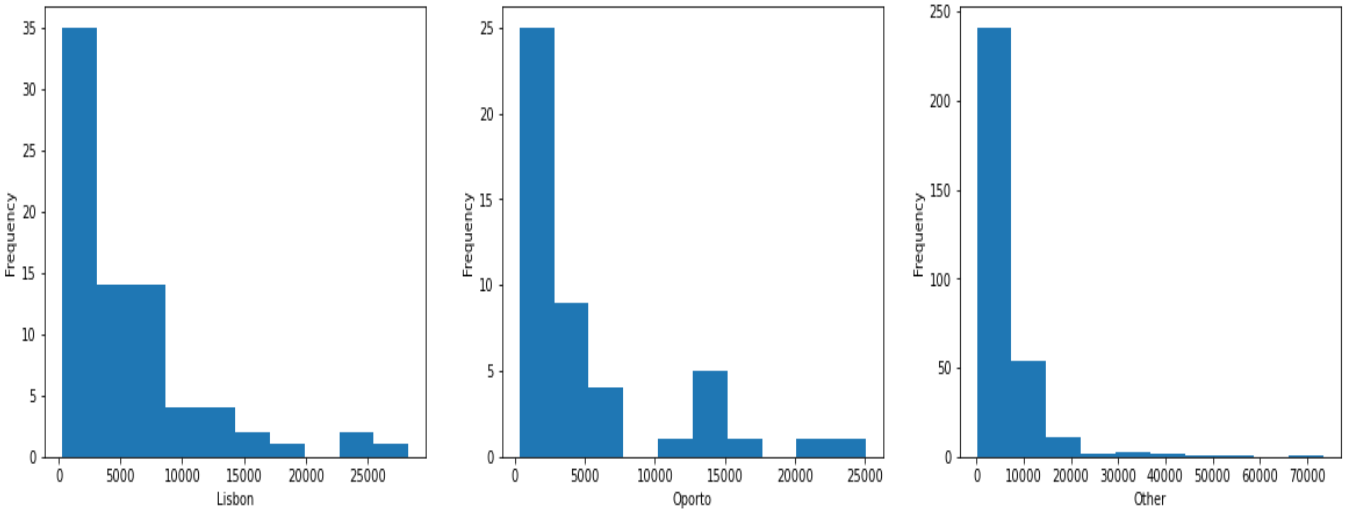
From the descriptive statistics (Table -1.3) we can calculate the spread of Fresh across different Channels.On an average a retailer spends 13475.5 through ‘Hotel’on fresh variety.Average spend of retailer on fresh variety through ‘Retail’ channel is 8904.32
From the plot we can state that the distribution of Fresh across Channels is right skewed

1.2.2Calculating the distribution of Milk variety across different Regions and Channels

Table1.4-Summary of Milk across regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	3.160000e+02
mean	5486.415584	5088.170213	5.977085e+03
std	5704.856079	5826.343145	7.935463e+03
min	258.000000	333.000000	5.500000e+01
25%	1372.000000	1430.500000	1.634000e+03
50%	3748.000000	2374.000000	3.684500e+03
75%	7503.000000	5772.500000	7.198750e+03
max	28326.000000	25071.000000	7.349800e+04
Total	422454.000000	239144.000000	1.888759e+06
CV	1.039815	1.145076	1.327648e+00

fig-1.5-Distribution of Milk across Regions



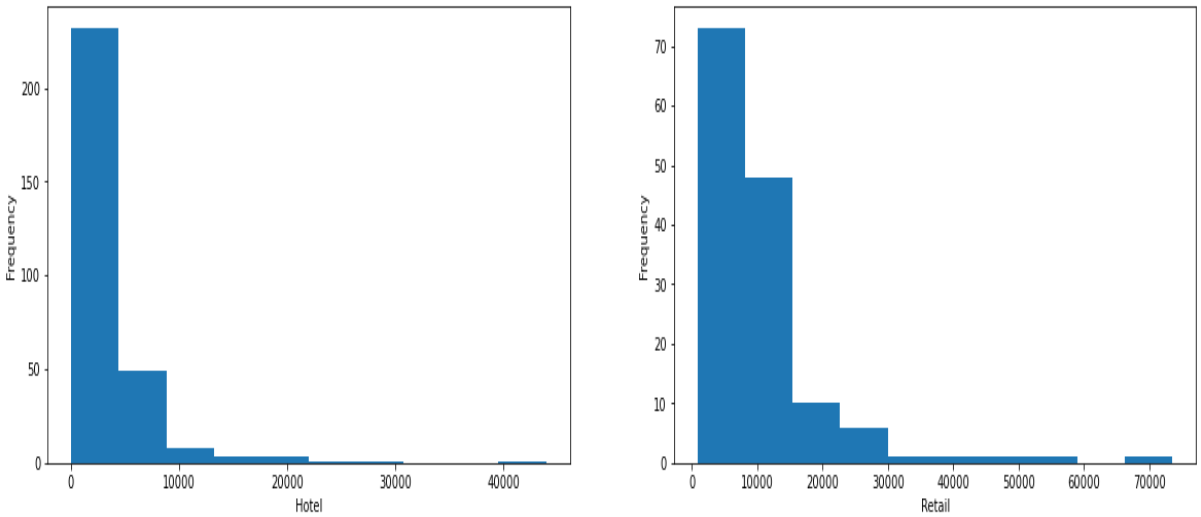
From the descriptive statistics(Table-1.4) the spread of Fresh across different regions is calculated.On an average a retailer from Lisbon spends 5486.415 on milk variety.Average spend of retailer on milk variety from Oporto is 5088.17 and that of retailer from Other region is 5977.085

From the plot we can state that the distribution of Milk across regions is right skewed.

Table1.5-Summary of Milk across Channels

Channel	Hotel	Retail
count	2.980000e+02	1.420000e+02
mean	3.451725e+03	1.071650e+04
std	4.352166e+03	9.679631e+03
min	5.500000e+01	9.280000e+02
25%	1.164500e+03	5.938000e+03
50%	2.157000e+03	7.812000e+03
75%	4.029500e+03	1.216275e+04
max	4.395000e+04	7.349800e+04
Total	1.028614e+06	1.521743e+06
CV	1.260867e+00	9.032456e-01

fig1.6 Distribution of Milk across Channels



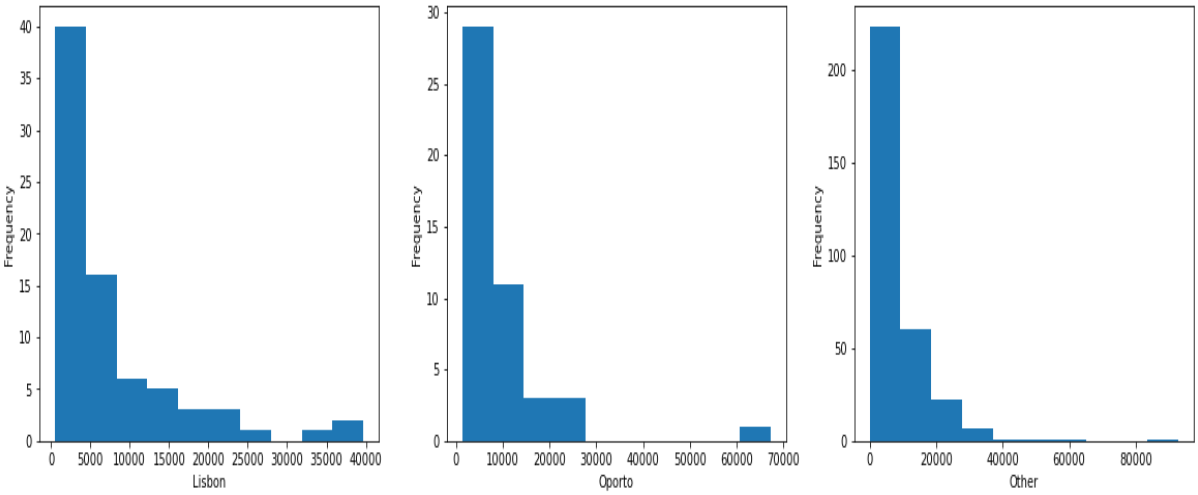
From the descriptive statistics (Table -1.5) we can calculate the spread of Milk across different Channels.On an average a retailer spends 3451.725 through 'Hotel'on milk variety.Average spend of retailer on milk variety through 'Retail' channel is 10716.50
From the plot we can state that the distribution of Milk across Channels is right skewed.

1.2.3Calculating the distribution of Grocery variety across different Regions and Channels

Table1.6-Summary of Grocery across Regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	3.160000e+02
mean	7403.077922	9218.595745	7.896364e+03
std	8496.287728	10842.745314	9.537288e+03
min	489.000000	1330.000000	3.000000e+00
25%	2046.000000	2792.500000	2.141500e+03
50%	3838.000000	6114.000000	4.732000e+03
75%	9490.000000	11758.500000	1.055975e+04
max	39694.000000	67298.000000	9.278000e+04
Total	570037.000000	433274.000000	2.495251e+06
CV	1.147670	1.176182	1.207808e+00

fig -1.7 Distribution of Grocery across Regions



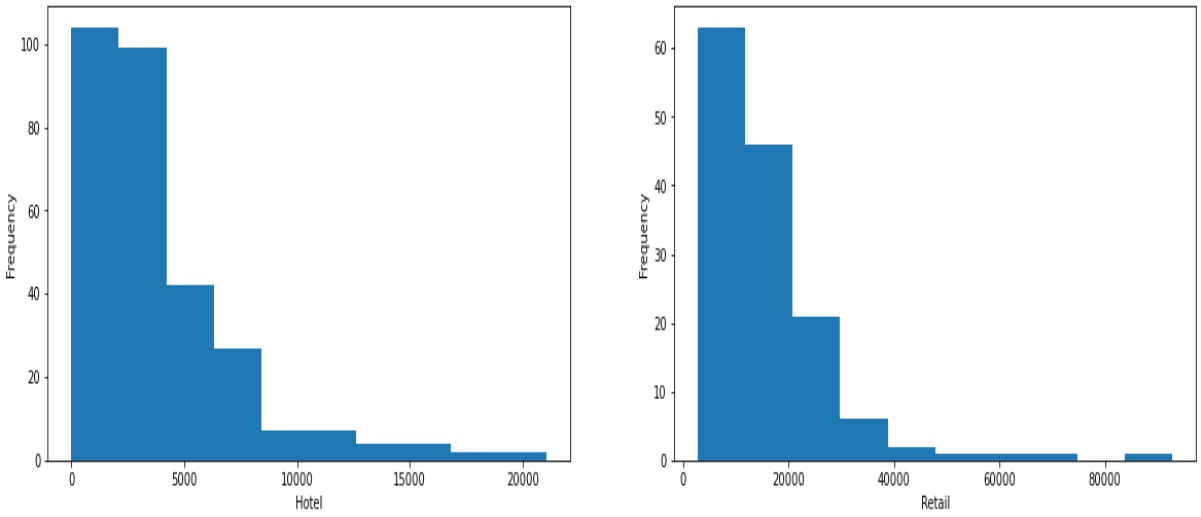
From the descriptive statistics(Table-1.6) we can calculate the spread of Grocery across different regions.On an average a retailer from Lisbon spends 7403.077 on Grocery variety.Average spend of retailer on grocery variety from Oporto is 9218.595 and that of retailer from Other region is 7896.364

From the plot we can state that the distribution of grocery across regions is right skewed.

Table1.7-Summary of Grocery across Channels

Channel	Hotel	Retail
count	2.980000e+02	1.420000e+02
mean	3.962138e+03	1.632285e+04
std	3.545513e+03	1.226732e+04
min	3.000000e+00	2.743000e+03
25%	1.703750e+03	9.245250e+03
50%	2.684000e+03	1.239000e+04
75%	5.076750e+03	2.018350e+04
max	2.104200e+04	9.278000e+04
Total	1.180717e+06	2.317845e+06
CV	8.948486e-01	7.515426e-01

fig1.8-Distribution of Grocery across Channels



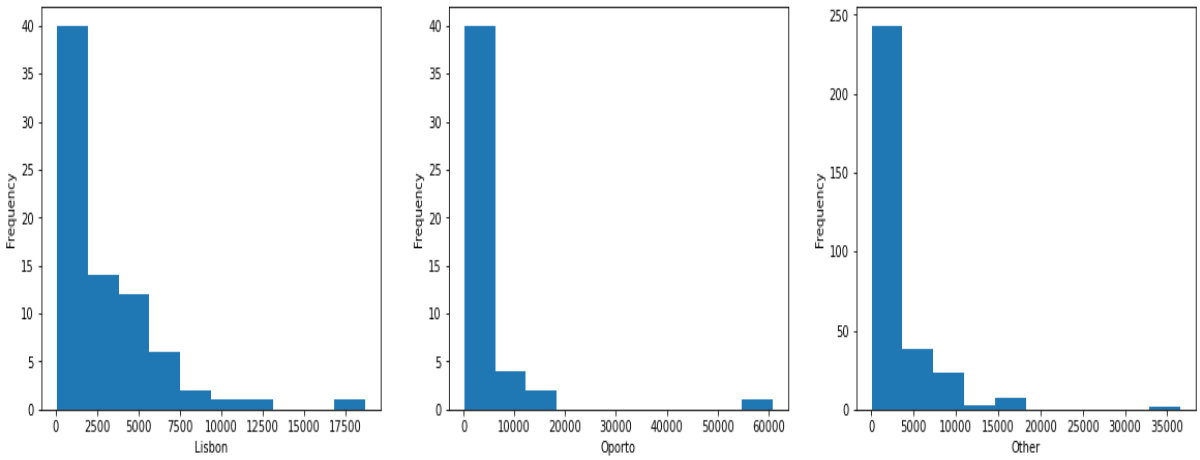
From the descriptive statistics (Table -1.7) we can calculate the spread of Grocery across different Channels.On an average a retailer spends 3962.138 through 'Hotel'on grocery variety.Average spend of retailer on grocery variety through 'Retail' channel is 16322.
From the plot we can state that the distribution of Grocery across Channels is right skewed.

1.2.4Calculating the distribution of Frozen variety across different Regions and Channels

Table1.8-Summary of Frozen across Regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	3000.337662	4045.361702	2944.594937
std	3092.143894	9151.784954	4260.126243
min	61.000000	131.000000	25.000000
25%	950.000000	811.500000	664.750000
50%	1801.000000	1455.000000	1498.000000
75%	4324.000000	3272.000000	3354.750000
max	18711.000000	60869.000000	36534.000000
Total	231026.000000	190132.000000	930492.000000
CV	1.030599	2.262291	1.446761

fig-1.9 Distribution of Frozen across Regions



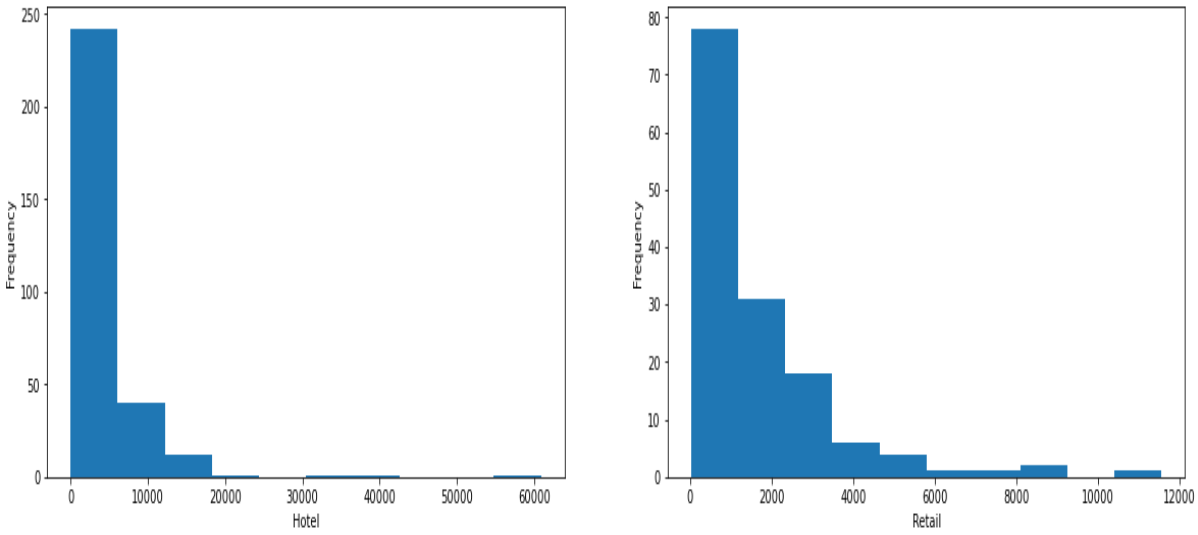
From the descriptive statistics(Table-1.8) we can calculate the spread of Frozen across different regions.On an average a retailer from Lisbon spends 3000.337 on grocery variety.Average spend of retailer on grocery variety from Oporto is 4045.36 and that of retailer from Other region is 2944.59

From the plot we can state that the distribution of grocery across regions is right skewed.

Table1.9-Summary of Frozen across Channels

Channel	Hotel	Retail
count	2.980000e+02	142.000000
mean	3.748252e+03	1652.612676
std	5.643913e+03	1812.803662
min	2.500000e+01	33.000000
25%	8.300000e+02	534.250000
50%	2.057500e+03	1081.000000
75%	4.558750e+03	2146.750000
max	6.086900e+04	11559.000000
Total	1.116979e+06	234671.000000
CV	1.505745e+00	1.096932

fig1.10-Distribution of Frozen across Channels



From the descriptive statistics (Table -1.9) we can calculate the spread of Frozen across different Channels.On an average a retailer spends 3748.252 through 'Hotel'on frozen variety.Average spend of retailer on frozen variety through 'Retail' channel is 1652.612

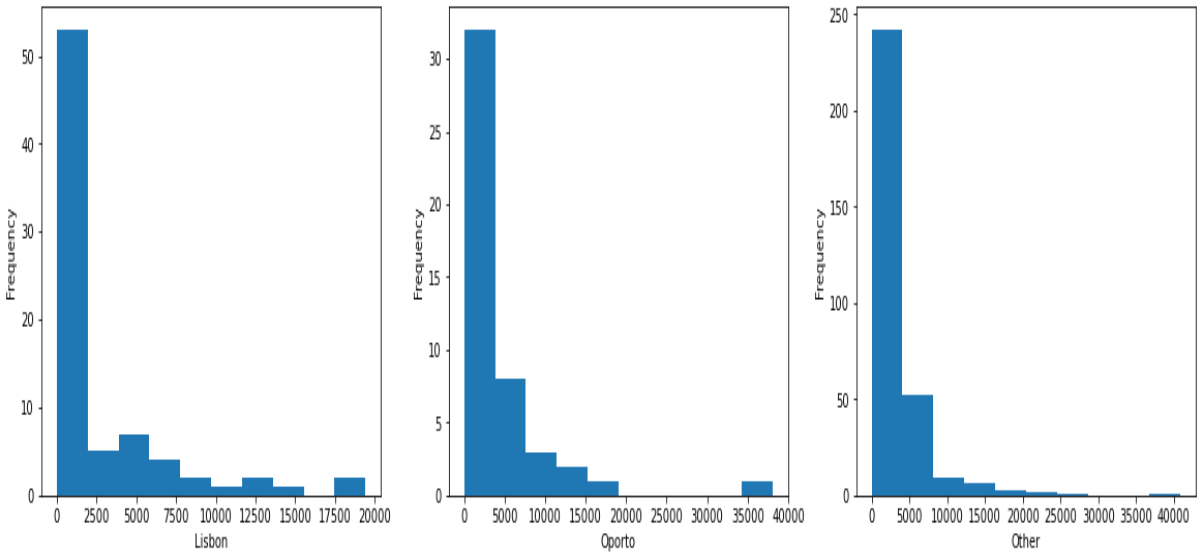
From the plot we can state that the distribution of Frozen across Channels is right skewed.

1.2.5 Calculating the distribution of Detergents paper variety across different Region and Channels

Table1.10-Summary of Detergents Paper across Regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	2651.116883	3687.468085	2817.753165
std	4208.462708	6514.717668	4593.051613
min	5.000000	15.000000	3.000000
25%	284.000000	282.500000	251.250000
50%	737.000000	811.000000	856.000000
75%	3593.000000	4324.500000	3875.750000
max	19410.000000	38102.000000	40827.000000
Total	204136.000000	173311.000000	890410.000000
CV	1.587430	1.766718	1.630040

fig-1.11-Distribution of Detergent Paper across Regions



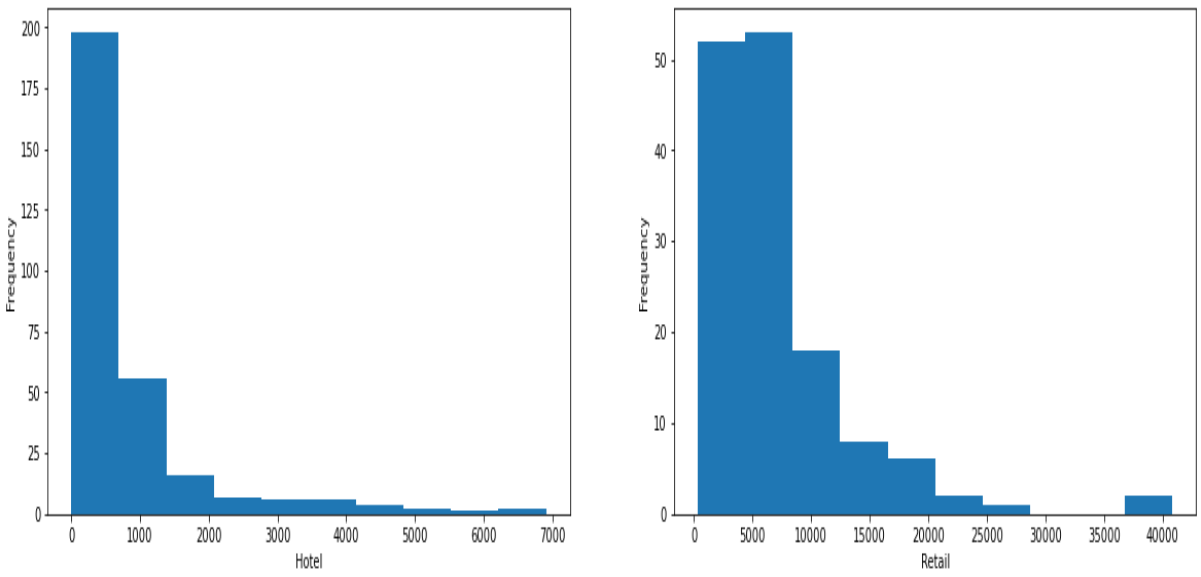
From the descriptive statistics(Table-1.10) we can calculate the spread of Detergent paper across different regions.On an average a retailer from Lisbon spends 2651.116 on Detergent paper.Average spend of retailer on detergent paper variety from Oporto is 3687.468 and that of retailer from Other region is 2817.753

From the plot we can state that the distribution of grocery across regions is right skewed.

Table1.11-Summary of Detergents Paper across Channels

Channel	Hotel	Retail
count	298.000000	1.420000e+02
mean	790.560403	7.269507e+03
std	1104.093673	6.291090e+03
min	3.000000	3.320000e+02
25%	183.250000	3.683500e+03
50%	385.500000	5.614500e+03
75%	899.500000	8.662500e+03
max	6907.000000	4.082700e+04
Total	235587.000000	1.032270e+06
CV	1.396596	8.654080e-01

fig1.12-Distribution of Detergent Paper across Channels



From the descriptive statistics (Table -1.11) we can calculate the spread of Detergents paper across different Channels.On an average a retailer spends 790.560 through 'Hotel'on detergents paper variety.Average spend of retailer on detergent paper variety through 'Retail' channel is 7269.507

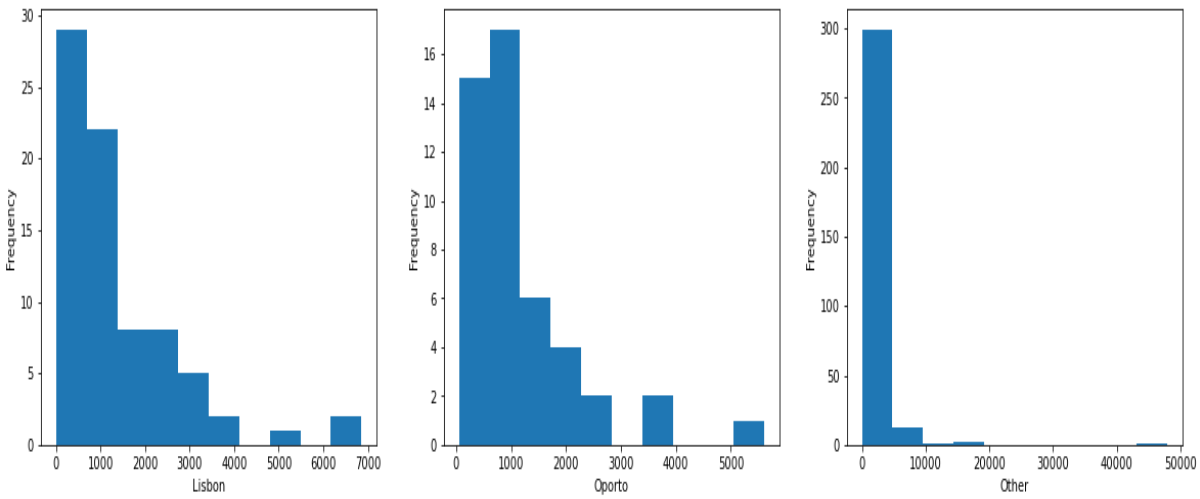
From the plot we can state that the distribution of Detergent paper across Channels is right skewed

1.2.6 Calculating the distribution of Delicatessen variety across different Region and Channels

Table1.12-Summary of Delicatessen across Regions

Region	Lisbon	Oporto	Other
count	77.000000	47.000000	316.000000
mean	1354.896104	1159.702128	1620.601266
std	1345.423340	1050.739841	3232.581660
min	7.000000	51.000000	3.000000
25%	548.000000	540.500000	402.000000
50%	806.000000	898.000000	994.000000
75%	1775.000000	1538.500000	1832.750000
max	6854.000000	5609.000000	47943.000000
Total	104327.000000	54506.000000	512110.000000
CV	0.993008	0.906043	1.994680

fig1.13-Distribution of Delicatessen across Regions



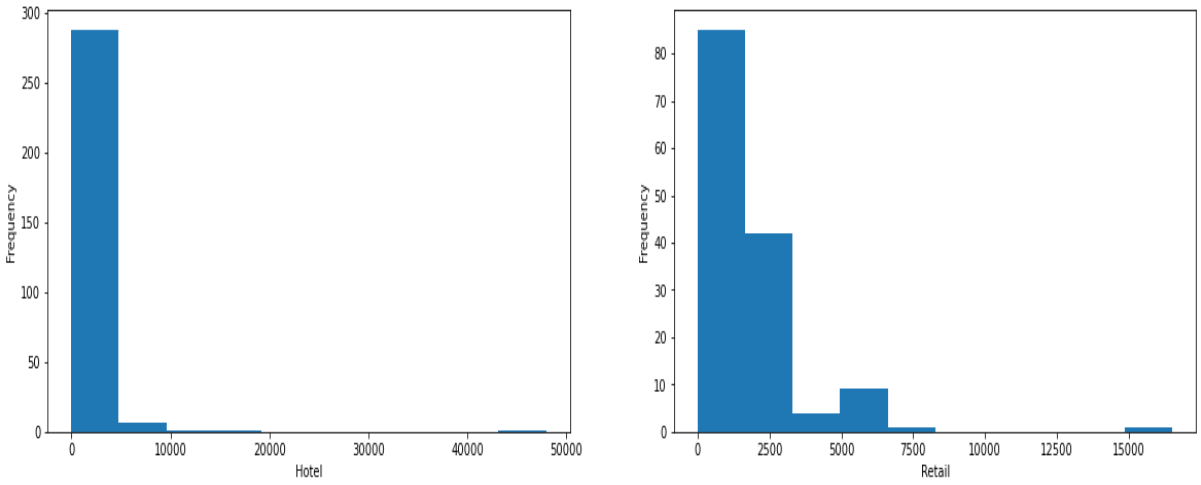
From the descriptive statistics(Table-1.13) we can calculate the spread of Delicatessen across different regions.On an average a retailer from Lisbon spends 1354.896 on Delicatessen.Average spend of retailer on delicatessen variety from Oporto is 1159.702 and that of retailer from Other region is 1620.601

From the plot we can state that the distribution of delicatessen across regions is right skewed.

Table1.13-Summary of Delicatessen across Channels

Channel	Hotel	Retail
count	298.000000	142.000000
mean	1415.956376	1753.436620
std	3147.426922	1953.797047
min	3.000000	3.000000
25%	379.000000	566.750000
50%	821.000000	1350.000000
75%	1548.000000	2156.000000
max	47943.000000	16523.000000
Total	421955.000000	248988.000000
CV	2.222828	1.114267

fig-1.14-Distribution of Delicatessen across Channels



From the descriptive statistics (Table -1.13) we can calculate the spread of Delicatessen across different Channels.On an average a retailer spends 1415.956 through 'Hotel'on delicatessen variety.Average spend of retailer on delicatessen variety through 'Retail' channel is 1753.436

From the plot we can state that the distribution of Delicatessen across Channels is right skewed

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

Table1.14-Summary of Data with Measures of Variability

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000
Cv	1.053918	1.273299	1.195174	1.580332	1.654647	1.849407

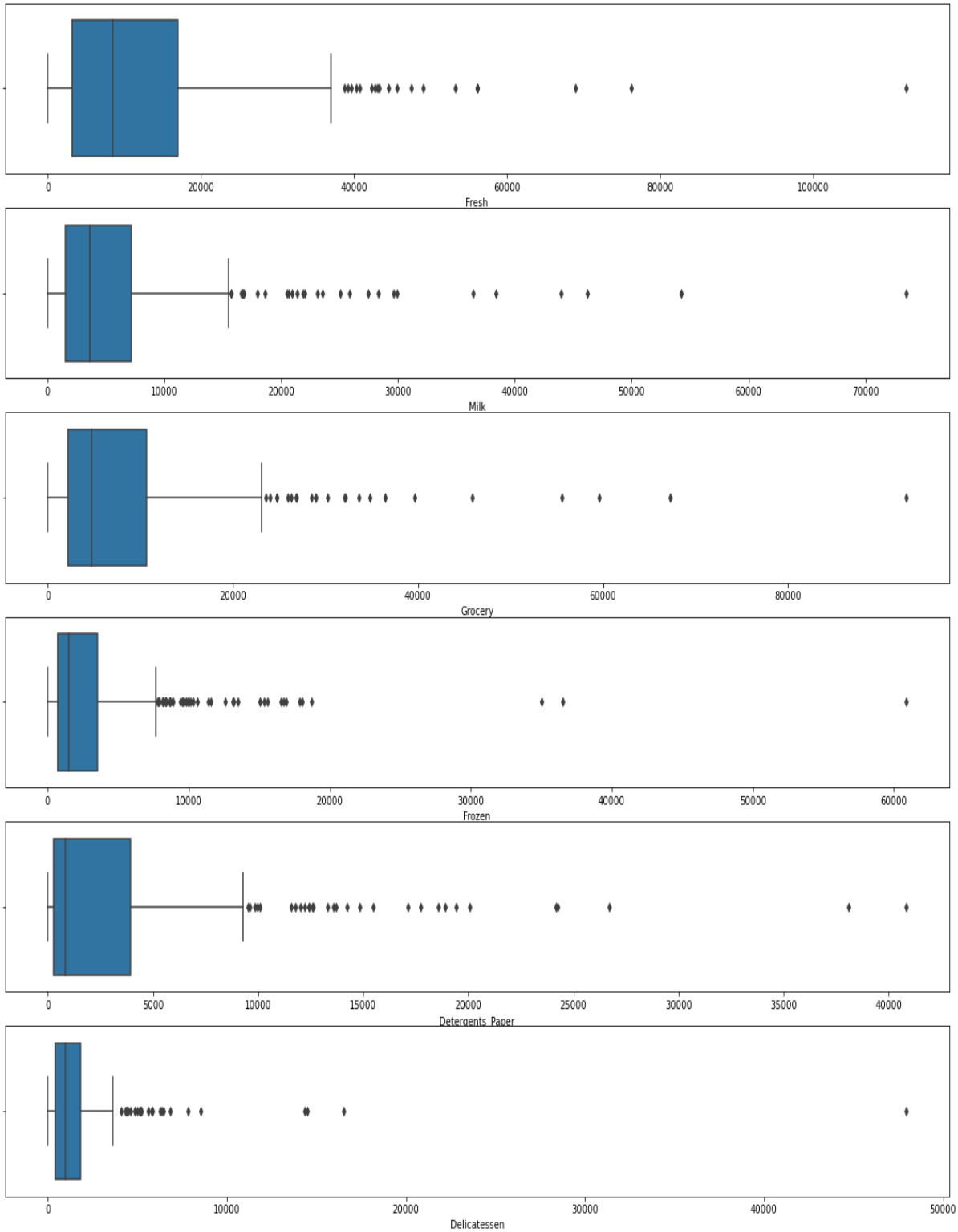
From the table we find that Co-efficient of Variance (Cv=1.8494) is higher for **Delicatessen** making it **the most inconsistent variety**

The Co-efficient of Variance (Cv = 1.053) is lower for **Fresh** making it **the least inconsistent variety.**

1.4 Are there any outliers in the data?

The most commonly adopted method to find the presence of outliers in the dataset is by constructing **Box plots** for the dataset.The values present above or below the whiskers are considered as **outliers**.

fig-1.15-Boxplot of Variables



Based on the result of boxplot we can conclude that the dataset contains outliers among all varieties.

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

Based on the analysis of the dataset we found that among the observed 3 regions ‘Other’ region is the most spent region.The Distributor can improve the sales by strengthening the most preferred sales channel ‘Hotel’. Also the demand for Fresh variety was consistent and higher than the other product varieties .New products can be introduced under fresh variety to improve the sales.The demand for delicatessen is most inconsistent.The inventory levels of such variety with higher inconsistency can be reduced .

Problem 2

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)

Executive Summary

The Student News Service at Clear Mountain State University(CMSU) gathers data from 62 undergraduate students of CMSU through a survey containing 14 questions and recorded their responses.Here the different probabilities of Gender,Major,Grad intention,Employment etc are considered to figure out the correlation between different variables.Contingency tables are constructed to find out such correlations.

Introduction

The dataset contains information about UG students of CMSU. Contingency tables are constructed to establish the correlation between different variables.The probability of different events are to be investigated.

Sample Dataset

Table-2 Sample 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

Exploratory Data Analysis

Let us check the type of Variables

ID	int64
Gender	object
Age	int64
Class	object
Major	object
Grad Intention	object
GPA	float64
Employment	object
Salary	float64
Social Networking	int64
Satisfaction	int64
Spending	int64
Computer	object
Text Messages	int64

The Dataset contains 62 rows and 14 columns.Of these 14 columns 2 columns are of float data type, 6 columns are integer data type and other 6 columns are object data type.

Let us check for the missing Values

ID 62 non-null int64

Gender	62 non-null	object
Age	62 non-null	int64
Class	62 non-null	object
Major	62 non-null	object
Grad Intention	62 non-null	object
GPA	62 non-null	float64
Employment	62 non-null	object
Salary	62 non-null	float64
Social Networking	62 non-null	int64
Satisfaction	62 non-null	int64
Spending	62 non-null	int64
Computer	62 non-null	object
Text Messages	62 non-null	int64

There are no null values present in dataset.

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

Contingency Table: A table showing the distribution of one variable in rows and another in columns.It is used to study the correlation between the two variables.This table usually shows the frequency for particular combination of variables.Keeping Gender as row variable following contingency tables are constructed.

2.1.1.GenderandMajor

Contingency Table:2.1 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

2.1.2. Gender and Grad Intention

Contingency Table:2.2 Gender vs 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

Contingency Table:2.3 Gender vs 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

Contingency Table:2.4 Gender vs 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 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?

Probability of choosing male student=(Total no.of male students)/(Total no. of students)

Prob_male=29/62=0.46774

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

Probability of choosing female student=(Total no.of female students)/(Total no. of students)

Prob_female=33/62= 0.53225806451612

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

Referring to the contingency table(table-2.1) the following probabilities are derived.

Contingency Table:2.1 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

Cond_Prob of different majors among male students = (No.of male students in a major) / (Total no. of male students).

Cond_Prob of Accounting among male students= 4/29 = 0.1379

Cond_Prob of CIS among male students= 1/29 = 0.0344

Cond_Prob of Economics/Finance among male students= 4/29 = 0.1379

Cond_Prob of International Business among male students= 2/29 = 0.0689

Cond_Prob of Management among male students= 6/29 = 0.2068

Cond_Prob of Other among male students= 4/29 = 0.1379

Cond_Prob of Retail/Marketing among male students= 5/29 = 0.1724

Cond_Prob of Undecided among male students= 3/29 = 0.1034

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

Cond_Prob of different majors among female students = (No.of female students in a major) / (Total no. of female students).

Cond_Prob of Accounting among female students= 3/33 = 0.090

Cond_Prob of CIS among female students= 3/33 = 0.090

Cond_Prob of Economics/Finance among female students= 7/33 = 0.212

Cond_Prob of International Business among female students= 4/33 = 0.1212

Cond_Prob of Management among female students= 4/33 = 0.1212

Cond_Prob of Other among female students= 3/33 = 0.090

Cond_Prob of Retail/Marketing among female students= 9/33 = 0.272

Cond_Prob of Undecided among female students= 0/33 = 0.0

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.

Referring to Contingency table

Contingency Table:2.2 Gender vs Grad Intention

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

P(Graduate ∩ Male) = P (Graduate| Male) x P (male) = (17/29)*(29/62)

P(Graduate ∩ Male) = 0.274

The probability that a randomly chosen student is a male and intends to graduate is **0.274**

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

Referring to Contingency table

Contingency Table:2.4 Gender vs Computer

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

P(NoLaptop ∩ Female) = P (NoLaptop| Female) x P (Female) = (4/33)*(33/62)

The probability that a randomly chosen student is a female without a laptop is **0.0645**

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 a male or has full-time employment?

Referring to Contingency table

Contingency Table:2.3 Gender vs Employment

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

P(Male U Full time) = P (Male) + P (Full time) - P(Male ∩ Full time) =(29/62)+(10/62)-(7/62) = 0.5161

The probability that a randomly chosen student is a male or has full-time employment is **0.516**

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

Referring to Contingency table

Contingency Table:2.1 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

P ((IB U Mg)| Female) = P (IB | Female) + P (Mg | Female)- P((IB ∩ Mg)|Female)
= (4/33)+(4/33)-0 = 0.242

The conditional probability that given a female student is randomly chosen, she is majoring in international business or management is **0.242**

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?

Contingency Table 2.5 Gender vs Graduate at 2 levels

Grad Intention / Gender	No	Yes
Female	9	11
Male	3	17

If the two events are independent the following condition must be satisfied

P(Female ∩ GI yes)= P(Female)*P(GI yes)

P(Female)*P(GI yes) = (20/40)*(28/40) = 0.35

P(Female ∩ GI yes) = (11/40) = 0.275

This is not independent events as probability multiplication of both events is not equal to the probability of combined event.
so being a female and graduate intention are **not independent events.**

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

Answer the following questions based on the data

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

Table 2.6 Dataset for GPA less than 3

	index	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
2	3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop	250
3	4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop	100
4	5	6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78.0	3	2	700	Laptop	30
5	10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50.0	2	5	400	Laptop	200
6	23	24	Male	22	Senior	Undecided	Yes	2.6	Full-Time	45.0	1	5	400	Laptop	600
7	27	28	Female	20	Junior	International Business	Yes	2.9	Part-Time	50.0	3	1	900	Laptop	100
8	31	32	Male	20	Junior	Other	Yes	2.9	Part-Time	47.0	3	1	300	Laptop	300
9	33	34	Male	22	Senior	Retailing/Marketing	Yes	2.6	Full-Time	40.0	1	4	1400	Laptop	800
10	37	38	Female	21	Sophomore	Accounting	Yes	2.5	Part-Time	60.0	2	3	500	Laptop	600
11	38	39	Male	24	Junior	Economics/Finance	Yes	2.8	Part-Time	50.0	1	6	600	Laptop	50
12	39	40	Male	19	Sophomore	Retailing/Marketing	Yes	2.5	Unemployed	50.0	2	5	300	Laptop	100
13	47	48	Male	19	Sophomore	Undecided	Undecided	2.5	Part-Time	80.0	2	4	500	Laptop	150
14	57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
15	58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	2	4	350	Laptop	250
16	59	60	Female	20	Sophomore	CIS	No	2.5	Part-Time	55.0	1	4	500	Laptop	500

From the table,we know that 17 students out of 62 has scored GPA less than 3.
The probability of randomly chosen student's GPA is less than 3 **= (**17/62**) = 0.274**

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.

To find the conditional probability of a person male/female earning 50 0r more the following dataset is obtained by slicing from the given data set.

Table 2.7-Sample Dataset of Salary 50 or more

	index	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
0	0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
1	5	6	Female	22	Senior	Economics/Finance	Undecided	2.3	Unemployed	78.0	3	2	700	Laptop	30
2	6	7	Female	21	Junior	Other	Undecided	3.0	Part-Time	50.0	1	3	500	Laptop	50
3	7	8	Female	22	Senior	Other	Undecided	3.1	Full-Time	80.0	1	2	200	Tablet	300
4	10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50.0	2	5	400	Laptop	200

Descriptive Analysis of sample dataset

Table 2.8 Summary of Sample Dataset

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer
count	32.000000	32	32.000000	32	32	32	32.000000	32	32.000000	32.000000	32.000000	32.000000	32
unique	NaN	2	NaN	3	8	3	NaN	3	NaN	NaN	NaN	NaN	3
top	NaN	Female	NaN	Junior	Economics/Finance	Yes	NaN	Part-Time	NaN	NaN	NaN	NaN	Laptop
freq	NaN	18	NaN	15	8	16	NaN	21	NaN	NaN	NaN	NaN	28
mean	30.437500	NaN	21.062500	NaN	NaN	NaN	3.071875	NaN	57.781250	1.406250	3.625000	493.593750	NaN
std	16.823443	NaN	1.702702	NaN	NaN	NaN	0.353995	NaN	8.597578	0.756024	1.338029	195.951913	NaN
min	1.000000	NaN	18.000000	NaN	NaN	NaN	2.300000	NaN	50.000000	0.000000	1.000000	200.000000	NaN
25%	18.750000	NaN	20.000000	NaN	NaN	NaN	2.900000	NaN	50.000000	1.000000	3.000000	337.500000	NaN
50%	27.500000	NaN	21.000000	NaN	NaN	NaN	3.100000	NaN	55.000000	1.000000	4.000000	500.000000	NaN
75%	42.250000	NaN	22.000000	NaN	NaN	NaN	3.300000	NaN	60.000000	2.000000	4.000000	600.000000	NaN
max	62.000000	NaN	26.000000	NaN	NaN	NaN	3.800000	NaN	80.000000	3.000000	6.000000	1100.000000	NaN

From the table we know that

No. of Female earning 50 or more = 18

No. of male earning 50 or more = 14

Adopting **Baye's theorem** for calculating Conditional Probability

The Conditional Probability P1(Salary>=50 | male)

=P (male | Salary>=50) * P (Salary>=50) / P (male)

=(14/32)*(32/62) / (29/62) = 0.482

The Conditional Probability P2(Salary>=50 |female)

=P (female | Salary>=50) * P (Salary>=50) / P(female)

=(18/32)*(32/62)/(33/62) = 0.545

The conditional probability that a randomly selected male earns 50 or more is **0.482**

The conditional probability that a randomly selected female earns 50 or more is **0.545**

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.

Test to check for normal distribution

The distribution of variable can be checked if it follows normal distribution or not by **Shapiro Wilk test for normality.**

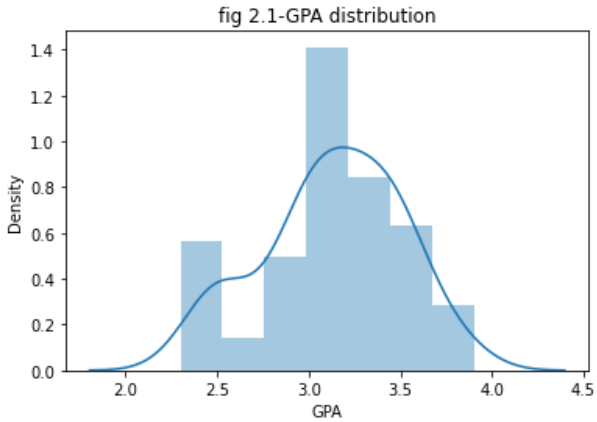
Let us assume the null and alternate hypothesis for the test as follows

Null hypothesis H0: The sample population is normally distributed at 95% confidence

Alternate hypothesis H1: The sample population is not normally distributed

Let significance of test Alpha=0.05

Test on GPA to check for Normal distribution:



Shapiro Test on GPA distribution

Result of shapiro test

Statistic = 0.968

p value = 0.112

Since the p value > alpha we don't have sufficient evidence to reject null hypothesis

The variable **GPA is normally distibuted**

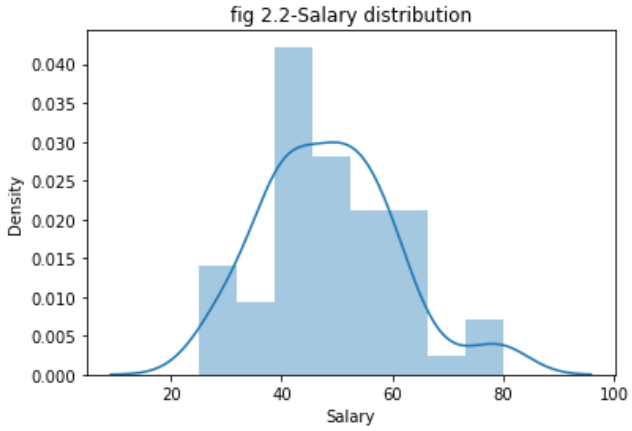
Measures of centre

Mean=3.129

Median=3.15

Mode=3.0,3.1,3.4. Here the measures are nearly equal confirming our hypothesis.

Test on Salary to check for Normal distribution:



Shapiro Test on Salary distribution

Result of shapiro test

Statistic = 0.956

p value = 0.028

Since the p value < alpha we have sufficient evidence to reject null hypothesis

The variable **Salary is not normally distributed**

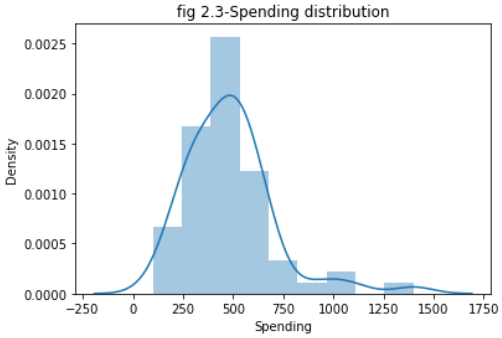
Measures of centre

Mean=48.54

Median=50

Mode=40. Here the measures are not equal confirming our hypothesis.

Test on Spending to check for Normal distribution:



Shapiro Test on Spending distribution

Result of shapiro test

Statistic = 0.8777

p value = $1.6854e^{-05}$

Since the p value < alpha we have sufficient evidence to reject null hypothesis

The variable **Spending is not normally distributed**

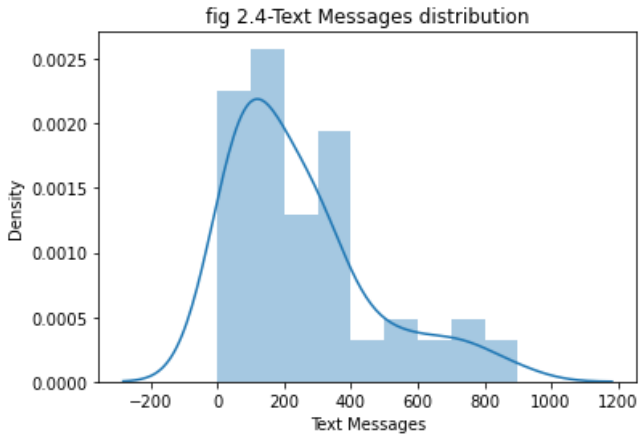
Measures of centre

Mean=482.016

Median=500

Mode=500. Here the measures are not equal confirming our hypothesis.

Test on Text messages to check for Normal distribution:



Shapiro Test on Text messages distribution

Result of shapiro test

Statistic = 0.859

p value = $4.32e^{-06}$

Since the p value < alpha we have sufficient evidence to reject null hypothesis
The variable **Text Messages is not normally distributed**

Measures of centre

Mean=246.209

Median=200

Mode=300.Here the measures are not equal confirming our hypothesis.

Problem 3

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 are calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.

Executive Summary

A company manufacturing ABC asphalt shingles is in need to ensure the quality of the shingles produced which can be done by performing moisture tests.The dataset contains moisture test results of sample A and B.In this problem we will be examining moisture content to check whether the mean moisture content is within permissible limit.

Introduction

The given dataset contains observations of moisture test estimating moisture per 100 square feet in samples of Shingle A and Shingle B.36 observations of sample A and 31 observations of sample B are taken. The permissible moisture content is less than 0.35 pounds per 100 square feet. Exploratory Data Analysis is to be done and the measures of cental tendency are calculated.

Sample Dataset

Table3.1 Sample Dataset

	A	B
0	0.44	0.14
1	0.61	0.15
2	0.47	0.31
3	0.30	0.16
4	0.15	0.37

Exploratory Data Analysis

Let us check the types of variables in the data frame

A float 64

B float 64

There are 36 rows and 2 columns in dataset .All the entries are of float data type.

Check for the missing values

RangeIndex: 36 entries, 0 to 35

Data columns (total 2 columns)
A 36 non-null float64
B 31 non-null float64
B 5 null

Descriptive Analysis

Table3.2 Summary of Dataset

	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

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.

A null hypothesis is a hypothesis is the hypothesis that the researcher is trying to disprove.

An alternative hypothesis simply is the inverse, or opposite, of the null hypothesis. In testing the company would like to show that the mean moisture content of samples A and B exceeds permissible limits.

Null hypothesis: The mean moisture content is within permissible level

Alternative hypothesis: states that the mean moisture content exceeds the permissible limit.

Hypothesis Testing for sample A

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

Null hypothesis $H0 \leq 0.35$

Alternate hypothesis $H1 > 0.35$ at 95% confidence

Let's assume significance of test (Alpha) as 0.05

By performing individual one sample t test on sample A

We get

p value for sampleA= 0.9252236685509249

here p value >Alpha
we have no evidence to reject null hypothesis

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

Null hypothesis $H0 \leq 0.35$

Alternate hypothesis $H1 > 0.35$ at 95% confidence

Let's assume significance of test (Alpha) as 0.05

By performing individual one sample t test on sample B

We get

p value for sampleB= 0.9979095225996808

here p value >Alpha
we have no evidence to reject null hypothesis

Thus the mean moisture contents in both types of shingles A & B are within the permissible limits The claim made by the company that the mean moisture content is less than 0.35 pounds per 100 square feet is found to be true. Thus the manufacturing process is as per the standards and no interference is required.

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?

For Normal distributions with unknown variance the difference between means follows a student's t-distribution.The Student t-test is one of the oldest and widely used hypothesis test. As per the t-test, under null hypothesis the test statistic follows a Student t-distribution.

Here the population variance remains unknown.To test whether the population mean for shingles A & B are equal, **t-test** can be used as it is the commonly used method to assess the hypothesis.. A t test can be estimated for: 1) One sample t test 2) Two sample t test (including paired t test)

Assumption: We assume that the *samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances***.**

First decide the level of significance:The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P (Type I error) = α. Confidence level: The level of significance 0.05 is related to the 95% confidence level. Level of Significance α = 0.05

The dataset has 36 measurements for Shingle A & 31 mesurements for shingle B.

The Null and Alternate hypothesis are considered as follows

H0 : population mean for shingles A=population mean for shingles B

H1 : population mean for shingles A != population mean for shingles B

Here let's assume significance of test (Alpha) as 0.05

By performing **Levene test** the population variance for both samples are found to be equal.Independent Two sample t test is to be done on A & B samples

t stat of two sample t test=1.289

pvalue of two sample t test =0.2017496571835328 at the level of 5% significance.

here p value >Alpha

we have no evidence to reject null hypothesis since pvalue > Alpha

fig 3.1-Sample Moisture content in shingle A

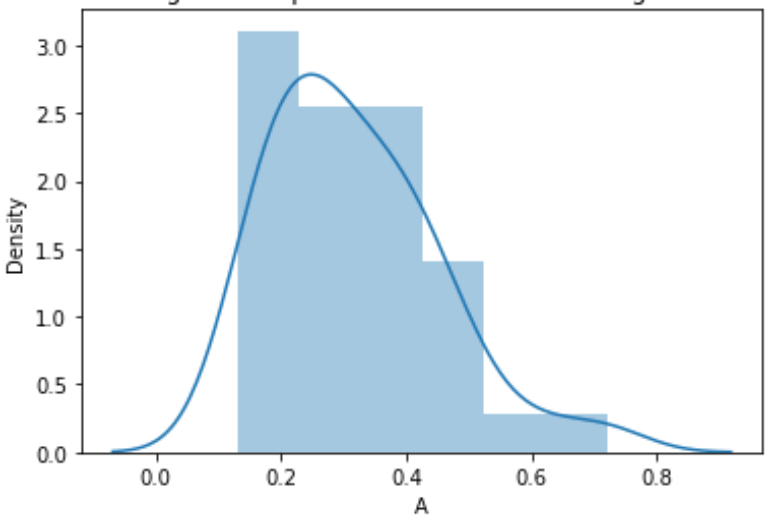
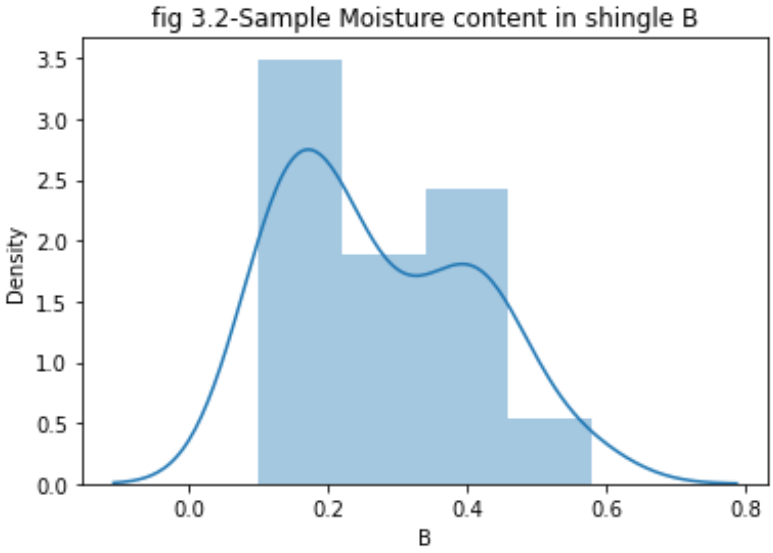


fig 3.2-Sample Moisture content in shingle B



The End…