SMDM Project

Table of Contents

1 Project Objective

The objective of the report is to explore the project data sets in Python and generate insights about the data sets. This exploration report will consist of the following:

- Importing the dataset in Python
- Understanding the structure of dataset
- Graphical exploration
- Descriptive statistics
- Insights from the dataset

2 Assumptions

- The data provided is accurate
- Data was collected by random
- Normally distributed

3 Exploratory Data Analysis

Data exploration consists of the following steps:

1. Environment Set up and Data Import
2. Variable Identification
3. Univariate Analysis

4. Bi-Variate Analysis
5. Outlier Identification

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

The following are the packages installed for this project

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Scipy
- Statsmodels

```
In [26]: import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         %matplotlib inline
         import scipy.stats as stats
         from scipy.stats import ttest_1samp, ttest_ind, mannwhitneyu, levene, shapiro
         from statsmodels.stats.power import ttest_power
         from statsmodels.stats.weightstats import ztest
```

3.1.2 Set up working Directory

The following is the working directory for the project
C:\Projects\SMDM\

3.1.3 Import and Read Dataset

The data for each problem has been imported using the command ‘read.csv’ or ‘read.excel’

Problem 1

```
In [5]: customer_data = pd.read_excel('Wholesale customers data-1.xlsx', sheet_name=1)
```

Problem 2

```
In [2]: survey_data = pd.read_csv('Survey-1.csv')
```

Problem 3

```
In [13]: shingles_data = pd.read_csv('A & B shingles-1.csv')
```

3.2 Variable Identification

In each of the three datasets the following are the variables identified

Problem 1

- Channel
- Region
- Fresh
- Milk
- Grocery
- Frozen
- Detergents Paper
- Delicatessen

Problem 2

- Gender
- Major
- Grad Intention
- Employment
- Salary
- Spending
- Computer
- Text messages

Problem 2

- A
- B

3.2.1 Variable Identification - Inferences

The following are the inferences made on each variable for the three datasets

Problem 1

- Channel - is a distribution variable which consists of specific distribution channels for the items
- Region - is a location variable
- Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen - are all varieties of products showing spending made my each retailer

Problem 2

- Gender - gender of student
- Major - the category the student has majored in
- Grad Intention - the student's intent to graduate
- Employment - the student's employment status
- Salary - the income of the student
- Spending - expenses made by the student
- Computer - type of computer that each student has
- Text messages - the number of text messages sent by the student

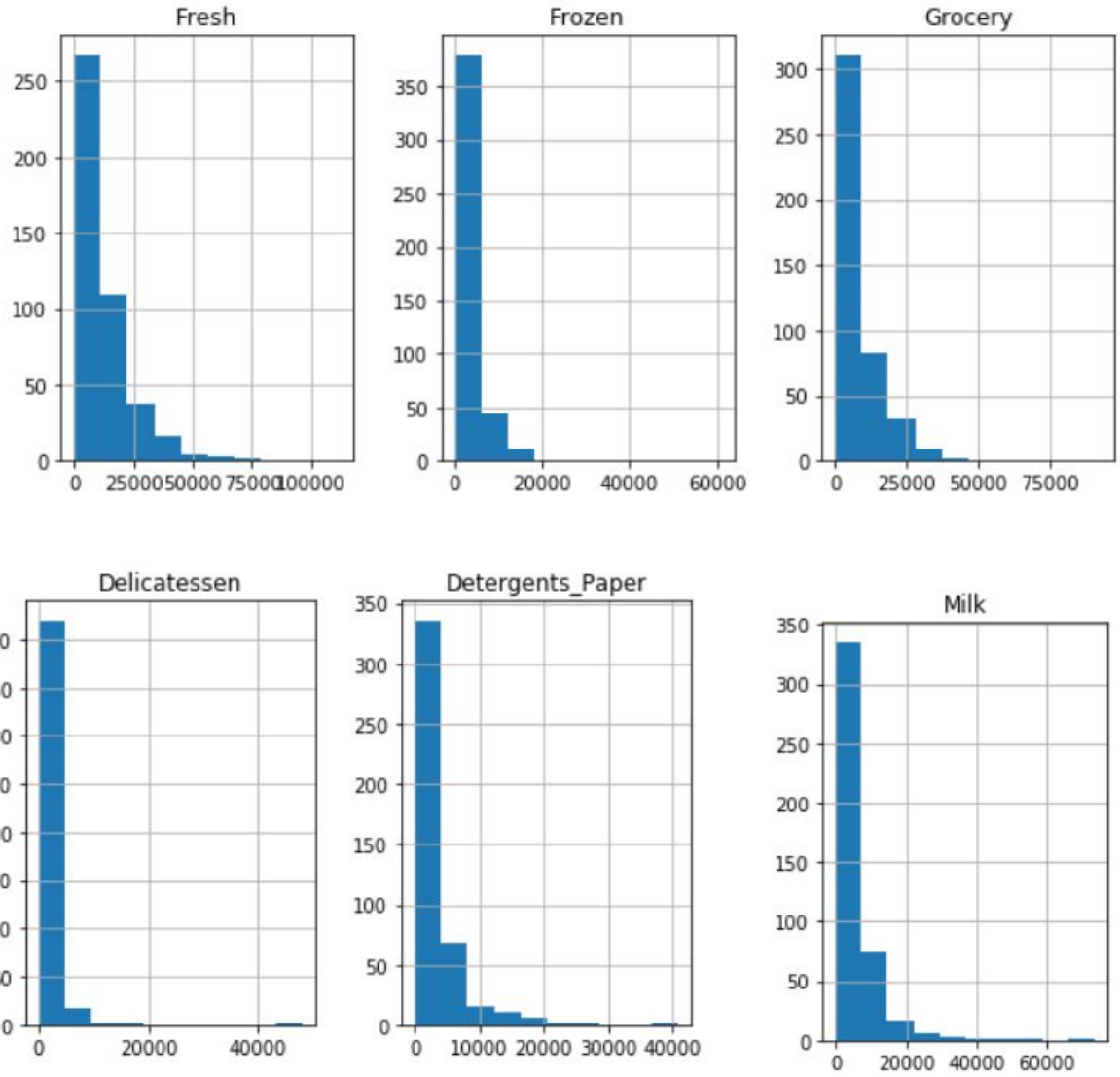
Problem 3

- A - first set of Shingles measured for moisture
- B - second set of Shingles measured for moisture

3.3 Univariate Analysis

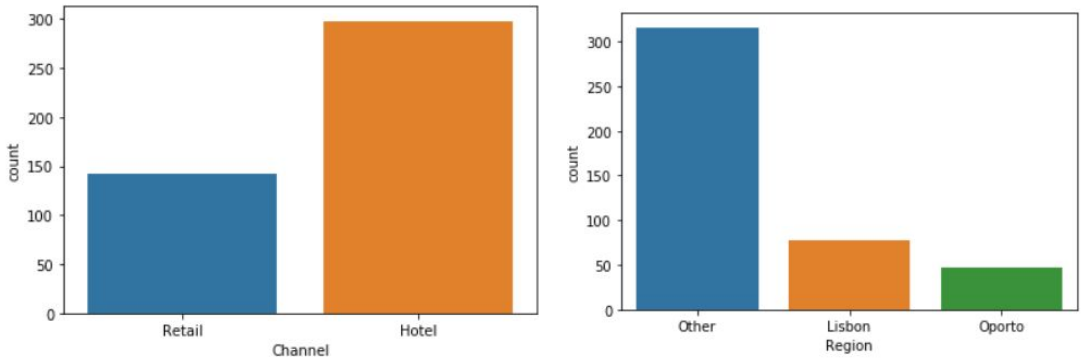
The following shows the univariate analysis done for each dataset

Problem 1
The following shows the histograms prepared for each variety of products



- All the histograms are positively skewed

  The following is the count plot prepared for variables Region and Channel



- Hotel and Other is the Channel and Region which spend more respectively
- Retail and Oporto is the Channel and Region which spend less respectively

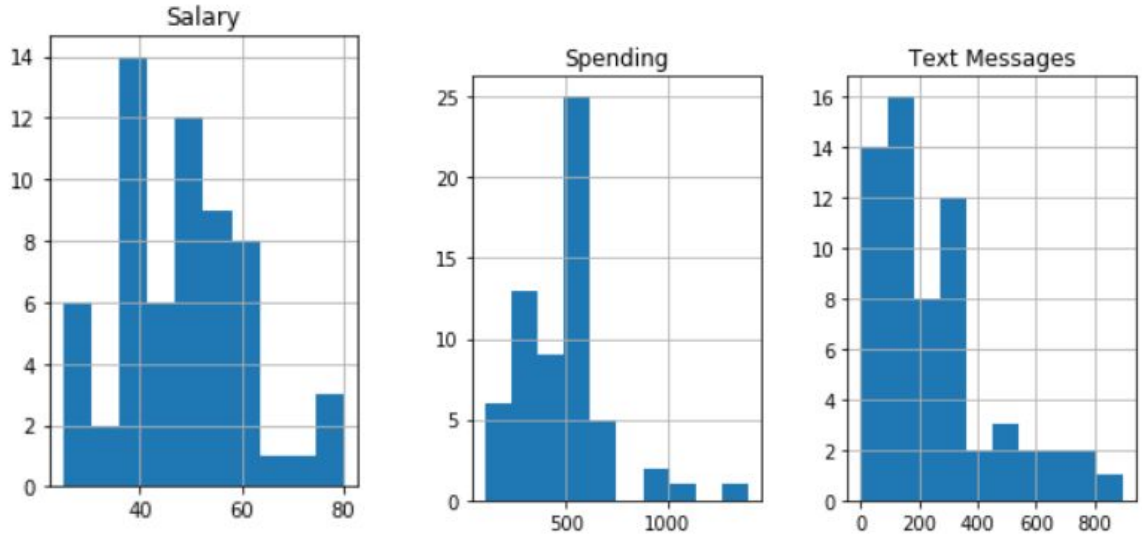
Coefficient of Variation

Variety of product	Variation
Fresh	1.05
Milk	1.27
Grocery	1.19
Frozen	1.58
Detergents Paper	1.65
Delicatessen	1.85

- Delicatessen shows the most inconsistent behaviour and Fresh shows the least inconsistent Behaviour

Problem 2

The following shows the histograms prepared for Salary, Spending and Text Messages



- The Histogram for salary shows a nearly bell shaped curve and the skewness is close to 0 and thus can say that salary is normally distributed
- The Histogram for Spending is not bell shaped and is right skewed this is not normally distributed
- similarly the Histogram for Text messages is also not bell shaped and is right skewed, thus not normally distributed

Problem 3

Hypothesis Testing for A

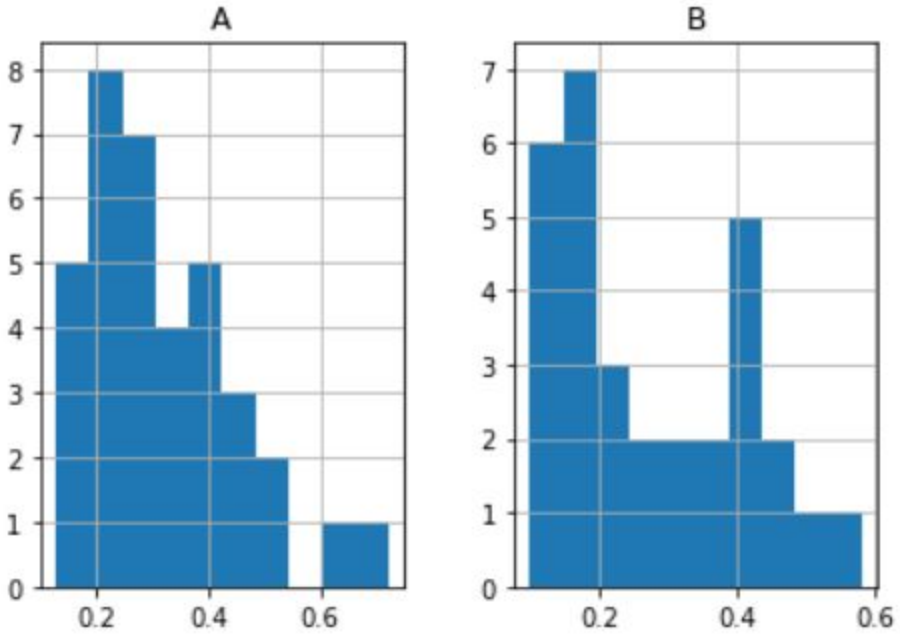
- The Null Hypothesis is that the mean moisture content for group A is greater than or equal to 0.35 pound per 100 square feet
- The Alternate Hypothesis is that the mean moisture content for group A is less than 0.35 pound per 100 square feet

- The statistical test conducted is Z test since the sample size is greater than 30
- Upon testing the P-value calculated is 0.88
- At the 0.05 level of significance we fail to reject the Null hypothesis
- There is no evidence to prove that the mean moisture content is less than 0.25 pound per square feet

Hypothesis Testing for B

- The Null Hypothesis is that the mean moisture content for group B is greater than or equal to 0.35 pound per 100 square feet
- The Alternate Hypothesis is that the mean moisture content for group B is less than 0.35 pound per 100 square feet
- The statistical test conducted is Z test since the sample size is greater than 30
- Upon testing the P-value calculated is 0.88
- At the 0.05 level of significance we fail to reject the Null hypothesis
- There is no evidence to prove that the mean moisture content is less than 0.25 pound per square feet

The following is the histograms prepared for groups A and B



- The data is not normally distributed

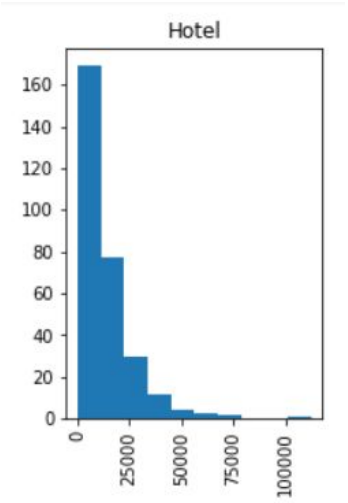
3.4 Bi-Variate Analysis

The following shows the Bi-variate analysis done for each dataset

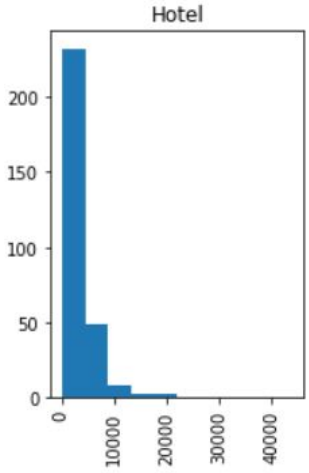
Problem 1

The following are the histograms prepared between Channel and the variety of products

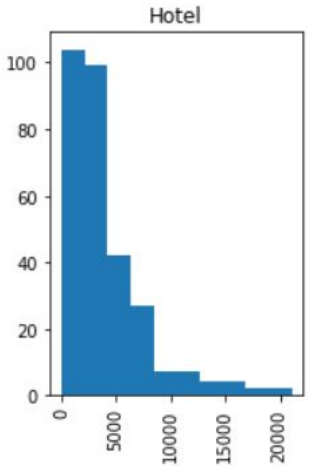
Hotel:



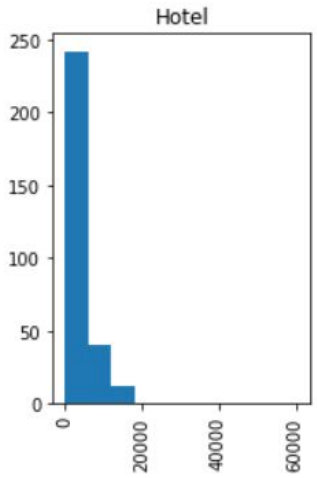
Fresh



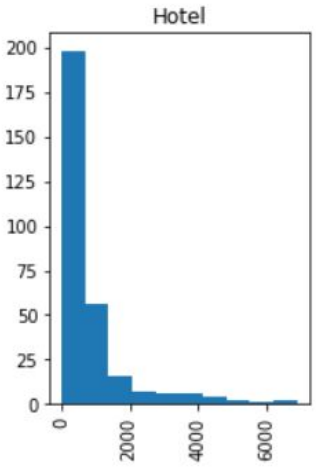
Milk



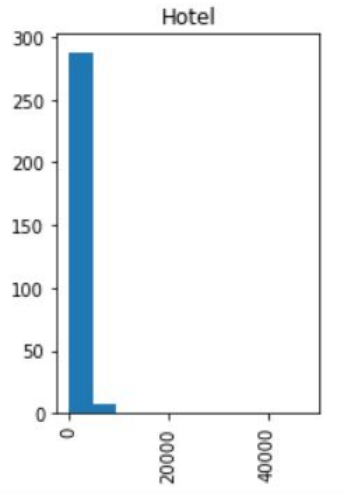
Grocery



Frozen

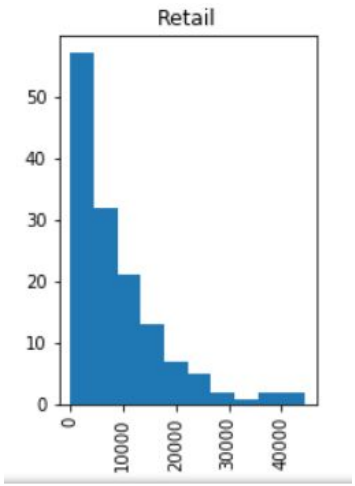


Detergents Paper

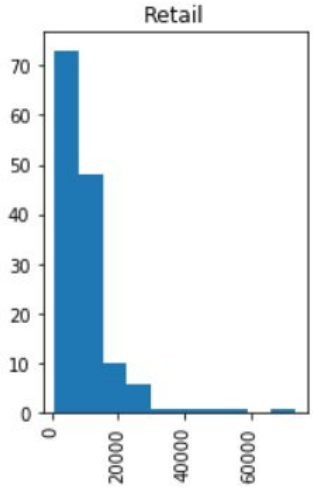


Delicatessen

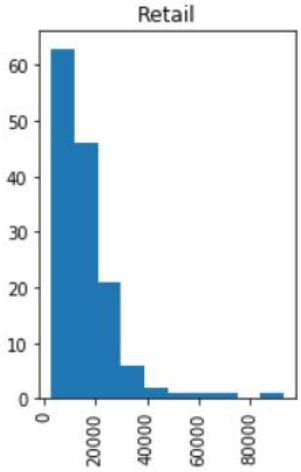
Retail:



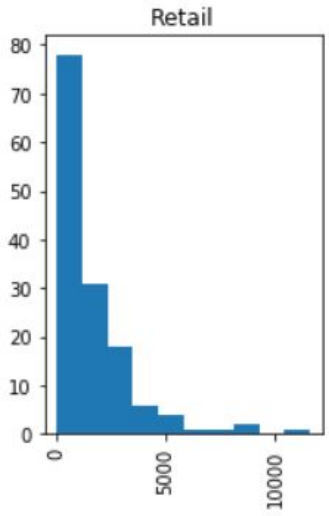
Fresh



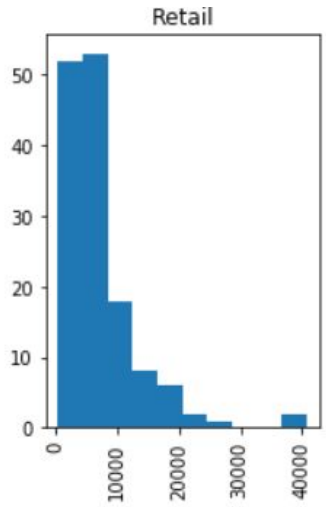
Milk



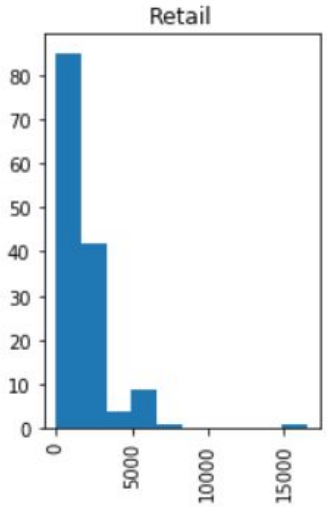
Grocery



Frozen

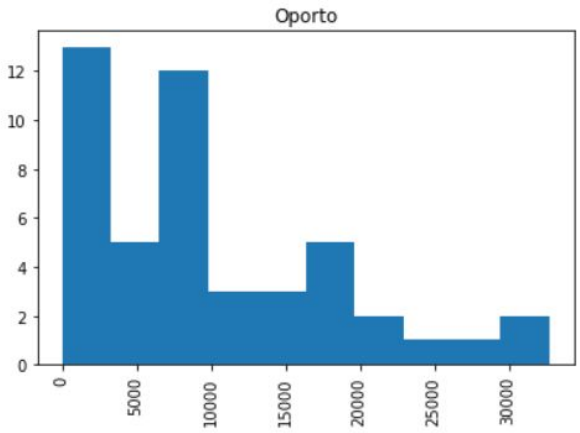


Detergents Paper

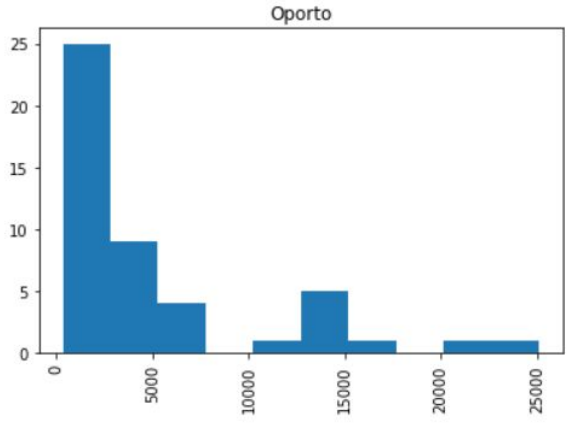


Delicatessen

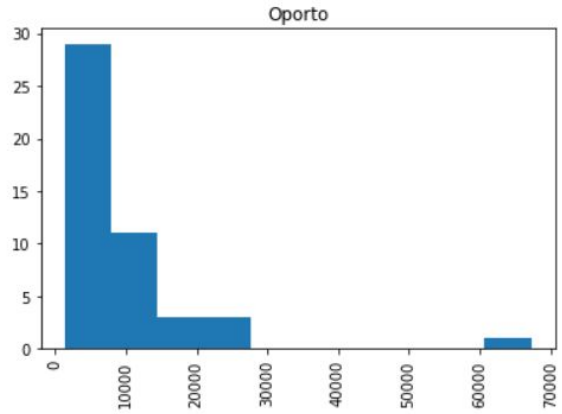
The following are histograms prepared Region and Variety of Products Oporto:



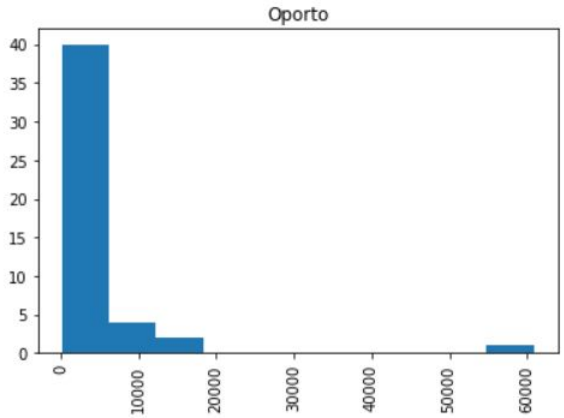
Fresh



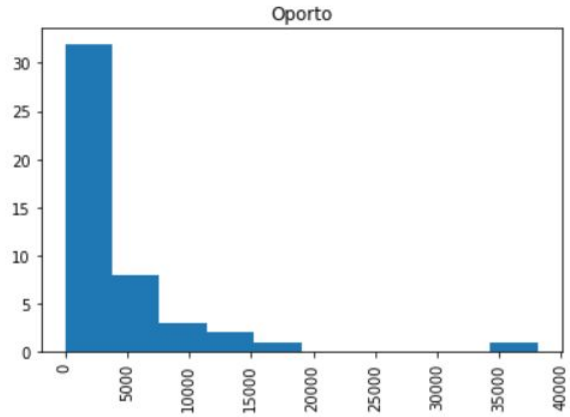
Milk



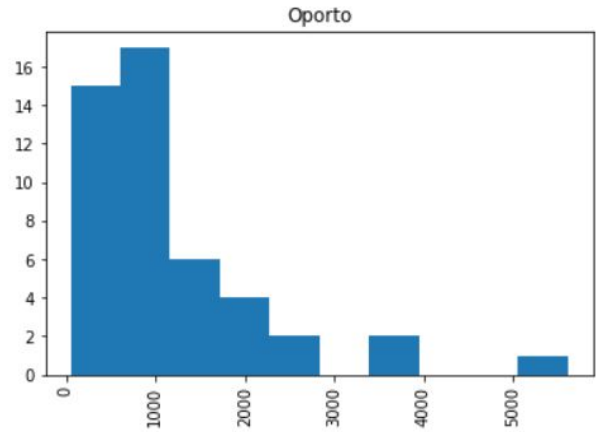
Grocery



Frozen

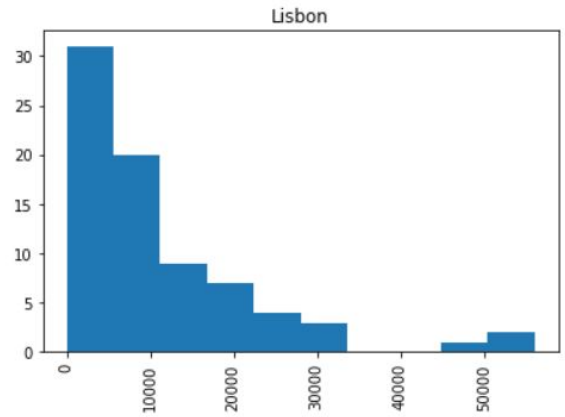


Detergents Paper

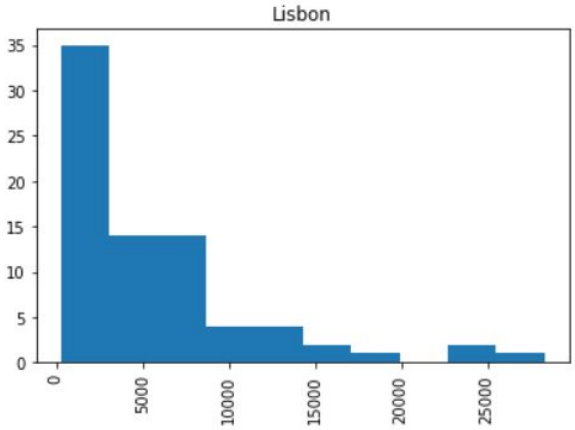


Delicatessen

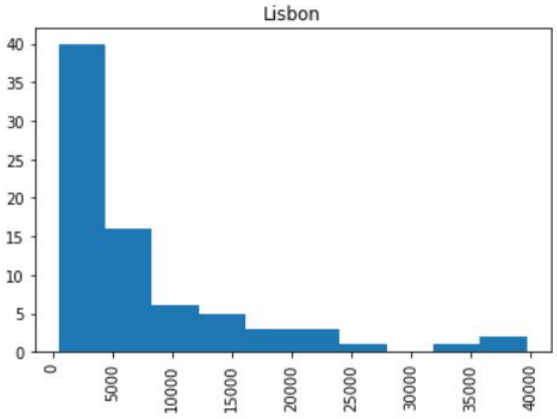
Lisbon:



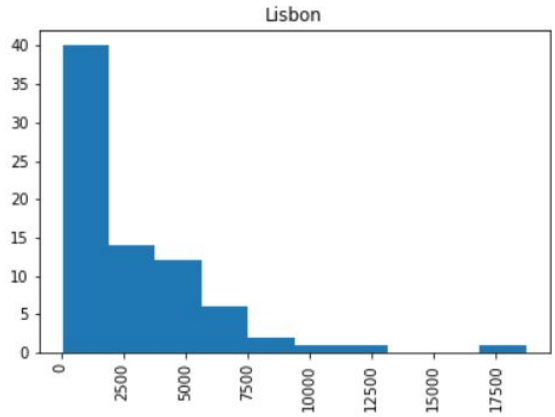
Fresh



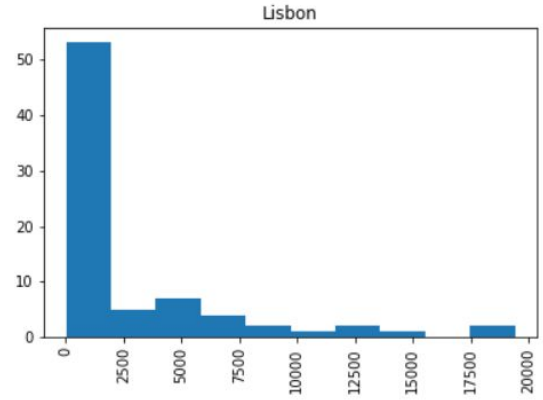
Milk



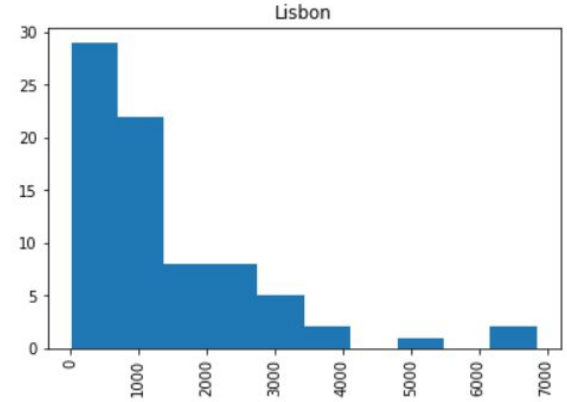
Grocery



Frozen

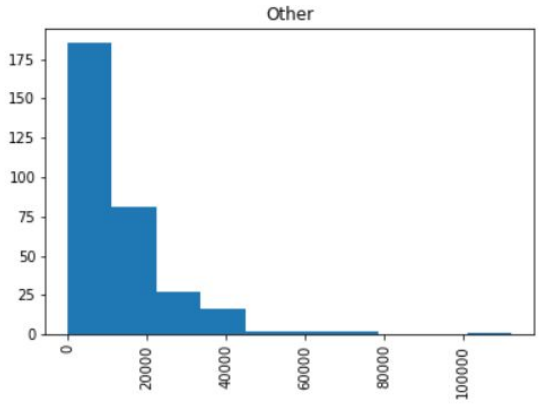


Detergents Paper

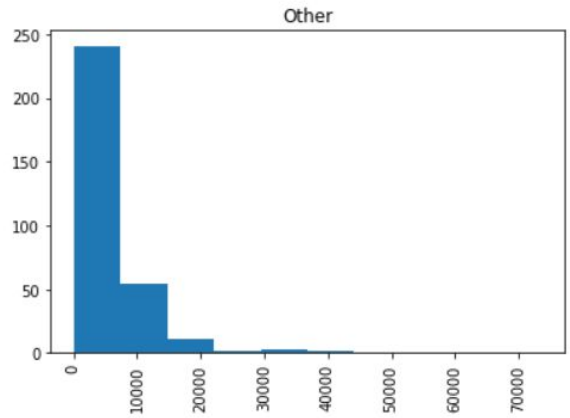


Delicatessen

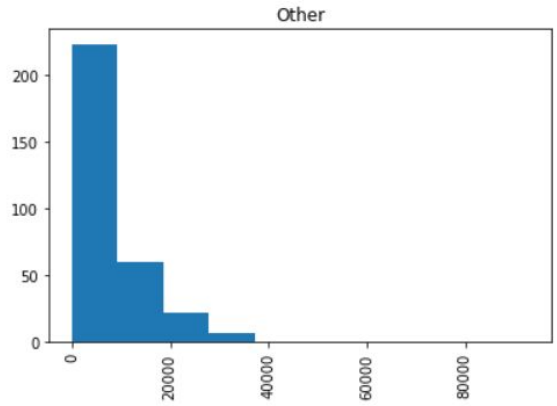
Other:



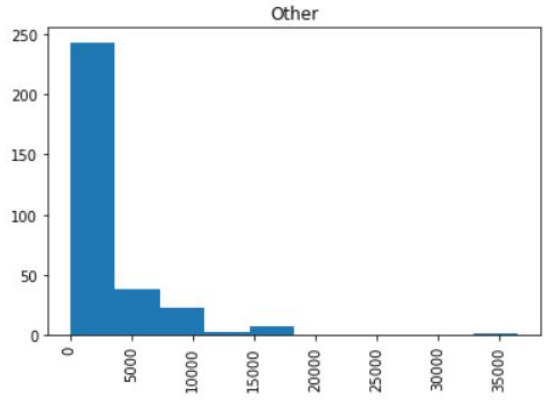
Fresh



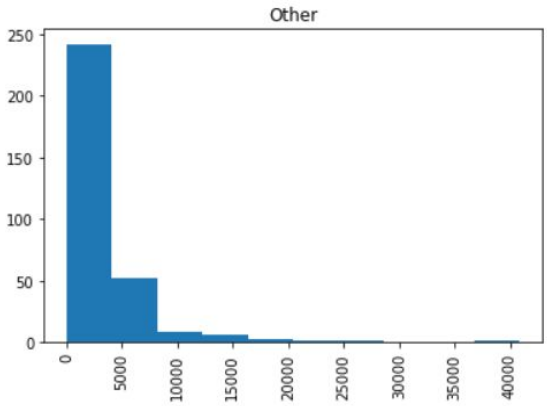
Milk



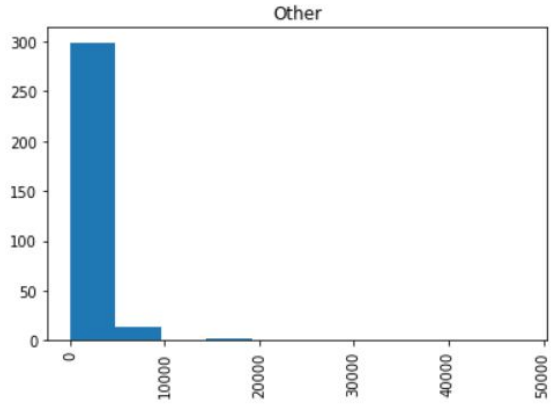
Grocery



Frozen



Detergents Paper



Delicatessen

- All the histograms show similar behaviour as they all are highly skewed to the right, although in some exceptional cases customers buy expensive products

Problem 2

The following are the contingency tables created between Gender and Major, Grad Intention, Employment, Computer

Gender and Major

Gender	Accounting	CIS	Eco/Finance	IB	Manage-ment	Other	Retail/ Market-ting	Undecided	Total
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29

Total	7	4	11	6	10	7	14	3	62

Gender and Grad Intention

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

Gender and Employment

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

Gender and Computer

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

The following shows a table of probabilities under specific conditions

Gender

Condition	Probability
Probability that a random CMSU student will be Male	46.8
Probability that a random CMSU student will be Female	53.2

Gender and Majors

Condition	Probability
Probability that a student chosen at random is an Accounting Major given that it's a male	13.8
Probability that an Accounting Major chosen at random is a male	57.1
Probability that a student chosen at random is an Accounting Major given that it's a female	9.1
Probability that a student chosen at random is a CIS Major given that it's a male	3.4
Probability that a CIS Major chosen at random is a male	25
Probability that a student chosen at random is a CIS Major given that it's a female	9.1
Probability that a CIS Major chosen at random is a female	75
Probability that a student chosen at random is an Economics/Finance Major given that it's a male	13.8
Probability that an Economics/Finance Major chosen at random is a male	36.4
Probability that a student chosen at random is an Economics/Finance Major given that it's a female	21.2
Probability that an Economics/Finance Major chosen at random is a female	63.6
Probability that a student chosen at random is an International Business Major given that it's a male	6.9
Probability that a student chosen at random is an International Business Major given that it's a female	12.1
Probability that an International Business major chosen at random is a female	66.7

Probability that a student chosen at random is a Management Major given that it's a male	20.7
Probability that a Management major chosen at random is a male	60
Probability that a student chosen at random is a Management Major given that it's a female	12.1
Probability that a Management major chosen at random is a female	40
Probability that a student chosen at random is an Other Major given that it's a male	13.8
Probability that an Other major chosen at random is a male	57.1
Probability that a student chosen at random is an Other Major given that it's a female	9.1
Probability that an Other major chosen at random is a female	42.9
Probability that a student chosen at random is a Retailing/Marketing Major given that it's a male	17.2
Probability that a student chosen at random is a Retailing/Marketing Major given that it's a female	27.3
Probability that an Undecided major chosen at random is a male	100
Probability that an Undecided major chosen at random is a female	0

Gender and Grad Intention

Condition	Probability
Probability that a male student chosen at random has intent to graduate	60.7
Probability that a female student chosen at random has intent to graduate	33.3
Probability that a male student chosen at	10.3

random has no intent to graduate	
Probability that a student chosen at random is a male given that he has no intent to graduate	25
Probability that a female student chosen at random has no intent to graduate	27.3
Probability that a male student chosen at random has undecided intent to graduate	31
Probability that a student chosen at random is a male given that he has undecided intent to graduate	40.9
Probability that a female student chosen at random has undecided intent to graduate	39.4

Gender and Employment

Condition	Probability
Probability that a male student chosen at random is a full time employee	24.1
Probability that a female student chosen at random is a full time employee	9.1
Probability that a full time employee chosen at random is a female	30
Probability that a male student chosen at random is a part time employee	65.5
Probability that a part time employee chosen at random is a male	44.2
Probability that a part time employee chosen at random is a female	55.8
Probability that a male student chosen at random is unemployed	10.3
Probability that a female student chosen at random is unemployed	18.2

Gender and Computer

Condition	Probability
Probability that a male student chosen at random has a desktop	10.3
Probability that a female student chosen at random has a desktop	6.1
Probability that a male student chosen at random has a laptop	89.7
Probability that a student chosen at random is a male given that he has a laptop	47.3
Probability that a student chosen at random is a female given that she has a laptop	52.7
Probability that a male student chosen at random has a tablet	0
Probability that a female student chosen at random has a tablet	6.1
Probability that a student chosen at random is a female given that she has a tablet	100

- Based on the above probabilities Grad Intention and Employment can be said to be dependent on gender.
- In Grad Intention the probability of females saying yes is less compared to males
- Similarly in Employment the probability of females employed full time is low while unemployment is vice-versa
- In Majors and computers the gender has no effect on the probabilities

Problem 3

Hypothesis testing

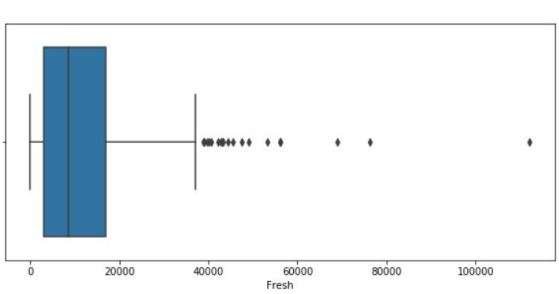
- The Null Hypothesis is that the mean of group A is equal to mean of group B
- The Alternate Hypothesis is that the mean of group A and mean of group B is not equal

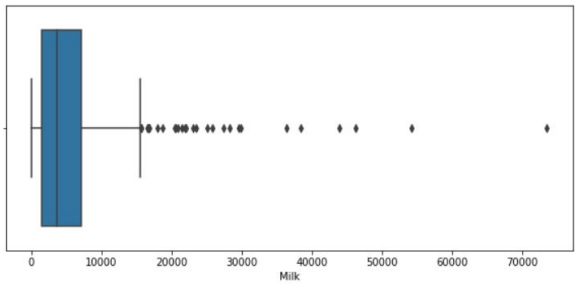
- The statistical test conducted is 2 sampleT test which can be conducted for two samples of different sizes and comparing means
- Upon testing the P-value calculated was 0.20
- At the 0.05 level of significance we fail to reject the Null hypothesis

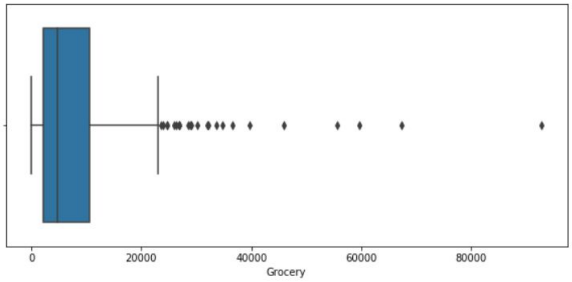
3.6 Outlier Identification

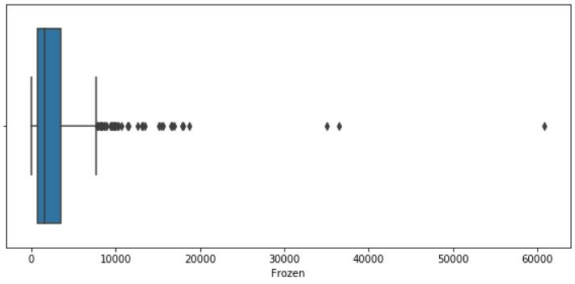
The datasets for problem 1 and 3 were checked for outliers

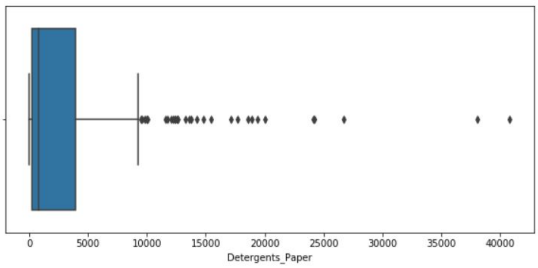
Problem 1

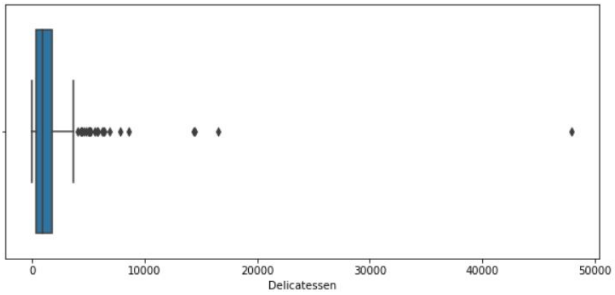






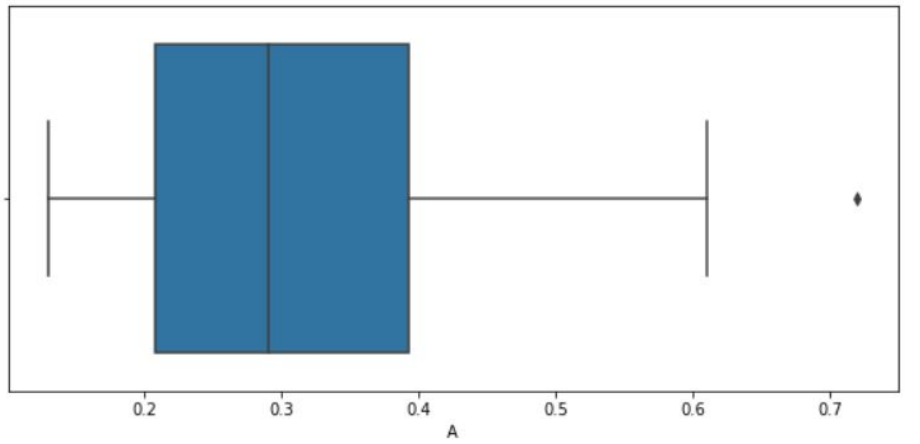


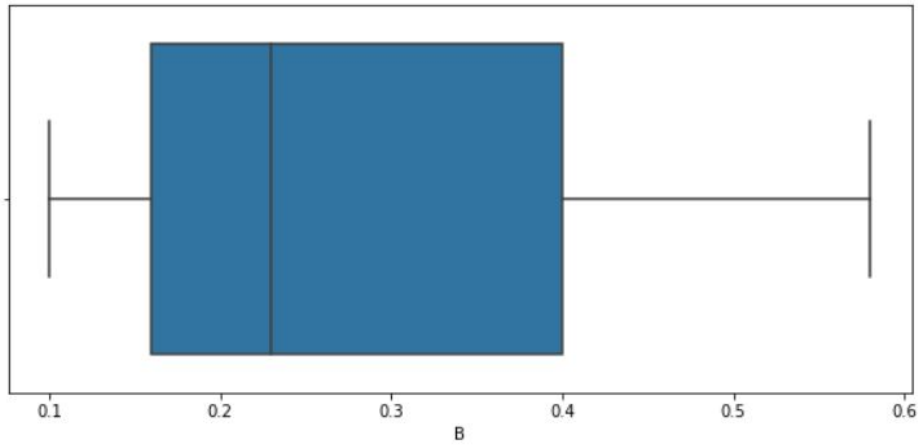




- All six items have outliers

Problem 3





- Group A has a single outlier

4 Conclusion

In Problem 1 the annual spending for each item has numerous outliers, and are all positively skewed, I would recommend that the Wholesale Distributor make Hotel as the primary channel and there is a lot of potential for the business in Other region

In Problem 2 various conditional probabilities were calculated from the survey data. It can be inferred that there is a nearly equal spread of students from various majors. Most of the students intend to graduate and are either employed full-time or partly.
The most commonly used computer is Laptop.

In Problem 3 Hypothesis Testing was conducted for both groups, which resulted in failure to reject the Null hypothesis in both cases. The data provided was too small for population testing, and thus may not be accurate.

5 Appendix A - Source Code

Refer Python file attached