Project

Karthiheswar

Table of Contents

1 Project Objective

The objective of the report is to explore all the projects data set in Python and generate insights about the data set. 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
- Exploring the outliers in dataset
- Probability
- Testing of hypothesis

2 Wholesale Customers Data

2.1 Descriptive statistics to summarize data

	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

Region and Channel which seems to spend more:

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	total
85	86	Retail	Other	16117	46197	92780	1026	40827	2944	199891

The retail channel in other region has spent more with total of 199891
Whereas, Other region has spent 10677599 and Hotel channel has spent 7999569

Region and Channel which seems to spend less:

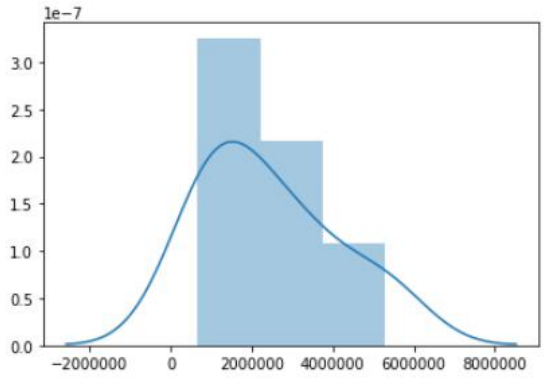
	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	total
154	155	Hotel	Other	622	55	137	75	7	8	904

The hotel channel in other region has spent less with total of 904
Whereas, Oporto region has spent 1555088 and Retail channel has spent 6619931

2.2 Behaviour of all varieties across Region and Channel

Fresh_total	Milk_total	Grocery_total	Frozen_total	Detergents_Paper_total	Delicatessen_total
5280131	2550357	3498562	1351650	1267857	670943

All varieties doesn't show any similar behaviour across Region and Channel



2.3 Descriptive measure of variability

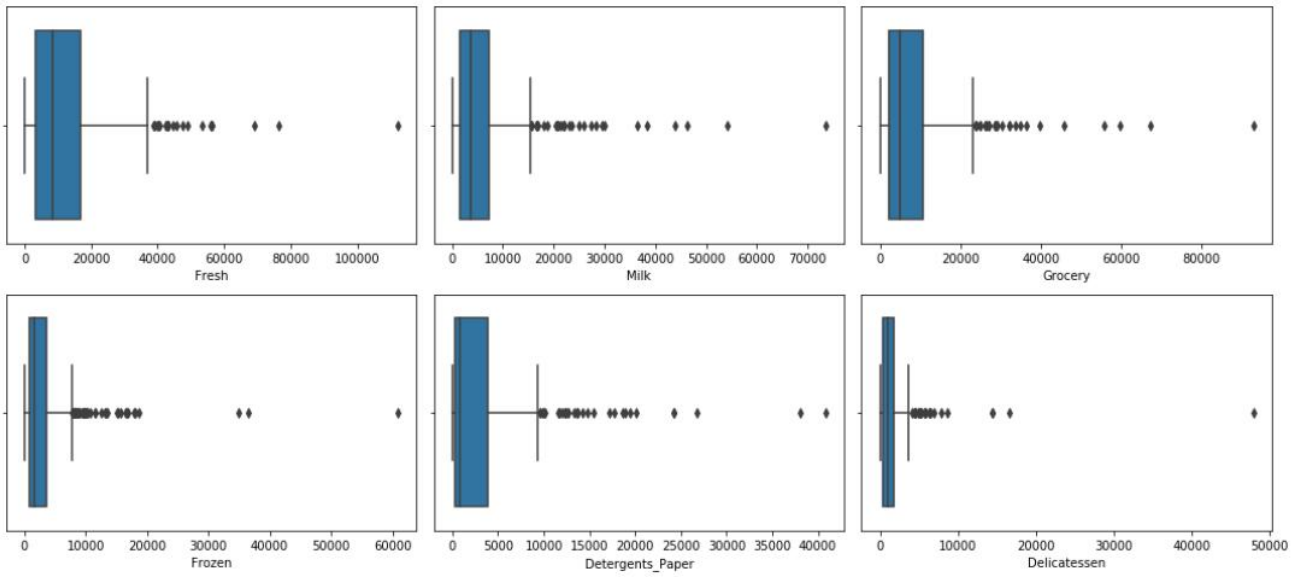
Item which shows the most inconsistent behaviour:

Fresh is the item which has most inconsistent behaviour with standard deviation of 12647.328865

Which items shows the least inconsistent behaviour:

Delicatessen is the item which has least inconsistent behaviour with standard deviation of 2820.105937

2.4 Outliers in data



The above graph displays the outliers in each variety of items

2.5 Recommendations

The frozen and detergent papers sales have to be increased across all channels and regions. Also fresh items shows a huge inconsistent behaviour which has to be corrected. The hotel channel in other region sales transaction has to be improved.

3 Clear Mountain State University (CMSU) Survey

3.1 Contingency tables by keeping gender as row

Gender and Major

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

Gender and Grad Intention

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

Gender and Employment

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

Gender and Computer

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

3.2 Sample of the population of CMSU

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer	Text Messages
13	14	Male	22	Senior	International Business	Undecided	3.1	Part-Time	40.0	1	3	400	Laptop	150
17	18	Male	21	Junior	Economics/Finance	Undecided	3.1	Part-Time	55.0	2	3	600	Laptop	300
49	50	Female	21	Senior	Economics/Finance	Undecided	3.0	Part-Time	45.0	1	3	520	Laptop	105
52	53	Female	21	Senior	Retailing/Marketing	Undecided	3.7	Part-Time	40.0	3	4	300	Laptop	700
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	1	3	1000	Laptop	10
22	23	Female	22	Senior	Retailing/Marketing	Undecided	3.0	Part-Time	55.0	0	4	300	Laptop	35
15	16	Male	24	Senior	Management	Undecided	3.4	Part-Time	45.0	4	4	500	Laptop	175
25	26	Male	24	Senior	Management	Yes	3.3	Full-Time	60.0	0	1	300	Laptop	40
7	8	Female	22	Senior	Other	Undecided	3.1	Full-Time	80.0	1	2	200	Tablet	300
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop	200
60	61	Female	23	Senior	Accounting	Yes	3.5	Part-Time	30.0	2	3	490	Laptop	50
11	12	Male	21	Senior	Undecided	No	3.5	Full-Time	37.0	2	3	500	Laptop	100
14	15	Male	21	Senior	Management	Yes	3.2	Part-Time	54.0	3	4	600	Laptop	400
54	55	Male	21	Senior	Other	Yes	3.4	Part-Time	50.0	1	4	250	Desktop	700
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop	200
35	36	Female	26	Junior	Accounting	Yes	3.3	Part-Time	60.0	1	4	450	Desktop	300
48	49	Female	21	Senior	Economics/Finance	Yes	3.2	Part-Time	47.5	2	4	220	Laptop	105
6	7	Female	21	Junior	Other	Undecided	3.0	Part-Time	50.0	1	3	500	Laptop	50
46	47	Female	20	Junior	Retailing/Marketing	Yes	3.5	Unemployed	60.0	1	3	350	Laptop	200
38	39	Male	24	Junior	Economics/Finance	Yes	2.8	Part-Time	50.0	1	6	600	Laptop	50
40	41	Male	22	Junior	Accounting	Yes	3.2	Full-Time	60.0	1	4	680	Desktop	200
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	0	4	500	Laptop	400

The above table is the 35% of sample data in the population of CMSU

- Probability that a randomly selected CMSU student will be male is 50% in the sample population of CMSU.
- Probability that a randomly selected CMSU student will be female is 50% in the sample population of CMSU.

Conditional probability of different majors among the male students and conditional probability of different majors among the female students in CMSU

```
Gender  Major
Female  Other                   0.250000
        Retailing/Marketing     0.250000
        Accounting              0.166667
        Economics/Finance       0.166667
        International Business  0.083333
        Management              0.083333
Male    Management              0.300000
        Economics/Finance       0.200000
        Other                   0.200000
        Accounting              0.100000
        International Business  0.100000
        Undecided               0.100000
Name: Major, dtype: float64
```

Conditional probability of intent to graduate, given that the student is a male and conditional probability of intent to graduate, given that the student is a female

```
Gender  Grad Intention
Female  Yes              0.500000
        Undecided        0.416667
        No               0.083333
Male    Yes              0.600000
        Undecided        0.300000
        No               0.100000
Name: Grad Intention, dtype: float64
```

Conditional probability of employment status for the male students as well as for the female students

```
Gender  Employment
Female  Part-Time     0.666667
        Full-Time     0.166667
        Unemployed    0.166667
Male    Part-Time     0.700000
        Full-Time     0.300000
Name: Employment, dtype: float64
```

Conditional probability of laptop preference among the male students as well as among the female students

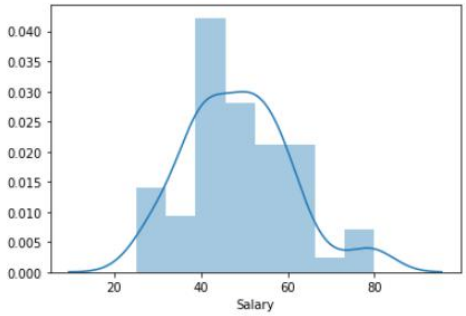
```
Gender  Computer
Female  Laptop      0.833333
        Desktop     0.083333
        Tablet      0.083333
Male    Laptop      0.800000
        Desktop     0.200000
Name: Computer, dtype: float64
```

3.3 Column variable in each case

Each column variable is dependent on gender in each case as each column variable shows a different behaviour across male and female gender

3.4 Normal distribution chart

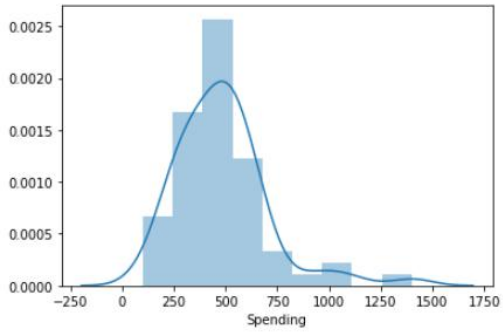
Salary



Mean of salary: 48.54
Median of salary: 50
Mode of salary: 40

Salary shows a normal distribution

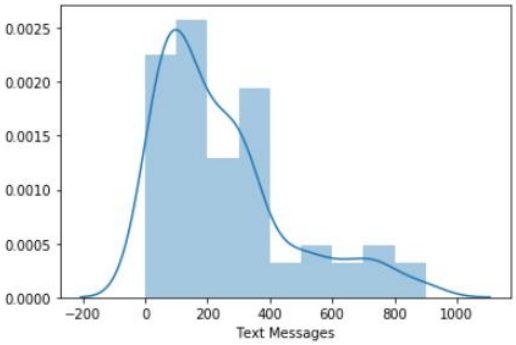
Spending



Mean of spending: 482.01
Median of spending: 500
Mode of spending: 500

Spending shows a normal distribution with right skewed

Text messages



Mean of Text messages: 246.2
Median of Text messages: 200
Mode of Text messages: 300

Text messages shows a normal distribution with right skewed

Summary

The mean of Salary is less than its median and greater than its mode, the shows a normal distribution as the their differences are less noticeable. Even though the mean of Spending is less than its median, the graph shows right skewed behaviour. The median of Text messages is less than its mean, so the graph shows right skewed behaviour.

4 ABC asphalt shingles

Note: Ho is null hypothesis, Ha is alternative hypothesis

4.1 Null and alternative hypothesis for A shingles

A is alternate hypothesis as 0.316<0.35

Ho:mu>=0.35

Ha:mu<0.35

4.2 Test of hypothesis for A shingles

one-sample t-test p-value= 0.14955266289815025

p_value 0.149>0.05

Ho is true

4.3 Null and alternative hypothesis for B shingles

B is alternate hypothesis as 0.273<0.35

Ho:mu>=0.35

Ha:mu<0.35

4.4 Test of hypothesis for B shingles

one-sample t-test p-value= 0.004180954800638363

p_value 0.004180954800638363<0.05

Ho is false

4.5 Test of hypothesis for A&B shingles

Mean of A is 0.316

Mean of B is 0.273

muA!=muB, so it is alternate hypothesis

T_statistic is 1.289628271966112 and p value is 0.2017496571835328

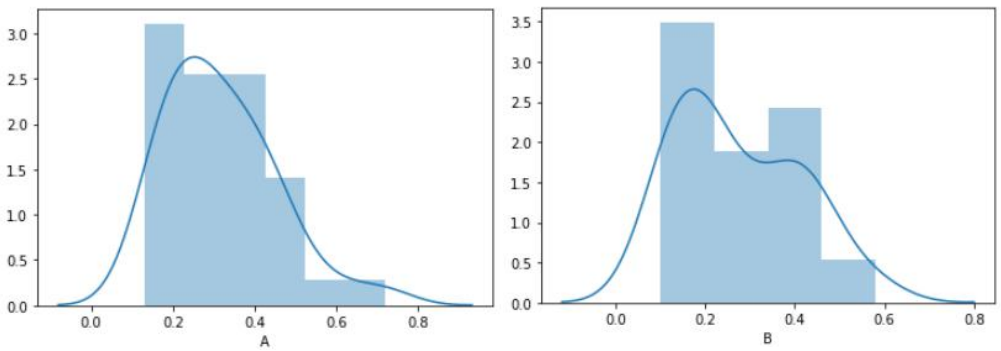
It has to be assumed that if muA=muB, then null hypothesis is true

4.6 Population distribution

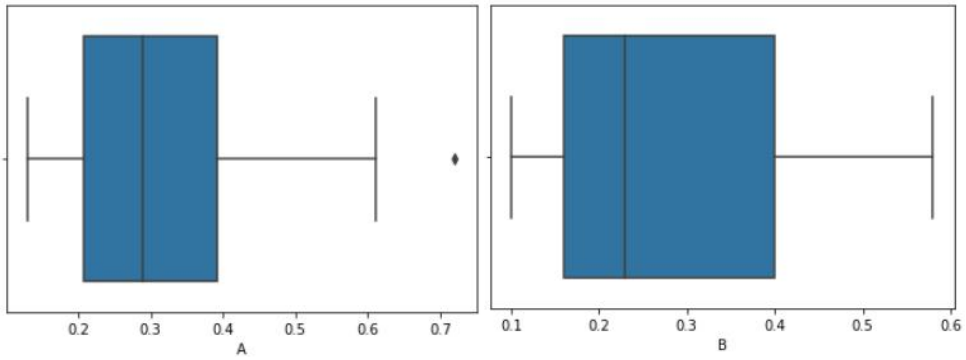
It has to be assumed as normal distribution in order to conduct the hypothesis tests

4.7 Assumptions

Histogram of A and B:



Box plot of A and B:



Empirical rule for A:

One standard deviation (μ ± σ): (0.316 - 0.135) to (0.316 + 0.135), or 0.181 to 0.451

Two standard deviations (μ ± 2σ): 0.316 - (2 x 0.135) to 0.316 + (2 x 0.135), or 0.046 to 0.586

Three standard deviations (μ ± 3σ): 0.316 - (3 x 0.135) to 0.316 + (3 x 0.135), or -0.089 to 0.721

Empirical rule for B:

One standard deviation (μ ± σ): (0.273 - 0.137) to (0.273 + 0.137), or 0.136 to 0.41

Two standard deviations (μ ± 2σ): 0.273 - (2 x 0.137) to 0.273 + (2 x 0.137), or -0.001 to 0.547

Three standard deviations (μ ± 3σ): 0.273 - (3 x 0.137) to 0.273 + (3 x 0.137), or -0.138 to 0.684

4.8 Assumption on hypothesis test

Assumptions are basic ideas on which direction the test has to be carried out. Based on our assumptions it is decided that we reject null hypothesis or we fail to reject null hypothesis.

5 Appendix A – Source Code

Karthiheswar_Stat.ipynb