SMDM Project Report

Abstract

Below is the analysis done for SMDM project. I have analyzed 3 different data sources and published my result in this document. Have analyzed the following data sources
Wholesale Customer data, **Survey** data, **Shingles** data.

Understandings from the Analysis

Wholesale Customer Data:

First we have analysed the Wholesale customer data and published the understandings below.

- The mean of the data is

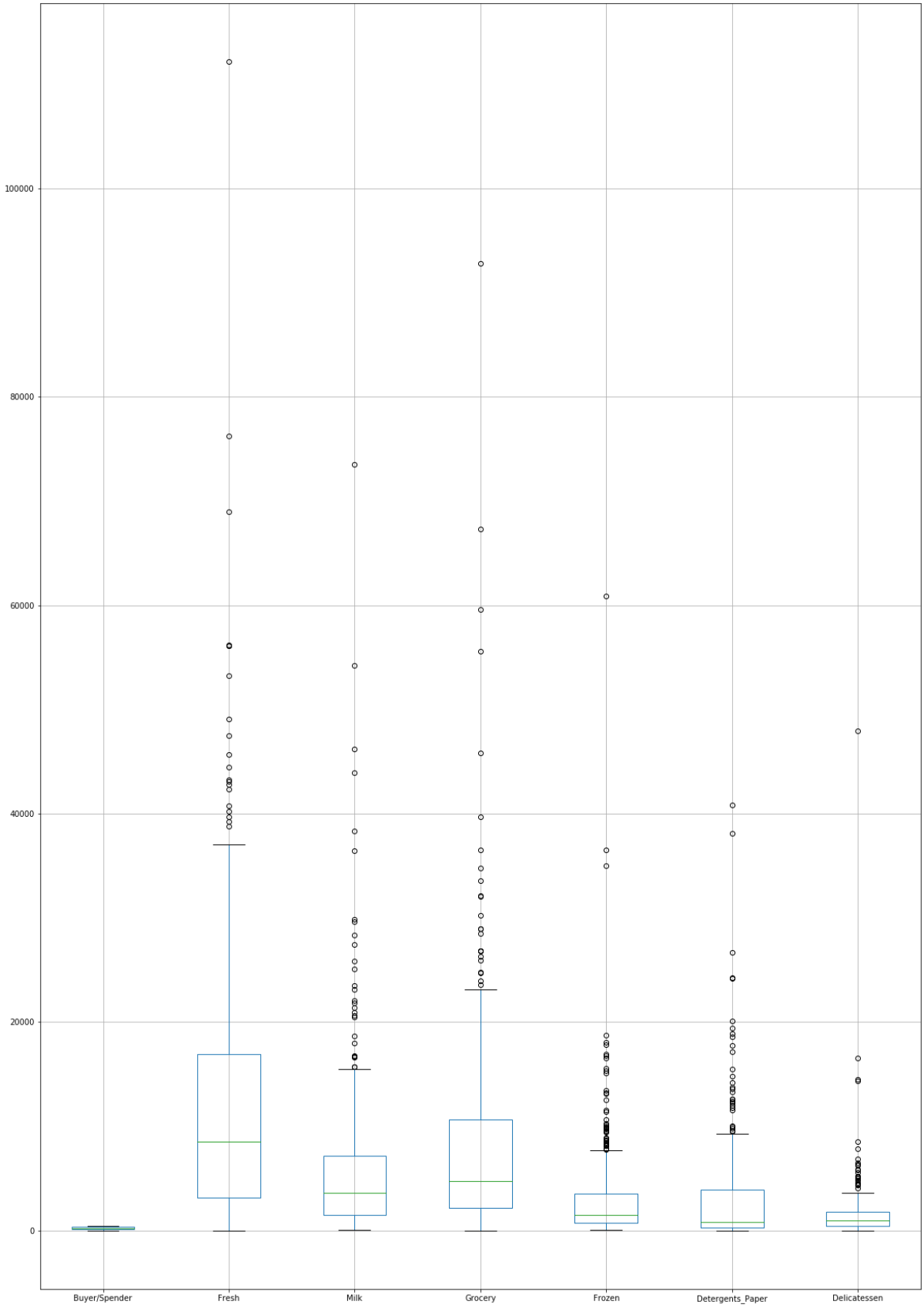
Fresh	**12000.297727**
Milk	**5796.265909**
Grocery	**7951.277273**
Frozen	**3071.931818**
Detergents_Paper	**2881.493182**
Delicatessen	**1524.870455**

- We have identified that **Retail channel** have made **maximum** spend and the **Hotel channel** have made the **minimum** spend
- Also we have identified the **maximum and minimum** spend have been made in **other regions**
- We have also identified the data is not behaving normal across regions and channels

Fresh	12647.328865
Milk	7380.377175
Grocery	9503.162829
Frozen	4854.673333
Detergents_Paper	4767.854448
Delicatessen	2820.105937

- The following item shows the most consistent behaviour across region and channel Delicatessen
- The following item shows the most inconsistent behaviour across region and channel Fresh

- There are few outliers in the data which can be clearly seen through the box plot put below. The **Fresh** item has a significant outliers



- The recommendation from the analysis is there is huge sale happening in Other regions in Retail channel. There are few

outliers in the data which states the sale is happening in an inconsistent way. Also the other regions are not performing as well as other regions. Need to concentrate more on the other 2 regions.

Survey Data:

Next we have analysed the survey data and put down all over identifications below,

- When we looked into the data we identified there is **no null record in the dataset**
- We have built a *contingency* table with **Gender** and other data and below are the identification
    - ***For Gender and Majors***

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided
Female	3	3	7	4	4	3	9	0
Male	4	1	4	2	6	4	5	3

    - ***Gender and Grad Intention***

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

    - ***Gender and Employment***

Employment Gender	Full-Time	Part-Time	Unemployed
Female	3	24	6
Male	7	19	3

- ***Gender and Computer***

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

- Then we started calculating probability and we have identified below things
  - probability that a **randomly selected** CMSU student will be **male** is **46.78%**
  - probability that a **randomly selected** CMSU student will be **female** is **53.22%**
  - **conditional probability** of different **majors** among the **male** students in CMSU

Gender Major	Female	Male
Accounting	3	4
CIS	3	1
Economics/Finance	7	4
International Business	4	2
Management	4	6
Other	3	4
Retailing/Marketing	9	5
Undecided	0	3

Major	Conditional Probability (Male) in %
Accounting	13.80
Economics/Finance	13.80
CIS	3.40
International Business	6.89
Management	20.68
Other	13.79
Retailing/Marketing	17.24
Undecided	10.34

Major	Conditional Probability (Female) in %
Accounting	9.09
Economics/Finance	21.21
CIS	9.09
International Business	12.12
Management	12.12
Other	9.09
Retailing/Marketing	27.27
Undecided	0

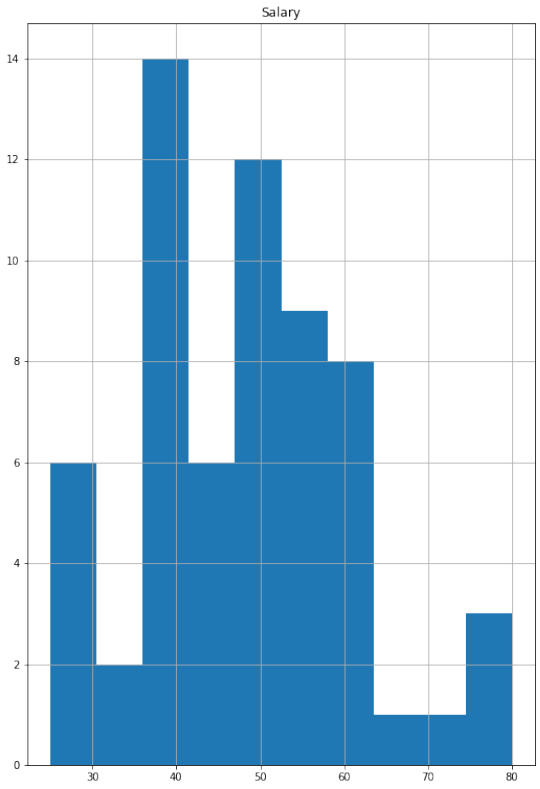
- probability of **intent to graduate**, given that the student is a **male** is **58.62%**
- **c**onditional probability of **intent to graduate**, given that the student is a **female** is **33.33%**
- conditional probability of **employment status** for the **male** students **Fulltime** is **24.13%**
- conditional probability of **employment status** for the **male** students **part-time** is **65.51%**
- conditional probability of **employment status** for the **male** students **Unemployed** is **10.34%**
- conditional probability of **employment status** for the **female** students **Fulltime** is **9.09%**

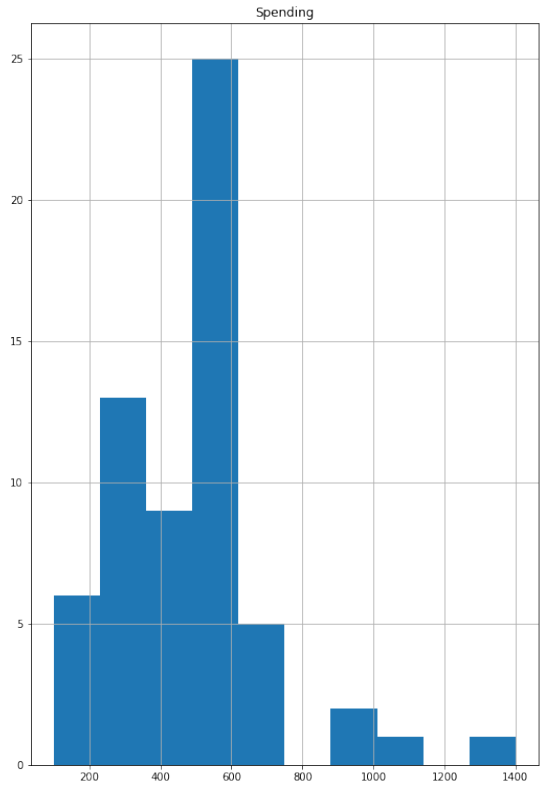
- conditional probability of **employment status** for the **female** students **part-time** is **72.72%**
- conditional probability of **employment status** for the **female** students **Unemployed** is **18.18%**
- **laptop** preference among the **male** students is **89.65%**
- **laptop** preference among the **female** students is **87.87%**
- By Analysing the above record we can get to conclusion that the data is **dependent** on **gender** as you can clearly see the conditional probability based on gender differ from each attributes and they are strongly dependant on the gender info. For (eg) **Male** students are **more intent** to **graduate** compared to **female** students and also **Female** students prefer **Economics/Finance more** compared to **male** students

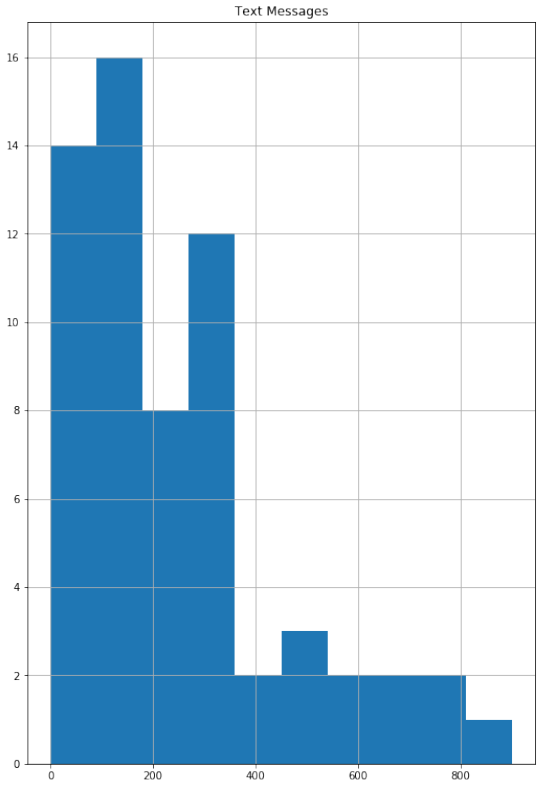
- The three columns data set **Salary, Spending and Text Messages** we tried to identify whether they are **normally distributed** or not. The results are below

	Skewness
Salary	0.521677
Spending	1.547285
Text Messages	1.264245

Salary: salary is uniformly distributes hardly any skew in it
Spending: Spending is highly skewed hence its not normally distributed
Text Messages: Text messages are also highly skewed and its also not normally distributed.


Salary
0
2
4
6
8
10
12
14
30
40
50
60
70
80


Spending
0
5
10
15
20
25
200
400
600
800
1000
1200
1400


Text Messages
0
2
4
6
8
10
12
14
16
0
200
400
600
800

Shingles Data:

Next we analysed **shingles** data for **hypothesis** testing. We have got 2 samples of A & B shingles moisture content and we are going to perform hypothesis testing on it

- For **A Shingles** we conducted **null and alternative hypothesis** to test whether the population mean moisture content is less than **0.35** pound **per 100 square feet**. We performed **a 1 sample T test** and identified the **P value as 0.119** . On **5%** significance level we failed to **reject** the **null hypothesis** and say the moisture content from sample B is **less** than **0.35** pound per 100 sqft
- For **B Shingles** we conducted null and alternative hypothesis to test whether the population mean moisture content is less than **0.35** pound per 100 square feet. We performed a **1 sample T test** and identified the **P value** as **0.004** . On **5%** significance level we reject the null hypothesis and say the moisture content from **sample B** is **not less** than **0.35** pound per 100 sqft
- For the test for equality of means is performed we need to make sure
    - Both the samples needs to be **normally distributed**.
    - **Mean** and **Median** Values should not have much difference.
- To identify population means for shingles A and B are equal we have performed **individual T test** and found that the **P** value is **0.32** which is good at 5% significance level
- The common assumptions made when doing a t-test include those regarding the scale of measurement, random sampling, normality of data distribution, adequacy of sample size and equality of variance in standard deviation.