SMDM Project

09.05.2021

Vinay Santosh

Post Graduate Program in Data Science and Business Analytics (Online)

Mar_B 2021

Contents:

List Of Tables:

List Of Figures:

Project Overview

Solving problems based on the concepts covered in the module: Statistical Methods for Decision Making. The key concepts used to identify and solve the problems are Descriptive Statistics, Inferential Statistics, Hypothesis Testing and Estimation.

The tools used to solve these problems are Python and the various libraries that the language supports such as Pandas, Numpy, Scipy, Matplotlib etc. These libraries allow us to do various arithmetic operations and help in visualizing the data. The findings of the problems and it's code are extensively laid out in the python file attached to this assignment.

Goals

- ❏ To read the dataset and understand the underlying problem.
- ❏ Implement various tools and practices to scrutinize the data.
- ❏ To identify patterns and irregularities in the dataset.
- ❏ Calculating probabilities and constructing hypotheses to predict outcomes.
- ❏ Visualize the data for better understanding of the problem.
- ❏ Make inferences from the findings and derive calculated solutions.
- ❏ Provide a detailed report and make suggestions to implement a better business approach.

Problem 1: Wholesale Distributor

In this problem we look at a dataset with the annual spendings of 440 different retailers across different regions of Portugal. The two channels through which the retailers spend are Hotels and Retail Stores. In the dataset, there is information on six items that the retailers spend on annually.

1.1 Data Summary:

By summarizing the given dataset, we are able to observe key values such as the mean, minimum and maximum spending of items across all Regions and Channels.

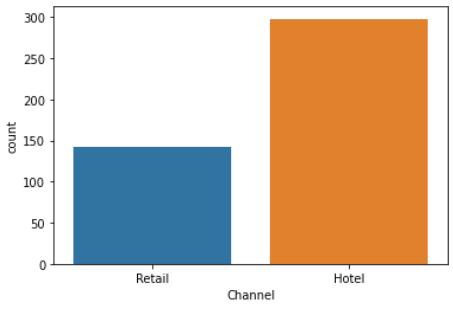
	count	mean	std	min	25%	50%	75%	max
Buyer/Spender	440.0	220.500000	127.161315	1.0	110.75	220.5	330.25	440.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

Table 1.1

We find that the average spending of the variable “Fresh” is higher than any of the other items and the lowest average spending being “Delicatessen”.

Categorical Total:

By Channel:



By Region:

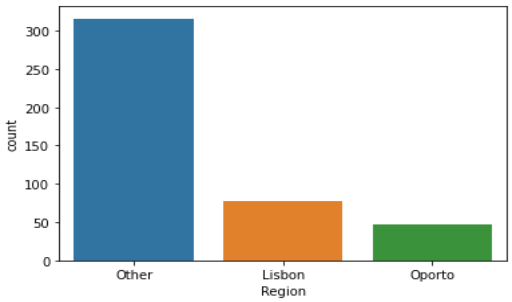


Fig 1.1

- Hotel : 298
- Retail : 142
- Other : 316
- Lisbon : 77
- Oporto: 47

The categorical total of the regions and channels are represented as above. It is evident that there are more hotels that buy from the distributor and the majority of the spenders are from parts outside of Lisbon and Oporto.

Highest & Lowest Spenders:

By Channel:

1.1.2: From the information above we can clearly see that Hotels spend the most with a total of 7,999,569 whereas Retail spends a total of 6,619,931 on all six items.

Channel	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Hotel	71034	4015717	1028614	1180717	1116979	235587	421955	7999569
Retail	25986	1264414	1521743	2317845	234671	1032270	248988	6619931

Tabel:1.1.2

By Region:

1.1.3: We can also find out that regions outside of Lisbon and Oporto contribute the highest in spending (10,677,599), Oporto having spent the least(1,555,088)

Region	Buyer/Spender	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen	Total
Lisbon	18095	854833	422454	570037	231026	204136	104327	2386813
Oporto	14899	464721	239144	433274	190132	173311	54506	1555088
Other	64026	3960577	1888759	2495251	930492	890410	512110	10677599

Tabel:1.1.3

1.2 Distribution of the varieties of items:

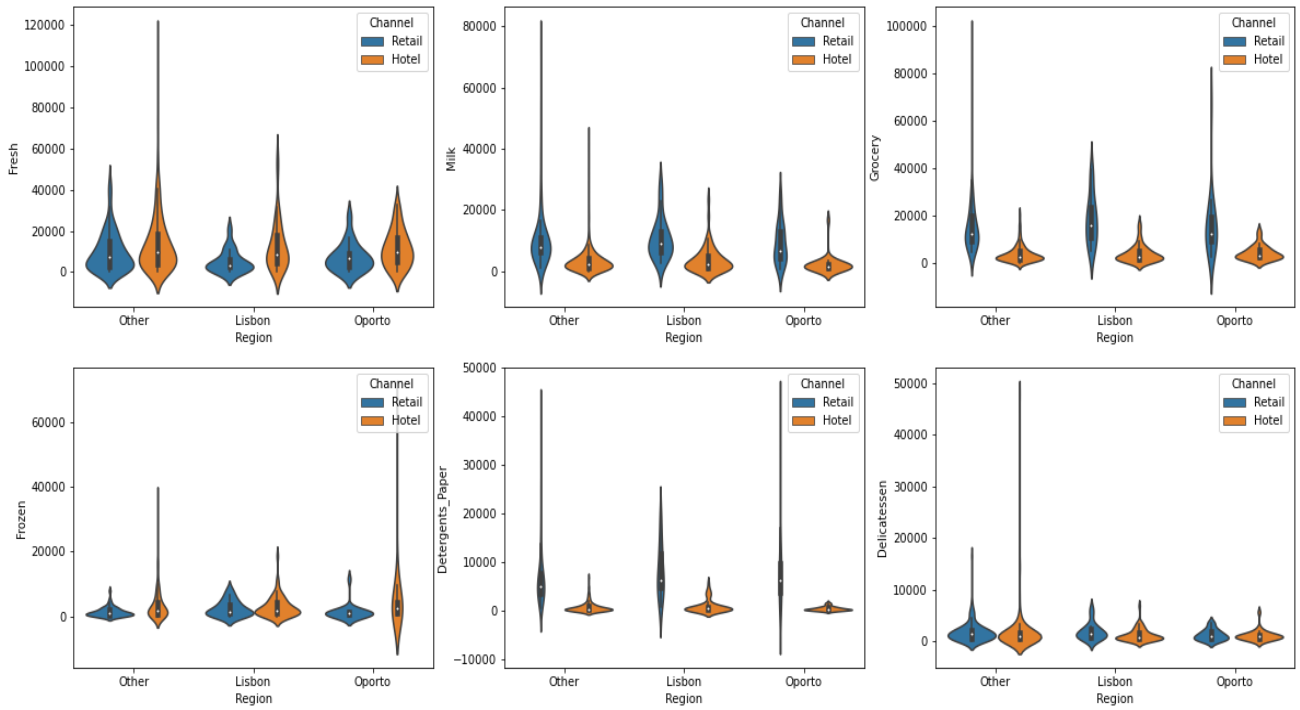


Fig: 1.2

* **Fresh:** We know that this variety has the highest contribution across all regions with an Avg of 12000 euros spent. We can see that Hotels spend more on fresh produce as compared to retail stores across all regions.

* **Milk** : We can see that Retail clients spend more on "Milk" as compared to hotels across all regions.

* **Grocery** : We can see that Retail clients spend more on the variety "Grocery" as compared to Hotel clients across all regions.

* **Frozen:** We can see from the plots above that Hotel clients spend more on "Frozen" goods as compared to retail stores.

* **Detergents_Paper:** We can see that Retail clients spend more on Detergents_Paper as compared to hotels across all regions.

* **Delicatessen:** We know that this variety contributes to the lowest spending across all regions with an Avg of 1525 euros spent. We can see that the spending is uniform between the channels as compared to the other variables.

We can thus infer that the purchases made by the channels Retail and Hotel fluctuates based on the item. "Fresh" being the item that all of the clients seem to spend on regularly and generates the most sales. Retailers spend more on Milk, Grocery and Detergents/Paper. Hotels spend more on Frozen and Delicatessen items.

1.3 Consistency of Variables:

To measure the consistency of a variable, we calculate it's coefficient of variance which gives us an idea of how consistent the items are being purchased from the distributor. The coefficient of variance (cov) is the ratio of the standard deviation to it's mean. The table below shows the cov of all the items in the dataset along with its range and Interquartile range.

	count	mean	std	min	25%	50%	75%	max	range	iqr	cov
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0	112148.0	13806.00	1.053918
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0	73443.0	5657.25	1.273299
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0	92777.0	8502.75	1.195174
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0	60844.0	2812.00	1.580332
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0	40824.0	3665.25	1.654647
Delicatessen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0	47940.0	1412.00	1.849407

Tabel:1.3

We can thus conclude that the item "Delicatessen" has the highest coefficient of variance (1.85) and hence displays the least inconsistent behavior while the item "Fresh" displays the most inconsistent behavior (cov = 1.05) in the dataset. We can plot the values of the variables to visually see it's inconsistency in distribution.

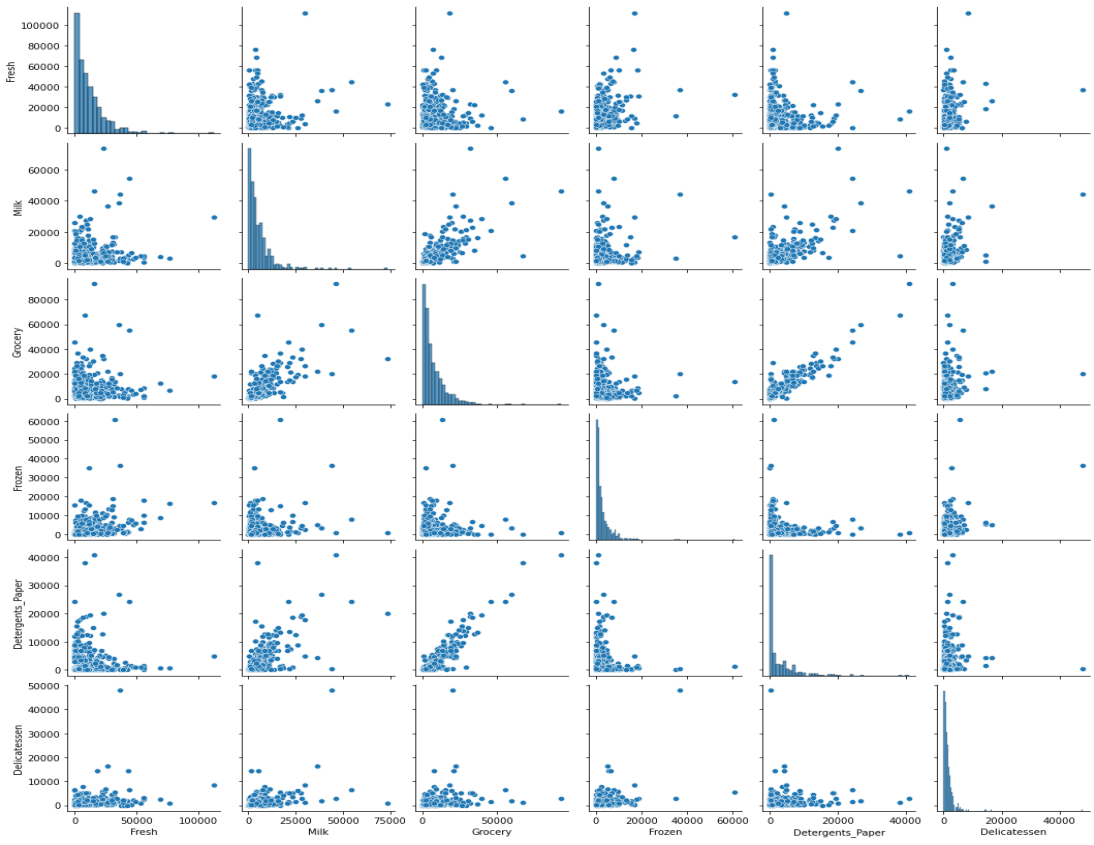


Fig:1.3

1.4 Identifying Outliers:

From the above box plots we can clearly see there are outliers present in the dataset across all regions and channels. It is evident that the items purchased by Retail outlets, in general, have the least amount of inconsistency. The outliers indicate that most of there are multiple purchases made that are above the third quartile. This indicates that the mean expenditures do not reflect on the actual sales.

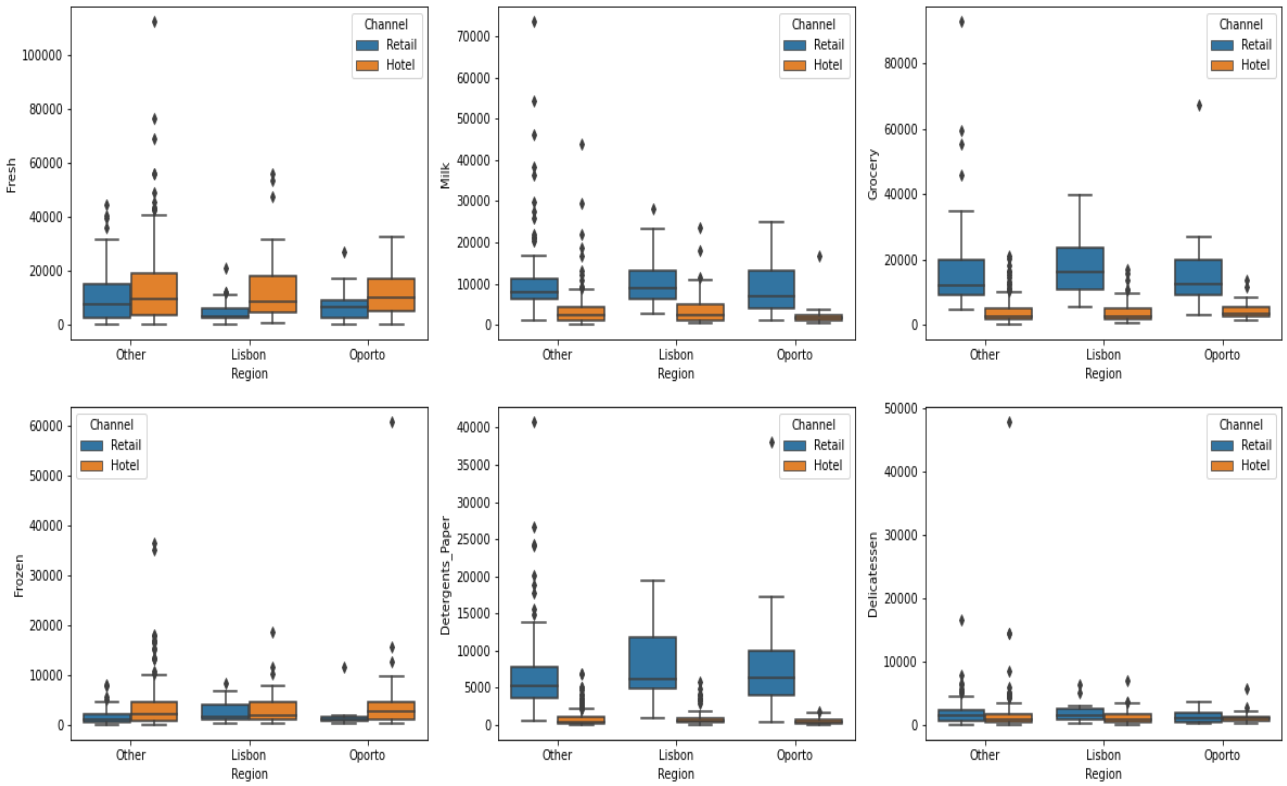


Fig1.4

1.5 Analysis, Insights and Business Recommendations:

- The items “Delicatessen” and “Frozen” seem to be consistent in sales with Hotel clients. Introducing more of these products especially in Retail stores will help boost the overall sales.
- The item “Fresh” generates the highest sales but lacks consistency. Improving the quality of these products and making it easily accessible to the clients will help improve sales.
- Marketing “Grocery”, “Milk” and “Detergent” products to suit the needs of Hotels will drastically impact sales as they seem to be the highest spending channel.

Problem 2: CMSU Student Survey

_____In this problem, we have a dataset with the information of undergraduate students who attend CMSU. The dataset is made up for information collected from a survey of 14 questions answered by a total of 62 students.

	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

Table 2.1

Contingency Tables:

By constructing contingency tables, we are able to predict the behavior of students based on the information they have provided. It also helps us compare the behaviors of male and female students.

2.1.1: Table for 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

Tabel 2.1.1

2.1.2: Table for Gender and Graduation Intention:

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

Tabel 2.1.2

2.1.3: Table for Gender and Employment:

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

Tabel 2.1.3

2.1.4: Table for Gender and Computer:

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

Tabel 2.1.4

2.2 Calculating probabilities of selecting a random student:

2.2.1 : The probability that a randomly selected student will be male is: **46.77 %**

2.2.2 : The probability that a randomly selected student will be female is: **53.23 %**

2.3 Calculating Conditional probability of different majors given gender of the student:

2.3.1 , 2.3.2 :

The conditional probability among male students based on their majors are :

Majors	Male	Female
Accounting:	**13.8 %**	**9.1 %**
CIS:	**3.5 %**	**9.1 %**
Economics/Finance:	**13.8%**	**21.1%**
International Business:	**7%**	**12.1%**
Management:	**20.7%**	**12.1%**
Other:	**13.8%**	**9.1%**
Retailing/Marketing:	**17.2%**	**27.3%**
Undecided:	**10.3%**	**0%**

Table2.3.1

2.4 When we assume that the given sample of 62 is the representation of the population, we can make further predictions:

2.4.1 The probability that a randomly selected student is a male who intends to graduate is : **28.4 %**

2.4.2 The probability that a randomly selected student is a female and does not have a laptop is: **6.45 %**

2.5.1 The probability that a randomly chosen student is male or has a full time job: **38.77 %**

2.5.2 The conditional probability that the randomly chosen female student is majoring in International Business or Management is: **48.48 %**

2.6 Independency of events :

Grad Intention	No	Yes
Gender		
Female	9	11
Male	3	17

Tabel 2.6

The two events are independent because a randomly selected student can be a female and can either have the intention to graduate or not.

2.7 Calculating probability of numerical variables in the dataset:

2.7.1: Probability of a student with GPA less than 3:

GPA	2.3	2.4	2.5	2.6	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
Gender																
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0

Table 2.7.1

The probability of the student's GPA being less than 3 is: **27.42 %**

2.7.2 Conditional probability of Salary given Gender:

Salary	25.0	30.0	35.0	37.0	37.5	40.0	42.0	45.0	47.0	47.5	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0
Gender																			
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1

Table 2.7.2

The conditional probability that a selected male earns 50 or more is: **43.75 %**
The conditional probability that a selected female earns 50 or more is: **56.25 %**

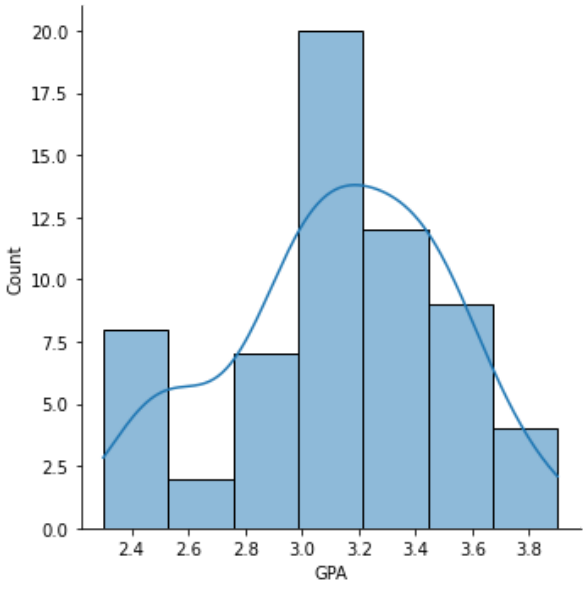
2.8.1 Distribution of continuous variables in the dataset

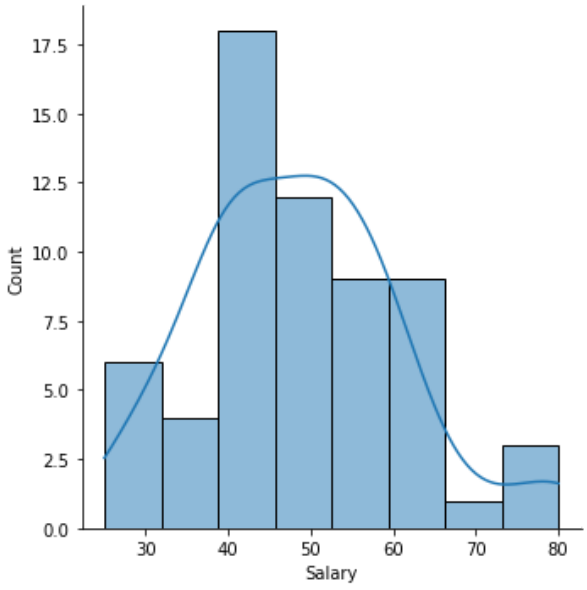
The four numerical variables in the dataset are GPA, Salary, Spending and Text Messages. To find it's distribution, we get the statistical description of these variables and plot the respective values.

	count	mean	std	min	25%	50%	75%	max
GPA	62.0	3.129032	0.377388	2.3	2.9	3.15	3.4	3.9
Salary	62.0	48.548387	12.080912	25.0	40.0	50.00	55.0	80.0
Spending	62.0	482.016129	221.953805	100.0	312.5	500.00	600.0	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.0	200.00	300.0	900.0

Table 2.8.1

From the below distribution plots, we can observe the distribution of the four variables. We can see that GPA and Salary follow a normal distribution while Spending and Text Messages show a bit of right skewness in data.





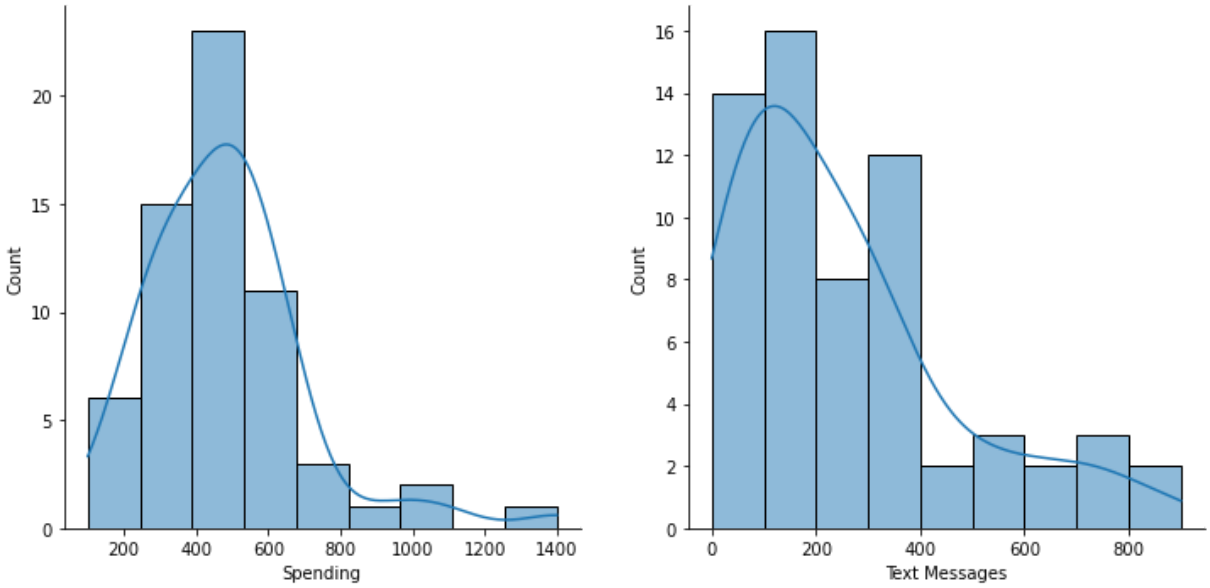


Fig 2.8.1

2.8.2 Problem Summary:

From the information of the students in the dataset, we can draw the following conclusions:

- More students from Retailing/Marketing answered the survey than any other major.
- The majority of students who have the intention to graduate are male with 35% of all students who remain undecided.
- Majority of the students (69%) who attend the university have part-time jobs.
- 88% of the students have laptops.
- The average GPA of all the students is 3.1and the average salary of the students is 48.5.
- The average spending of the students is 482 with a maximum spending of 1400 and a minimum of 100.

Problem 3: Moisture content in ABC Asphalt Shingles:

In this problem we have information from a test conducted by ABC asphalt shingles manufacturing company to find out the moisture content in the two types of shingles manufactured. The company has set a moisture content limit of 0.35 pound per 100 square feet post the tests. The dataset has two samples of uneven sample sizes.

	count	mean	std	min	25%	50%	75%	max
A	36.0	0.316667	0.135731	0.13	0.2075	0.29	0.3925	0.72
B	31.0	0.273548	0.137296	0.10	0.1600	0.23	0.4000	0.58

Tabel 3.1

3.1 Hypothesizing the moisture contents in both types of shingles:

Step1: Constructing the hypothesis:

Null hypothesis :

H0 : Mean moisture contents of shingles = 0.35 (or >= 0.35)

Alternative hypothesis :

H1 : Mean moisture content of shingles < 0.35

Step 2: Determining the level of significance:

Since there is no information given on the level of significance, we consider it to be 0.05

I.e alpha = **0.05**

Step 3: Identifying the statistical test to be used:

Since the dataset has independent samples with varied sample sizes and the population standard deviation is unknown, we conduct a separate One Sample T- statistic test on each of the types of shingles to find out the test statistic value and the p-value. This is a one tail test.

Step 4: Test statistic and p-value:

T-stat for Shingle A : **-1.6005252585398313**
p-value for Shingle A : **0.11996170801033942**

T-stat for Shingle A : **-3.1003313069986995**
p-value for Shingle A : **0.004180954800638363**

Step 5: Acceptance or Rejection of Null Hypothesis:

Since the p-value of Shingle **type A** is greater than alpha(0.05), we fail to reject the null hypothesis. Hence we have enough evidence to prove that the mean moisture content of the shingle type A is not within the permissible limit of 0.35.

Since the p-value of Shingle **type B** is less than alpha(0.05), we reject the null hypothesis. Hence we can say that at 95% there is evidence to prove that the mean moisture content of the shingle type A is within the permissible limit of 0.35 (proves right the alternative hypothesis).

3.2 Test for equality of means:

Step1: Constructing the hypothesis:

Null hypothesis :

H0 : uA = uB

Alternative hypothesis :

H1 : uA ≠ uB

Step 2: Determining the level of significance:

Since there is no information given on the level of significance, we consider it to be 0.05

I.e alpha = **0.05**

Step 3: Assumptions and Identifying the statistical test to be used:

Let us assume that the population variance is equal and the data is normally distributed. Since the dataset there are two samples and the population standard deviation is unknown. The sample size is not very large and so we conduct a Two Sample T- statistic test.

Step 4: Test statistic and p-value:

T-stat : **1.289628271966112**
p-value : **0.2017496571835328**

P-value > alpha

Step 5: Acceptance or Rejection of Null Hypothesis:

Since the p-value is greater than alpha(0.05), we fail to reject the null hypothesis. Hence we can say that at 95% the alternative hypothesis is proved right which means the **population means** of the shingle types A and B **are equa**l.