STATISTICAL METHODS FOR DECISION MAKING

Project Report

TABLE OF CONTENTS

*Please refer Appendix for source code

1. PROJECT OBJECTIVE

The objective of the report is to explore the data set for each of the 3 problem sets respectively in Python and generate insights and recommendations for the same. This report will consist of the following:

- Importing the dataset in python;
- Understanding the structure of the data set;
- Graphical exploration;
- Descriptive statistics; and
- Insights from the dataset and recommendations.

2. ASSUMPTIONS

We will make the following assumptions about the data set:

- The population follows the normal distribution; and
- The sample is representative of the population.

3. EXPLORATORY DATA ANALYSIS

A Typical Data exploration activity consists of the following steps:

(a) Environment Set up and Data Import;
(b) Variable Identification;
(c) Univariate Analysis;
(d) Bi-Variate Analysis;
(e) Missing Value Treatment (Not in scope for our project);
(f) Outlier Treatment (Not in scope for our project);
(g) Variable Transformation / Feature Creation; and
(h) Feature Exploration.

We shall follow these steps in exploring the provided datasets.

Although Steps (e) and (f) are not in scope for this project, a brief about these steps (and other steps as well) is given, as these are important steps for Data Exploration journey

(a) Environment Setup and Data Import

(i) Importing the libraries

Before analysing the problem, the first step is to import all the required libraries such as, numpy, pandas, seaborn, matplotlib etc, based on our requirements. Doing this in the beginning helps the readability of the code

(ii) Setting up Working Directory

Setting up a working directory at the start of the Python session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project. You can also upload the required files directly into jupyter notebook and import the required files from there

*Please refer Appendix for source code

Since we have 3 different problems, we will look at each of them individually for all the steps, from exploratory data analysis, to addressing the questions, and drawing insights to conclude and give recommendations

4. ANALYSIS OF 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 Customers Data.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/Restaurant/Café HoReCa, Retail).

(a) Import the Data Set

We use the command 'read.csv' to import the file in python– Wholesale customers' data.

(b) Variable Identification

(i) Data view

After importing the data, we use the df.head() function to view the data to see if the data has been imported properly.

	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

(ii) Checking the summary of the data

We use the df.describe() and df.info function to view the summary of the data. This includes the 5 point summary and other specific details regarding the data. With the info function we can also view the data type and if there are any null values

*Please refer Appendix for source code

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Buyer/Spender	440	NaN	NaN	NaN	220.5	127.161	1	110.75	220.5	330.25	440
Channel	440	2	Hotel	298	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Region	440	3	Other	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fresh	440	NaN	NaN	NaN	12000.3	12647.3	3	3127.75	8504	16933.8	112151
Milk	440	NaN	NaN	NaN	5796.27	7380.38	55	1533	3627	7190.25	73498
Grocery	440	NaN	NaN	NaN	7951.28	9503.16	3	2153	4755.5	10655.8	92780
Frozen	440	NaN	NaN	NaN	3071.93	4854.67	25	742.25	1526	3554.25	60869
Detergents_Paper	440	NaN	NaN	NaN	2881.49	4767.85	3	256.75	816.5	3922	40827
Delicatessen	440	NaN	NaN	NaN	1524.87	2820.11	3	408.25	965.5	1820.25	47943

(iii) Checking for null values

We use the df.isnull() function to check for any blank or NA cells in the data. By using the isnull function by itself we only get a Boolean output. We use the df.isnull().sum() function to get an output of the count of null values in each column. This is helpful in cleaning the data. The rows with NA values can be removed and then one can move further with the analysis. However, we need to be mindful of cases when there are a large number of NA values and the treatment of the same needs to be carried out.

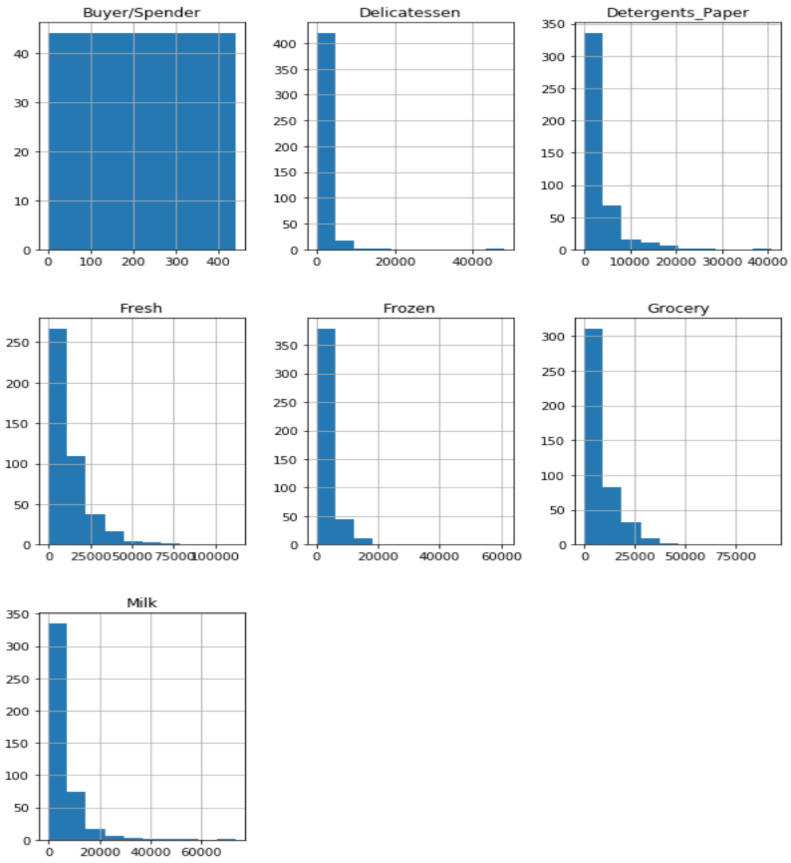
```
Buyer/Spender       0
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicatessen        0
dtype: int64
```

There are no missing values in the dataset

(c) Univariate Analysis

We use the df.hist() function to plot a histogram for all the variables. This helps us to understand the distribution of each variable, whether it is normally distributed, skewed, or uniformly distributed.

*Please refer Appendix for source code

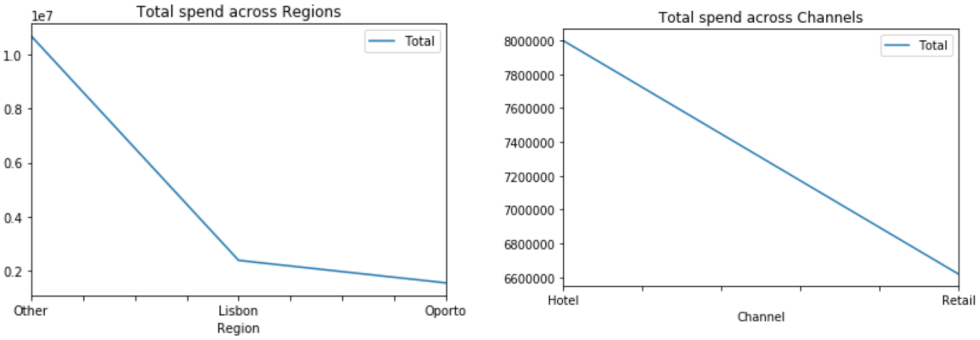


From the histograms we can see that all the variables are highly right skewed. The buyer data is uniformly distributed as all values appear only once.

(d) Bivariate Analysis

Q1.1 Use methods of descriptive statistics to summarize data.
Which Region and which Channel seems to spend more?
Which Region and which Channel seems to spend less?

We use the '.plot' function with kind=line to plot the spend vs region/Channel graphs

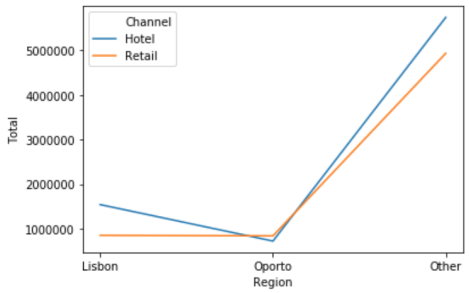


Interpreting separately for Region and Channels

1. Other has the highest spend among region
2. Hotel has the highest spend among channels

*Please refer Appendix for source code

We create a subset of the data using the group.by function. We group the data by Channel and Region. Further we plot line graph with hue= Channel to observe a consolidated output of the spend across region and channel



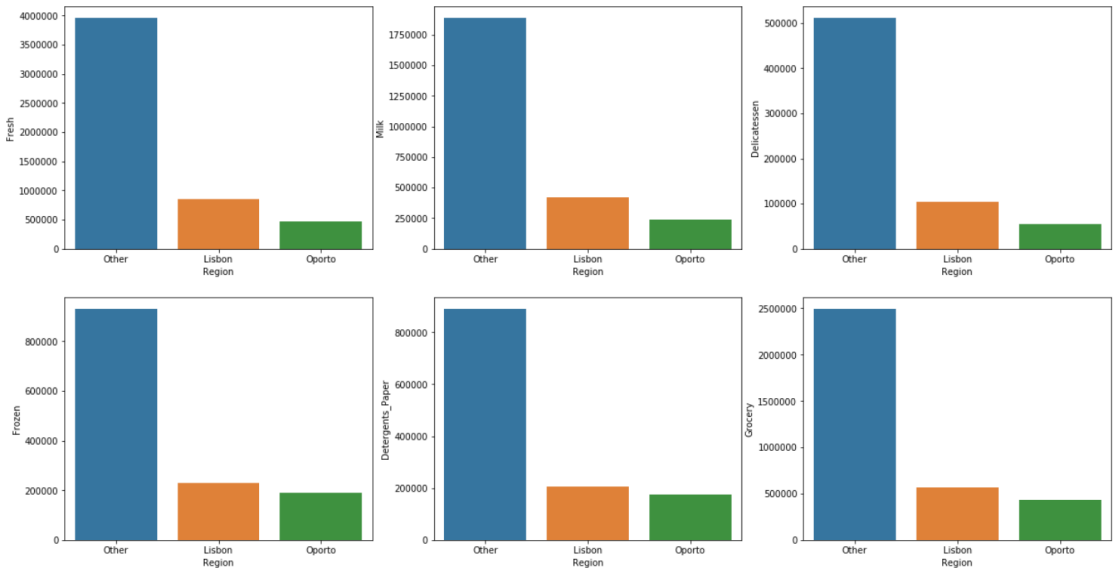
We can observe from the chart that -

- Among regions Other has the highest annual spend, whereas Oporto has the lowest annual spend for all 6 different varieties of items
- Between channels, both Hotel and Region have high spend although Hotel has a higher spend than Region
- Overall, Other region with hotel has the highest spend, whereas, Lisbon-Retail has the lowest spend

1.2. There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

To check the behaviour of all 6 items, we again use the group by function. We further used the plt.bar functions to plot bar charts to view the behaviour of each item among regions and channel and also to compare the behaviour of all 6 items with each across regions and channel.

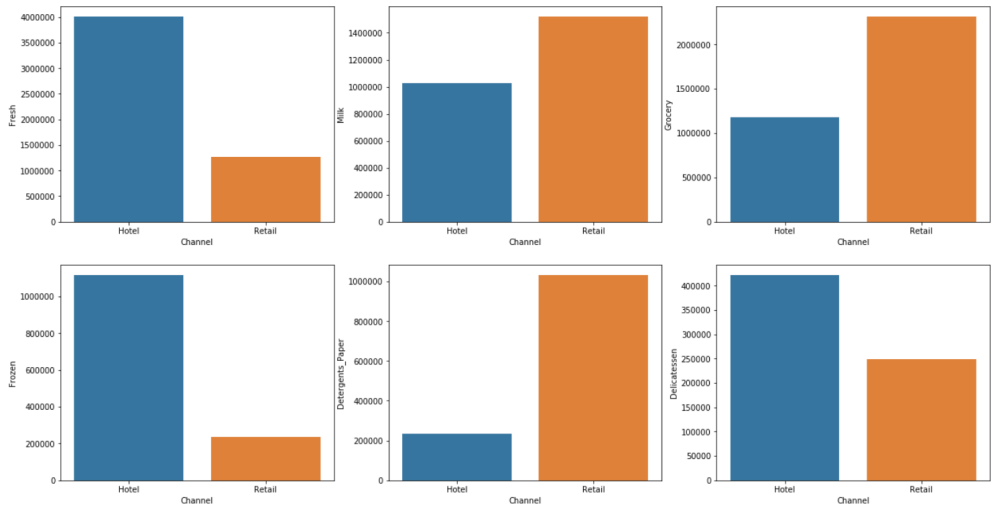
Region wise:



We can see from the bar plots above that all 6 varieties of items have the same behaviour across regions however each variable individually does not behave the same way over regions. The highest annual spend is in Other region for all variables. The spend in Lisbon and Oporto is much lesser as compared to Other region, however the spend in Lisbon is higher than Oporto.

*Please refer Appendix for source code

Channel wise:

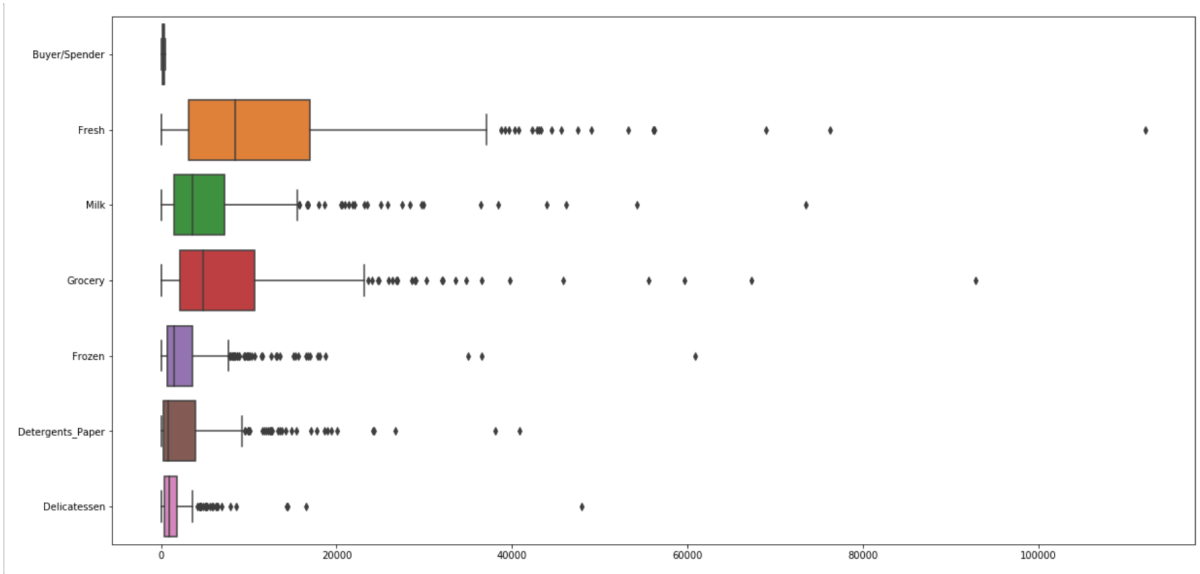


Visually we can observe that the items do not have the same behaviour across channels. If we compare the behaviour of the variables with each other we can see that spending on items such as Milk, Grocery and Detergents_paper, follow the same behaviour across Channels where Retail has a much higher spend than Hotel. For the other 3 categories - Delicatessen, Fresh and Frozen, the opposite behaviour is exhibited wherein the spend in higher for the Hotel channel and lower for Retail

1.3. On the basis of the descriptive measure of variability, which item shows the most inconsistent behaviour? Which items shows the least inconsistent behaviour?
Measures of variability include:

1. Range;
2. Interquartile range;
3. Variance; and
4. Standard deviation.

We can check for consistent and inconsistent behaviour with the help of a boxplot as it depicts the IQR of each variable graphically. We use the ‘sns.boxplot’ function to plot the same.

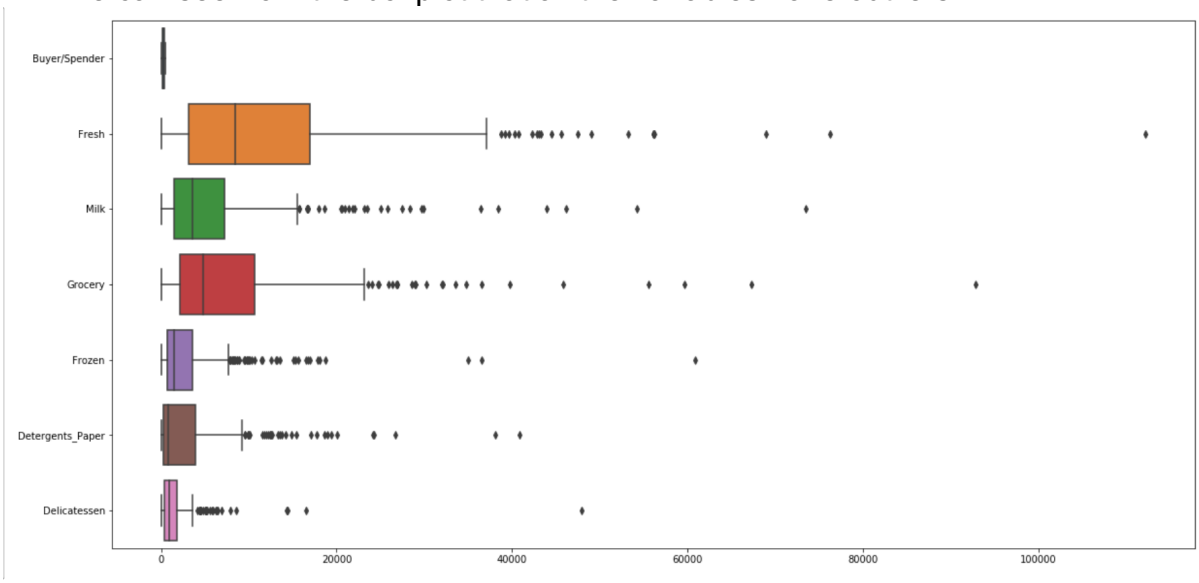


*Please refer Appendix for source code

We can see that Fresh category has the largest IQR which leads us to the conclusion that the variable has the most inconsistent behaviour.
The smallest IQR is for Delicatessen variable which means it has the most consistent behaviour.

(e) Outlier Identification

1.4. Are there any outliers in the data?
We can see from the boxplot that all the variables have outliers



Conclusion

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

1. Among the 3 regions, only Oporto has more spends in retail as compared to hotel. Diving deeper into this, we see that out of the 6 categories, Detergents_Paper, Grocery and milk have lower spends in Hotel. We can try to identify the reasons for this and try and improve sales for these categories in Oporto region as we know the demand exists basis behaviour for other two regions
2. Delicatessen has the most consistent behaviour across regions and categories. This can be exploited to increase sales in this category by trying to understand the trends and putting in the necessary resources
3. Looking at the inconsistent behaviour of Fresh items, we can maybe be sub categorize the items in order to make the behaviour more consistent and then drive sales basis the trends noted so far in order to increase revenue
4. Overall the spend in hotel channel is much higher than the retail channel. Perhaps, more opportunities and partnerships can be explored in the retail chain across all regions to increase sales

*Please refer Appendix for source code

5. ANALYSIS OF 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.csv file).

(a) Import the Data Set

We use the command 'read.csv' to import the file Survey.csv into python

(b) Variable Identification

(i) Data view

After importing the data, we use the df.head() function to view the data to see if the data has been imported properly.

	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

(ii) Checking the summary of the data

We use the df.describe() and df.info function to view the summary of the data. This includes the 5-point summary and other specific details regarding the data. With the info function we can also view the data type and if there are any null values.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
ID	62	NaN	NaN	NaN	31.5	18.0416	1	16.25	31.5	46.75	62
Gender	62	2	Female	33	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	62	NaN	NaN	NaN	21.129	1.43131	18	20	21	22	26
Class	62	3	Senior	31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Major	62	8	Retailing/Marketing	14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad Intention	62	3	Yes	28	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GPA	62	NaN	NaN	NaN	3.12903	0.377388	2.3	2.9	3.15	3.4	3.9
Employment	62	3	Part-Time	43	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	62	NaN	NaN	NaN	48.5484	12.0809	25	40	50	55	80
Social Networking	62	NaN	NaN	NaN	1.51613	0.844305	0	1	1	2	4
Satisfaction	62	NaN	NaN	NaN	3.74194	1.21379	1	3	4	4	6
Spending	62	NaN	NaN	NaN	482.016	221.954	100	312.5	500	600	1400
Computer	62	3	Laptop	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Text Messages	62	NaN	NaN	NaN	246.21	214.466	0	100	200	300	900

(iii) Checking for null values

We use the df.isnull() function to check for any blank or NA cells in the data. By using the isnull() function by itself we only get a Boolean output. We use the df.isnull().sum() function to get an output of the count of null values in each column. This is helpful in cleaning the data. The rows with NA values can be removed and then one can move further with the analysis. However we need to be mindful of cases when there are a large number of NA values and the treatment of the same needs to be carried out.

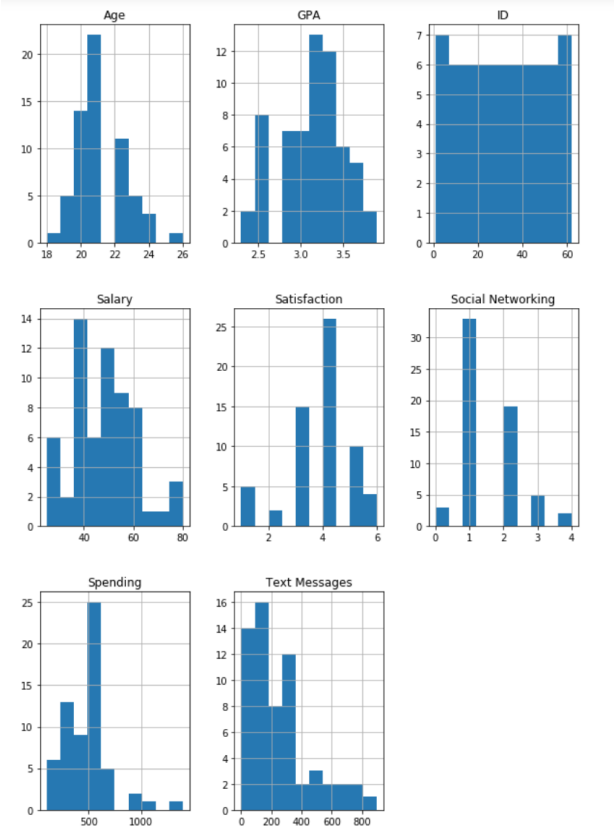
*Please refer Appendix for source code

```
ID                  0
Gender              0
Age                 0
Class               0
Major               0
Grad Intention      0
GPA                 0
Employment          0
Salary              0
Social Networking   0
Satisfaction        0
Spending            0
Computer            0
Text Messages       0
dtype: int64
```

There are no null values in the data set.

(c) Univariate Analysis

We use the df.hist() function to plot a histogram for all the variables. This helps us to understand the distribution of each variable, whether it is normally distributed, skewed, or uniformly distributed.



Interpretation of histogram

1. The distribution of Text messages is highly right skewed
2. All variables are continuous variables apart from Satisfaction and Social Networking
3. Satisfaction and Social Networking are not continuous variables hence there are breaks in the histogram

(d) Bivariate Analysis

Part I

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

*Please refer Appendix for source code

2.1.1. Gender and Major
2.1.2. Gender and Grad Intention
2.1.3. Gender and Employment
2.1.4. Gender and Computer
For this we use the crosstab function. The crosstab function helps us form the contingency table by building a cross tabulation of two variables in the data set and computes the their frequency

Following are the outputs of the contingency tables for 2.1.1 through 2.1.4

2.1.1.

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
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

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

2.1.3

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

2.1.4

Computer	Desktop	Laptop	Tablet	All
Gender				
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 questions:
2.2.1. What is the probability that a randomly selected CMSU student will be male? What is the probability that a randomly selected CMSU student will be female?

Based on the values calculated in the contingency tables we assign variables to each required probability. Further using the formula of conditional probability we calculate the same

P(Male student)=No. of male/Total students = 29/62
P(Female student) = No. of female/Total students = 33/62

Probability of randomly selecting a male CMSU student is 46.8%
Probability of randomly selecting a female CMSU student is 53.2%

*Please refer Appendix for source code

2.2.2. Find the conditional probability of different majors among the male students in CMSU.
Find the conditional probability of different majors among the female students of CMSU.

P(Major | Male) = P (Major ∩ Male)/P(Male)

P(Major | Female) = P (Major ∩ Female)/P(Female)

Probability of Major given Gender (in%)		
Gender	Male	Female
Major		
Accounting	13.8	9.2
CIS	3.5	9.1
Economics and Finance	13.8	21.2
International Business	6.9	12.1
Management	20.7	12.1
Other	13.8	9.1
Retail	17.2	27.3
Undecided	10.3	0.0

2.2.3. Find the conditional probability of intent to graduate, given that the student is a male.
Find the conditional probability of intent to graduate, given that the student is a female.

P(Grad Intention | Male) = P(Grad Intention ∩ Male)/P(Male)

P(Grad Intention | Female) = P (Grad Intention ∩ Female)/P(Female)

Probability of intent to graduate given student is male is 17.0%
Probability of intent to graduate given student is female is 11.0%

2.2.4. Find the conditional probability of employment status for the male students as well as for the female students.

P(Employment Status | Male) = P (Employment Status ∩ Male)/P(Male)

P(Employment Status | Female) = P (Employment Status ∩ Female)/P(Female)

Probability of employment status given Gender (in%)		
Gender	Male	Female
Employment Status		
Full time employment	24.1	9.1
Part time employment	65.5	72.7
Unemployed	10.3	18.2

2.2.5. Find the conditional probability of laptop preference among the male students as well as among the female students.

P(Laptop | Male) = P (Laptop ∩ Male)/P(Male)

P(Laptop | Female) = P (Laptop ∩ Female)/P(Female)

*Please refer Appendix for source code

Probability of laptop preference given student is male is 89.7%
Probability of laptop preference given student is female is 87.9%

2.3. Based on the above probabilities, do you think that the column variable in each case is independent of Gender?
Justify your comment in each case.

We know that the probability of independent is calculated by P(A∩B)=P(A).P(B). We will calculate both P(A∩B) and P(A).P(B) and conclude that variables are independent if both are equal to each other. For the purpose of calculation we will consider the male gender P(A)=P(Particulars) , P(B)= probability of Male

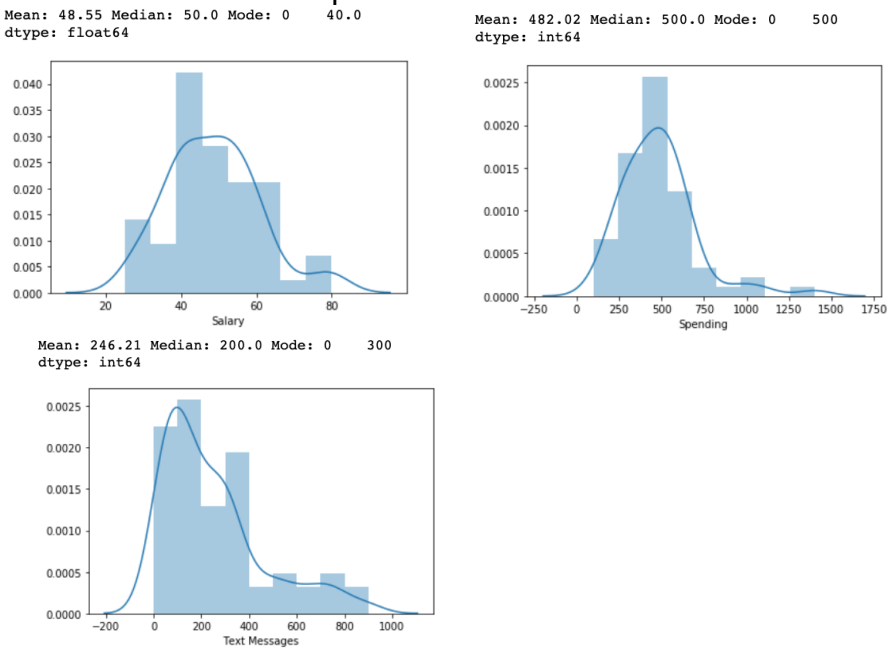
Particulars	P(A).P(B)	P(A∩B)	Independent?
Major (Accounting)	0.053	0.064	No
Grad Intention (Yes)	0.21	0.27	No
Employment Status (Unemployed)	0.068	0.048	No
Computer (Laptop)	0.41	0.41	Yes

Part II

2.4. Note that there are three numerical (continuous) variables in the data set, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions. [Recall that symmetric histogram does not necessarily mean that the underlying distribution is symmetric]

To check whether a variable follows a normal distribution, we can use various methods:

1. Comparing the value of mean median and mode along with looking at the spread of the distribution. For normal distribution, we would have a symmetric bell shaped curve, and the value of mean=median=mode
   We can use the df['column name'].mean(), df['column name'].median(), df['column name'].mode() functions to find the mean, median and mode of the column respectively.
   We can use the sns.distplot function to view the distribution curve of the data



*Please refer Appendix for source code

We can see from the distribution plot that distribution for salary is closest to the normal distribution, however it is not normally distributed.
Further we can also observe from the mean median and mode values, that the difference between these is the least for Salary. For the other cases the difference is more and hence the variables do not follow the normal distribution.

2. Checking the skewness of the distribution of the variable. For a normal distribution the skewness would be 0.
We use the stats.skew function to check the skewness of the variable.

	Skewness
Text Messages	1.264245
Spending	1.547285
Salary	0.521677

Basis the values of Skewness we can infer the following:
(i) Text Messages - The value of skewness is 1.2 which means the distributions is right skewed and hence does not follow normal distribution
(ii) Spending - The value of skewness is 1.5 which means the distribution is right skewed and hence does not follow normal distribution
(iii) Salary - The value of skewness is 0.5 which means the distribution is slightly skewed to the right. This variable is the closest to the normal distribution. However, as there is skewness it also does not follow normal distribution

3. Using the Shapiro test to test for normality

First we formulate the null and alternative hypothesis -

Ho: Variable follows normal distribution
Ha: Variable does not follow normal distribution

For each variable we perform the shapiro test using the shapiro function and interpret basis the p-value whether to reject or fail to reject the null hypothesis for alpha 0.05

Variable	Statistic Value	p value	Null hypothesis
Spending	0.878	0.000017	Reject
Salary	0.957	0.028001	Fail to Reject
Text Message	0.859	0.000004	Reject

Spending:
With a confidence interval of 0.05, we can infer from the p-value that Spending variable does not follow normal distribution
Salary:
With a confidence interval of 0.05, we can infer from the p-value that Salary variable follows normal distribution
Text Messages:
With a confidence interval of 0.05, we can infer from the p-value that Text Messages variable does not follow normal distribution

*Please refer Appendix for source code

6. ANALYSIS OF 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 colouring 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 is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

(a) Import the Data Set

We use the command 'read.csv' to import the file– A&B shingles.csv into python.

(b) Variable Identification

(i) Data view

After importing the data, we use the df.head() function to view the data to see if the data has been imported properly. We can view the number or rows we desire by entering the number in the function for eg. df.head(5) which will give us the following output.

	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

(ii) Checking the summary of the data

We use the df.describe() and df.info function to view the summary of the data. This includes the 5 point summary and other specific details regarding the data. With the info function we can also view the data type of all the variables and if there are any null values.

*Please refer Appendix for source code

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 2 columns):
A    36 non-null float64
B    31 non-null float64
dtypes: float64(2)
memory usage: 704.0 bytes
```

(iii) Checking for null values

We use the df.isnull() function to check for any blank or NA cells in the data. By using the isnull() function by itself we only get a Boolean output. We use the df.isnull().sum() function to get an output of the count of null values in each column. This is helpful in cleaning the data. The rows with NA values can be removed and then one can move further with the analysis. However, we need to be mindful of cases when there are a large number of NA values and the treatment of the same needs to be carried out.

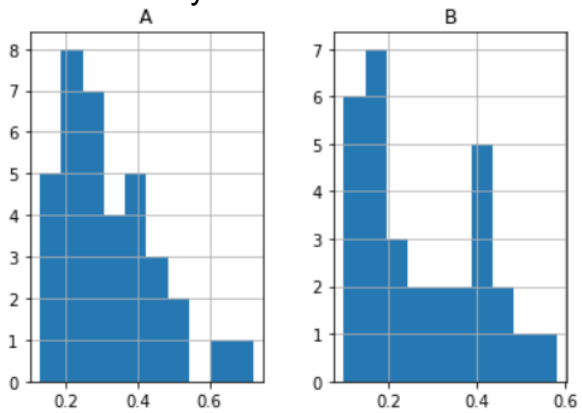
```
A    0
B    5
dtype: int64
```

As we can see column B has 5 null values.

We use the drop.na function to remove these null values and then proceed with our analysis. We also give the command inplace=True in order to permanently reflect this change in the dataset

(c) Univariate Analysis

We use the df.hist() function to plot a histogram for all the variables. This helps us to understand the distribution of each variable, whether it is normally distributed, skewed, or uniformly distributed.



Both variables are right skewed.

(d) Bivariate Analysis

3.1 For the A shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet. While forming the null and alternative hypothesis,

*Please refer Appendix for source code

Ho : μA ≥ 0.35 - Population mean moisture content of A shingles is greater than or equal to 0.35 pound per 100 square feet

Ha : μB <0.35 - Population mean moisture content of A shingles is less than 0.35 pound per 100 square feet

3.2 For the A shingles, conduct the test of hypothesis and find the p-value. Interpret the p-value. Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

To conduct the hypothesis we use the t test as population variance is unknown. We use the code for a 1 sample t test to find the t statistic and p- value.
The result of the test is –
T-Statistic: -1.6005252585398313 p-value: 0.11996170801033942

We are asked to draw an inference based on the p-value for a confidence level of 0.05. We reject the null hypothesis if the p-value < 0.05 and fail to reject the null hypothesis if p-value is > 0.05

In this case, the p value calculated is 0.12 which is greater than 0.05, hence we fail to reject the null hypothesis that the population mean moisture content is greater than equal to 0.35 pound per 100 square feet

3.3. For B shingles, form the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet.

Ho : μB ≥ 0.35 - Population mean moisture content of B shingles is greater than equal to 0.35 pound per 100 square feet

Ha : μB <0.35 - Population mean moisture content of B shingles is less than 0.35 pound per 100 square feet

3.4. For the B shingles, conduct the test of the hypothesis and find the p-value. Interpret the p-value. Is there evidence at the 0.05 level of significance that the population mean moisture content is less than 0.35 pound per 100 square feet?

T-Statistic: -3.1003313069986995 p-value: 0.004180954800638363

In this case, the p value is 0.0041 which is less than 0.05, hence we reject the null hypothesis that the population mean moisture content is equal to 0.35 pound per 100 square feet

3.5. Do you think that the population means for shingles A and B are equal? Form the hypothesis and conduct a test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Ho : μA = μB - Population mean moisture content of A shingles is equal to Population mean moisture content of B shingles

Ha : μA ≠ μB - Population mean moisture content of A shingles is not equal to Population mean moisture content of B shingles

*Please refer Appendix for source code

We need to check the assumption that both population variances are equal before performing the test for equality of means. We use the levene test to check the same

Ho: Var(A)=Var(B)

Ha: Var(A)≠Var(B)

LeveneResult(statistic=0.08443854431268433, pvalue=0.77237221788485)

In this scenario, our p-value is greater than 0.05, hence we fail to reject our null hypothesis that our variances are equal. Basis this we assume that the population variances are equal and carry out the test for equality of means

Further, we calculate the t statistic and p-value to test our hypothesis
T-Statistic: 0.985249977839441 p-value: 0.3284577916404776

Our p value is 0.33, which is greater than 0.05. Basis this we can conclude that we fail to reject the null hypothesis which is - Population means of Shingles A and B are equal to each other

3.6. What assumption about the population distribution is needed in order to conduct the hypothesis tests above?

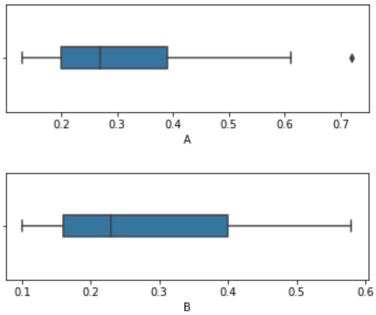
We need to assume that the population follows the normal distribution in order to conduct the hypothesis tests above

3.7. Check the assumptions made with histograms, boxplots, normal probability plots or empirical rule.

To check the assumption whether the distribution follows normal distribution we can check the same using various methods both statistically and graphically.

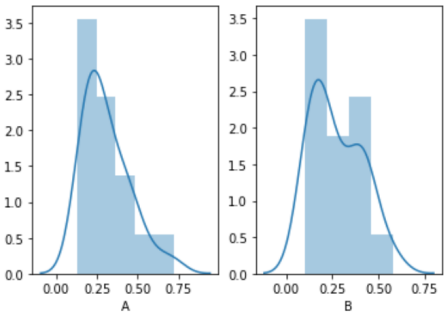
To check statistically we can check using the empirical rule. If a variable follows normal distribution then as per empirical rule its 68% of the value will lie within 1 standard deviation of the mean. Source code is in appendix

Further to check this graphically we can use histogram or boxplots. For a normal distribution, when represented graphically, the plots are symmetric.



We can view from the distribution of the box plot both variables seem to be skewed to the right as the right is much longer. we can also observe that the distance of Q3 is also further away from Q2 as compared to the distance between Q1 and Q2

*Please refer Appendix for source code



Basis the distribution plot, we can see that the distribution is right skewed.

3.8. Do you think that the assumption needed in order to conduct the hypothesis tests above is valid? Explain.

Yes, the assumption that the population follows normal distribution is valid as we know that based on the central limit theorem - ***Given random and independent sample of N observations each, the distribution of sample mean approaches normality as the size of N increases***. This means that when N increases the sample mean will follow normal distribution. In this case we are trying to hypothesize about the population mean. Here the population size N is unknown, however it would ideally be a large number due to which its distribution will be the normal distribution

7. Appendix

Attached separately

*Please refer Appendix for source code