STATISTICS PROJECT

1. WHOLESALE CUSTOMER ANALYSIS:

We have imported the “Wholesale customer data” in python using pandas read_csv function to analyze the spend across various regions and channels.

1.1.1 Use methods of descriptive statistics to summarize data.

- data.describe()
- It gives the basic descriptive statistics of the wholesale customer data.

1.1.2 Which Region and which Channel spent the most? 1.1.3 Which Region and which Channel spent the least?

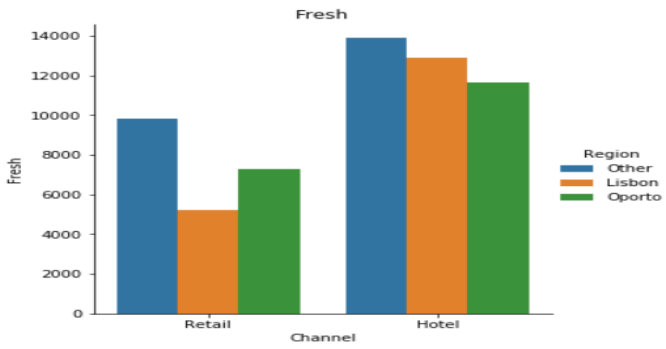
region_channel_totalspend = data.groupby(['Region','Channel'])['Total_Spend'].sum()

print(region_channel_totalspend)

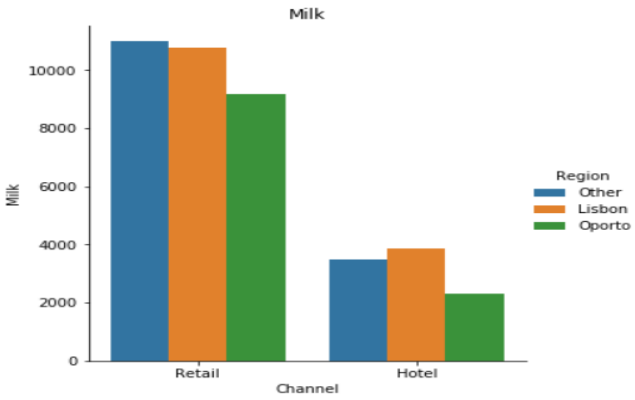
```
Region  Channel
Lisbon  Hotel      1538342
        Retail      848471
Oporto  Hotel       719150
        Retail      835938
Other   Hotel      5742077
        Retail     4935522
Name: Total_Spend, dtype: int64
```

- Using the groupby and sum function with Region and Channel, it is found that Highest Spend in the Region and channel is Others in Hotel
- Comparing to retail, Hotel channel spends the highest amount.
- In regions Others spend the highest amount.
- Lowest Spend in the Region is Oporto and Lowest spend in the channel is Retail.

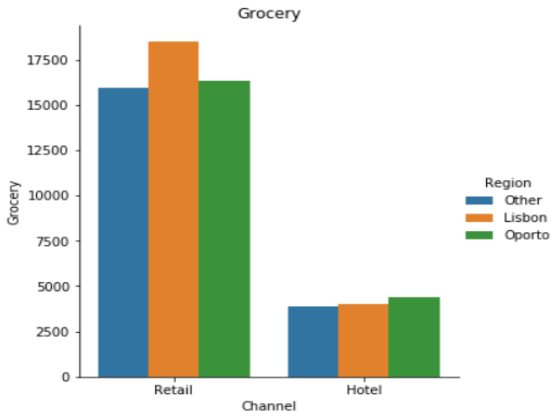
1.2. There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer



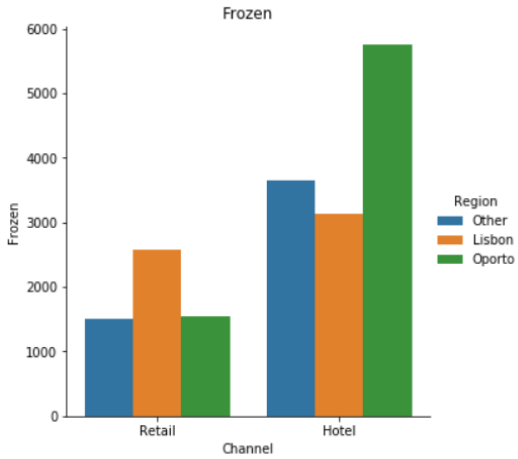
In the Fresh Item, Others is spent more on the Hotel channel



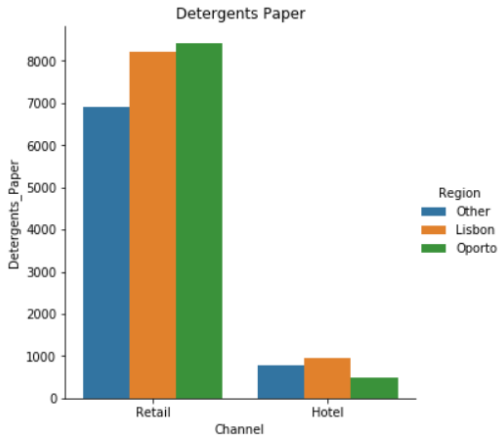
In the Milk Item, Others is spent more on Retail Channel



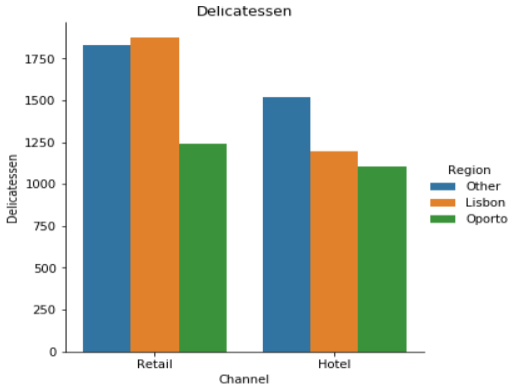
In the Grocery Item, Lisbon is spent more on Retail channel



In the Frozen Item, Oporto is spent more on Hotel channel



In the Detergents Paper, Oporto is spent more on Retail channel



In the Delicatessen Item, Lisbon is spent more on Retail channel

Fresh and Frozen items have more spend on Hotel channel than Retail across all Regions. Other Items have more spend on Retail.

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

cv_fresh = np.std(data['Fresh']) / np.mean(data['Fresh'])

cv_fresh - 1.0527196084948245

cv_milk = np.std(data['Milk']) / np.mean(data['Milk'])

cv_milk – 1.2718508307424503

cv_grocery = np.std(data['Grocery']) / np.mean(data['Grocery'])

cv_grocery – 1.193815447749267

cv_frozen = np.std(data['Frozen']) / np.mean(data['Frozen'])

cv_frozen - 1.5785355298607762

cv_detergents = np.std(data['Detergents_Paper']) / np.mean(data['Detergents_Paper'])

cv_detergents - 1.6527657881041729

cv_delicatessen = np.std(data['Delicatessen']) / np.mean(data['Delicatessen'])

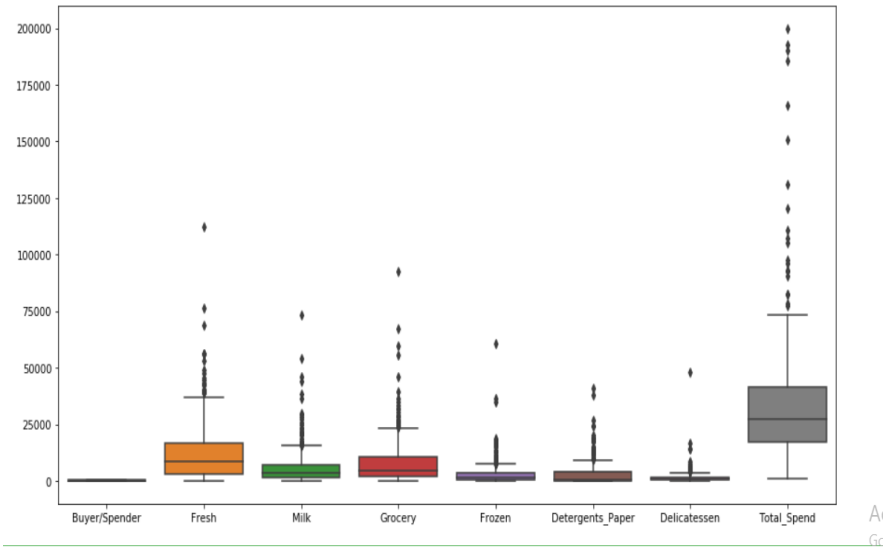
cv_delicatessen – 1.8473041039189306

We compute consistency based on "**Coefficient of variation**". From the above measures, it is shown that the most inconsistent behavior is "Delicatessen" and least value is of "Fresh" category.

1.4 Are there any outliers in the data?

plt.figure(figsize=(15,8))

sns.boxplot(data=data)



Using Boxplot it is identified that all the data has outliers. There are outliers in the data in the following items (Fresh, Milk, Grocery, Frozen, Detergents_Paper & Delicatessen)

1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

From the analysis, we find out there are inconsistencies in the different items which should be minimized and make the data consistent. The spending in the different channels and Regions are seen different which should be similar. The spend for all the items should be almost equal.

2. A & B Shingles:

We have imported “A+&+B+shingles.csv” using pandas read_csv function.

3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Ho = the mean moisture content is no less than 0.35 pound per 100 square feet.

H1 = the mean moisture content is less than 0.35 pound per 100 square feet.

```
from scipy import stats

t_stat,p_value = stats.ttest_1samp(data.A,0.35)

print("The value for t_statistics is:{0} and p_value is:{1}".format(t_statB,p_valueB/2))
```

```
The value for t_statistics is:-1.4735046253382782 and p_value is:0.0747763
3144907513
```

Since alpha = 0.05 and p value is greater than 0.05, we accept Ho.

```
print("The value for t_statistics is:{0} and p_value is:{1}".format(t_stat,p_value/2))

t_statB,p_valueB = stats.ttest_1samp(data.B,0.35,nan_policy='omit')
```

```
The value for t_statistics is:-3.1003313069986995 and p_value is:0.0020904
774003191826
```

Since alpha = 0.05 and p value is lesser than 0.05, we reject Ho.

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis an d conduct the test of the hypothesis. What assumption do you need to check before the test for eq uality of means is performed?

Ho = Population mean of A and B are equal, H1 = Population mean of A and B are not equal

```
t_stat2,p_value2 = stats.ttest_ind(data['A'],data['B'],equal_var = True,nan_policy='omit')

print("The value for p_value is:{}".format(round(p_value2,3)))
```

```
The value for p_value is:0.202
```

Since alpha = 0.05 and p value is greater than 0.05, we accept Ho. It means both A and B shing les are equal.

The assumptions like whether it has normal distributions for both A and B and the variances are same for both are to be checked.

3. SURVEY ANALYSIS FOR CMSU:

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

2.1.1. Gender and Major

```
data_crosstab1 = pd.crosstab(data['Gender'], data['Major'], margins = False)

print(data_crosstab1)
```

```
Major   Accounting  CIS  Economics/Finance  International Business  \
Gender
Female           3    3                  7                       4
Male             4    1                  4                       2

Major   Management  Other  Retailing/Marketing  Undecided
Gender
Female           4      3                    9          0
Male             6      4                    5          3
```

2.1.2. Gender and Grad Intention

```
data_crosstab2 = pd.crosstab(data['Gender'], data['Grad Intention'], margins = False)

print(data_crosstab2)
```

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

2.1.3. Gender and Employment

```
data_crosstab3 = pd.crosstab(data['Gender'], data['Employment'], margins = False)

print(data_crosstab3)
```

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

2.1.4. Gender and Computer

```
data_crosstab4 = pd.crosstab(data['Gender'], data['Computer'], margins = False)

print(data_crosstab4)
```

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

2.2. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.2.1 What is the probability that a randomly selected CMSU student will be male?

2.2.2 What is the probability that a randomly selected CMSU student will be female?

```
count = 0

count1 = 0

for i in range(0,len(data['Gender'])):

   if data.iloc[i,1]=='Male':

      count = count+1

   else:

      count1=count1+1

prob_male = (count/len(data['Gender']))*100

prob_female = (count1/len(data['Gender']))*100

print("The probability that a randomly selected student will be male is",prob_male)

print("The probability that a randomly selected student will be female is",prob_female)
```

```
The probability that a randomly selected student will be male is 46.774193
548387096
The probability that a randomly selected student will be female is 53.2258
064516129
```

2.3. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

```
pd.crosstab(data['Major'],(data['Gender']),margins=True,margins_name="Total")
```

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

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

```
Male_Accounting = ((4/62)/(29/62))*100
print("The Probaility of Accounting Major among the male students is",Male_Accounting)
```

- **The Probaility of Accounting Major among the male students is 13.793103448275861**

```
Female_Accounting = (3/62)/(33/62)*100

print("The Probaility of Accounting Major among the Female students is",Female_Accounting)
```

- **The Probaility of Accounting Major among the Female students is 9.090909090909092**

```
Male_CIS = (1/62)/(29/62)*100
print("The Probaility of CIS Major among the male students is",Male_CIS)
```

- **The Probaility of CIS Major among the male students is 3.4482758620689653**

```
Female_CIS = (3/62)/(33/62)*100
print("The Probaility of CIS Major among the Female students is",Female_CIS)
```

- **The Probaility of CIS Major among the Female students is 9.090909090909092**

```
Male_EconomicsFinance = (4/62)/(29/62)*100
print("The Probaility of EconomicsFinance Major among the male students is",Male_Economics
Finance)
```

- **The Probaility of EconomicsFinance Major among the male students is 13.793103448275861**

```
Female_EconomicsFinance = (7/62)/(33/62)*100
print("The Probaility of EconomicsFinance Major among the Female students is",Female_Econo
micsFinance)
```

- **The Probaility of EconomicsFinance Major among the Female students is 21.21212121212121**

```
Male_InternationalBusiness = (2/62)/(29/62)*100
print("The Probaility of InternationalBusiness Major among the male students is",Male_Internat
ionalBusiness)
```

- **The Probaility of InternationalBusiness Major among the male students is 6.896551724137931**

```
Female_InternationalBusiness = (4/62)/(33/62)*100
print("The Probaility of InternationalBusiness Major among the Female students is",Female_Int
ernationalBusiness)
```

- **The Probaility of InternationalBusiness Major among the Female students is 12.121212121212121**

```
Male_Management = (6/62)/(29/62)*100
print("The Probaility of Management Major among the male students is",Male_Management)
```

- **The Probaility of Management Major among the male students is 20.689655172413794**

```
Female_Management = (4/62)/(33/62)*100
print("The Probaility of Management Major among the female students is",Female_Manageme
nt)
```

- **The Probaility of Management Major among the female students is 12.121212121212121**

Male_Other = (4/62)/(29/62)*100

print("The Probaility of Other Major among the male students is",Male_Other)

- **The Probaility of Other Major among the male students is 13.793103448275861**

Female_Other = (3/62)/(33/62)*100

print("The Probaility of Other Major among the Female students is",Female_Other)

- **The Probaility of Other Major among the Female students is 9.090909090909092**

Male_RetailingMarketing = (5/62)/(29/62)*100

print("The Probaility of Retail/Marketing Major among the male students is",Male_RetailingMarketing)

- **The Probaility of Retail/Marketing Major among the male students is 17.24137931034483**

Female_RetailingMarketing = (9/62)/(33/62)*100

print("The Probaility of Retail/Marketing Major among the Female students is",Female_RetailingMarketing)

- **The Probaility of Retail/Marketing Major among the Female students is 27.272727272727**

Male_Undecided = (3/62)/(29/62)*100

print("The Probaility of Undecided Major among the male students is",Male_Undecided)

- **The Probaility of Undecided Major among the male students is 10.344827586206897**

Female_Undecided = (0/62)/(33/62)*100

print("The Probaility of Undecided Major among the Female students is",Female_Undecided)

- **The Probaility of Undecided Major among the Female students is 0.0**

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1 Find the probability That a randomly chosen student is a male and intends to graduate.

pd.crosstab(data['Gender'],(data['GradIntention']),margins=True,margins_name="Total")

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

Male_Graduate = (17/29)*100

print("The probability that a randomly choosen person is Male and intends to Graduate is", Male_Graduate)

- **The probability that a randomly choosen person is Male and intends to Graduate is 58.620689655172406**

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop

pd.crosstab(data['Gender'],(data['Computer']),margins=True,margins_name="Total")

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

Female_NoLaptop = (4/29)*100

print("The probability that a randomly choosen person is Female and does not have a laptop is" , Female_NoLaptop)

- **The probability that a randomly choosen person is Female and does not have a laptop is 13.793103448275861**

2.5. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.5.1 Find the probability that a randomly chosen student is either a male or has a full-time employment

pd.crosstab(data['Gender'],(data['Employment']),margins=True,margins_name="Total")

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

Male_Fulltime = ((29/62)+(10/62)-(7/62))*100

print("The probability that a randomly choosen person is either male or has full time", Male_Fulltime)

- **The probability that a randomly choosen person is either male or has full time 74.19354838709677**
- **The probability that a randomly choosen person is either male or has full time 51.61290322580645**

2.5.2 Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

pd.crosstab(data['Major'],(data['Gender']),margins=True,margins_name="Total").T

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

Female_Business_Management = Female_InternationalBusiness+Female_Management

print("The Conditional Probability that given a female student is randomly chosen, she is majoring in international business or management is", Female_Business_Management)

- **The Conditional Probability that given a female student is randomly chosen, she is majoring in international business or management is 24.242424242424242**

2.6 Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think graduate intention and being female are independent events?

Gender	No	Yes
Female	9	11
Male	3	17

Probability that student being Grad Intention is Yes : 0.7

Probability that student is female and grad intention is yes : 0.55

Since these probabilities are not equal, it is said that these two events are independent.

2.7 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. Answer the following questions based on the data

2.7.1 If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

pd.crosstab(data['Gender'],(data['GPA']),margins=True,margins_name="Total")

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	Total
Gender																	
Female	1	1	2	0	1	3	5	2	4	3	2	4	1	2	1	1	33
Male	0	0	4	2	2	1	2	5	2	2	5	2	2	0	0	0	29
Total	1	1	6	2	3	4	7	7	6	5	7	6	3	2	1	1	62

GPA_Lessthan3 = (14/62)*100

print("The probability that a randomly choosen student GPA is less than 3 is",GPA_Lessthan3)

- **The probability that a randomly choosen student GPA is less than 3 is 22.58064516129032**

2.7.2 Find conditional probability that a randomly selected male earns 50 or more. Find conditional probability that a randomly selected female earns 50 or more.

pd.crosstab(data['Gender'],(data['Salary']),margins=True,margins_name="Total")

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	Total
Gender																				
Female	0	5	1	0	1	5	1	1	0	1	5	0	0	5	5	0	1	1	1	33
Male	1	0	1	1	0	7	0	4	1	0	4	1	1	3	3	1	0	0	1	29
Total	1	5	2	1	1	12	1	5	1	1	9	1	1	8	8	1	1	1	2	62

Male_Earnsmorethan50 = (10/62)/(29/62)*100

print("The probability that a randomly selected male earns 50 or more is",Male_Earnsmorethan50)

Female_Earnsmorethan50 = (10/62)/(33/62)*100

print("The probability that a randomly selected Female earns 50 or more is",Female_Earnsmorethan50)

- **The probability that a randomly selected male earns 50 or more is 34.48275862068966**
- **The probability that a randomly selected Female earns 50 or more is 30.303030303030305**

2.8.1 Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending and Text Messages. For each of them comment whether they follow a normal distribution. 2.8.2 Write a note summarizing your conclusions.

import matplotlib.pyplot as plt

data['GPA'].hist()

plt.show()

data['Salary'].hist()

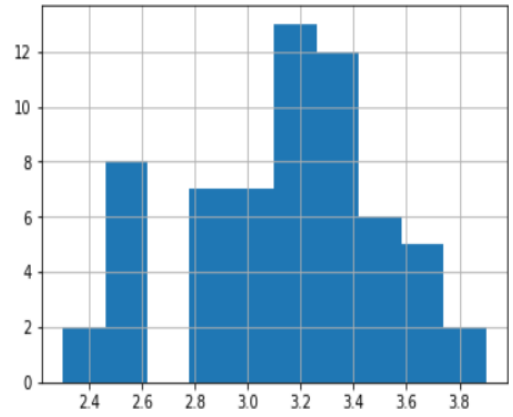
plt.show()

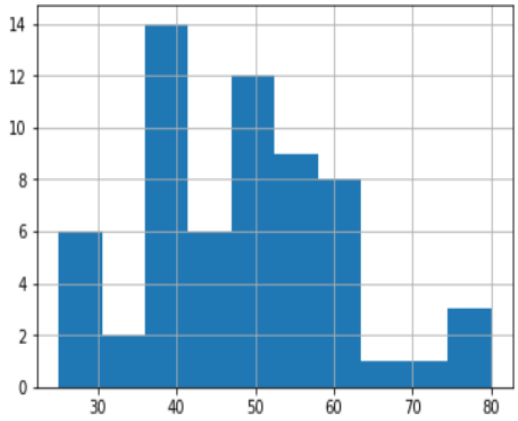
data['Spending'].hist()

plt.show()

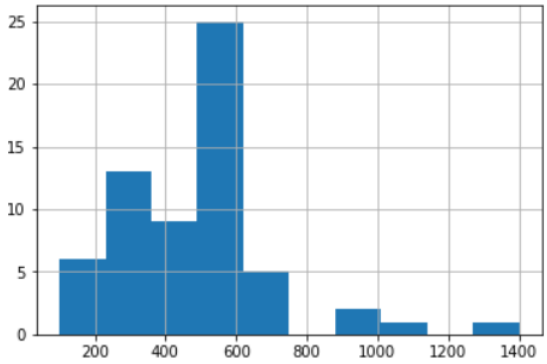
data['Text Messages'].hist()

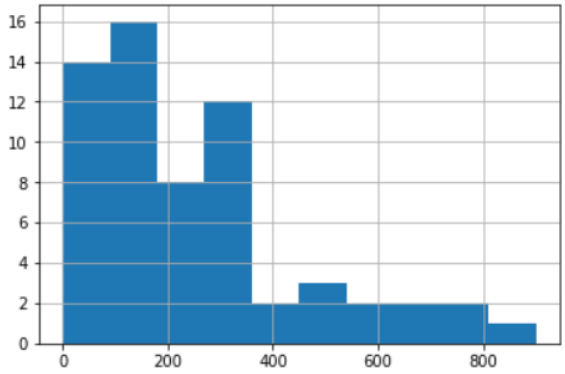
plt.show()





By using histogram or distplot, it is found that GPA and Salary follows normal distribution





By using histogram or distplot, it is found that Spending and Text messages does not follow normal distribution