Exploratory Data Analysis for Machine Learning - Coursera

EDA and Hypothesis Test

On Direct Marketing data

About the data

The data set includes data from a direct marketer who sells his products only via direct mail. He sends catalogs with product characteristics to customers who then order directly from the catalogs. The marketer has developed customer records to learn what makes some customers spend more than others. The data set includes n = 1000 customers and the following variables:

1. Age (of customer; old/middle/young)
2. Gender (male/female)
3. OwnHome (whether customer owns home; yes/no)
4. Married (single/married)
5. Location (far/close; in terms of distance to the nearest brick and mortar store that sells similar products)
6. Salary (yearly salary of customer; in dollars)
7. Children (number of children; 0–3)
8. History (of previous purchase volume; low/medium/high/NA; NA means that this customer has not yet purchased);
9. Catalogs (number of catalogs sent)
10. AmountSpent (in dollars).

Plan for exploration

The objective of the marketer is to explain the amount spent by customers using these Characteristics of the customers so that he can improve the targeting process. So we can analyse the influence of each characteristic on the target attribute.

Data Preprocessing and Feature Engineering

Missing Values :

The only column with NA values was History and it implies that the customer has no history of purchase so we cannot avoid that information. Therefore, it was added as a fourth category in the column.

Encoding Categorical Variables :

Label Encoding : on columns: Gender,Married,OwnHome
One-Hot Encoding : Age and History , since they are ordinal

Feature scaling : Standard scaler was used on the numerical columns to scale.

Data Analysis

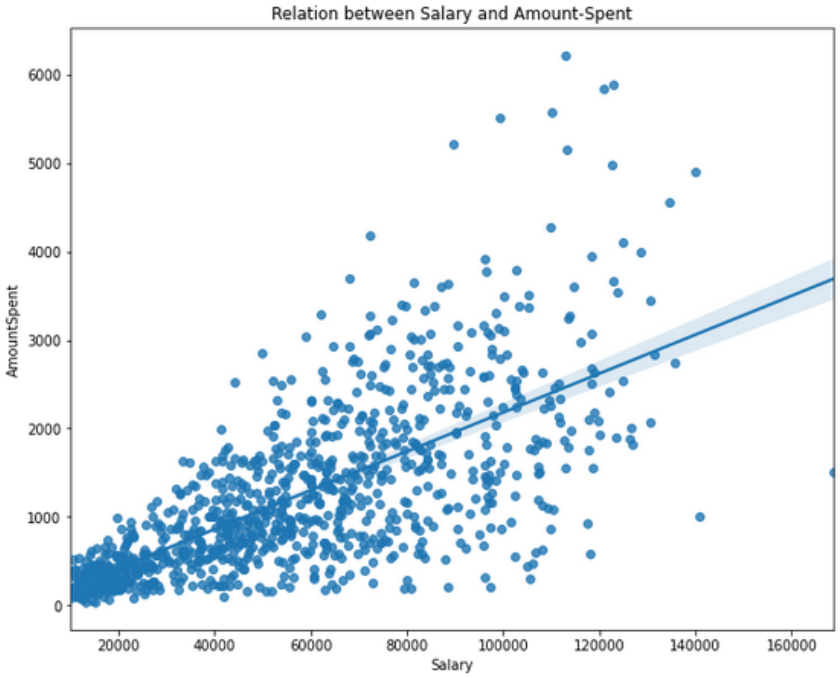
Data :

	Age	Gender	OwnHome	Married	Location	Salary	Children	History	Catalogs	AmountSpent
0	Old	Female	Own	Single	Far	47500	0	High	6	755
1	Middle	Male	Rent	Single	Close	63600	0	High	6	1318
2	Young	Female	Rent	Single	Close	13500	0	Low	18	296
3	Middle	Male	Own	Married	Close	85600	1	High	18	2436
4	Middle	Female	Own	Single	Close	68400	0	High	12	1304

Five - point summary of numerical variables :

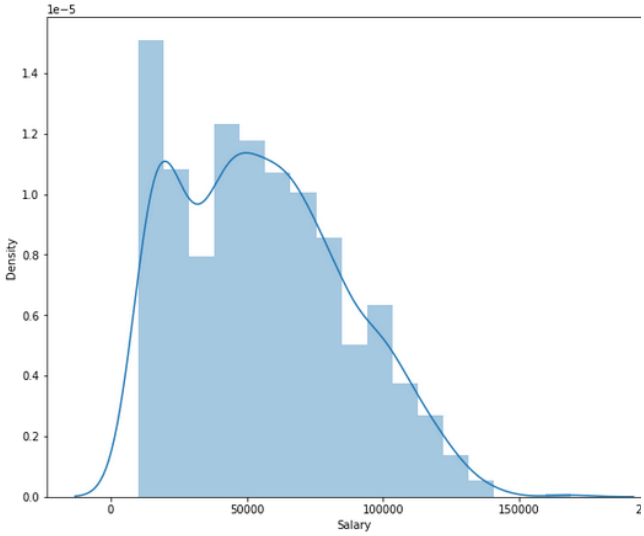
	Salary	Children	Catalogs	AmountSpent
count	1000.000000	1000.00000	1000.000000	1000.000000
mean	56103.900000	0.93400	14.682000	1216.770000
std	30616.314826	1.05107	6.622895	961.068613
min	10100.000000	0.00000	6.000000	38.000000
25%	29975.000000	0.00000	6.000000	488.250000
50%	53700.000000	1.00000	12.000000	962.000000
75%	77025.000000	2.00000	18.000000	1688.500000
max	168800.000000	3.00000	24.000000	6217.000000

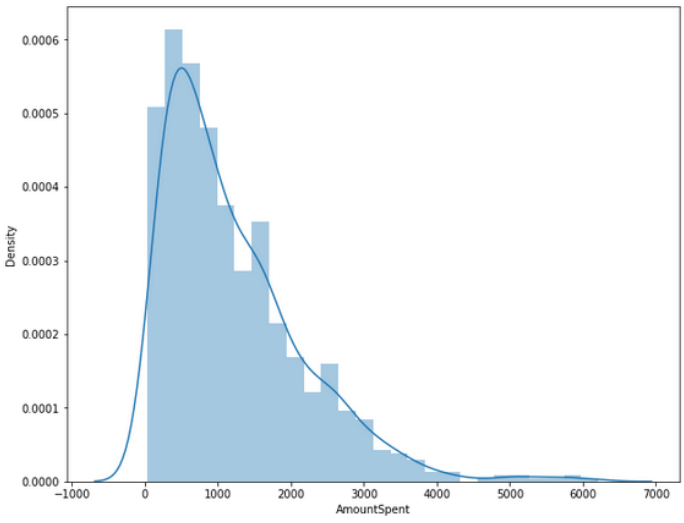
Relationship between Salary of customer and AmountSpent:



Looking at the scatterplot, There is clearly some positive linear relationship between the variables,
And we can see the regression line capturing most part of the data accurately and the number of outliers are clearly affecting the regression line.

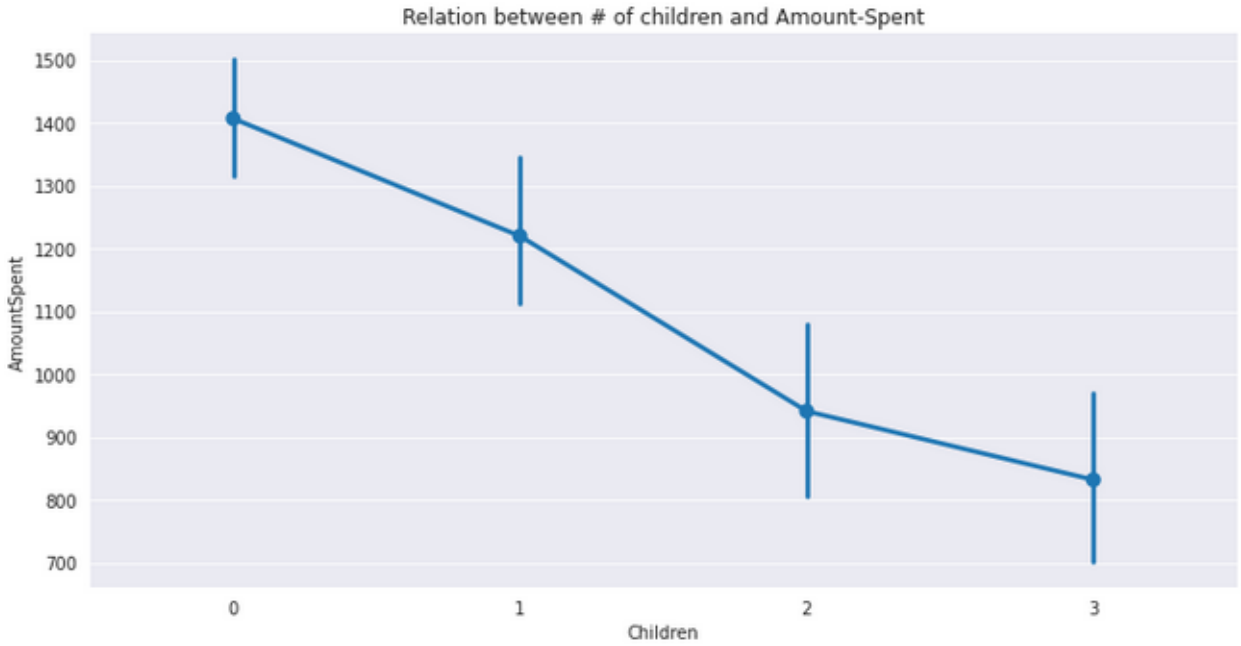
Since, there is an influence of outliers let us look into the distribution of these variables.





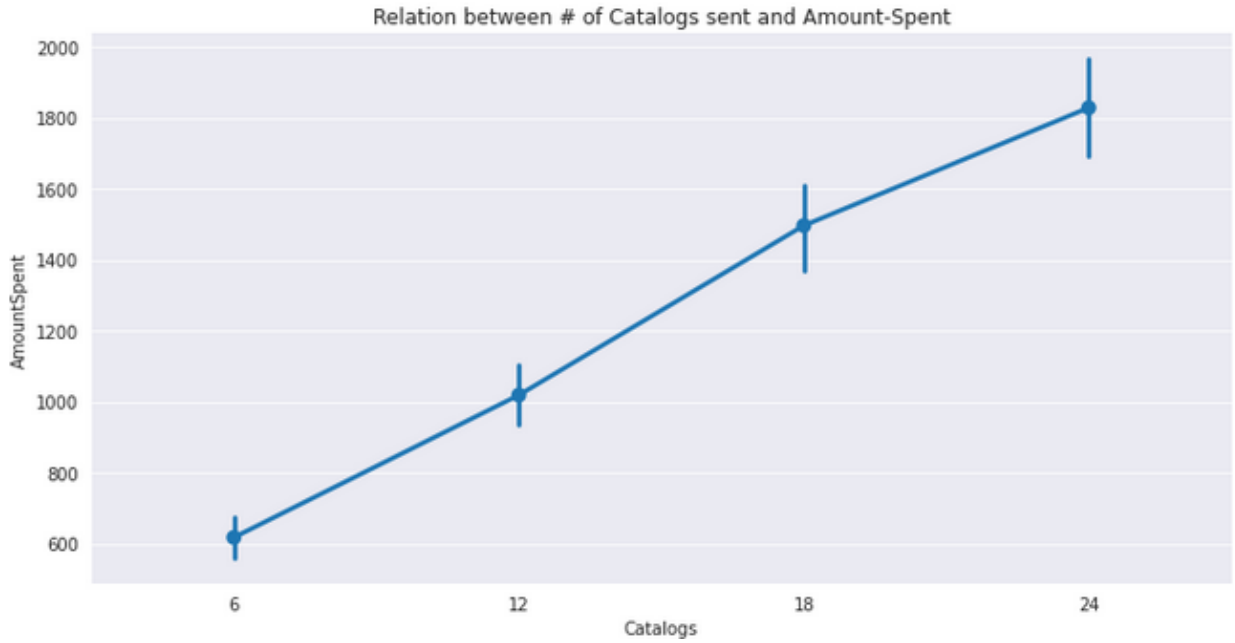
Both Salary and AmountSpent are positively-skewed variables, This can affect our prediction since there is a presence of extreme values.

Is there a relation with the number of children a customer has and the amount they spend ?



It does, we can see that with increase in number of children, the amount decreases and even the confidence interval has a significant difference with customers with no children or single

What about the number of catalogs sent by the marketer ?

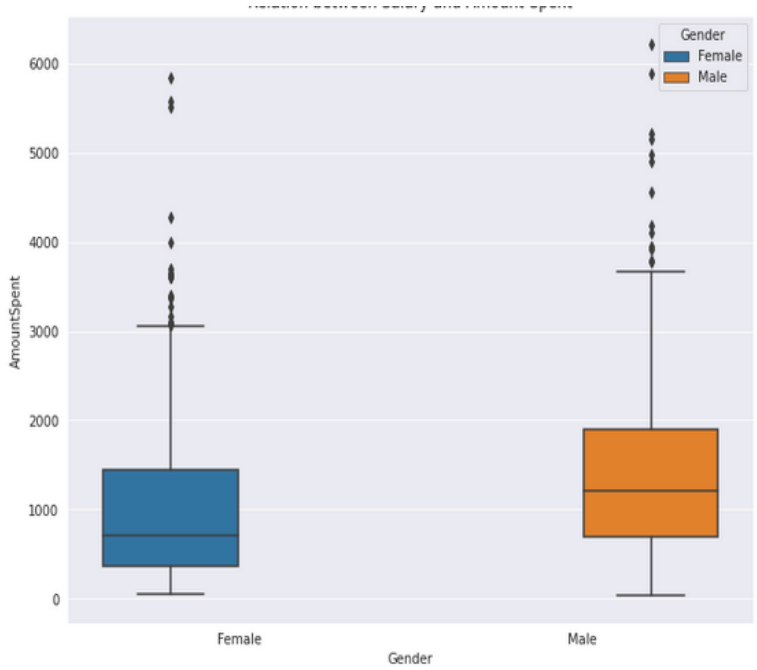


This can imply that his history of target is pretty much accurate, The customers he sent more catalogs were actually this highest spenders in his store.

Let us take a look at the categorical columns:

```
Age   : ['Old' 'Middle' 'Young']
Gender   : ['Female' 'Male']
OwnHome   : ['Own' 'Rent']
Married   : ['Single' 'Married']
Location   : ['Far' 'Close']
History   : ['High' 'Low' 'Medium' 'NH']
```

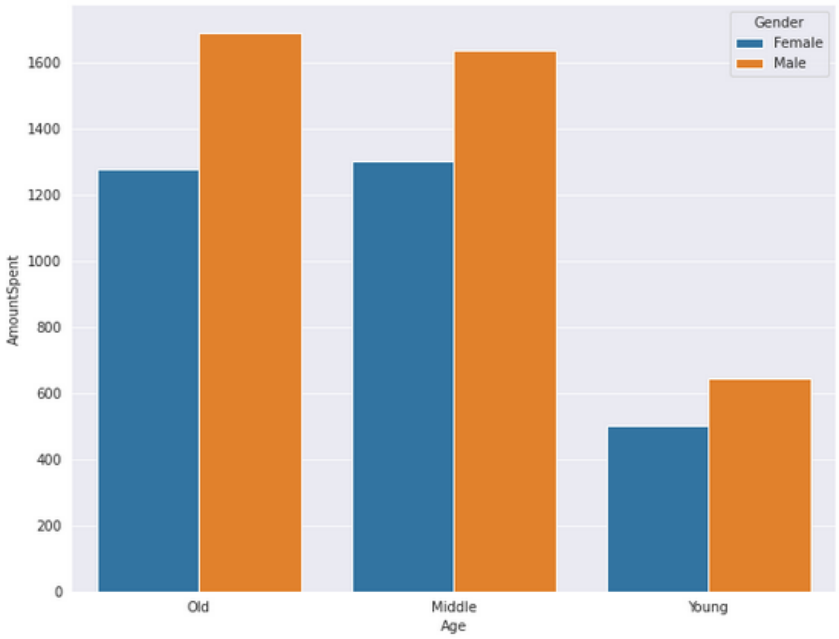
Unique values in each categorical column.



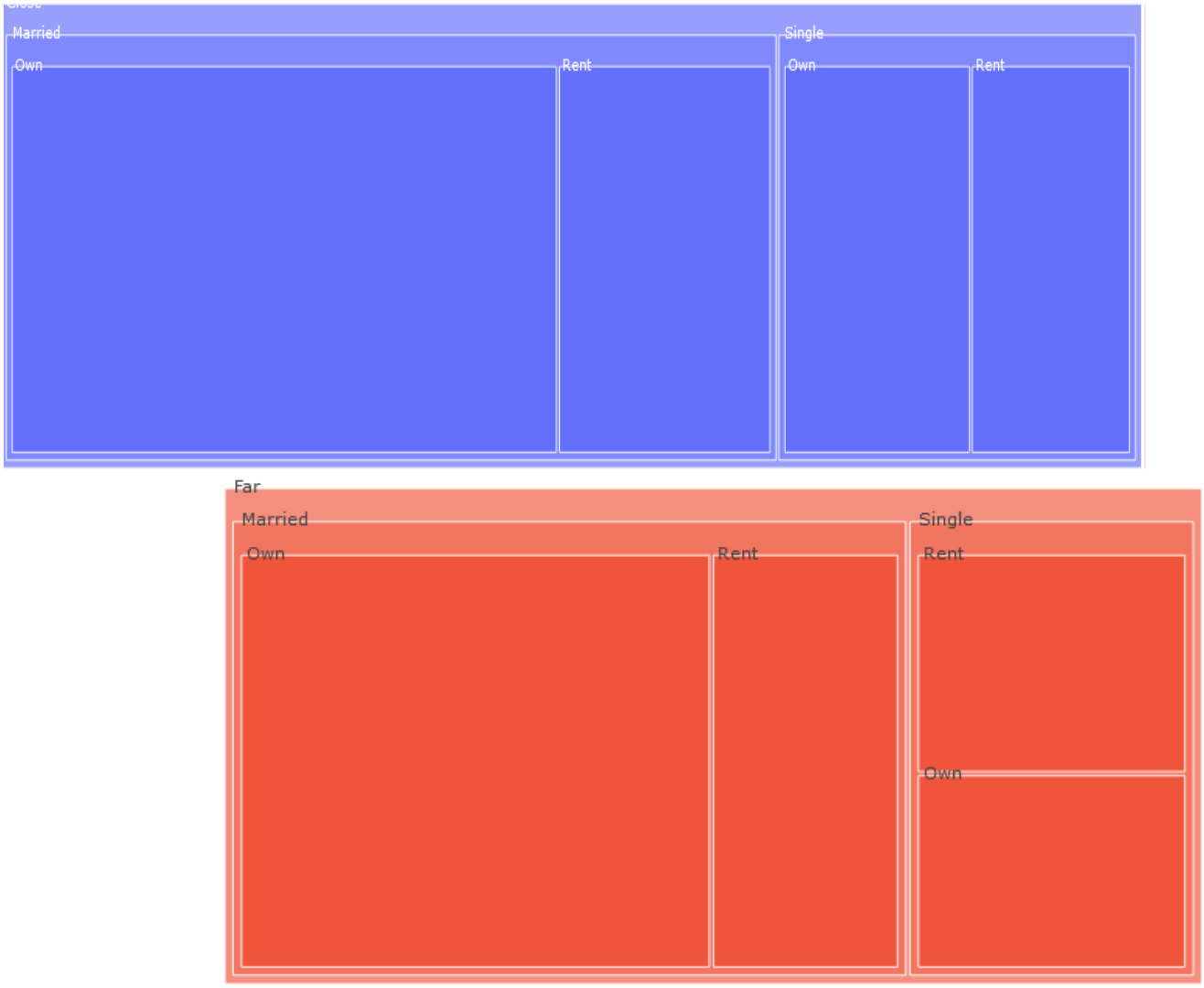
How the Amount-spent vary with gender ?

From the box-plot,
Even though they are pretty close, we can see that the average of male is slightly greater than females and IQR is also wider for male.
Taking this into concern,
Let us also add the age also in this comparison.
The Age is classified into four categories, we'll differentiate each category based on gender and see the results.

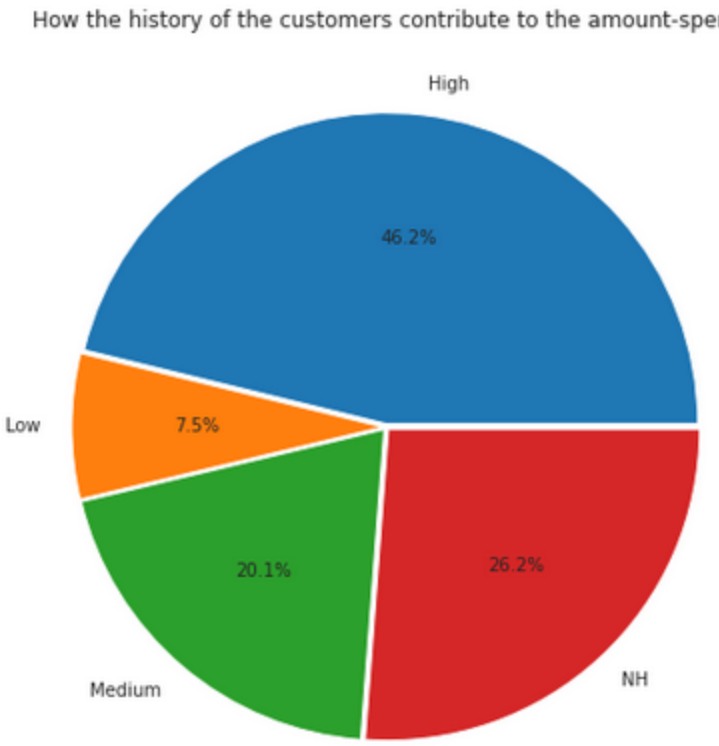
From the plot we can see that Across all categories of age, men are spending a lot in the store.
So initiative should be taken to engage more women to spend to increase the profit.



Tree Map : With a hierarchical classification in the order :
Home (Close : Blue ; Far : Red) => Marital Status => Own_home (Own/Rent)
The Area of each square denotes the amount spend , more the area higher spent in the specific classification.



From the first level, We can say that the people who are closer to the branch of the store tend to spend more than the people who live far away. Likewise we can identify a class of people who should be targeted more to improve sales.



The insight is somewhat obvious, People who have a history of high-volume purchase is tending to maintain a high average amount-spent,
But an interesting point is people who don't have a history also contribute a little over ¼ th of the time.
And we can see how different the average of Low volume is compared to Medium volume.

Hypothesis Test.

1. The variance of the salary of customers was studied using samples of 20 men and 24 women, and the info obtained was :
   sample of 20 men: mean = 64590.0 variance = 1023062000.0
   sample of 24 women: mean = 44979.164 variance = 798585199.2753624
   Test with 5% significance whether there is a significant difference in variance salary of customers with respect to their gender

   Ho : **σ**1^2 = **σ**2 ^2 : There is no significant difference
   Ha : **σ**1^2 > **σ**2 ^2 : variance of salary of men is greater than women

Solution :

F stat = S_1^2 / S_2^2

, Where S1 and S2 are variances of Men and Women respectively

$\mathbf{F}_{\alpha, 19,23} = \mathbf{0.47}$

F stat

```
F = S1_sqr / S2_sqr
F

1.2810931143331084
```

= 1.281 ,

Since the F statistic is greater than tabulated value we fail to accept the null hypothesis, Therefore there is sufficient evidence to prove that there is a significant difference in the variance two populations.

2. Fit a Simple Linear regression model taking salary as the regressor and with a sample of n = 50 to the data and test the significance of regression :

   Ho :- B1 = 0 There is no significance of regression

   Ha :- B1!=0 There is significance of regression , where B1 is coefficient

Solution:

The statistic we use in this case is also F, but the parameters are :

$$F = \frac{SSR\ /\ K}{SSE/(n-(k+1))}$$

Where k is number of regressors

SSR = ∑ (y_pred - Y_mean)^2

SSE = ∑ (y_actual - y_pred)^2

```
SSR = sum((y_pred - Y.mean())**2)
SSE = sum((Y - y_pred)**2)

f_stat = (SSR/1) / (SSE/(50-2))
f_stat
```

```
42.218881760597625
```

```
import scipy.stats as ss
ss.f.ppf(0.05,1,50-2)
```

```
0.0039734750402595645
```

F > Fα,1,48,

This means that we fail to reject the null hypothesis ,

Therefore we don't have sufficient evidence to conclude that there is no significance of regression.

3. From a sample of 25 observations of amount spent, test the hypothesis that :

   Ho : Population mean is equal to 1216$: μ = 1216$

   Ha : Population mean is not equal to 1216$: μ ≠ 1216$

   Use 0.05 significance level

$$t = \frac{xbar - \mu}{S/\sqrt{n}}$$

+t α/2,24 = 0.68

-t α/2,24 = -0.68

```
sample = df.AmountSpent.sample(25)
x_bar = sample.mean()
mu = df.AmountSpent.mean()
s = sample.std()
print('x bar :',x_bar)
print('pop mean :',mu)
```

```
x bar : 1326.36
pop mean : 1216.77
```

t = -0.2912

```
-0.6848496281662989 < -0.29126495842482547 < 0.6848496290936935
```

The calculated t falls in the acceptance region therefore null hypothesis is accepted.

Further Analysis.

- *The Dataset has some categorical columns in it. We can use the chi square test for finding out the independence of the variables.*
- *From the significance of regression test one for simple linear regression, We can extend towards significance of regression test for Multiple linear regression.*
- *We saw two columns with skewed distribution and some outliers, residual analysis using various errors such as r2 score , standardised residual , studentized residual can be done.*

Summary.

This dataset is limited to a single marketer or store owner, we can extend this idea to a broader aspect, for example : Sales of a particular brand, corporate network, etc.
The features present in the data are able to capture the most important qualities we need to predict, analyze and summarise our results. But it is limited, there are other factors in which sales of a company or a product depends on which may be dependent on the place or the company but the presence of some generic features and extension of places will also improve the quality of the data.