Dhirubhai Ambani Institute of Information and Communication Technology

CS306 - Data Analysis and Visualization Course Project

Assigned By: Prof. Pankaj Kumar

Prepared By:
Dishant Goti (201801449)
Siddharth Moradiya (201801470)

DAIICT Winter 2021

Contents

1 | Analyze E-commerce Customer Dataset using Linear Regression

An E-commerce company sells clothes online, but they have a physical store where they can come to receive advice on styles. They try to identify if a company should focus on improving mobile app experience or on website experience for their customers. This report can provide the information about whether the company owner should invest more money on the online app or website for their customers.

Here we are going to use E-commerce company Customer dataset which contains 500 details of 500 different customers about 8 different features. Dataset provides the information of customers like Email Address, Residential Address, Avatar. Dataset provides mathematical information of the customers like Avg. Session Length, Time on App, Time on Website, Length of Membership, Yearly Amount Spent. Here time is given in minutes for all numerical columns except for Length of Membership (years) and Yearly Amount Spent (Dollars)

Now using this parameter we have to conclude that the Clothes company should invest money to improve their App or company should invest money to improve their Website. Using Linear Regression technique we will analyze the effects of different parameters like Avg. Session Length, Time on App, Time spent on Website, Length of Membership (Premium card), Yearly Amount Spent effects on customers. Also, we will analyze how these parameters affect each other.

Time on App and Time on Website are important in company's growth and we want to see how they effect on yearly amount spent parameter. So, we can propose the hypothesis like,

Hypothesis: Time on App and Time on Website are the two factors that drive Yearly Amount Spent.

First we have to explore the given dataset to test this hypothesis. We have to check if any null value present in the dataset, outliers available or not, how the parameters are distributed and correlation between different parameters etc.

1.1 Exploratory Data Analysis (EDA)

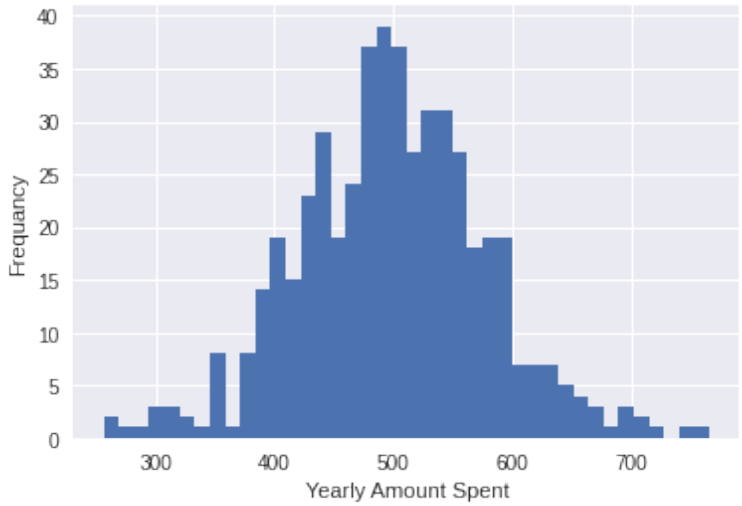

Figure 1: Histogram for yearly amount spent by customers

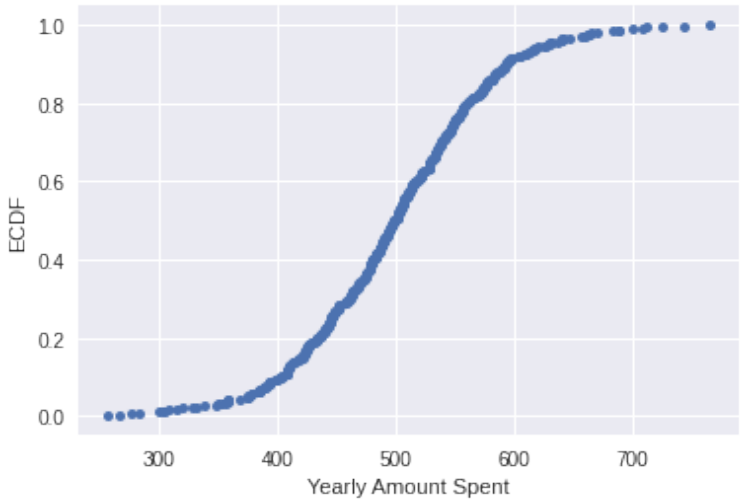

Figure 2: ECDF for yearly amount spent by customers

Here, from the figure(1) we can observe that their is no outliers available on the dataset and from the figure(2) ECDF gives us a sense for the probabilitistic distribution of the data.

1.1.1 correlations in the data

Figure 3: Correlation matrix

Here from the above heat map we can see that correlation between any two parameter. Correlation value is between -1 to 1. If value is near to 1 than we can say that the parameters are highly correlated (Positive Correlation). If value is near to -1 than we can say that the parameters are inversely proportional (negative correlation). If value is 0 than the parameters are uncorrelated.

Here, "Yearly Amount Spent" and "Time on App" is positive (0.5) and significant but correlation between "Time on Website" and "Yearly Amount Spent" is negligible (close to 0). "Yearly Amount Spent" is also positively correlated to the "Length of Membership" parameters.

1.1.2 Scatter plot of Yearly amount spent vs Time on Website

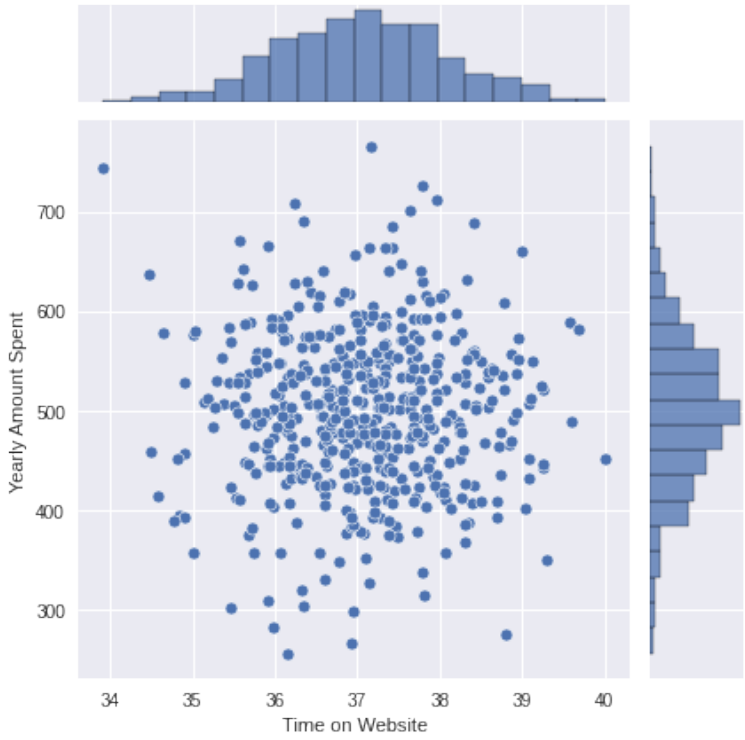

Figure 4: Yearly amount spent vs Time on Website

Correlation of "Time on Website" and "Yearly Amount Spent" is : -0.0026

We can see that the correlation is very low, at almost 0, and also the plot tells us the same. Here we have expected that there to be some correlation between time spent on the website and the yearly amount spent but here correlation value is nearly 0. So, even without any further analysis, we can say that Time on Website is not a driver for Yearly Amount Spent.

1.1.3 Scatter plot of Yearly amount spent vs Time on App

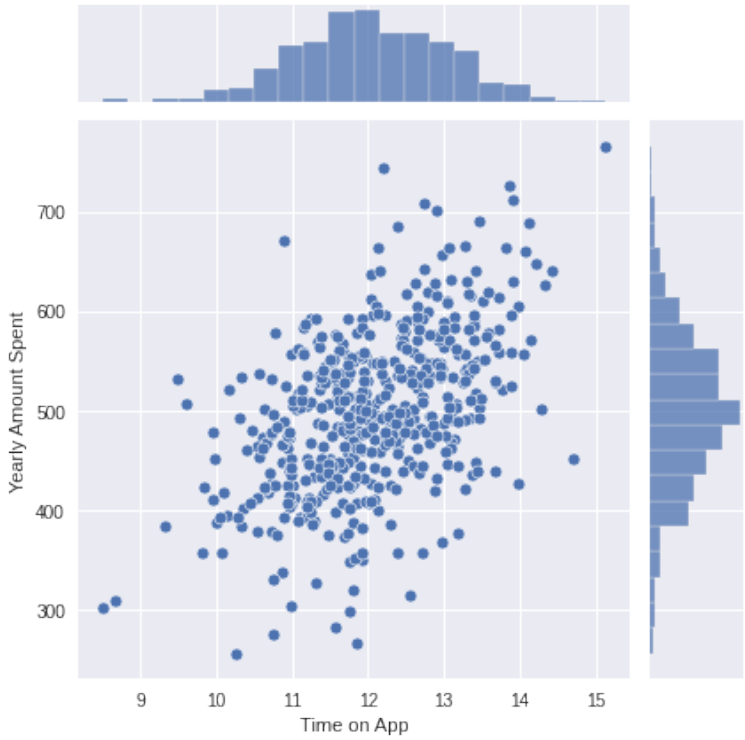

Figure 5: Yearly amount spent vs Time on App

Correlation of "Time on App" and "Yearly Amount Spent" is : 0.499

We see a larger correlation coefficient and the plot shows a more positive correlation, specially in comparison to the Time on Website plot.
Now, we want to see the comparison of each other parameter through pair plot.

1.1.4 Pair Plot

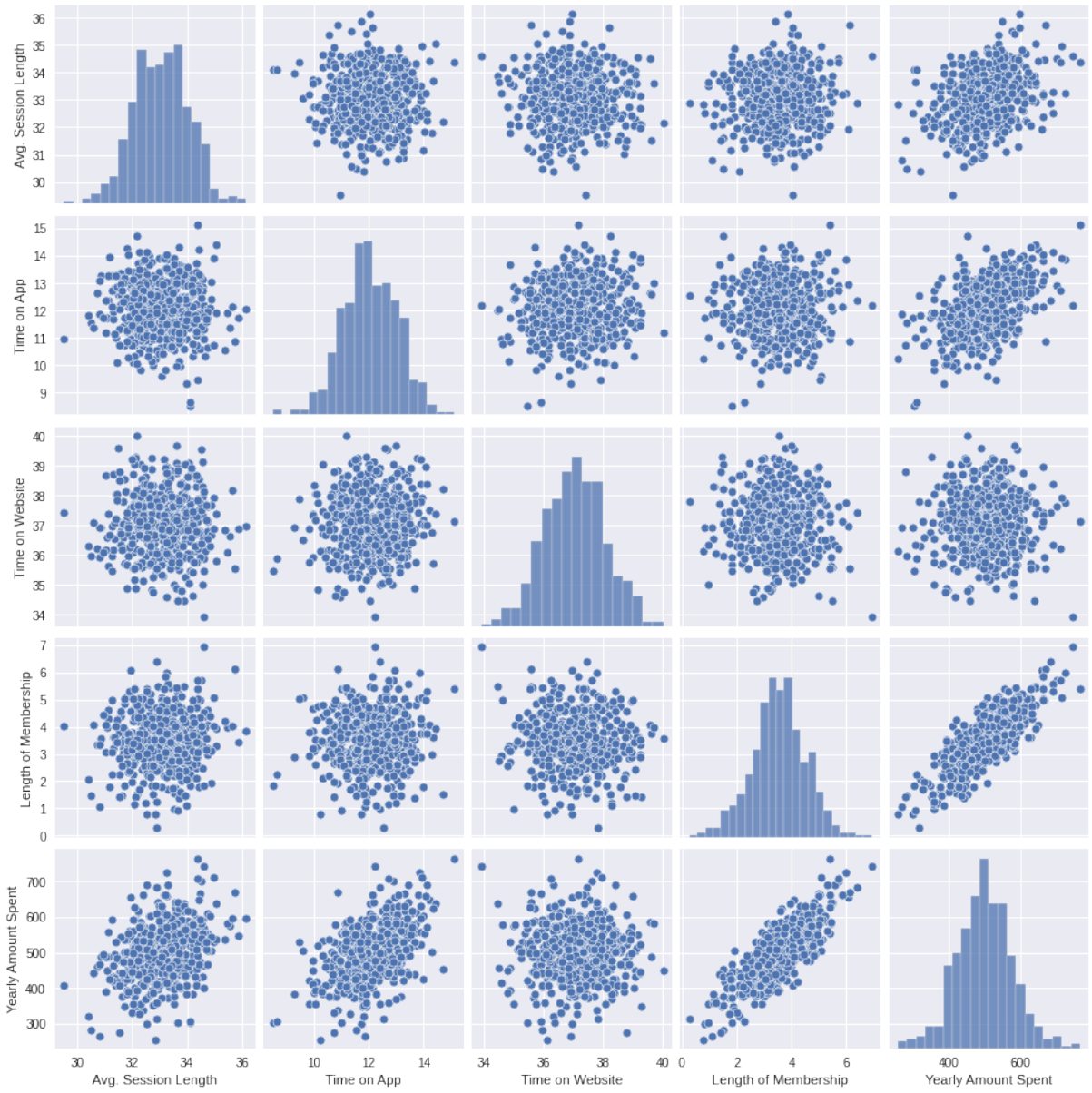

Figure 6: Pair Plot

From the above pairplot we can see that yearly amount spent feature is correlated with the features Avg. Session Length, Length of Membership and Time on App. Here we can see that avg. Session Length, Time on App is slightly correlated and Length of Membership is heavily correlated with the Yearly Amount Spent.

Here from the pair plot we can also observe that all of our other categories do not appear obviously correlated with one another, making a linear regression analysis model and try to predict Yearly Amount Spent.

From the above pair plot Length of membership looks like a linear relationship with the yearly amount spent, so let's do Linear Regression analysis for both parameter.

1.2 Linear Regression of Yearly amount spent and Length of Membership

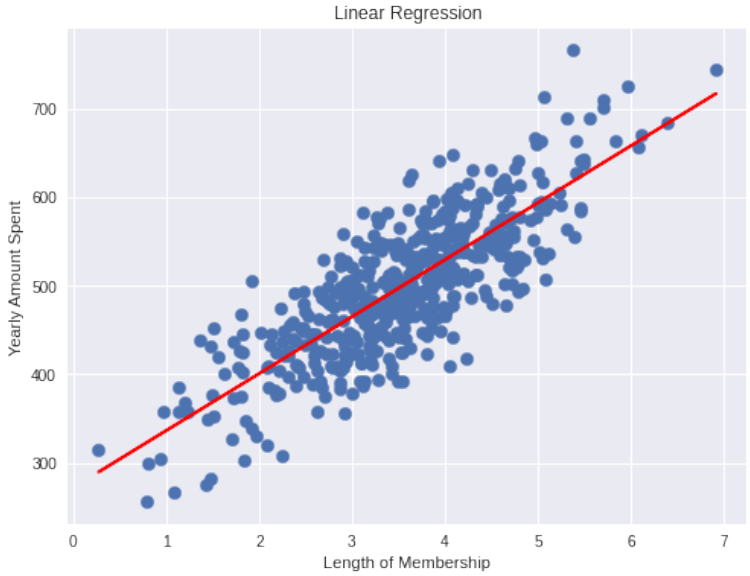

Figure 7: Linear Regression

$$\beta_0 = 272.399786058034$$
$$\beta_1 = 64.2186843155828$$
Y(Yearly Amount Spent) = β_0 + β_1*X(Length of Membership)

Here we can clearly see that this simple linear fit is good for the parameter "Yearly Amount Spent" and "Length of Membership". The longer you stay a member, the larger your Yearly Amount Spent. Also we can relate this parameter with the real life example for any kind of membership (Here membership in clothing company).

We have taken only one parameter effect on yearly amount spent. Now let's add other features (Avg. Session Length, Time on App, Time on Website, Length of Membership) and see how it effects on yearly amount spent.

1.3 Multiple Linear Regression Model

In this multiple linear regression analysis we want to show that the Yearly Amount Spent feature is linearly depend on the different features like Avg. Session Length, Time on App, Time on Website, Length of Membership. Now, we will going to Train the model, Predicting the Model with test data, and evaluating the model. We have divided our main dataset into two parts: Training dataset and Testing dataset. We have taken 30% data as a testing data and 70% data as training data.

1.3.1 Training the Model

we are interested in creating a model that can help us make prediction decisions for the parameter yearly amount spent.
Our Model is:

$$Y(\text{Yearly Amount Spent}) = \beta_0 + \beta_1*(\text{Avg.Session Length}) + \beta_2*(\text{Time on App}) + \beta_3(\text{Time on Website}) + \beta_4*(\text{Length of Membership})$$

We applied multiple linear regression technique and found the coefficient value for different parameter as shown below:

Coefficients of:
Intercept (β_0) = -1047.9327822502387
Avg. Session Length (β_1) = 25.981549723495792,
Time on App (β_2) = 38.59015875311409 ,
Time On Website (β_3) = 0.19040527751100633,
Length of Membership (β_4) = 61.27909654482186

1.3.2 Predicting Test Data

Now we are going to use testing data and feed this data to the linear regression model and compare it with its original value.

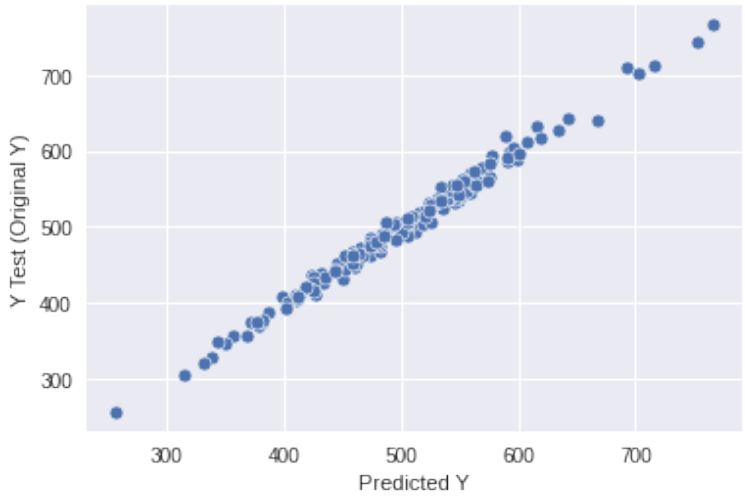

Figure 8: Real vs Predicted Value of yearly amount spent

Here from the above figure(8) shows the plot of actual values versus predicted values. we can see the mostly straight diagonal line being a perfect prediction of the testing data that means our model predicts data well.

1.3.3 Evaluating the Model

Now we want to evaluate the linear regression model before drawing any conclusions. The explained variance score (R^2) is used to determine how much variance the model explains. R^2 value is in between 0 and 1. The closer to the 1 the better the model. Other common calculations to measure the performance of a regression model are Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Here we used RMSE to calculate the error between Actual value and predicted value of the yearly amount spent.

$$R^2 : 0.9891$$
$$RMSE : 8.9338$$

Our R^2 value is almost 99% which is very good, as our model describes almost 99% of the variance in the sample.

We want to explore the residuals and we're hoping to see something that is mostly normally distributed.

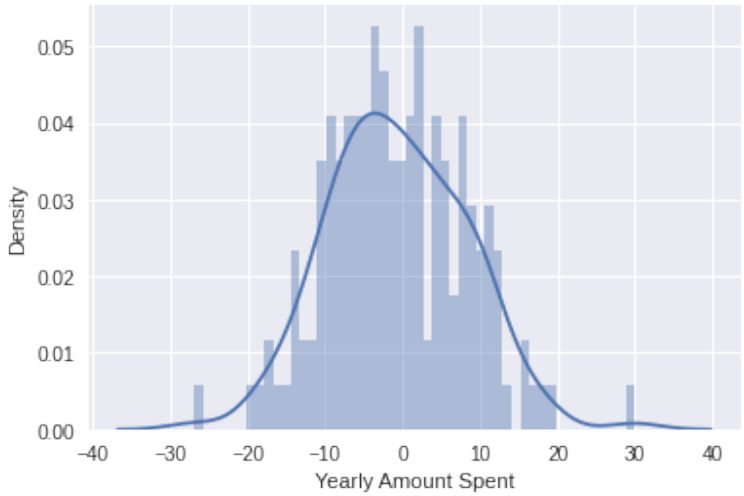

Figure 9: Histogram of Real-Predicted Value of yearly amount spent

Here we plotted the histogram of (Actual value - predicted value) of yearly amount spent feature and we found that with a low average error compared to the magnitude of the values we are working with, and the residuals plot looking normally distributed, this can be considered a good model.

2 | Conclusion

Attribute	Coefficient(β_i)
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

From the above table we can see different coefficient value for different feature. From this table company can decide what should they do to increase their customers. Here we can see that an increase in the Length of Membership (in years) would affect more on the value of our customers for Yearly Amount Spent ($). Next parameter which affects on yearly amount spent is Time on the App.

Time spent on App has more impact than Time spent on Website so, here we can say that company should focus on invest money on App because the app provides greater profitability for one more minute increase compared to the time on the website but it also depends on what the costs of developing the app vs. the website are. It's clear that the website needs more work compared to the time on the app, but it may be more cost effective to continue working on the App instead of bringing the website up to speed.

Different economic factors would determine which course of action to take, but at least we now have knowledge of the state of our website and our app in terms of yearly spend per customer.

With addition to this, however, Length of Membership was the greatest impact in the amount a customer spent yearly, meaning that the longer the customers stay with the company, the more money the company will make in the long run. So, length of Membership should be included as the economic factors to decide between focusing on their app experience or their website.

Click To view the Code: **Google Colab DAV PROJECT**.

Click To view Github Repository: **GITHUB DAV PROJECT**.