**Name:** Matheus Vinicius Ferreira Figueiredo Teixeira

**Student Number:** 301236904

Assignment 3 – Linear Regression

# Exercise 1

## Introduction

First, we were asked to generate a scatter graph for a given linear function: y = 12x – 4. 100 values for x were generated ranging uniformly from -1 to 1, and the respective values for y were calculated using the given function than the 100 ordered pairs (points) where plotted.

Second, we were asked to add some noise to each point. The noise was generated using the standard normal curve with a mean equal to 0, and a standard deviation equal to 1.

## Conclusion

In conclusion, we have an original plot with 100 points over the line segment (-1, -16) to (1, 8); these points were generated using an equation (“theoretical prevision”); however, in order to emulate real data, we have added a small noise (“random error”) to each (“predicted”) y value, add that is why on the second plot each point is a little bit above or a little bit bellow the theoretical line. Figure 1 presents the original graphic, and Figure 2 the graphic with noise addition.

Chart, scatter chart

Description automatically generated

Figure 1: Scatter plot of a linear function for 100 points.

Chart, scatter chart

Description automatically generated

Figure 2: Scatter plot of a linear function for 100 points with noise addition.

# Exercise 2

## Introduction

This exercise is about Linear Regression. From a given database (CSV file), we were expected to select the appropriate features and generate a linear model to predict an output.

First, we were given tabular data saved on a CSV file. Then, we were asked to explore the data; this step is important to get familiarized with the data and have an idea of what we are working with. After getting familiarized with the databank, we got to know that there were some categorical variables such as gender.

We cannot plug values like “male” or “female” into a linear mathematical equation; hence, we were asked to transform the categorical variables into numerical values. For instance, we have converted a single column (Gender) holding words like “male” and “female” into 2 columns, namely Gender\_Male and Gender\_Female. If the customer's gender is male, the column var\_male is marked as 1, and the column var\_female is marked as 0. The next step was dropping the Transaction ID because this is a non-numerical value with no real meaning; it is not a category; it is only an identifier.

At this point, all our data were now numerical. The last step before being able to use all this data to generate a mathematical equation is normalizing it, which means bringing all values to a compatible scale. As we can notice in Figure 3, after normalizing the data, all of them now range from a 0 to 1 scale. In addition, we can see in Figure 3 that:

1. we have 5 binary variables;
2. ‘Total Spend’ approximately follows the normal curve; and
3. the other 5 continuous variables are approximately uniform, despite some picks and valleys.

We also plotted a scatter matrix (Figure 4) that gave us a hint that there is a linear relationship between the “Monthly Income” and the “Total Spend”.

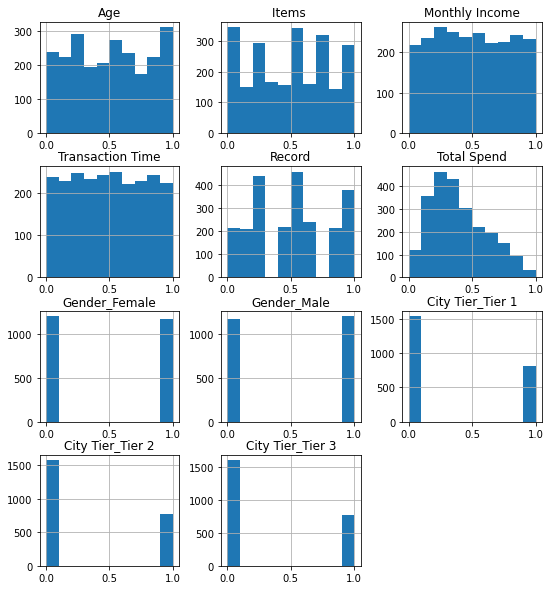


Figure 3: Histogram of each variable.

A picture containing chart

Description automatically generated

Figure 4: Scatter Matrix.

Finally, we should choose some variables to use as parameters and generate the linear regression model. The exercise asked us to split the data into 65% train data and 35% test data. It also asked us to generate the linear model using 2 different sets of variables and compare them.

The first chosen set of features was:

['Monthly Income', 'Transaction Time', 'Gender\_Female', 'Gender\_Male', 'City Tier\_Tier 1', 'City Tier\_Tier 2', 'City Tier\_Tier 3']

The second chosen set of features was:

['Monthly Income', 'Transaction Time', ‘Record’, 'Gender\_Female', 'Gender\_Male', 'City Tier\_Tier 1', 'City Tier\_Tier 2', 'City Tier\_Tier 3']

## Results

### First Model

Using the first set of features we got the following weights for our Linear Model:

[ 3.27189813e-01, -5.59612829e-03, -1.65228186e+13, -1.65228186e+13, -1.03982886e+12,

-1.03982886e+12, -1.03982886e+12]

And a coefficient of determination (R²) equals to 0.19515366679700408.

Looking at these results, we notice, for example, that the 'Transaction Time' weight is very small (almost insignificant) if we compared its order of magnitude with the order of magnitude of the 'Monthly Income' weight. On top of that, all the weights for discrete data are close to -∞ (extremely big negative numbers). All of this is an indication that this is not a good model. In fact, the coefficient of determination is suggesting that roughly only 20% of the variation in the output (Total Spend) can be explained by this model; this confirms that we have generated a very poor model.

### Second Model

Using the second set of features we got the following weights for our Linear Model:

[ 0.32566986, 0.01741505, 0.59634381, -0.01115803, 0.01115803, 0.0051521, 0.00365729,

-0.00880938]

And a coefficient of determination (R²) equals to 0.9175710979534945.

With the addition of 1 single feature (‘Record’), we got to generate a much better model. All weights now have a nice order of magnitude, and the coefficient of determination is pointing that about 92% of the variation in the target (Total Spend) can be explained by this model.

### Conclusion

Comparing both models, we can notice how important is to make good feature choices. One single feature difference can determine if a model is very good or very poor.