**TechPro Academy**

**Retail Sales Analysis**

**Dimitris Katos**

[dimitriskatos6@gmail.com](mailto:dimitriskatos6@gmail.com)

You can find my work on the two provided links.

* Github: <https://github.com/DimitrisKatos/Retail_Sales_Amount>.
* Google Drive:

<https://drive.google.com/drive/folders/1Gh84HHPBWkYKdYSzw3FMjTcBTfcfuJIK>.

**Analysis Goals**

**Data Cleaning and Preparation**

The first goal of the assignment is to prepare the dataset for the upcoming analysis steps like, descriptive statistics and other. The preparation of the dataset can be found in the following link: <https://colab.research.google.com/drive/1Zp6IqFrgWk4HA2oumvqsRFbsFPyn6mjU>.

Let’s move forward to the details of the data cleaning and preparation. We have three main goals in this section of the assignment:

1. We must ensure that every column of the dataset is in the provided format. For example, SalesDate column must be in date time format.
2. Detect outliers, duplicates or null values.
3. Ensure that every column have consistent data.

First of all, we print the use the function info of pandas module. This function returns the non-null values of the dataset and the type of values in every column. We can check that some columns isn’t in the right format, so we will try to convert into the desired format. Also we can check if we have some duplicates rows in the dataset.

The SalesDate is in object format, but we want to be in date time format. We use pandas function to convert the type of the values to the desired format. In the next step, we focusing on the SalesAmount column. In this column we can see that some values we have some inconsistent data like ‘-’ or ‘%’. So, we must remove all this characters and then we convert the type of the values to integers.

After that we move forward to the ProductCategory and CustomerGender columns. We use a pandas function to return the unique values of the columns. The results gives as, that in the ProductCategory we have ‘Female’ in one row. This value is clearly wrong and we must do something about this value. Also in the CustomerGender column we can see that we have some inconsistent data like ‘Clothing’, ‘Unknown’ and ‘Did not answer’.

After finding this inconsistencies, we must do something about them. We find the row which the ProductCatgory has the value ‘Female’. In this row we can easily detect that CustomerGender is ‘Clothing’ which is completely wrong. So, we change the values of this two columns. After that, we drop all the rows from the CustomerGender column which values are ‘Did now Answer’ and Unknown’.

In the next we focusing on the ProductRatings and CustomerLocation columns. In the first column we can check some inconsistent data. We detect the value 10, but the column has ratings from one to five so we drop those columns. Also the values of the columns must be in strings format but the stings must be integers like ‘1’ or ‘2’. Also we detect a string value ‘two’ and we convert this value to ‘2’. Finally, we convert the type of the data to strings. Also we ensure that the CustomerLocation column has only consistent countries. If we detect some inconsistent data we drop this row.

The remaining column to check is the CustomersAge column. The first step in this column is to detect some outliers by creating a boxplot. In this step we also check for outliers in Sales Amount column

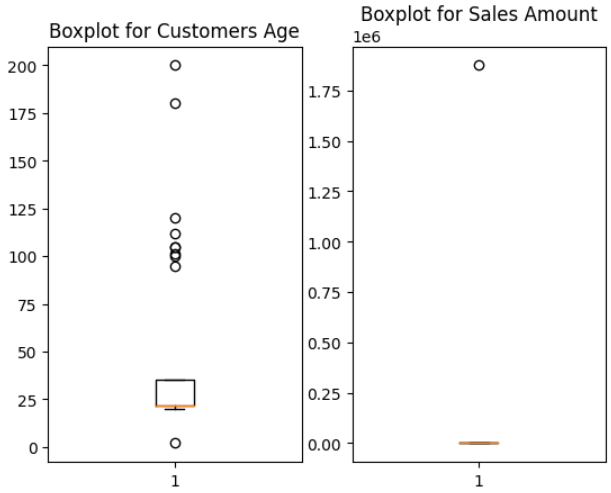


Image 1. Outliers from Boxplot.

From the above image we can see that we have many outliers in the Customers Age column. We apply a z- score to detect all those outliers and then we drop those values. Also we detect the outliers from the Sales Amount column and also drop them. We convert the values to string format and finally we extract the clean dataset to use it for the following steps of this assignment.

**Descriptive Statistics and Visualization**

In this chapter we will try to calculate some basic statistics measurements to understand sales trends, customer demographics and product performance. Then we will visualize the results using Matplotlib and Seaborn. The code for descriptive statistics and visualization can be found in the following link: <https://colab.research.google.com/drive/18U8fLUDbc3NjVV2t9GIX37r4Z7eZRtlr#scrollTo=VP6xW2tewrBF>.

For understanding the Sales trends we extract the Quarter from every year of the dataset. After that, we will use the pivoting method to understand the Sales Amount per quarter and per product. We calculate the mean values for every Product Category. Also we will count the amount of sales every product category was sold per quarter.

In the following plot we can see, the mean sales amount per product per quarter.

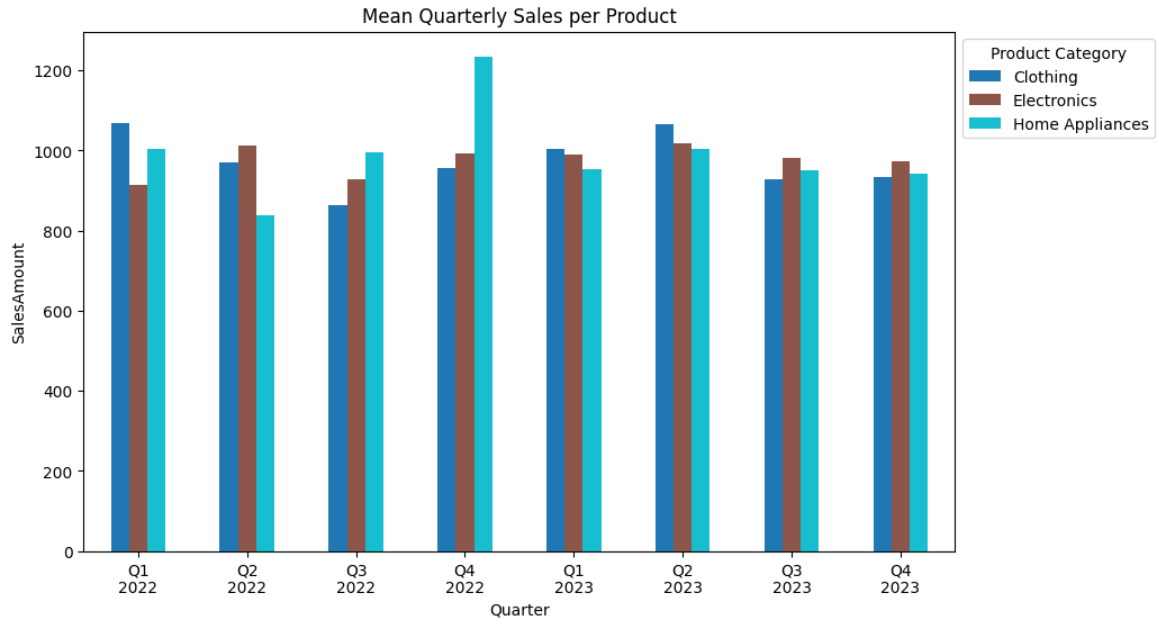


Image 2. Mean Quarterly Sales per product.

We can also visualize the amount of sales of every product in every quarter. We use a time series plot to understand the sales per quarter.

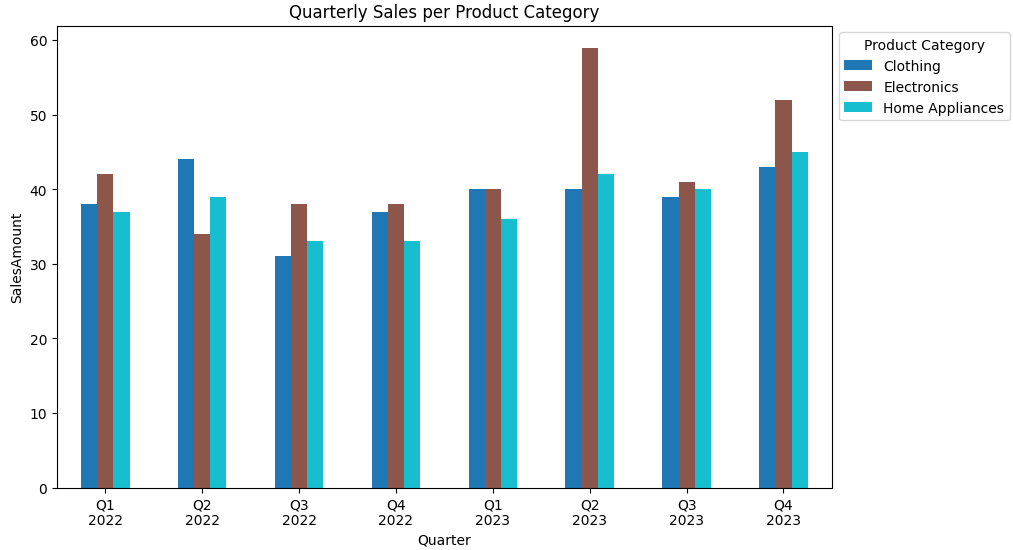


Image 3. Time Series for Product Category per Quarter.

We move forward to extract information about customer demographics. In the following plot we visualize how many customers from every country we have.

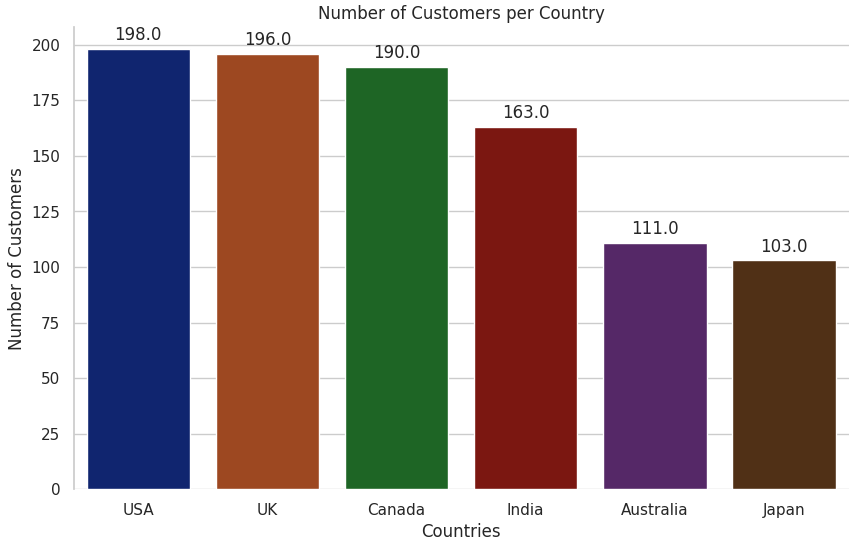


Image 4. The customers per country.

Now we can visualize how many customers per gender and per country we have.

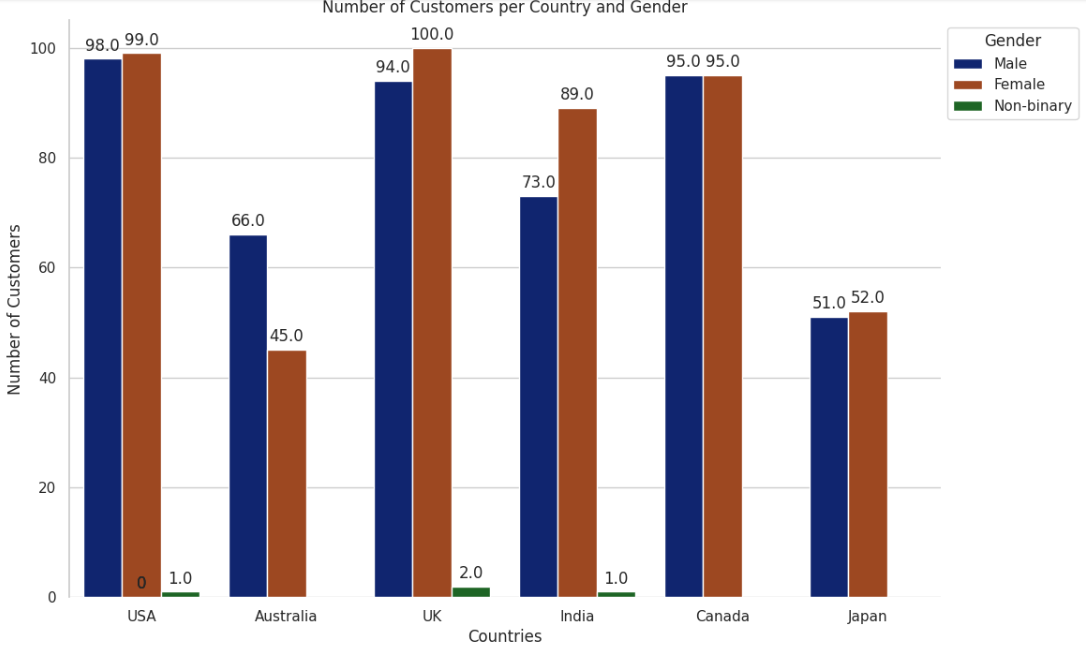


Image 5. Customers gender per country.

We can also visualize what product every country bought.

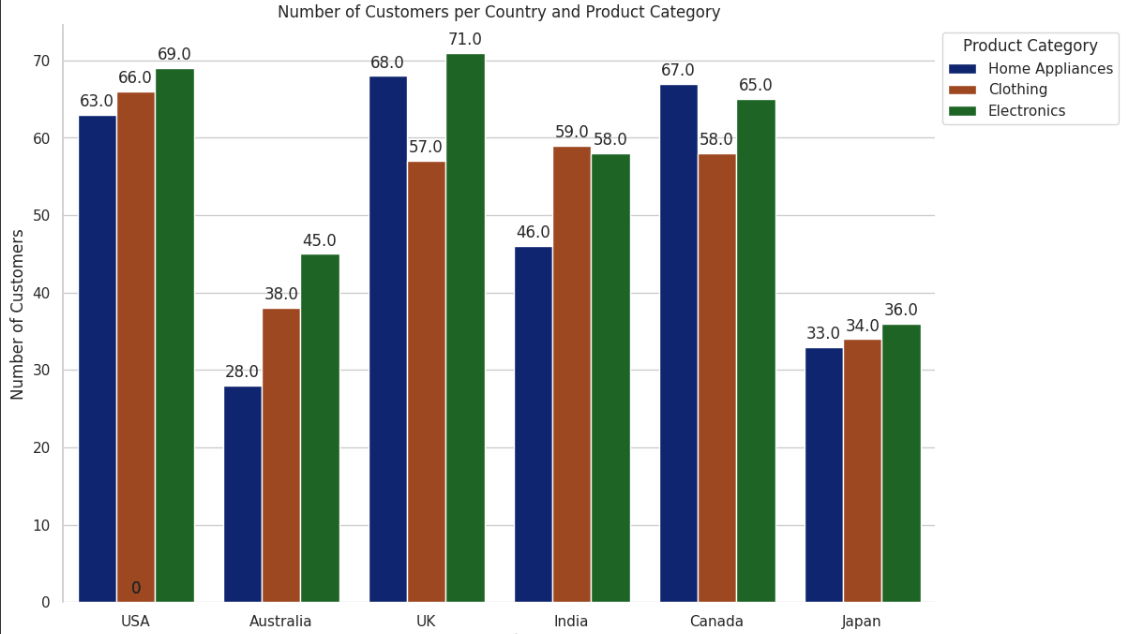


Image 6. Product sales per country.

Let’s move forward to calculate some statistics for product performance. Also we will visualize the results. We use grouped by method to extract information for the product performance. After extracting the information we visualize the mean sales amount per product category and per customer Gender.

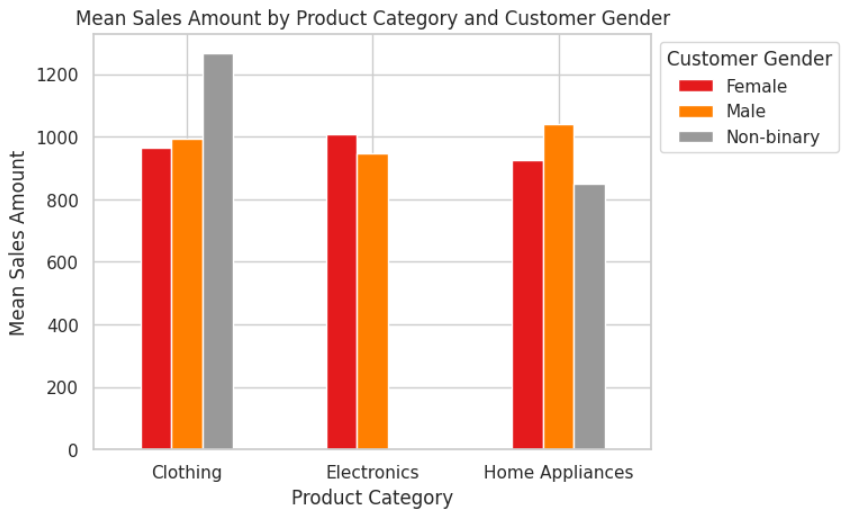


Image 7. Mean Sales amount per product per gender.

We also visualize the mean product rating value per product and per Gender.

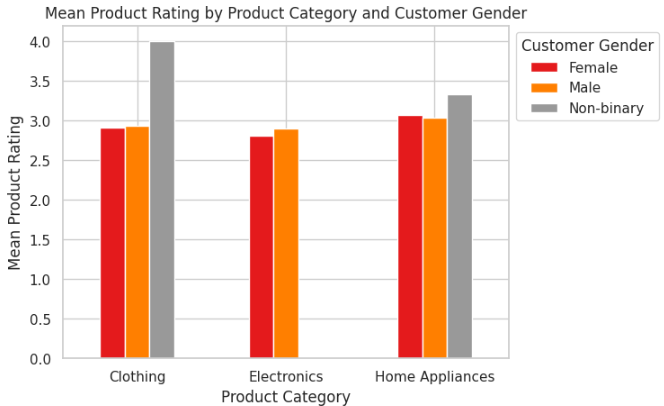


Image 8. Mean Product Rating per product and per gender.

**Inferential Statistics**

The main goal in this chapter is to perform some hypothesis testing to understand the dataset and draw conclusions from the data. The code for this chapter can be found in the following link: <https://colab.research.google.com/drive/1RtGIjxXDPzL4SNpBAmdB-wpMRYTsuftO>.

We will perform two tests, one for product category and one for Customer’s locations. We will perform a one way Anova test for product category. The anova test is a statistical test used to analyzed the difference between the means of more than two groups. The test will tell us if the dependent variable changes according the level of the independent variable. In our case independent variable is the product category and the dependent is Sales Amount.

* Null Hypothesis (H0): The Sales amount has no difference between the products categories.
* Alternative Hypothesis (Ha): There is difference between product categories.

After performing the test we fail to reject the null hypothesis because the p- value equals 0.97. So, there is no significant difference between product categories.

Now we perform an Anova test to test if average sales differ significantly between regions.

* Null hypothesis (H0) : There is no difference between the locations.
* Alternative hypothesis (Ha): There is difference between locations.

After performing the Anova test we reject the null hypothesis because of the fact that the p-value equals 0.03. So there is a significant difference between Customer’s locations.

**Advanced components**

In this chapter we try to build some predictive models. We will try predict sales amount by applying simple linear regression and multiple linear regression. The code can be found in the following link: <https://colab.research.google.com/drive/1UxAfJu77O8wJSjCb0q93qky-1334EZG0>.

The first for the prediction model is to select the column which will be used for the model. In fact we will use all the column apart from the SalesDate. The regression analysis doesn’t allow the categorical variable, so we will use the pandas function get\_dummies to convert the data. This function is a very powerful tool. The function create new columns and fill the properly with 0 or 1.

Now we can check the correlation between the columns.

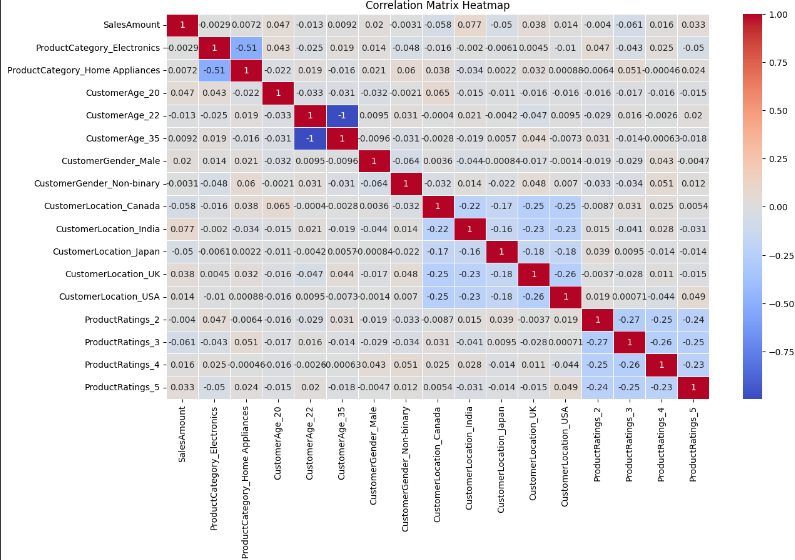


Image 9. Correlation between columns.

From the above heatmap, we can check that the correlation between Sales Amount and the others variables is very weak. So we expect that the regression analysis won’t make good predictions.

Let’s try a simple linear regression model between the Product category and Sales Amount. At first we define the depended and independent variable. The we split the dataset to training and testing values. The we create and train our model. Our train score equals 0.004, which is significantly low. Also the mean squared error which is a value to measure the amount of error in statistical models is very high 285260.

So, we can easily say that our prediction model is not good and doesn’t help us predict the SalesAmount based on the product category. As mentioned before, the result of the model doesn’t surprised us because there is weak correlation between the two variables.

We can also visualize the actual and the predicted Sales Amount to see that the model is not good.

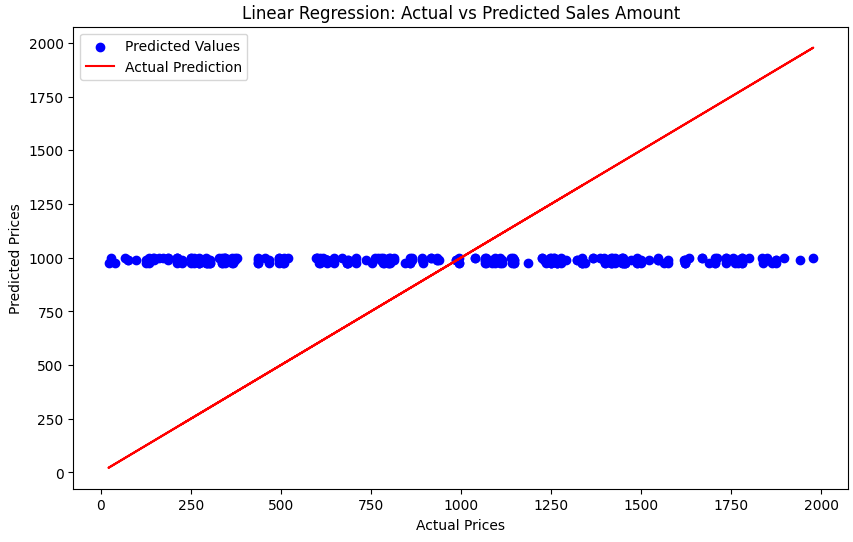


Image 10. Actual vs predictions.

Also the multivariable regression analysis doesn’t work properly. We can visualize the results of the new model.

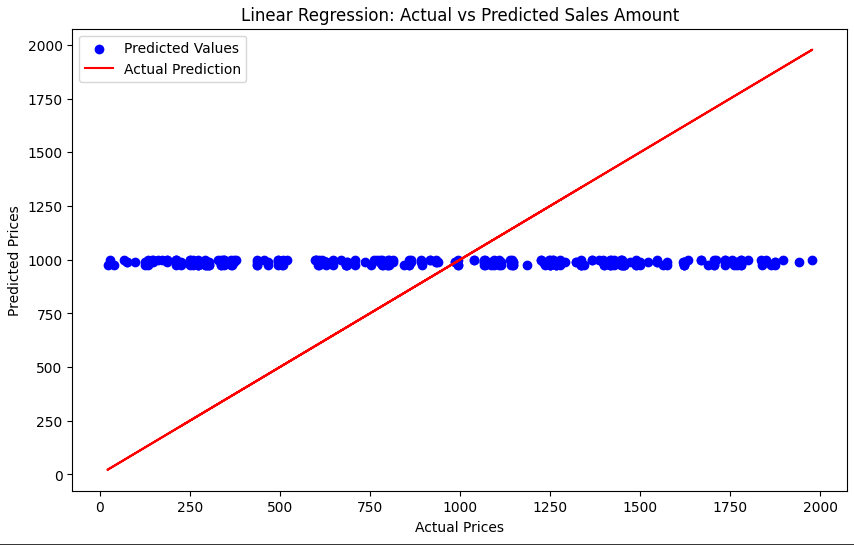


Image 11. Multi- variable regression analysis.