

|  |  |
| --- | --- |
| Subject Code: | COSC2789 |
| Subject Name: | Practical Data Science |
| Title of Assignment: | Assessment 3: Group Project |
| Student name and student number: | Vo Minh Thien An (s3916570)  Do Ha Minh Long (s3634734)  Nguyen Phuong Nam (s3877256) |
| Teachers Name: | Ms Thuy Nguyen |
| Number of pages including this one: | 11 |

I declare that in submitting all work for this assessment I have read, understood and agreed to the content and expectations of the Assessment Declaration.

# **Data Modeling for Online Retail II Dataset**

# 

# *Group 10*

# Abstract

This report is about modeling the sales dataset from a retailer from late 2009 to the end of 2010. In this report, we will use some techniques to clean the data and use different methods to classify, cluster, and regress the data. Then we will see the accuracy rate between each method while bringing out our observations about this dataset.

In this report, we will mainly focus on Regression and Clustering - two techniques that we used in this assignment

Table of Contents:

[Assessment 3: Group Project 1](#_Toc125053074)

[**Data Modeling for Online Retail II Dataset** 1](#_Toc125053075)

[1](#_Toc125053076)

[*Group 10* 1](#_Toc125053077)

[Abstract 2](#_Toc125053078)

[Table of Contents: 2](#_Toc125053079)

[1. Introduction 3](#_Toc125053080)

[2. Methodology 3](#_Toc125053081)

[2.1 Retrieving and Preparing the Data 3](#_Toc125053082)

[3. Result 10](#_Toc125053083)

[3.2 Clustering 10](#_Toc125053084)

[3.3 Regression 10](#_Toc125053085)

[Conclusion 10](#_Toc125053086)

[References. 11](#_Toc125053087)

[Team Member Contributions 11](#_Toc125053088)

# Introduction

E-commerce companies, such as online retailers, need to segment their customer base in order to have a deeper understanding of their customers' routines. The clientele are segmented into subgroups by the system in accordance with their demographics and the patterns of their purchases, as demonstrated in the code and this report later.

There is an extremely large quantity of files at this location. Before carrying out any type of analysis on the data, it is necessary to first completely clean them. If these enormous data sets are not managed appropriately, it is possible that the results will not be what they seem to be. We will go to the next level using both the information we already knew and the new data we have.

# 2. Methodology

We decided to build some clustering and regression models with several differences’ methods and techniques.

2.1 Retrieving and Preparing the Data

The goal of this assignment is to model the data from a UK-based retailer so that it can be used to make predictions and financial decisions in the future. For that, the accuracy of the data is paramount.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Invoice** | **StockCode** | **Description** | **Quantity** | **InvoiceDate** | **Price** | **Customer ID** | **Country** |
| **0** | 489434 | 85048 | 15CM CHRISTMAS GLASS BALL 20 LIGHTS | 12 | 2009-12-01 07:45:00 | 6.95 | 13085.0 | United Kingdom |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **525460** | 538171 | 21931 | JUMBO STORAGE BAG SUKI | 2 | 2010-12-09 20:01:00 | 1.95 | 17530.0 | United Kingdom |

525461 rows × 8 columns

The data contains eight attributes and over 500 000 entries. The first step is to identify any *null* entries and drop them.

|  |
| --- |
| Invoice 0  StockCode 0  Description 2928  Quantity 0  InvoiceDate 0  Price 0  Customer ID 107927  Country 0  dtype: int64 |

The sum of all null values, group by attribute

All null value entries are dropped, resulting in 107 927 entries being removed, which line up with all entries that were missing in customer ID column.

Running data.info() command, we confirmed all null entries have been removed

|  |
| --- |
| Invoice object  StockCode object  Description object  Quantity int64  InvoiceDate datetime64[ns]  Price float64  Customer ID float64  Country object  dtype: object |

Next, we check the various attributes for typos, illogic, and other mistakes.

First we check the “Country” attribute, by counting and displaying all values within that group

|  |
| --- |
| Australia 654  Austria 537  Bahrain 42  Belgium 1054  Brazil 62  … …  West Indies 54  Name: Country, dtype: int64 |

No major spelling or logical mistake was found here.

Next, we check the item's description.

|  |
| --- |
| DOORMAT UNION JACK GUNS AND ROSES 53  3 STRIPEY MICE FELTCRAFT 117  4 PURPLE FLOCK DINNER CANDLES 17  ANIMAL STICKERS 12  … …  ZINC WILLIE WINKIE CANDLE STICK 272  Name: Description, dtype: int64 |

After manually going over all entries, we identify major spelling inconsistencies throughout the list. All of these errors were caused by inconsistent word spacing and the use of special characters such as '>, / >,, >,... To clean up these mistakes, we decided to strip all spaces and special characters, resulting in the entries being counted correctly. Although this resulted in the entries being turned into a long continuous string of characters, the fact that human readability is not a major concern meant that this was of minimal issue.

Afterward, we check the “Price” attributes.

|  |
| --- |
| 0.000 31  0.001 14  0.010 1  0.030 4  0.040 1  ..  7044.790 1  8985.600 2  10468.800 3  10953.500 3  25111.090 1  Name: Price, Length: 664, dtype: int64 |

Despite the large number of small values, as they are mostly used for manual entry of discount, refunds and service fees, those value a left in the data

2.2 Feature Engineering

We created several new columns dubbed “revenue”(describe total price, ”month”, ”year”, “month\_year” and “Last Active” to help organize some of the data so that it could be more easily worked with. ”month”, ”year”, “month\_year” are data extracted from the “InvoiceDate” into separate attributes containing the month, year and the full date without timestamp respectively. The “revenue” is calculated using the “Price” and “Quantity” attributes. "Last Active" is worked out by looking at the date of the last entry in the list and comparing each entry to that date.

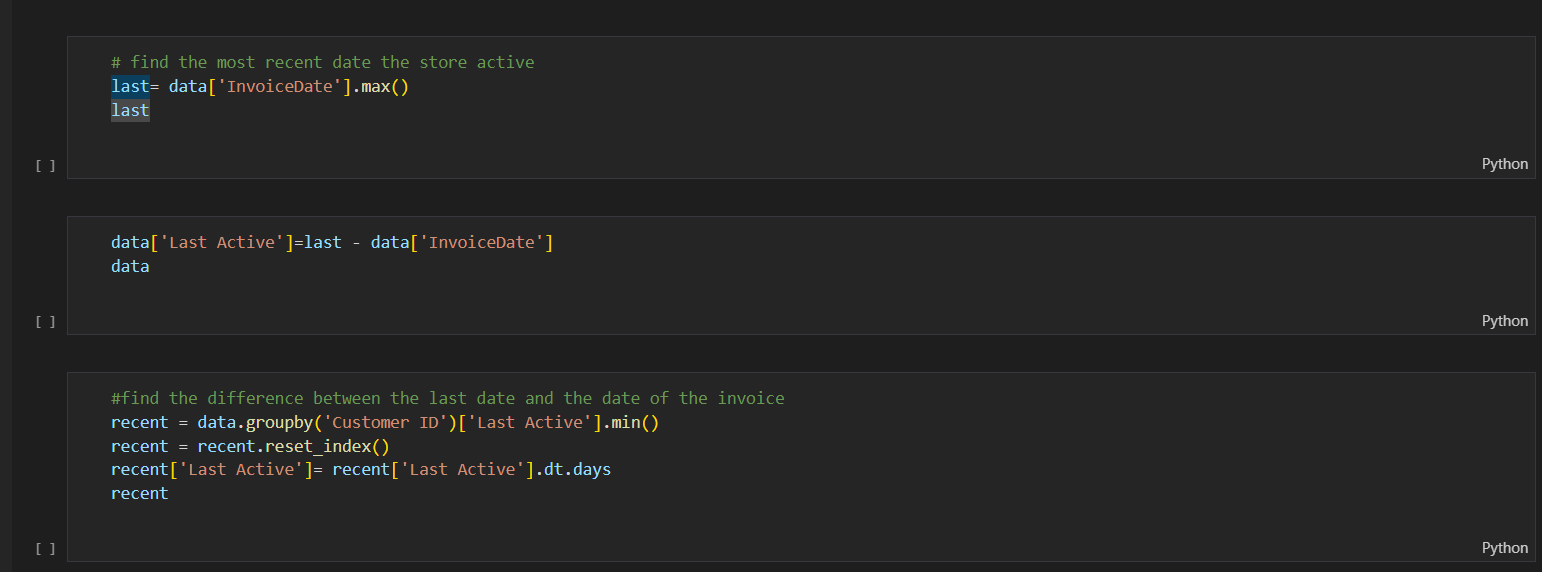
2.3 Data Modelling

* Clustering

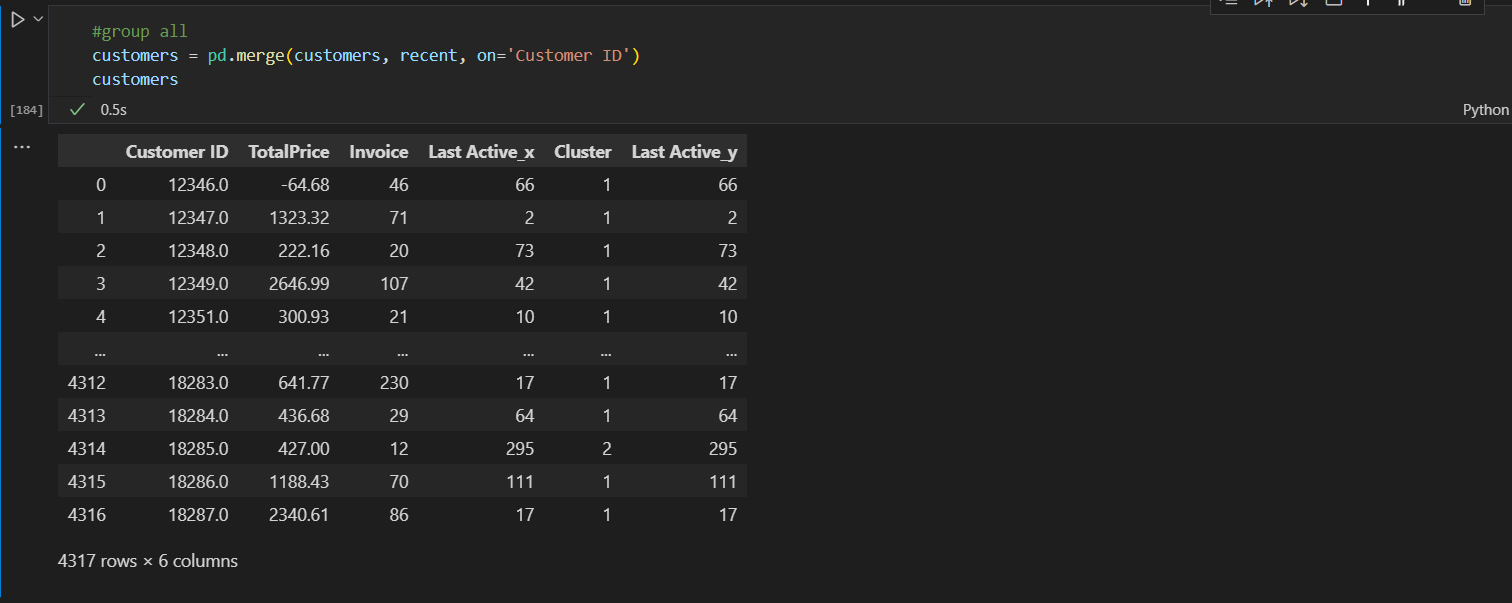
For this task, we use the K-means clustering technique with the elbow method. Furthermore, silhouette score is calculated to find the best number of cluster(s)

* KMean cluster: KMean clustering is one of the most simple unsupervised techniques being used in this task. The purpose of K-means is straightforward: to group similar data points and uncover underlying patterns. To accomplish this purpose, K-means searches a dataset for a predetermined number (k) of clusters.

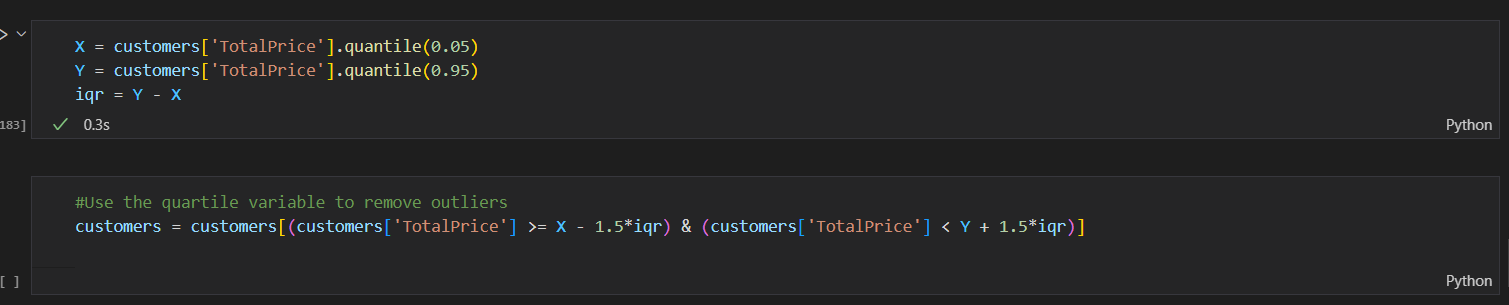
First, we calculated the customer's last active date. By calculating that, we can cluster customers who have not visited the store.



Then we merged the inspected columns (CCustomerID, Total Price, Involve, and Last Active).

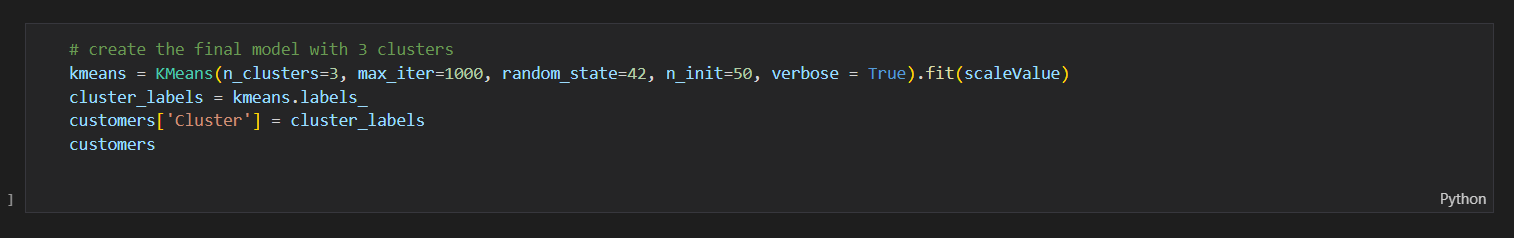


We deleted outliers in order to scale using a quartile value, since the data standards varied greatly across columns. The quantile function is very important because it gives us the value for the first 5% of the Distribution on data in Amount column and same for the rest 95%.



because the Kmeans cluster requires a value of K for the algorithm to work properly. Therefore, in order to get the best possible value for "K," we decide to use the elbow graph. Finally, we created the model with 3 clusters and visualized the clustered data using boxplot

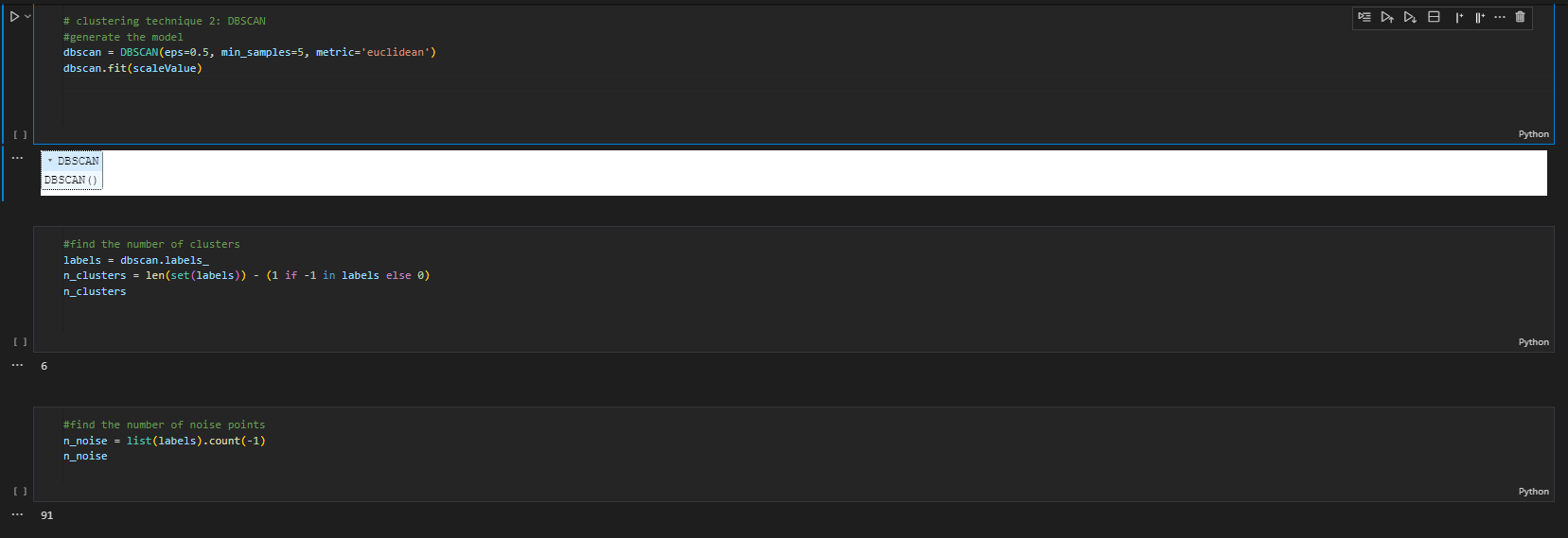




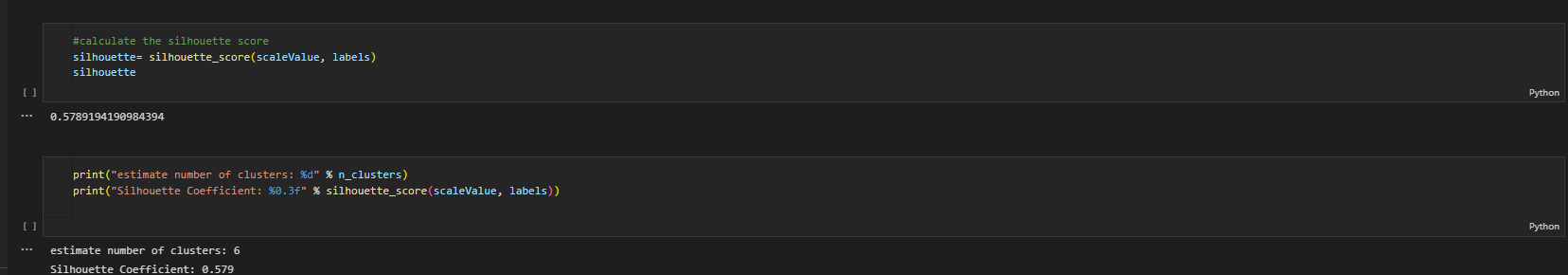
* DBSCAN:

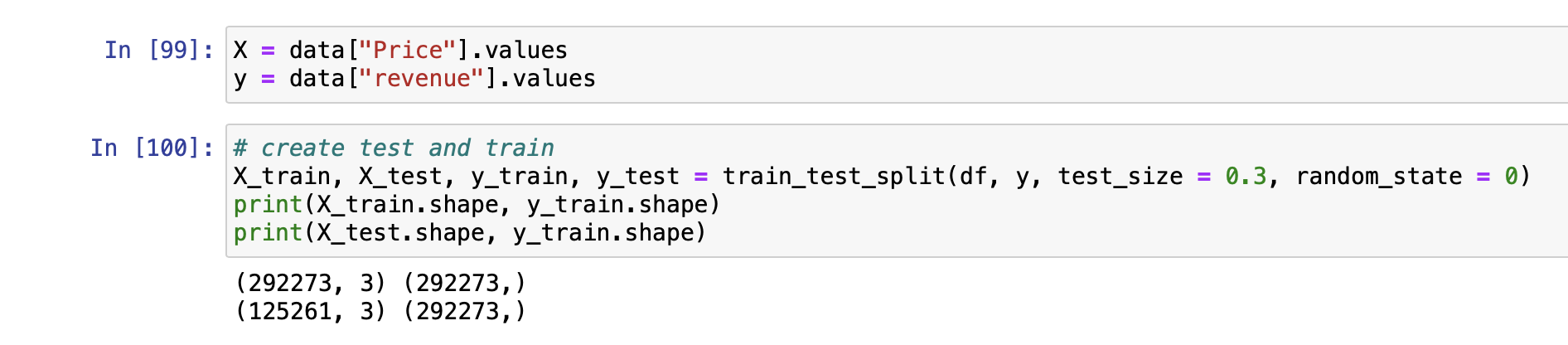
DBSCAN is a density-based clustering technique that believes clusters are dense regions in space separated by regions with a lower data point density. By examining the local density of data points, we may discover clusters in massive datasets.

First, we generated the model



After that, calculate silhouette score to compare with the previous technique



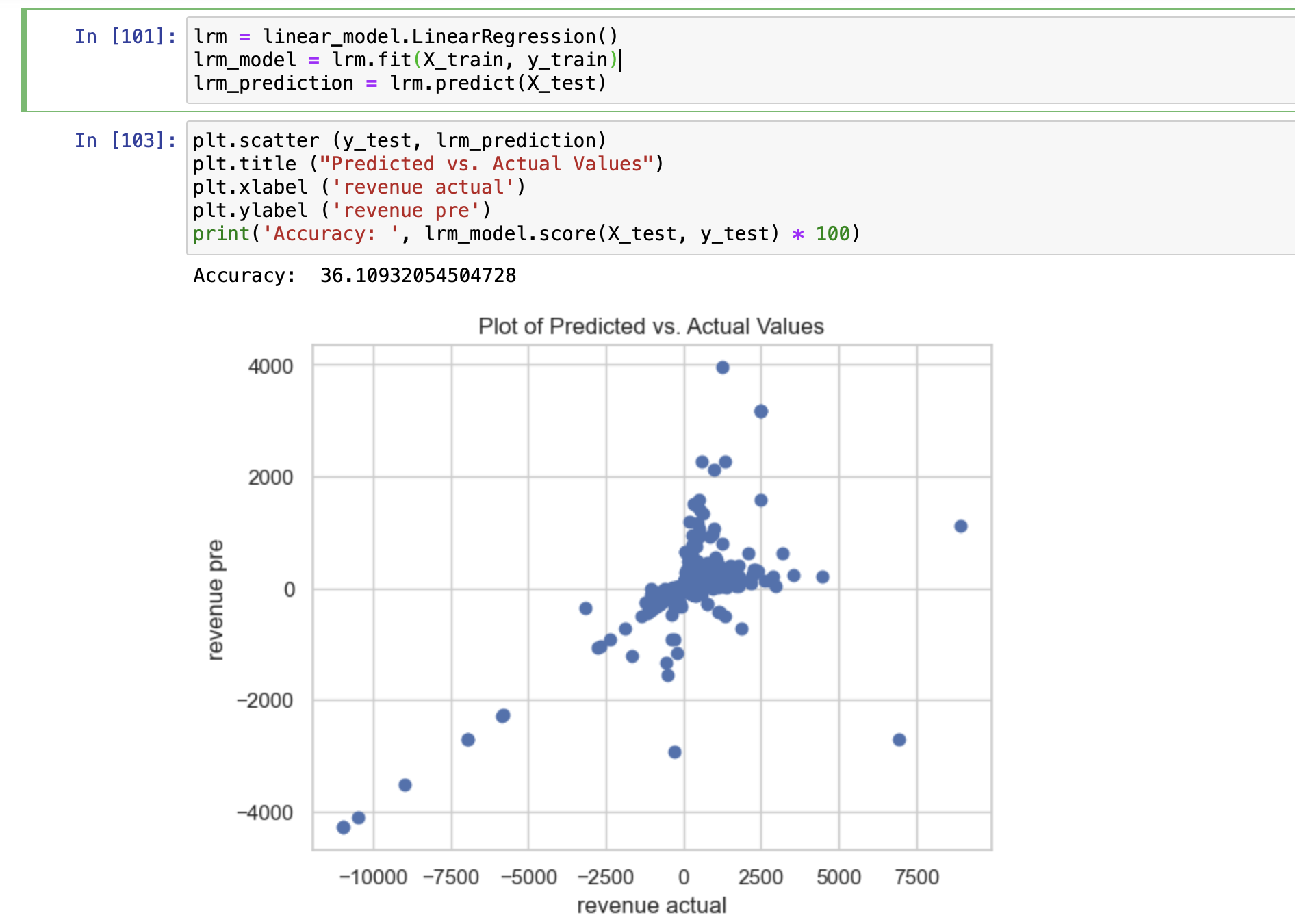
* Linear Regression: this is the most simple method that used to predict the revenue by month in this case

In this method I used “price” and “revenue” as X and y, respectively.

Price is an independent variable that can affect revenue, and revenue is a dependent variable.

Then use train\_test\_split to split the data into train and test data. The ratio here is 70-30, random\_state = 0 means that the data will remain the same each time.

Then use linear regression to train and predict the data. The prediction is then used to create a graph to show us the accuracy of this method. And it's almost 36%.



Multiple Linear Regression: In this method I used multiple independence data to calculate the only revenue. It make the accuracy go very high nearly 100%

# 3. Result

3.2 Clustering

Since the data didn't show the right number of clusters, we ran tests with different numbers of clusters and made predictions about the silhouette coefficient based on what we learned. A data point's silhouette coefficient shows how much it looks like the other points in the cluster to which it belongs. This coefficient can range from -1 (highly dissimilar) to 1 (very similar). You may also figure out the total number of clusters by utilizing the elbow method. Because there are only small differences in silhouette score between the two methods, the models give results that are a little different from each other.

Because of these problems, the k-means algorithm might not work as well as you would expect for many different datasets, especially low-dimensional datasets. However, the trained model are well suited for our initiall targets.

3.3 Regression

The results show that there is a huge difference between linear regression and multiple linear regression. The accuracy percentage is 36% to 99%, and the graph shows a big difference between the two methods.

# Conclusion

The final outcome shows that we can use these methods to visualise and predict the target that we need. The model aligned with the testing data. We already get the customer value separated in three different groups, and also have predicted the revenue for next year. Even though the classification did not work very well, we did understand which method is good for this kind of dataset, by understanding that we can apply it to further dataset easier.

The process of data modeling is an effective method for visualizing and preserving the information related to a company. The process of data warehousing requires system analysis, which is an essential part of the process. The goal of a data model is to serve as a source of instructions on how the information included in a company's database should be arranged. Having the same vision, however, has to be a straightforward process for all parties involved. “Data modeling must depict the structure of the data system serving various purposes such as identification of some of basic aspects of the system regarding the entities and their role in the organization”(Sargo, 2021).

# References.

# Team Member Contributions

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Vo Minh Thien An | Do Ha Minh Long | Nguyen Phuong Nam |
| Work | Regression, report and presentation | Retrieving and preparing data, report | Clustering, DBscan report and presentation |
| Contribute(%) | 33% | 33% | 33% |
| Sign | An | Long | Nam |