# Super Market Data Analysis

Introduction :

*How can we predict the profit income of the products of the supermarket based on its dataset?*

By visualizing data of supermarket dataset in 3 months to get insights on its data,

We tried to find and implement the relations between the data and each other and to predict the profit income of the supermarket in these 3 months.

We tried to make decisions based on insights extracted from the dataset to improve the supermarket situation in the market.

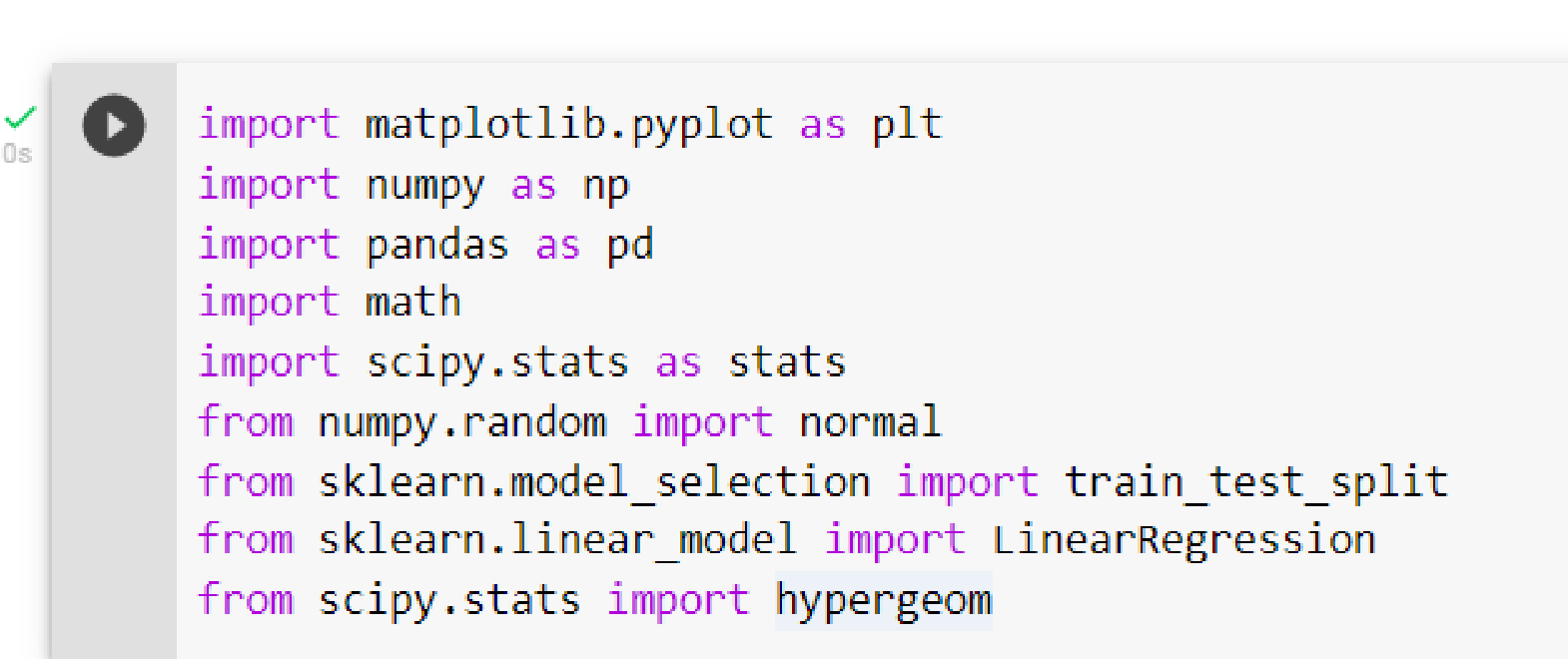
Summary of Research :

*Data introduction*

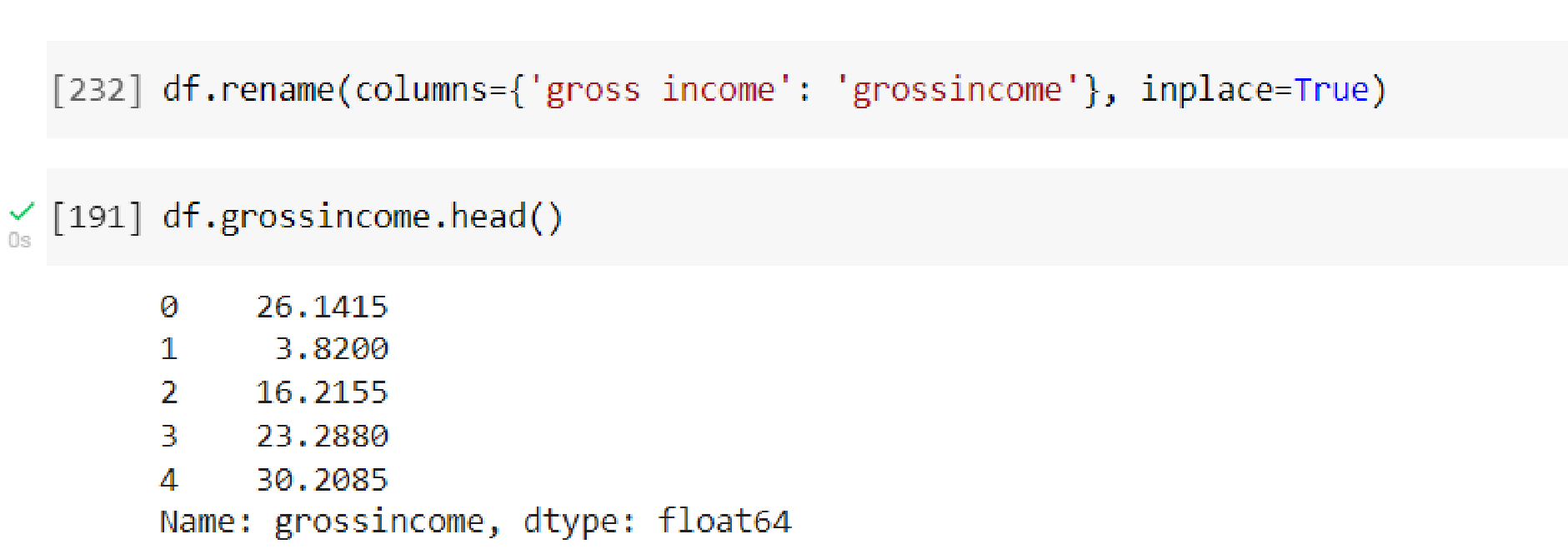
We searched for supermarket datasets and we found one on Kaggle, which is a three months dataset that is recorded in three different branches globally and we started working on it.

*Data preprocessing*

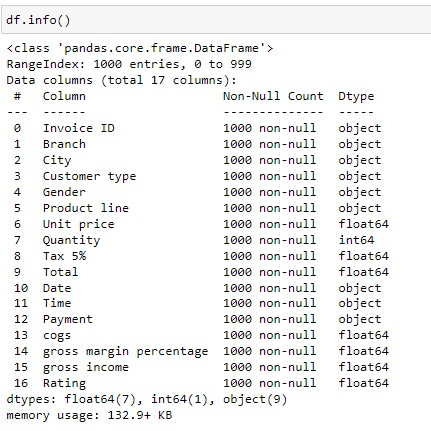
1. Loading related libraries needed for analyzing data



1. Columns renaming



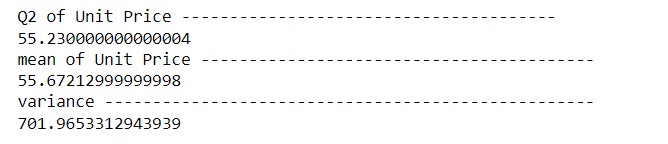
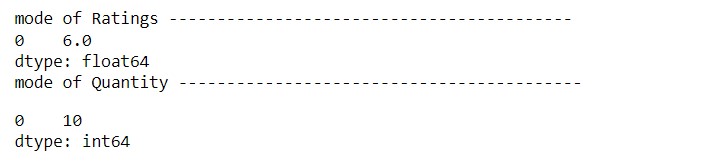
1. Checking for data information



Since there no null values in the dataset, no adjustment is required on data

*Data Analysis* Numerically:

Extracting mean, Mode, median and the variance of the numerical columns in data.



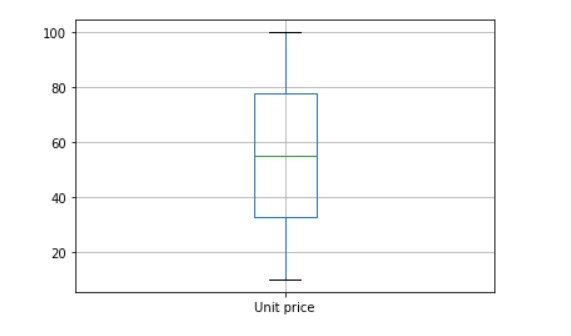
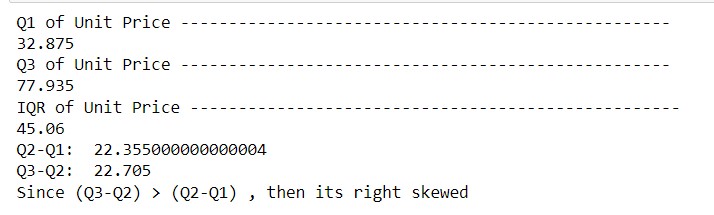
Now we can determine from the mode of Ratings that the most common rating of this supermarket is 6,

And the most common purchased quantity of products is 10.

The average of prices in this supermarket is 55.67.

The median(Q2) is extracted to help in identifying if the Unit price data column is right or left skewed.

Extracting Q3 and Q1 to get Interquartile Range.



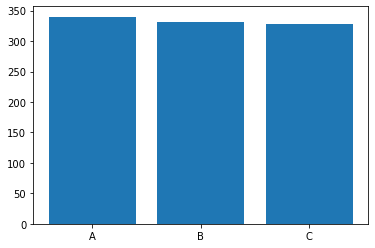
We can determine from the boxplot and the IQR that this data is right skewed.

The range of prices in this supermarket lies between

32.875 and 77.935

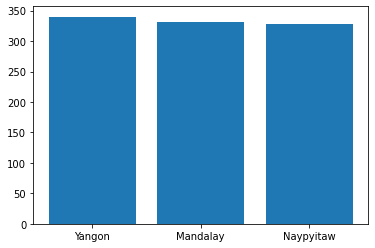
Graphically:

* Plotting the branches bar chart to know what is the most visited branch of the three.



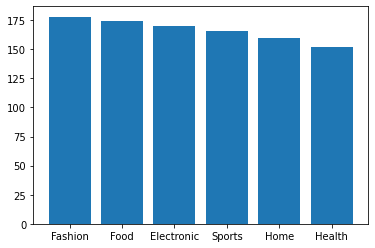
Branch A is the most visited branch of all three.

Plotting the cities graph to know where is the supermarket most popular.



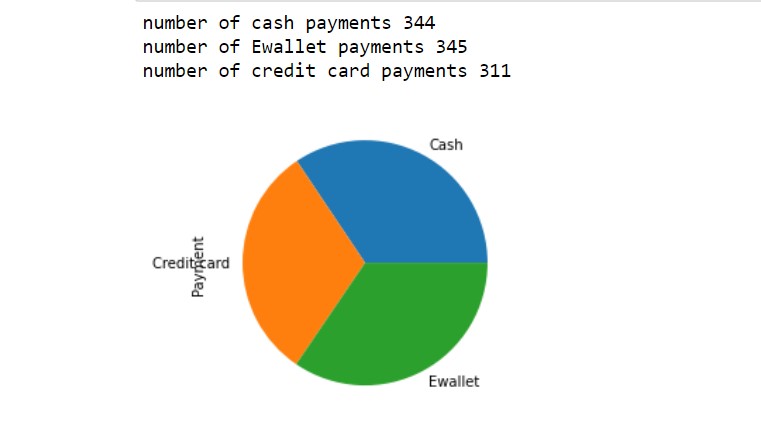
The supermarket is most popular in Yangon.

* Plotting product line bar chart to know what is the most purchased category of products.



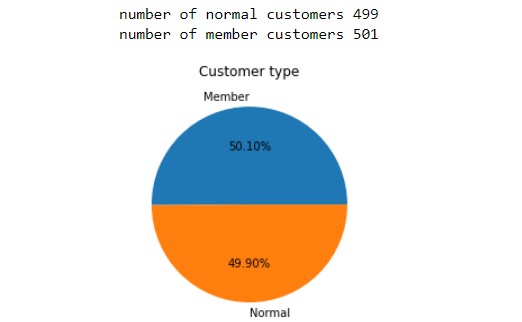
Fashion products are the most purchased.

* Plotting the pie chart of the most used payment method



Ewallet payment method is the most used by costumers

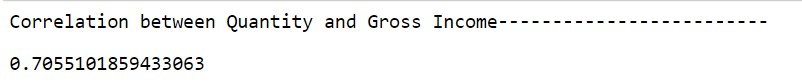
* Plotting the pie chart of the customer types to identify the most visiting customers.



Member customers are the most visiting but both percentages are not very far from each other.

*Correlation and Regression analysis for two dimensional data*

*Correlation* between Quantity and Gross Income to determine the relationship between them



There’s a strong correlation between them.

*Regression:*

Applying regression between gross income and Unit price.

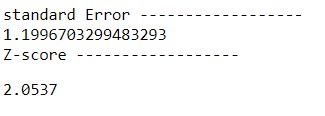
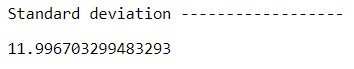
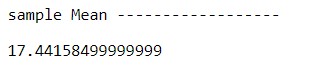
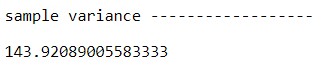


Determining the predicted Y : ŷ

And the coefficient of determination, which is how well the regression equation fits the data.

*Performing estimation and inference on our dataset:*

Getting a sample of size 100 from our data and then getting the sample mean, variance, standard deviation, alpha (1 – confidence Interval), Standard Error and the Z-score.



*Now we can conclude that:*

95% of confidence interval of growth income lies between lowerlimit 14.977822043385105 and upperlimit 19.905347956614875.

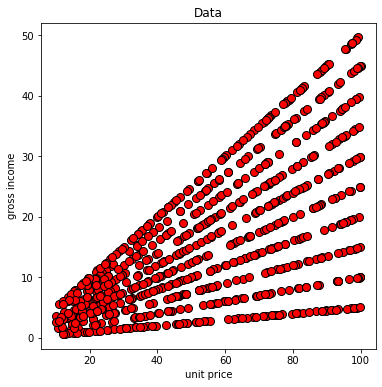
*Machine learning model:*

*Can we predict the Gross Income knowing the Unit price of product?*

We identified our dataset into training and testing

And then applied machine learning techniques on this data.

Graphical Representation of the training data:



Graphical Representation of training data with the predictive line:

