# Sales Prediction of Big Mart

##### **Ruba Alharbi 1, Ghada Aldabayan 2, Khuzama Alsalem 3**

Department Information Technology, Qassim University

***Abstract-***Sales forecasting is an important aspect of different companies engaged in retailing, logistics, manufacturing, marketing and wholesaling. It allows companies to efficiently allocate resources, to estimate achievable sales revenue and to plan a better strategy for future growth of the company. In this paper, produces better predictive performance compared to any of the popular single model predictive learning algorithms. The approach is performed on Big Mart Sales data of the year 2013. Data exploration, data transformation and feature engineering play a vital role in predicting accurate results. The result demonstrated that the statistical approach performed better than a single model approach as the former provided more information that leads to better prediction.

1. **Introduction**

**1.1 Background**

BigMart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store. BigMart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business, Also it used machine learning to predict BigMart sales enables the data scientist to do so, as it studies the various patterns per store and per product to give accurate results, The development of the Internet of Things [IOT] has been primarily driven by needs of large corporations that stand to benefit greatly from the foresight and predictability afforded by the ability to follow all objects through the commodity chains in which they are embedded. BigMart allow companies to become more efficient, speed up processes, reduce error, and incorporate complex and flexible organizational systems through IOT. Also, it has ability to solve this problem in coding by Python (is fast becoming the preferred language in data science).

**1.2 Problem specification and goal**

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. The data also includes certain attributes of each product and store.

**There is some problem for examples:**

*1)Traditional storage can cost lot of money to store big data.*

*2)Lots of big data is unstructured and Lots of effort.*

*3)BigMart analysis is not useful in short run. It needs to be analyzed for longer duration to leverage its benefits.*

*4)BigMart analysis results are misleading sometimes.*

*5)Speedy updates in big data can mismatch real figures.*

The goal is to build a predictive model to find out the sales of each product at a particular store so that it would help the decision makers at BigMart to find out the properties of any product or store, which play a key role in increasing the overall sales.

1. **Literature review**

A literature review surveys books, scholarly articles, and any other sources relevant to a particular issue, area of research, or theory, and by so doing, provides a description, summary, and critical evaluation of these works in relation to the research problem being investigated. Literature reviews are designed to provide an overview of sources you have explored while researching a particular topic and to demonstrate to your readers how your research fits within a larger field of study, may consist of simply a summary of key sources, often within specific conceptual categories , a summary is a recap of the important information of the source. But a synthesis is a re-organization, or a reshuffling, of that information in a way that informs how you are planning to investigate a research problem.

(Summary of three recent Literature Review):

**[2.1] A STUDY ON SELECTION OF LOCATION BY RETAIL CHAIN: BIG MART**

(Article in International Journal of Research - GRANTHAALAYAH · January 2019)

The modern-day projects, such as privatized infrastructure projects, have project life that is spread over many years. Projects are becoming larger and more complex. These projects involve the large capital investment, generate unbalanced cash flows, and involve complex contractual agreements. They encounter changing economic and financial situation, face unpredictable political environmental changes. The stability of modern projects is thus, constantly subjected to certain sensitive and volatile, external and internal environments (Mishra and Mallik, 2017). The main aim for this research it to understand how retail chains like Big Mart choose their location in order to maintain its existence in a competitive environment as well as maximize its profit. It is well known that the selection of location is vital for successful operation of stores. Especially in case of retail stores the location has huge importance. A knowhow of the approach towards selection of location can help policy makers in making land use plan for commercial purpose to attract investors. It may also aid similar commercial enterprises during location selection process.

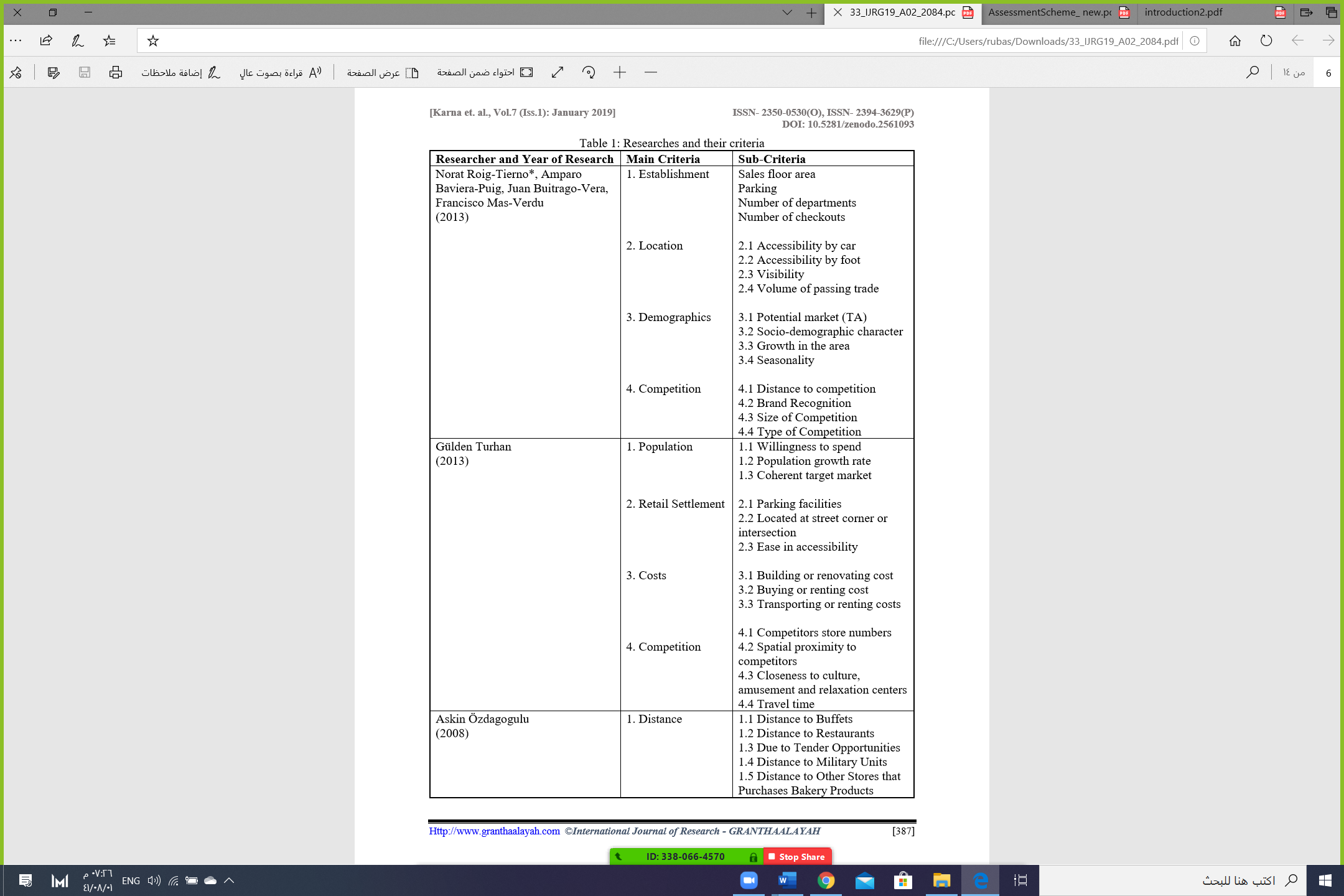
**A) Research Question -**How do different factors in combination affect the selection of location by retail chain (Big Mart) and what is the predominant factor?

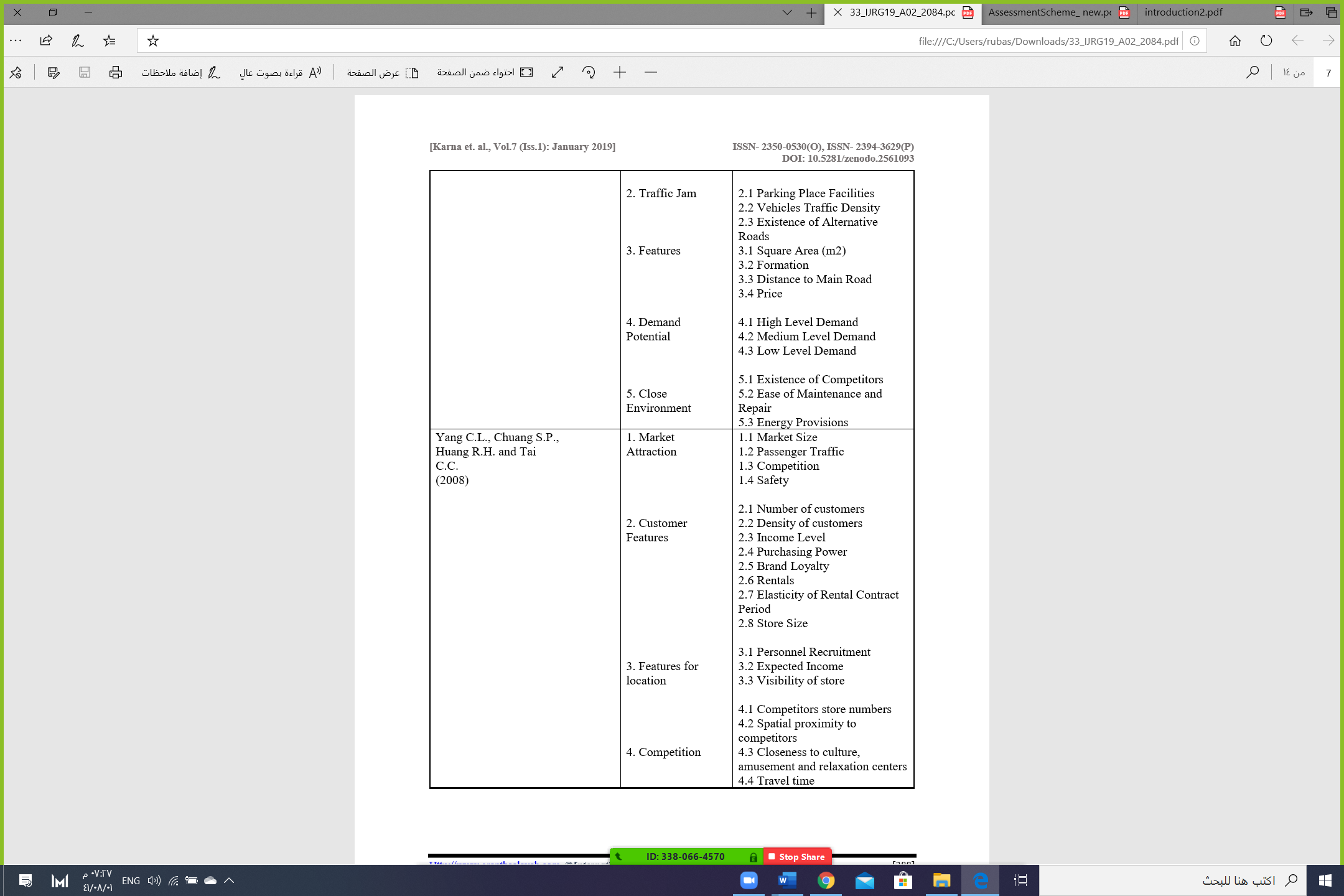
**B) Scope and Limitation-**The results are overwhelming based on observation as there was no other source of reliable data. Because of the concerned authorities not entertaining any type of questions, the results are highly subjective. Since the results are based on observations of supply chain of Big Mart, it might not be sufficient enough to explain the location selection criteria for other chain outlets.

**C) Literature Review** -As a well-known real estate mantra is “location, location, location” and this seems to hold particularly true for retail chains as it is cited as one of the most important variable factors affecting “the profitability and sales performance of the management”. Although there may be other factors that affect the success or failure of retail stores, the adverse effects of a poor location choice are formidable and non-removable. The selection of a new location for an existing store is a strategic decision requiring a long-term investment. Since new location is an additional cost there is significant financial burden on chains looking at expansion as well as the danger of their image getting damaged if the right decision in store location selection is not implemented (Turhan, 2013).

**D) Parameters of Location Selection-** In terms of population, the amount of money that people are willing to spend for buying retailers’ goods, population growth rate and coherent target market in terms of demographics such as gender, education, age, occupation etc are studied. Retailers desire to reach people who are willing to spend money for buying goods. Since the rate of retail expenditures per capita are expected to increase by a rise in population density, the population growth is a good indicator. In addition, the coherent target market is a key to eventual success of any one location as its target market is hugely defined by where it is established.

Table 1: Researches and their criteria





**E) Methodology -**For this research the Analytical Hierarchy Process (AHP) has been used. This method is used for solving multi-criteria decision problems and can effectively handle both qualitative and quantitative data. The primary purpose of the research is to understand the most suitable retail location selection for Big Mart by analysis of the various criteria and attributes involved which has been made easy by the AHP method. It is necessary to compare various factors in the location selection problem. It has been possible to make comparisons and calculations of predetermined criteria. Different criteria and its attributes are arranged in a hierarchal structure. Thus, designed hierarchal structures are compared with each other in sequential order. All factors are divided into four major criteria with the help of literature study i.e. Demography, Competition, Economic factors and Features. Further these criteria are divided into three sub –criteria. These four major criteria is compared with each other to obtain a weighted values for each. Similar method is followed for sub-criteria. Relevant comparisons and results that have been obtained with the help of literature review have been prepared.

**F) Analysis and Finding -**As discussed in the previous section of this paper the comparison and calculations have been done using AHP by forming hierarchy for main criteria and attributes in sequential order.

1) Demography

-Income of people.

-Growth Rate of people.

-Target Population.

2) Competition

-Distance (Physical and communication).

-Brand recognition.

-Size and type of competition.

3) Economic Factors

-Building and renovating cost.

-Buying or renting cost.

-Transportation.

4) Features

-Accessibility & visibility.

-Traffic.

-Services.

**[2.2] A FORECAST FOR BID MART SALES BASED ON RANDOM FORESTS AND MULTIPLE LINEAR REGRESSION**

(IJEDR 2018 | Volume 6, Issue 4 | ISSN: 2321-9939)

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.

**A) Literature Survey -**The method for long term electric power forecasting using long term annual growth factors was proposed. Prediction and analysis of aero-material consumption based on multivariate linear regression model was proposed by collecting the data of basic monitoring indicators of aircraft tire consumption from 2001 to 2016.

**B) Proposed System-**We propose below methodology for solving the problem. Raw data collected at big mart would be pre-processed for missing data, anomalies and outliers. Then an algorithm would be trained on this data to create a model. This model would be used for forecasting the final results.

BigMart’s data scientists collected sales data for the year 2013 of 1559 products across 10 stores in different cities. Also, they provided definitions for certain attributes of each product and store. They are as follows-:

* Item\_Identifier - Unique identifier for each product.
* Item\_Weight – Product weight.
* Item\_Fat\_Content – Fat content of the product.
* Item\_Visibility – Percentage of total display area in a store allocated to the product.
* Item\_Type – Product category.
* Item\_MRP – List price of the product.
* Outlet\_Identifier - Unique identifier for each store.
* Outlet\_Establishment\_Year – Establishment year for each store.
* Outlet\_Size - The size of the store.
* Outlet\_Location\_Type - The type of city in which the store is located.
* Outlet\_Type - Whether the store is a grocery store or a supermarket.
* Item\_Outlet\_Sales - Sales of the product in each store.

**C) Multiple Linear Regression-** Multiple linear regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). It is represented by an equation.

Y=a+b\*X + e

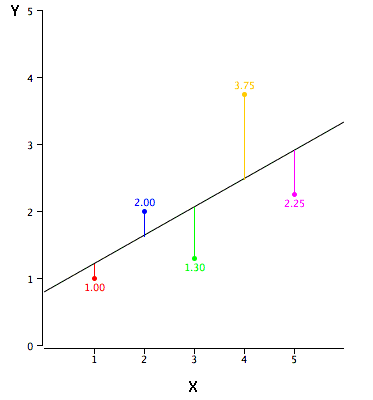


Fig.2 Linear regression

Where a is intercept, b is slope of the line and e is error term. Using this method, an accuracy can be found out. Multiple linear regression is very famous method for prediction and analysis but one drawback is it gives less accuracy.

**[2.3] A COMPARATIVE STUDY OF BIG MART SALES PREDICTION**

(Conference Paper · September 2019)

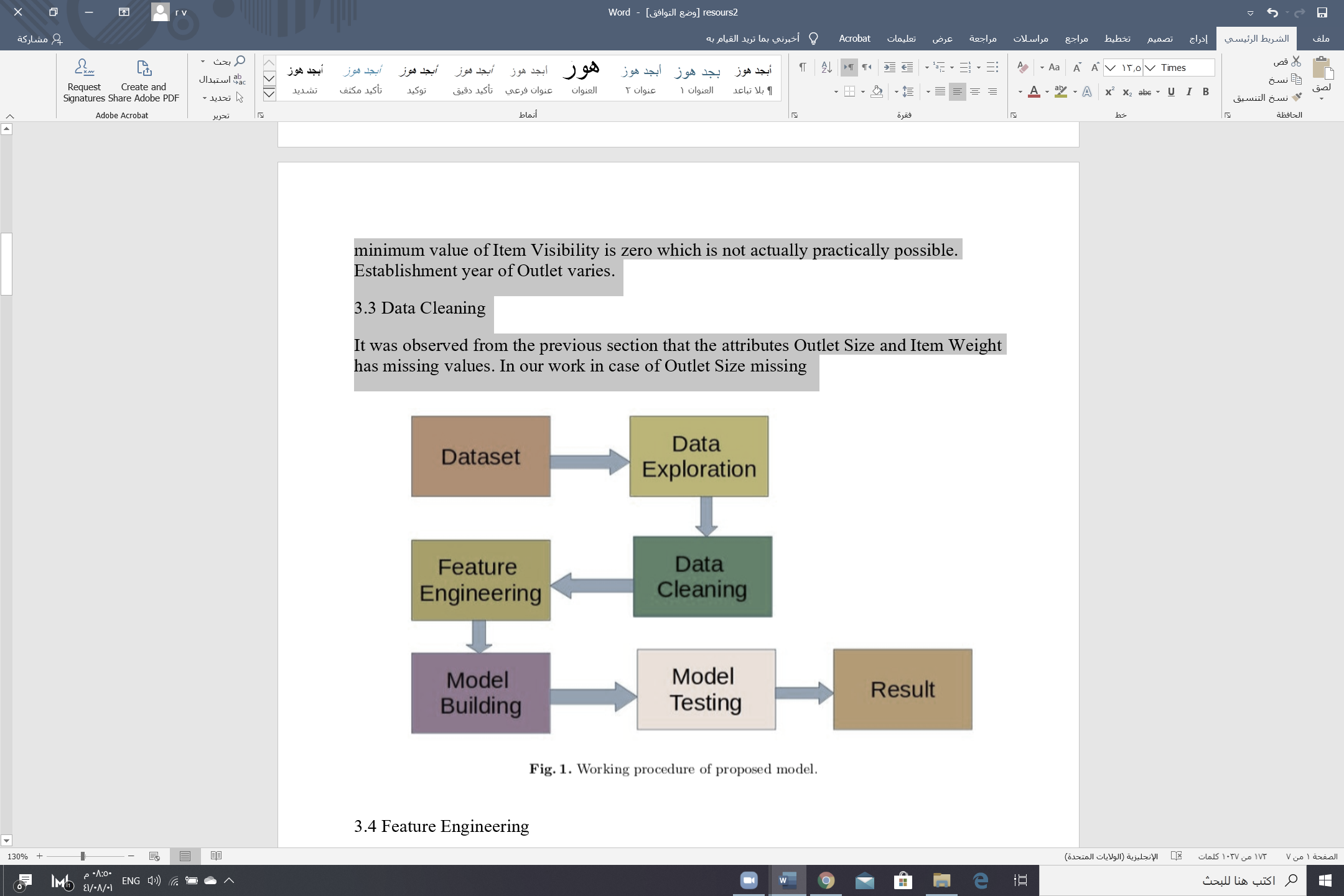
Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personal- ized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service. Sales forecasting as well as analysis of sale forecasting has been conducted by many authors as summarized: The statistical and computational methods are studied in also this paper elaborates the automated process of knowledge acquisition. Machine learning is the process where a machine will learn from data in the form of statistically or computationally method and process know- edge acquisition from experiences. Various machine learning (ML) techniques with their applications in different sectors has been presented in.

**A) Proposed System**-For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequences of steps as mentioned in Figure 1 and in this work, we propose a model using Xgboost technique.

**- Dataset Description of Big Mart -** Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales.

**- Data Exploration -** In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. Establishment year of Outlet varies.

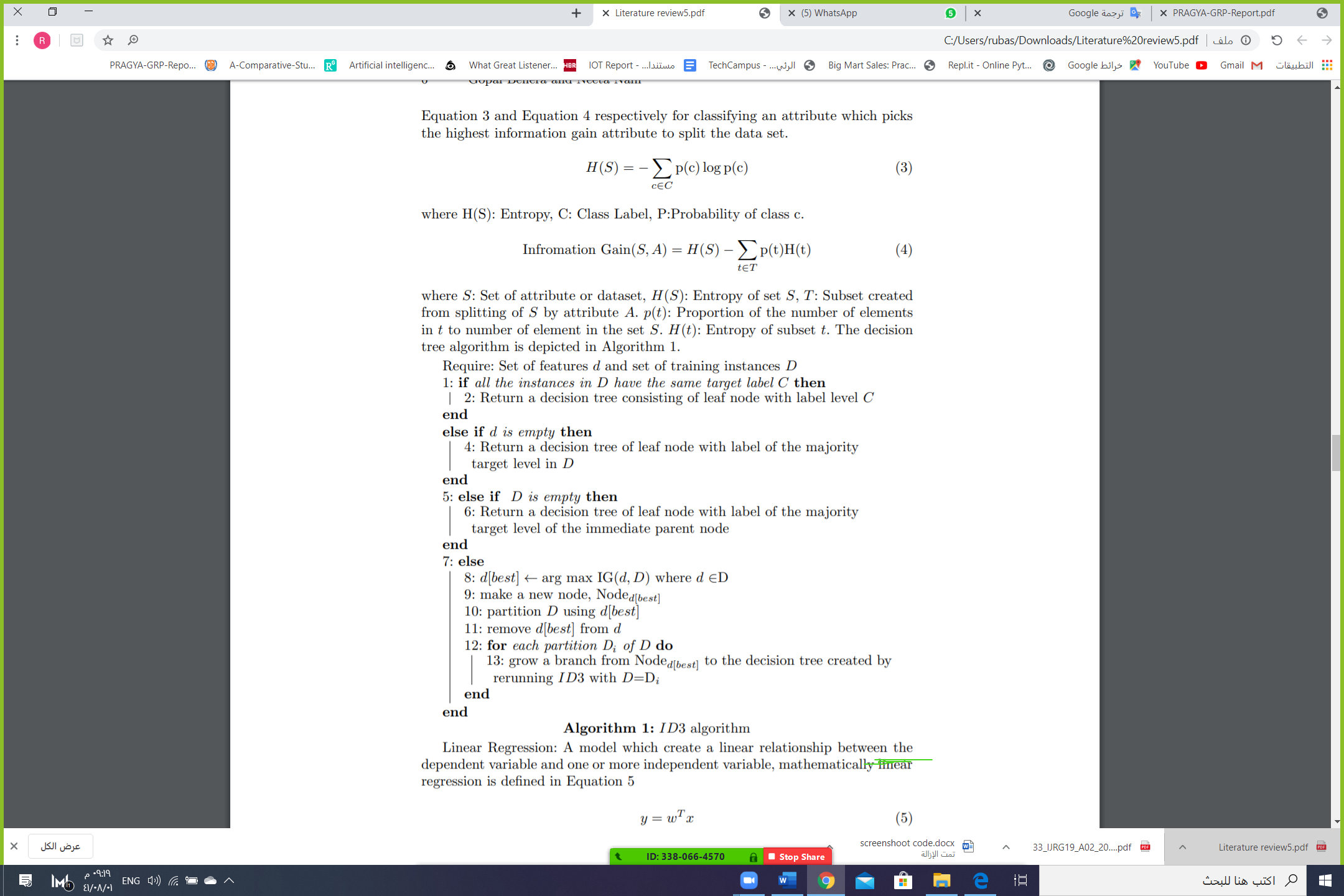
**- Data Cleaning -** It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing.



**- Feature Engineering -** Some nuances were observed in the data-set during data exploration phase. So, this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So, the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell.

**B) Model Building-**After completing the previous phases, the dataset is now ready to build proposed model. Once the model is built it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model using Xgboost algorithm and compare it with other machine learning techniques like Linear regression, Ridge regression, Decision tree.

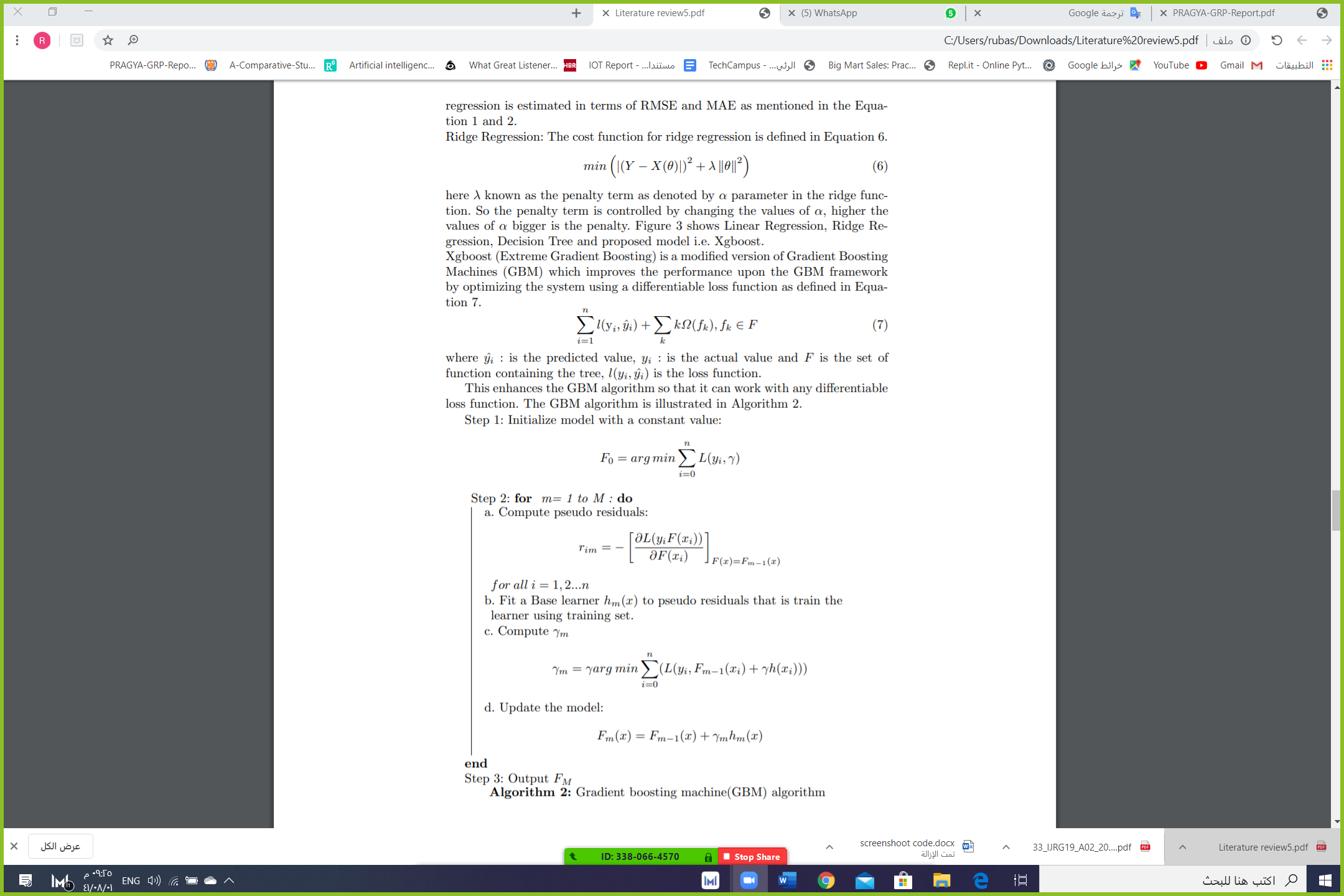
**Decision Tree:** A decision tree classification is used in binary classification problem and it uses entropy and information gain as metric and is defined in, Equation 3 and Equation 4 respectively for classifying an attribute which picks the highest information gain attribute to split the data set.



**Linear Regression:** A model which create a linear relationship between the dependent variable and one or more independent variable, mathematically linear regression is defined in Equation 5

**y = (**5)

where y is dependent variable and x are independent variables or attributes. In linear regression we find the value of optimal hyperplane w which corresponds to the best fitting line (trend) with minimum error. The loss function for linear regression is estimated in terms of RMSE and MAE.

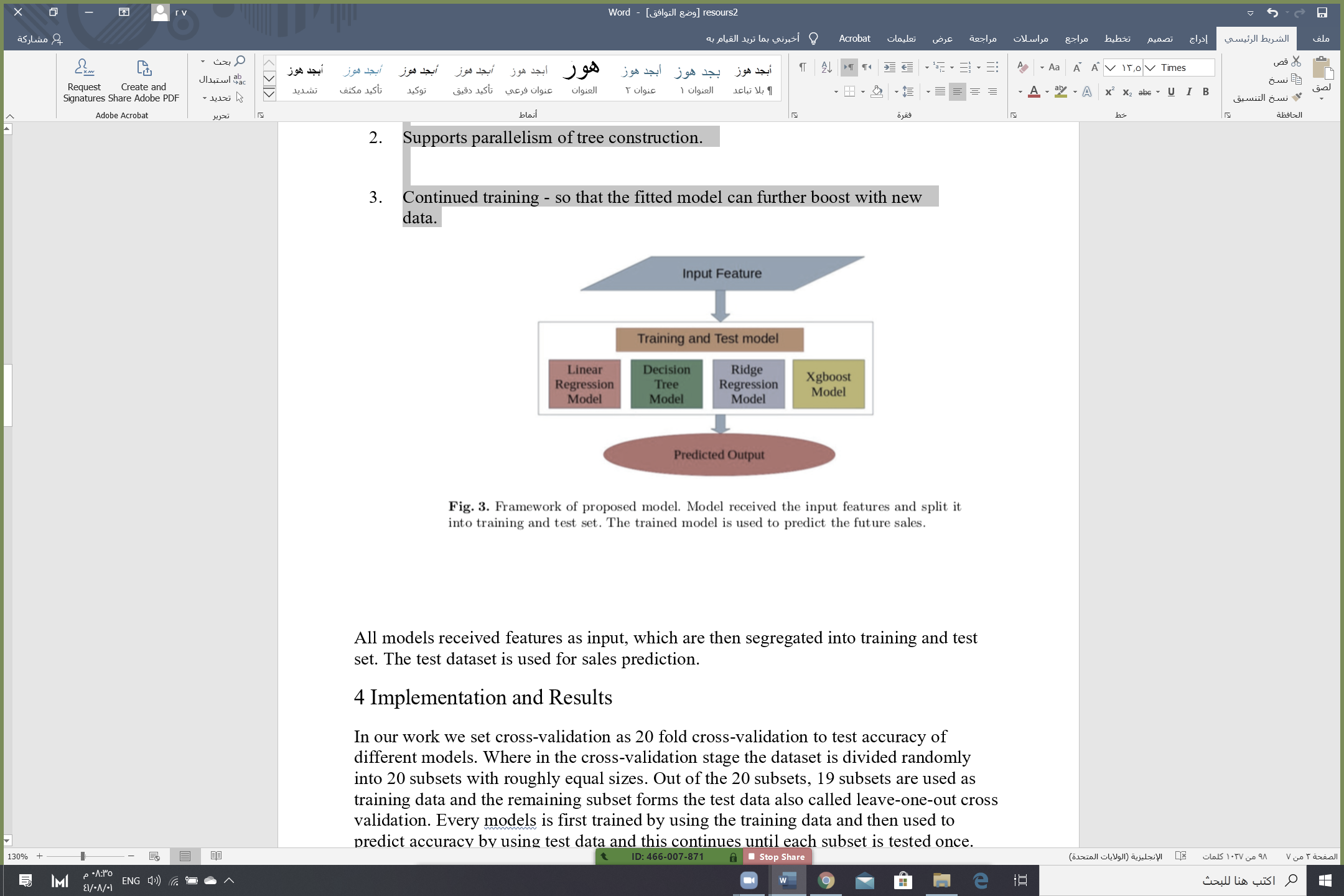


The Xgboost has following exclusive features:

- Sparse Aware - that is the missing data values are automatic handled.

- Supports parallelism of tree construction.

- Continued training - so that the fitted model can further boost with new data.



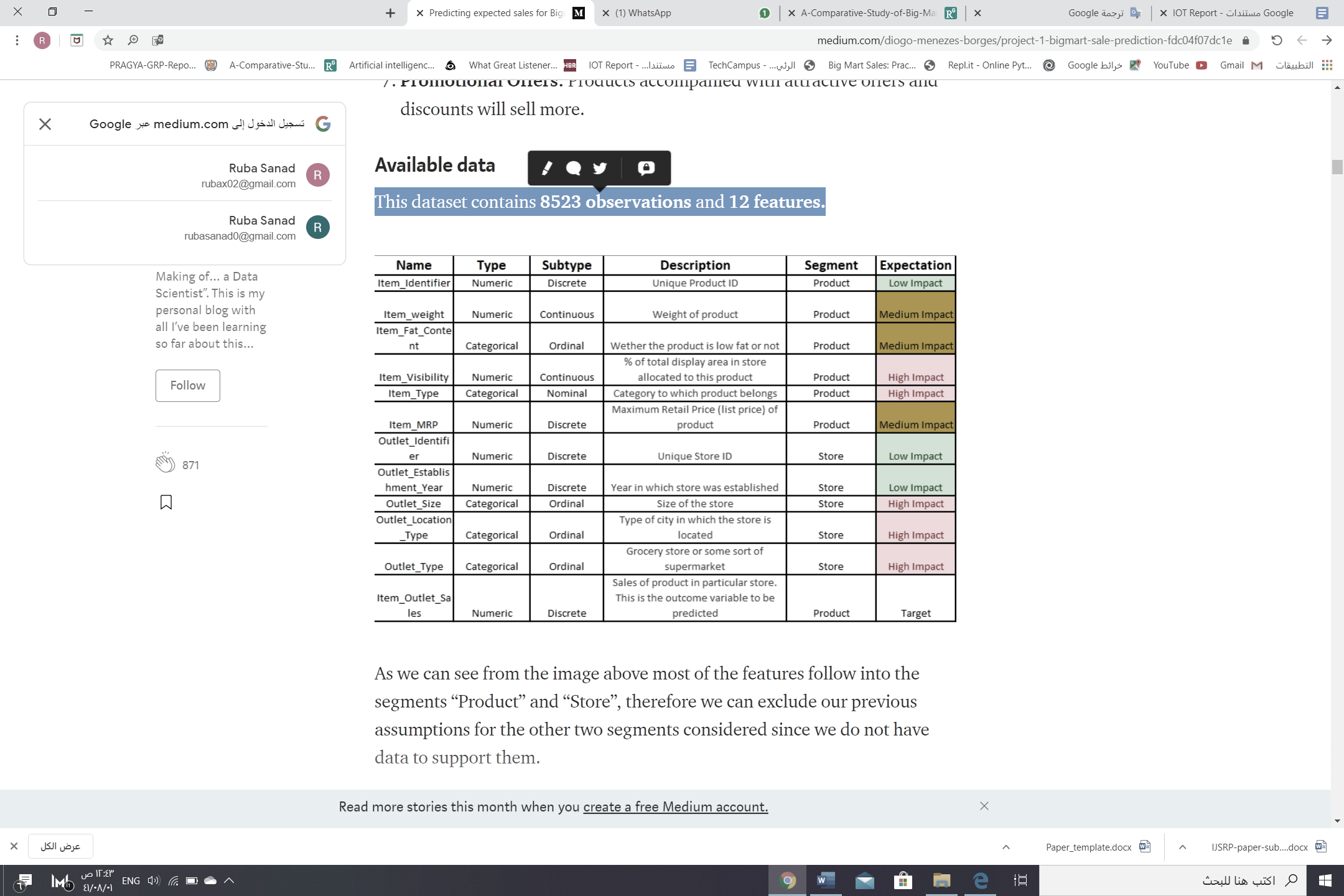
All models received features as input, which are then segregated into training and test set. The test dataset is used for sales prediction.

1. **DATA PREPARATION**

**3.1 Dataset Description of Big Mart**

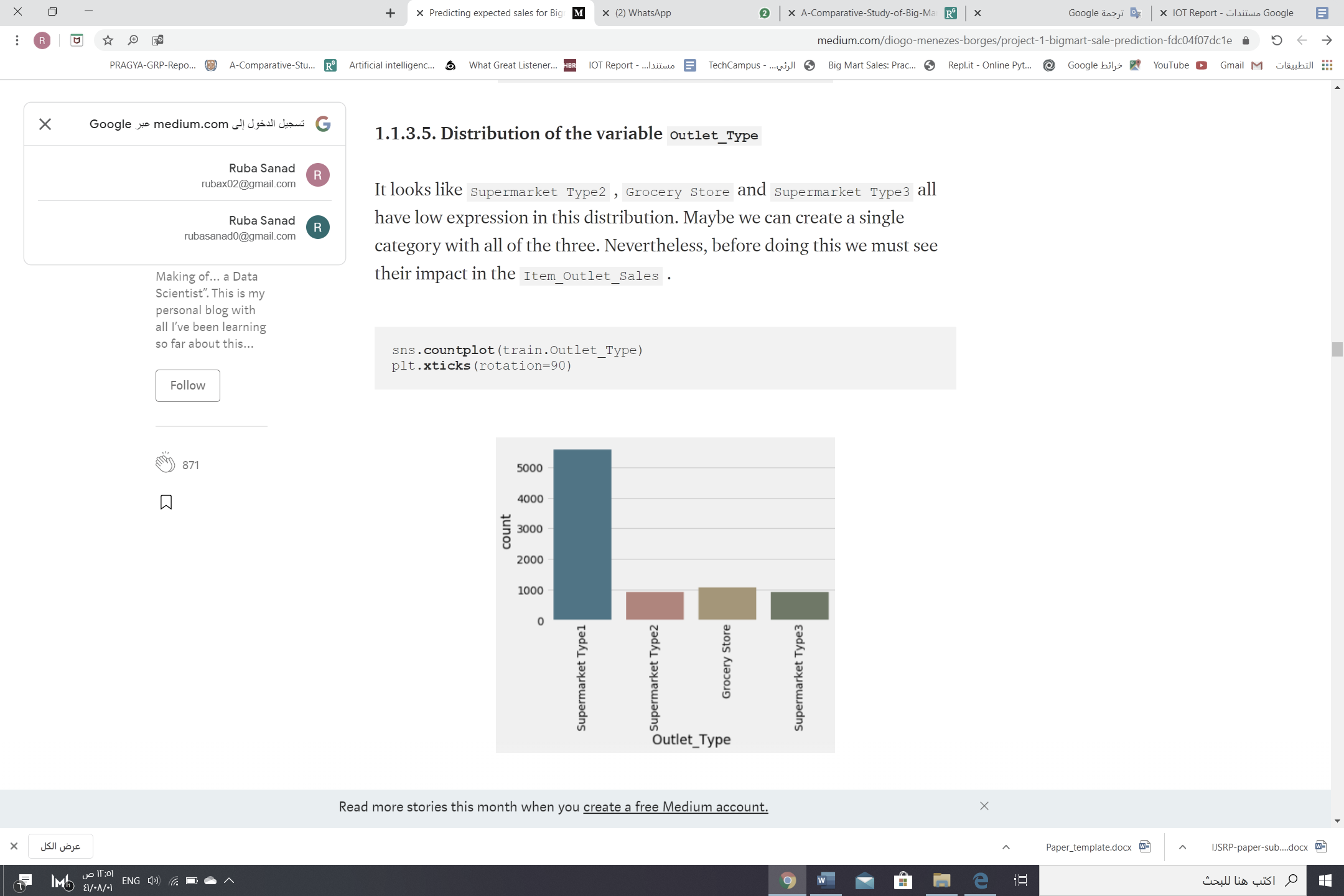
## In our work we have used 2013 Sales data of Big Mart as the dataset. Where the dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales. Out of these attributes' response variable is the Item Outlet Sales attribute and remaining attributes are used as the predictor variables. The data-set consists of 8523 products across different cities and locations. The data-set is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store capacity, location, etc, and the product level.

This dataset contains **8523 observations**and **12 features.**

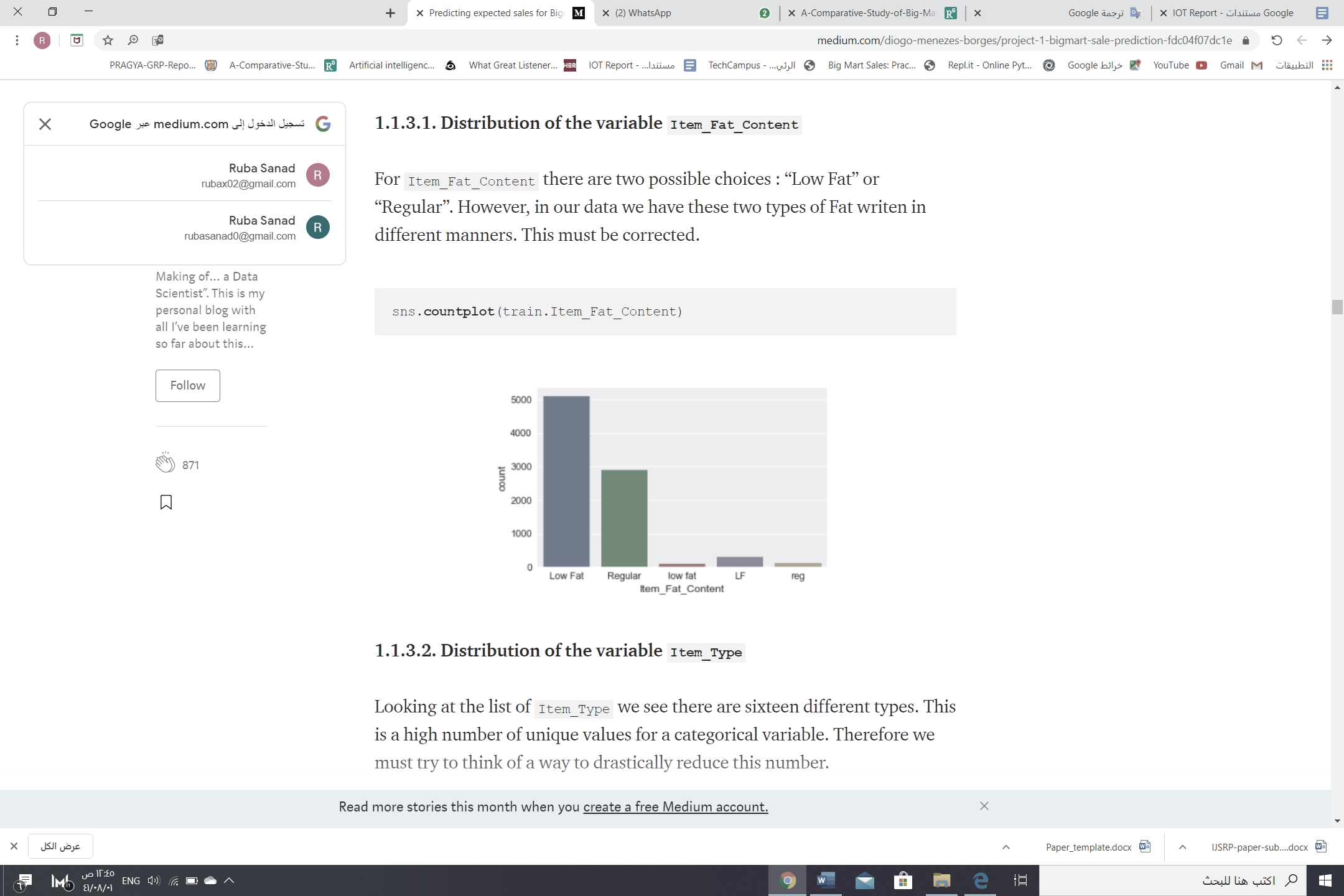


**(Graphical presentation of some attributes from dataset)**

(Distribution of the variable Outlet\_Type)



(Distribution of the variable Item\_Fat\_Content)



(Distribution of the variable Outlet\_Location\_Type)



(Distribution of the variable Item\_Type)



## **3.2 Data preprocessing**

***Data cleaning***- It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing, Data cleaning can be applied to remove noise and correct inconsistencies in data.

*1) If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied.*

*2) Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust.*

*3) Therefore, a useful preprocessing step is to run your data through some data cleaning routines.*

*4) Real-world data tend to be incomplete, noisy, and inconsistent.*

***Data cleaning routines attempt to:***

*1) Fill in missing values*

*2) Smooth out noise*

*3) Identifying outliers*

*4) Correct inconsistencies in the data.*

***-MISSING VALUES***

a) Use global constant to fill the missing value:

Replace all missing attribute values by the same constant such as a label like “Unknown”,” none” or “null”.

|  |  |  |  |
| --- | --- | --- | --- |
| **Item\_Identifier** | **Item\_Weight** | **Item\_Fat\_Content** | **Item\_Type** |
| fdw58 | 20.75 | Low fat | Snack Foods |
| fdw14 | 8.3 | Regular | Dairy |
| ncn55 | 14.6 | Low fat | Fruits and Vegetables |
| ncc54 | 24.5 | Hight fat | Health and Hygiene |
| NULL | 16.1 | NULL | Drinks |
| ncn15 | NULL | NULL | NULL |

b) Use the attribute mean or median for all samples belonging to the same class as the given tuple.

**(24.5+16.1)/2 =20.3**

|  |  |  |  |
| --- | --- | --- | --- |
| **Item\_Identifier** | **Item\_Weight** | **Item\_Fat\_Content** | **Item\_Type** |
| fdw58 | 20.75 | Low fat | Snack Foods |
| fdw14 | 8.3 | Regular | Dairy |
| ncn55 | 14.6 | Low fat | Fruits and Vegetables |
| ncc54 | 24.5 | Hight fat | Health and Hygiene |
| NULL | 16.1 | Regular | Drinks |
| ncn15 | **20.3** | Hight fat | Breads |

In some cases, a missing value may not imply an error in

the data.

◼ For example, in supermarket, some producer like toy or frish bread those who don’t have one leave this field in item\_fat\_content

◼ Forms should allow respondents to specify values such as “not applicable.”, “don’t know,” “?” or “none”

◼ Fields may also be intentionally left blank if they are to be provided in a later step.

◼ The design of a good database and data entry procedure should help minimize the number of missing values or

errors in the first place.

1. **MODEL PLANNING**

Machine learning is a large field of study that overlaps with and inherits ideas from many related fields, Types of Learning:

1. Supervised Learning
2. Unsupervised Learning

Supervised learning describes a class of problem that involves using a model to learn a mapping between input examples and the target variable, Models are fit on training data comprised of inputs and outputs and used to make predictions on test sets

where only the inputs are provided and the outputs from the model are compared to the withheld target variables and used to estimate the skill of the model.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.

**Classification:** Supervised learning problem that involves predicting a class label.

**Regression:** Supervised learning problem that involves predicting a numerical label.

Some algorithms may be specifically designed for classification (such as logistic regression) or regression (such as linear regression) and some may be used for both types of problems with minor modifications (such as artificial neural networks).

Why we use Regression Analysis?

The importance of regression analysis is that it is all about data: data means numbers and figures that actually define your business. The advantages of regression analysis is that it can allow you to essentially crunch the

numbers to help you make better decisions for your business currently and into the future. The regression method of forecasting means studying the relationships between data points, which can help you to:

Predict sales in the near and long term.

Understand inventory levels.

Understand supply and demand.

Review and understand how different variables impact all of these things.

**Regression model:** is a predictive modelling technique. It estimates the relationship between a dependent (target) and an independent variable(predictor). It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used. Three major uses for regression analysis are-Determining the strength of predictors, Forecasting an effect, Trend for forecasting.

**Comparison between Linear and Logistic Regression:**

|  |  |  |
| --- | --- | --- |
| **LOGISTIC REGRESSION** | **LINEAR REGRESSION** | |
| Logistic Regression is categorical variables. | Linear Regression is continuous variables. | |
| In Logistic Regression, we predict the value by 1 or 0. | In Linear Regression, we predict the value by an integer number. | |
| Here solve Classification problems. | Here solve Regression problems. | |
| Here we use precision to predict the ne | Here we calculate Root Mean Square Error (RMSE) to predict the next weight value. |

**The linear regression model can be represented by the following equation:**

**The linear regression model equation **where,

Y is the predicted value

θ₀ is the bias term.

θ₁,…,θn are the model parameters

x₁, x₂,…,xn are the feature values.

The above hypothesis can also be represented by

**The above hypothesis**

Where, θ is the model’s parameter vector including the bias term θ₀; x is the feature vector with x₀ =1

**Y (pred) = b0 + b1\*x**

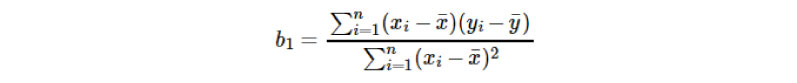
The values b0 and b1 must be chosen so that the error is minimum. If sum of squared error is taken as a metric to evaluate the model, then the goal is to obtain a line that best reduces the error.

**Error Calculation in Linear Regression**

If we don’t square the error, then the positive and negative points will cancel each other out.

For a model with one predictor,

Intercept Calculation in Linear Regression



**Exploring ‘b1’** - If b1 > 0, then x (predictor) and y(target) have a positive relationship. That is an increase in x will increase y.

If b1 < 0, then x (predictor) and y(target) have a negative relationship. That is an increase in x will decrease y.

**Exploring ‘b0’** - If the model does not include x=0, then the prediction will become meaningless with only b0. For example, we have a dataset that relates height(x) and weight(y). Taking x=0 (that is height as 0), will make the equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.

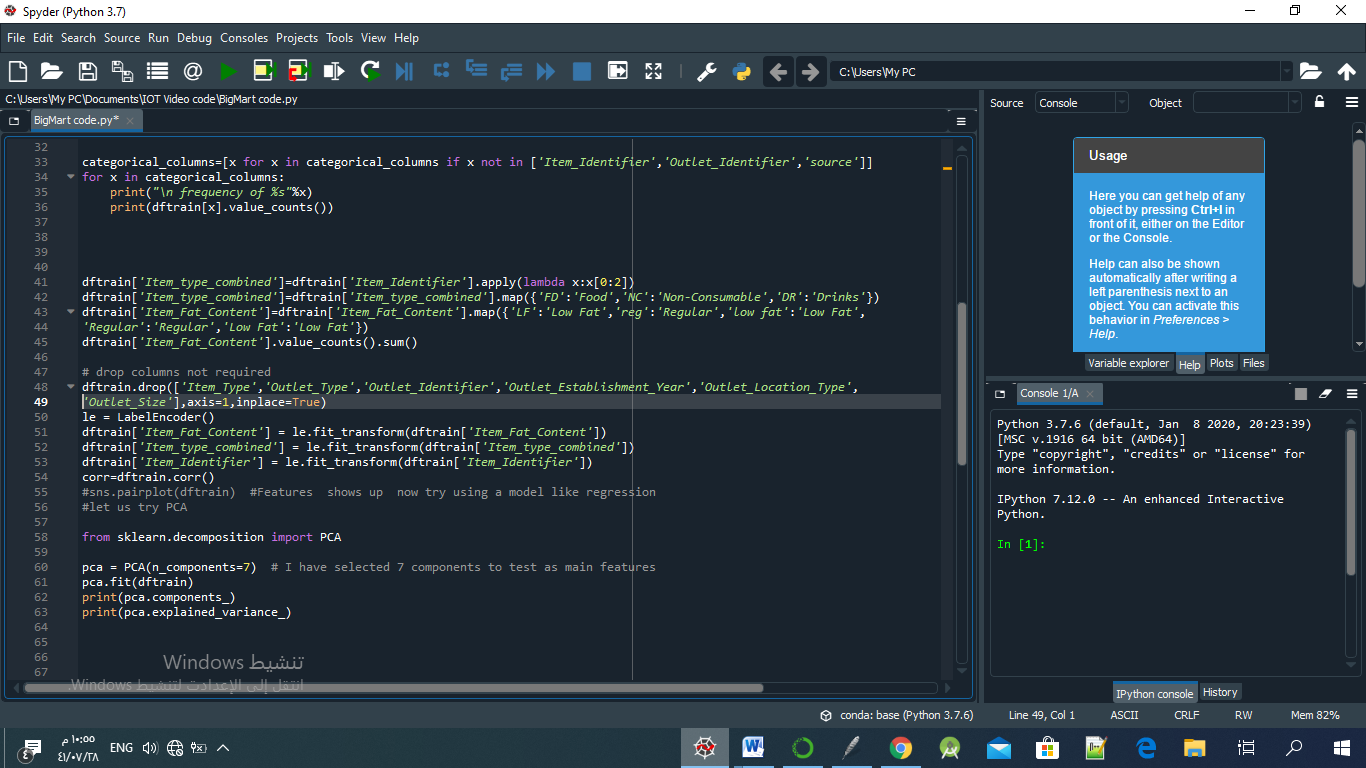
If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

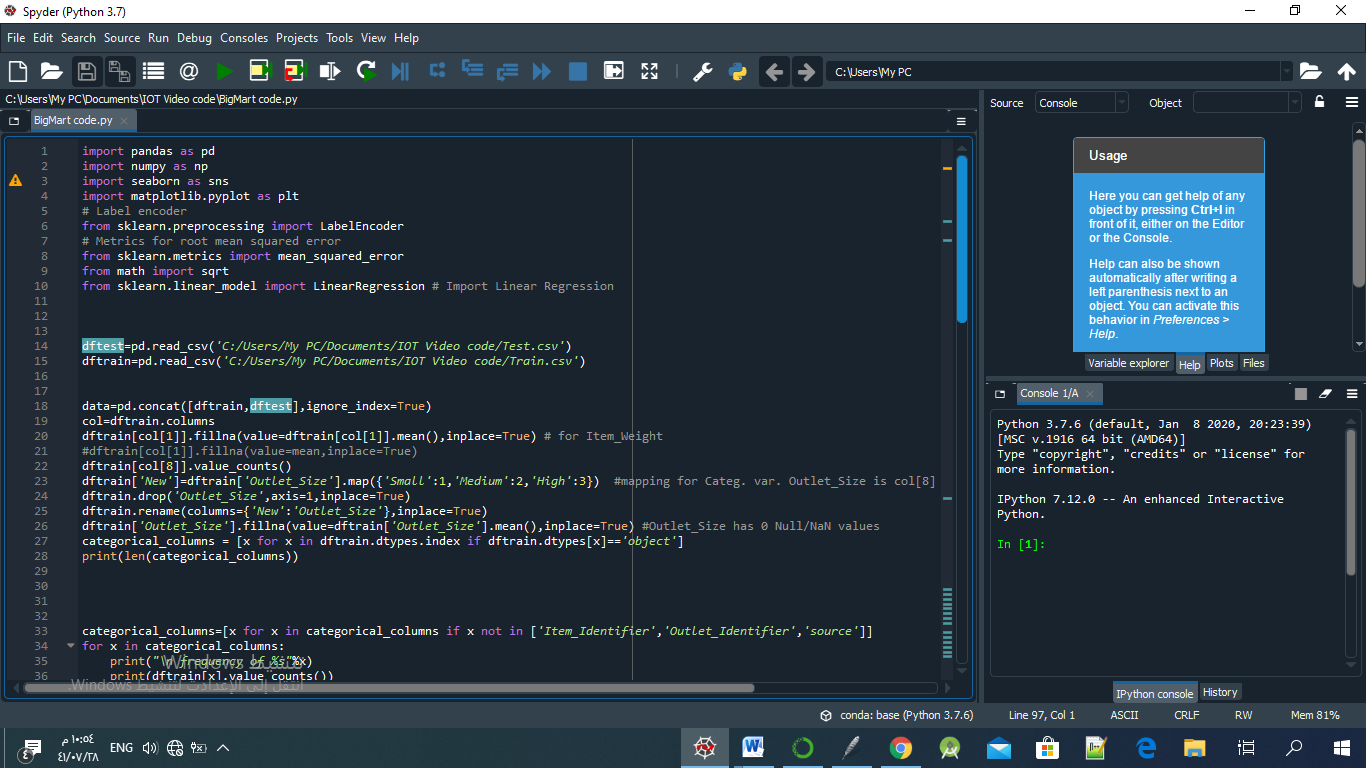
The value of b0 guarantees that the residual will have mean zero. If there is no ‘b0’ term, then the regression will be forced to pass over the origin. Both the regression coefficient and prediction will be biased.

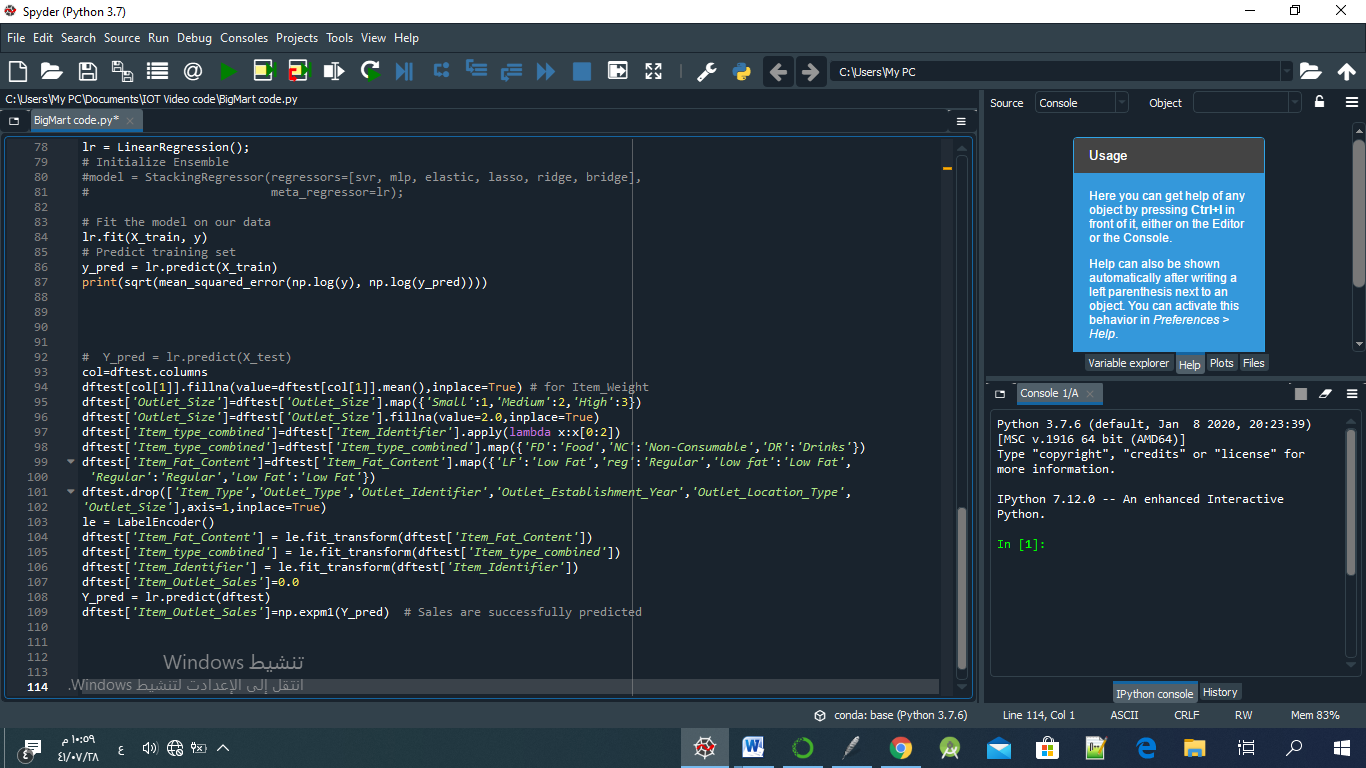
**v. MODEL BUILDING**

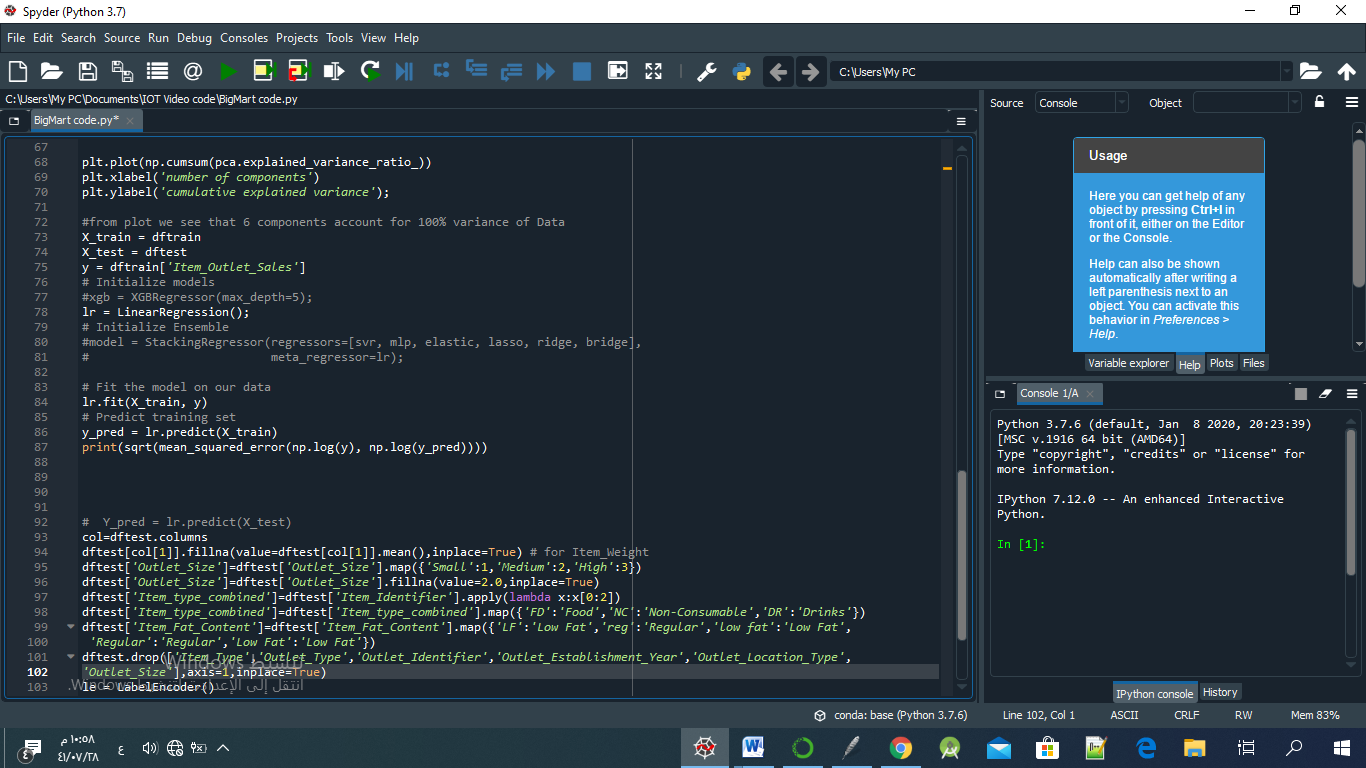
(implementation)

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is built it is used as predictive model to forecast sales of Big Mart.Here we will import the Linear Regression. We chose, because want to predict the sales of BigMart and it based on supervised learning that mean: helps you to predict outcomes for unforeseen data.



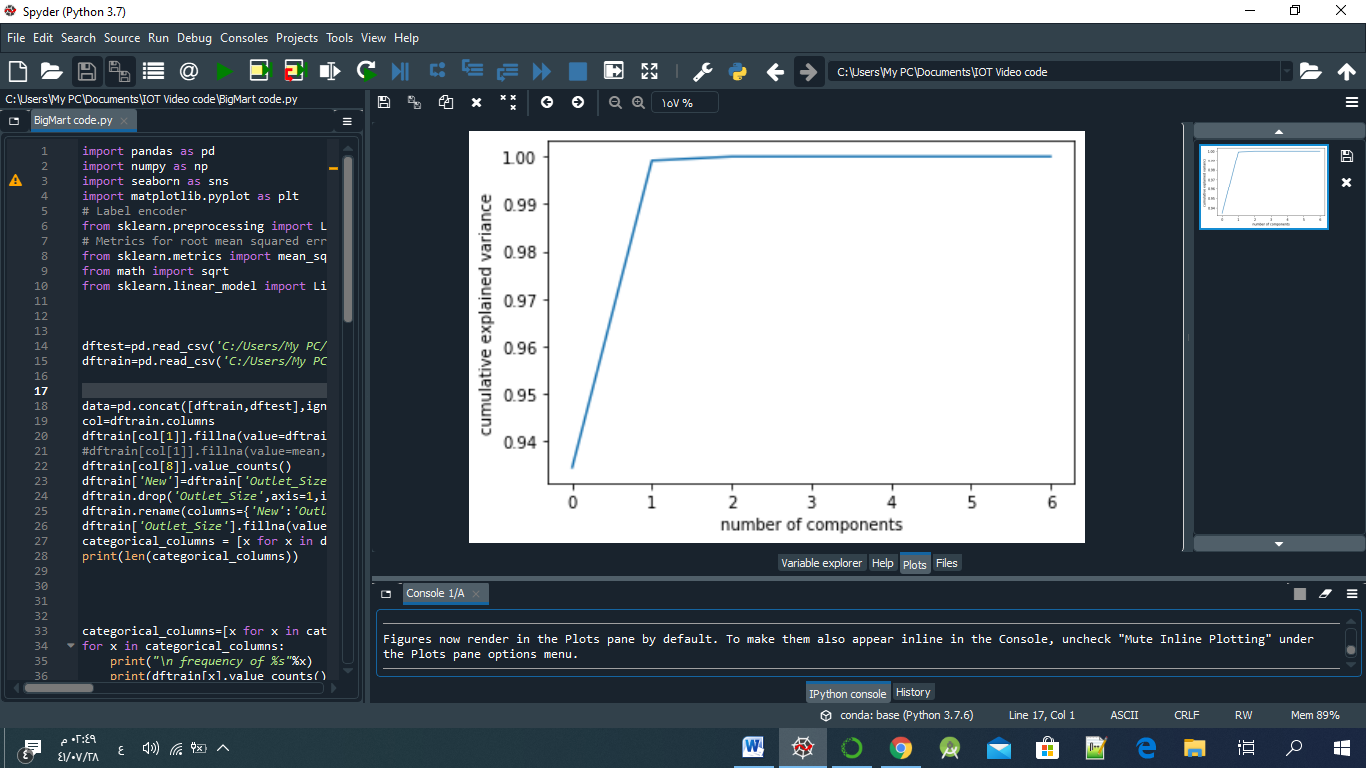
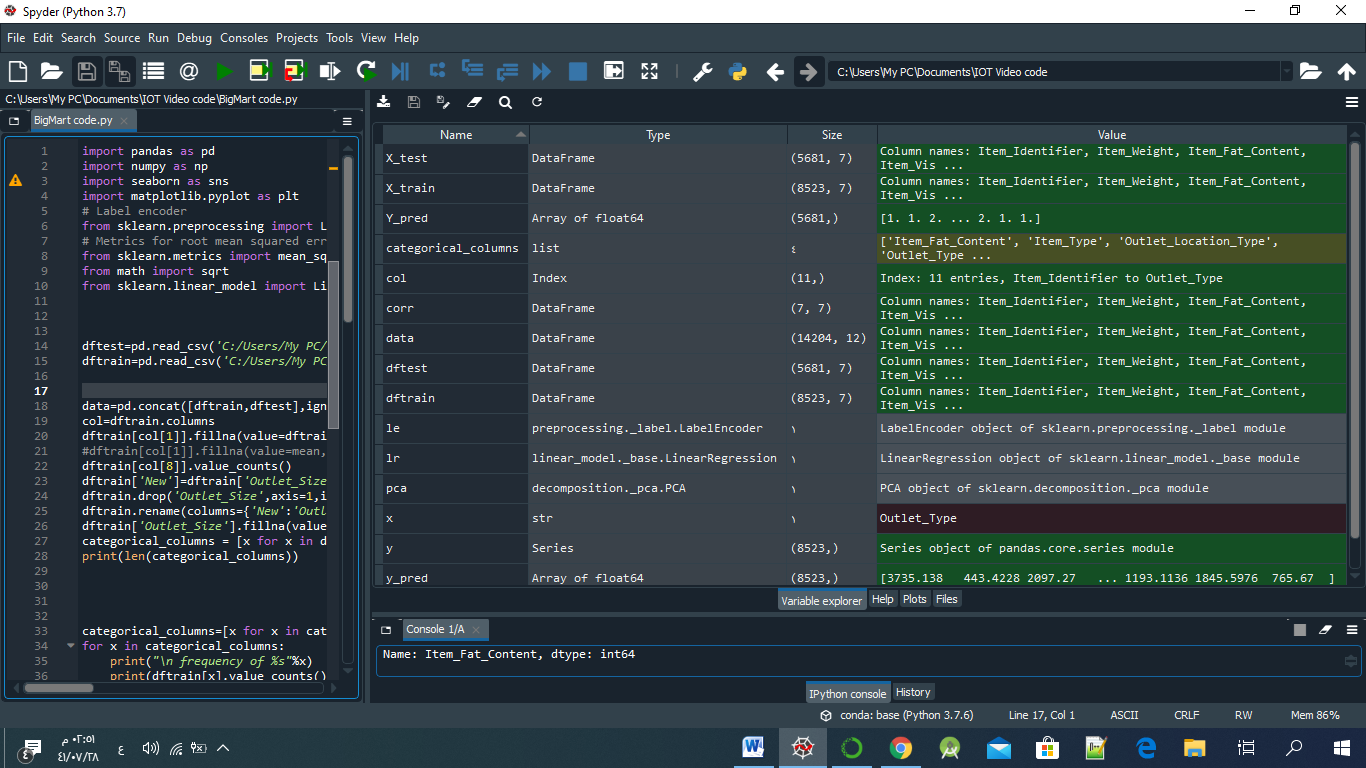


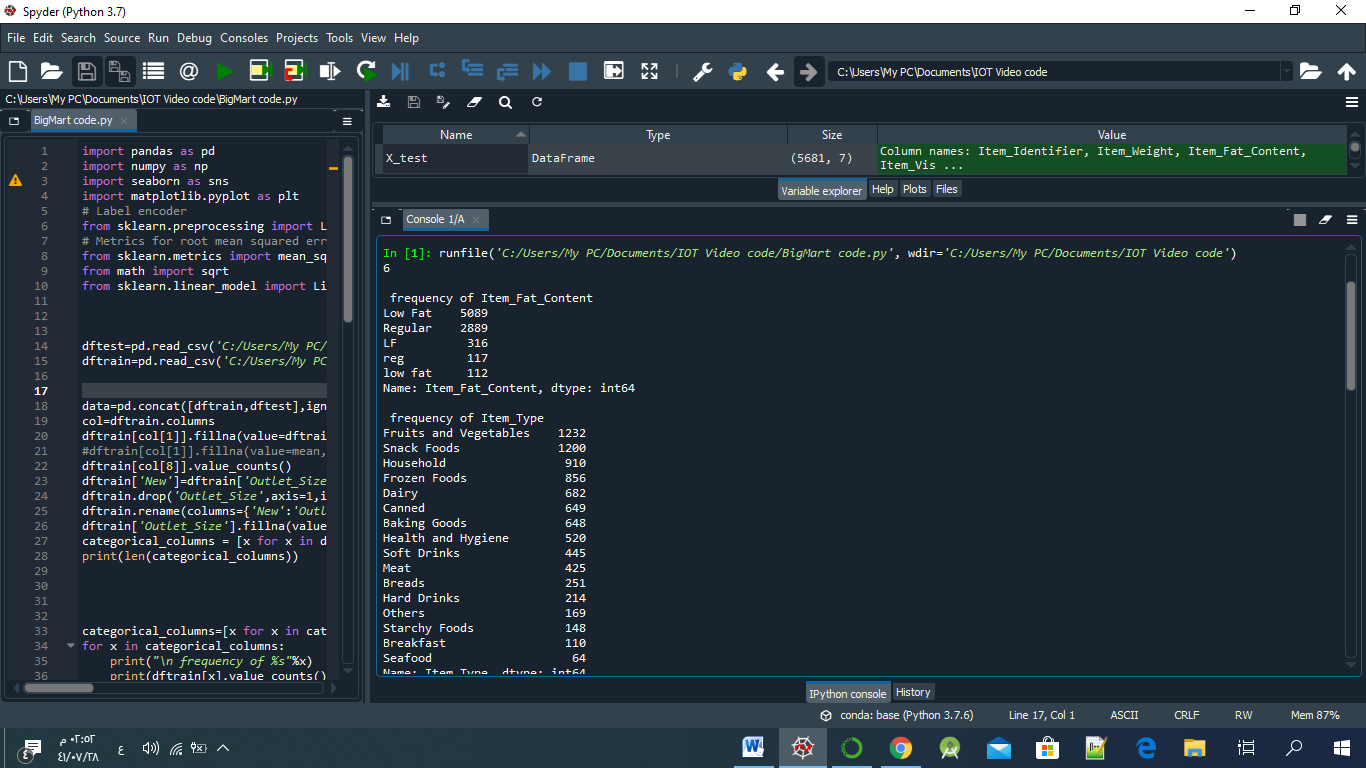
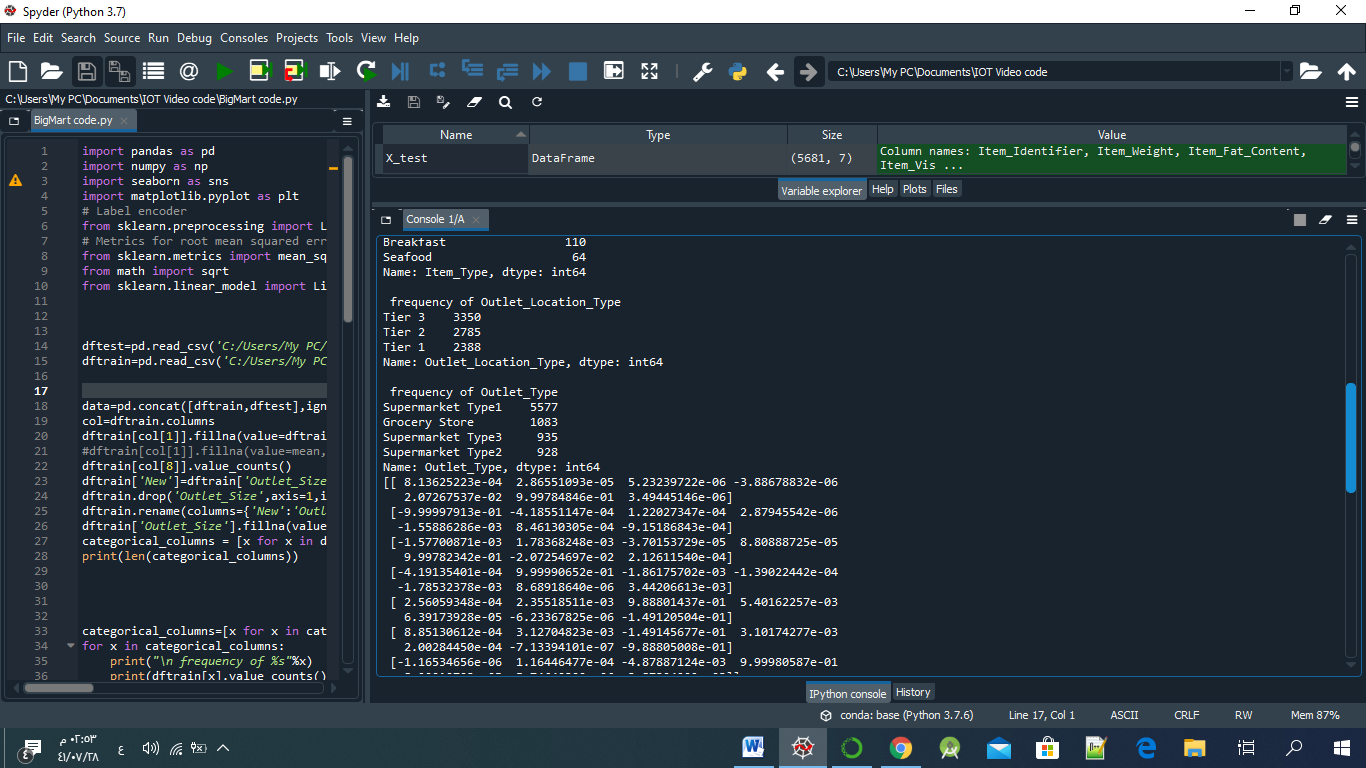


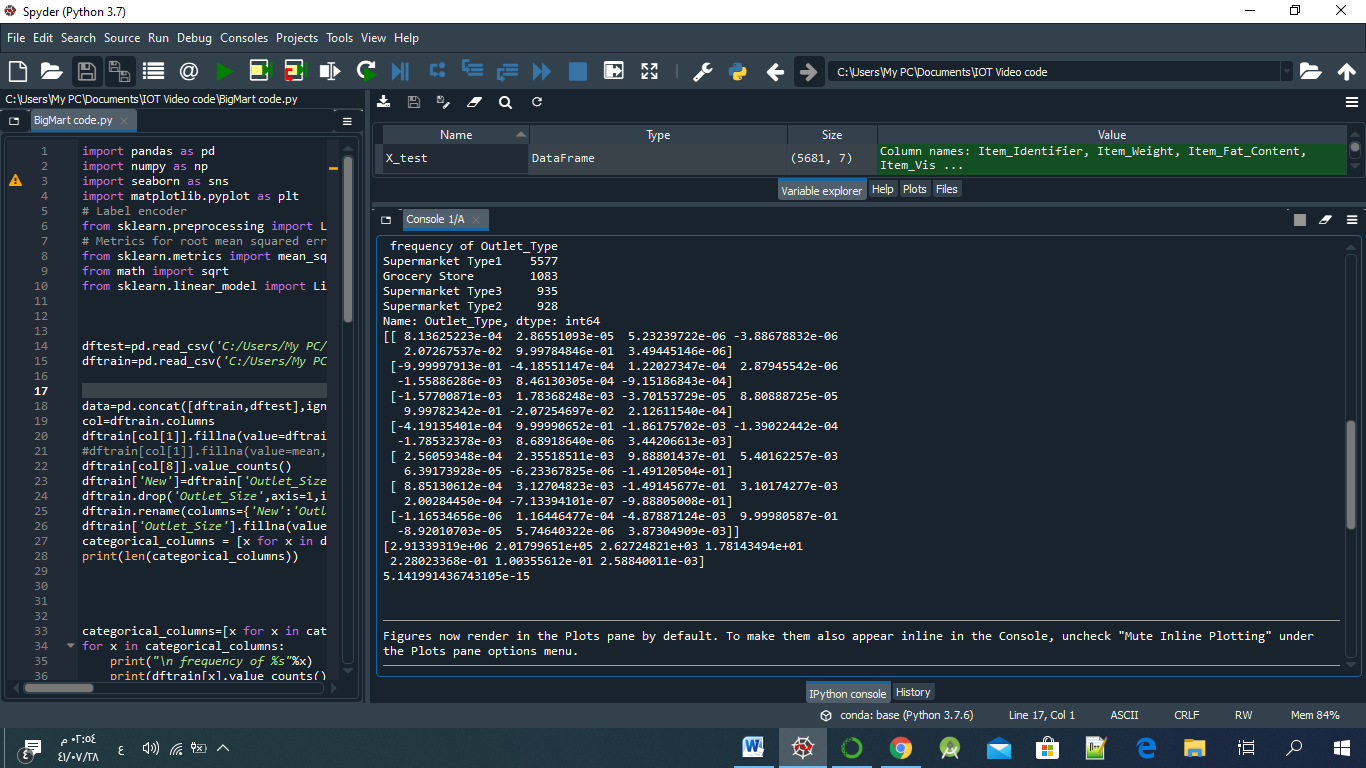


**(Here to out the prediction using RMSE)**

* **OUTPUT:**





**VI**. **MODEL Evaluation**

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results, we applied the RMSE in above code.

**VII.Conclusion**

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items. This helped from corporate profits and higher economic income. Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Extensive research in this area at enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses. In this paper, we use: linear regressions experimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet. Finally, experiments that our technique produce more accurate prediction compared to than other available techniques.

**References**

1. https://www.researchgate.net/publication/330994981\_A\_STUDY\_ON\_SELECTION\_OF \_LOCATION\_BY\_RETAIL\_CHAIN\_BIG\_MART
2. <https://www.ijedr.org/papers/IJEDR1804010.pdf>
3. https://www.researchgate.net/publication/336530068\_A\_Comparative\_Study\_of\_Big \_Mart\_Sales\_Prediction

1. <https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning>
2. <https://libguides.usc.edu/writingguide/literaturereview>
3. <https://www.statisticshowto.datasciencecentral.com/rmse/>
4. <https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/?ref=rp>
5. https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-prediction

**Names of Students**

* Ghada Aldabayan , 371201914
* Khuzama Alsalem , 362206020
* Ruba Sanad Alharbi , 362216679

**Under the supervision of -** Dr. Amal Alshargabi