**Sales Prediction Concept.**

**Dataset.**

**Dataset Background and Source.**

The dataset used was put together by a group of data scientists from **Bigmart** in the year 2013. It contains records of sales data for 1559 products from 10 over BigMart stores located in different cities.

The dataset downloaded contains structured data and was downloaded from the website, **AnalyticsVidhya.com**.

**Insights and Objectives.**

Basing on the available attributes(in the dataset), which include a set of properties of the products and of the stores in which they are being sold, an analysis is to be done so as to identify which of these properties affect the amount of sales of a specific product more. This will enable the manipulation of these properties so as to increase on the sales of a given product.

To achieve this goal, we will build a predictive model and find out the sales of each product at a particular store.

Below is a list and description of the each of the twelve attributes of the dataset to be used during the analysis of sales.

1. Item\_Identifier: This is a unique identifier for each product sold.
2. Item\_Weight: This is the weight of a given product.
3. Item\_Fat\_Content: A categorical attribute that holds information on whether a product contains either a “low” or “regular” fat composition. It has only two values, which are, “Low Fat” and “Regular”.
4. Item\_Visibility: This contains the total amount of visibility that a product is given on the shelves of the store in which it is being sold expressed as a percentage.
5. Item\_Type: This contains information on which food category a specific product falls.
6. Item\_MRP: This contains the maximum retail price of a specific product in a store.
7. Outlet\_Identifier: This is a unique identifier for each store in which products are sold.
8. Outlet\_Establishment\_Year: This is the year in which the specific store was established.
9. Outlet\_Size: This is the size of a specific store in terms of the area it covers.
10. Outlet\_Location\_Type: A categorical attribute which holds information about the type of city in which a store is located. It has three values, which are, “Tier 1”, “Tier 2” and “Tier 3”.
11. Outlet\_Type: A categorical attribute which determines which type of establishment a specific store is. It has four values, which are, “Supermarket Type1”, “Supermarket Type2”, “Supermarket Type3” and “Grocery Store”.
12. Item\_Outlet\_Sales: This attribute contains the sales of a product in a particular store. This is the value that will need to be predicted.

**Data Analytics Approach (Data Pipeline).**

The steps that are to be followed during the data analysis process are explained below:

1. **Data Collection:**

This step was already done and completed. The dataset used was downloaded as a whole from the website, **AnalyticsVidhya.com**. This dataset was put together by a group of **BigMart Data Scientists** who collected the sales data for over 1559 products from 10 BigMart stores located in different cities in the year 2013. This data was already in a structured format and, therefore, did not require to be organized in rows and columns.

1. **Data Cleaning.**

In the dataset, there is expected to be a set of errors, duplicates or incomplete areas that need to be corrected. These will be corrected in the following ways:

* In the column of **Item\_Fat\_content**, there are several values that are being used to represent one value, for example, “Low Fat” and “LF” are both used to represent **Low Fat Content**. All the values that are referring to the same thing will all be unified to be represented in the same way, for example, all products with a low fat content will have their value in the **Item\_Fat\_Content** column as **Low Fat** and the rest will have this value as **Regular**.
* The missing values in the column of **Item\_Weight** will be handled by computing the mean value of all the instances of the weight of the specific product that is missing a value and assigning the calculated mean value as the new value for the weight of the product with a missing weight value.
* The missing values in the column of **Outlet\_Size** will be handled by computing the mode of the outlet size of the outlet type to which the specific outlet which is missing a value for its size belongs. This mode value will be used replace its missing value for the **Outlet\_Size** column.
* If a product has a value of **0** (**zero**) in the column of **Item\_Visibility**, the mean value of the visibility of all the instances of that particular product will be computed. This computed mean value will be used to replace the zero value of visibility for the product.

**3. Feature Engineering.**

Because some of the relevant libraries(scikit-learn) only accept numerical values, all the nonnumerical variable, with the exception of **Item\_Identifier** and **Outlet\_Identifier** will be converted into a binary, numerical format by using a **Label Encoder.**

**4. VISUALIZATION.**

The data visualization methods that are expected to be used are mentioned below:

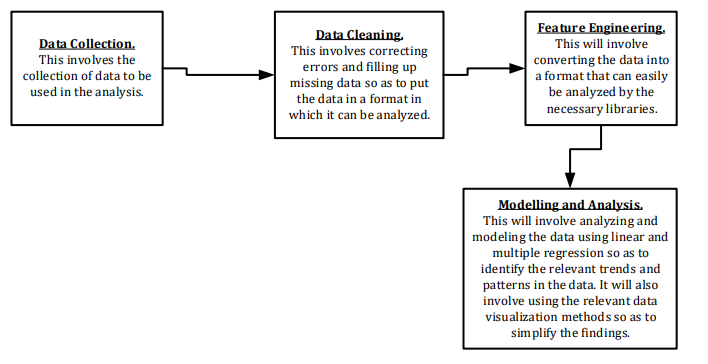
* Bar graphs.
* Scatter plots.
* Histogram

**5. Model Building.**

* Using the training and test data-set we shall use linear regression model to predict the sales for each store since its the best model for performing predictions. We shall perform the linear regression on all items.

The accuracy of the linear regressionmodel used will be tested using the root mean square error(rmse).

**Picture Representation of the Data Pipeline.**



**Sales Analysis and Prediction System Design Specification Document.**

**Summary.**

This document contains a detailed description of the components of the sales analysis and prediction data pipeline. In the description, each component of the pipeline is described by explaining what is means, what it involves, how it is done and why it is done. It also contains a diagram to explain the data pipeline.

Below is a description of each of the key components of the data pipeline.

1. **Data Loading.**

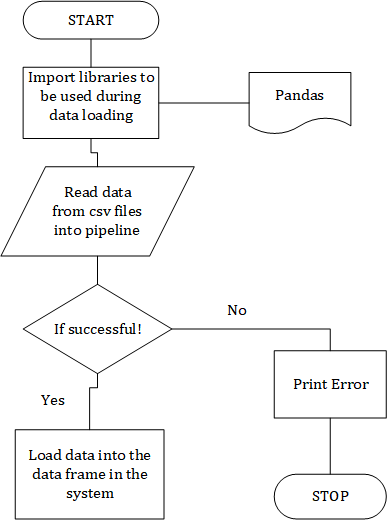
This step involves picking the data from the csv file in which it is stored and loading it into the system for analysis to be carried out on it. This is done to load the data into memory so that it can be cleaned and then analyzed by the system.

**Libraries to be used.**

There is one python library that are used in this process and these are:

* Pandas.

The process of loading the data from the csv file into the system is done by the use of the function, **read\_csv(“filename.csv”)**, which is found in the **pandas library**. This method will load the data from the csv file in which it is stored into a data frame in the system.



1. **Data Cleaning.**

This is the component in the data pipeline in which the data that has been loaded into the pipeline is prepared, corrected and put into the right format that can be analyzed by the use of the relevant libraries, methods and algorithms so as to get the right results from the analysis.

This is done because, initially, the data that has been loaded into the pipeline contains a number of errors, missing values and some of the values are in formats and datatypes in which the relevant libraries cannot handle analysis on them. Therefore, these errors and irregularities need to be handled so as to prepare the data for analysis which will yield the correct results.

**Methods to be used.**

There are three python libraries that are going to be used in this component in the pipeline. These are;

* Pandas.
* Numpy.
* Mode.

**Handling Missing values.**

This is the section of the data pipeline in which the columns containing missing values will be identified and all the records missing values will be filled up with the appropriate values.

In this component, the columns in the loaded data that are containing missing values will be identified using the function, **sum(DataFrame[‘Column\_Name’].isnull())**. This function will return the sum of all the missing values in the specified column in the data frame.

Then all the columns that will have been identified to have more than **zero** rows with missing values will have the data types of their expected values determined so as to determine the right method to be used to fill in their missing values.

In case a row in a numerical column is missing a value for a given product, the average of the values in that column for that specific product will be determined using the method, **DataFrame.pivot\_table(values=' Column\_name\_for\_column\_with\_missing\_values ', index='Column\_name\_for\_Unique\_Identifier')**. This method will return all the average values of each of the products in the specified column which is containing missing values.

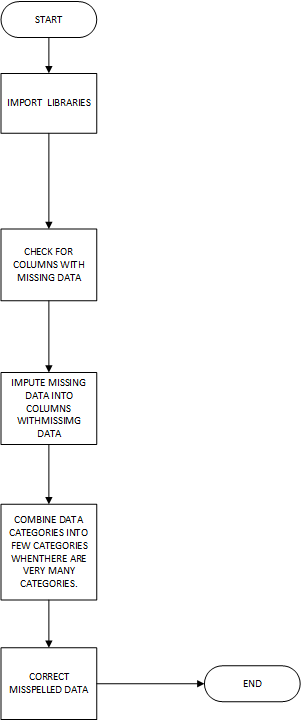
In case a row in a non-numerical column is missing a value for a given product, the mode of the values in that column for that specific product will be determined using the method, **DataFrame.pivot\_table(values='Column\_name\_for\_column\_with\_missing\_values', columns='Name\_of\_column\_which\_is\_the\_determinant',aggfunc=(lambda x:mode(x).mod e[0]))**. This method will return the modal value of the values of the determinant column in terms of the column which is containing missing values.

**Handling misspelled values.**

This involves identifying columns which contain values which have been input in different ways (for example, containing more than one value that represent the same thing). In these columns, all the different values that are meaning the same thing will be identified and then replaced by one value which will refer to that specific meaning.

**Label Encoding.**

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.



1. **Data Visualization.**

This involves the types of graphs that are to be used during the visualization of the data to enable the analyst to identify the different patterns and make the relevant conclusions.

**Libraries to be used.**

The libraries that are to be used in this part of the pipeline include;

* Matplotlib.
* Pandas.

The different graphs that are to be used to visualize the loaded data during the data analysis process are explained below.

* Histogram.

The histogram is a graphical representation of the distribution of a given set of data.

The histogram will be used to identify the distribution of some of the properties of the products that are being sold in the different store outlets. This will be done to identify the categories under which the different properties fall as far as affecting the volumes of sales of the different products fall.

The histogram will be plotted using the function, **hist()**.

* Line graph.

The line graph is a graphical representation of the loaded data which will show the rate of change and relationship between two properties of a given product.

The line graph will be used to determine how some of the properties change in relation to each other.

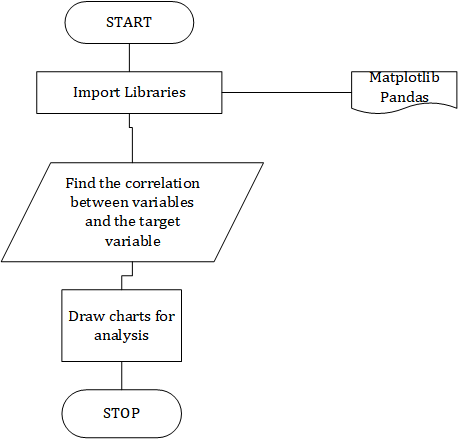
The line graph will be plotted using the function, **plot()**.

* Bar graph.

The bar chart is a graphical representation of data which can be used to determine the relationship between two categories of data.

The bar chart will be used to determine the relationship between some of the properties of the products being sold.

This chart will be plotted using the function, **bar()**.



1. **Apply Algorithms and Modeling Techniques.**

In this section, a set of models will be applied so as to identify the relationships between some of the properties of a product so as to make predictions of the sales of the products.

**Libraries to be used.**

The libraries that are to be used are listed below:

* Pandas.
* Matplotlib.
* Seaborn.

**Models.**

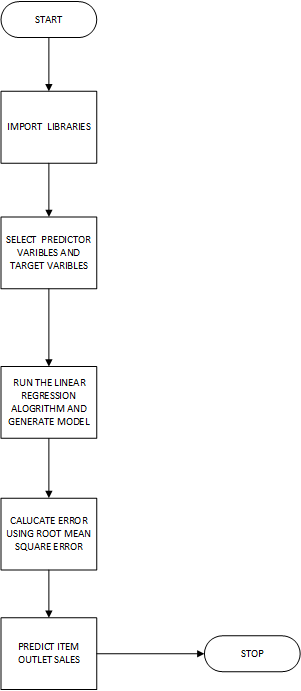
These are the algorithms that are to be used so as to analyze the data and come up will a set of conclusions and predictions.

* **Linear Regression.**

The linear regression model will be used so as to determine the way the different properties affect the volume of sales of a given product and to predict the sales volumes of the product.

* **Correlation.**

The correlation analysis will be done so as to determine the relationship between some of the product properties.

****

**Systems Requirements Specification.**

**Document Description.**

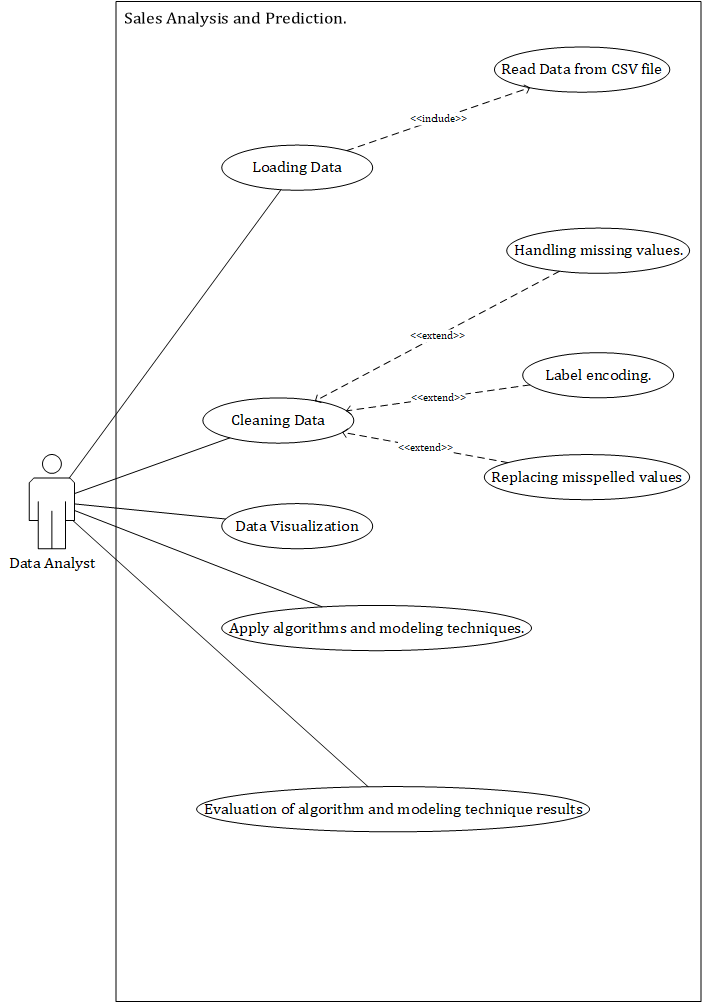
This chapter contains a detailed use case (and its description) of the sales analysis and prediction data pipeline through which the data will pass when being analyzed. It states the primary users of the pipeline and also how they will be interacting with it in order to get the desired outcome from the data analysis process.

**Primary User.**

The data analyst is the primary user of the data pipeline whose use case is described below.

**Use Case Diagram.**

Below is the use case diagram of the sale analysis and prediction data pipeline.



**Description of the Use Cases in the Use Case Diagram.**

Below are the descriptions of the key use cases in the use case diagram.

1. **Loading Data.**

In this module, the data, which is stored in a csv file, will be copied and loaded into the system. This will be done so that an analysis can be done on the copied data. In this use case, there is a sub-module of **Read Data from CSV file** in which the data will be copied from a csv file into the system.

**Outcome.**

At the end of this module, the data which is to be analyzed will have been copied from the csv file in which it is stored and loaded into the system for analysis and evaluation.

1. **Cleaning Data.**

In this module, the data which will have been loaded into the system will be cleaned and prepared before the analysis process can be carried out on the data. In this module, there will be three sub-modules, **Handling missing values**, **Replacing misspelled values** and **Label encoding** in which the missing values in the data will be filled, all the misspelled in the data will be corrected and label encoding will be carried out on the loaded data respectively.

**Outcome.**

At the end of this module, the loaded data will have been converted into a state in which an analysis can be comfortably carried out on the data without having any errors or wrong results.

1. **Data Visualization.**

In this module, the data which will have been cleaned will be visualized using the selected visualization tools and graphs/plots so as to enable the analyst to get some of the required deductions from the data.

**Outcome.**

The end result of this module will be a set of graphs and plots generated from the data which will enable the analyst to acquire some of the required deductions concerning the data analysis process.

1. **Apply algorithms and modeling techniques.**

In this module, the necessary data analysis algorithms will be applied to the data so as to identify which product properties affect the volumes of sales of the specific product most as well as predict the sales of that product. The model to be used in this module is a **Linear Regression Model**.

**Outcome.**

The end result of this module will be the predicted volume of sales of each product as well as a summary of which product properties affect the volumes of sales of all the products the most.

1. **Evaluation of algorithm and modeling technique results.**

In this module, an evaluation of the results of the model used will be done so as to establish its level of accuracy.

**Implementation Report.**

**Introduction.**

This chapter contains the findings that resulted from the analysis of one of the Big-Mart sales datasets. The analysis was made to determine which product properties affect the outcome volume of sales of each product/item in each sales outlet the most and how they can be manipulated in order to increase the sales of theses items. It contains all the relevant data visualizations, detailed explanations of all the charts used together with the insights and conclusions that are made from them. In addition to the insights made, some recommendations, in addition to the insights acquired, are also included that will be used to increase on the sales of every specific product.

**Correlation Analysis.**

An analysis is carried out to establish the correlation coefficient of between the variables, **Item\_MRP**, **Item\_Weight**, **Item\_Visibility** and **Outlet\_Age** and the target variable, **Item\_Outlet\_sales**.

These correlation coefficients are listed below.

|  |  |
| --- | --- |
| **PREDICTOR VARIABLES** | **Item\_Outlet\_Sales (TARGET VARIABLE)** |
| **Item\_MRP** | 0.567574 |
| **Item\_Weight** | 0.013261 |
| **Item\_Visibility** | -0.128453 |
| **Outlet\_Age** | 0.049135 |

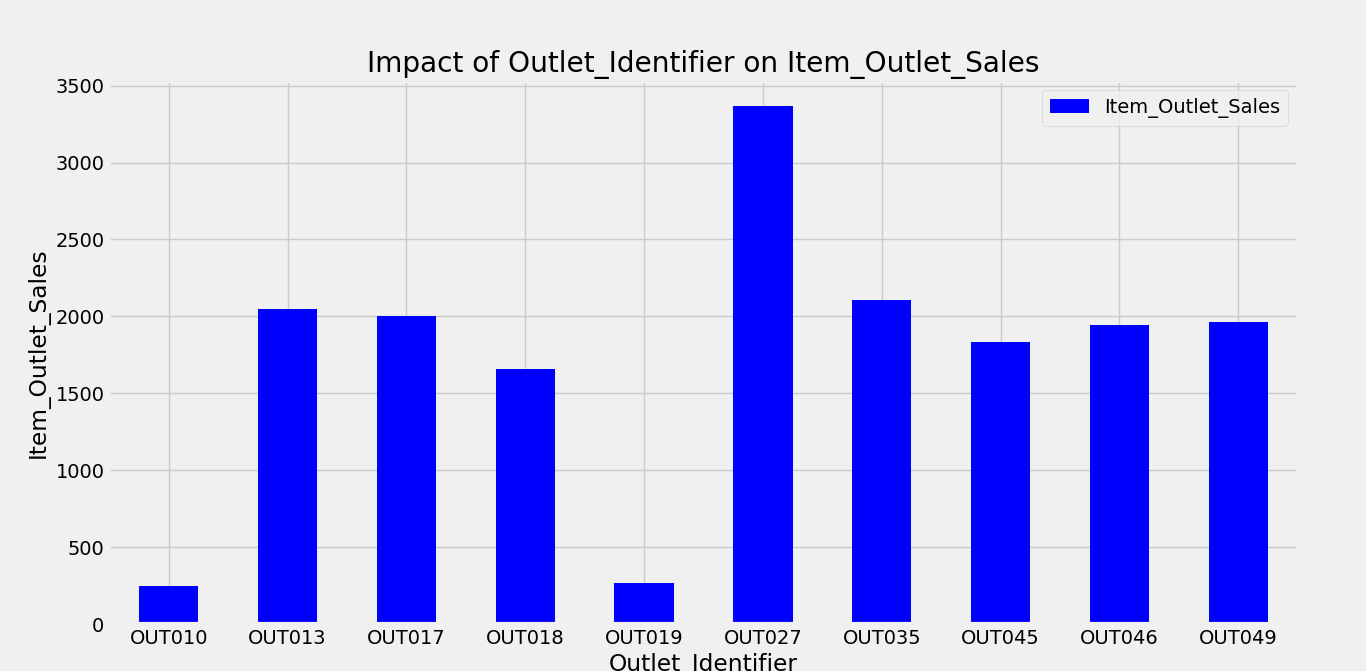
**Conclusion from the Correlation Coefficients.**

Basing on the correlation coefficients that are showing the relationship between the variables (**Item\_MRP**, **Item\_Weight**, **Item\_Visibility** and **Outlet\_Age**) and the target variable (**Item\_Outlet\_Sales**), it is observed that the variables, **Item\_Weight** and **Outlet\_Age** have a very weak linear relationship with target variable and, therefore, do not make any effect on the target variable as they change.

Therefore, the **Item\_Weight** and **Outlet\_Age** variables that are not to be used during the analysis.

**Data Visualization Analysis.**

1. **The Impact of the Outlet/Store on the Sales of an Item.**



**Figure 1.1: An analysis of the outlet identifier with the maximum item sales.**

**The description of the above graph.**

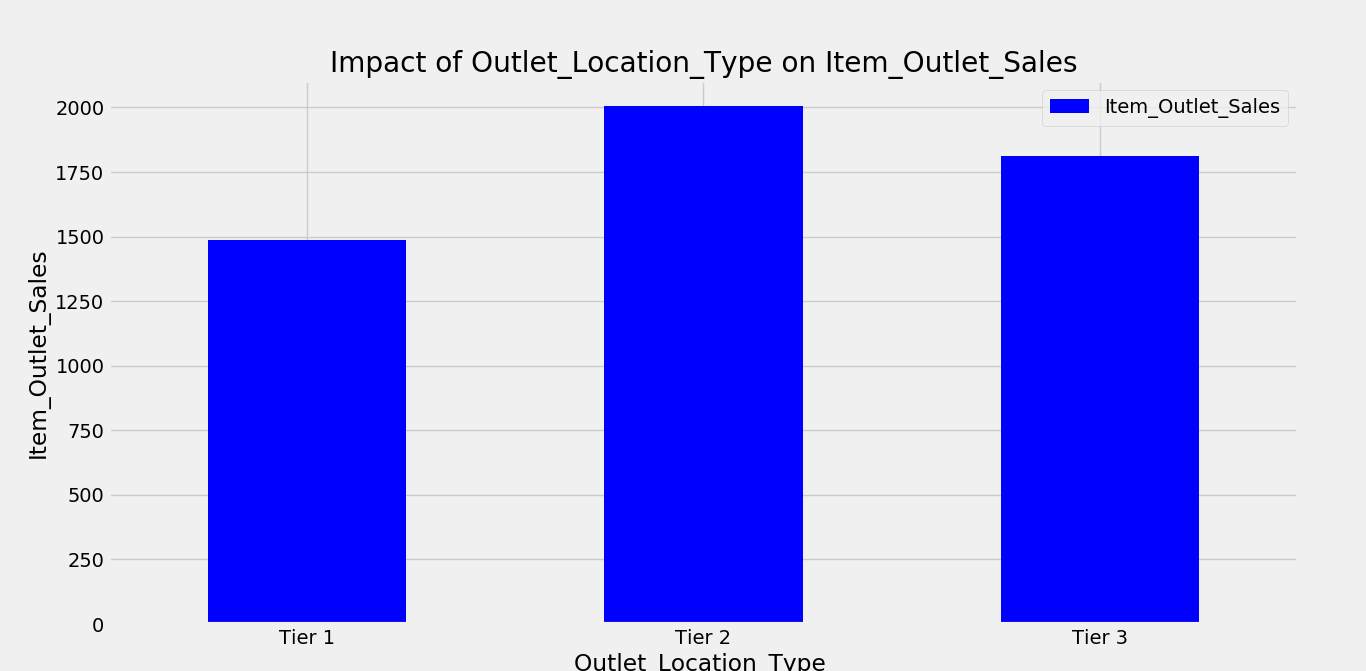
The above graph, **Figure 1.1**, is a bar chart showing the maximum sales that were made in all ten Big-mart outlets. The outlets have been plotted on the horizontal axis, while the maximum sales made at the outlets have been placed on the vertical axis. Each of the outlets is represented by a unique **Outlet Identifier**.

**Insights and observations.**

From the above graph, **Figure 1.1**, it was observed that as of the year 2013(when the analyzed dataset was compiled) the outlet at which the most sales were made was **OUT027** and the two outlets at which the least sales were made were **OUT010** and **OUT019**.

It was also observed that the outlet at which the most sales were made, **OUT027**, is the **only** outlet of type, **Supermarket Type3** and the two outlets at which the least sales were made, **OUT010** and **OUT019**, are the **only** outlets of the type, **Grocery Store**. Furthermore, the outlet whose sales were less than those of all the other outlets and only greater than those of the two grocery stores, **OUT018**, is of the type, **Supermarket Type2**. All the other outlets are of type, **Supermarket Type1**. This leads to the conclusion that the **Grocery Store** outlets make the least sales, followed by the **Supermarket Type2** outlets, which are then followed by the **Supermarket Type1** outlets and then the **Supermarket Type 3** outlets make the most sales.

It has also been observed that most of the outlets are of the type **Supermarket Type1**.



**Figure 1.2: An analysis of the outlet location type with the maximum item sales.**

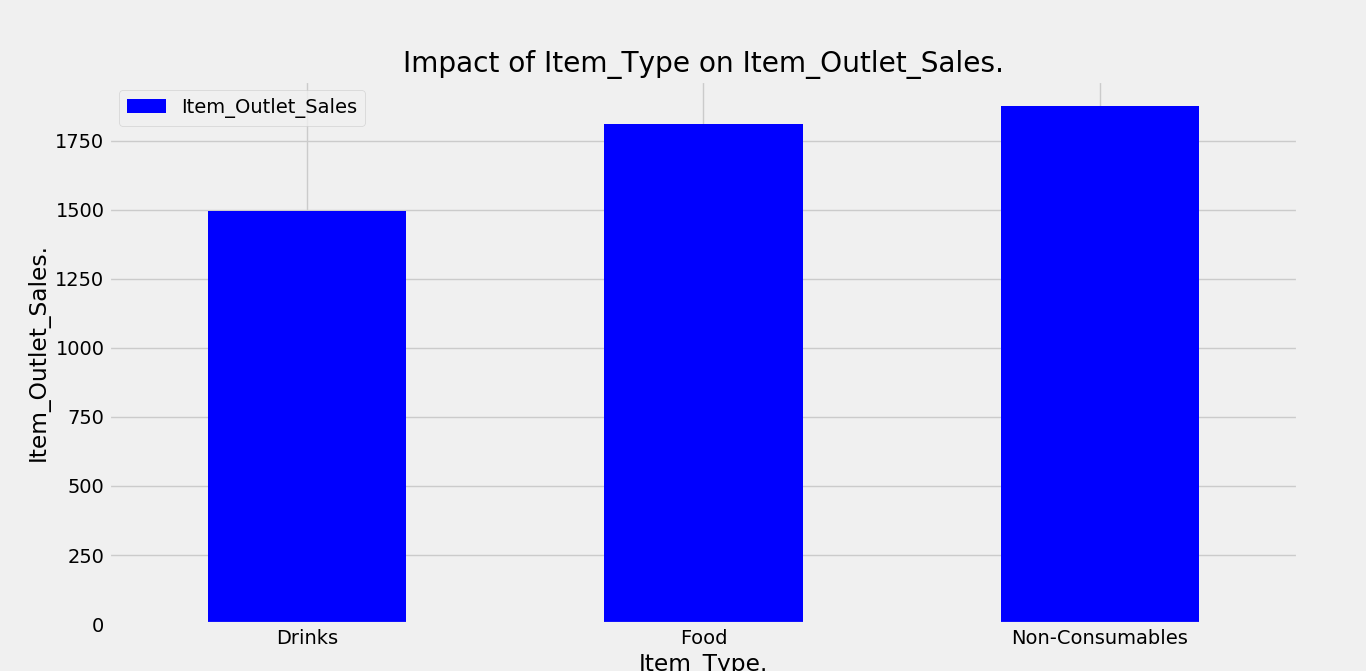
**The description of the above graph.**

The above graph, **Figure 1.2**, is a bar chart showing the maximum sales that were made in all the three location types in which the ten outlets are situated. The outlet location types have been plotted on the horizontal axis, while the maximum sales made in the location types have been plotted on the vertical axis. There are three outlet location types, i.e., **Tier 1**, **Tier 2** and **Tier 3**.

**Insights and observations.**

From the graph, **Figure 1.2**, it was observed that the outlets in **Tier 2** locations make the most sales, followed by those in **Tier 3** locations and then the outlets in the **Tier 1** locations make the least sales.

1. **The impact of the item specifications/properties on the item sales.**



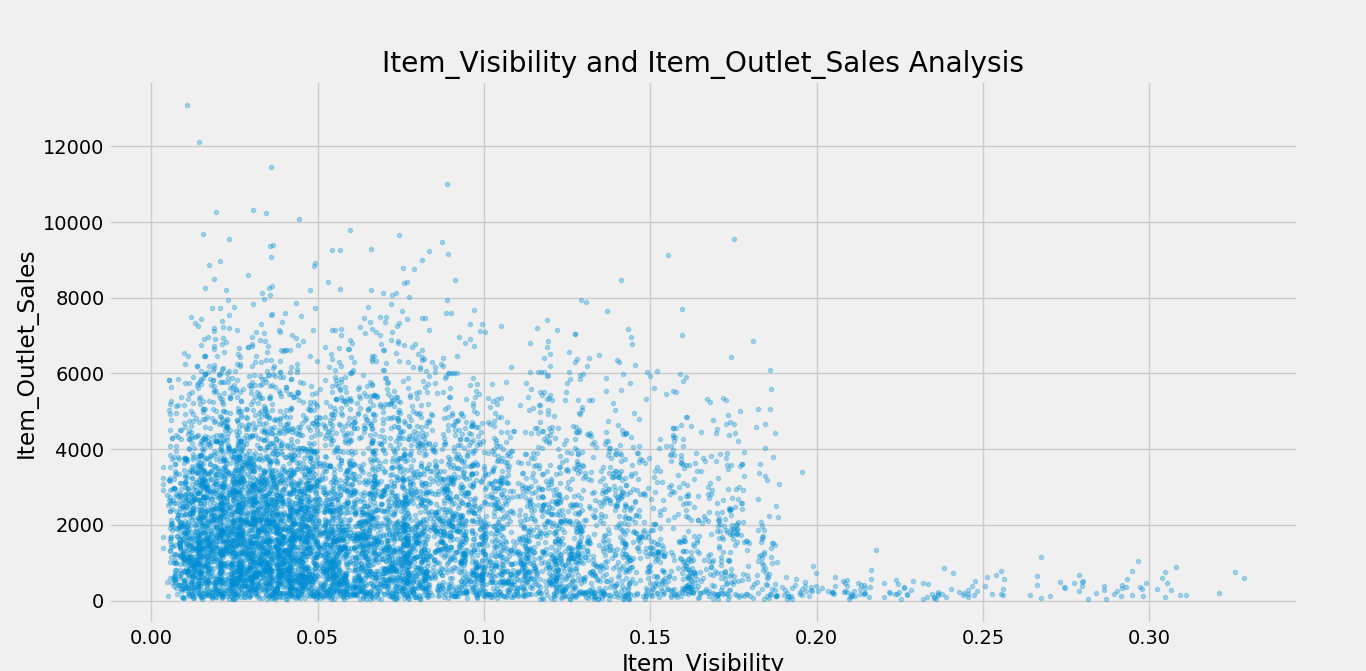
**Figure 2.1: A graphical representation of the relationship between the item type and the maximum item sales.**

**A description of the above graph.**

The above graph, **Figure 2.1**, is a bar chart showing how the item type affects the maximum sales of the items. The item type has been plotted on the horizontal axis, while the maximum sales of the items have been plotted on the vertical axis.

**Insights and observations.**

From the graph, **Figure 2.1**, it is observed that the **non-consumable items** generate the most sales when compared to the other two types of items. The **drinks** generate the least sales followed by the **foods** which generate the second most sales.



**Figure 2.2: Graphical analysis of the relationship between the maximum item sales and the item visibility.**

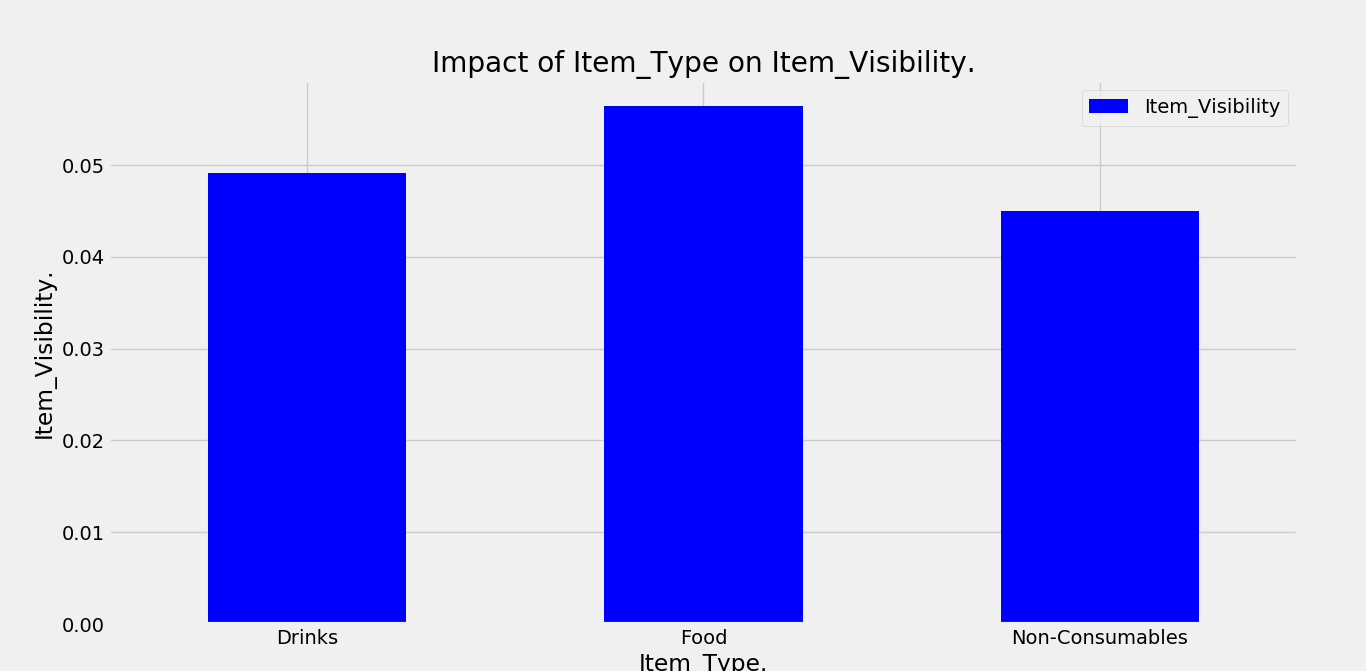
**A description of the above graph.**

The above graph, **Figure 2.2**, is a scatter plot showing how the maximum item sales change with respect to the total visibility given to an item towards the customers in the item in which it is being sold. The item visibility has been plotted on the horizontal axis, while the maximum sales of the items have been plotted on the vertical axis.

**Insights and observations.**

In addition to the insights generated from **Figure 2.1**, **Figure 2.2**(in support of the correlation coefficient of the item visibility and item sales, which is **-0.128453**) shows that the items that are given less visibility towards the customers in the stores generate more sales than those that are given more visibility. This shows that the items that tend to have less market are given more visibility so as to attract customers to purchase them and those with more market are given less visibility because they are already on demand.

It is also observed that there are more items with a lower visibility and only a few items that are given a very large amount of visibility towards the customers and these generate a few sales.

****

**Figure 2.3: A graphical representation of the relationship between the item type and the item visibility.**

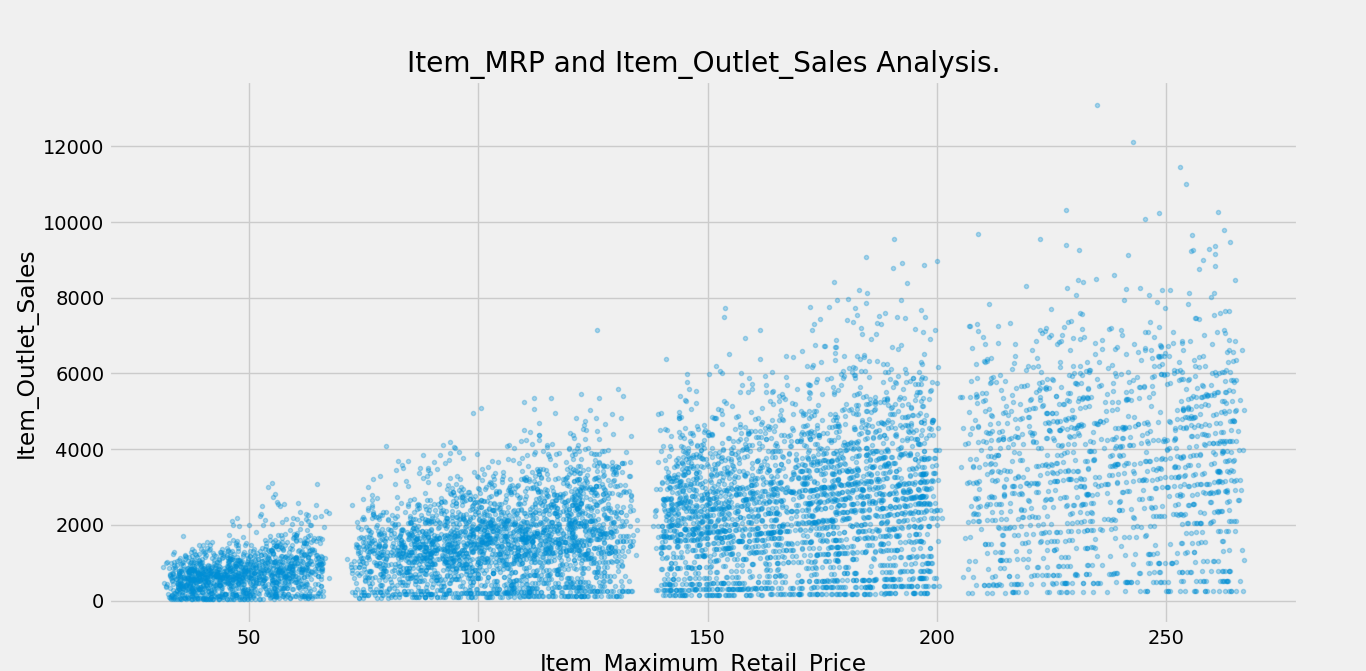
**A description of the above graph.**

The above graph, **Figure 2.3**, is a bar chart showing the total visibility that is given to each of the three types of items sold in all the outlets. The item visibility has been plotted on the vertical axis, while the item type has been plotted on the horizontal axis. There are three item types, i.e., **Drinks**, **Food** and **Non-Consumables**.

**Insights and observations.**

From **Figure 2.3**, it is observed that **non-consumable** items are given the least visibility when compared to the two other item types. The **food** items are given the most visibility followed by the **drinks** which have the second highest amount of visibility.

Basing on the graphs, **Figure 2.1**, **Figure 2.2** and **Figure 2.3**, the items with less visibility generate the most sales and those with more visibility generate the least sales for the outlets. The **non-consumable goods**, which have the least visibility, generate the most sales. However, the **food** items, which have the highest amount of visibility, generate the second highest amount of sales and the **drinks**, which have the second highest amount of visibility generate the least amount of sales. This must be because the **food** items tend to be more perishable than the **drinks**. And even if they have more market, they need to be purchased very quickly so as to avoid making losses.

****

**Figure 2.4: Item\_MRP analysis against item sales.**

**Conclusions and recommendations.**

Basing on the findings in the plotted charts, in order to increase sales, the management need to stock more **non-consumable** goods because they are more marketable.

The administration is also encouraged to stock more **Food** goods while giving them **more visibility** with increase in the amount of stock so as to increase sales and avoid making losses from perishing of these goods.

It is also recommended that in order to increase sales totals, the **Maximum Retail Price(MRP)** of an item can be increased.

It is also advised that they stock more items into outlets of **Supermarket Type3** because they have been seen to make more sales than the other outlet types.

**Applying the regression algorithm.**

Linear regression is used to analyze and predict the target variable.

When the linear regression is used, the accuracy of the model was measured using the **Root Mean Square Error** (**RMSE**) method and it generated an accuracy value of **1128**.

**The New Github Repository.**

The link onto which the work is uploaded onto the repository is:

https://github.com/Data-Science1998/RecessDataScience.git