

Project Report

“BigMart Sales Prediction”

Phase 1: (Data Cleaning)

Name: Mohammad Hunain | Course: Machine Learning | Date: 16-Nov-2018

1. **What is the business problem?**

“Business problem is to increase the sales of any SuperMarts or BigMarts”

So the idea is to find out the properties of a product, and store which impacts the sales of a product.

1. **How business is working?**

A **BigMart** is a self-service shop offering a wide variety of food and household products, **organized** into aisles. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market

1. **Possible solutions:**

The possible solution are:

Store Level Solution:

* **City type:** Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.
* **Population Density:** Stores located in densely populated areas should have higher sales because of more demand.
* **Store Capacity:** Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place
* **Marketing:** Stores which have a good marketing division should have higher sales as it will be able to attract customers through the right offers and advertising.
* **Location:** Stores located within popular marketplaces should have higher sales because of better access to customers.
* **Customer Behavior:** Stores keeping the right set of products to meet the local needs of customers will have higher sales.
* **Ambiance:** Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

Product Level Solution:

* **Brand:** Branded products should have higher sales because of higher trust in the customer.
* **Packaging:** Products with good packaging can attract customers and sell more.
* **Utility:** Daily use products should have a higher tendency to sell as compared to the specific use products.
* **Display Area:** Products which are given bigger shelves in the store are likely to catch attention first and sell more.
* **Visibility in Store:** The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.
* **Advertising:** Better advertising of products in the store will should higher sales in most cases.
* **Promotional Offers:** Products accompanied with attractive offers and discounts will sell more.

1. **What are the goals of business problem?**

The Goal of a business problem is to find out the properties of a product and stores, which impacts in increasing the sales of a product.

1. **In what ways a Predictive Analytic model help to address the business problem?**

There are some basic descriptive features that help our predictive model to predict the sales of the product based on previous data available of that particular store, these features play a vital role in finding the solution. The model take these descriptive feature and check the relation between the descriptive features with the target feature hence to figure out the optimized solution. it find out the sales of each product at a particular store and give insights for future sales. So that they can increase their future sales.

1. **Data Availability and Data Constraints:**

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

1. **Organization constraints:**

Organizationalconstraintsof BigMart sales including:

* Capacity of Mart
* Their Product Management
* Location of Mart
* Competitors: Stores having similar establishments nearby should have less sales because of more competition.

1. **Data description:**

|  |  |
| --- | --- |
| **Item\_Identifier** | Unique product ID |
| **Item\_Weight** | Weight of product |
| **Item\_Fat\_Content** | Whether the product is low fat or not |
| **Item\_Visibility** | The % of total display area of all products in a store allocated to the particular product |
| **Item\_Type** | The category to which the product belongs |
| **Item\_MRP** | Maximum Retail Price (list price) of the product |
| **Outlet\_Identifier** | Unique store ID |
| **Outlet\_Establishment\_Year** | The year in which store was established |
| **Outlet\_Size** | The size of the store in terms of ground area covered |
| **Outlet\_Location\_Type** | The type of city in which the store is located |
| **Outlet\_Type** | Whether the outlet is just a grocery store or some sort of supermarket |
| **Item\_Outlet\_Sales** | Sales of the product in the particular store. This is the outcome variable to be predicted. |

1. **Loading Data, Analyzing and Visualizing:**

* Load the data.
* Correlation.
* Visualization and Analyze (Scatter Plot and Pie Chart).
* Histogram
* Statistics

1. **. Data Cleaning:**

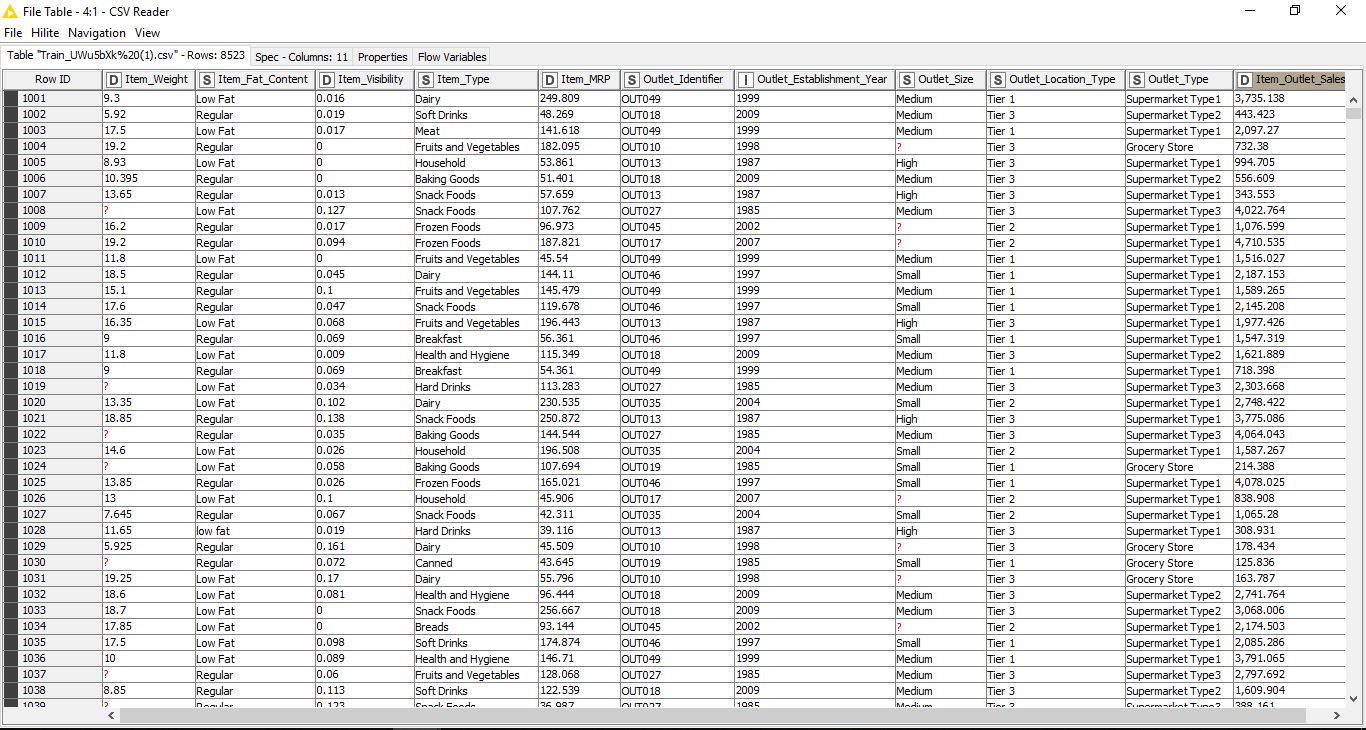
* Filter column missing value.
* Normalize the data.
* Find outliers.

1. **. ABT (Analytic Base Table):**
2. **. Workflow:**

(All Screenshots are below respectively)

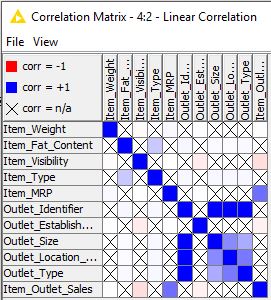
**Loading Data, Analyzing and Visualizing:**

* **Load the data.**

****

* **Correlation:**

**Correlation** helps me to find what columns might be related to each other and have a higher predictive power to help me in my final results.

****

**Visualization and Analyze:**

**One of the primary things I would like to know from my data would be that what item is sold the maximum out of the others.**

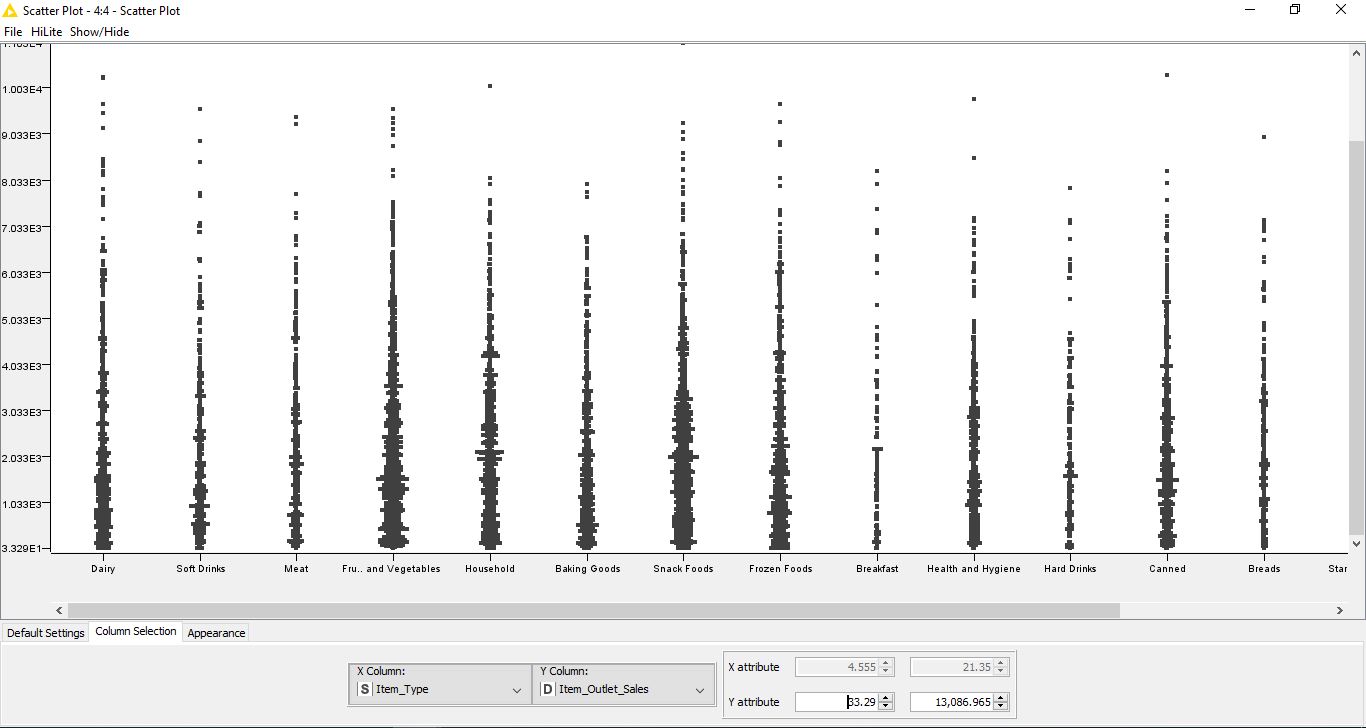
**There would be two ways to interpret the information:**

1. **Scatter Plot**

**Configure:**

**I configured my node to select how many rows of the data I need and wish to visualize. [I chose 9000]**

**I have selected the X axis to be Item\_Type and the Y axis to be Item\_Outlet\_Sales.The plot below represents the sales of each item type individually,**

****

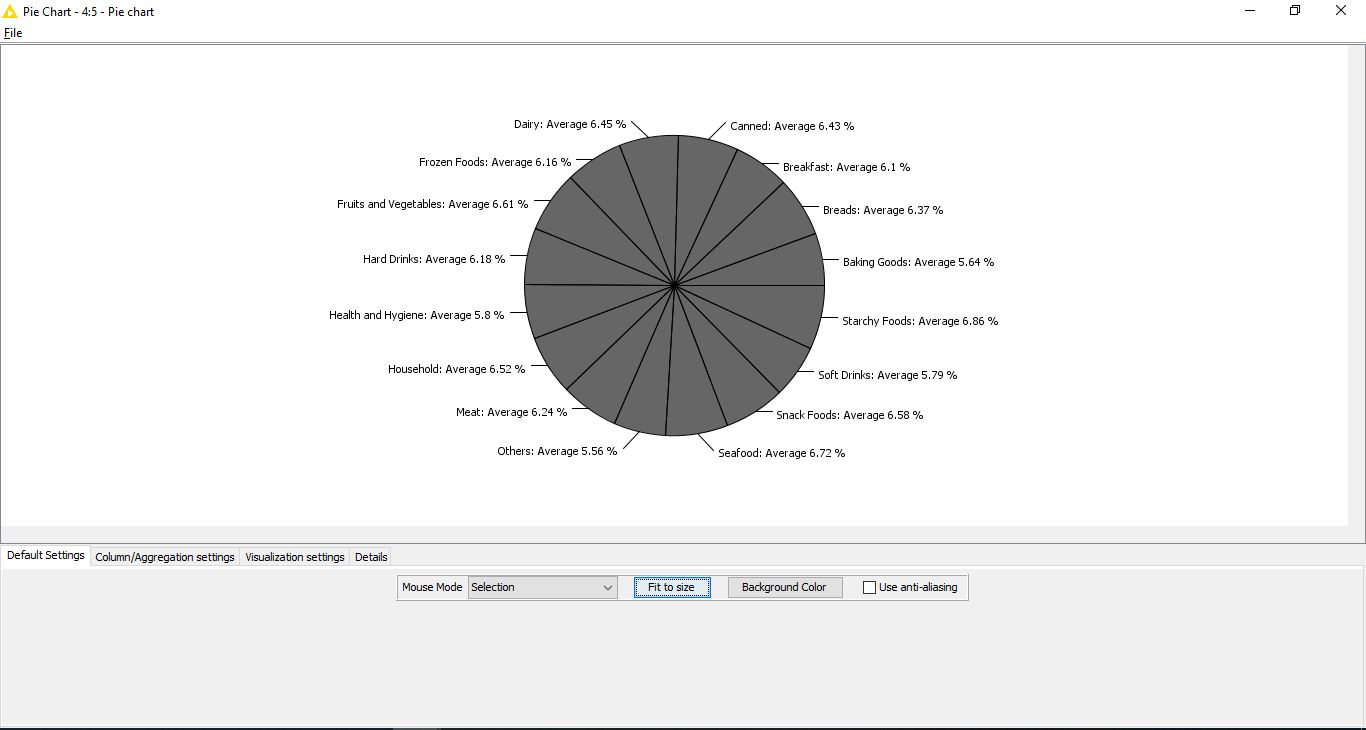
1. **Pie Chart**

To understand an average sales estimate of all product types in our database, I used a pie chart.

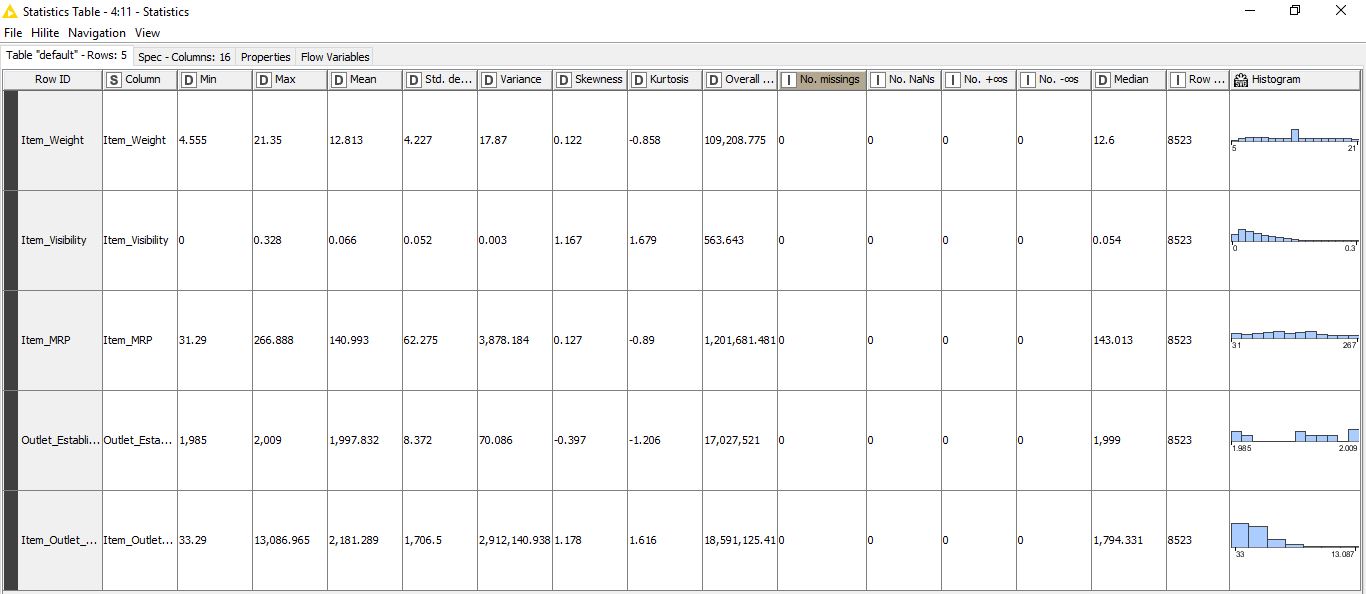
**Configure:**

Pie Column: Item\_Type  
Aggregation Column: Item\_Outlet\_Sales  
Aggregation Method: Average

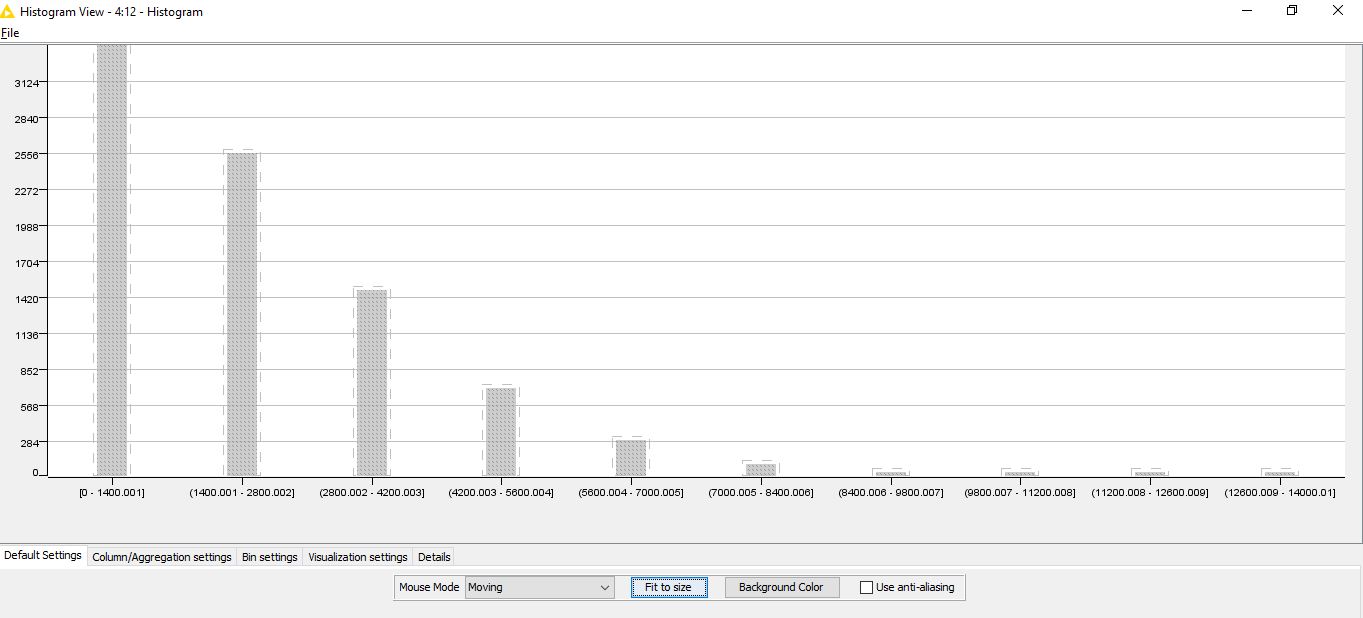
This chart shows me that sales were averagely divided over all kinds of products. “Starchy Foods” amassed the highest average sales of 6.86%.



**Statistical Table:**

****

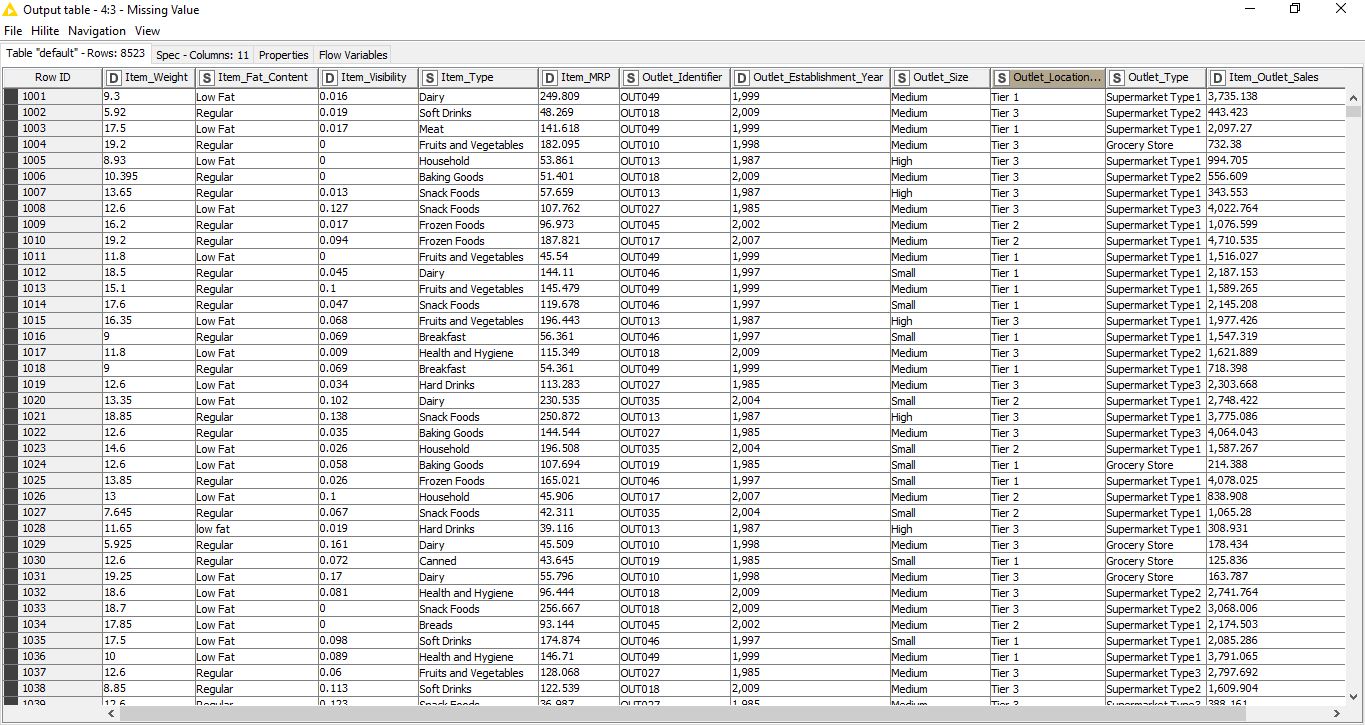
**Histogram:**

****

1. **. Data Cleaning:**

* **Missing Values.**

**Imputation:** Number (Double, integer) =Median , String =MostFrequentValue

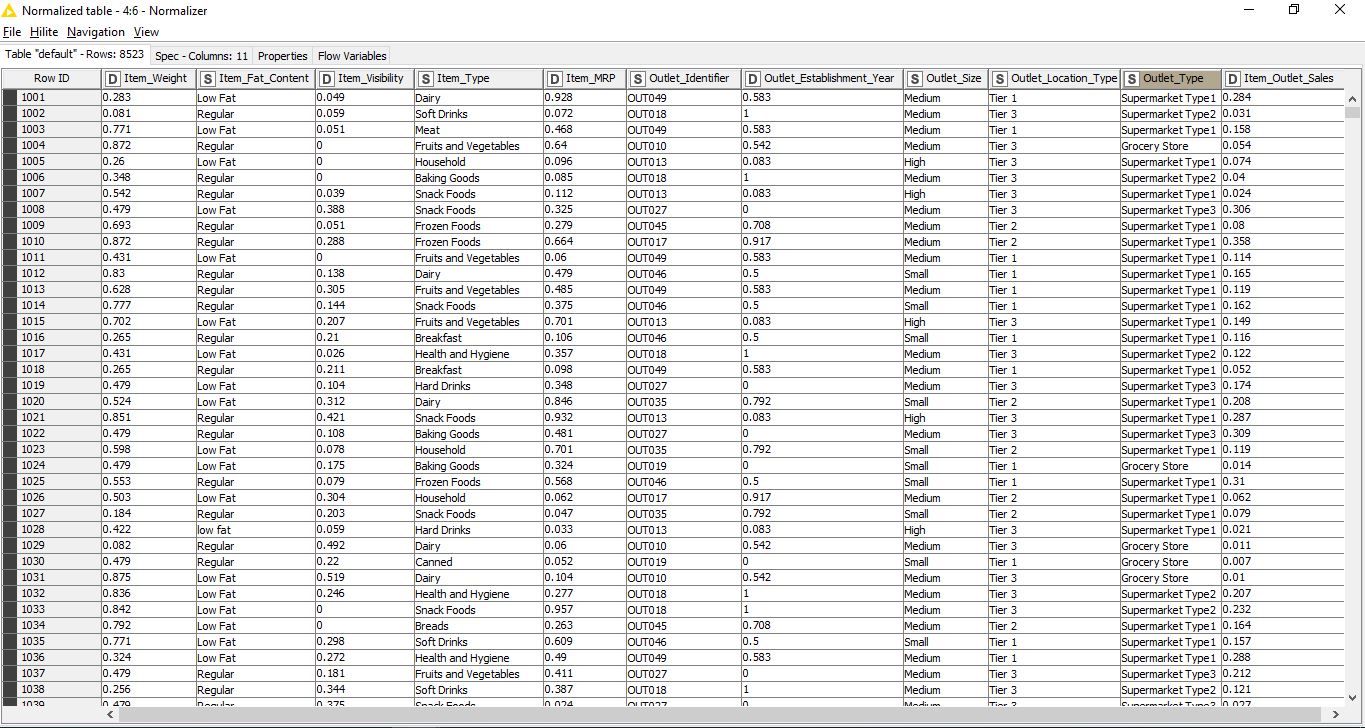


* **Normalize the data.**

**Configure :**

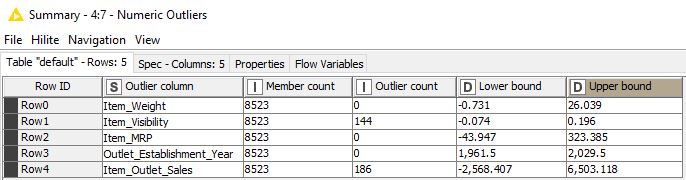
Min = 0 , Max = 1

**Result:**



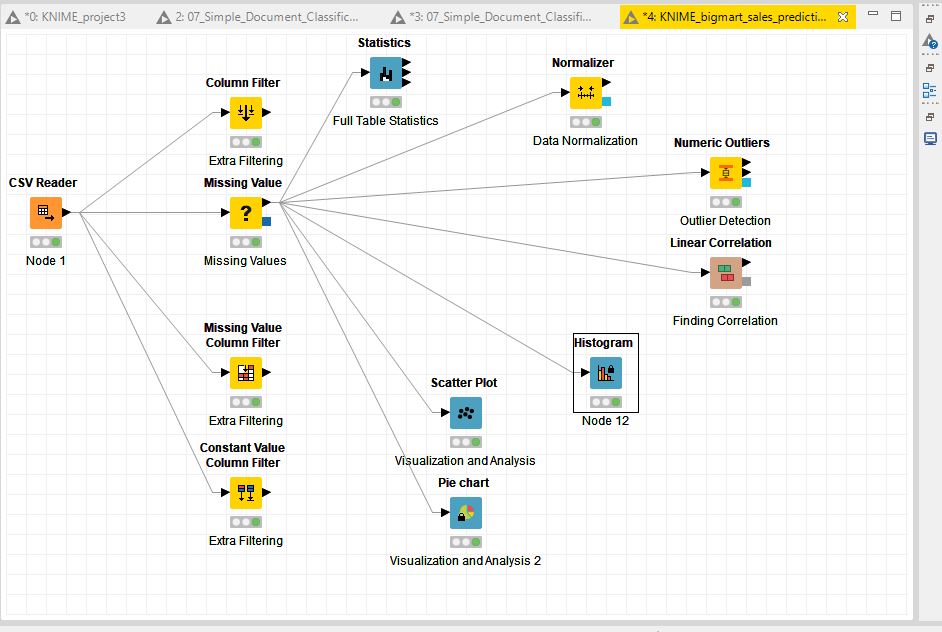
* **Finding Outliers:**

**Result:**

****

**Workflow:**

**Result:**



**ABT (Analytic Base Table):**

