

Course ALY6015 80797 Intermediate Analytics SEC 02

Spring 2019 CPS

# Module 6: Final Project

# BIGMART SALES INSIGHTS

Submitted by: Shivam Chauhan

Index

* Abstract
* Introduction
* About the dataset
* Data Cleaning
* Data pre-processing
* Analysis
* Conclusion
* Reference

Abstract

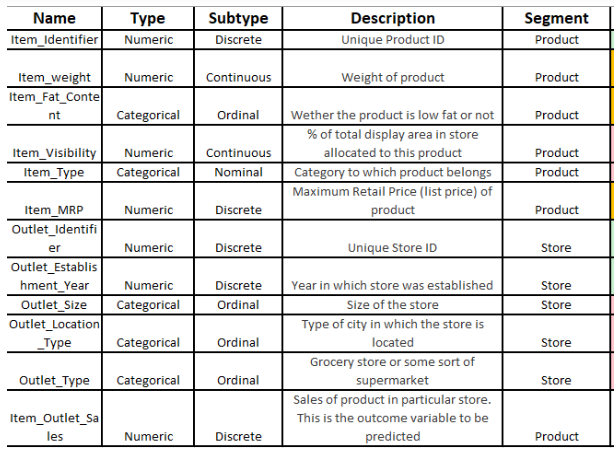
The report talks about the analysis which was performed on a dataset which contained the Big mart sales prediction. Parameters like data about store and product were some of the variables considered for the determining the sales. The first section gives a general introduction about how the project was envisioned and what are the datasets and methods used for the analysis. The second section gives a detailed explanation of the analysis which was carried out in R and the predictive model which was created for determining the cost of a random individual. The third section gives a brief conclusion of the project while answering some of the research questions mentioned before and the future scope of this analysis.

Introduction

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. The main objective is to understand whether specific properties of products and/or stores play a significant role in terms of increasing or decreasing sales volume. To achieve this goal, we will build a predictive model and find out the sales of each product at a store. This will help BigMart to boost their sales by learning optimized product organization inside stores.

About the data

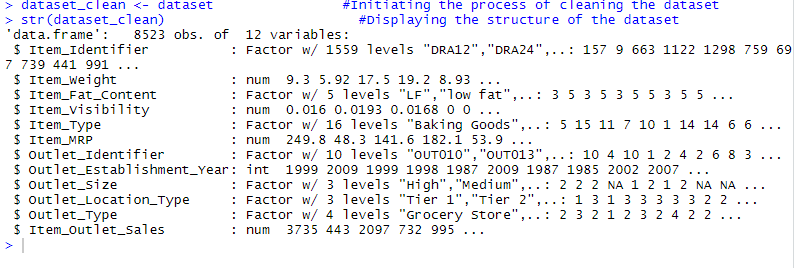
This dataset contains **8523 observations**and **12 features.**

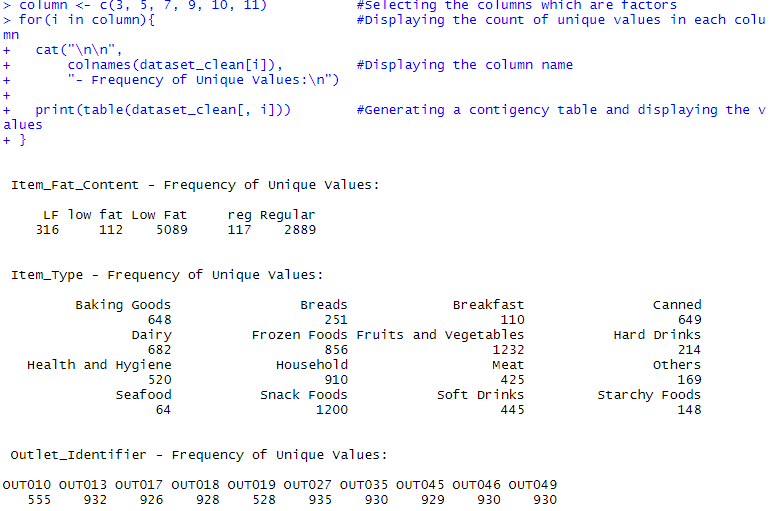


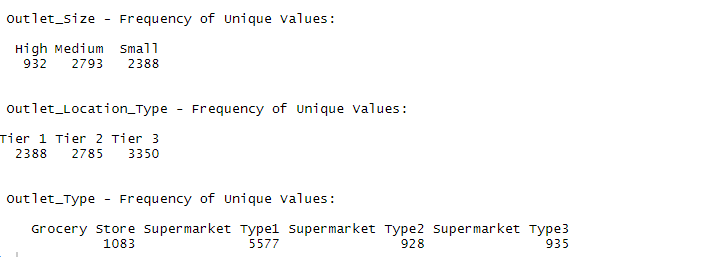
Data Cleaning

Data cleaning is very important process as we can observe that there are several missing values in the dataset. With an unclean data, we may get wrong insights which can lead to losses if decision is made on those insights. Data may have missing values as some store might not report all the data due to technical glitches. Here, it will be required to treat them accordingly.

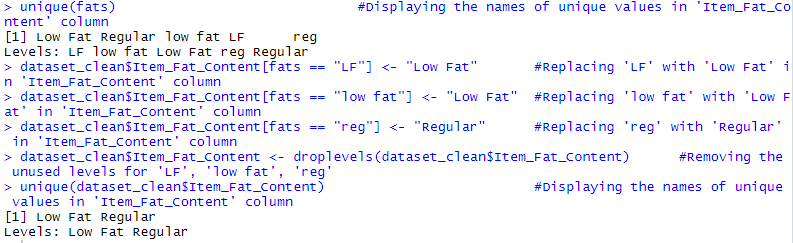
First, we will see the structure of the data and ggenerate a contingency table and displaying the values for all the columns which are factors.



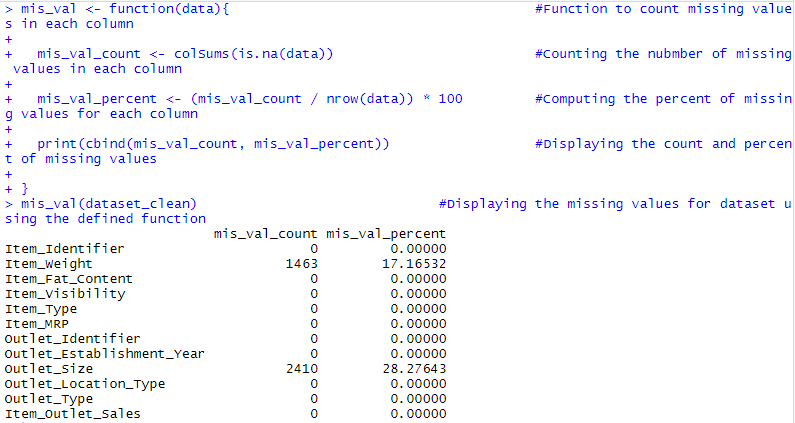




We can observe that data is not clean. For Item\_Fat\_Content there are only two categories, however it is divided into 5 categories which is not right. LF, low fat and Low Fat are all same. Regular and reg are all same.



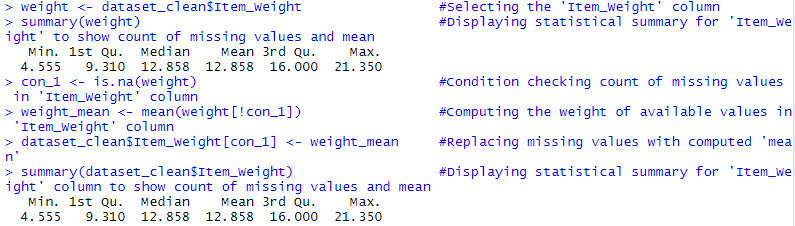
Now we will check how much data is missing.



We can observe that there are two columns which has missing values. 17.16% of data for Item\_Weight and 28.27% data of Outlet\_Size is missing.

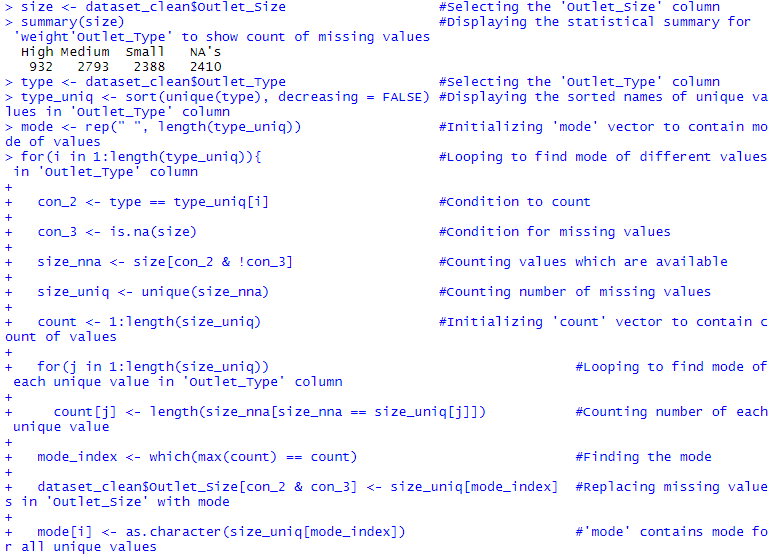
Removing all the missing data rows may affect the findings and can lead to wrong findings because there is higher number of data which is missing.

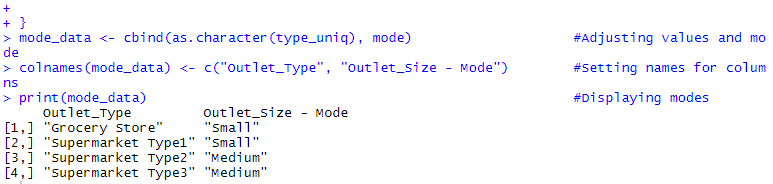
In this case we will replace all the missing values in Item\_Weight with the mean weight of the items.



After replacing the missing values with mean weight in Item\_Weight, we can observe the mean, median and quartiles are not changing.

Now we will replace all the missing values in Outlet\_Size column with the mode of each categories of Outlet\_Type because we think that outlet size is related to the outlet type.

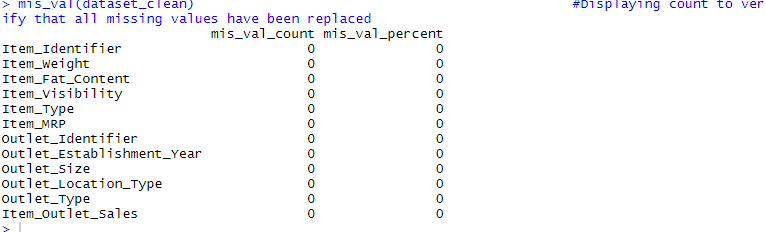




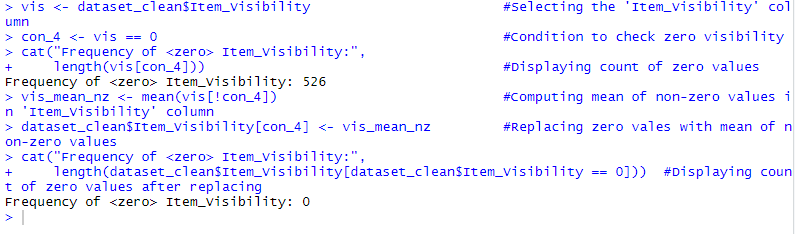
We found mode of outlet size for each outlet type and we have replaced the missing value in the Outlet\_Size column with respect to the missing value row Outlet Type.



Now we will check again how much data is missing.



We can see that no data is missing, however for the Item\_Visibility we have observed that large number of lows the value is 0. Visibility is the percentage of display area allocated to product. Logically if the item is sold, it should be available in store. We will replace the 0 value with the mean of the Item\_Visibility column.

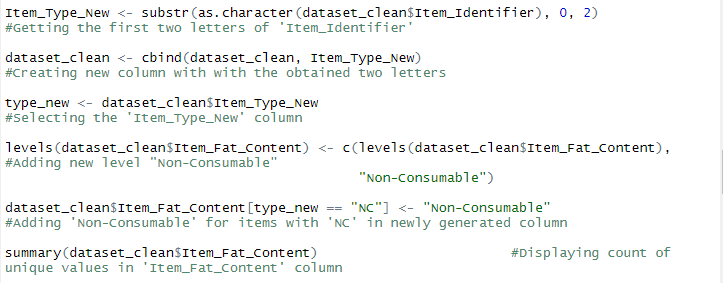


We can see that now our data is clean, and we can prepare the data for analysis.

Data Preprocessing

Before Analyzing the clean data, we will be creating the groups or diving the categories so that we can use this information to find more insights or find insights for a group separately.

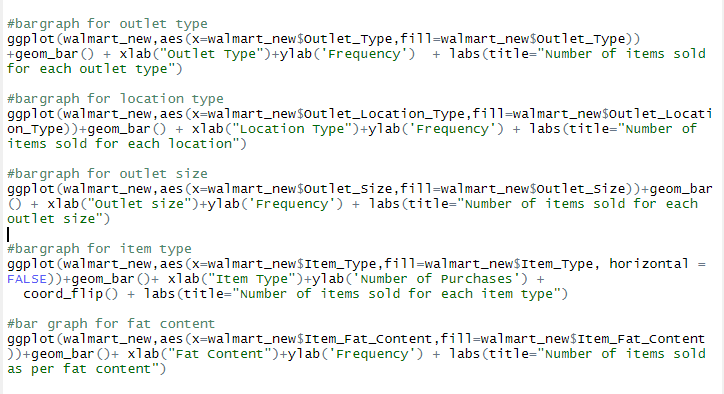
We will add a new level in the Item\_Type column as “non-consumable” for non-consumable items.



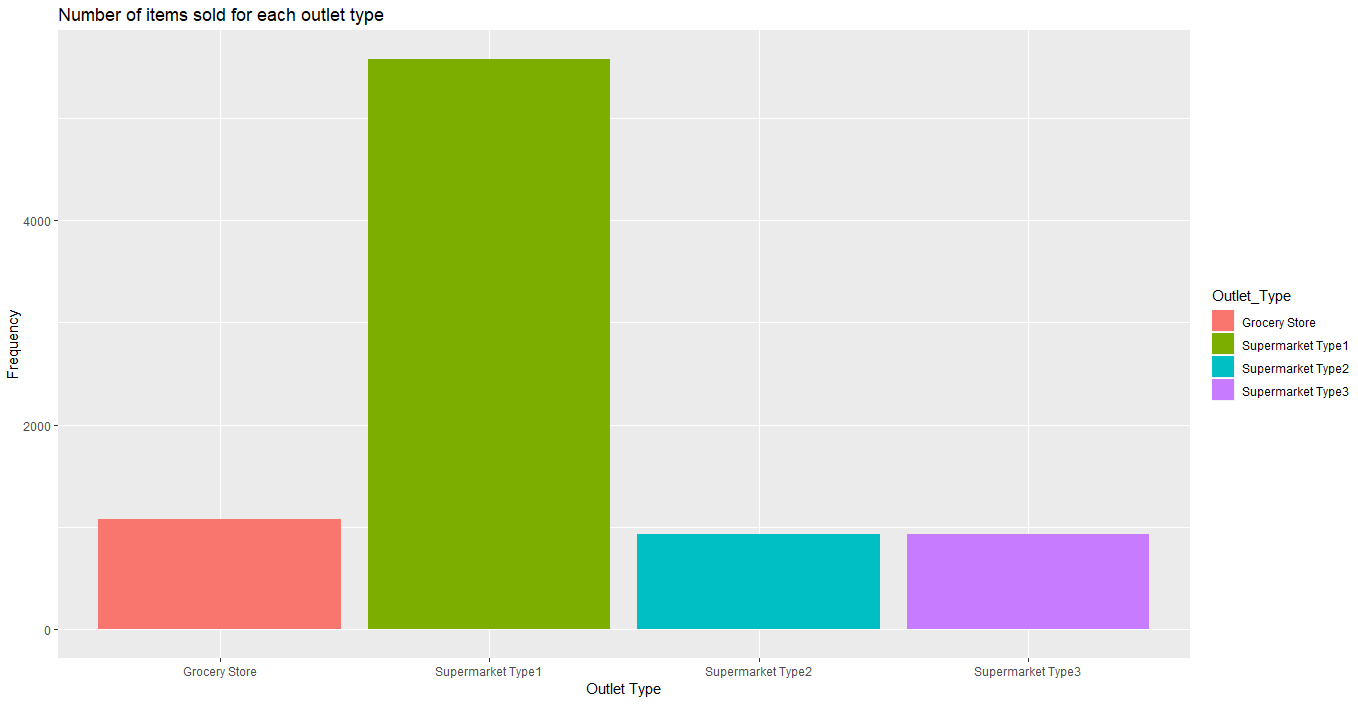


Analysis

Exploratory Data Analysis

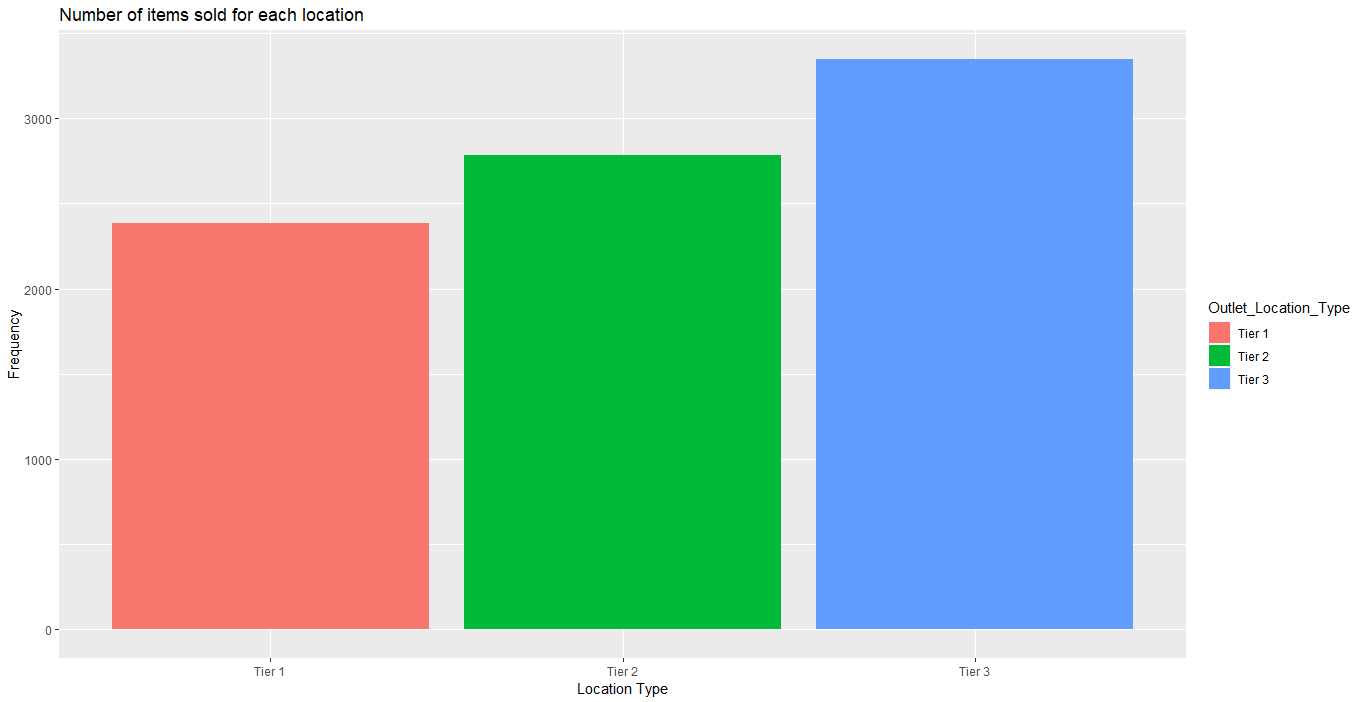
First, we will be analyzing the categorical variable.

1. Which outlet type records the greatest number of purchases.



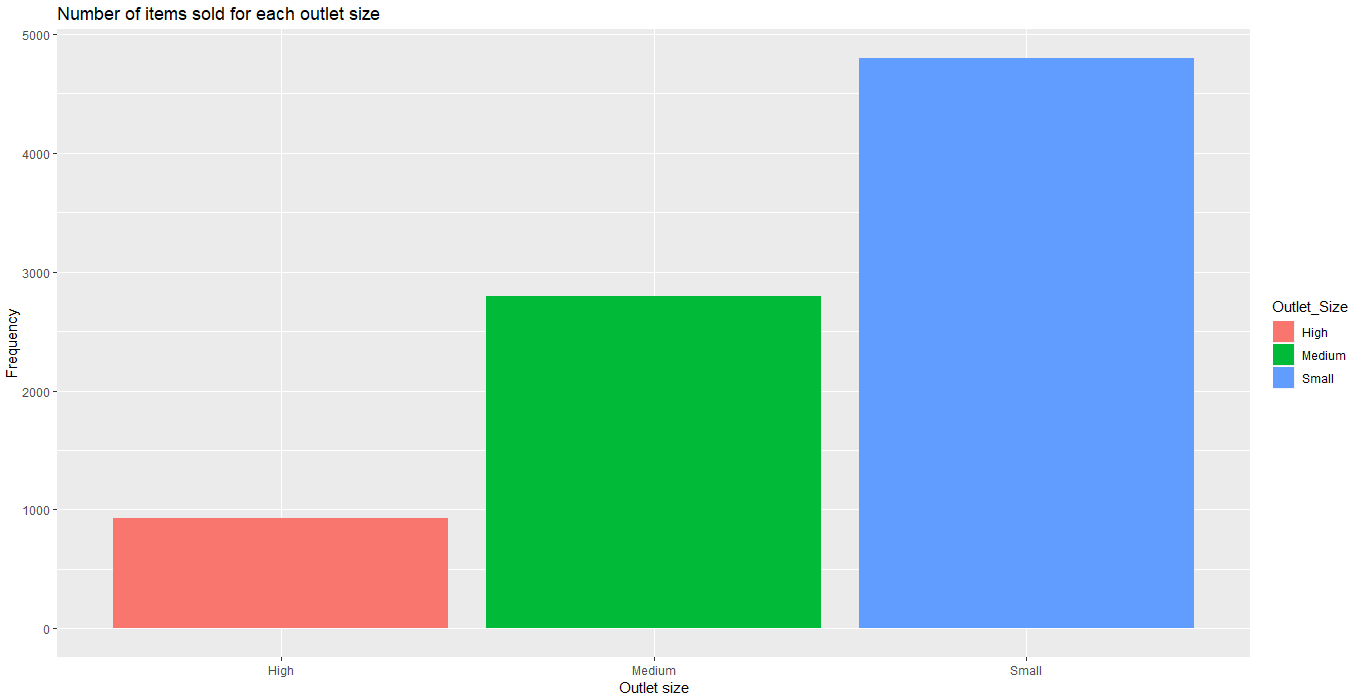
From this graph we can see Supermarket Type 1 records the greatest number of purchases. BigMart can start looking the stats for other outlet type and decide what can be done to increase the sales if possible. Meanwhile it can focus on Supermarket Type 1 because most of the sales revenue is coming from it.

1. From which location people buy most?



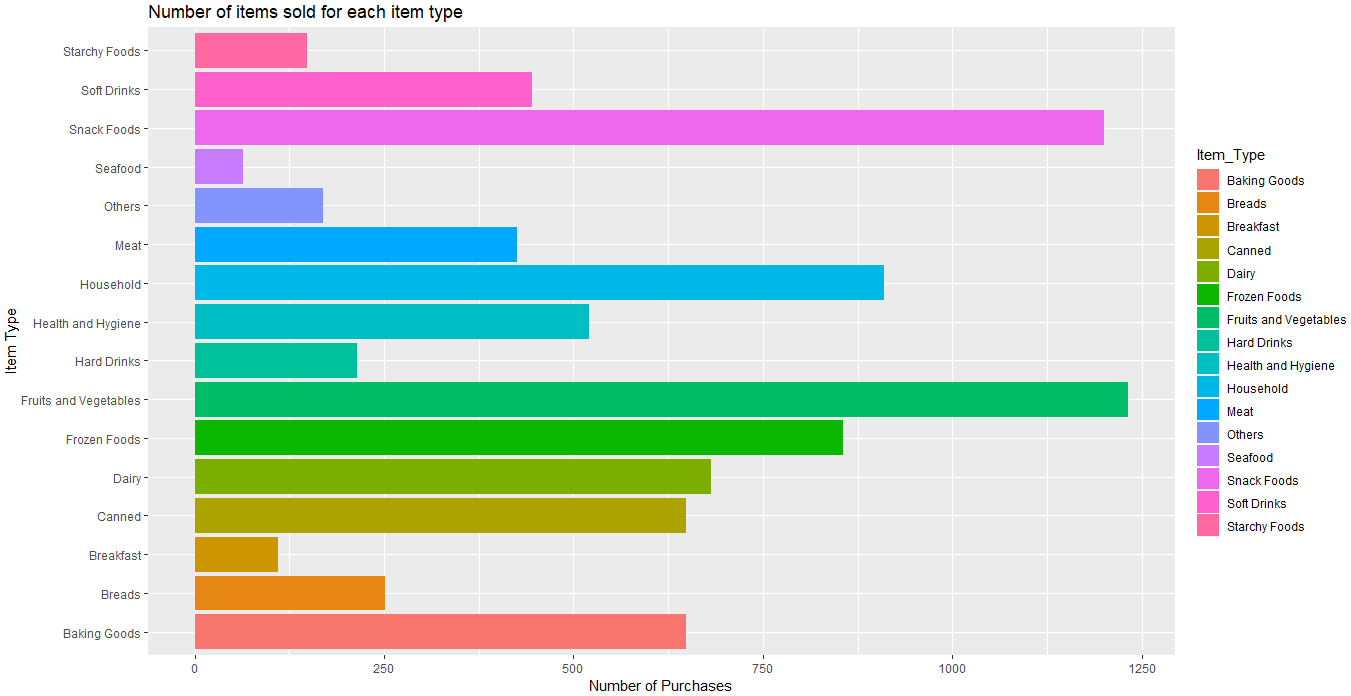
Tier 3 is the location which is contributing most to the sales. This may be because the population. Tier 3 location has a greater number of people, so it is expected to have a greater number of records.

1. Which outlet size records the greatest number of purchases?



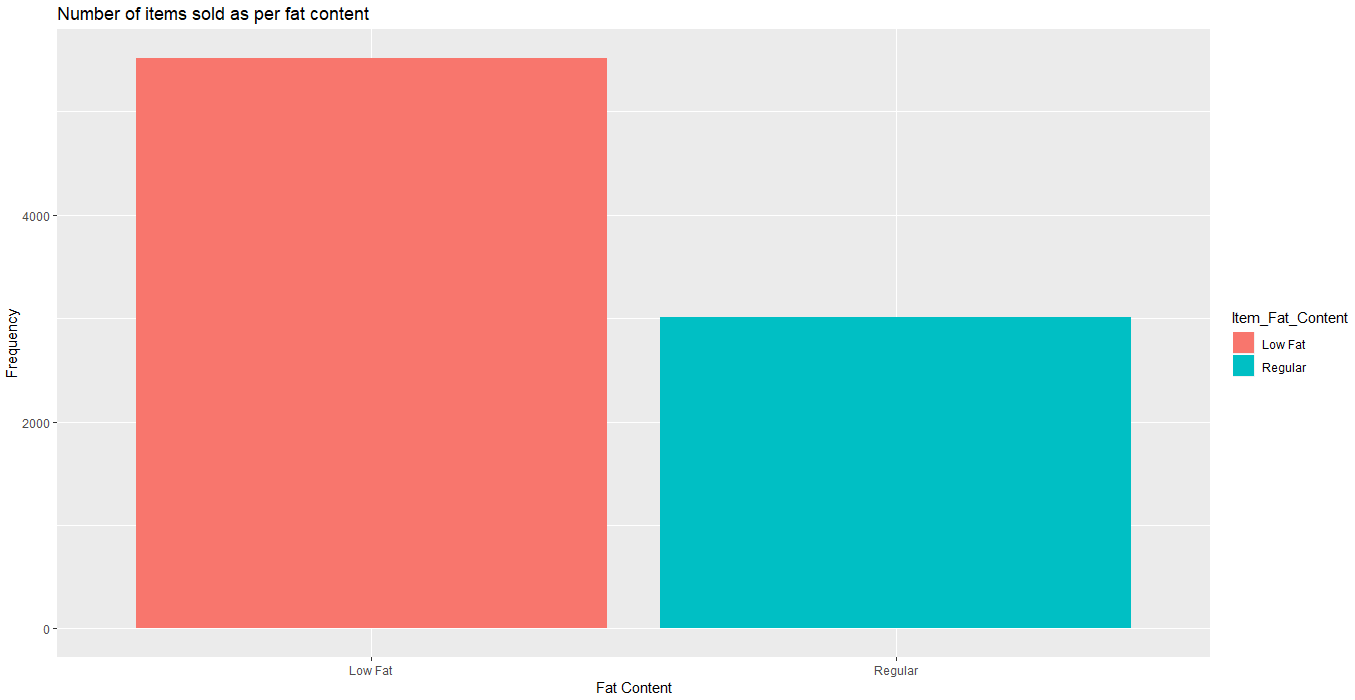
We can observe that small outlet size records most number purchases. With this data BigMart can focus on the small size outlets most because this is were people buy most.

1. Which item people buy most?



From this graph we can see that snacks food and fruits and vegetable are purchased most followed by household and frozen foods. Least selling item from BigMart is seafood. BigMart can start looking this data for its inventory management and plan to increase the sales by providing exclusive discount on products. They can also find why other food like seafood is sold very less and if they can stop selling this item, weather it may save money.

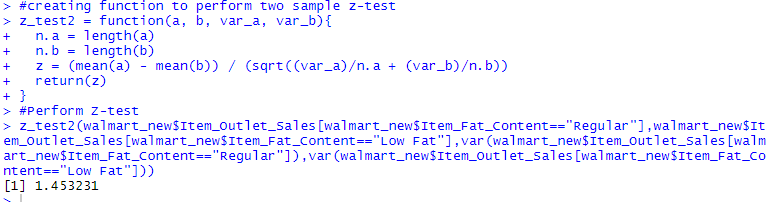
1. What Fat Content people prefer most?

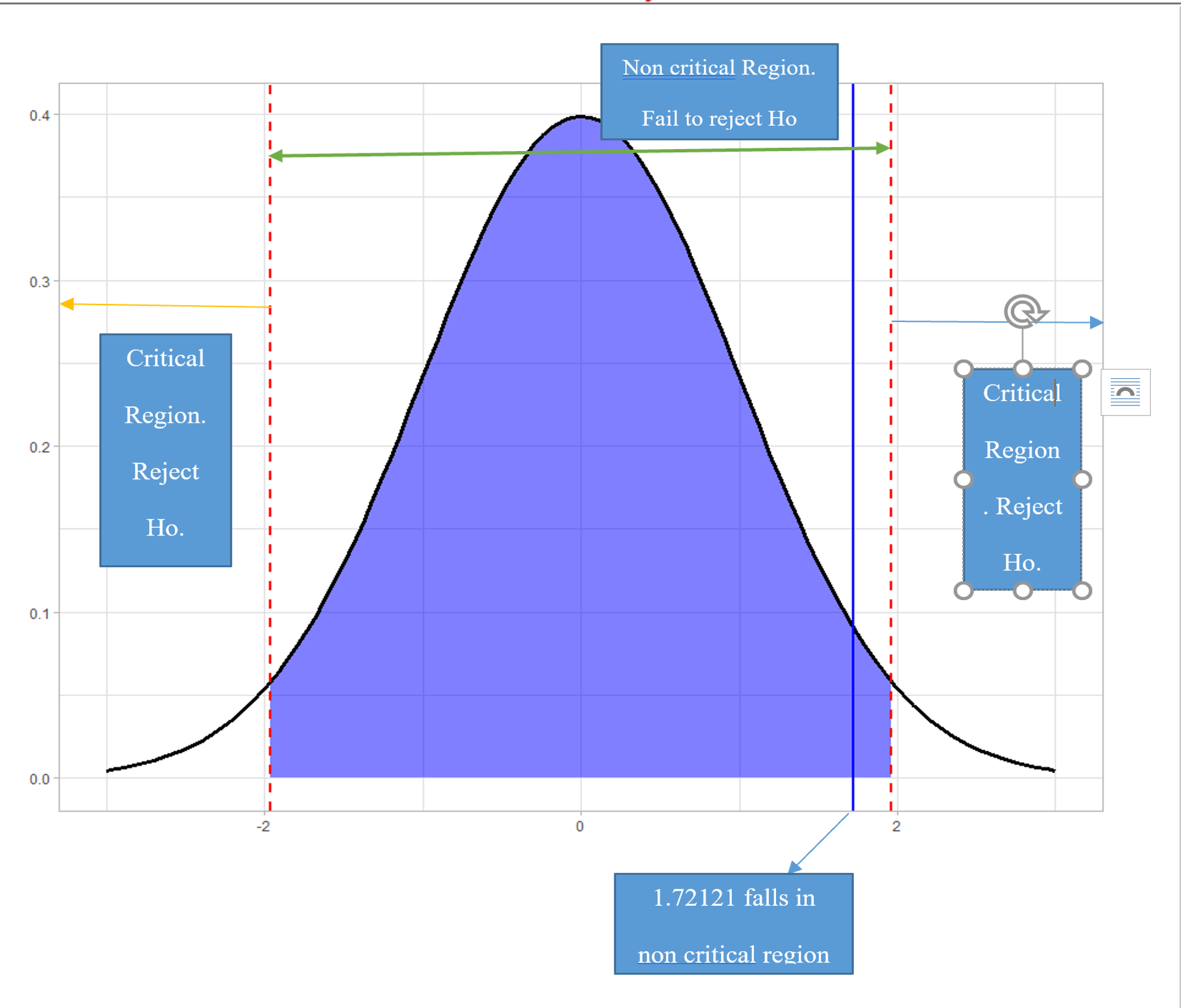


From this graph we can observe that people prefer to buy low fat food item more than regular food item. However, we do not know the if low fat food items contribute most to sales or regular fat items cost more and contributes most to sales. We can run a hypothesis test to check if this there is a difference in mean of sales for low fat and regular items.

Hypothesis Testing

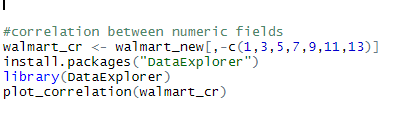
In this case our null hypothesis is there is no significant difference in mean of the sales for low fat and regular food items. Alternative hypothesis is there is significant difference in mean of the sales for low fat and regular food items.

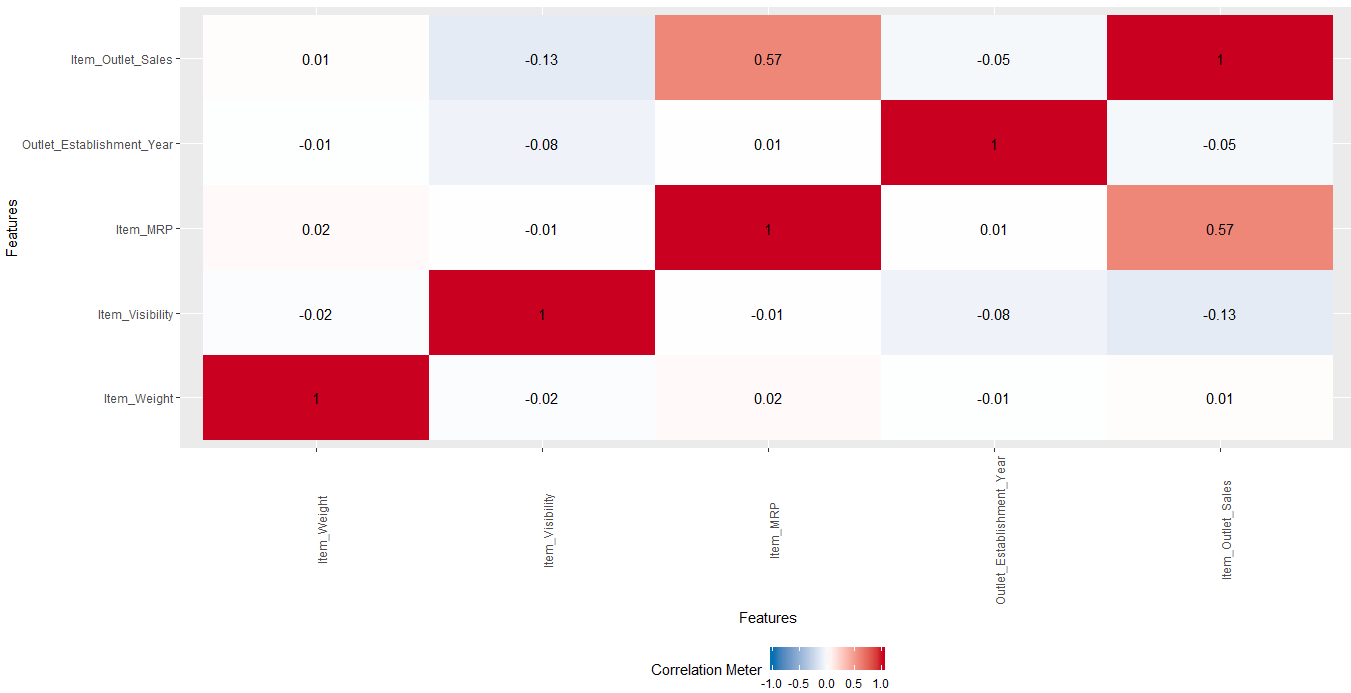




The z-value from this test is 1.453231. If we consider a significance level of α = 0.05, we will fail to reject the null hypothesis because the z-value lies in the range [−1.96, 1.96] (from z-tables). Thus, we fail to reject null hypothesis and conclude that we are 95% sure that there is no significant difference between mean of the sales for low fat and regular food items.

Correlation Test for numeric fields.



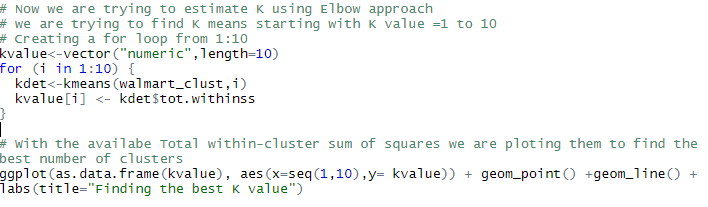


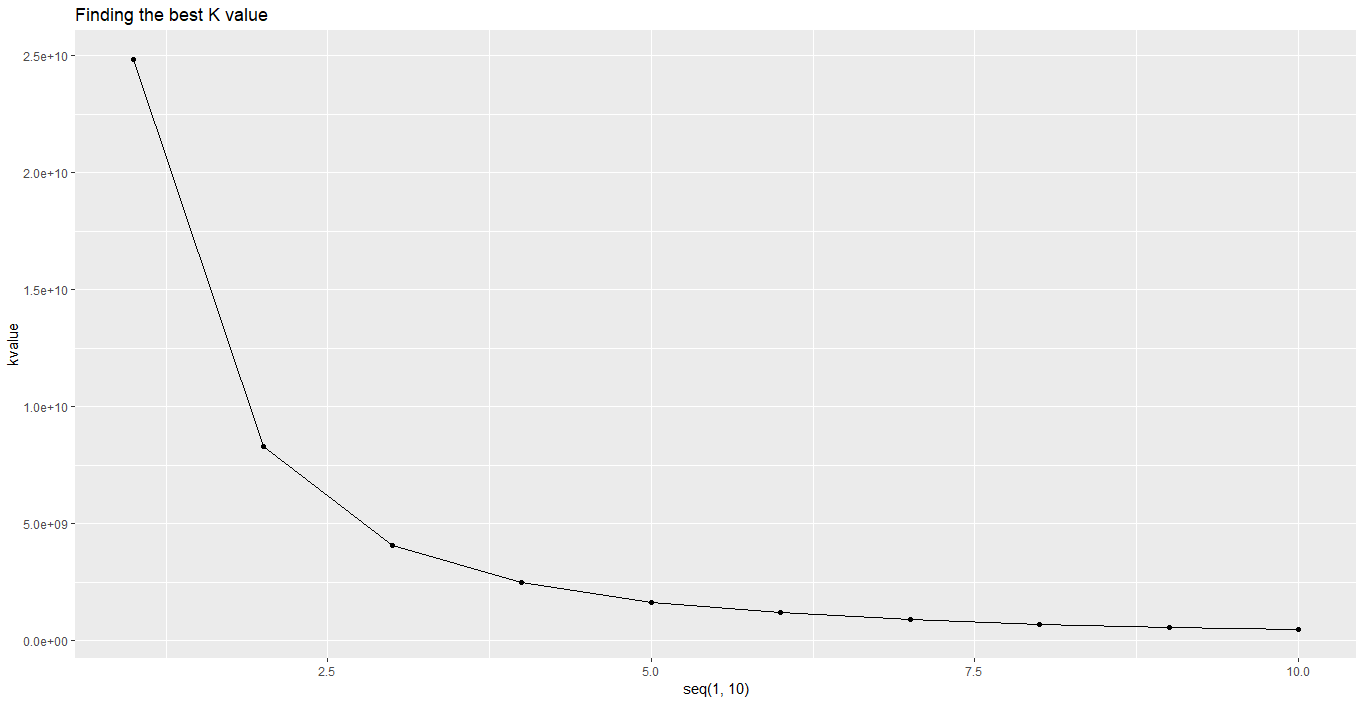
From this correlation matrix we can observe that there is positive correlation between MRP and Sales. There is also negative correlation between Item visibility and Sales.

Clustering for MRP and Sales

Now we are trying to estimate K using Elbow approach. We are trying to find K mean starting with K value =1 to 10.

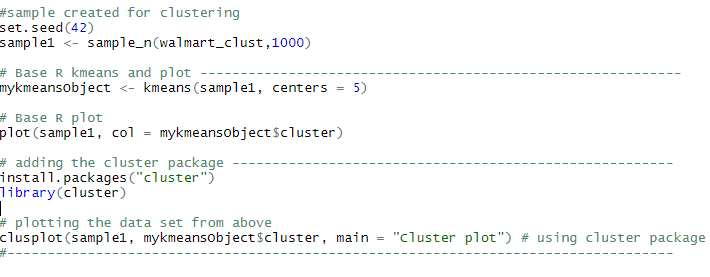
Creating a for loop from 1:10 and plotting the graph.

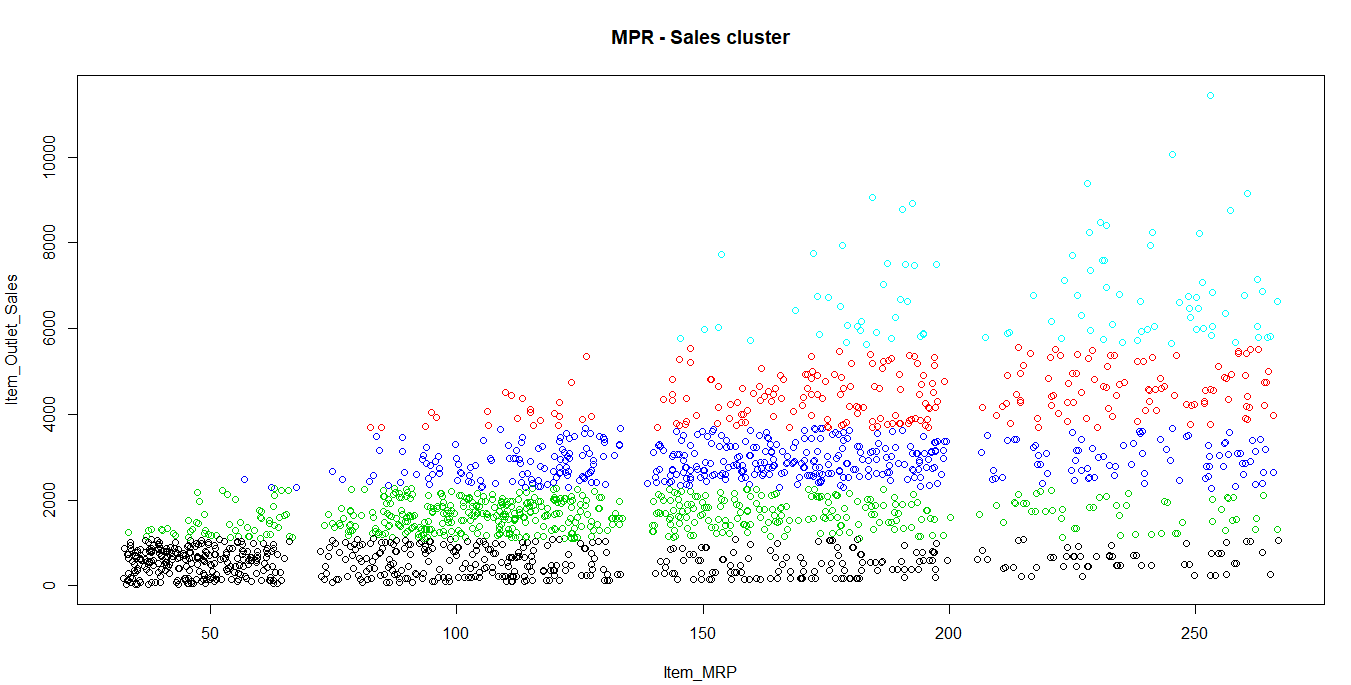




It can be seen from the graph above that a reasonable selection for the K value would be the k = 5. Hence, we are going to create 5 clusters to generate our segments.

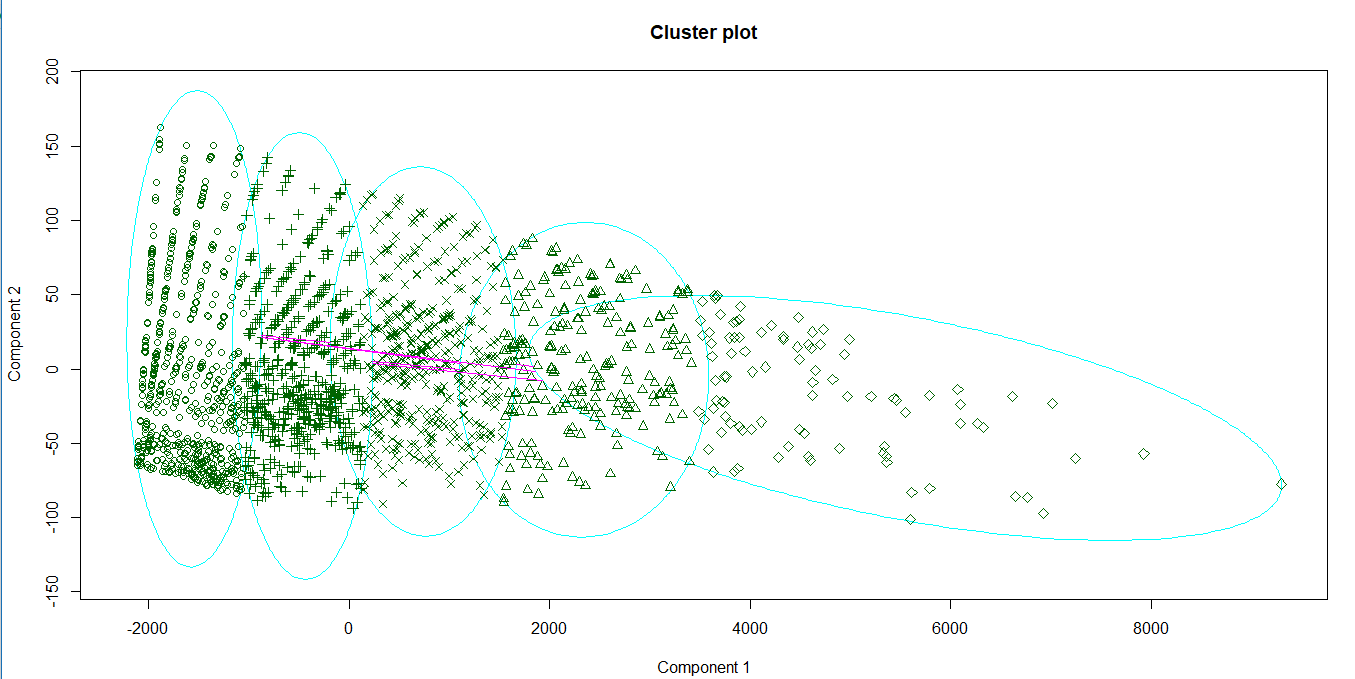
Now we will create a sample of 1000 observation and plot the K mean cluster and use the cluster package to same graph for better representation.





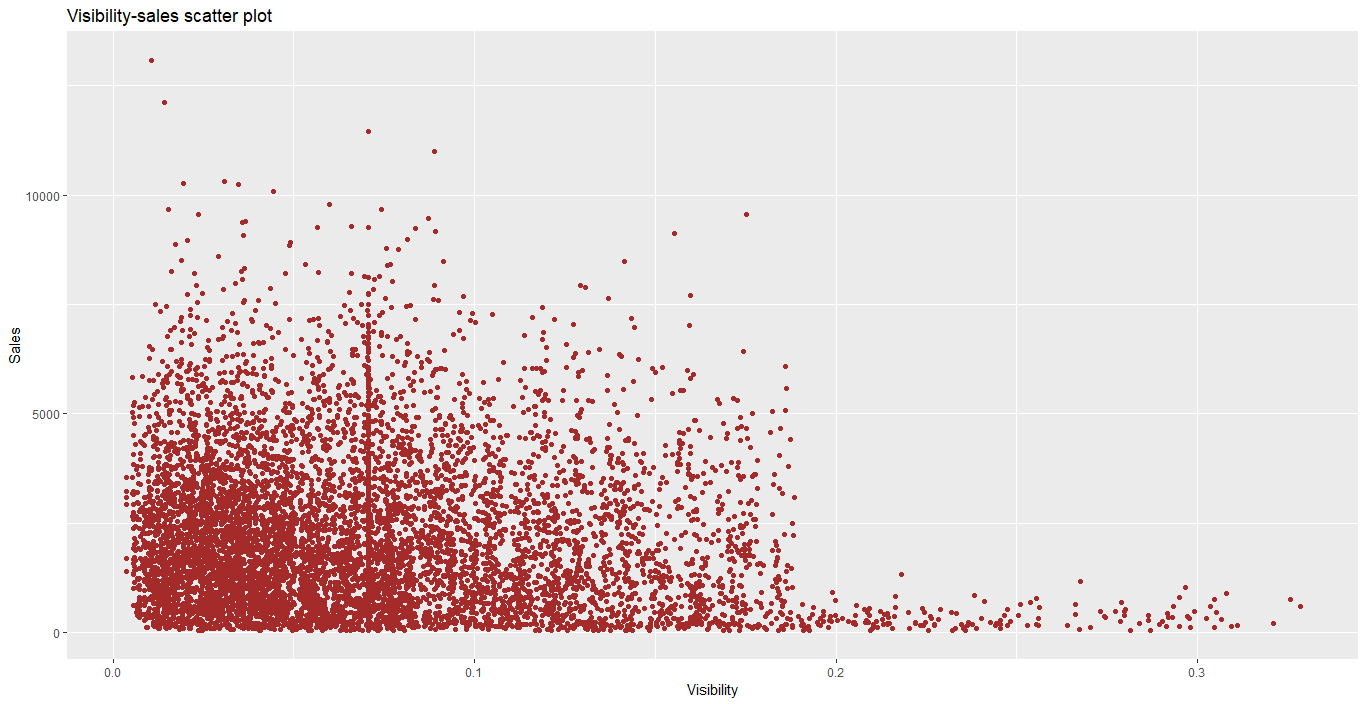
From this graph we can see that R has created 5 clusters which can be seen in different colors. For low sales one cluster is created in black color and for high sales cluster is represented in light blue color. We can also see that number of purchases for low MRP items is more however it is not contributing to total sales more because it is low price items. Concentration for purchases decreases when the MRP of the item increases but the total sales is increasing. This can be because large items are purchased less but it is a very high price item, so the sales are more.

Plotting using cluster package.



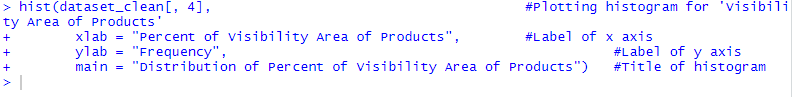
In this graph component 1 is sales and component 2 is MRP. This is a better representation of graph which also shows the region of the cluster. In this plot every cluster is represented in different shapes. Different regions are enclosed in light blue circles.

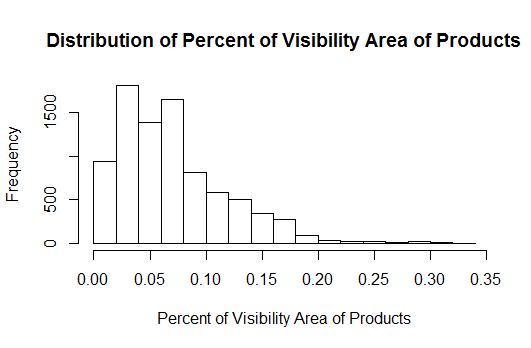
Visibility – sales scatter plot



Visibility is the percentage of display area allocated to product. The vertical line is because we have replaced the 0 value with the mean value of the Item\_Visibility column because if the item is sold its visibility cannot be 0. We can observe here that there is a negative correlation between sales and visibility. When the visibility of the item increases then sales are not increasing. The concentration is decreased when the visibility is increased more than 0.3.

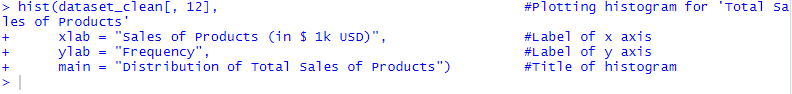
Histogram for distribution of percentage of visibility area of product.

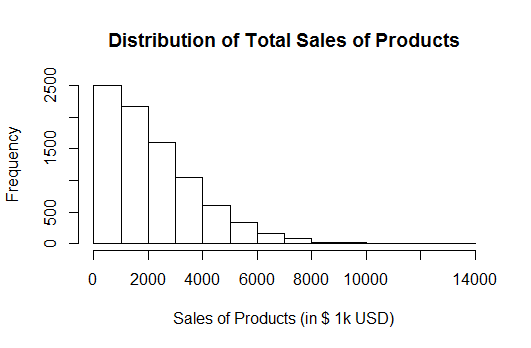




The chart shows distribution of percent of visibility area of products in Big Mart. It can be observed from the chart that most items have a display area between 0 to 0.1. Very few items have a display area percent of 0.2 or more. To improve sales, Big Mart may evenly distribute the visibility area among different items to enhance the sales of low selling products.

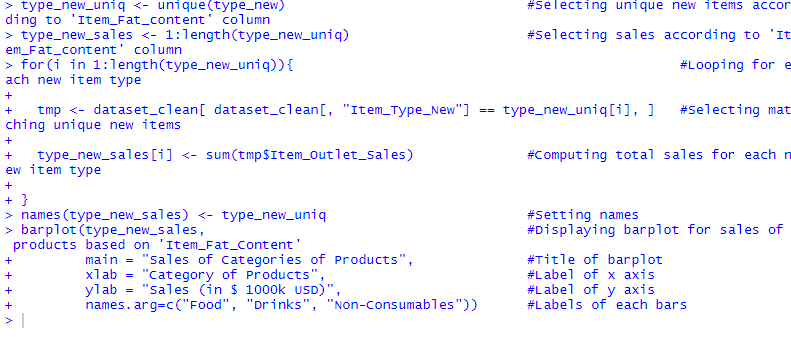
Histogram for distribution of total sales of product

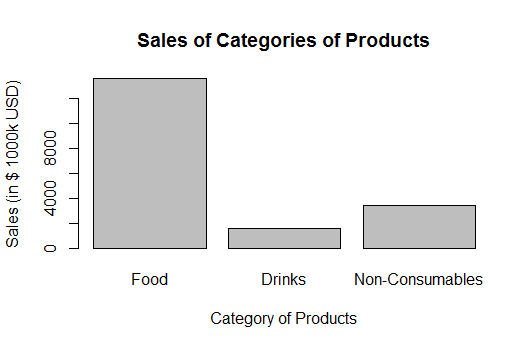




The chart shows distribution of total sales of products in Big Mart. It can be observed from the plot that the distribution is skewed to the right. This indicates that most items have total sales less than $4000k USD. Big Mart may employ different strategies to improve market of products with low total sales. Further, it may find out the causes for high sales of other products.

Histogram for sales of categories of products



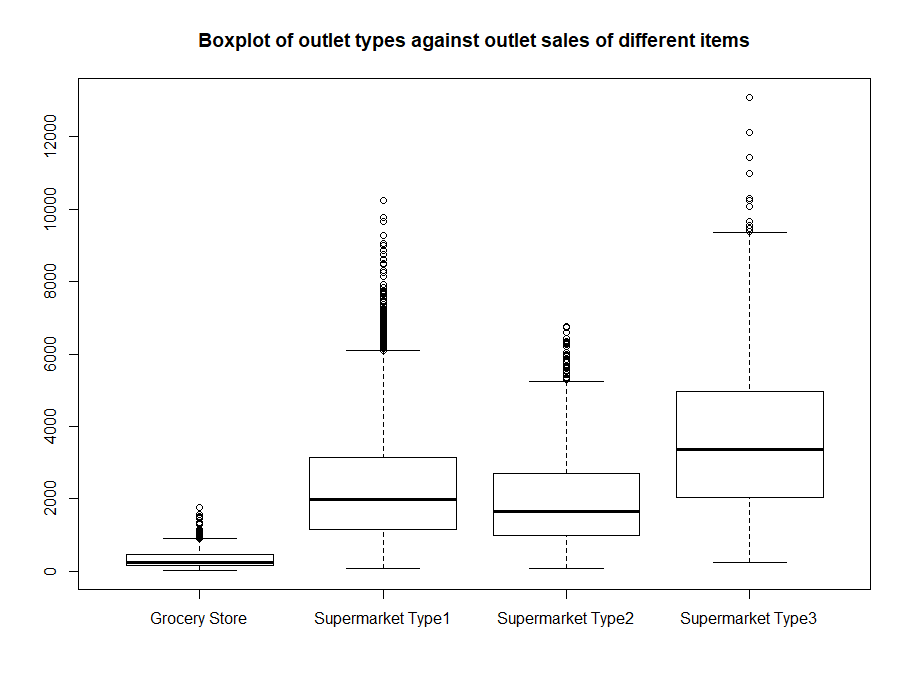


The bar-chart shows sales of different categories of products such as food, drinks, and non-consumables. It can be observed that the sales of food items are the highest, followed by non-consumables, and drinks. Big Mart may promote sales of drinks and non-consumables to enhance overall product sales.

Comparing Means

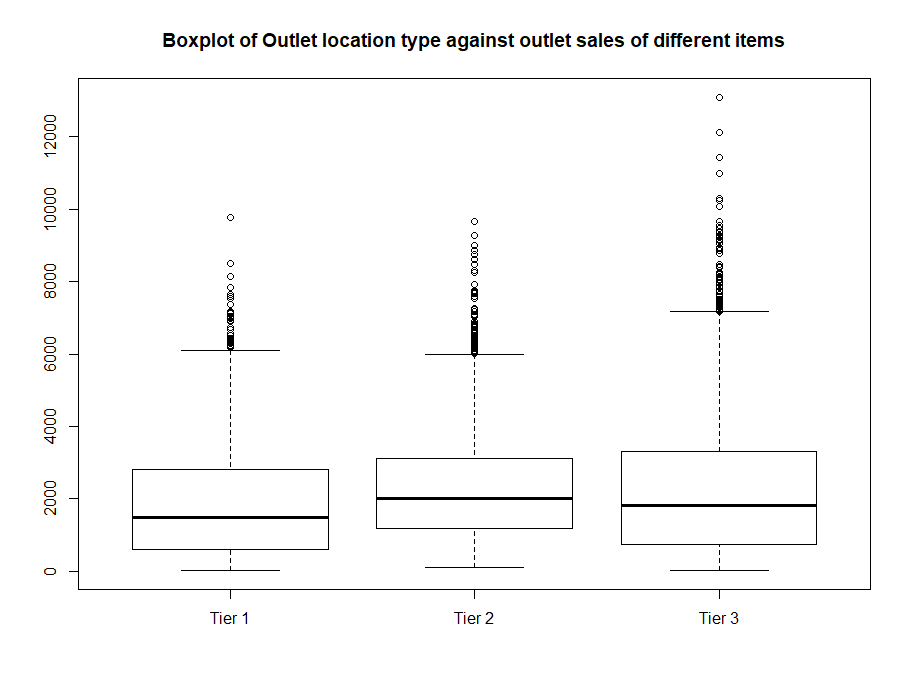
We thought of comparing distribution of sales for different types of market, Location type and fat content of the items using boxplot.

Distribution of sales for different types of market:



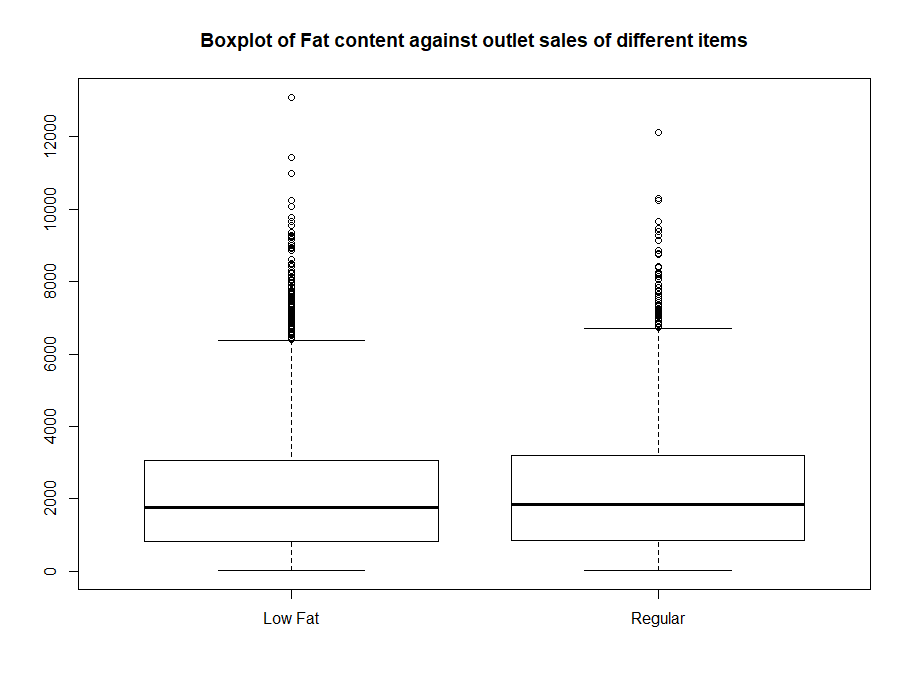
We can observe that there are many outliers in this boxplot. Outliers are more for Supermarket Type 1 because it has a greater number of records of item purchased. We can see that “Grocery store” has the least mean which almost less than 500 may be because the value of items available in grocery store are of less cost. The means of Supermarket “type 1” and “type 2” are almost equal. We observe that “Supermarket type 3” has the highest mean around 3000, we can infer that these types of markets are in well developed areas and the cost of living might be high and it also has the least outliers with items of extreme cost we only find a brunch of people buying the higher cost items even in a well-developed area.

Distribution of sales for different location types where outlets are located:



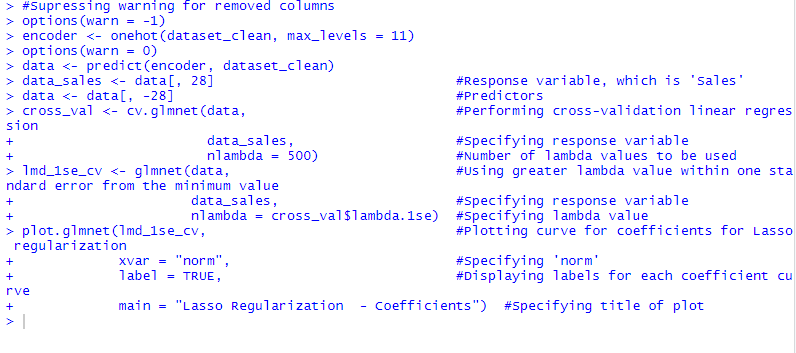
We can observe that the means of all the tiers are almost equal lying in the range 1500-2000. We find more outliers in Location Tier-3, hence, we can infer that the location is well developed, and people earn more as we can see many high cost items have been sold in that location.

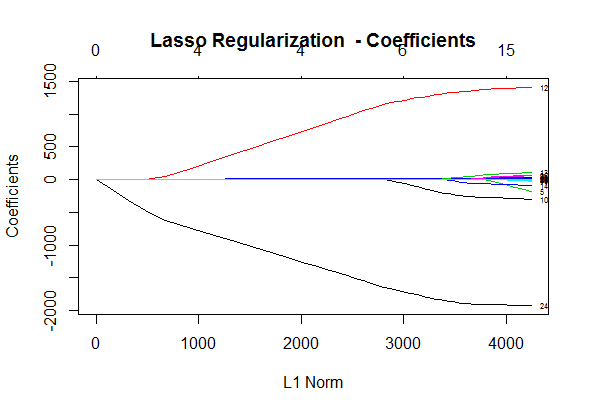
Distribution of sales with respect to fat content of items



We can observe that the means of sales of both low fat and regular fat content items is almost similar very near to 2000. We can see there are many outliers for both the boxplot. This may be because the price of some items is very high, and it contribute more to sales.

Lasso Regularization





The chart shows plot of coefficients for lasso regularization. It can be observed that four of the predictors corresponding to ‘Outlet 18’, ‘Outlet 27’, ‘Grocery Store’, and ‘Item MRP’ have coefficients which converge to zero earlier than other factors. This shows that the abovementioned features have a significant impact on the sales of products. The Big Mart outlets ‘Outlet 18’ and ‘Outlet 27’ appear to have substantially high sales. Moreover, ‘Item MRP’ also plays an important role in determining sales, along with ‘Grocery Store’ outlets. A probable explanation is the higher frequency of customers visiting grocery stores as compared to other outlet types.

Conclusion

Data Cleaning plays important role in Data Analysis Process. Exploratory Data Analysis summarizes characteristics. Correlation shows relationship between numeric fields. From this correlation matrix we can observe that there is positive correlation between MRP and Sales. There is also negative correlation between Item visibility and Sales. Hypothesis test shows there is no significant difference between mean of sales of Low-Fat content items and mean of sales of Regular fat content items. We compared mean for the numeric variable. Lasso regularization shows Outlet 18, outlet 27, item MRP, grocery store are the features which impact sales the most. With these insights BigMart can analyse its sales data and work on improving its performance to increase the sales.

Reference

### Data Source: walmart

#### *walmart*. (2019). *Kaggle.com*. Retrieved 17 May 2019, from <https://www.kaggle.com/gcarra/walmart>

### Data Description: Predicting expected sales for Bigmart’s stores

#### *Predicting expected sales for Bigmart’s stores*. (2018). *Medium*. Retrieved 17 May 2019, from <https://medium.com/diogo-menezes-borges/project-1-bigmart-sale-prediction-fdc04f07dc1e>