Introduction

In this assignment, we analyze and interpret the given data to us and use the appropriate techniques & tools to be able to visualize & understand the data. For this assignment, we were given a dataset called the BigMart Sales Dataset which is one of the popular datasets found on Kaggle website. The data scientists at BigMart have collected 8,523 sales data for 1,559 products across 10 stores in different cities. This dataset includes attributes for each products and stores.

The Data analytics life cycle has a total of 6 different phases and some of these phases are vital in helping us produce a data science report for BigMart. This report will explain whether it is possible to predict the sales of each product at a particular store automatically. This may be achieved by building a predictive model which may help the organization to understand the properties of products and stores which plays a role in increasing their sales.

This report will explain how we are able to reach our proposal and decide which model is the most suitable. Our team have decided to use R as the platform to help us deal with the data preparation process.

For this dataset we will list the dependent and independent variables below

Dependent variable

* Item\_outlet\_sales.

Independent Variables:

The first step that needs to be done is to explore the data given to us and to study them. The initial dataset given has many missing and meaningless values such as NA values for the item weight & outlet size. Another example is the Item visibility attribute which indicates the total display area of all products in a store allocated to the particular product. From the given dataset we can see that several of the product has a value of 0.0000 for their item visibility. This clearly is an error as it means that these products are not visible in store.

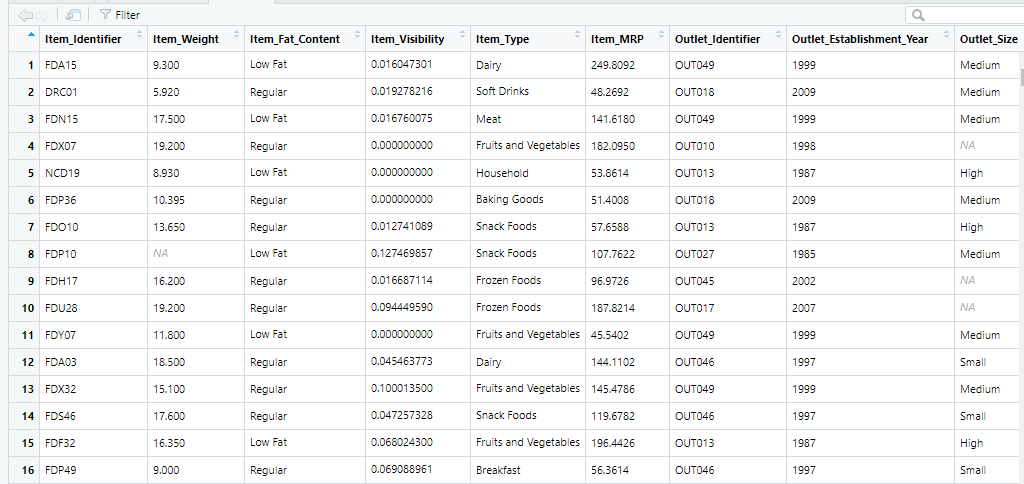
Some of this data even has no value at all as shown in the figure below:



The least we can do is to add N/A on these values which are missing. To do this,we can include the following commands:

big <-read.csv(file.choose(),header = T, row.names = NULL, na.strings = c(""))

Now when we look at the data, we can see they all have NA’s

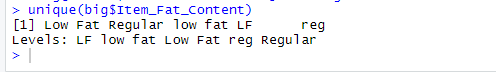


This indicates that the data must be cleaned and prepared for us to have an accurate result in the model.

First, we notice that the Item\_fat\_content has been labelled incorrectly, we have Low fat which is also coded as low fat or LF and Regular which is coded as reg.

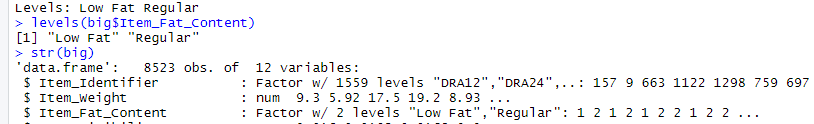
So, we can use the command

unique(big$Item\_Fat\_Content), we see below

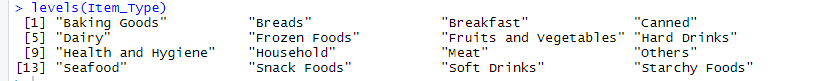


We can fix this issue by renaming LF and low fat -> Low Fat and reg -> Regular using below command

levels(big$Item\_Fat\_Content) <- list ('Low Fat' = 'LF','Low Fat' = 'low fat','Regular'='reg')

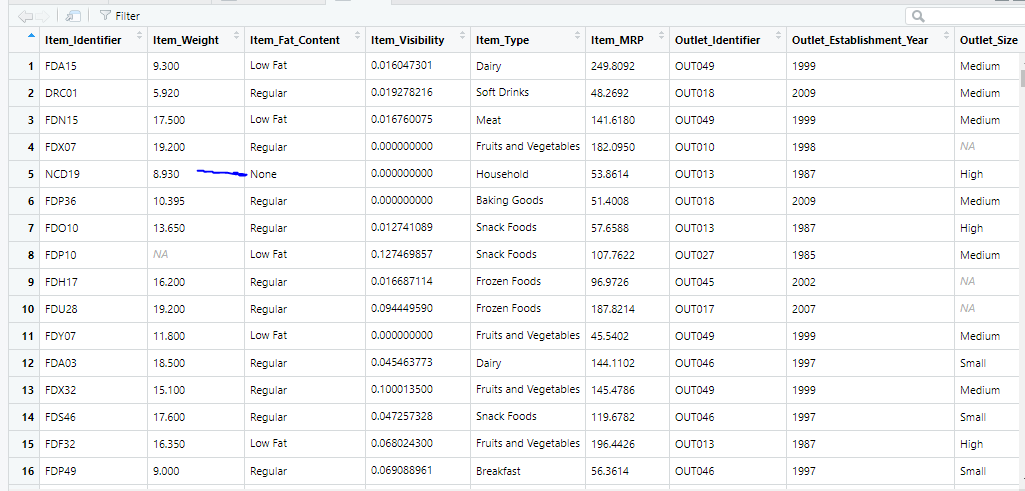


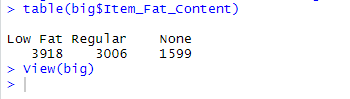
As we can see that the Item type has factors that are none-food, it would make no sense that the item fat content is just Low Fat and Regular as none-food item should contain no Fat.



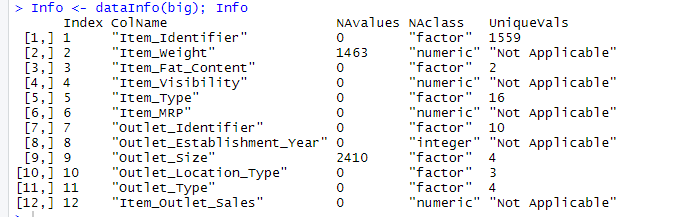
To fix this, we will create a new level for item fat content that will contain nonfood items. We will name this “None”.

Looking at item type, we can see that “Health and Hygiene”, “Household” and “Others” should not have a fat content so we will change their fat content to none below.

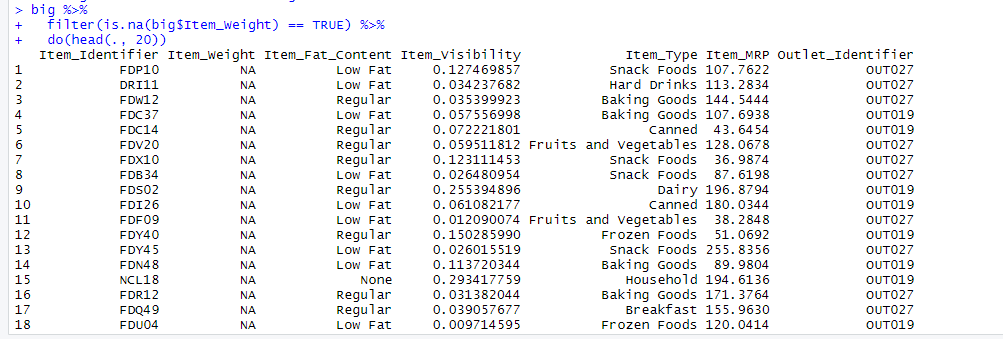




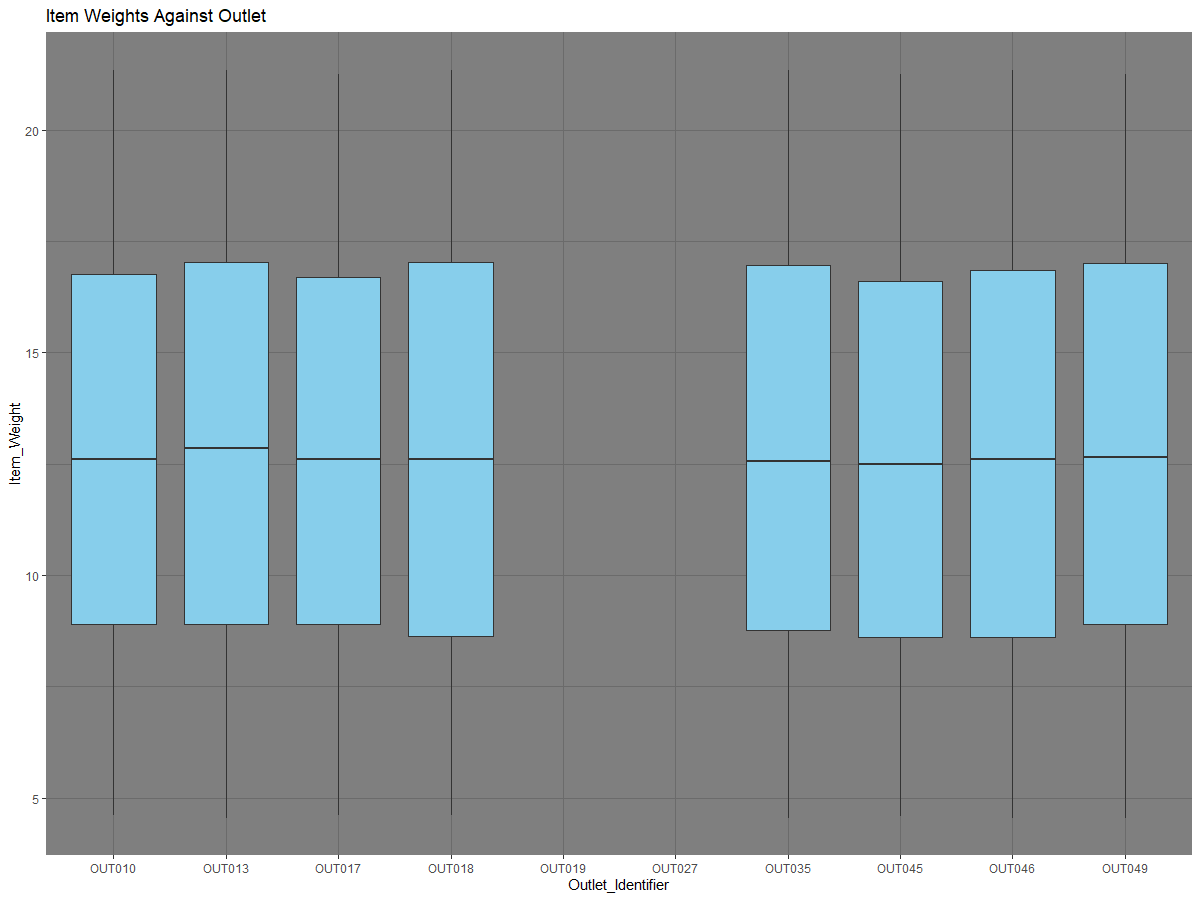
Now looking at missing values. There is a function to get an overall look at the variables and its missing data. Here we can see where we need to perform data cleaning.



As shown in the figure we can see that Item weight and outlet size has missing values. In this case we have a total of 3873 missing values, it would not be ideal to remove them from the dataset as it would reduce our dataset. We will first deal with Item\_weight. We can use Dplyr to filter where the NA’s are coming from.



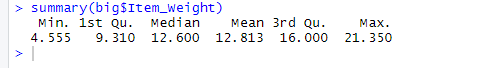
As seen above, we noticed that most of the NA’s are coming from outlets OUT019 and OUT027. We can see a clearer picture if we plot the item weights against outlet identifier.



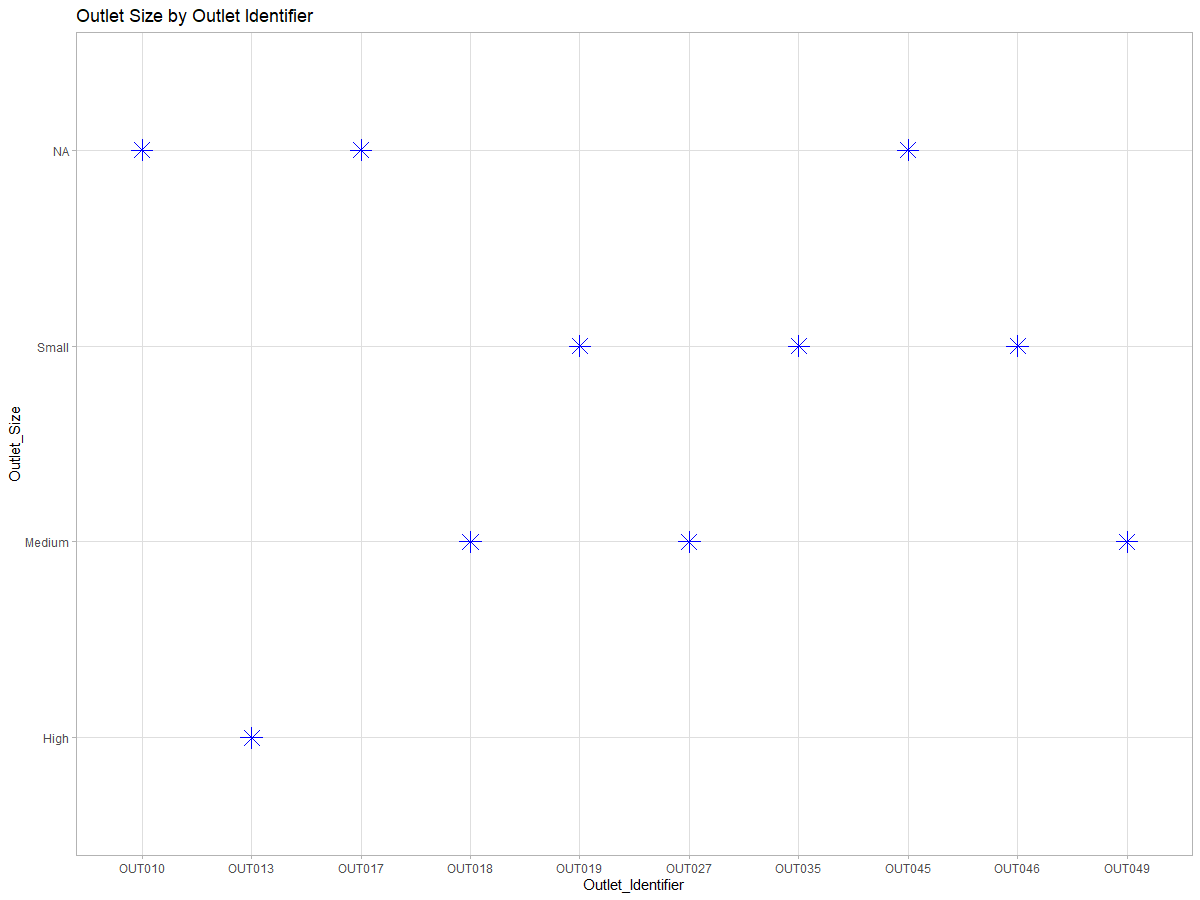
As can be seen above, the outlets 19 and 27 has just not reported any weight information on their items. Also, it seems the distribution of weights does not vary across outlet locations so we can impute the missing values with the median. Here is a distribution of the item weight before imputation.



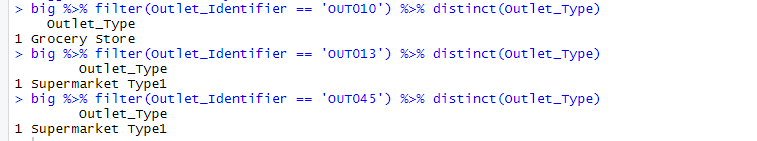
After the imputation we can observe the distribution as below.



We will now observe the second variable with missing values which is the outlet size. We can plot this against outlet identifier to find out where the NA’s are coming from.

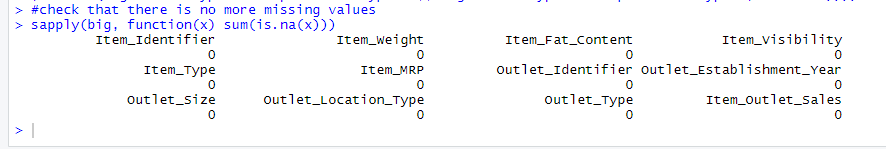


As can be seen from the diagram, NA’s are from outlet 10, 13 and 45. We can now filter the outlet type from the missing outlets.

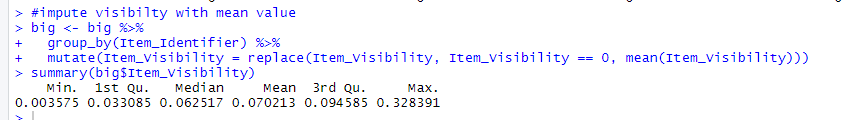


All we must do now is impute the missing values with the mode of the given outlet type.

We can now check to ensure there is no more missing values



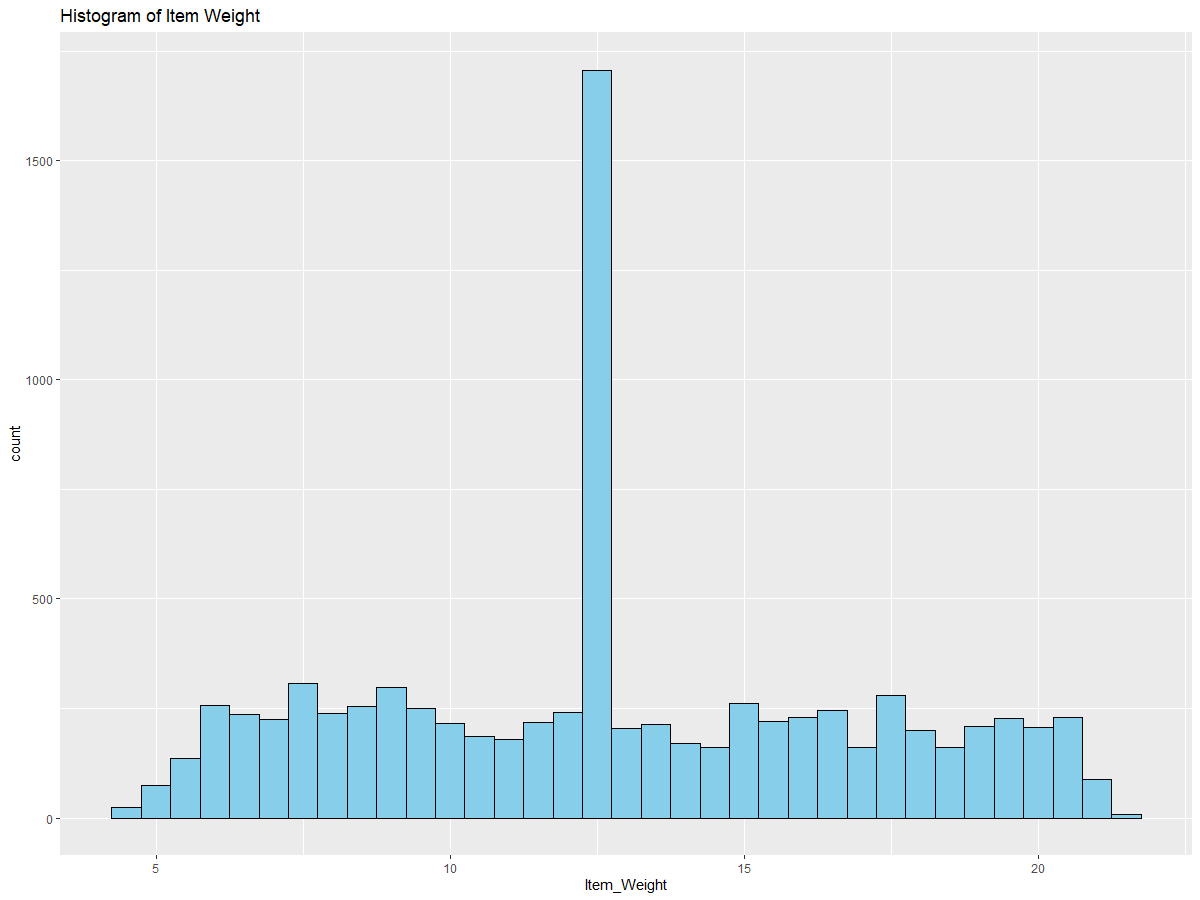
Lastly, we can observe that the visibility has some zero values which is not right so we will impute that with the mean visibility value.



Phase 2 is the data preparation phase. In this phase, we first learn about the data, Check for Null values and see the data and load it in our RStudio. Then, will be performing some data conditioning such as clean and transform the data. We will also identify the target column and talk about it.

Next phase is to perform some exploratory data analysis. We will address each variable one after the other.

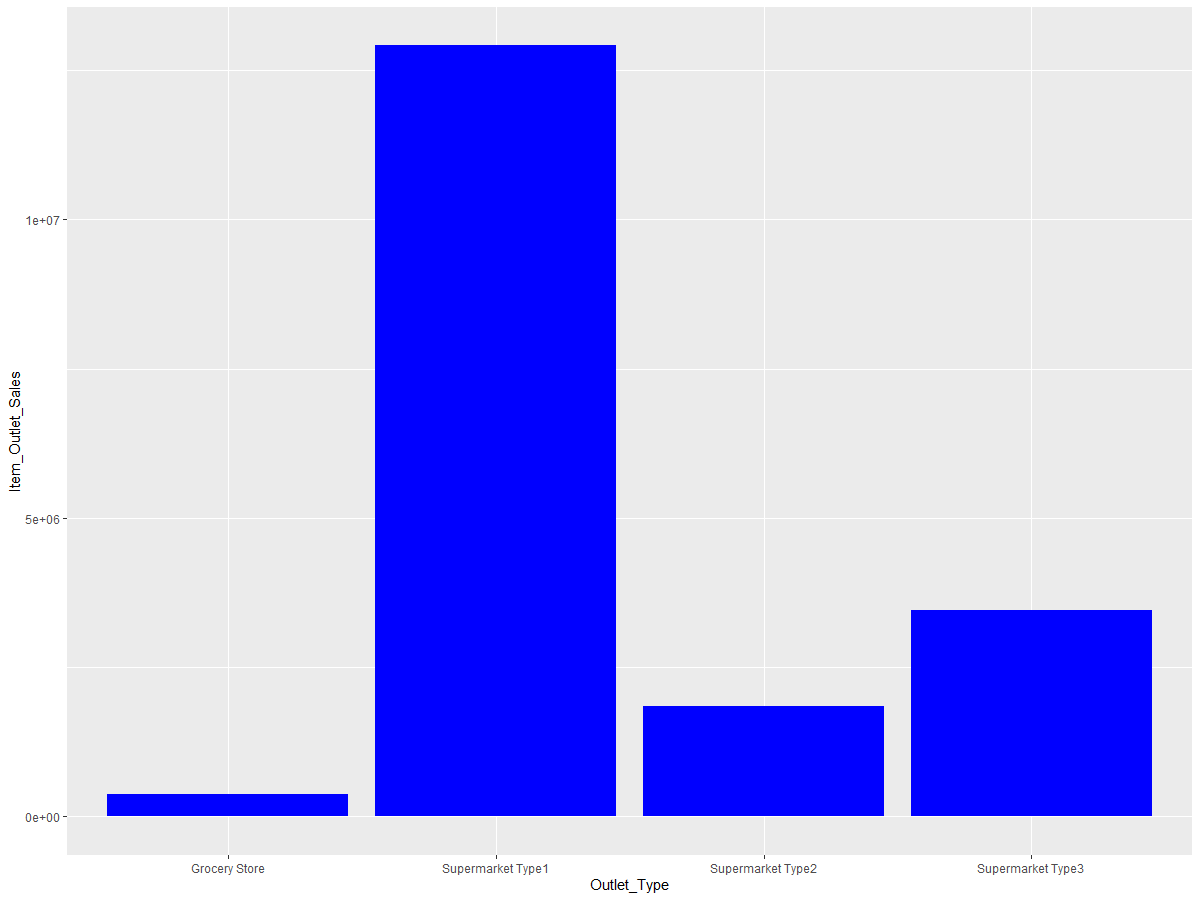
Let us take a look at the item weight. We can use a histogram to quickly explore this.



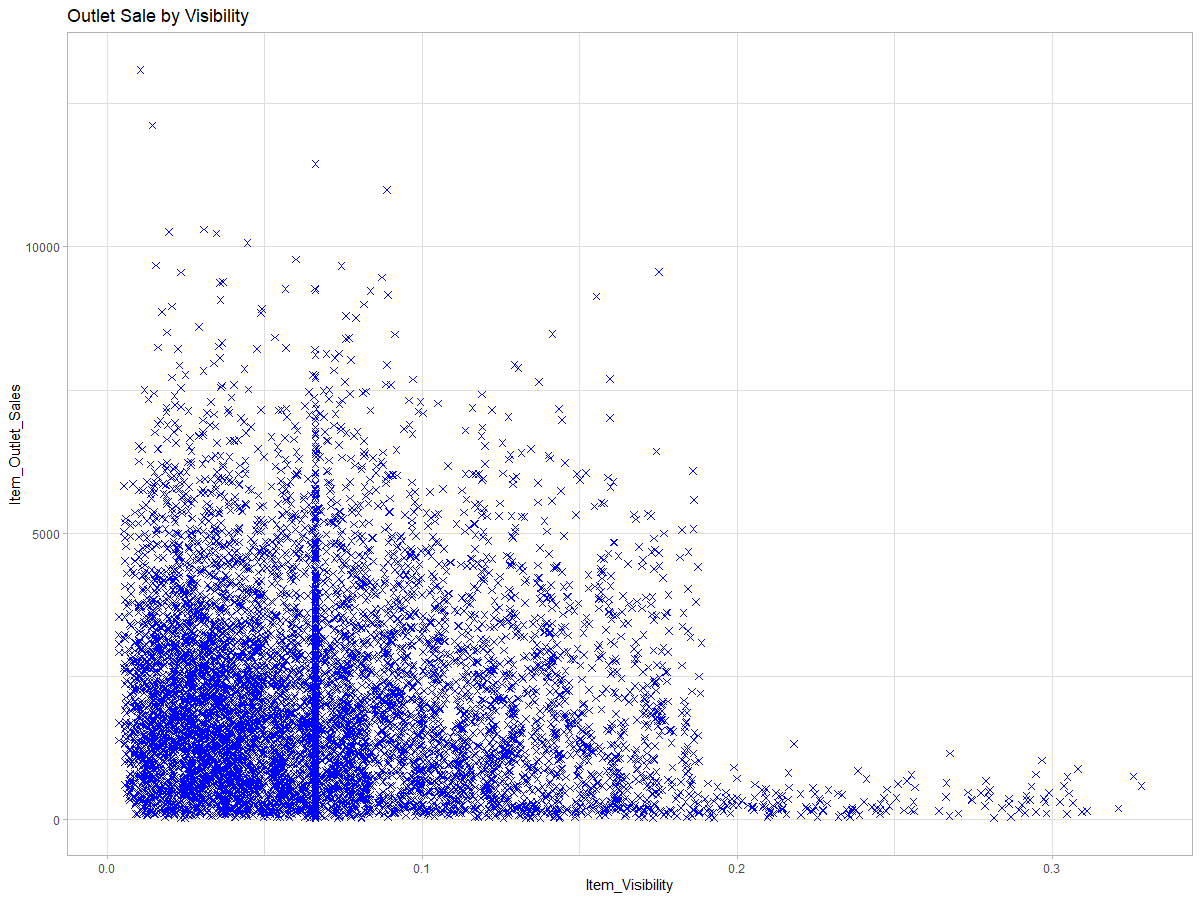
From the histogram, we can see that the graph has no skewness. This is good because we do not have to worry about any further transformation. The long bar indicates the median imputation we performed earlier.

Hypothesis testing.

We will develop 5 store level hypotheses based on our findings. The first hypothesis would be that supermarket type 1 has highest sales. From the figure below we can confirm this.



2. Items with lower visibility has a high number of sales. As can be seen a lot of sale has been made for items with visibility less than 0.2



3. Another interesting hypothesis is that items with a higher visibility have a Low-Fat content. I think this is because of the health benefits of low-fat products. As shown in the diagram, the higher the visibility, the lower the fat content.

A screenshot of a cell phone

Description automatically generated

Notes:

1) Item identifier

No error value

Distribution of work done by each member

|  |  |
| --- | --- |
| Michael | Jefferson |
| 50% | 50% |