**Contents**

* Abstract
* Introduction
* About Dataset
* Data Cleaning
* EDA (Exploratory data analysis)
* Analysis
* Conclusion
* Reference

**Abstract**

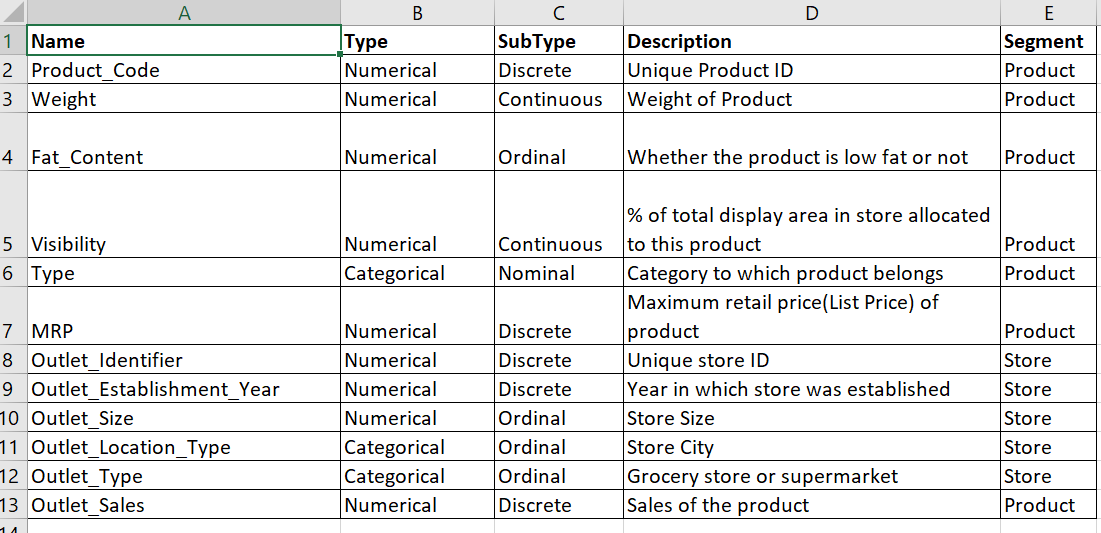
Thereport briefs us about the analysis that we have performed on Walmart dataset. We are using three techniques hypothesis testing, lasso regression and clustering to analyzing the sales with various other fields as a predictor. In the first section we have cleaned the data to remove all the anomalies, second is the analysis part in which we have analyzed the data and performed several techniques using R. The last section is the result of the all techniques which we have used. Several business question can be answered for the future scope of this analysis.

**Introduction**

Walmart Inc is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores, headquartered in Bentonville, Arkansas. Our Goal is to understand how the sales are impacted by various features of Walmart. To understand the impact of various features on sales of Walmart we have used three techniques which will be useful to understand the other factors impact on Walmart sales and thereby used our techniques to boost the sales. We have used hypothesis testing, lasso regularization and clustering techniques using the R platform. We have used onehot, cluster, ggplot, glmnet and dataexplorer packages for our analysis in R.

**About the Dataset**

The Dataset contains 8523 observations and 12 features



**Data Cleaning**

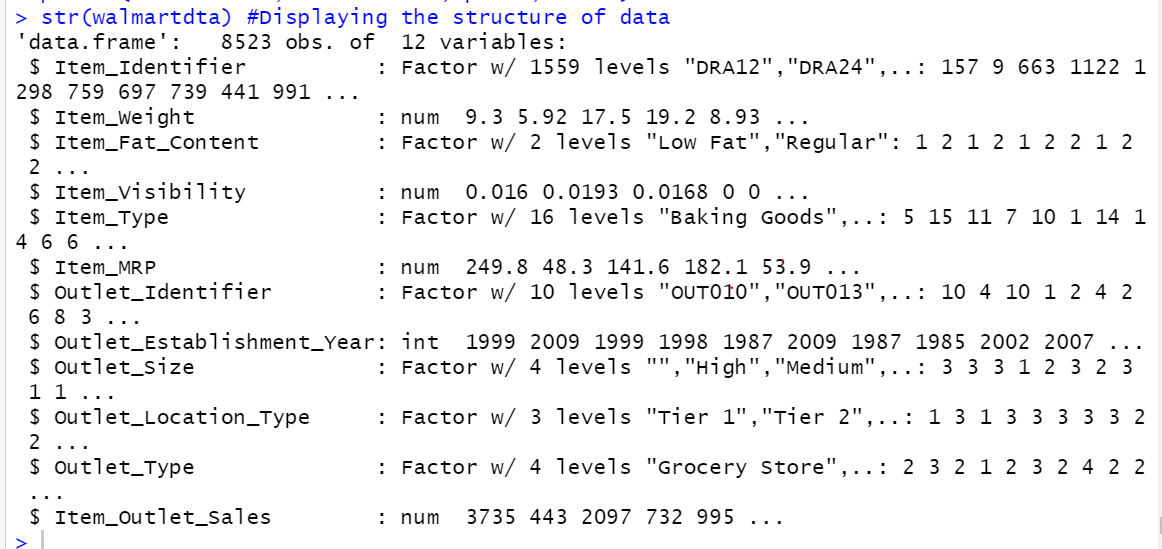
Data cleaning is a process of cleaning, detecting the anomalies like unused variables, dummy values, absence of data, violation of business rules, cryptic data and contradicting data.

Data Cleaning is one of those things that everyone does but no one talks about. However proper cleaning of data can make or break your project. Professional data scientist usually spends a very large portion of their time on this step.

Better data beats fancier algorithms.

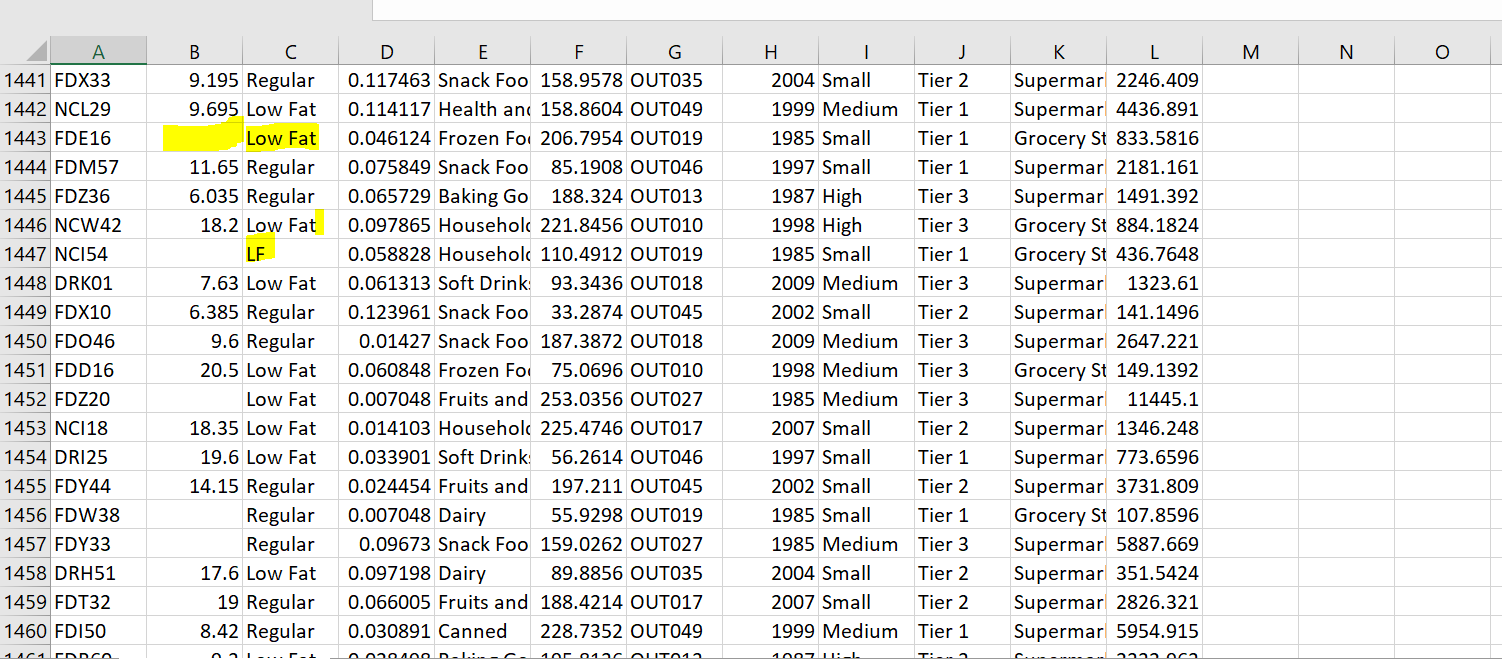
In fact, if you have a properly cleaned dataset, even simple algorithms can learn impressive insights from the data. Obviously, different types of data will require different types of cleaning. However, the systematic approach laid out in this lesson can always serve as a good starting point.

The structure of dataset:



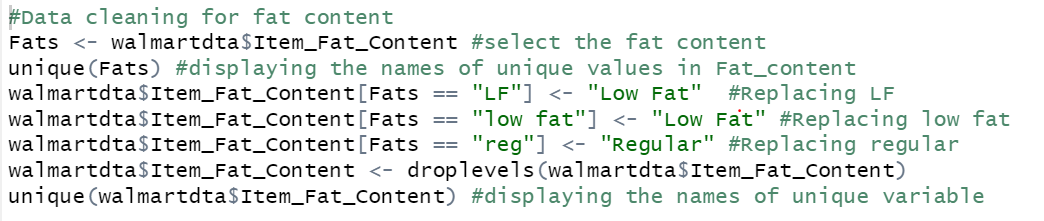
Weight, visibility, MRP and outlet sales are the numerical variables while the rest are categorical variables.

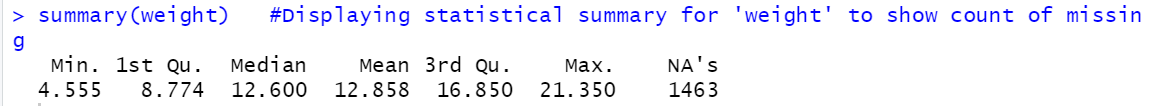
Dataset in our Excel sheet before cleaning:



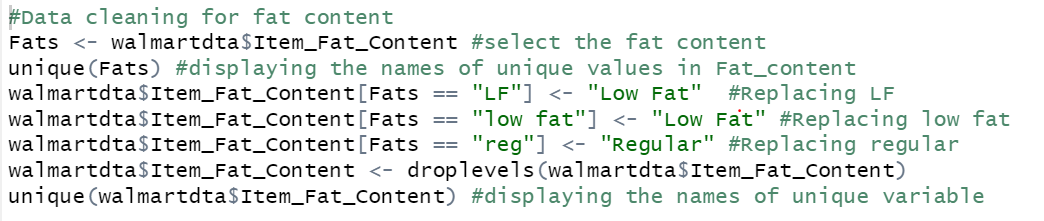
As you can see from the structure and the unclean data there are three values of low fat (Low fat, Lf, low fat) and regular has two values (reg and regular). Using the following code in R we categorized this data into Regular and Low Fat.

We fill the empty cells in weight column as every item has some weight.





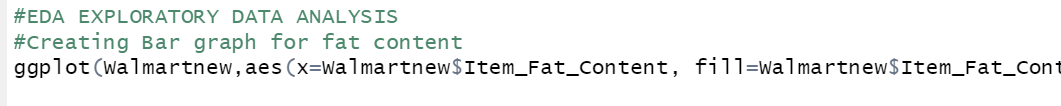
There are 1463 NA values which we will be replaced with mean of non-NA values.



After the data is cleansed, we exported the data into new Excel to perform the tests for accurate results.

**EDA (Exploratory Data Analysis)**

**Bar graph for fat content**

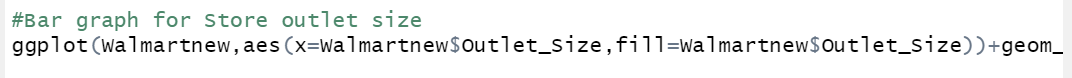


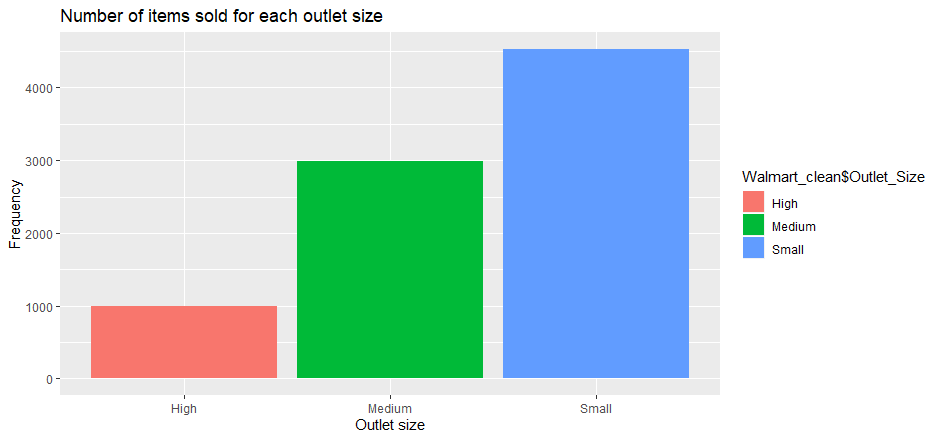
A picture containing screenshot

Description automatically generated

From the fat content, we can interpret that people prefer Low fat food more than regular foods.

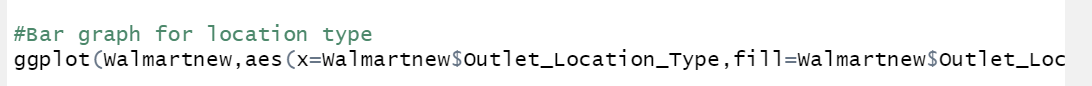
**Bar graph for store outlet size**

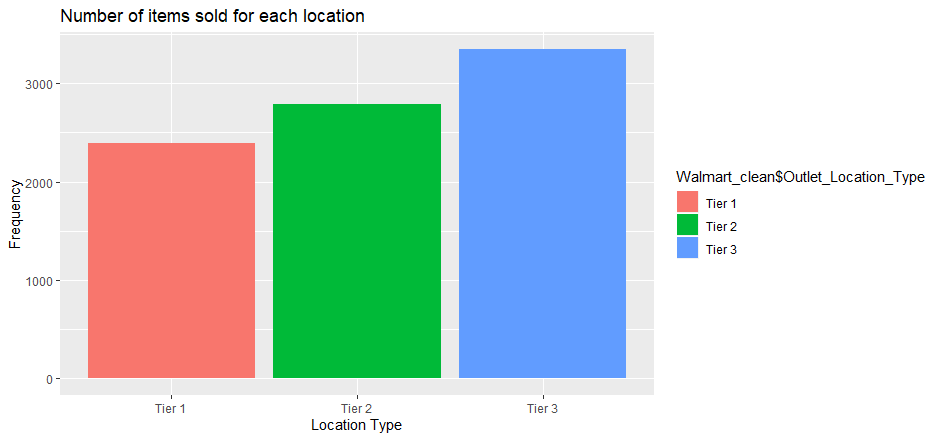




From the above graph e can interpret that the small sized outlets are generating more turnover than the medium and high sized outlets as people prefer BJ’s and Costco for the wholesale purchasing.

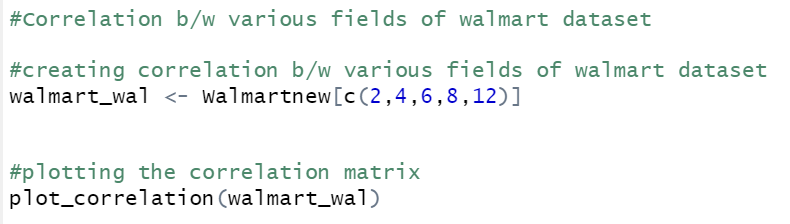
**Bar graph for location type**

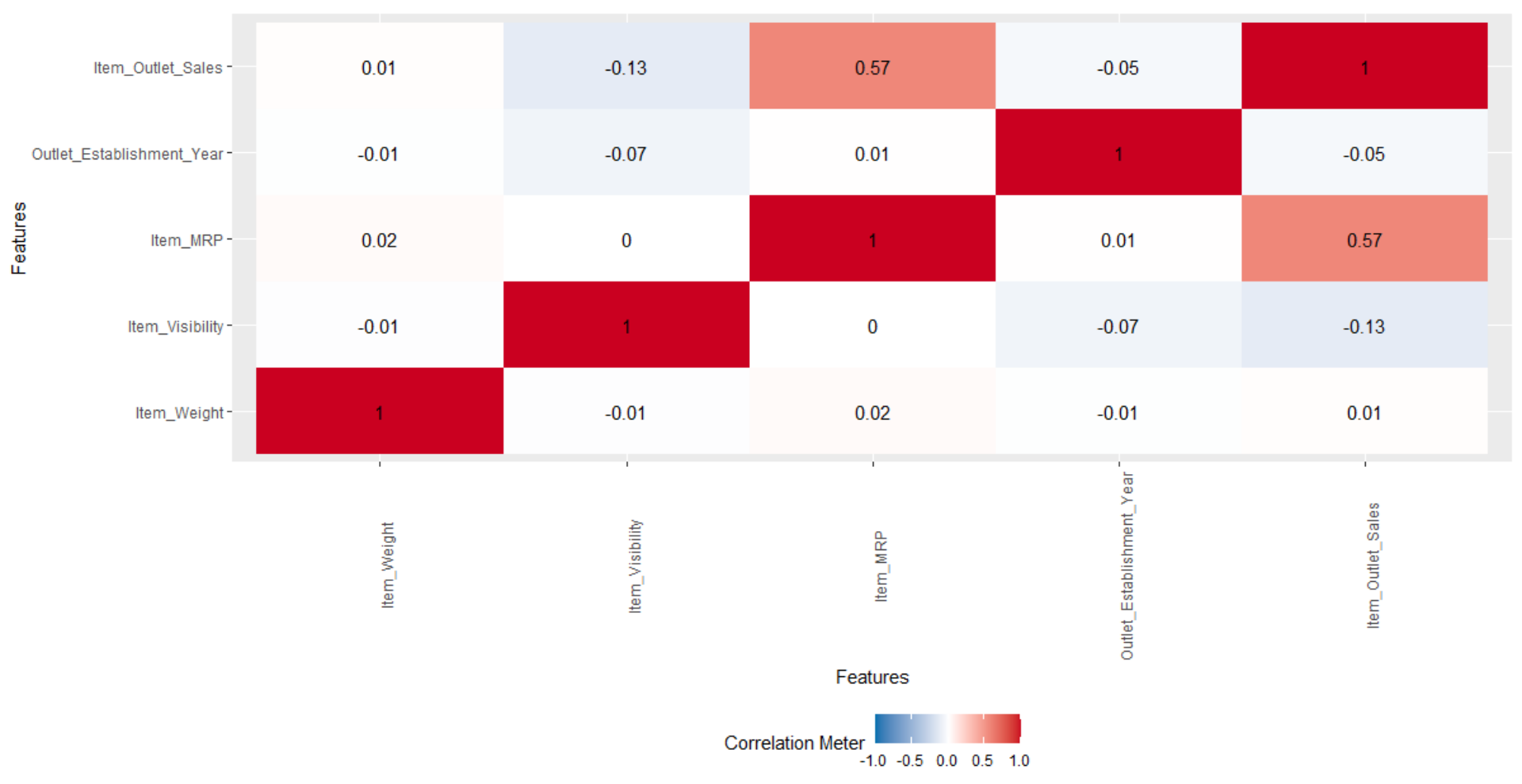




The classification of the tiers is based on the population in the location indicating that tier 1, tier 2 and tier 3 are big, medium and small cities respectively. Sales in tier 1 and tier 2 are low when compared to tier 3 as the location of Walmart are generally outside the city limits.

**Correlation between various fields of Walmart dataset**





The above matrix is a correlation matrix which is used for summarizing the data. It is matrix formed between the features of the data to tell how correlated one feature is with the other. From the matrix we can infer that outlet sales and Mrp are highly correlated. Depending on the outlet size purchases are made. We observe that outlet sales and outlet establishment year have a weak negative correlation with the visibility.

**Analysis**

**Hypothesis Testing**

Hypothesis testing refers to formal procedure to accept or reject statistical hypothesis. The best way to determine whether statistical hypothesis is true would be to examine entire population. Our main aim is to determine whether Walmart should increase the variety of low-fat items rather than regular items. For this instance, we will compare the variances of sale of both categories of product and will decide if the sale of low-fat items are more than regular items. For this purpose, we are using F test to determine test statistics.

So, my null and alternative hypothesis are as follows:

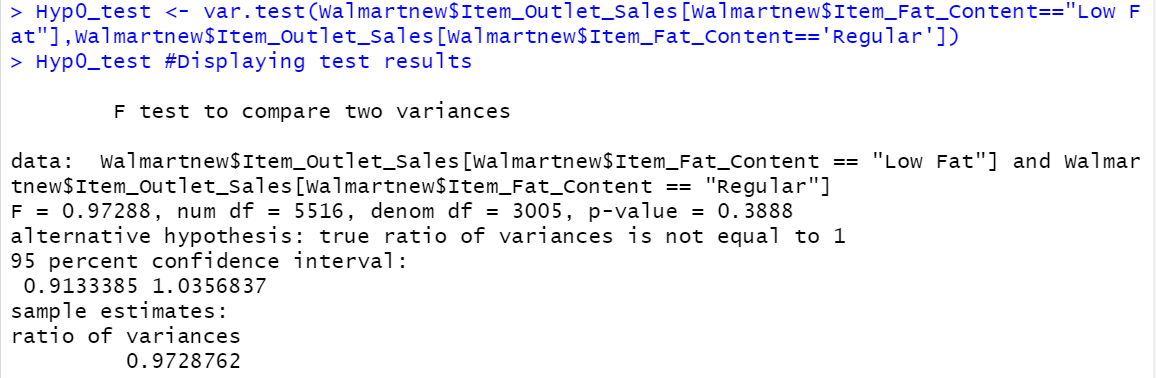
Null Hypothesis: Ratio of variances of sale of both low-fat items and regular items is equal to 1.

* Ho: σ12 /σ22 =1

Alternate Hypothesis: Ratio of variances of sale of both low-fat items and regular items is not equal to 1

* Ha: σ12 /σ22 ≠ 1

On using R codes of F test, we have received following results



From the above result we can see that

* Since our P-value (0.3888) < α (0.9133) and
* f value (σ12/ σ22 ) (5516/3005) = 1.8356 > F critic (0.9728) we fail to reject null hypothesis

An important characteristics of F test is that, since σ12 can always selected to be larger sample variance, then hypothesis testing problems about the ratio σ12/ σ22 are always of two types: right tailed problem and two tailed problem in our case it is right tailed problem.

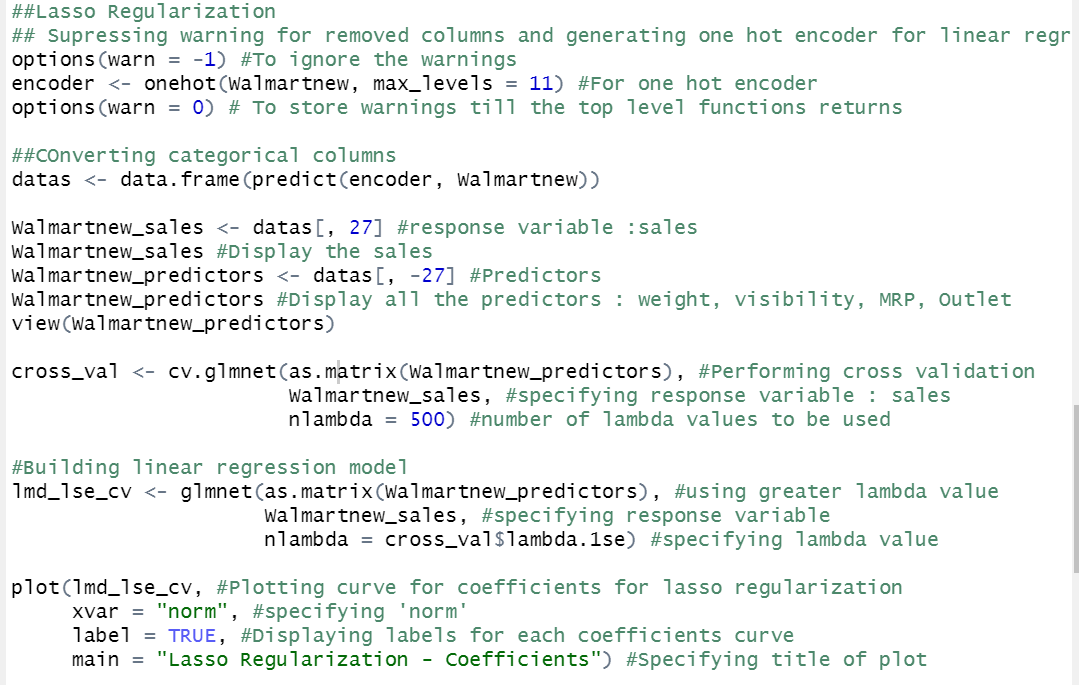
Therefore, we can say that ratio of variance of sales of low fat and regular items is equal to 1. But this statement cannot be true as sale of low-fat items is greater than regular items and this forms a type 2 error.

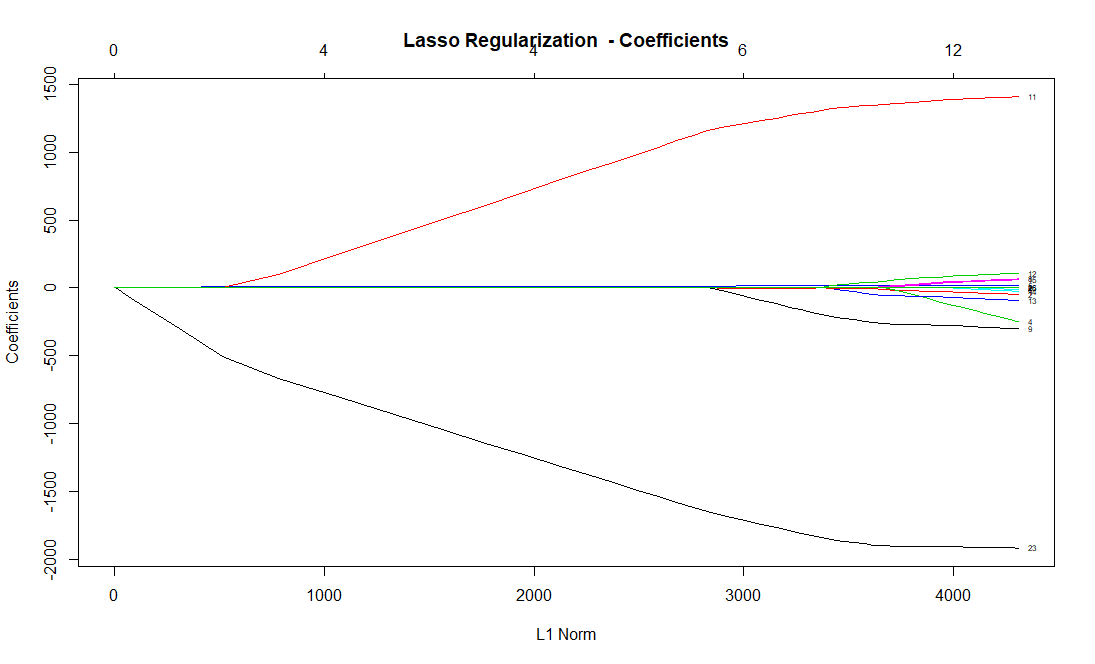
**Lasso Regression**

Regularization is a technique to solve the problem of overfitting in the datasets. The three regularization methods are L1 regularization (LASSO Regression), L2 regularization (Ridge Regression) and Elastic-net Regularization. These techniques are used to reduce the error by fitting the function to avoid overfit problem and are better alternatives to the traditional methods like cross-validation and stepwise regression.

In our group project we have chosen Lasso regularization to avoid the problem of overfitting. We have chosen lasso regularization over ridge as it cannot shrink the values of coefficient to zero. Lasso not only shrinks the unimportant variables to zero but also selects variables automatically (Wang,1960).

LASSO is the acronym for Least Absolute Shrinkage and Selection Operator. It is the method where the dataset is reduced to only the important features or predictors and uses the shrinkage where the L1 regularization adds penalty equal to the absolute value of the magnitude of coefficient. The strength of penalty is controlled by lambda.





This graph helps us in interpreting the variables appearance and the influence they have on the response variable. From the Lasso Regularization graph, we can observe that the coefficients for grocery store (variable 23) has entered the model first but is negative. Almost all the variables have been shrunk to zero or approaching zero except grocery store (variable 23), outlet 18(variable 9), outlet 27(variable 11) and visibility (variable 4) indicating that the lasso will drop those variables from the model. We observe that there is a positive effect from outlet 27 (variable 11) on the sales indicating that customers choose to prefer that outlet more. The rest important variables have a negative effect on the sales.

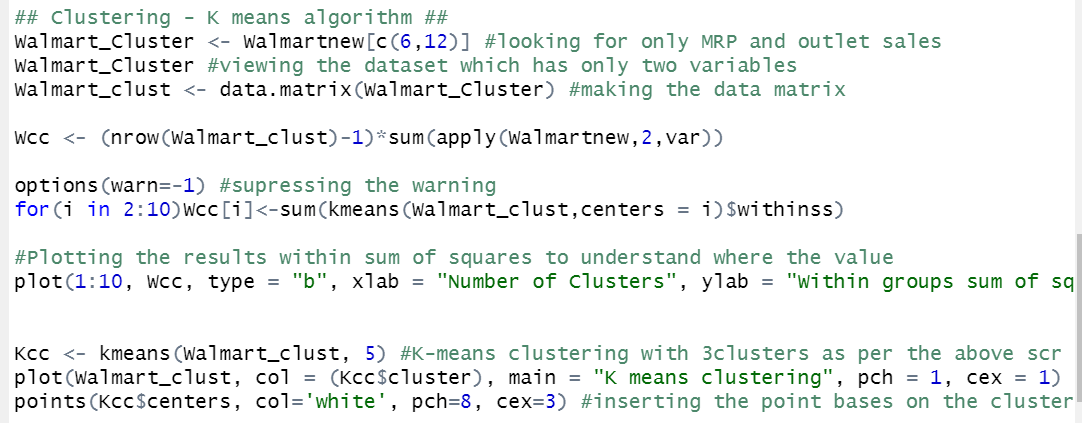
**Clustering**

Clustering involves dividing the data points into a specific number of groups such that the data points in one group are more similar to other. This is one of the unsupervised learning algorithms in machine learning useful to find unknown patterns in a dataset without pre-defined labels. The algorithm works in such a way that the clustering of data is done based on similarity that exhibits common traits. Here we are trying to discover and identify relationship based on distance measures. The types of clustering methods are

1.Hierarchical Clustering – Find clusters using previously available clusters

2.K means (Partitional clustering) – Helps in determining all clusters at once.

In our project, we have used k means clustering (preferred over hierarchical clustering as the data is big) as we are interested in grouping the entire data by minimizing the sum of squares with cluster centroid. K means is simple to confirm the business assumptions on MRP and Sales and about what types of groups exist.



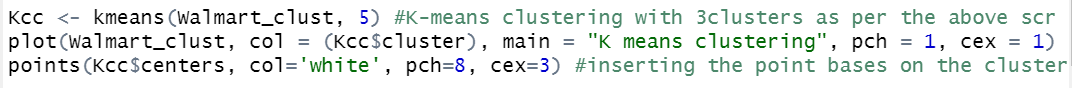
A close up of a map

Description automatically generated

Need to identify and specify the number of clusters:

This is done using elbow approach with a scree plot. The iteration for default of 1 to 10 values are run where for each k value a sum of square distances from each point to centroid is calculated. When plotting in the scree plot, the required number of clusters is identified at point where we do not find significant differences in the curve which conveys that beyond that point clustering is not going to be productive.

Here, after 5, the curve gets almost saturated and hence we fix the cluster value to 5 which means model is suggesting forming 5 clusters for sales and MRP.



A screenshot of a cell phone

Description automatically generated

The graph depicts the black color with low sales cluster and the high ones with dark blue cluster. It also conveys that the low MRP items are contributing to lesser sales even with high number of concentrations. Thus, at Walmart we find the pattern that large items are purchased less by customers. Yet due to their high MRP, the contribution to sales is increasing here. If the Walmart need to have more concentration of data points in high MRP items, they can come up with promotional or seasonal offers on these items to check if there are a greater number of high MRP items getting sold. This will give an idea of people’s behavioral pattern and they can group sales activity accordingly. The white stars in the graph give us an idea of cluster centroid for each of the 5 clusters.

**Conclusion**

In this project we have performed several steps of data analysis. Initially we have cleaned the data by changing the corrupted values and summarized the data through exploratory data analysis. Secondly, we have calculated the correlation between all the features to find which feature influences the dependent variable sales. Thirdly, we calculated variances for low and regular fat and checked it through hypothesis testing which resulted in the null hypothesis acceptance and a false negative error. Due to the false negative error we have conducted lasso regularization in order to find the predictors which impact the target variable, sales of Walmart. As the correlation matrix shows a moderate correlation between MRP and sales, we conduct clustering as the lasso regularization failed to identify the MRP impact on sales. Finally, from clustering we tried to analyses the pattern of purchasing behavior based on MRP of the items. This project has cleared the concepts of clustering, hypothesis testing, regularization, exploratory data analysis and how to do analysis on R platform.

**References**

The Ultimate Guide to Data Cleaning. (2020). Retrieved 11 February 2020, from <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>

Wang, L. W. L. (1960, July 1). When should I use lasso vs ridge? Retrieved from <https://stats.stackexchange.com/questions/866/when-should-i-use-lasso-vs-ridge>

“K Means Clustering in R.” *R*, 28 Dec. 2015, [www.r-bloggers.com/k-means-clustering-in-r/](http://www.r-bloggers.com/k-means-clustering-in-r/).

parry prabhu Follow. “K Mean-Clustering Algorithm.” *LinkedIn SlideShare*, 29 July 2015, [www.slideshare.net/parryprabhu/k-meanclustering-algorithm](http://www.slideshare.net/parryprabhu/k-meanclustering-algorithm)

F-Test: Compare Two Variances in R - Easy Guides - Wiki - STHDA. (2020). Retrieved 11 February 2020, from <http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r>