Sales Forecasting for New Releases in Fast Fashion Retailing.

Submitted to

Mr. Chongshou Li

MH6151 Data Mining, Analytics (MSc) Program, School of Physical and Mathematical Sciences, Nanyang TechnologicalUniversity, Singapore 637371, Singapore, cs.li@ntu.edu.sg

By

Balaji Sri Raj. D.U *(G1600775B)* **& Gautham Tinnium Raju** *(G1601020E)*

NTU - School of Physical & Mathematical Science.

Contents

Introduction:

The aim of this project is to analyse the given data set we need to build a model that could predict the Sales for the future. The Dataset given to us was sales database of Fast fashion products. By using the given dataset, we need to predict a regression model and predict the sales for the new releases. Before we get into the analysis part of the project let us see in brief about the business context of the problem and how our prediction will actually impact the business and through which business decisions can be made. In below lines, the descriptive extract of the project, its business impact, the analysis is explained.

Business Analytics:

What is Business Analytics?

Business analytics (BA) is the practice of iterative, systematic assessment of an organization's data with prominence using statistical analysis. Business analytics is used by corporations committed to data-driven decision making. Successful business analytics depends on data quality, skilled analysts who understand the technologies and the business and an organizational commitment to data-driven decision making.

The BA is mainly concerned with exploring the data to find new patterns and relationships, explaining the reason behind the results in particular period of time, and also can be used to experiment to test previous results.

Sales Forecasting:

Sales forecasting is the process of estimating future sales. Precise sales projections enable companies to make informed business decisions and predict short-term and long-term performance. Organisations can base their forecasts on historical sales data, industry-wide evaluations, and economic drifts.

About the Organisation & Dataset:

The sales forecasting analysis is done for the Multinational Fast Fashion retailer in Singapore, where they have multiple stores across various geographies. The datasets consist of mainly two parts 1) Attributes that are mainly associated with the Store keeping Units (SKU), 2) Sales transactions from the point of Sales (POS). The data has been pre-processed already and given as a single dataset. The dataset actually contains forty-five attributes and one response variable.

The explanatory attributes contain various information about the product's appearance like colour, size and the heel height and it also provides us with information such as the Designer's name, the month when the product is launched and the store where the SKU is sold.

The information which is provided here is so important in predicting the Sales. For example, the rate of demand for a particular product can be found using the period information between the launch of the product and the sale point. That is if the period between the launch of the product and the sale point is less this could mean that demand for the product is high. If the period is too long then there can be various reasons that the demand for the product is low, there is very poor marketing for that particular product, the designer of that particular product can be of less fame, there can be an issue with the store location.

Let's assume we are selling a premium product with very high price margin and the store is located in a place with middle/low-income group.

Eventually, the probability towards selling the product is minimised. These things can be predicted more precisely using the dataset.

Note: *Due to privacy issues the data provided to us for the project is being masked such that no information on the retailer can be found.*

The sales dataset is given for the period of 11 weeks across 30 different stores. The dataset is divided into two sets, training set, and test set. The training set is used to train the model and the test is used to verify the prediction. Since we do not have access to the original test set, we have divided the training set into two 1) Training set 2) Test set (training). Through which we can assume the righteousness of our model before sending our prediction for validation.

Software used for Analysis:

The software which is used for analysis purpose is R (3.4.1). We chose R as it can be easily programmable and the availability of various packages from global community will yield us good predictions. Most of the statistical analysis was carried out using R, and we have used Tableau for visualising purposes. The charts and Graphs which are mentioned below was generated using both R & Tableau. We have used machine learning algorithm which was used for modelling was ***Random Forest in the H2O package.***

Random Forest

The Random forest is one of the best machines learning algorithms for feature selection. Since our data has so many attributes, we have used Random forest. It also takes care of the overfitting by default.

Random forest is nothing but a way of averaging various decision trees, which are trained with a different part of the same dataset.

In simple words, in a random forest, to classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Data Cleaning:

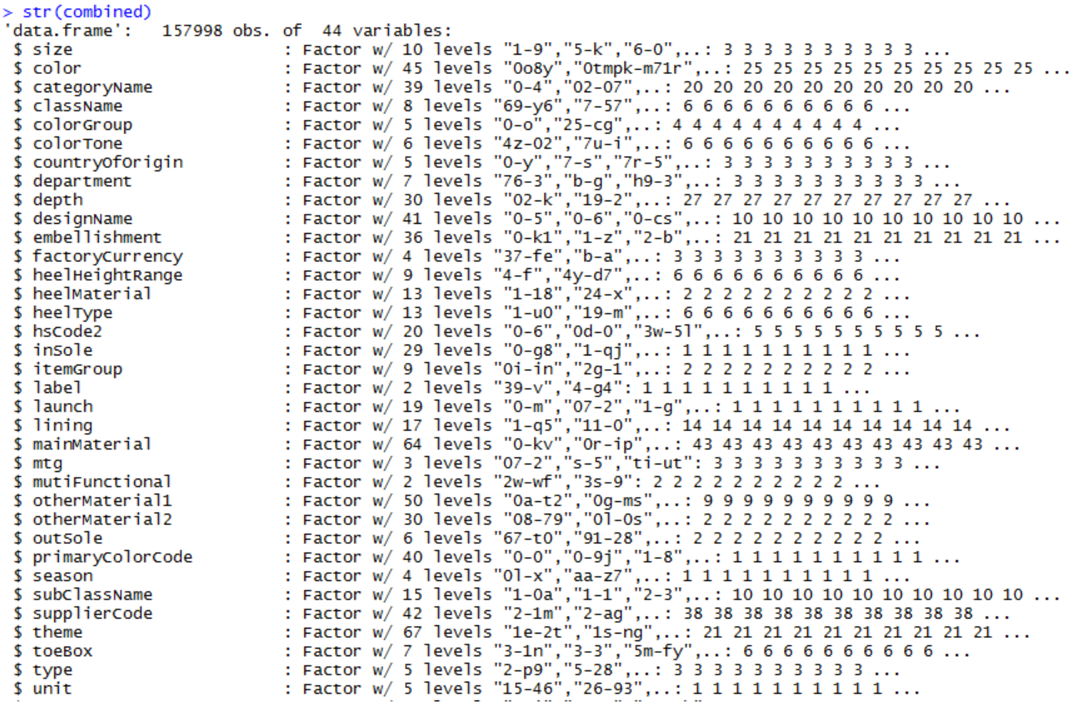
We noticed that certain columns were not useful in prediction and will adversely affect the model, hence we have removed the following columns: -

1. Article Number
2. Bar Code
3. ID
4. Version

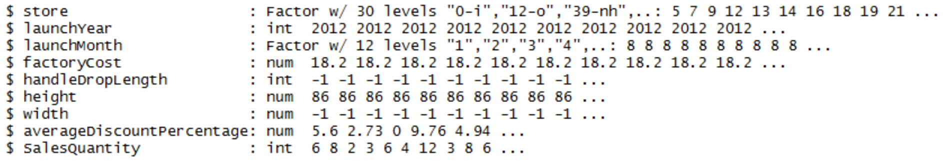
The launch month was changed to numeric values as we noticed irregularities when it was a string

Another issue was that there were certain factors present in the test dataset that was not present in the training dataset. Hence, we combined the data in order to include all factors and then divided it once again into training and test data.

After cleaning and combining the data, the data is as follows:-



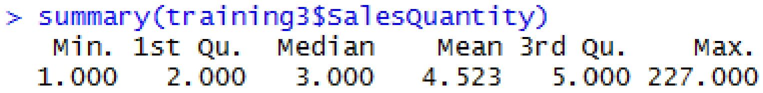
```
> str(combined)
'data.frame':   157998 obs. of  44 variables:
 $ size                     : Factor w/ 10 levels "1-9","5-k","6-0",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ color                    : Factor w/ 45 levels "0o8y","0tmpk-m71r",..: 25 25 25 25 25 25 25 25 25 25 ...
 $ categoryName             : Factor w/ 39 levels "0-4","02-07",..: 20 20 20 20 20 20 20 20 20 20 ...
 $ className                : Factor w/ 8 levels "69-y6","7-57",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ colorGroup               : Factor w/ 5 levels "0-o","25-cg",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ colorTone                : Factor w/ 6 levels "4z-02","7u-i",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ countryOfOrigin          : Factor w/ 5 levels "0-y","7-s","7r-5",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ department               : Factor w/ 7 levels "76-3","b-g","h9-3",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ depth                    : Factor w/ 30 levels "02-k","19-2",..: 27 27 27 27 27 27 27 27 27 27 ...
 $ designName               : Factor w/ 41 levels "0-5","0-6","0-cs",..: 10 10 10 10 10 10 10 10 10 10 ...
 $ embellishment            : Factor w/ 36 levels "0-k1","1-z","2-b",..: 21 21 21 21 21 21 21 21 21 21 ...
 $ factoryCurrency          : Factor w/ 4 levels "37-fe","b-a",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ heelHeightRange          : Factor w/ 9 levels "4-f","4y-d7",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ heelMaterial             : Factor w/ 13 levels "1-18","24-x",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ heelType                 : Factor w/ 13 levels "1-u0","19-m",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ hsCode2                  : Factor w/ 20 levels "0-6","0d-0","3w-5l",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ inSole                   : Factor w/ 29 levels "0-g8","1-qj",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ itemGroup                : Factor w/ 9 levels "0i-in","2g-1",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ label                    : Factor w/ 2 levels "39-v","4-g4": 1 1 1 1 1 1 1 1 1 1 ...
 $ launch                   : Factor w/ 19 levels "0-m","07-2","1-g",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ lining                   : Factor w/ 17 levels "1-q5","11-0",..: 14 14 14 14 14 14 14 14 14 14 ...
 $ mainMaterial             : Factor w/ 64 levels "0-kv","0r-ip",..: 43 43 43 43 43 43 43 43 43 43 ...
 $ mtg                      : Factor w/ 3 levels "07-2","s-5","ti-ut": 3 3 3 3 3 3 3 3 3 3 ...
 $ mutiFunctional           : Factor w/ 2 levels "2w-wf","3s-9": 2 2 2 2 2 2 2 2 2 2 ...
 $ otherMaterial1           : Factor w/ 50 levels "0a-t2","0g-ms",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ otherMaterial2           : Factor w/ 30 levels "08-79","0l-0s",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ outSole                  : Factor w/ 6 levels "67-t0","91-28",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ primaryColorCode         : Factor w/ 40 levels "0-0","0-9j","1-8",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ season                   : Factor w/ 4 levels "0l-x","aa-z7",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ subClassName             : Factor w/ 15 levels "1-0a","1-1","2-3",..: 10 10 10 10 10 10 10 10 10 10 ...
 $ supplierCode             : Factor w/ 42 levels "2-1m","2-ag",..: 38 38 38 38 38 38 38 38 38 38 ...
 $ theme                    : Factor w/ 67 levels "1e-2t","1s-ng",..: 21 21 21 21 21 21 21 21 21 21 ...
 $ toeBox                   : Factor w/ 7 levels "3-1n","3-3","5m-fy",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ type                     : Factor w/ 5 levels "2-p9","5-28",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ unit                     : Factor w/ 5 levels "15-46","26-93",..: 1 1 1 1 1 1 1 1 1 1 ...
```



```
$ store                    : Factor w/ 30 levels "0-i","12-o","39-nh",..: 5 7 9 12 13 14 16 18 19 21 ...
$ launchYear               : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
$ launchMonth              : Factor w/ 12 levels "1","2","3","4",..: 8 8 8 8 8 8 8 8 8 8 ...
$ factoryCost              : num  18.2 18.2 18.2 18.2 18.2 18.2 18.2 18.2 18.2 18.2 ...
$ handleDropLength         : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
$ height                   : num  86 86 86 86 86 86 86 86 86 86 ...
$ width                    : num  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
$ averageDiscountPercentage: num  5.6 2.73 0 9.76 4.94 ...
$ SalesQuantity            : int  6 8 2 3 6 4 12 3 8 6 ...
```

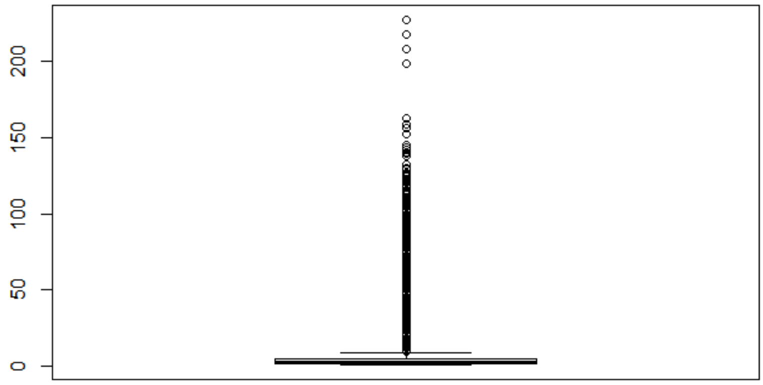
The data now consists of 44 variables having a total of ***157998*** *observations.*

```
> describe (training$SalesQuantity)

training$SalesQuantity

Mean    .05    .10    .25    .50    .75    .90    .95
4.523    1      1      2      3      5      8     12
```



```
> summary(training3$SalesQuantity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   2.000   3.000   4.523   5.000 227.000 
```

The above summary statistics shows that the data contains outliers.

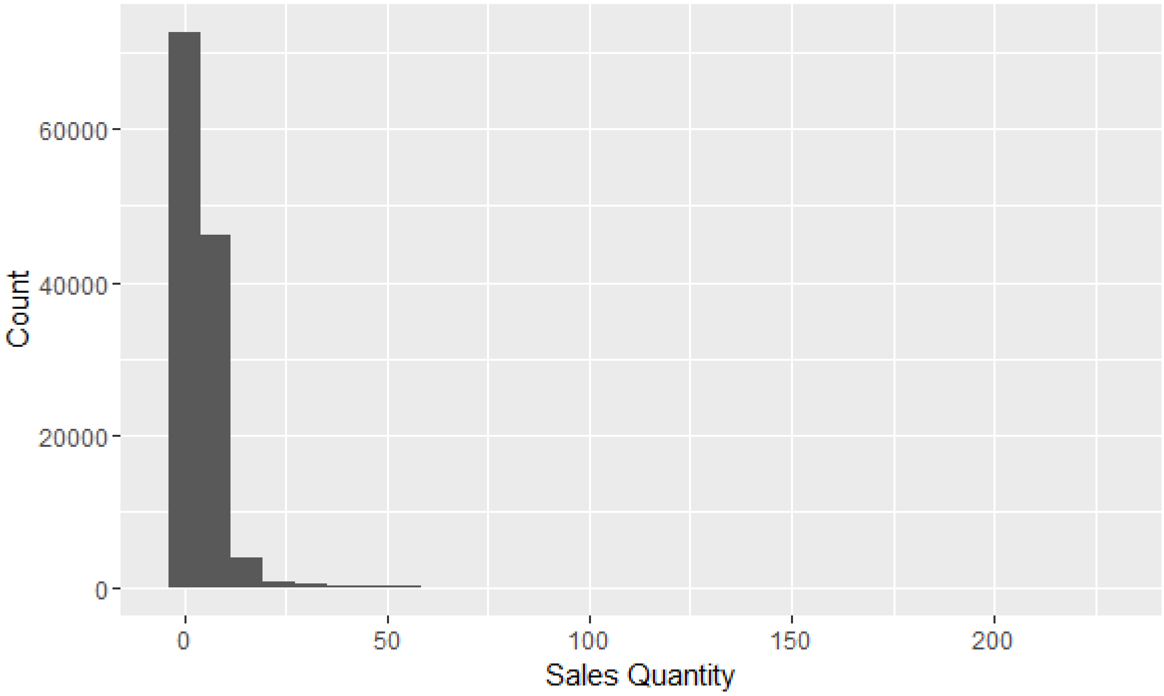


What are outliers and they how affect the model?

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. This can really add so many issues when it comes to data modelling. Because when you model a particular dataset with outliers that prediction capacity of the particular model is diminishing due to the characteristic of the outliers. So, it is always good to remove the outliers and define the model appropriately. But there are also chances this might lead to improper modelling. When you remove outliers there is a great chance to loose sensitive information about the particular data set when in turn again affect the data modelling.

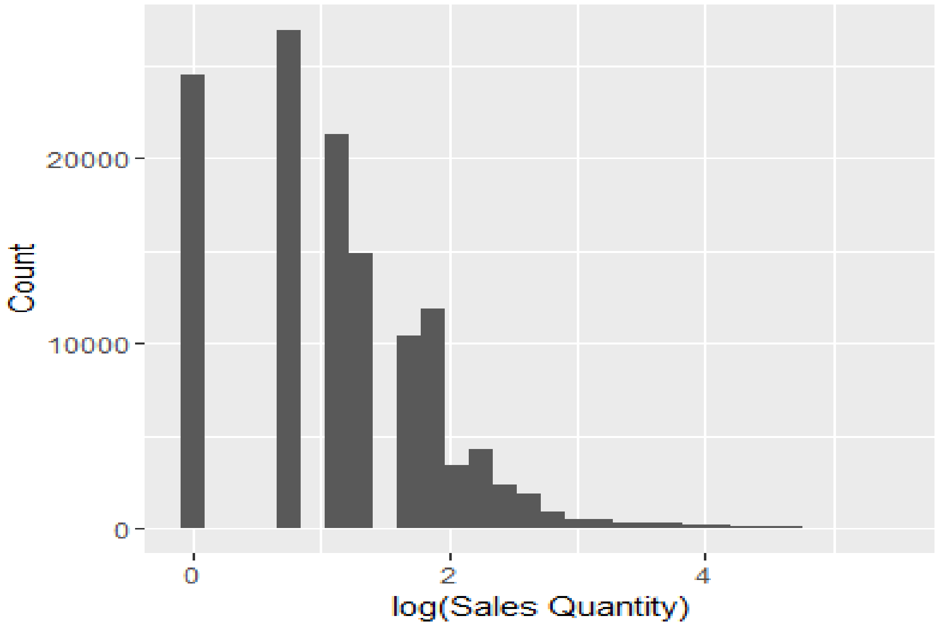
So, before removing the outliers it is a mandatory condition to understand what is abnormal about the data, and what are the chances of losing information due to the removal of the outliers. In certain cases, the removed outliers can be grouped into a separate subset and can be modelled accordance to their characteristics.

Data Exploration

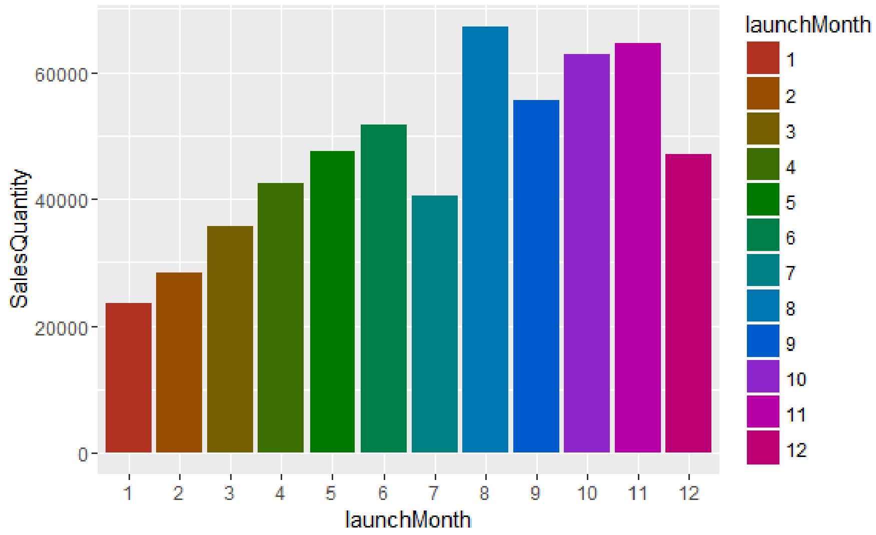


If we plot a graph between the Sales Quantity and Count, we notice that the data is highly left-skewed. It does not follow a normal distribution.

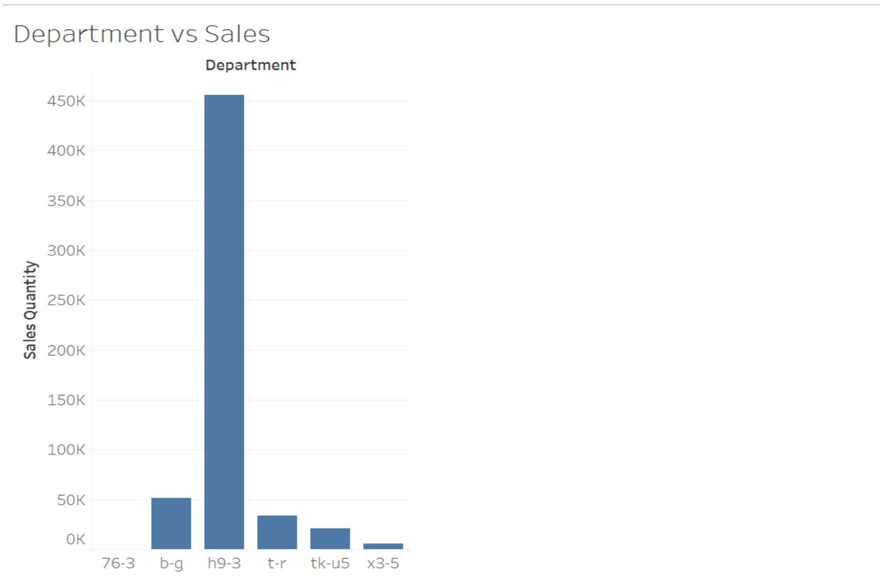
In order to normalise the data to some extent, we perform Log Transformation on the Sales Quantity. The graph of the data after normalising is as follows: -



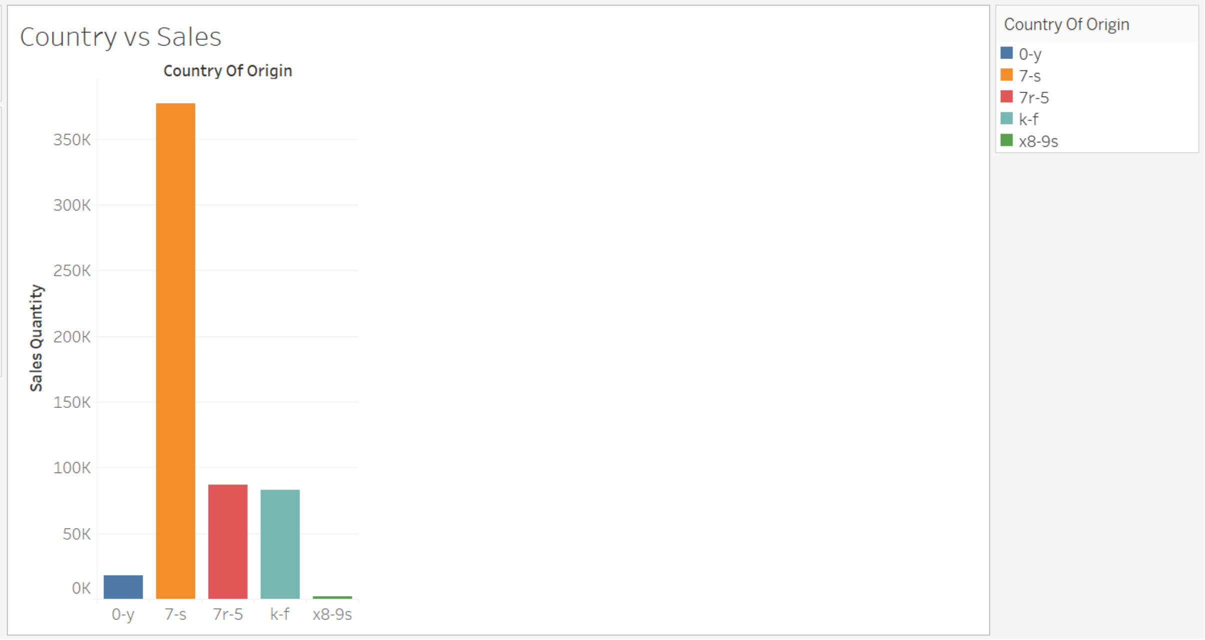
Apart from predicting the sales quantity, we tried to understand the data further.



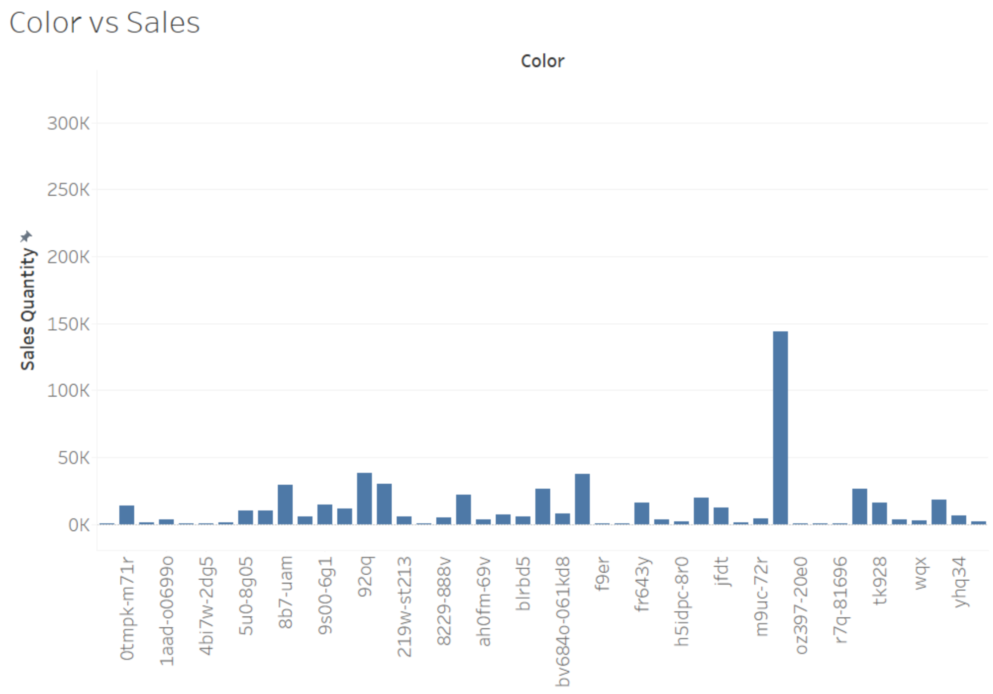
We notice that between August (8) to November (11), the sales is at the highest. During the beginning of the year, the sale is comparatively low and gradually increases as time goes on.



From the above graph, we can notice that the department "h9-3" contributes the most to the sales and the contribution of "76-3" is negligible. Hence the resources allocated for "76-3" can further be allocated to "h9-3".

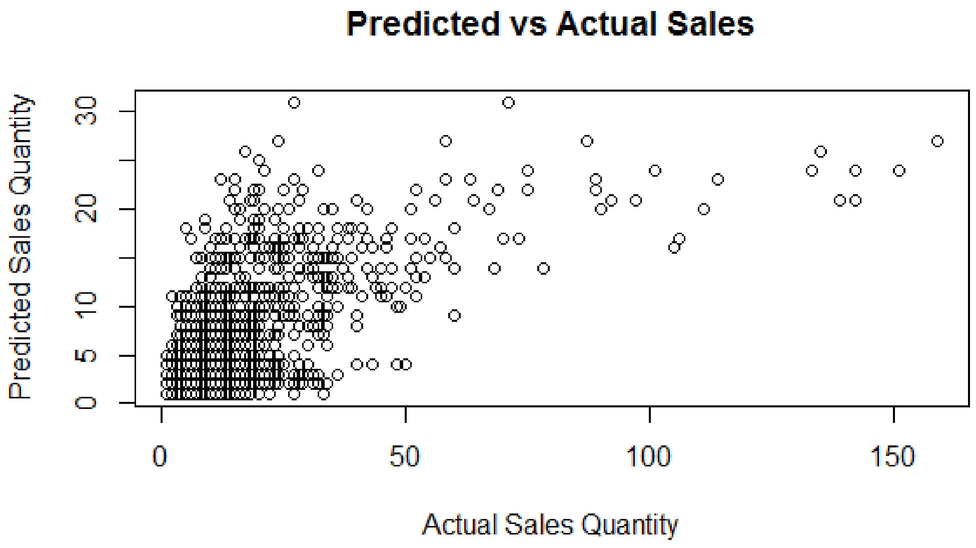


We can see that the Country of origin plays a major role in Sales. When the product is produced in “7-s”, the sales are excellent. This might indicate that the quality of goods produced in that country is of good quality. It might also mean that the number of goods manufactured in that country is high. This might indicate that the cost of produce is less in that country.



We notice that a specific colour is far more popular than other colours among the crowd. The business can concentrate on targeting their products based on these colours.

Results & Conclusion

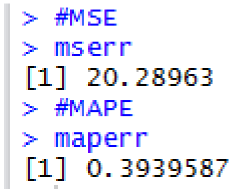


From the scatter plot of Predicted sales vs Actual sales, we can see that an approximate straight line can be fit.

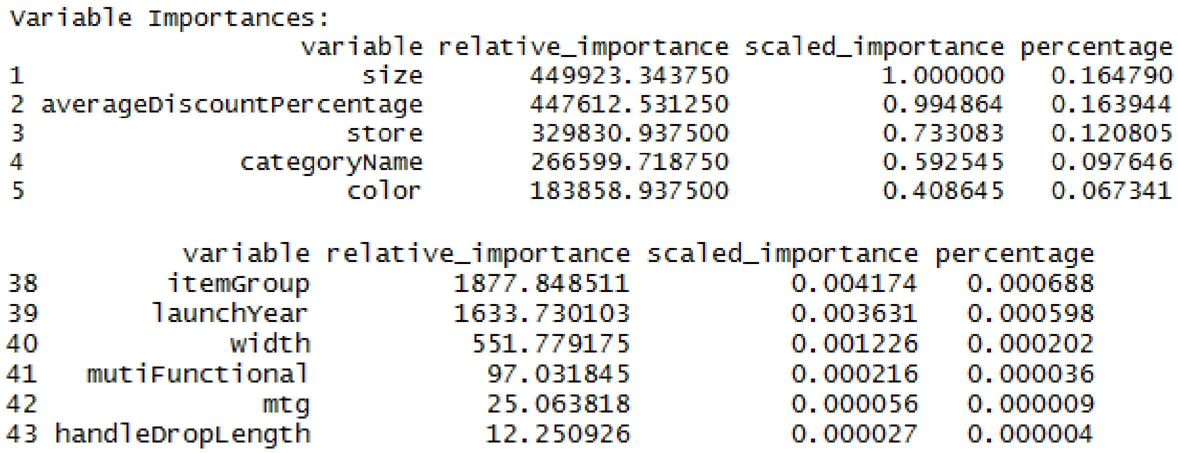
The MSE and MAPE are as follows: -

MSE = 20.28963

MAPE = 0.3939587



```
> #MSE
> mserr
[1] 20.28963
> #MAPE
> maperr
[1] 0.3939587
```



```
Variable Importances:
                     variable relative_importance scaled_importance percentage
1                        size       449923.343750          1.000000   0.164790
2 averageDiscountPercentage       447612.531250          0.994864   0.163944
3                       store       329830.937500          0.733083   0.120805
4                categoryName       266599.718750          0.592545   0.097646
5                       color       183858.937500          0.408645   0.067341

            variable relative_importance scaled_importance percentage
38         itemGroup         1877.848511          0.004174   0.000688
39        launchYear         1633.730103          0.003631   0.000598
40             width          551.779175          0.001226   0.000202
41    mutiFunctional           97.031845          0.000216   0.000036
42               mtg           25.063818          0.000056   0.000009
43 handleDropLength           12.250926          0.000027   0.000004
```

The above images show the importance of each attributes in predicting the Sales quantity of the dataset.