**Analytics Project – II**

**Cereal nutrition analysis using R**

**2202 MSA 6702 1 5001 PRA 13390 GC**

**Project Report**

**Venkata Sai Suneeth Ravilla**

**Academic Supervisor:**

**Dr. John Chen**

**Spring 2020**

**Bowling Green State University, Ohio**

**Dataset Background:**

Dataset represents cereal products with their dietary characteristics. The response variable is to predict Rating of the product. The dataset contains 80 records with 16 variables These 16 variables contain 12 numerical and 4 categorical variables.

**Variables:**

**Numerical:** Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugars, Potassium, Vitamins and Minerals, Weight, Cups, Rating.

**Categorical:** Type, Shelf, Manufacturer, Name of cereal

**Goal:**

In this analysis, we will disclose the main attributes that influence cereal ratings. We will also be predicting ratings using a simple linear regression model and followed with analyzing variables into clusters using K – Means clustering.

**Approach:**

In this report, we are first analyzing the data background how the patterns of the target variable lie and fitting the model to predict the cereal based on other factors. Using K means clustering to cluster the data for classifying the manufactures.

**Major Findings:**

From the analysis in the report, based on the regression model fitted using step wise we have achieved adj R square of 90% which shows that the variables affecting this result will help in predicting the rating of the product. Clustering analysis helped in identifying Nabisco as good manufacturer as the cluster rating is high with many Nabisco products.

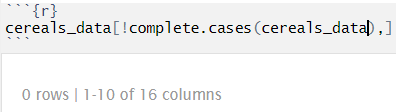
Below report gives detailed analysis and approach followed for analyzing Rating of the product.

**Data Preprocessing**

Below steps are performed to ensure the dataset is cleaned and ready for building the models. That is because while gathering the data, it is not always perfect, and it will help in advance to find any missing values and irrelevant data.

**Missing data:**

Aim in this task is to analysis the data whether the data has any missing values to make sure the data has the values completely for the analysis.



No missing values are found in the given dataset. This ensures that there is no need to pass any mean or median values for the empty cells.

**Data Transformation:**

Given above that preprocessing helps in identifying irrelevant data and missing values, data transformation helps to transform the data into appropriate justification for the analysis.

**Normalization:**

This technique is performed in preprocessing which provides linear transformation on original range of the data

**Categorical data:**

The categorical data present in the dataset has various levels which belongs to specific sets and these need to be scaled for better fit of the model.

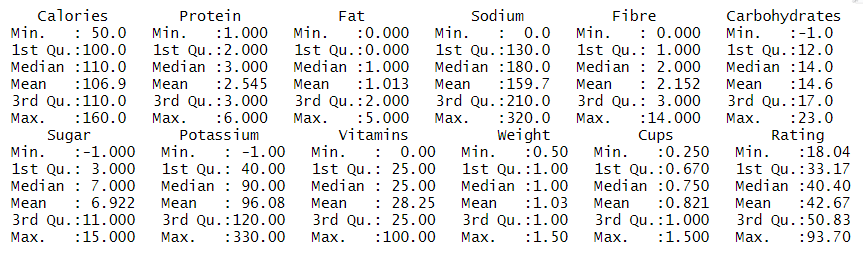
**Data Configuration:**

The target predictor Rating given as continuous variable, for better analysis of the target variable we are creating a new column Average Rating with 3 levels. Group 3 – Ratings less than 35, Group 2 – Ratings between 35 and 54, Group 1 – Ratings greater than 54

Changing Manufacturer Name and Type from alphabets to actual name.

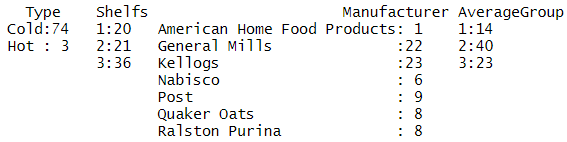
**Summary:**

Analyzing summary of the numerical variables,

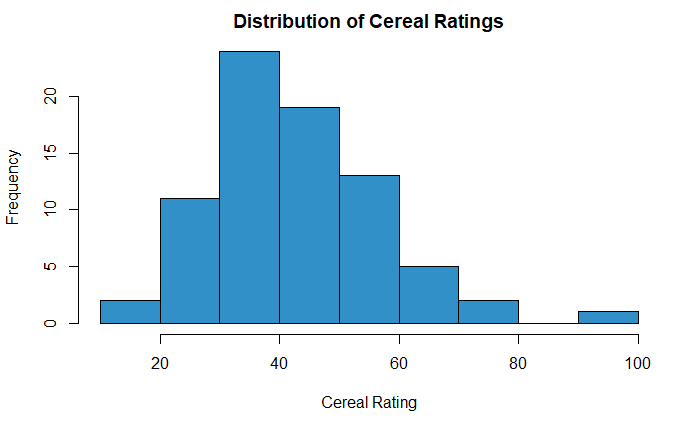
****

From the summary of observations, we can see the average rating of the products is 42.67. Sodium and Calories are high with average of 159.7 and 106.9 respectively per serving.

Analyzing summary of the numerical variables,

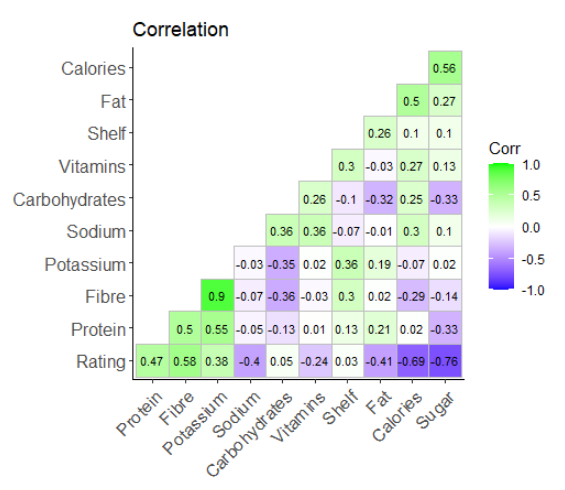


From categorical variable summary, we can see that General Mills and Kellogg's are producing different type of cereals than other 5 manufacturers. Most of the cereals are placed on Shelf 3 assuming Shelfs are decided based on age group and height. Cold consumption of cereals is more in number with 74 which shows that analyzing using Type cannot benefit the results. Average Group 2 contains a greater number of data with 40 records.

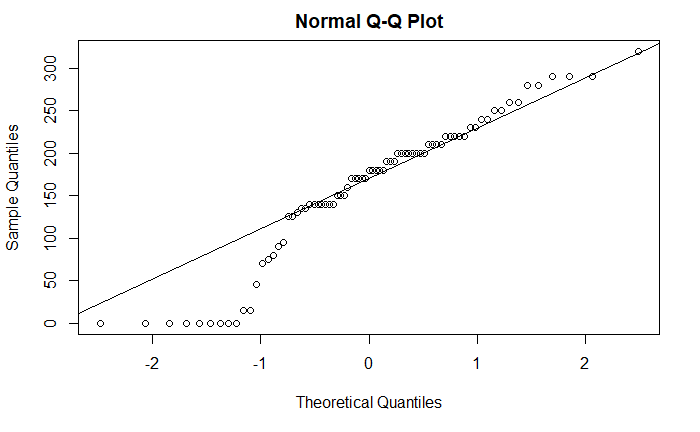


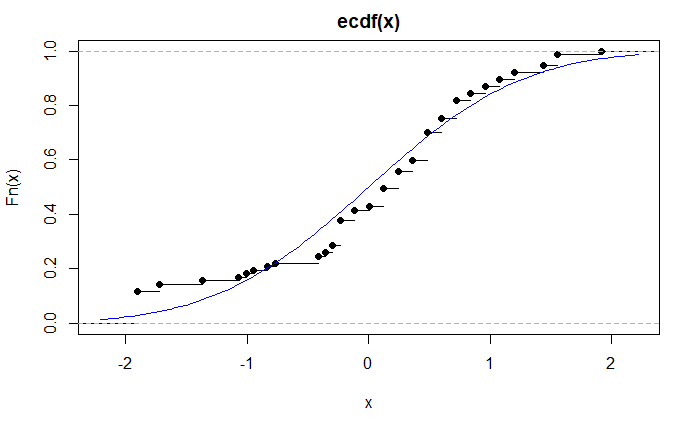
**Distribution:**

Initially, visualizing the distribution of the Ratings of all the products, we can see ratings are distributed towards right causing right skewed distribution. Most of the ratings fall between the range of 30 and 50. This shows on average the customer is neither satisfied nor unhappy with the product.

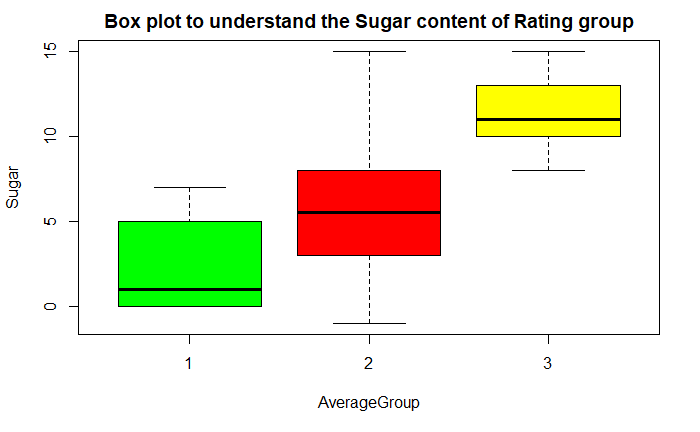
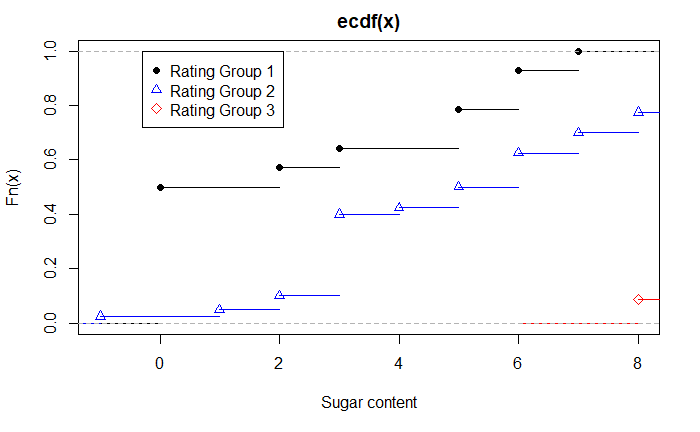
**Correlation:**

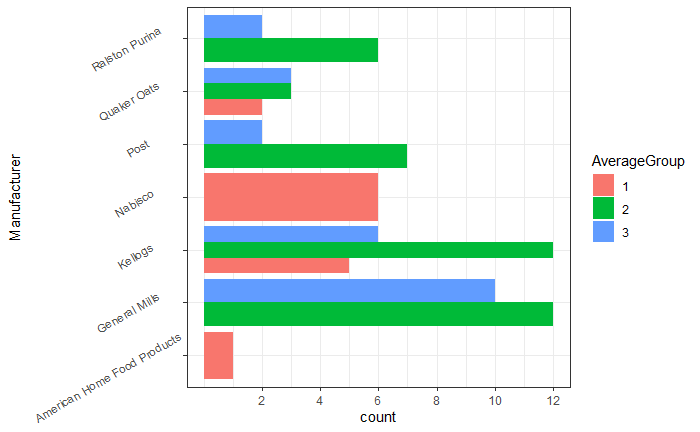
To check which variable is having relation with other variables, performing correlation with the numerical variables which are affecting each other. From the correlation plot, Fibre and Potassium have high correlation with 0.9 which is having positive relation each other. Considering Rating, Sugar and Calories are highly correlated 0.76 and 0.69 respectively with Rating negatively. Fibre is positively correlated with Rating of 0.58

**Normality:**

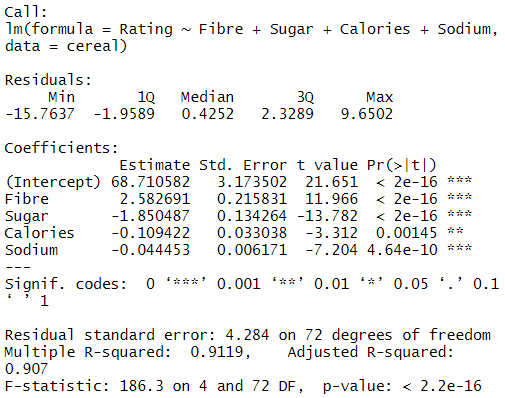
****Sodium has high portion of consumption per serving, so assuming sodium has normality and checking whether it follows normality. Performing Q-Q plot for sodium we can observe the data is slightly skewed which shows light tail distribution of the graph. Extending the analysis using the ECDF function, the plot shows the blue line which passes along the points assumed in the normality distribution. We can still see at the start of the graph that data has slight skewness.

**Average Group vs Sugars (ECDF)**

****Analyzing the data using the average group values which are grouped by 3 levels with sugar content. Group 3 which has less ratings contains a greater number of sugar content. We can also observe the sugar content in group 2 is almost distributed normally. Considering sugar content more than 7 we can only see group 2 and 3 are having sugar content which shows at sugar below 7 are likely to have better rating.

**Manufacturer Ratings:**

Performing analysis for manufacturer with the average group which are grouped by 3 levels. From the plot, we can observe only two manufacturers, Kellogg’s and Quaker Oats are fallen under 3 groups of average. Out of 23 products of General Mills 10 products ratings are below 35 which is poor rating. Kellogg’s and General Mills has same number of products in group 2 with 12 each, ut Nabisco and American Food is exception, it has all its products in group 1 which is highest rating group of the data. Kellogg’s has 7 products in group 1 but General Mills failed to have single product rating over 54

**Fitting Regression Model:**

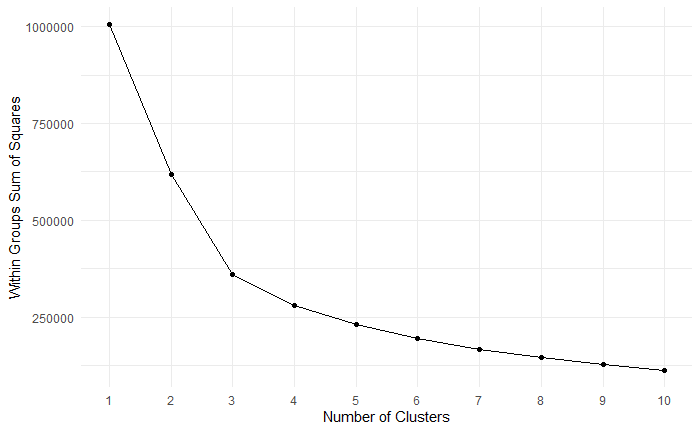
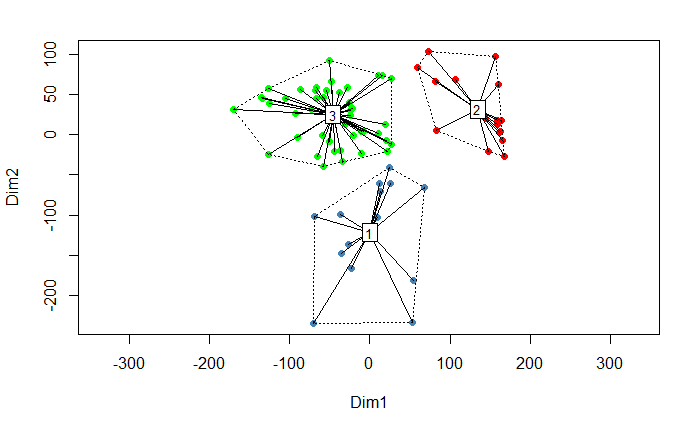
Building model with target variable Rating and fitting all the variables in the dataset. Later performing stepwise regression to drop the variables which are not significant to the model. Give image is the best model fitted using stepwise regression. All the variables, Fibre, Sugar, Calories, Sodium are significant to the model with p value less than 0.05. The model has highest Adjusted R Square 90% which shows that with these variables we can predict the rating of the product.

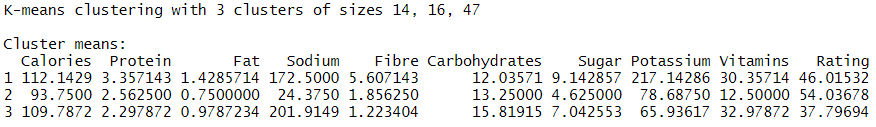
Regression model output,

Rating = 68.71 + (Fibre) 2.58 – (Sugar) 1.85 – (Calories) 0.10 – (Sodium) 0.04

**K – Means Clustering:**

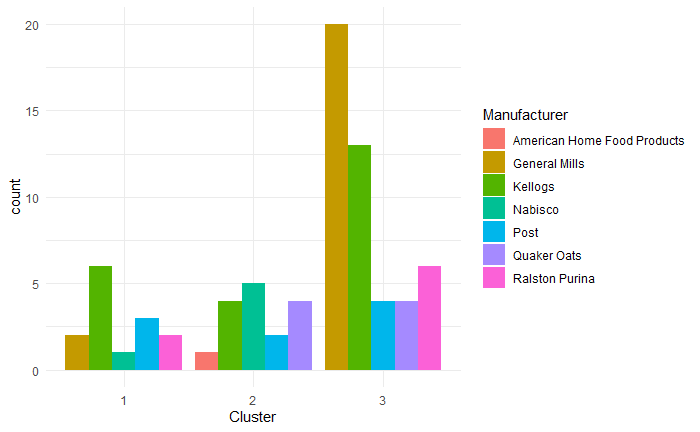
Analyzing the variables for better grouping of the nutrients of cereal and rating, choosing k means clustering will help in grouping the variables. For choosing right number of clusters, performed Sum of Squares of the distances of each data point in the cluster to their centroid. From elbow point graph, the optimal k value for the cluster is chosen as 3.





Clusters are group into 3 and the sizes of 3 clusters are of 14, 16 and 47. From the table we can see Cluster 3 has least rating of 37.7. Highest average Rating is obtained in cluster 2 with 54 which is not so good but better than the average Rating of the complete dataset. We can observe less amount of sugar and calories in cluster 2. Clustering is plotted using centroid method which takes the average of all values in the group and it gives a point of centroid which calculates distance between other clusters and the values in the cluster. By looking at the cluster plot, we can see cluster 2 is on positive side which shows it is positive group out of 3. From the cluster 2 we can determine the factors which are affecting the rating and need to improve accordingly for better rating of the product.

**Analyzing clusters on manufacturer:**



Clusters are analyzed with the manufacturer to identify who is performing better in producing better rating products. From the plot, we can see out of 7 companies, 5 Manufacturer lie in cluster 3 which we clustered as less Rating earlier. Manufacturer like Nabisco and American is in cluster 2 which has better rating and less sugars. Even though General Mills has multiple products it failed to gain spot in better ratings cluster. Using this clustering we can define which manufacturer has produced high rating products and we can conclude Nabisco stays top on the list with most of their products lying in cluster 2.

**Conclusion:**

* Rating can be predicted with the variables Fibre, Sugar, Calories and Sodium.
* Average Cereal Rating is 42.6
* Potassium is highly correlated with Fibre, but Calories and Sugar are correlated with Rating.
* Group 3 contains high amount of sugars based on ECDF plot but on average it also has less ratings.
* Almost all the manufacturer products are in Group 2 and 3 except Nabisco who seems to focus on Ratings.
* Sodium variable seem to follow the normality assumptions with slight skewness.
* Data is clustered into 3 groups using K-Means Clustering and on average 3rd cluster has more data with least average of rating than other two clusters.
* Analyzing clusters with manufacturer we can see Nabisco and American Home stands Cluster 2 which has better Rating and on average contains less amount of sugar.