GROUP PROJECT

**Forecasting future ratings of a cereal using statistical modelling in R**

**Team**

Akash deep konda

Harshitha Varimadugu

Mahesh Reddy

Lakshmi Ramya Marineni

Venkata Subramanya Sai Sasank Prakki

Business Analytics and Project Management

OPIM 5603: Statistics in Business Analytics

Prof. Pankaj Prakash.

**Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page No.** |
| 1 | Introduction | 3 |
| 2 | Problem Statement | 3 |
| 3 | Data Dictionary | 4 |
| 4 | Methodology | 5 |
| 5 | Modeling | 8 |
| 6 | ANOVA | 9 |
| 7 | Pearson’s Chi-Squared test | 10 |
| 8 | Conclusion and Recommendations | 11 |
| 9 | References | 12 |

**Introduction**

The dataset includes the names of various cereals produced by the five separate companies. Calories, Protein, Fat, Sodium, Fiber, carbohydrates, Sugars, Potassium, and Vitamin Content are all listed for each cereal. There are 80 different types of cereals in this collection. The dataset also includes the weight in ounces, the number of cups used in serving, the display shelf on which the product is shown, and customer ratings. Our endeavor entails forecasting future ratings based on the information we have. It assists manufacturers in maintaining the contents of a cereal that customers enjoy. As a result, cereal manufacturing enterprises will be more profitable.

**Problem Statement**

In this study, we have employed modeling techniques to assess the ratings of the different cereals. It assists the company in obtaining information such as which cereals consumers prefer. In today's environment, one of the most important jobs is to preserve good health, and one method to do so is to keep excellent eating habits. As a result, clients are quite picky about what they eat. A customer's rating is influenced by the substance of any products in it, as well as the flavor. As a result of this, information on these cereals and their ratings have been closely monitored to formulate the appropriate formula while manufacturing a cereal.

**Data Dictionary**

**Target Variable:**   
**Rating:** It is the rating we give to the cereal based on the input variables.

**Exploratory Variables:**   
**Name:** Name of the Cereal  
**mfr:** Manufacturer of the Cereal  
**type:** Hot cereal or Cold cereal  
**Calories:** Represents the number of calories present per serving  
**Protein:** Represents the amount of protein present per serving  
**Fat:** Represents the amount of fats present per serving  
**Sodium:** Represents the amount of sodium present per serving  
**Fiber:** Represents the amount of fiber present per serving  
**Carbo:** Represents the amount of carbohydrates present per serving  
**Sugars:** Represents the amount of sugars present per serving  
**Potass:** Represents the amount of potassium present per serving  
**Vitamins:** Represents the amount of vitamins present per serving  
**Shelf:** Position of the cereal placed in a store  
**Weight:** Represents weight in ounces  
**Cups:** Represents number of cups to take for a serving

**Methodology**

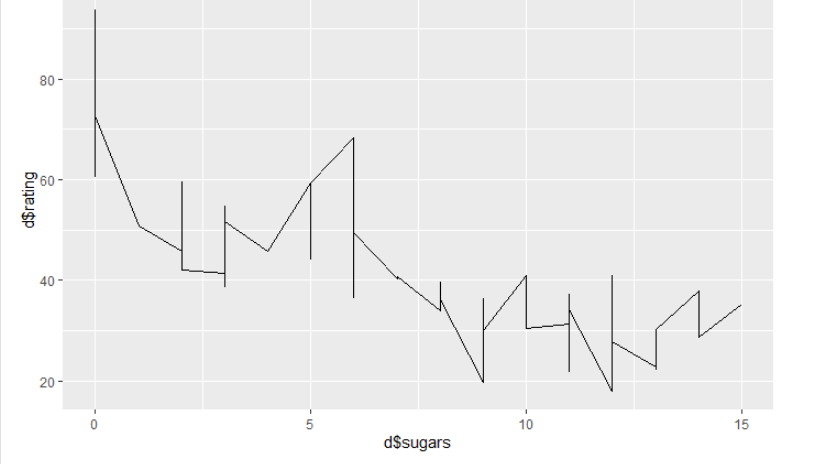
The data is first sampled, then we did exploration, Modeling to predict the future rating of the cereals is done and a anova test to find if the cereals are in same terms of the ratings and chi square test to find if there is any relation between the cereal manufacturers and cereals fat level.

**Data cleaning:**

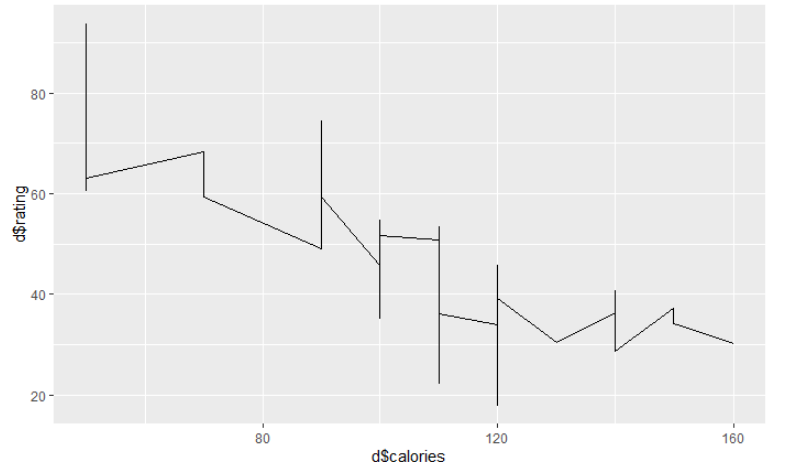
Our initial data consists of missing values and outliers. So we with the help of R found that there are 4 missing values in the Form NA’s. So, we deleted the NA’s from the dataset.

**Exploring:**

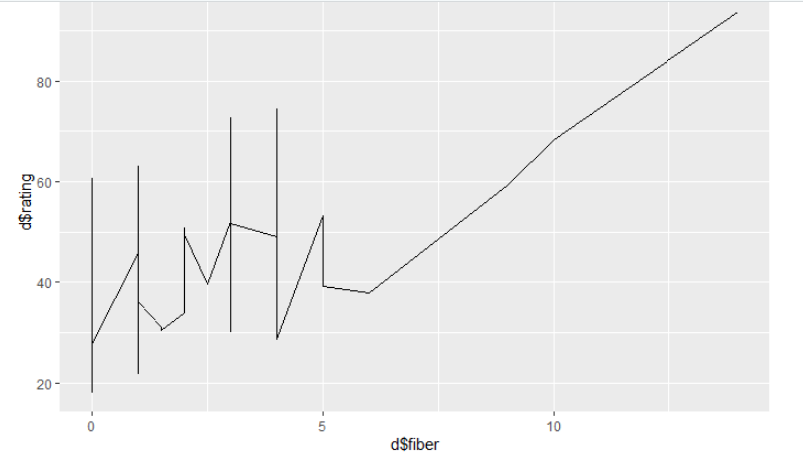
Our dataset contains 16 rows. Since our target variable is rating. We plotted a scatterplot between rating and each of the other variables in the dataset.



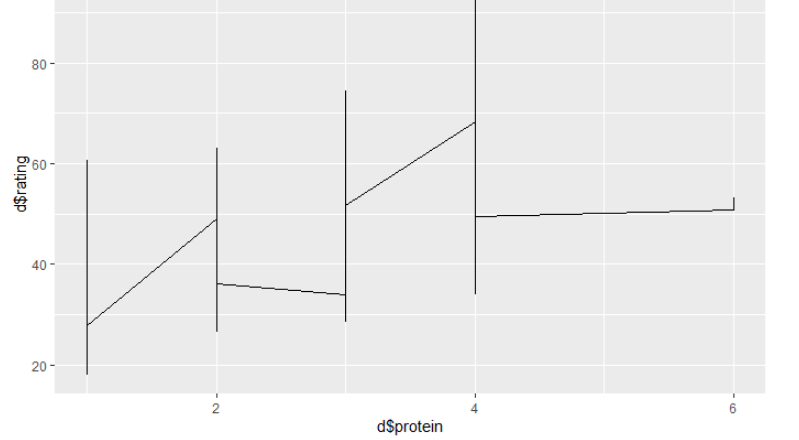
**#Chart 1 - As sugar is more, the rating decreases (Sugar and ratings are inversely proportional)**



**#Chart-2 -When calorie increases, the rating decreases (Calories and ratings are inversely proportional)**



**#chart3 - When fiber content increases, the rating increases (fiber and ratings are directly proportional)**

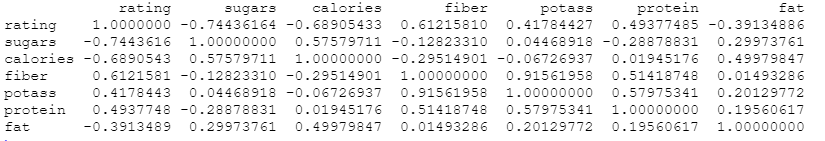


**#Chart4 - When protein content increases, the rating increases and there were two outliers for protein value = 6(protein and ratings are directly proportional)**

Note: The plots between the fat, potass, carbo, sodium and rating doesn’t have any well-defined relation between them.

**Correlation matrix analysis:**

We further did correlation matrix analysis among all the variables using the R.

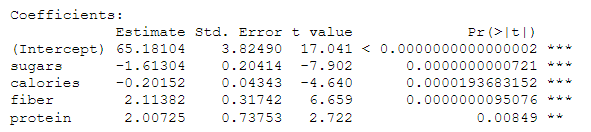


* Rating and sugar variables exhibit the highest negative correlation.
* Fiber and ratings, protein and ratings are also positively correlated according to the correlation matrix.

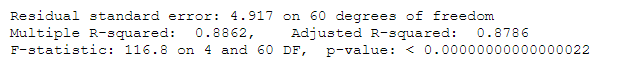
So based on this we concluded that sugars, calories, fiber and protein are correlated with the ratings. Hence, we selected these variables for modeling.

**Modeling:**

After sampling and data exploration we were able to eliminate missing variables and we have selected the variables that are highly correlated with rating i.e., sugars, calories, fiber and protein. With the help of these variables, we have developed a multiple regression model. The output of the multiple regression model is as follows.



The p value for each corresponding variable is less than the significant value (α = 0.05). So, we conclude that the variables which we have selected have significance in the multiple regression model we built.



**Summary of the model:**

* The R square value for our regression model is 0.8862.
* The statistical significance of this model is 99%.
* For every unit increase in sugars there is a decrease of 1.633 in the rating.
* For Every unit increase in calories there is a decrease of 0.2 in the rating.
* For every unit increase in the fiber and protein is an increase of 2.11 and 2.007 respectively in the rating.

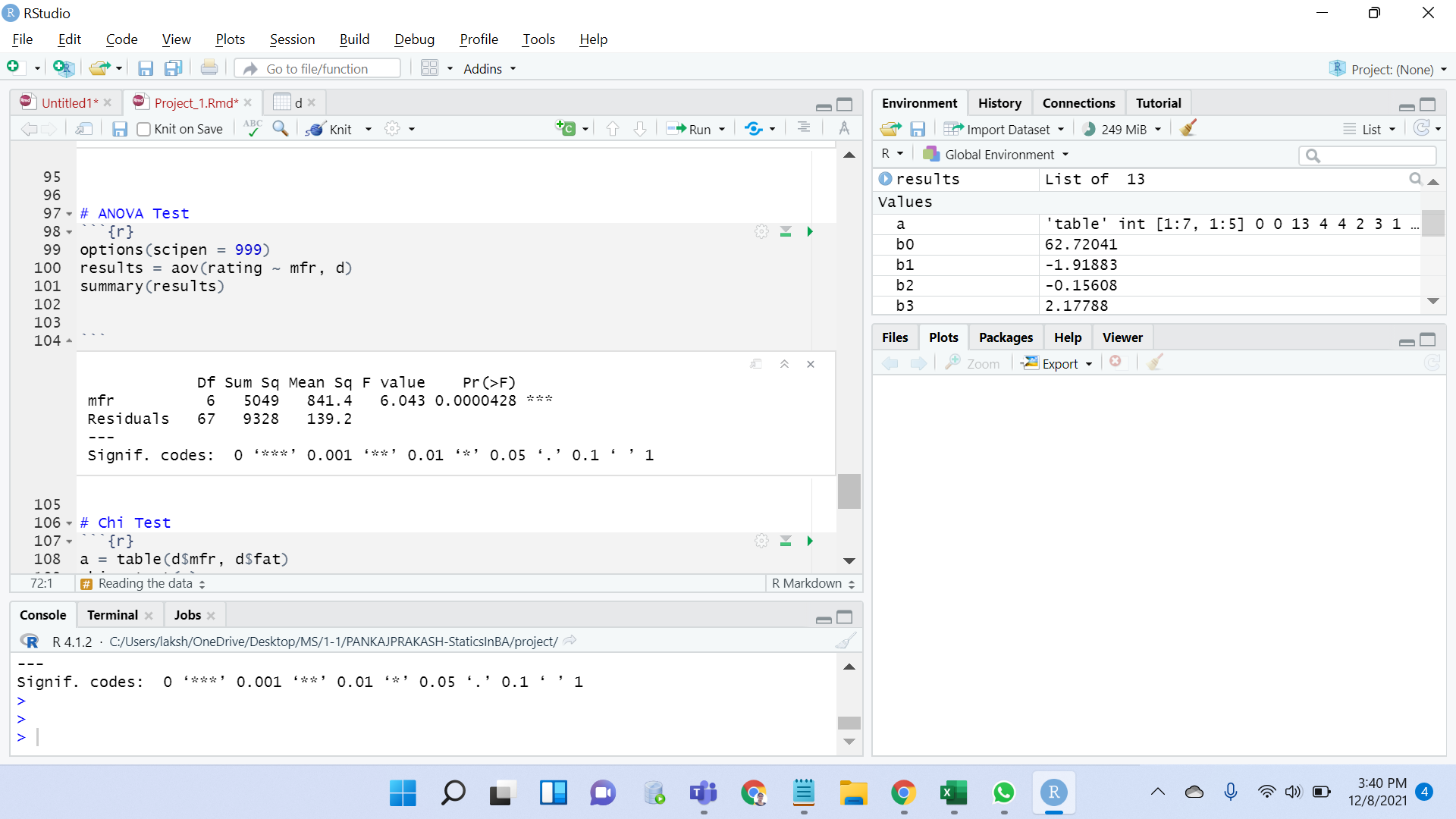
**ANOVA**

ANOVA is a statistical test for estimating how a quantitative dependent variable change according to the levels of one or more categorical independent variables. ANOVA tests whether there is a difference in means of the groups at each level of the independent variable.

**The null hypothesis (H0):** There is no difference in means

**Alternate hypothesis (Ha):** The means are different from one another

We perform an ANOVA in R using the aov() function. This will calculate the test statistic for ANOVA and determine whether there is significant variation among the groups formed by the levels of the independent variable.



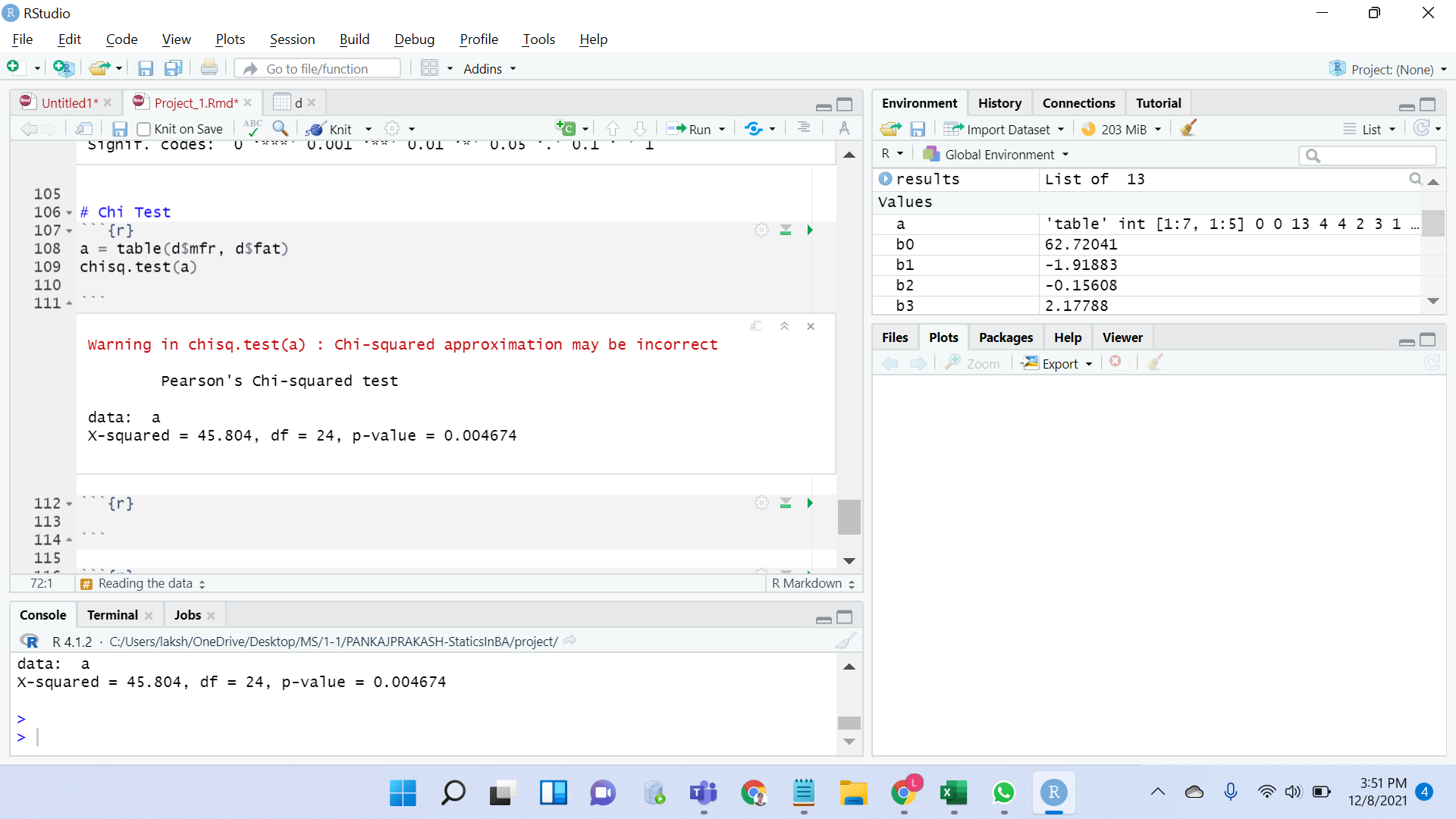
The model summary first lists the independent variables being tested in the model (in this case we have ‘mfr’) and the model residuals (‘Residual’). The p-value of the mfr variable is low (p < 0.01), so it appears that the mfr has a real impact on the cereal rating.

**Pearson's Chi-squared test**

Chi Square test is performed to estimate whether two variables (mfr and fat) are associated or related by a function.

**H0:** The two variables (mfr, fat) are independent.

**H1:** The two variables (mfr, fat) relate to each other.



We have a chi-squared value of 45.804. Since we get a p-Value less than the significance level of 0.05, we reject the null hypothesis and conclude that the two variables mfr and fat are in fact dependent.

**Conclusion**

“Cereal” is a dataset that shows data gathered from different types of cereals. We performed data cleaning, exploratory data analysis to extract information from the data and then we performed data visualization to gain a clear idea on the information by giving it visual context through graphs. We have performed multiple regression analysis to find the significance of variables in the model. We did ANOVA test and checked whether the mean is the same or not by doing hypothesis testing and did Chi-Squared test to compare the observed results with the expected results. By knowing the trends in ratings, it can help in predicting the future forecast of cereal ratings that helps us to assist companies on what to invest to get good profits.

**Recommendations**

We can suggest that Manufacturers that want to bring in high ratings in future should create cereals that are high in fiber, protein and avoid creating cereals with high calorie counts or lots of sugar or fat. We can extend our recommendations by saying Cereals with high ratings are more likely to be placed on the first or third shelf, because that is generally where the consumers’ eyes gravitate. Hence the collection of scatterplots is extremely useful in determining how certain nutritional factors determine the end rating of a cereal. And also By applying Statistical models to the data we are investigating, we were able to understand and interpret the information more strategically.

**References**

1. [**https://www.kaggle.com/jeandsantos/breakfast-cereals-data-analysis-and-clustering/data**](https://www.kaggle.com/jeandsantos/breakfast-cereals-data-analysis-and-clustering/data)
2. [**https://stats.libretexts.org/Courses/Las\_Positas\_College/Math\_40%3A\_Statistics\_and\_Probability/11%3A\_Chi-Square\_and\_Analysis\_of\_Variance\_(ANOVA)**](https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_and_Probability/11%3A_Chi-Square_and_Analysis_of_Variance_(ANOVA))
3. [**http://r-statistics.co/Linear-Regression.html**](http://r-statistics.co/Linear-Regression.html)
4. [**http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/**](http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/)