★ Exploratory Data Analysis(EDA) of Cereals data using R

•••

A report by Group no 5
Janhvi Pardeshi, Aadesh Motghare and Shruti Ramteke

Contents

➔ **Various fields in the Dataset**

- Name of cereal
- Manufacturer of cereal
- Type: cold or hot
- Calories per serving
- Grams of protein
- Grams of fat
- Milligrams of sodium
- Grams of dietary fiber
- Grams of complex carbohydrates
- Grams of sugars
- Display shelf (1, 2, or 3, counting from the floor)
- Milligrams of potassium
- Percentage of vitamins and minerals
- Weight in ounces of one serving
- Number of cups in one serving

Introduction to EDA...

EDA stands for Exploratory Data Analysis. It is the process of investigating the dataset to discover patterns, and anomalies(outliers), and form hypothesis based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.

It's approach can be used to gather knowledge about the following aspects of data:

- Main characteristics of features of data.
- The variables and their relationships.
- Finding out the important variables that can be used in our problems.

★ Methodology Applied

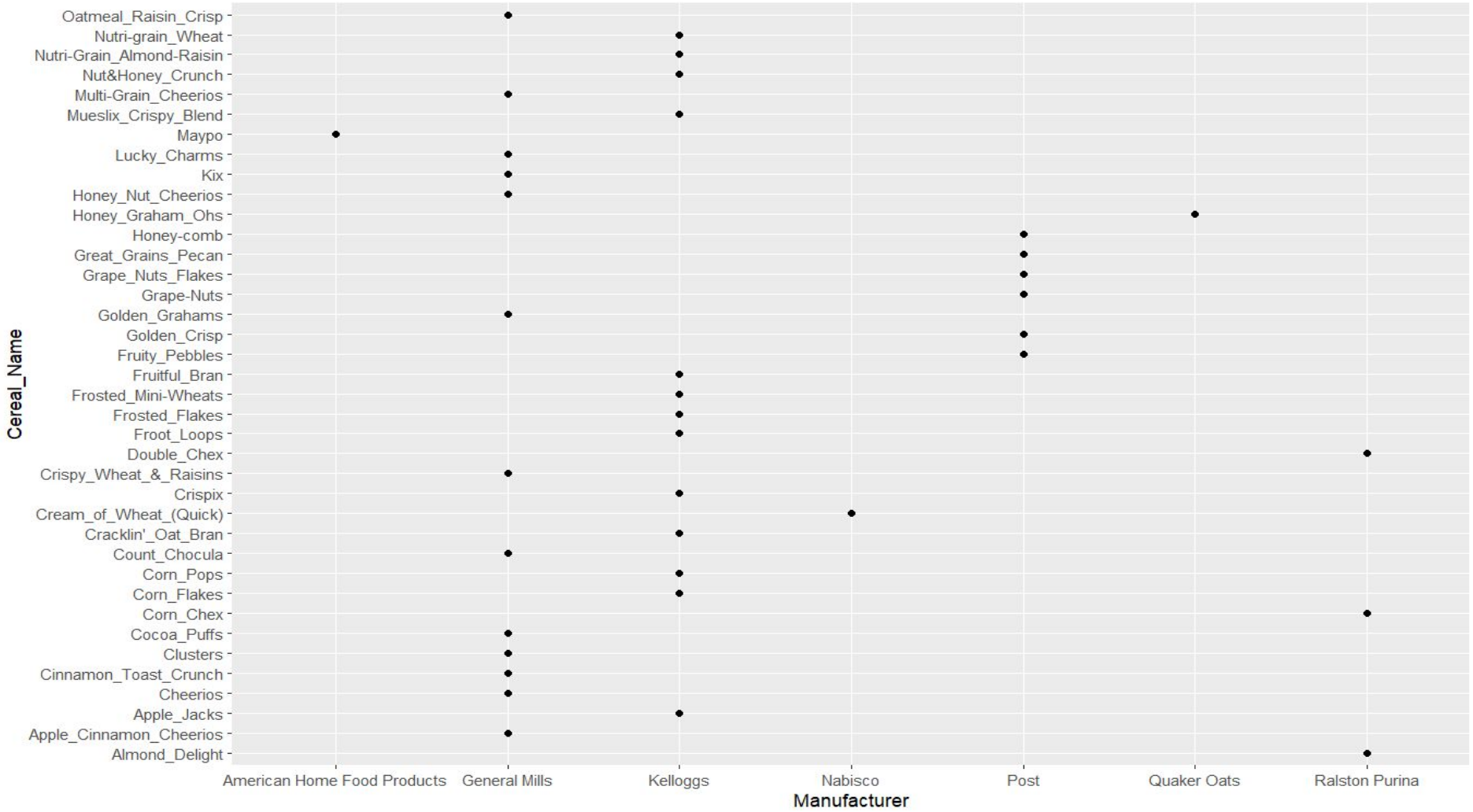
- Importing packages like ggplot, dplyr, moments, epiDisplay.
- Variable analysis on every nutrients over the manufacturer to find that which manufacturer is producing the high nutritional value of cereal.
- Plotting the graphs of analysis of categorical variable and analysis of numerical variable using appropriate graphs.
- All this plotting is performed using R studio.
- We are going to perform EDA under two broad conditions:

  1)Descriptive Statistics, which includes mean, median, mode, interquartile range and so on.

  2)Graphical methods, which includes histogram, box plot, scatter plot and so on.

- **Graph of Cereal Name *Vs* Manufacturers**

  Here we plot the scatter plot for identifying the count for a specific manufacturer produce how much cereals....


Oatmeal_Raisin_Crisp
Nutri-grain_Wheat
Nutri-Grain_Almond-Raisin
Nut&Honey_Crunch
Multi-Grain_Cheerios
Mueslix_Crispy_Blend
Maypo
Lucky_Charms
Kix
Honey_Nut_Cheerios
Honey_Graham_Ohs
Honey-comb
Great_Grains_Pecan
Grape_Nuts_Flakes
Grape-Nuts
Golden_Grahams
Golden_Crisp
Fruity_Pebbles
Fruitful_Bran
Frosted_Mini-Wheats
Frosted_Flakes
Froot_Loops
Double_Chex
Crispy_Wheat_&_Raisins
Crispix
Cream_of_Wheat_(Quick)
Cracklin'_Oat_Bran
Count_Chocula
Corn_Pops
Corn_Flakes
Corn_Chex
Cocoa_Puffs
Clusters
Cinnamon_Toast_Crunch
Cheerios
Apple_Jacks
Apple_Cinnamon_Cheerios
Almond_Delight
Cereal_Name
American Home Food Products
General Mills
Kelloggs
Nabisco
Post
Quaker Oats
Ralston Purina
Manufacturer

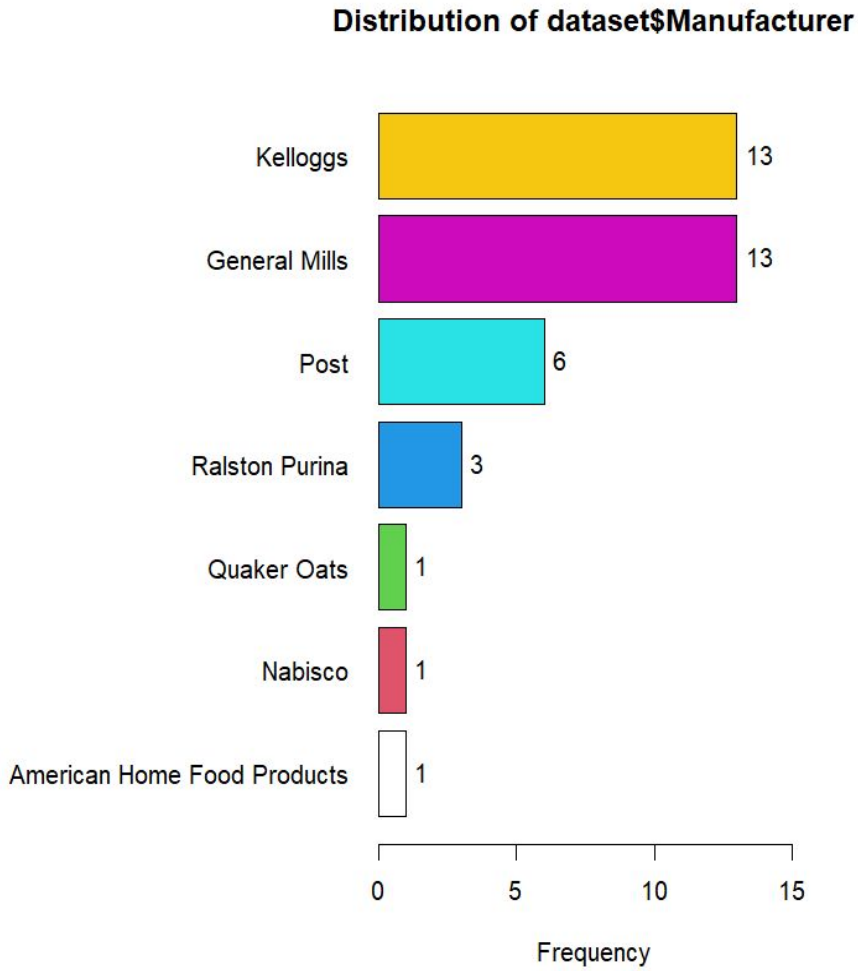
Variable Analysis

- Analysis of Categorical Variable

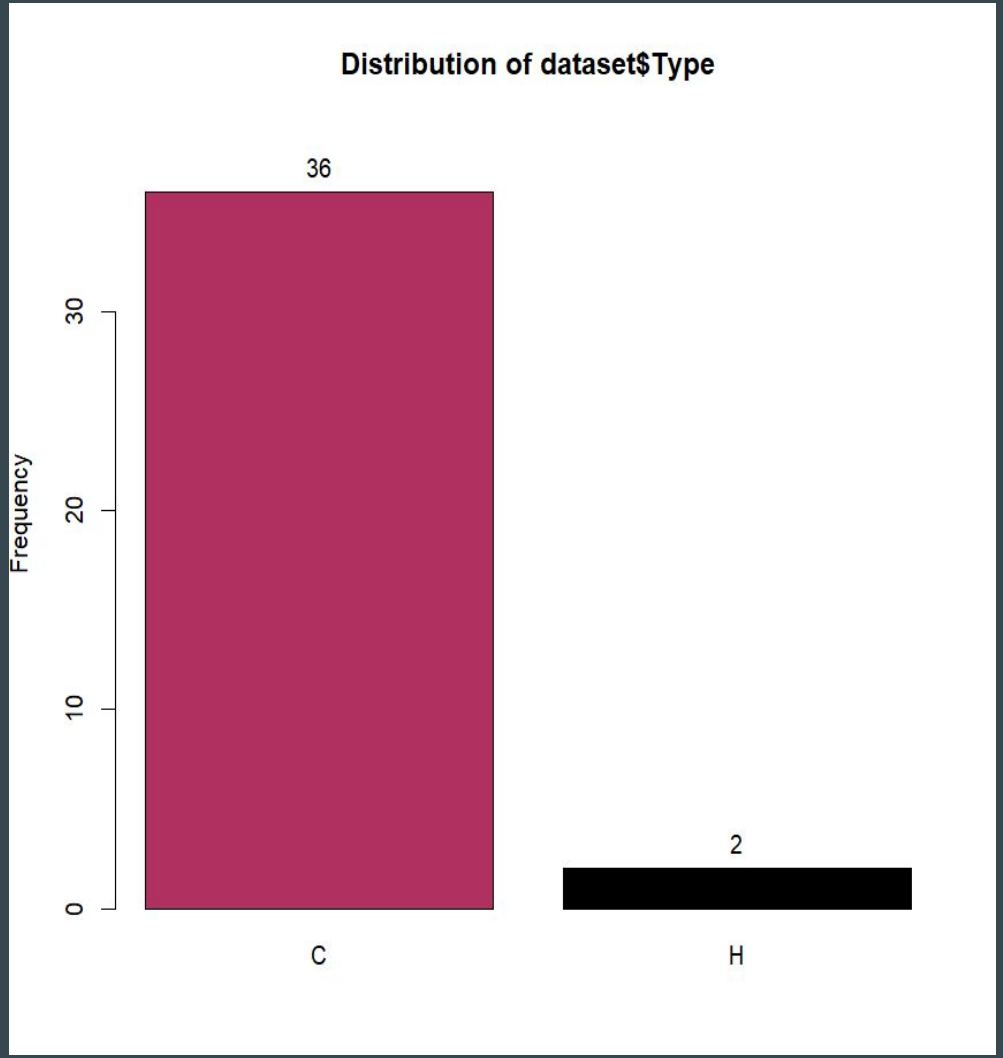
➔ For Manufacturer

➔ For Type

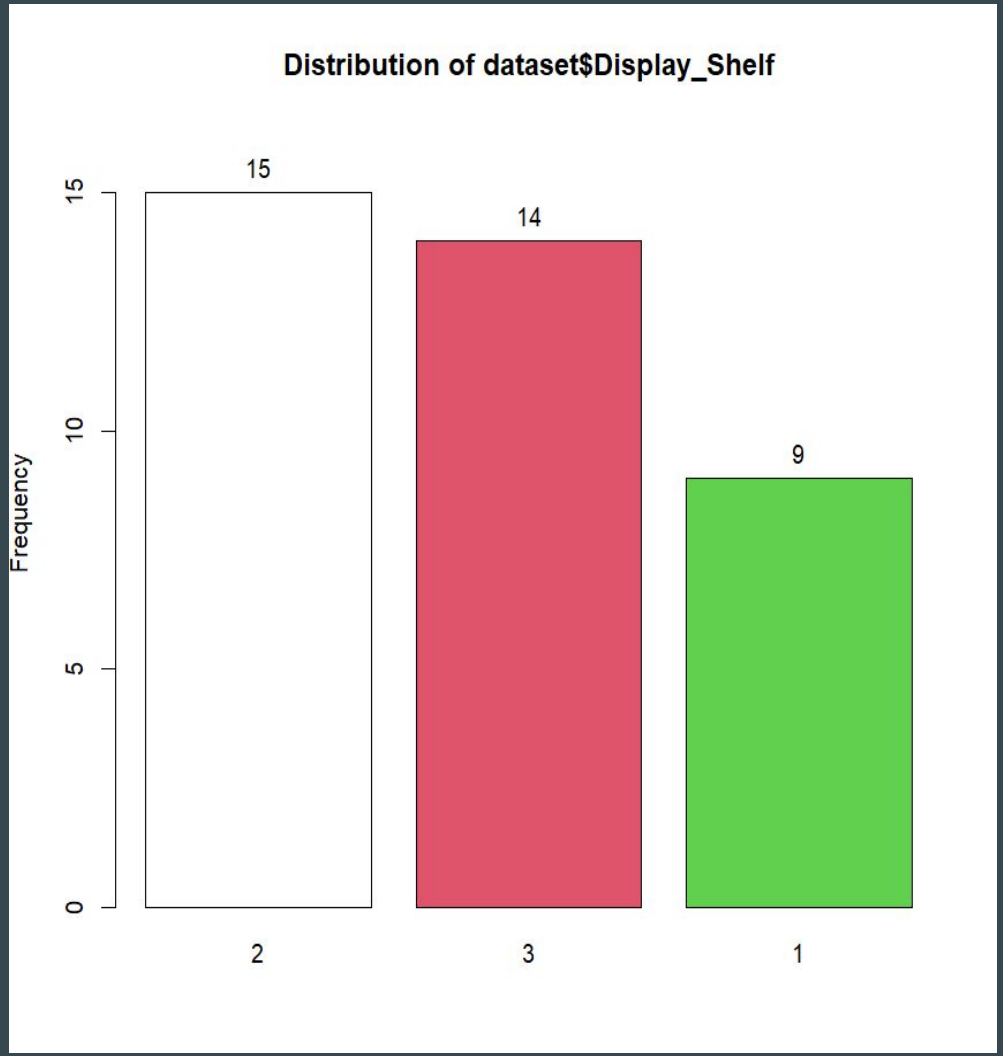
➔ For Display Shelf



- This is the distribution of manufacturer over the different cereal.
- Here, Kellogg's and General Mills manufacturers are producing higher amount of cereal products.
- While, American Home Food , Nabisco and Quaker Oats Manufacturers producing lower amount of cereal product.



- After analysing the data, we seen that only 2 types of cereal product i.e. Maypo and Cream of wheat are Hot (H) Type of cereal.
- While rest of all are Cold (C) Type of cereal.

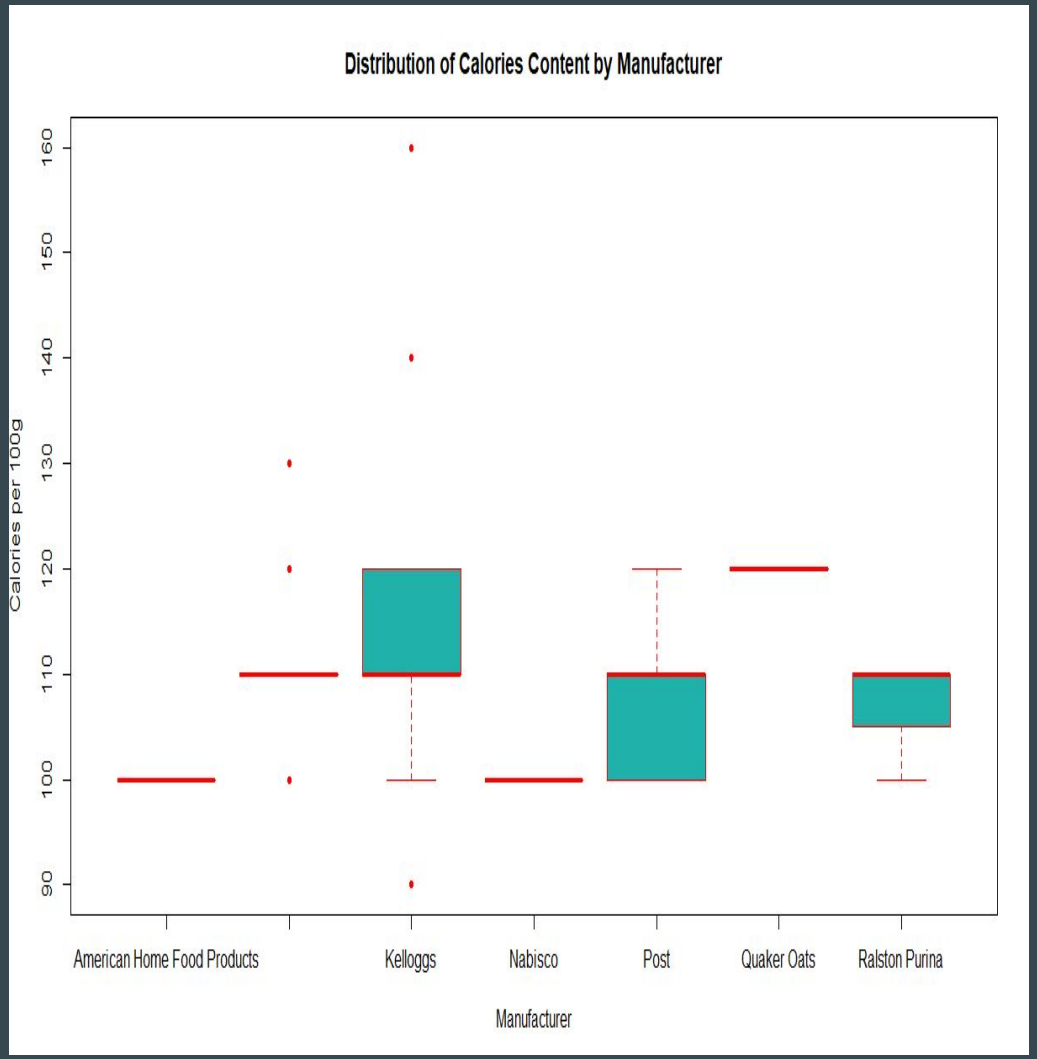


- Here, we see that there are three types of display shelf i.e. 1, 2 and 3.
- Most of the cereal contains 2 Display shelf which is higher amount of shelf amongst other.
- Least number of cereal contain the 1 Display shelf.

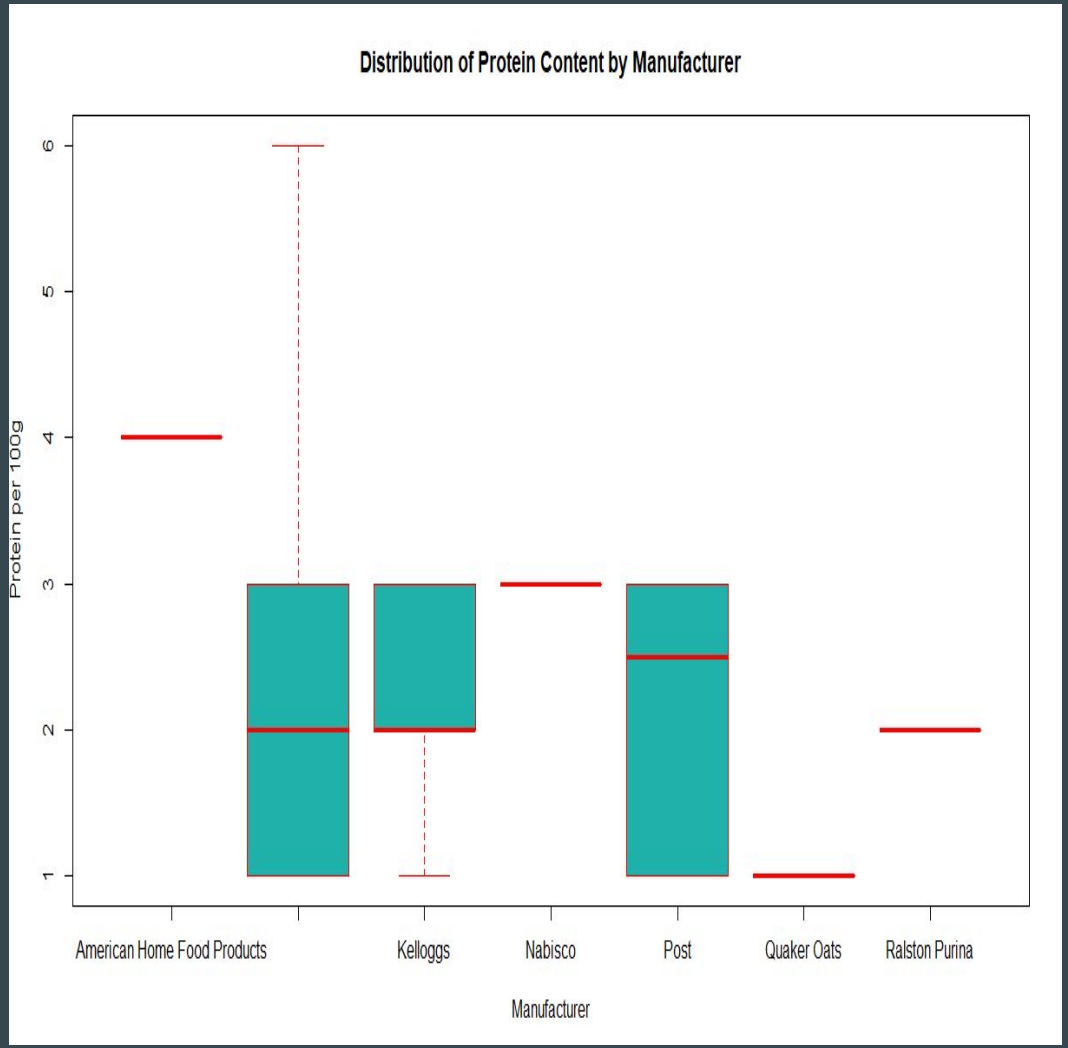
- **Analysis of Numerical Variable**

- ➔ Distribution of Calories content per 100g
- ➔ Distribution of Protein content in g per 100g
- ➔ Distribution of Fat content in g per 100g
- ➔ Distribution of Sodium content in mg per 100g
- ➔ Distribution of Dietary Fiber content in g per 100g
- ➔ Distribution of Carbohydrates content in g per 100g
- ➔ Distribution of Sugars content in g per 100g
- ➔ Distribution of Potassium content in mg per 100g
- ➔ Distribution of Vitamins and Minerals content in g per 100g

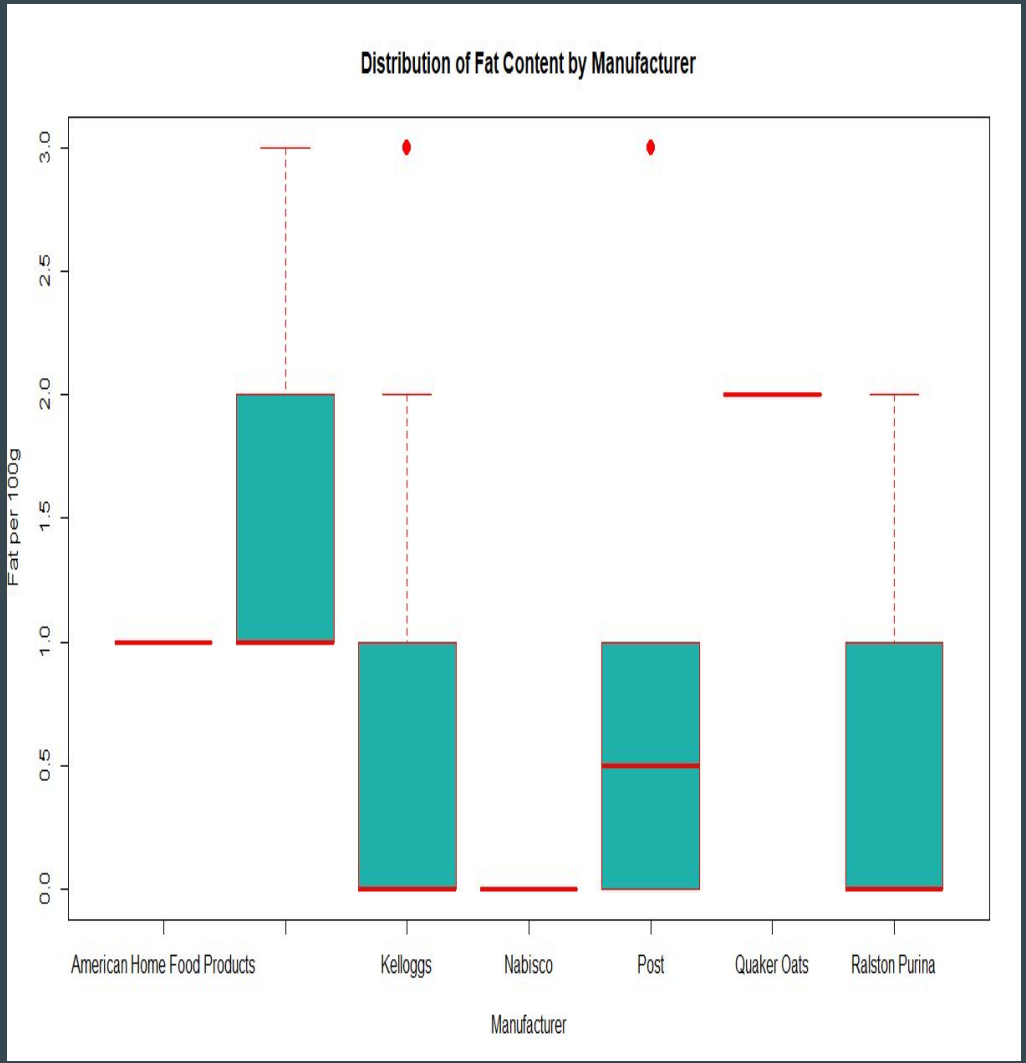
- In fig , Box plot of 'Distribution of Calories content by Manufacturer' is shown.
- In this plot, we analyse that Kellogg's, Post and Ralston purina manufacturer's cereals products contain average number of calories because they have large interquartile range.
- Here, some of the cereals contain highest calories about 160 g which are produced by Kellogg's manufacturer this will be known by the outliers.



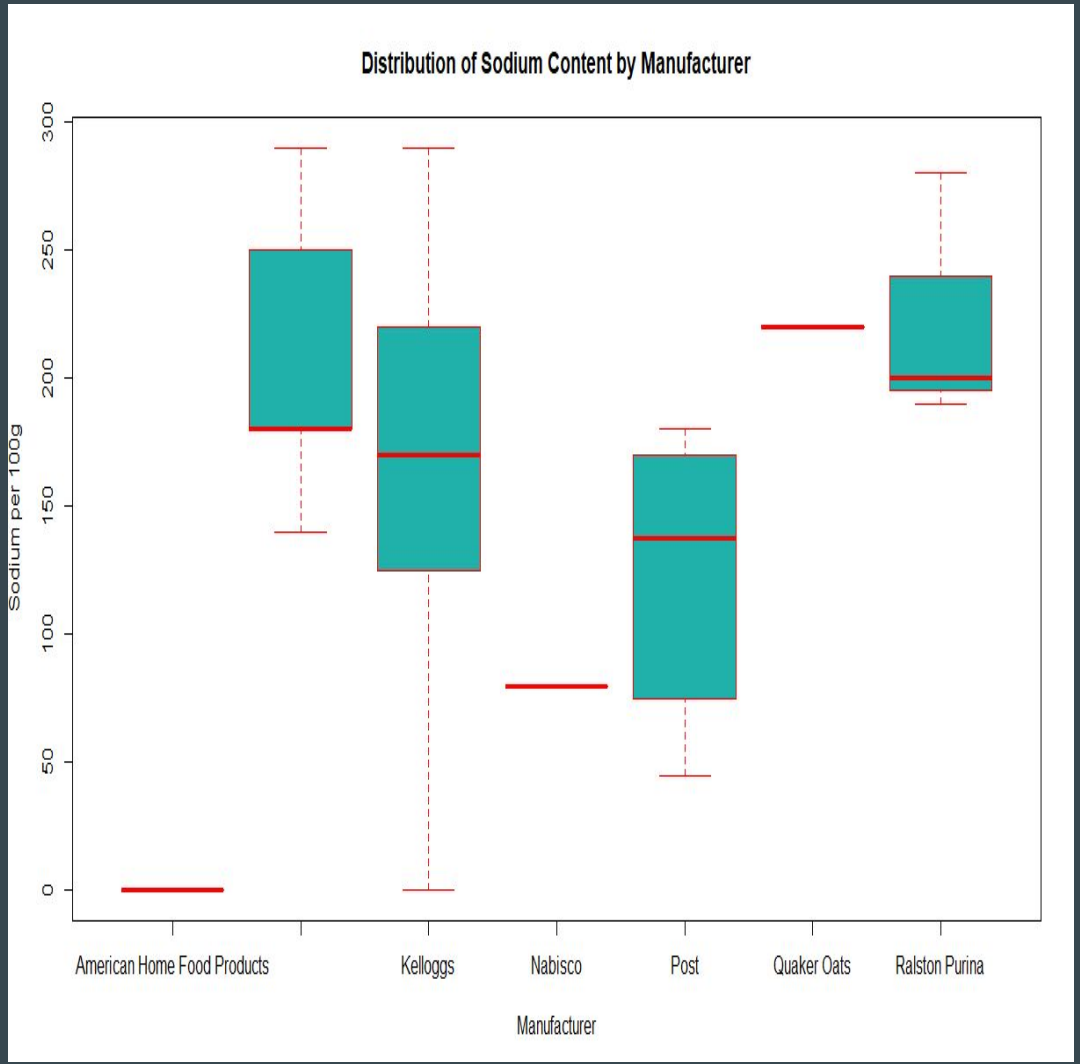
- In fig , Box plot of 'Distribution of Protein content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's and Post manufacturer's cereal product contain average or mean number of protein values in because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of protein about 6 g which are produced by General mill manufacturer and we identify it by Quartile range.



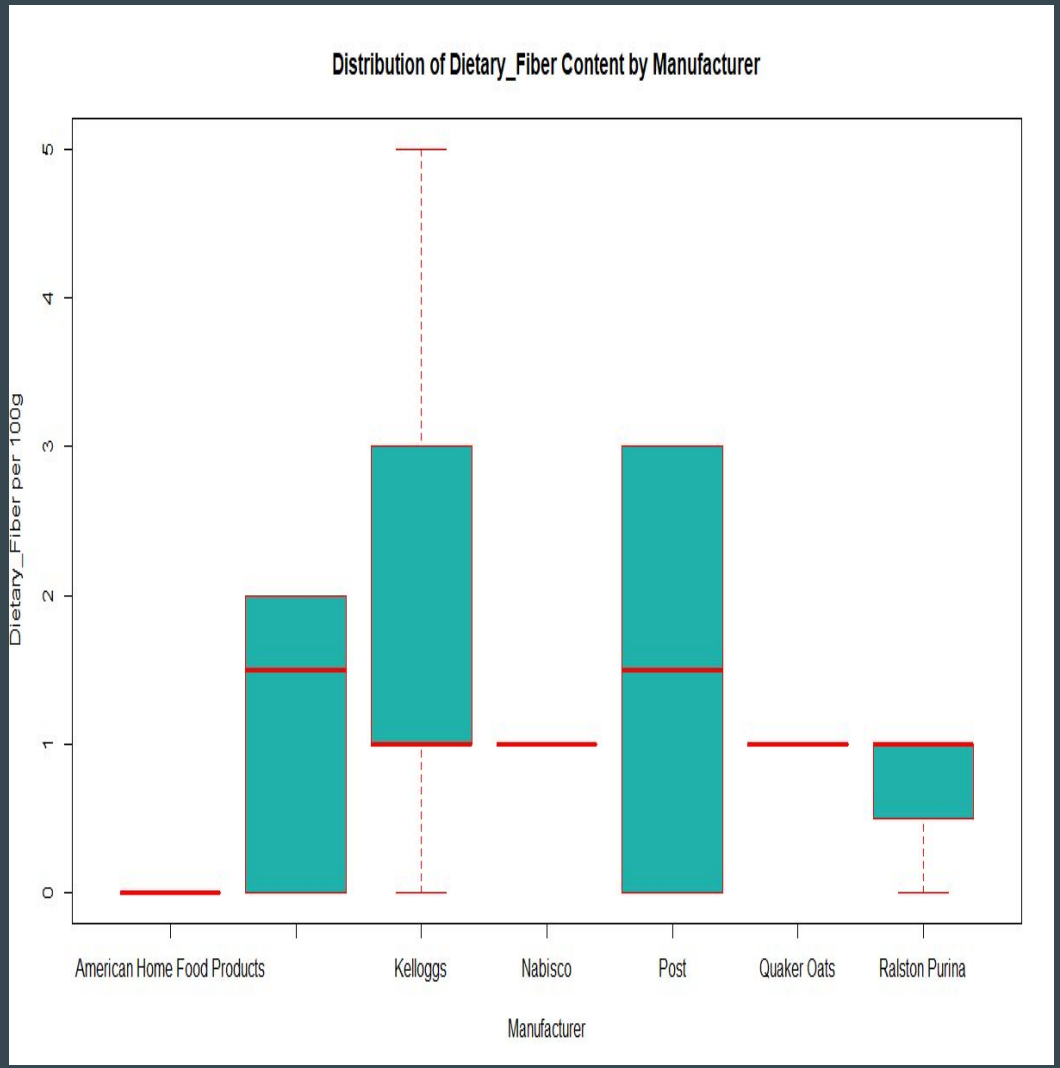
- In fig , Box plot of 'Distribution of Fat content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of fat content because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of fat about 30 g which are produced by General mill, Kellogg's and Post manufacturer and we identify it by Quartile range as well as by outliers.



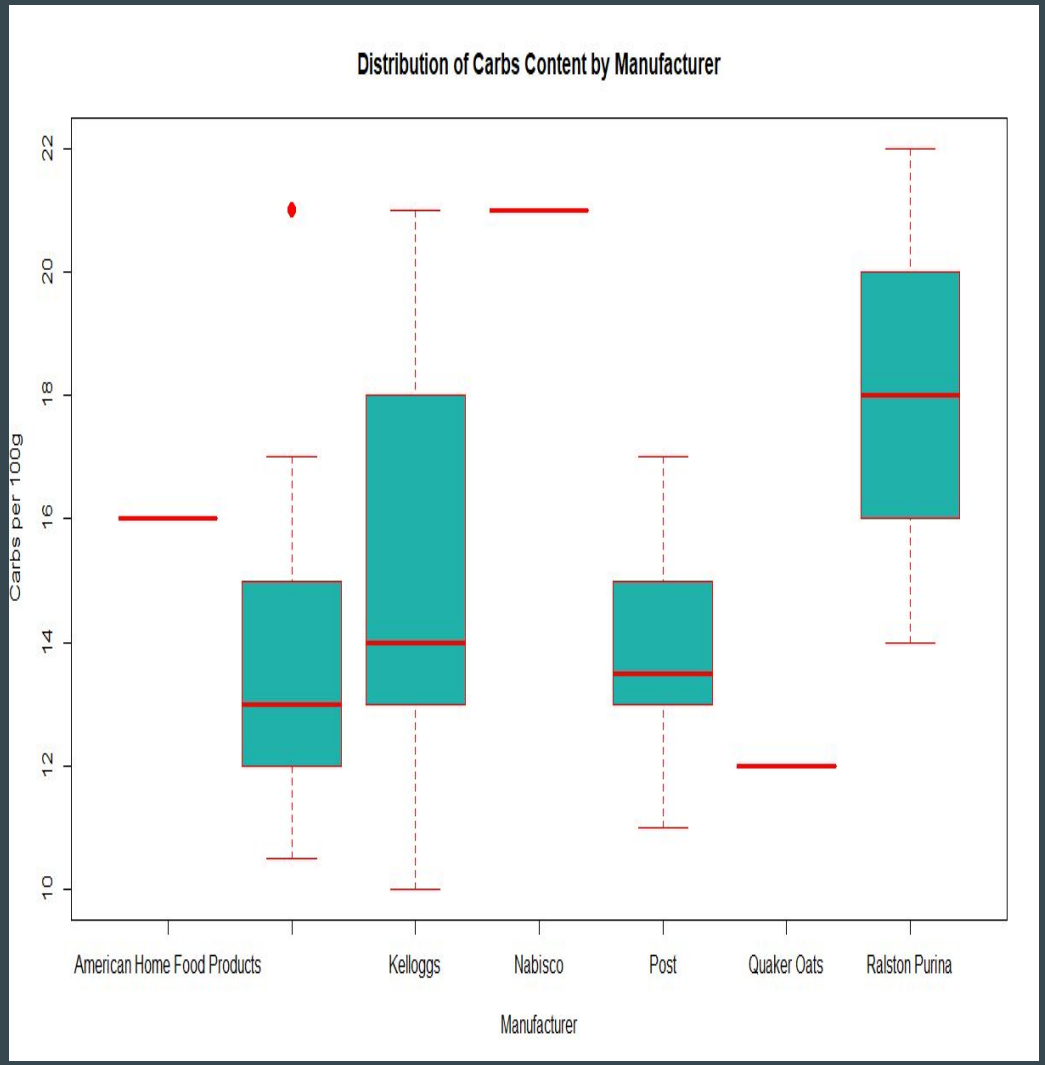
- In fig , Box plot of 'Distribution of Sodium content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of sodium content because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of sodium about 290 mg which are produced by General mill and Kellogg's manufacturer and we identify it by Quartile range.



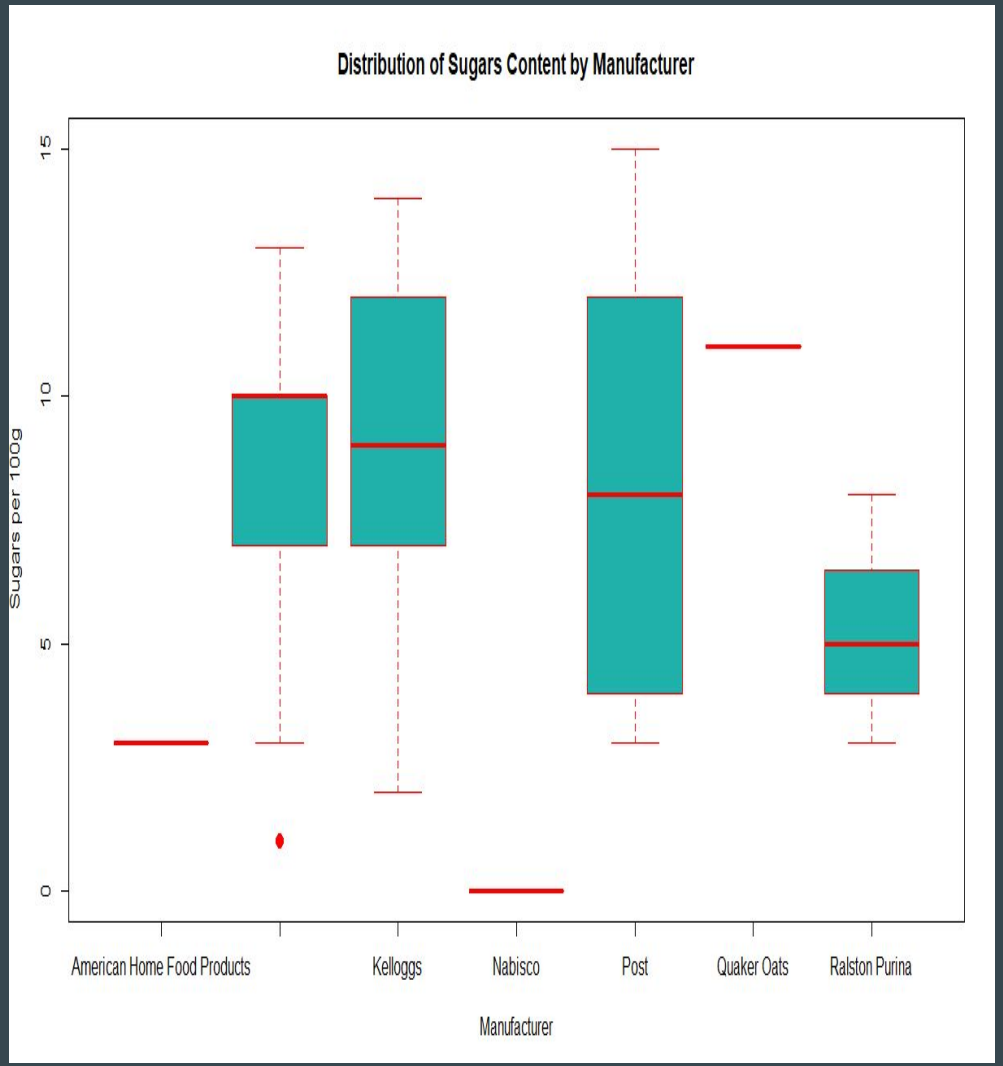
- In fig , Box plot of 'Distribution of Dietary Fiber content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of dietary fiber content because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of dietary fiber about 5 g which are produced by Kellogg's manufacturer and we identify it by Quartile range.



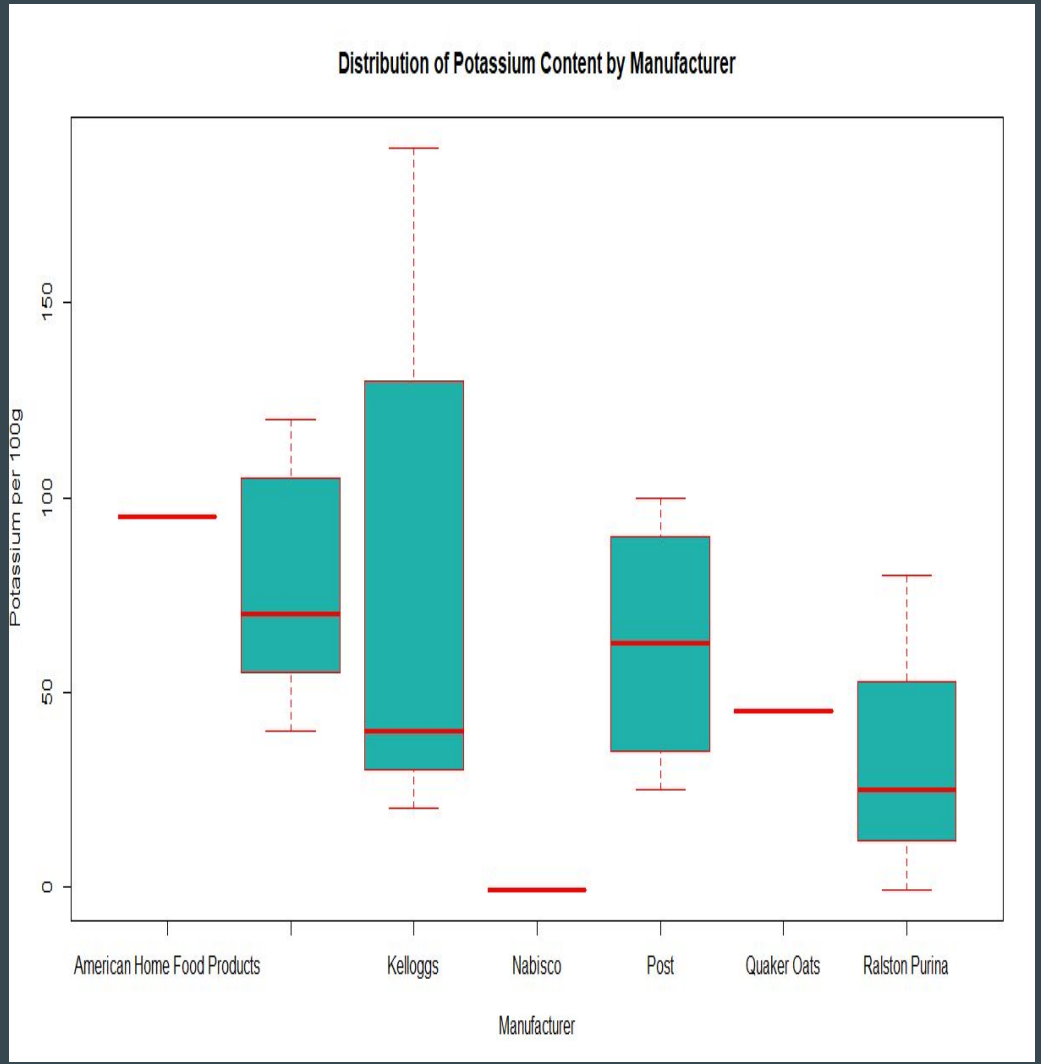
- In fig , Box plot of 'Distribution of Carbohydrates or Carbs content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of carbohydrates because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of carbohydrates about 22 g which are produced by Ralston purina manufacturer and we identify it by Quartile range.



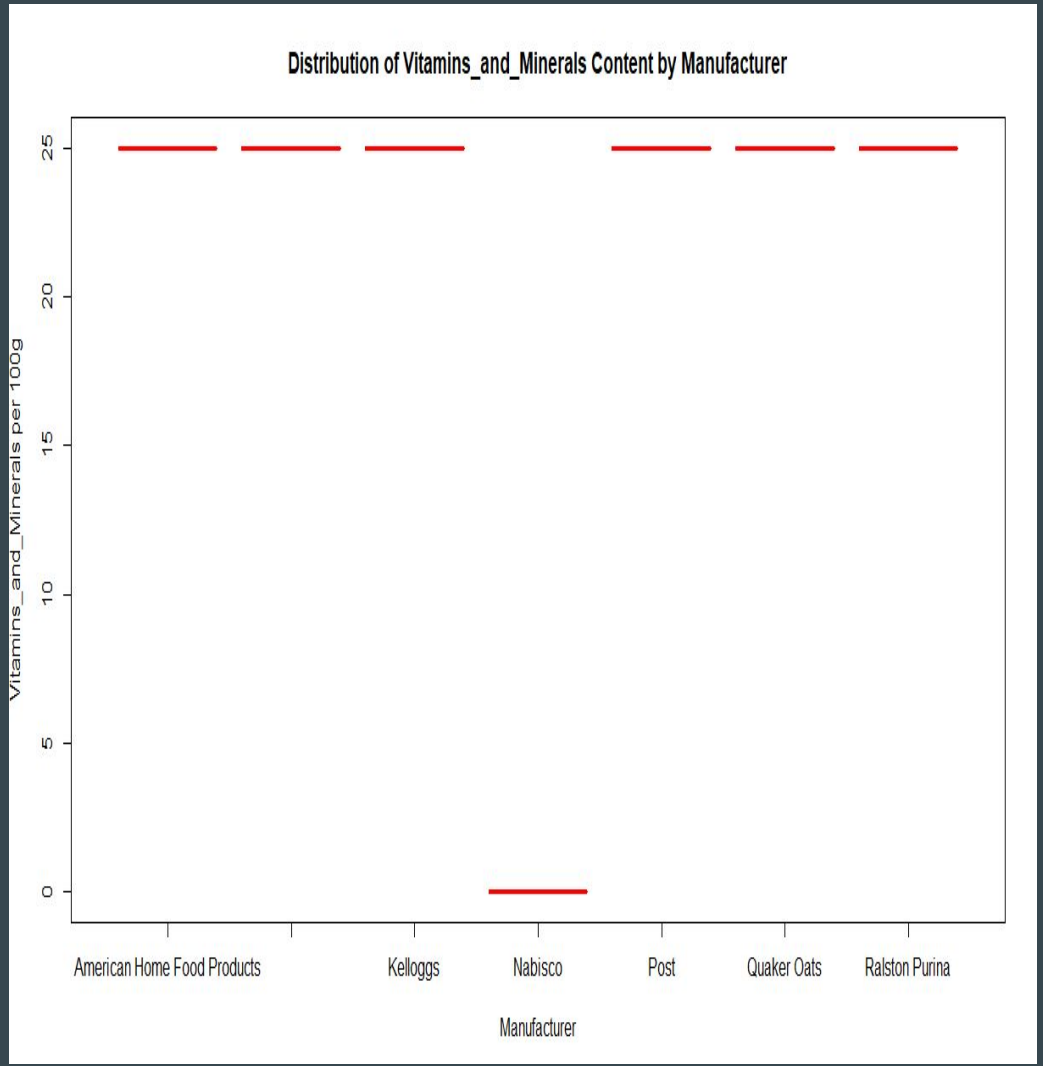
- In fig , Box plot of 'Distribution of Sugars content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of sugars content because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of sugars about 15 g which are produced by Post manufacturer and we identify it by Quartile range.



- In fig , Box plot of 'Distribution of Potassium content by Manufacturer' is shown.
- In this plot, we analyse that General mill, Kellogg's, Post and Ralston purina manufacturer's cereal products contain average or mean number of potassium content because they have large number of interquartile range.
- Here, some of the cereals contain highest amount of potassium about 190 mg which are produced by Kellogg's manufacturer and we identify it by Quartile range.



- In fig , Box plot of 'Distribution of Vitamins and Minerals content by Manufacturer' is shown.
- Here, all manufacturer products contain



★ Results and Discussions

- After analysing whole data, we understood that EDA is very helpful to analysing any type of dataset.
- Here we explored cereal data using two methods i.e by analysis of categorical variable and by analysis of numerical variable.

★ Conclusion

- ➔ We conclude that the Kellogg's manufacturer is a Most convenient manufacturer to produce nutritional cereals.
- ➔ After Kellogg's, General Mills products also produces the cereal products having good amount nutritional values.
- ➔ So, here we conclude that nutritional cereals are good for health and if you want to buy cereals, then go for Kellogg's and General Mills.