STAT 35500 Project Report

Matthew Sindac

December 1, 2023

1 Summary

This technical report delves into the exploration of the Cereal Dataset, which includes various nutritional components of cereals from different manufacturers and shelves. With the data set containing multiple nutritional variables for a range of cereals, the objective is to uncover patterns, differences, and potential relationships within the data set.

The analyses conducted in this project offer key insights into the nutritional content cereals. Manufacturer-specific and shelf-specific trends were explored, showcasing the variations in nutritional content. Distribution patterns across manufacturers and shelves have been visualized to provide a visually pleasing view of the data set.

In summary, this report serves as a thorough exploration of the Cereal Dataset, providing valuable insights into nutritional patterns and distribution across manufacturers and shelves. The comprehensive nature of this analysis contributes to the broader understanding of nutritional content in cereals.

2 Introduction

The Cereal Dataset contains nutritional information across a many different cereals. With the data set containing variables such as calories, protein, fat, sodium, fiber, complex carbs, sugars, potassium, and ratings, our objective is find relationships and variations within this data set. Cereals, being a staple in many households, play a significant role in dietary choices. Understanding the nutritional content across different manufacturers and shelves not only provides consumers with valuable insights for their own decisions but also offers manufacturers the opportunity to make products that consumers enjoy. Through our exploration of this data set, we will create interpretations that can contribute to the broader field of cereal.

3 Methodology

In this project, I explored a data set containing nutrition data for various cereal products. The data set is stored in a file named "cereal.csv" and includes essential nutritional components such as calories, protein, fat, sodium, fiber, complex carbohydrates, sugars, potassium, and consumer ratings.

3.1 Initial Exploration

I began by examining key variables in the data set, providing descriptions, summary statistics, and visual representations.

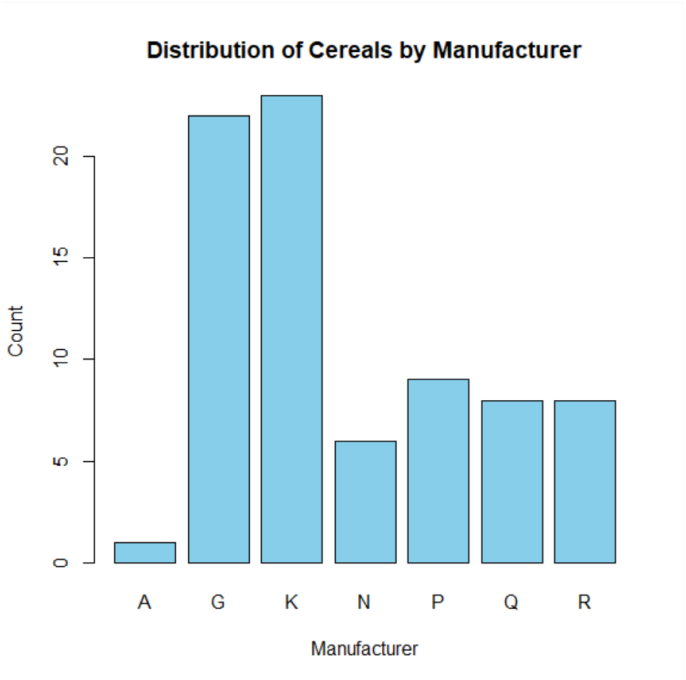
3.1.1 Variable: 'name'

This variable contains the name of each cereal. There are 77 unique cereal names within this data set.

3.1.2 Variable: 'mfr'

Count of Cereals Per Manufacturer

A	G	K	N	P	Q	R
1	22	23	6	9	8	8



To understand the distribution of cereals among manufacturers, I examined the count of cereals per manufacturer. The manufacturers ('mfr') and their corresponding codes are as follows:

- A = American Home Food Products
- G = General Mills
- K = Kelloggs
- N = Nabisco
- P = Post
- Q = Quaker Oats
- R = Ralston Purina

It's notable that General Mills and Kelloggs appear to produce a more extensive variety of cereals, while American Home Food Products produces the least unique types. Nabisco, Post, Quaker Oats, and Ralston Purina show a relatively similar production volume.

3.1.3 Variable: 'type'

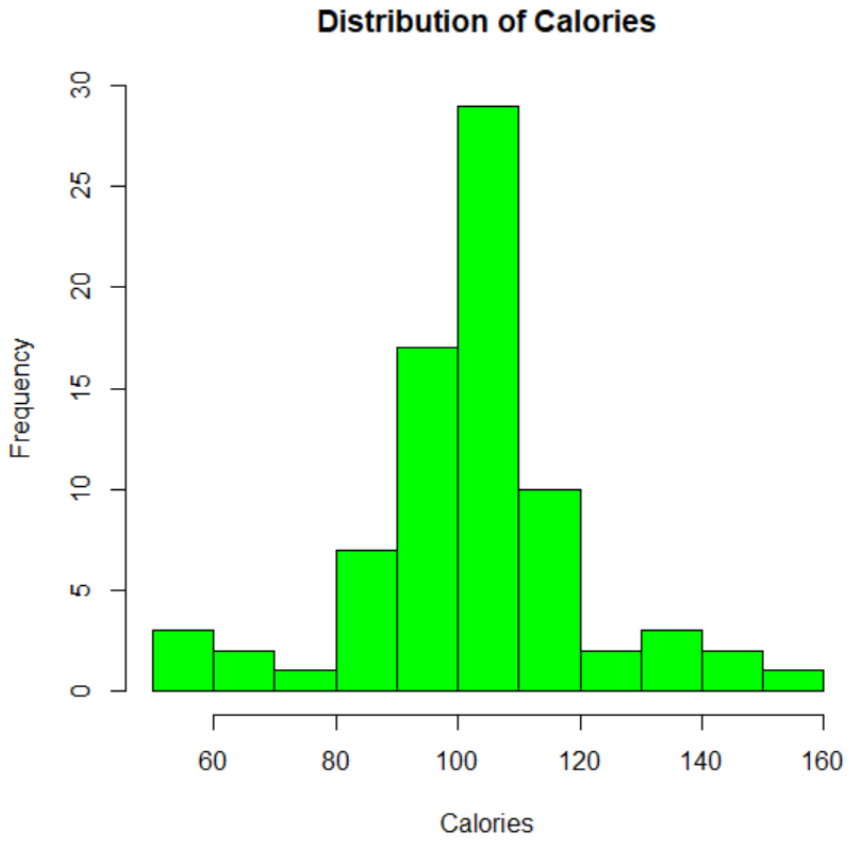
C	H
74	3

The 'type' variable categorizes cereals as either cold (C) or hot (H). Out of the analyzed cereals, 74 are cold cereals, while only 3 are hot cereals.

3.1.4 Variable: 'calories'

Calories Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
50.0	100.0	110.0	106.9	110.0	160.0

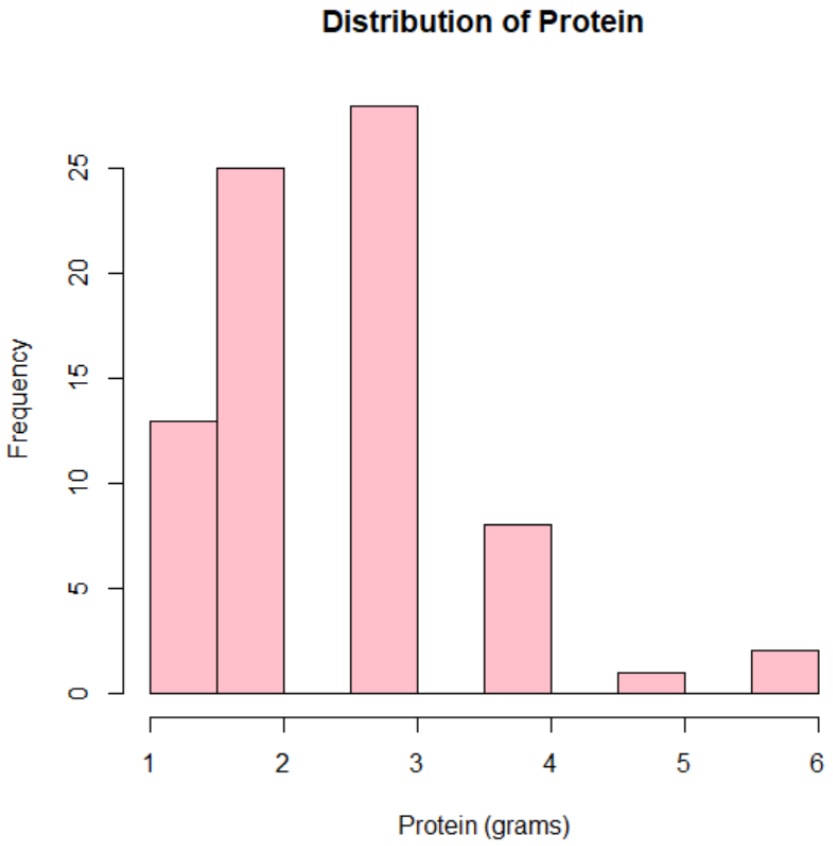


The 'calories' variable represents the calories per serving for each cereal. The distribution of calories appears to follow a nearly normal pattern, ranging from a minimum of 50 calories per serving to a maximum of 160 calories per serving. The median (110.0) and mean (106.9), provide insights into the centrality of the calorie values.

3.1.5 Variable: 'protein'

Grams of Protein Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
1.000	2.000	3.000	2.545	3.000	6.000

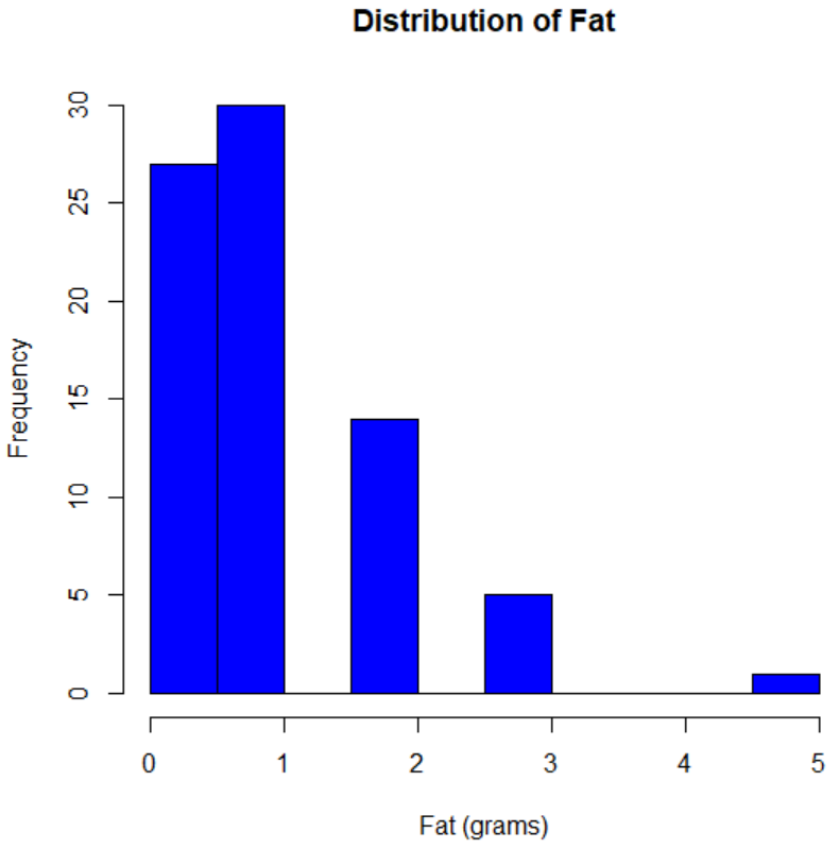


The 'protein' variable represents the grams of protein per serving for each cereal. The distribution of protein content appears to be right-skewed, with values ranging from a minimum of 1 gram to a maximum of 5 grams per serving. The median (3.000) and mean (2.545), show us the central tendencies of the protein values. The histogram visually illustrates the distribution which is skewed towards the lower end.

3.1.6 Variable: 'fat'

Grams of Fat Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.000	0.000	1.000	1.013	2.000	5.000

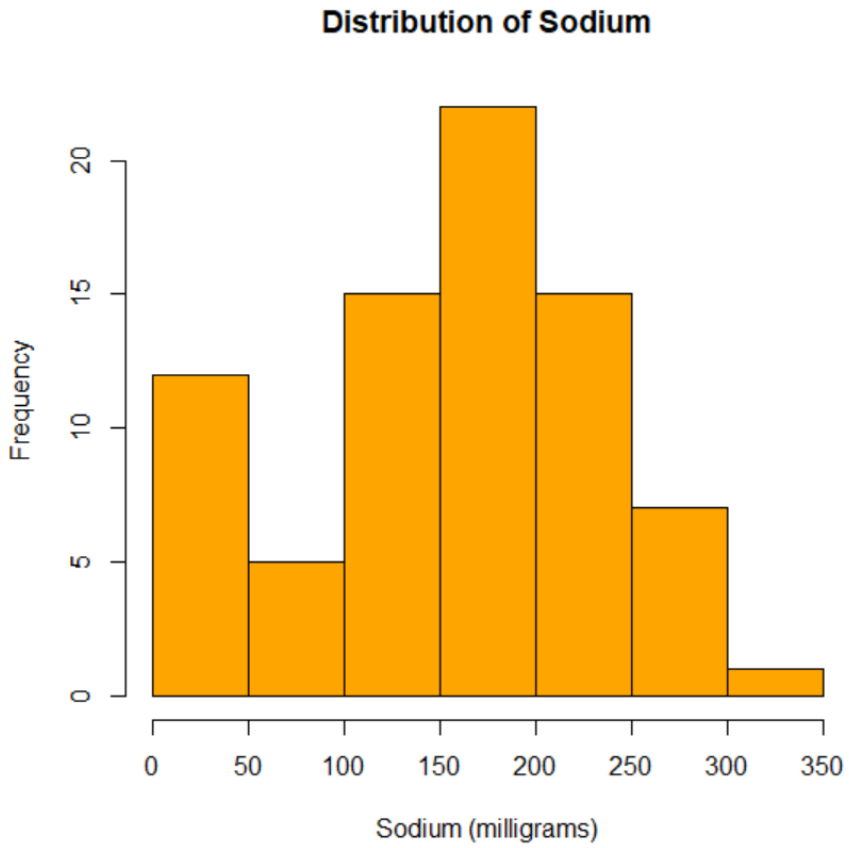


The 'fat' variable represents the grams of fat per serving for each cereal. The distribution of fat content appears to be right-skewed, ranging from a minimum of 0 grams to a maximum of 6 grams per serving. The median (1.000) and mean (1.013), provide insights into the centrality of the fat values. The histogram visually depicts the distribution which is skewed towards the lower end.

3.1.7 Variable: 'sodium'

Milligrams of Sodium Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.0	130.0	180.0	159.7	210.0	320.0

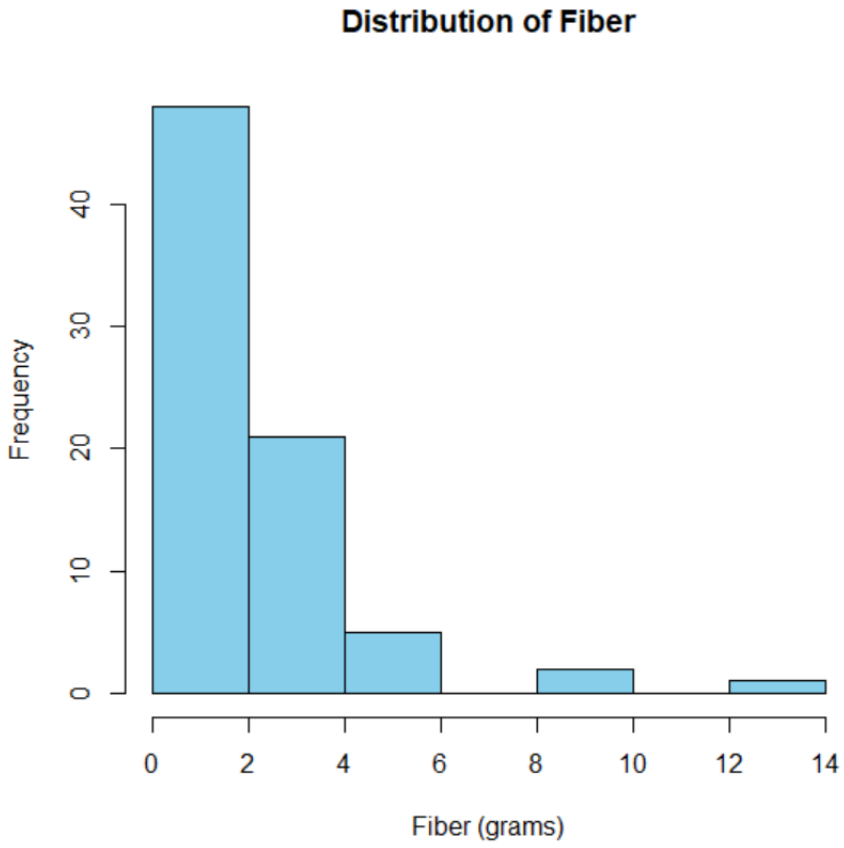


The 'sodium' variable represents the milligrams of sodium per serving for each cereal. The distribution of sodium content appears to be approximately normal, with a peak in the 0-50 milligrams bin. The values range from a minimum of 0 milligrams to a maximum of 320 milligrams per serving. The median (180.0) and mean (159.7), which helps us understand the centrality of the sodium values. The histogram portrays the distribution which emphasizes a concentration in the lower milligram range.

3.1.8 Variable: 'fiber'

Grams of Dietary Fiber Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.000	1.000	2.000	2.152	3.000	14.000

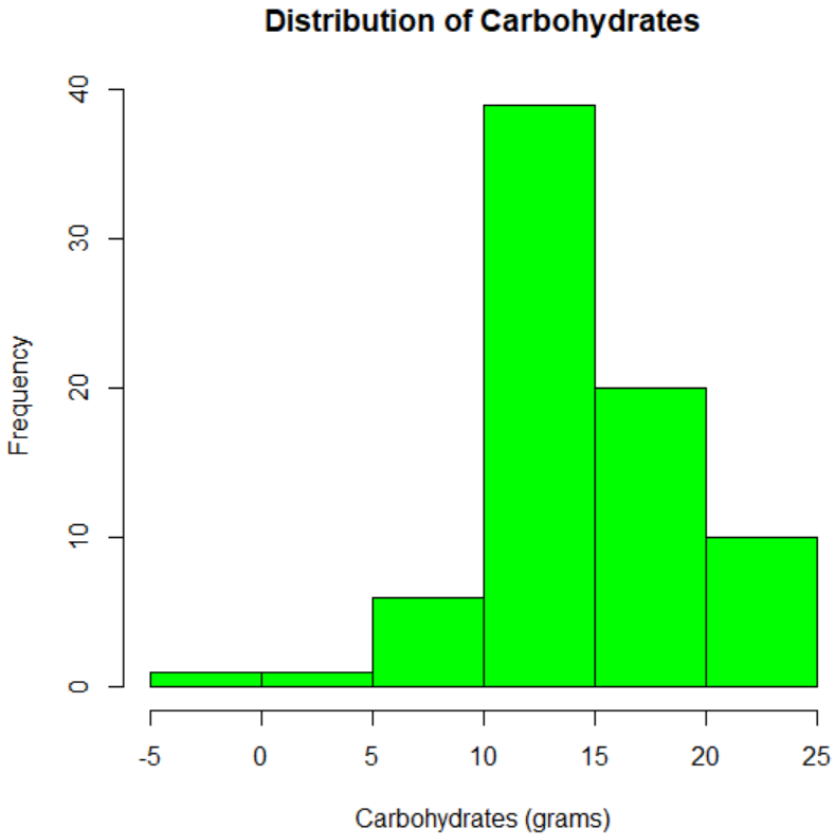


The 'fiber' variable represents the grams of dietary fiber per serving for each cereal. The distribution of fiber content appears to be right-skewed, ranging from a minimum of 0 grams to a maximum of 14 grams per serving. The median (2.000) and mean (2.152), provide insights into the central tendency of the fiber values. The histogram illustrates the distribution which is skewed towards the lower end.

3.1.9 Variable: 'carbo'

Grams of Complex Carbohydrates Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
-1.0	12.0	14.0	14.6	17.0	23.0

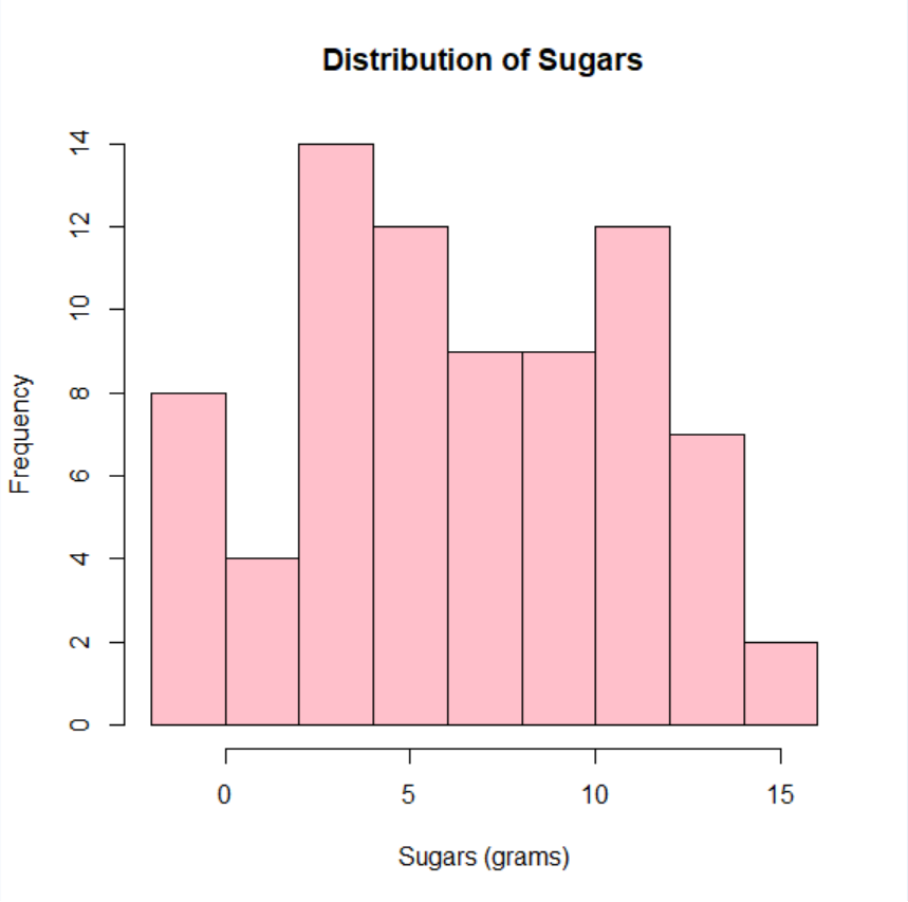


The 'carbo' variable represents the grams of complex carbohydrates per serving for each cereal. The distribution of carbohydrate content appears to have a left skew, ranging from a minimum of -1 gram to a maximum of 23 grams per serving. The presence of a minimum value of -1 throughout the dataset is to be interpreted as an extremely small amount, approaching zero but not exactly zero. The median (14.0) and mean (14.6), provide insights into the centrality of the carbohydrate values. The histogram visually depicts the distribution which is shown to skew towards the higher end.

3.1.10 Variable: 'sugars'

Grams of Sugars Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
-1.000	3.000	7.000	6.922	11.000	15.000

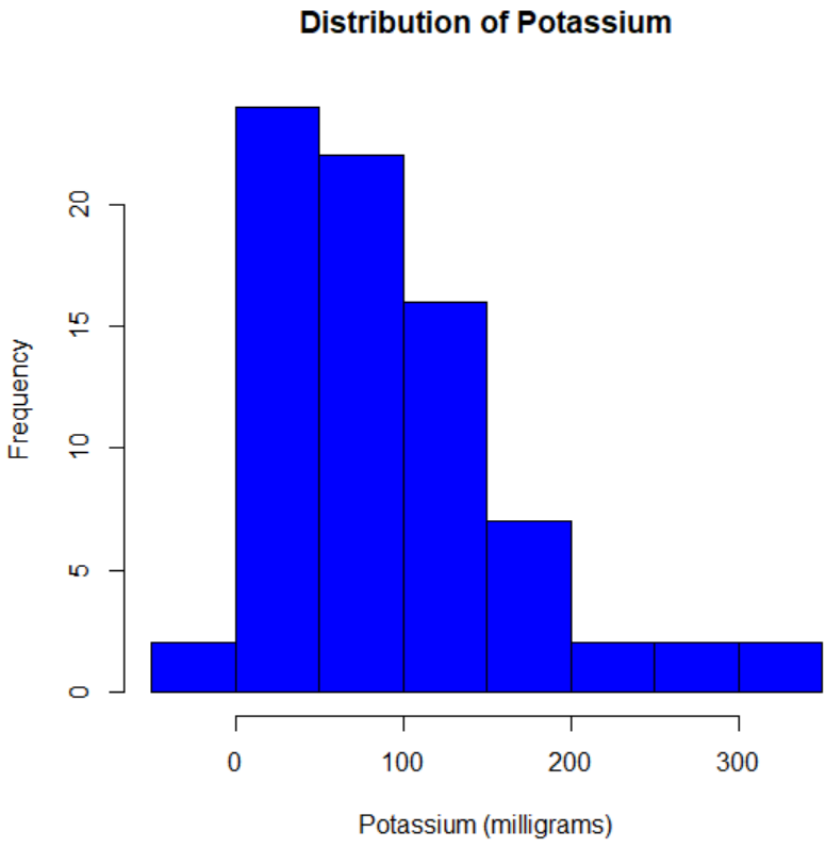


The 'sugars' variable represents the grams of sugar per serving for each cereal. The distribution of sugar content appears to be approximately normal, featuring several peaks throughout the dataset. The values range from a minimum of -1 gram to a maximum of 15 grams per serving. Similar to previous variables, the minimum value of -1 is to be interpreted as an extremely small amount, almost zero but not exactly zero. The median (7.000) and mean (6.922), provide insights into the centrality of the sugar values. The histogram illustrates the distribution, highlighting the wide range of sugar content, showing multiple peaks within the data.

3.1.11 Variable: 'potass'

Milligrams of Potassium Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
-1.00	40.00	90.00	96.08	120.00	330.00

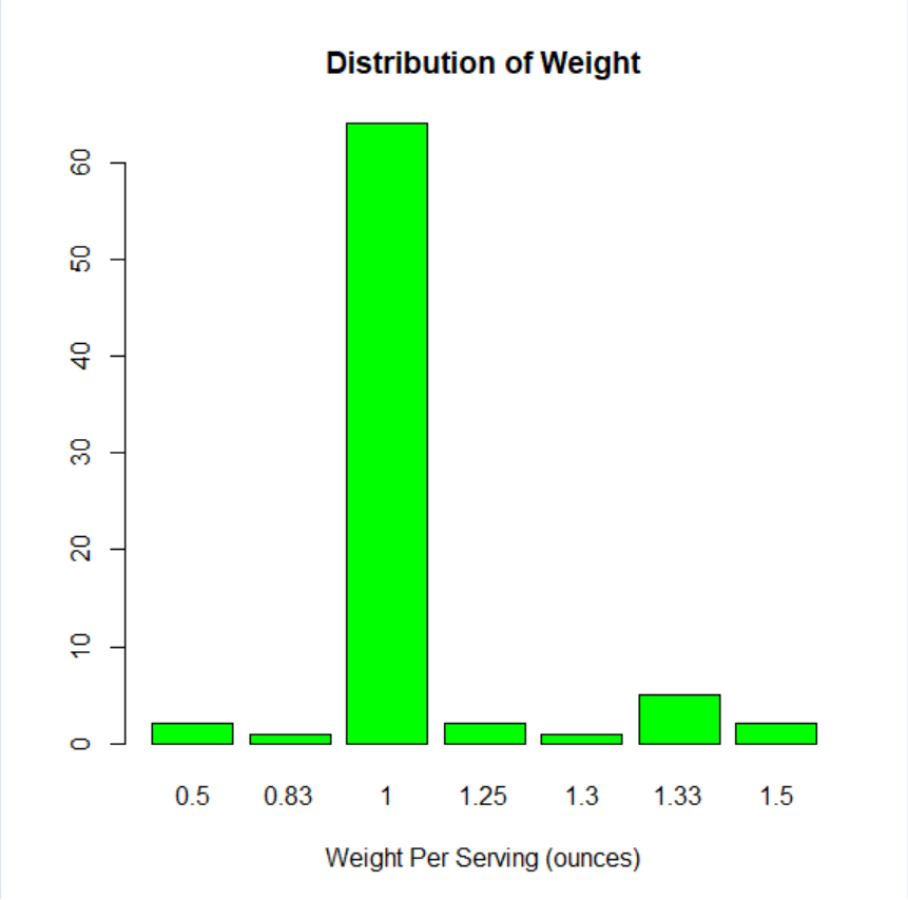


The 'potass' variable represents the milligrams of potassium per serving for each cereal. The distribution of potassium content appears to have a right skew, ranging from a minimum of -1 milligrams to a maximum of 330 milligrams per serving. Similar to previous variables, the minimum value of -1 is to be interpreted as an extremely small amount, approaching zero but not exactly zero. The median (90.00) and mean (96.08) showcase how the potassium values are centered. The histogram showcases the distribution, emphasizing the various levels of potassium content which is skewed towards the lower end.

3.1.12 Variable: 'weight'

Weight Per Serving (in ounces)

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.50	1.00	1.00	1.03	1.00	1.50

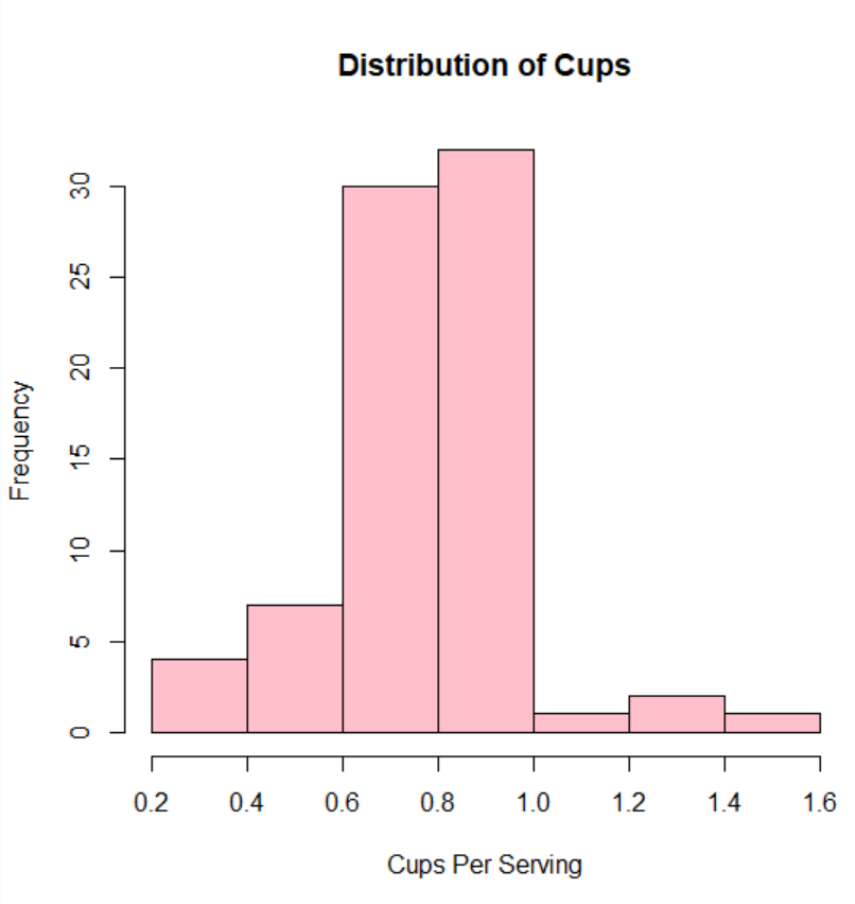


The 'weight' variable represents the weight per serving in ounces for each cereal. The distribution of serving weights appears to be relatively consistent, with values ranging from a minimum of 0.50 ounces to a maximum of 1.50 ounces per serving. The median (1.00) and mean (1.03), indicate that the majority of cereals have a serving size of about 1 ounce. The bar plot visually represents the distribution where the general trend of serving sizes center around the 1-ounce mark for each cereal.

3.1.13 Variable: 'cups'

Cups Per Serving

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
0.250	0.670	0.750	0.821	1.000	1.500

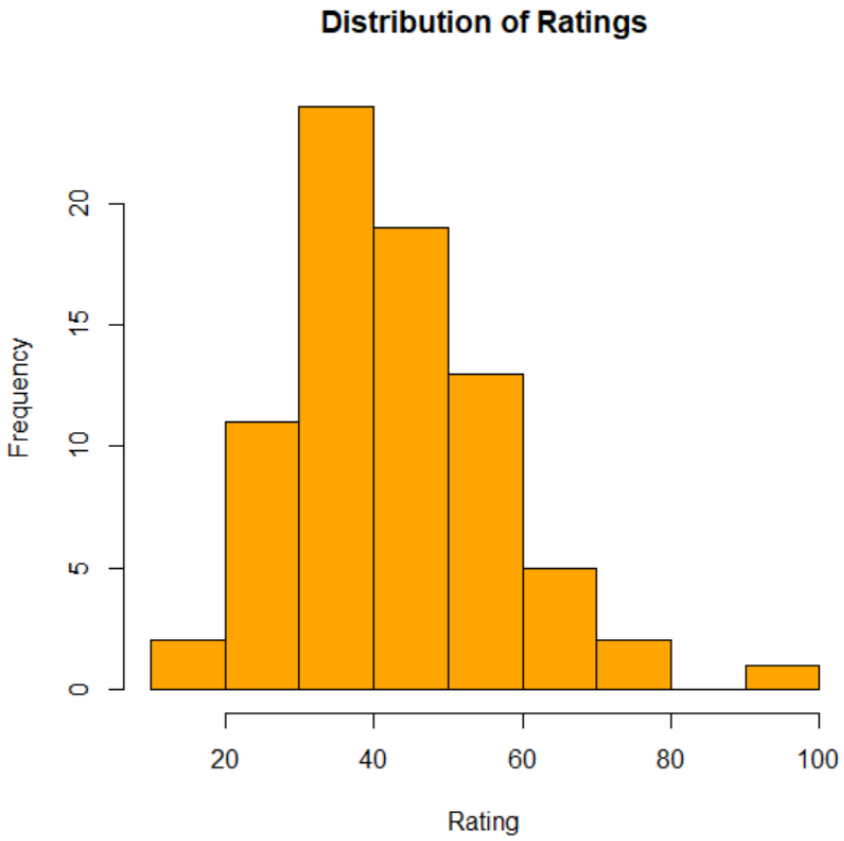


The 'cups' variable represents the amount of cups per serving for each cereal. The distribution of cups per serving appears to be approaching normal, with values ranging from a minimum of 0.250 cups to a maximum of 1.500 cups. The median (0.750) and mean (0.821) provide insights into how this data is centered. The histogram visually illustrates the distribution, indicating a trend towards a normal distribution with a range of cup quantities per serving for each cereal.

3.1.14 Variable: 'rating'

Ratings of Cereals (Consumer Reports)

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
18.04	33.17	40.40	42.67	50.83	93.70

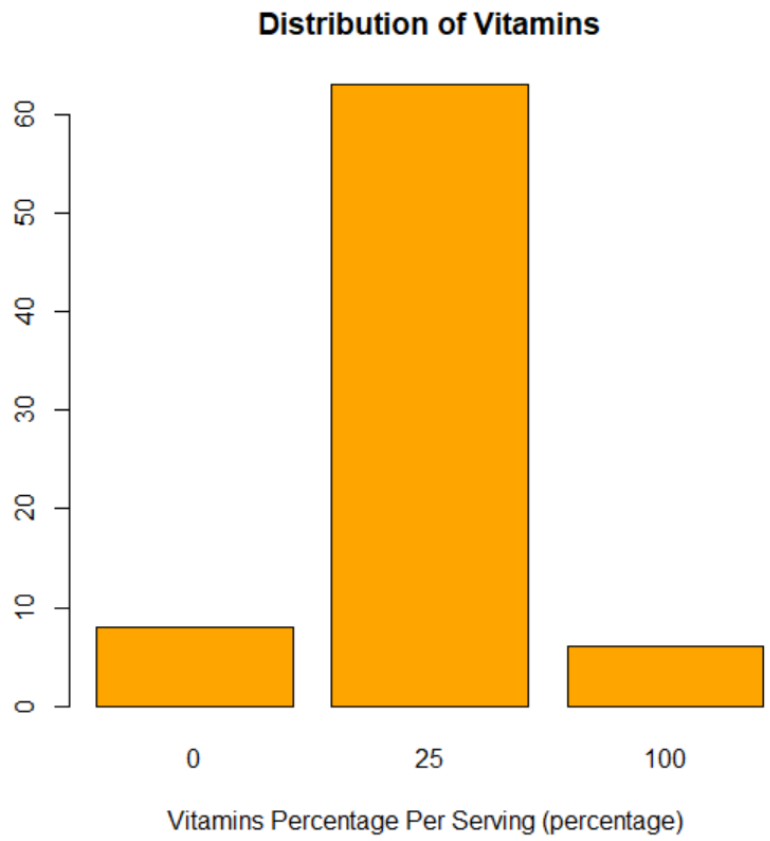


The 'rating' variable represents the consumer ratings of each cereal based on Consumer Reports. The distribution of ratings appears to be approximately normal, with values ranging from a minimum of 18.04 to a maximum of 93.70. The median (40.40) and mean (42.67) show the central tendencies of the cereal ratings. The histogram showcases the distribution, highlighting the normal distribution pattern with a wide range of consumer ratings for the cereals.

3.1.15 Variable: 'vitamins'

Vitamins and Minerals Per Serving

0%	25%	100%
8	63	6

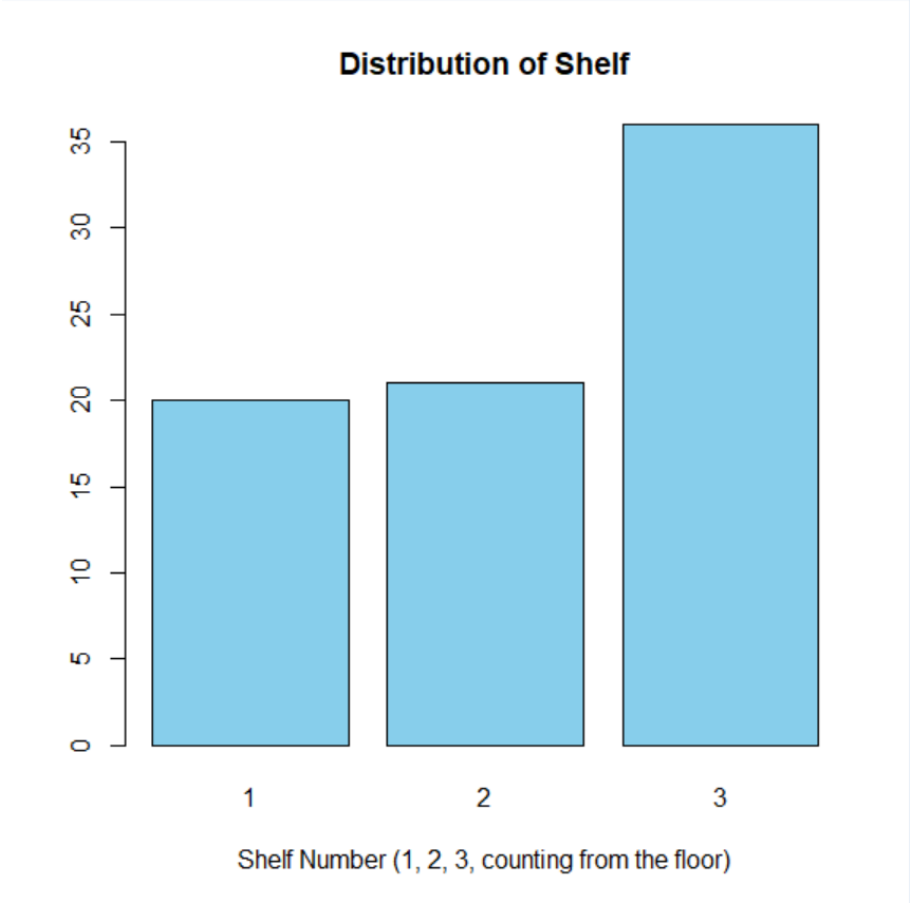


The 'vitamins' variable represents the percentage of vitamins and minerals per serving within each cereal, categorized as 0%, 25%, and 100% of the recommended daily value according to FDA standards. The distribution of these percentages is captured by the bar plot, showcasing values of 8, 63, and 6 for 0%, 25%, and 100% respectively. This variable provides insights into the nutritional content of the cereals, indicating how they contribute to the recommended daily intake of vitamins and minerals.

3.1.16 Variable: 'shelf'

Display Shelf (1, 2, or 3, counting from floor)

1	2	3
20	21	36



The 'shelf' variable represents the display shelf number (1, 2, or 3, counting from the floor) for each cereal. The distribution of cereals across shelves is visualized through the bar plot, indicating counts of 20, 21, and 36 for shelves 1, 2, and 3 respectively. This variable provides insights into the placement distribution of cereals on the shelves, with a higher concentration observed on shelf 3 compared to shelves 1 and 2.

3.2 Additional Exploration and Insights

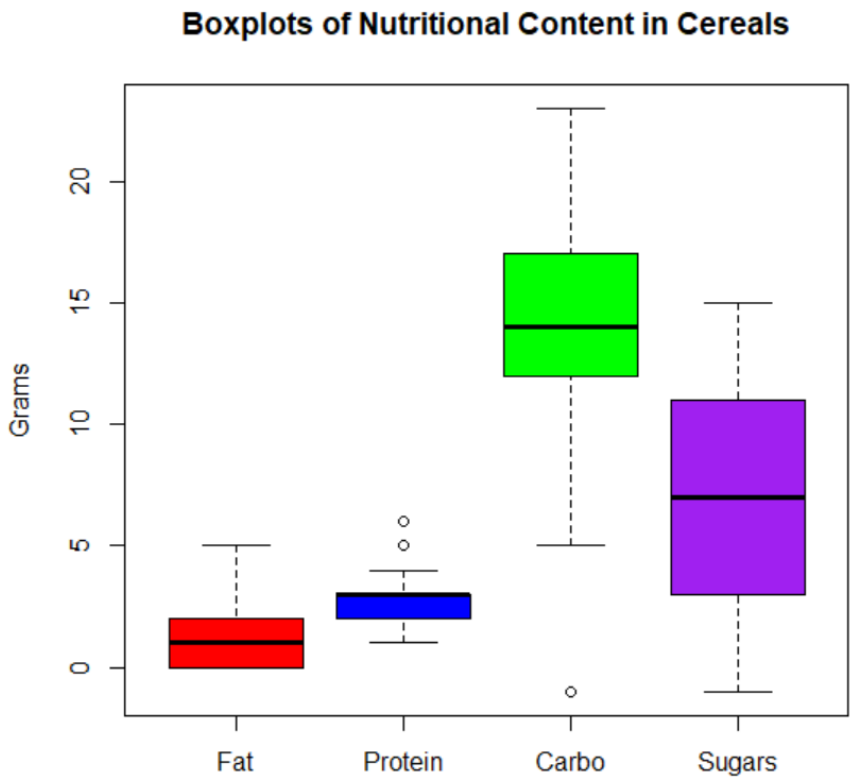
In this section, a more in-depth visual analysis was conducted to draw comparisons between various variables, shedding light on nuanced relationships within the dataset.

3.2.1 Comparison of Numerical Variables

A comparison of numerical variables is presented through a scatterplot, showcasing relationships and potential correlations. The various data points reveals patterns, and specific points of interest have been circled for further examination.

This visual representation helps in identifying potential trends and associations between different numerical features. Further analysis of these relationships could uncover insights into the features of the dataset.

3.2.2 Comparison of Nutritional Content in All Cereals



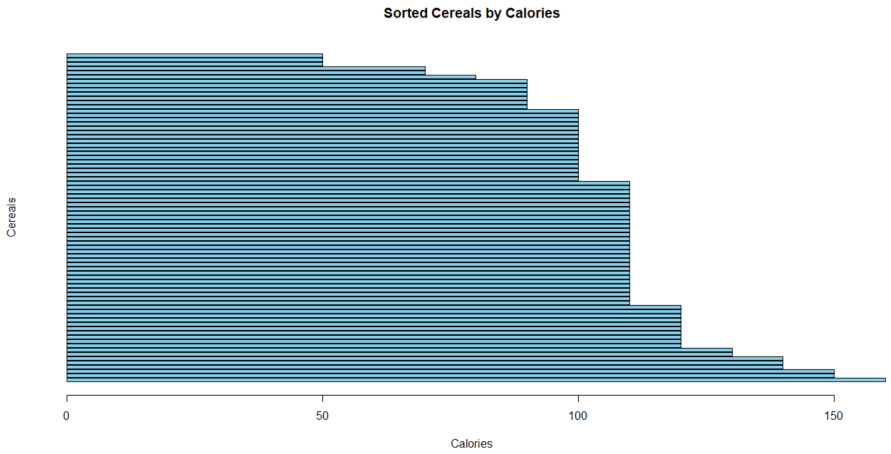
The boxplots provide a visual comparison of key nutritional components, including 'fat,' 'protein,' 'carbo,' and 'sugars' across all cereals

Observations:

- The boxplots highlights that both fat and protein content generally showcase lower variability compared to sugars and complex carbohydrates.
- There is a noticeable trend of a higher range in complex carbohydrates than in sugars, indicating a broader range of carbohydrate content in the cereals.

3.2.3 Barplot Sorted By Calories

The barplot below presents a sorted view of cereals based on their calories per serving, with the highest calorie counts at the bottom and the lowest at the top.



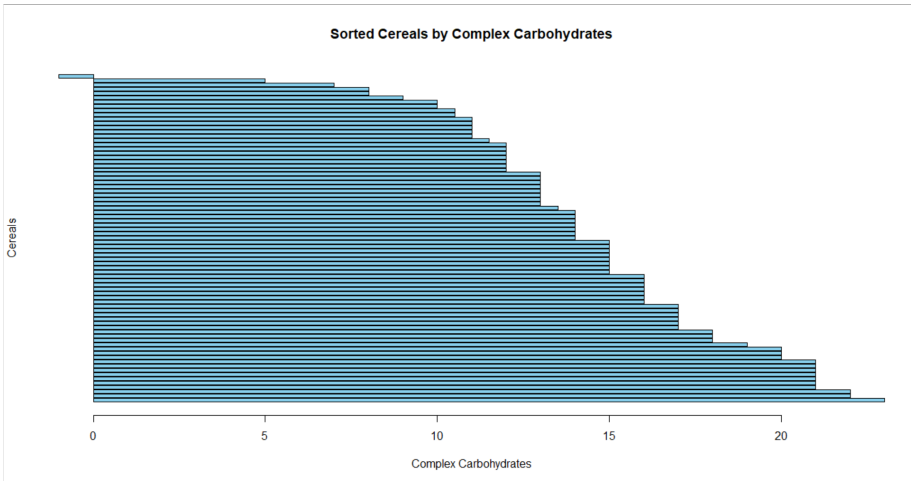
(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Calorie Cereals: Mueslix Crispy Blend, Muesli Raisins; Dates; & Almonds, Muesli Raisins; Peaches; & Pecans, Just Right Fruit & Nut, and Nutri-Grain Almond-Raisin emerge as cereals with the highest calorie content per serving.
- Lowest Calorie Cereals: Notable cereals with the lowest calorie content include 100% Bran, All-Bran, All-Bran with Extra Fiber, Puffed Rice, and Puffed Wheat.

3.2.4 Barplot Sorted By Complex Carbohydrates

The barplot below presents cereals sorted by their total complex carbohydrates, showcasing cereals with the highest complex carbohydrate counts at the bottom and the lowest at the top.



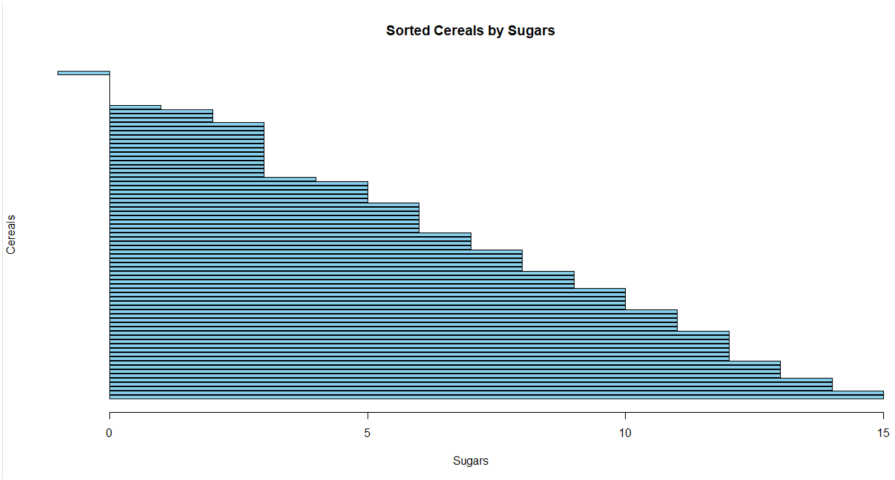
(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Complex Carbohydrate Cereals: Rice Chex, Corn Chex, Rice Krispies, Corn Flakes, and Cream of Wheat (Quick) are prominent for having the highest complex carbohydrate content per serving.
- Lowest Complex Carbohydrate Cereals: Cereals with the lowest complex carbohydrate content include 100% Natural Bran, All-Bran with Extra Fiber, All-Bran, 100% Bran, Quaker Oatmeal.

3.2.5 Barplot Sorted By Sugars

The barplot below displays cereals sorted by their total sugar content, showcasing cereals with the highest sugar counts at the bottom and the lowest at the top.



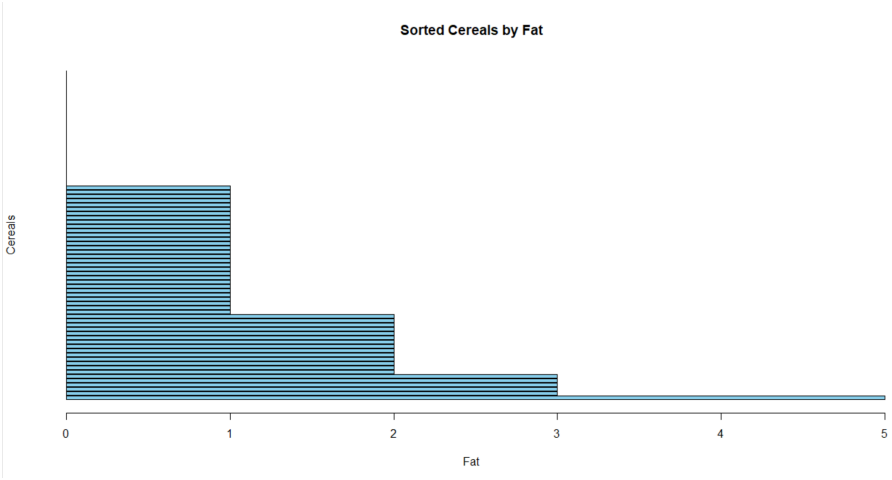
(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Sugar Cereals: Golden Crisp, Smacks, Apple Jacks, Post Nat. Raisin Bran, and Total Raisin Bran are notable for having the highest sugar content per serving.
- Lowest Sugar Cereals: Cereals with the lowest sugar content include Puffed Wheat, Shredded Wheat, Shredded Wheat 'n' Bran, Shredded Wheat spoon size, and Quaker Oatmeal.

3.2.6 Barplot Sorted By Fat

The barplot below illustrates cereals sorted by their total fat content, showcasing cereals with the highest fat counts at the bottom and the lowest at the top.



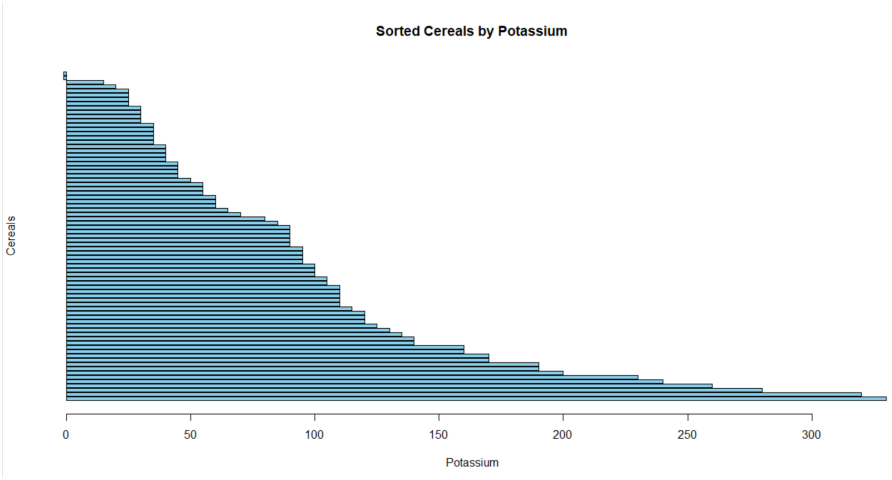
(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Fat Cereals: 100% Natural Bran, Cinnamon Toast Crunch, Cracklin' Oat Bran, Great Grains Pecan, and Muesli Raisins; Dates; & Almonds stand out for having the highest fat content per serving.
- Lowest Fat Cereals: Cereals with the lowest fat content include Shredded Wheat, Shredded Wheat 'n'Bran, Shredded Wheat spoon size, Special K, and Strawberry Fruit Wheats.

3.2.7 Barplot Sorted By Potassium

The barplot below presents cereals organized by their total potassium content, showcasing cereals with the highest potassium counts at the bottom and the lowest at the top.



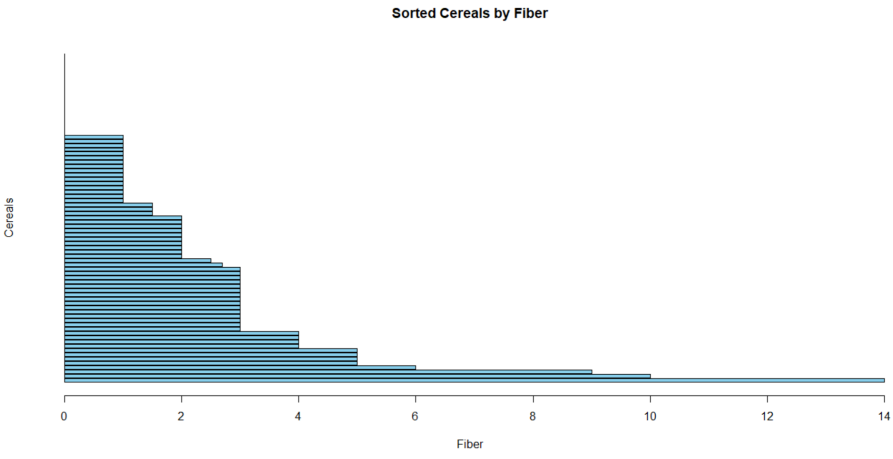
(highest is the bottom bar, to the top which is the lowest)

Observations

- Highest Potassium Cereals: All-Bran with Extra Fiber, All-Bran, 100% Bran, Post Nat. Raisin Bran, and Raisin Bran stand out for having the highest potassium content per serving.
- Lowest Potassium Cereals: Cereals with the lowest potassium content include Trix, Corn Pops, Puffed Rice, Almond Delight, Cream of Wheat (Quick).

3.2.8 Barplot Sorted By Fiber

The barplot below presents cereals organized by their total fiber content, showcasing cereals with the highest fiber counts at the bottom and the lowest at the top.



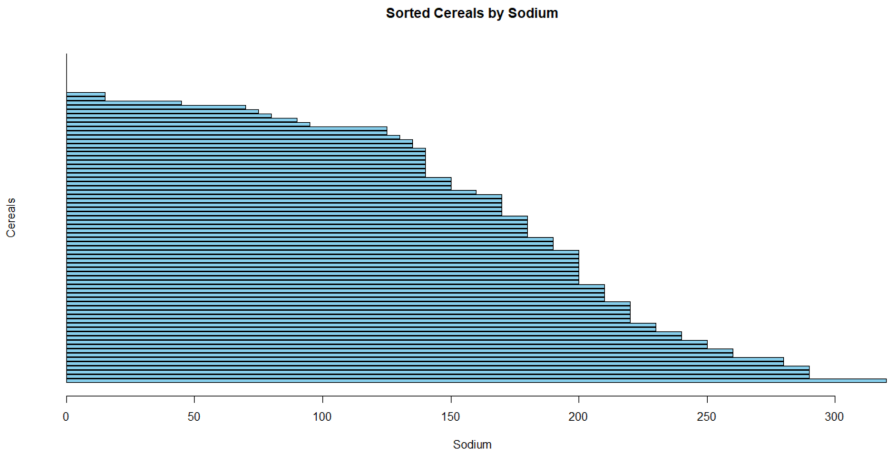
(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Fiber Cereals: All-Bran with Extra Fiber, 100% Bran, All-Bran, Post Nat. Raisin Bran, and Bran Flakes stand out for having the highest fiber content per serving.
- Lowest Fiber Cereals: Cereals with the lowest fiber content include Rice Chex, Rice Krispies, Total Corn Flakes, Triples, and Trix.

3.2.9 Barplot Sorted By Sodium

The barplot below displays cereals arranged by their total sodium content, showcasing those with the lowest sodium counts at the top and the highest at the bottom.

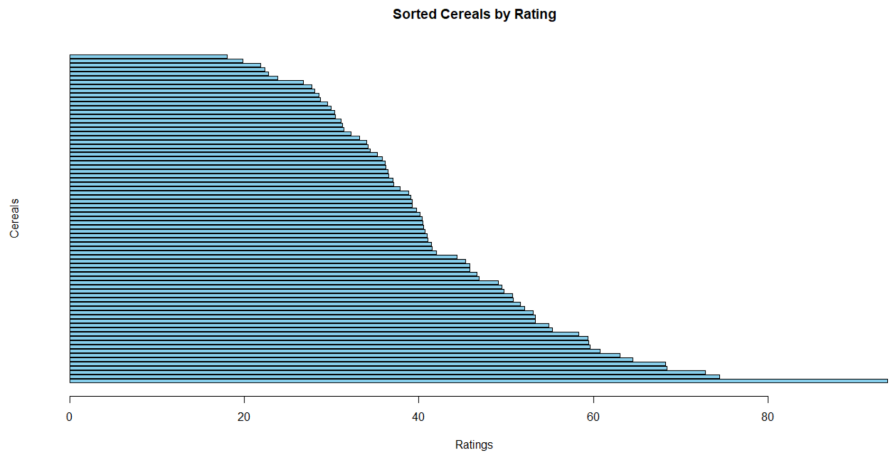


(highest is the bottom bar, to the top which is the lowest)

Observations:

- Highest Sodium Cereals: Cereals such as Product 19, Cheerios, Corn Flakes, Rice Krispies, and Corn Chex showcase the highest sodium content per serving.
- Lowest Sodium Cereals: Quaker Oatmeal, Raisin Squares, Shredded Wheat, Shredded Wheat 'n'Bran, and Shredded Wheat spoon size are among the cereals with the lowest sodium content.

3.2.10 Barplot Sorted By Ratings



(highest is the bottom bar, to the top which is the lowest)

This barplot presents cereals sorted from the highest to lowest ratings, offering a visual representation of the cereals' consumer report scores.

Observations:

- Top-Rated Cereals: All-Bran with Extra Fiber, Shredded Wheat 'n' Bran, Shredded Wheat spoon size, 100% Bran, and Shredded Wheat emerge as the top-rated cereals based on consumer reports.
- Lowest-Rated Cereals: Cocoa Puffs, Count Chocula, Honey Graham Ohs, Cinnamon Toast Crunch, and Cap'n'Crunch are among the cereals with the lowest consumer ratings.

3.2.11 Comparison of Top 30 Rated Cereals and Bottom 30 Rated Cereals

This section provides a detailed nutritional analysis comparing the top 30 and bottom 30 rated cereals, offering insights into potential differences in their nutritional content.

The Lowest 30 Rated Cereals Nutritional Content

	calories	protein	fat	sodium	fiber	complex carbs	sugars	potassium
Minimum	100	1	0	15	0	8	6	-1
1st Quartile	110	1	1	140	0	11.62	9	35
Median	110	2	1	180	1	13	11	50
Mean	117.3	1.933	1.333	165.5	1.05	13.32	10.93	68.80
3rd Quartile	120	2.750	2	200	1.50	15	12.75	93.75
Maximum	160	4	5	280	4	20	15	230

The Top 30 Rated Cereals Nutritional Content

	calories	protein	fat	sodium	fiber	complex carbs	sugars	potassium
Minimum	50	1	0	0	0	-1	-1	-1
1st Quartile	90	2	0	0	1.25	13	0.25	86.25
Median	100	3	0	140	3	15.5	3	100
Mean	92	3.133	0.6	124.2	3.19	14.6	3.033	112.63
3rd Quartile	100	4	1	200	3	17	5	113.75
Maximum	120	6	3	290	14	21	7	330

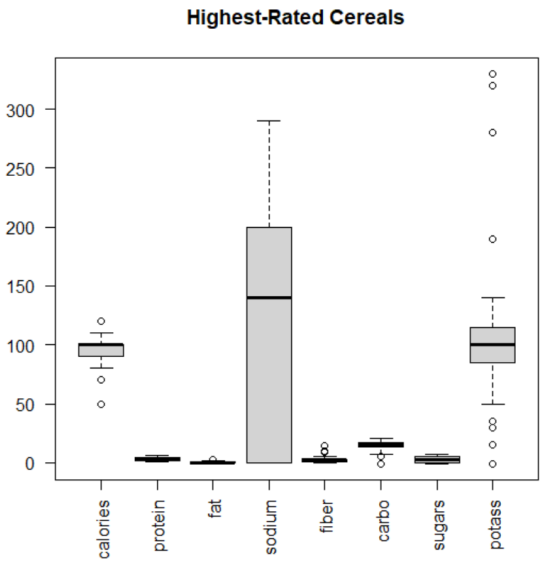
Top 30 Rated Cereals:

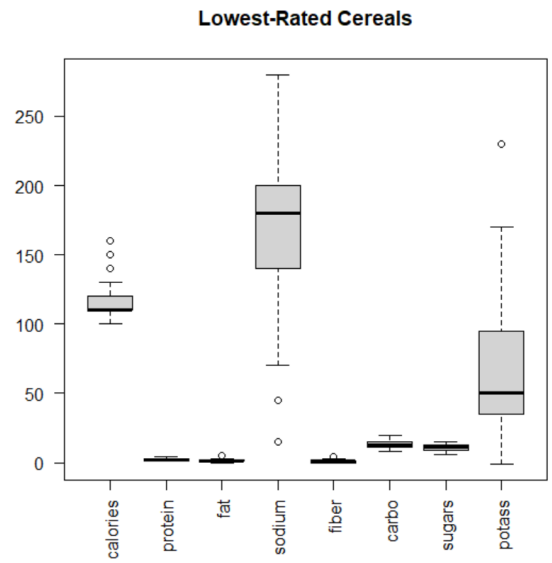
- Calories: Range from 50 to 120, with a mean of 92.
- Protein: Range from 1 to 6 grams per serving, with a mean of 3.133.
- Fat: Ranges from 0 to 3 grams, with a mean of 0.6.
- Sodium: Varies between 0 and 290 milligrams, averaging at 124.2.
- Fiber: Ranges from 0 to 14 grams, with an average of 3.19.
- Complex Carbs: Range from -1 to 21 grams, averaging at 14.6.
- Sugars: Ranges from -1 to 7 grams, with a mean of 3.033.
- Potassium: Varies between -1 and 330 milligrams, averaging at 112.63.

- -1 in these answers represents a value extremely close to 0 as in previous answers

Bottom 30 Rated Cereals:

- Calories: Range from 100 to 160, with a mean of 117.3.
- Protein: Ranges from 1 to 4 grams per serving, with a mean of 1.933.
- Fat: Varies between 0 and 5 grams, with a mean of 1.333.
- Sodium: Ranges from 15 to 280 milligrams, averaging at 165.5.
- Fiber: Varies between 0 and 4 grams, averaging at 1.05.
- Complex Carbs: Range from 8 to 20 grams, with an average of 13.32.
- Sugars: Varies between 6 and 15 grams, with a mean of 10.93.
- Potassium: Ranges from -1 to 230 milligrams, averaging at 68.80.
- -1 in these answers represents a value extremely close to 0 as in previous answers.





Observations:

- The top 30 rated cereals generally exhibit lower values in calories, fat, and sodium compared to the bottom 30 rated cereals.
- Fiber content appears to be higher in the top 30 rated cereals.
- There's a considerable range in nutritional content within both groups, and specific nutrients like protein and potassium show less distinct patterns.

3.2.12 Mean Nutritional Content By Manufacturer

This table provides an overview of the mean nutritional content across different manufacturers. The analysis offers insights into potential variations in nutritional components based on the cereal's manufacturer.

mfr	calories	protein	fat	sodium	fiber	complex carbs	sugars	potassium	ratings
A	100.00000	4.000000	1.0000000	0.0000	0.000000	16.00000	3.000000	95.00000	54.85092
G	111.36364	2.318182	1.3636364	200.4545	1.272727	14.72727	7.954545	85.22727	34.48585
K	108.69565	2.652174	0.6086957	174.7826	2.739130	15.13043	7.565217	103.04348	44.03846
N	86.66667	2.833333	0.1666667	37.5000	4.000000	16.00000	1.833333	120.66667	67.96857
P	108.88889	2.444444	0.8888889	146.1111	2.777778	13.22222	8.777778	113.88889	41.70574
Q	95.00000	2.625000	1.7500000	92.5000	1.337500	10.00000	5.250000	74.37500	42.91599
R	115.00000	2.500000	1.2500000	198.1250	1.875000	17.62500	6.125000	89.25000	41.54300

Observations:

- American Home Food Products (A): Appears to have the highest protein content, but this is based on only one cereal.
- Nabisco (N): Shows lower calorie and sodium content along with high potassium and receives the highest mean ratings.
- General Mills (G): Has the highest sodium count and the lowest mean ratings among the manufacturers.

3.2.13 Mean Nutritional Content By Shelf

This table provides an overview of the mean nutritional content based on the display shelf number. Analyzing patterns across different shelves may reveal interesting insights into how the placement of cereals on shelves correlates with their nutritional attributes.

shelf number	calories	protein	fat	sodium	fiber	complex carbs	sugars	potassium	ratings
1	102.5000	2.650000	0.60	176.2500	1.6850000	15.80000	4.800000	75.50000	46.14544
2	109.5238	1.904762	1.00	145.7143	0.9047619	13.61905	9.619048	57.80952	34.97283
3	107.7778	2.861111	1.25	158.6111	3.1388889	14.50000	6.527778	129.83333	45.22003

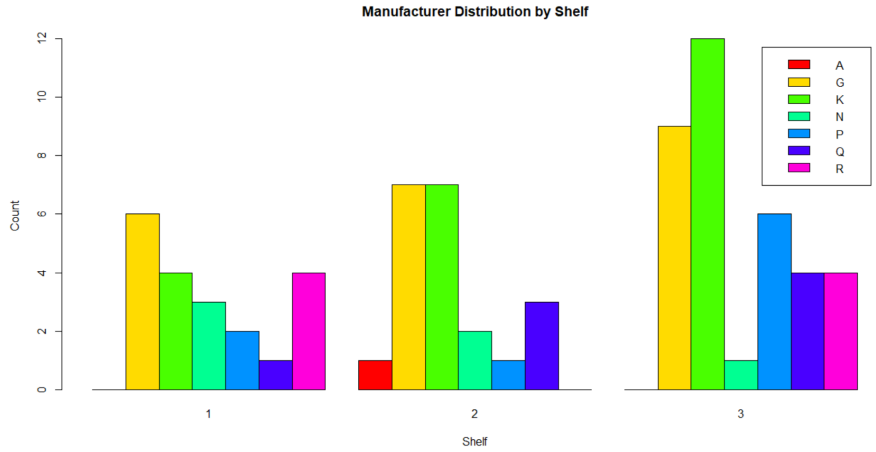
Observations:

- Shelf 2: Cereals on this shelf showcase the highest mean sugar content, yet they also have a low mean rating, suggesting a possible negative correlation between sugar content and consumer rating.
- Uniformity in Calories, Protein, Fat, and Complex Carbs: Across all shelves, these nutritional components appear to be relatively uniform, indicating consistent nutritional composition among cereals on different shelves.
- Potassium Variation: Shelf 3 stands out with a significantly higher mean potassium content compared to the other shelves, showcasing potential variability in the nutritional components of cereals placed on different shelves.

3.2.14 Distribution of Cereals By Manufacturer

This table and accompanying graph illustrate the distribution of cereals across different shelves for each manufacturer. Analyzing this distribution provides insights into how manufacturers choose to place their products on display shelves.

	1	2	3
A	0	1	0
G	6	7	9
K	4	7	12
N	3	2	1
P	2	1	6
Q	1	3	4
R	4	0	4



Observations:

- Manufacturer Placement Variation: Manufacturers exhibit variability in their choice of shelves. For example, Kelloggs and Post cereals are predominantly placed on shelf 3, while General Mills (G) cereals are spread across all shelves.
- Shelf Exclusivity: Certain manufacturers, such as American Home Food Products (A) and Ralston Purina (R), exclusively place their cereals on specific shelves (shelf 2 for A and shelves 1 and 3 for R).
- Shelf 3 Dominance: Shelf 3 appears to be a preferred location for Kelloggs (K) and Post (P) cereals, as they have a higher concentration on this shelf compared to others.

4 Research Questions

4.1 Question 1

Do cereals from different shelves have significantly different mean sodium content?

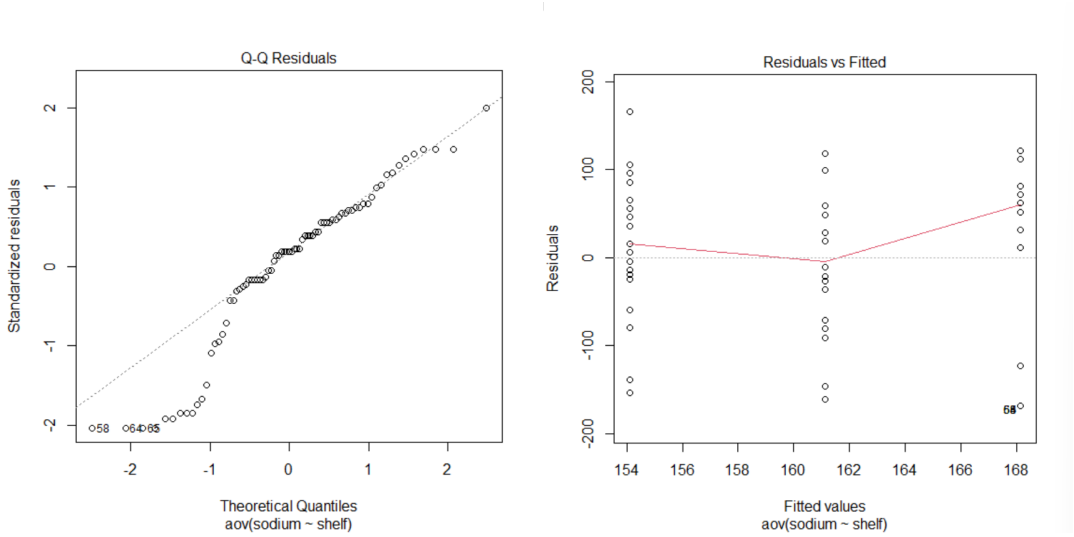
This question aims to uncover if cereals placed on different shelves have significantly different mean sodium content. The investigation seeks to understand the correlation between shelf placement and sodium levels, showcasing whether sodium content influences cereal placement.

4.1.1 Method Used

Method Used: ANOVA

ANOVA is the chosen method as it suits scenarios with three or more independent groups. It helps determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. Given the aim to identify statistically significant differences in mean sodium content among various shelves, ANOVA proves optimal for this analysis.

In order to check the validity of these assumptions, I first checked homogeneity of variances and the normality of the ANOVA:



The residuals closely align with the straight line with a skew in the plot, suggesting normality. The spread also seems to be roughly similar across groups suggesting homoscedasticity.

It seems that the residuals follow the straight line with a slight skew, as seen with the deviation of points. The spread also seems to be roughly similar across groups.

```
> summary(anova_model)
            Df Sum Sq Mean Sq F value Pr(>F)
shelf        1   2596    2596   0.366  0.547
Residuals   75 531521    7087
```

4.1.2 Statistical Results:

The ANOVA results show a non-significant F-value (0.366) with a p-value of 0.547 for the effect of shelf on sodium content.

4.1.3 Interpretation:

- The lack of statistical significance ($p > 0.05$) suggests that there is no substantial evidence to conclude that cereals placed on different shelves have significantly different mean sodium content.
- In practical terms, it implies that the placement of cereals on a particular shelf does not influence or contribute to variations in sodium content.
- The F-value being low indicates that the variability in sodium content among different shelves is not significant compared to the variability within each shelf.

4.1.4 Contextual Insight:

- Sodium content might not be a critical factor considered in the shelf placement decision for cereals.
- If sodium content were a significant consideration, we would expect to see more variability in sodium content among different shelves, which is not evident from the analysis.

4.2 Question 2

Is there a linear relationship between the amount of potassium in cereals and their overall fiber?

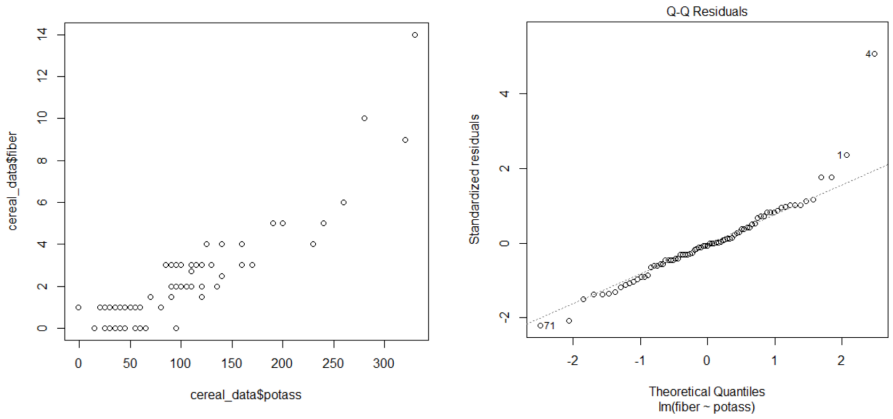
This question delves into the potential linear relationship between potassium and overall fiber content in cereals. The exploration seeks to understand if there's a straightforward connection between these nutritional aspects.

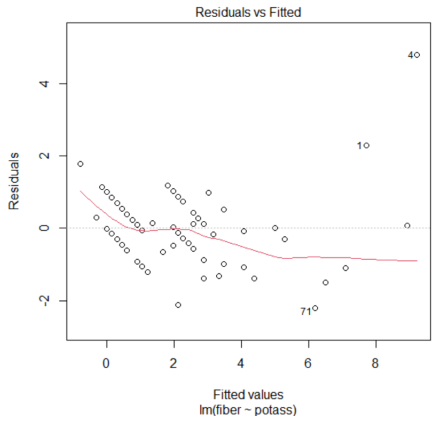
4.2.1 Method Used

Method: Simple Linear Regression

Simple Linear Regression is used here, as it is suitable for looking at linear relationships between two continuous variables. In this case, it helps determine the strength and direction of the relationship between potassium and fiber content in cereals. The goal of simple linear regression is to find the best-fitting straight line (linear equation) that describes the relationship between these two variables.

In order to check the validity of these assumptions, I first checked homogeneity of variances, normality, and linearity of the regression model:





There seems to be a linear relationship within the scatter plot. The residuals closely align with the straight line in the plot suggesting normality. The spread also seems to be roughly similar across groups suggesting homoscedasticity.

```
> summary(lm_model)

Call:
lm(formula = fiber ~ potass, data = cereal_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1968 -0.5724 -0.0623  0.5215  4.7829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.749884   0.197614  -3.795 0.000298 ***
potass       0.030203   0.001656  18.243  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.029 on 75 degrees of freedom
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8136
F-statistic: 332.8 on 1 and 75 DF,  p-value: < 2.2e-16
```

4.2.2 Statistical Results:

- The linear regression model reveals a strong positive relationship between fiber content and potassium content.
- The coefficient estimate for potassium (0.030203) indicates that, on average, for each unit increase in potassium, the fiber content increases by 0.030203 units.
- The p-value for the coefficient is highly significant ($<$ 2e-16), suggesting that the observed relationship is not due to random chance.
- The R-squared value is 0.8161, indicating that approximately 81.61% of the variability in fiber content can be explained by the variability in potassium content.

4.2.3 Residual Analysis:

- Residuals (the differences between observed and predicted values) have a mean close to zero, suggesting that the model is not systematically overestimating or underestimating fiber content.
- The residuals' standard error is 1.029, indicating the average distance between observed and predicted values.

4.2.4 Interpretation:

- The negative intercept (-0.749884) implies that when potassium content is zero, the predicted fiber content is expected to be -0.75. However, this value has no practical interpretation in this context.
- The positive coefficient for potassium indicates a positive linear relationship, suggesting that higher potassium content is associated with higher fiber content in cereals.

4.2.5 Contextual Insight:

- This insight is valuable for cereal manufacturers, showcasing that they can potentially predict fiber content by adjusting the potassium levels within their products.
- The high R-squared value suggests that potassium content is a significant factor in explaining the variability in fiber content.

4.3 Question 3

Are there significant differences in calories among cereals from different shelves?

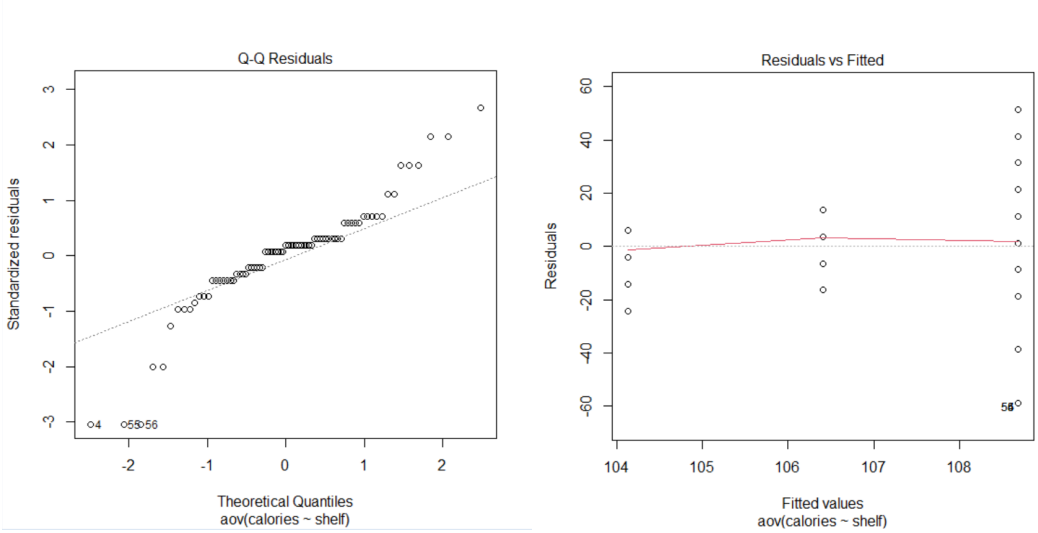
This question investigates whether the placement of cereals on different shelves is associated with significant differences in caloric content. By analyzing caloric content based on shelf placement, the aim is to see if caloric content influences the shelf location of a cereal.

4.3.1 Method Used

Method: ANOVA

ANOVA is chosen due to the presence of three or more independent groups. It helps determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. The goal is to identify statistically significant differences in mean caloric content among different shelves, making ANOVA the appropriate statistical test.

In order to check the validity of these assumptions, I first checked homogeneity of variances and the normality of the ANOVA:



The residuals closely align with the straight line with some skew on both ends in the plot, suggesting normality. The spread also seems to be roughly similar across groups suggesting homoscedasticity.

```
> summary(anova_model_calories)
            Df Sum Sq Mean Sq F value Pr(>F)
shelf        1    273   272.8   0.716    0.4
Residuals   75  28579   381.1
```

4.3.2 Statistical Results:

- The ANOVA results indicate an F-value of 0.716 with a corresponding p-value of 0.4.
- The p-value is above the alpha level of 0.05, suggesting that there is no significant difference in mean calorie content among cereals from different shelves.

4.3.3 Interpretation:

- The F-value tests the null hypothesis that all group means (caloric content for each shelf) are equal.
- A small F-value and a large p-value indicate that there is insufficient evidence to reject the null hypothesis.
- In this context, the non-significant p-value suggests that the differences in mean calorie content among shelves, as observed in the data, could be due to random chance.

4.3.4 Contextual Insight:

- The shelf placement does not appear to be a significant factor influencing the caloric content of cereals in this dataset.
- Manufacturers or consumers may not need to consider shelf placement as a relevant factor when evaluating or selecting cereals based on caloric content.

4.4 Question 4

Is the mean sugars of the lowest 30 cereals significantly different from the mean sugars of the top 30 cereals?

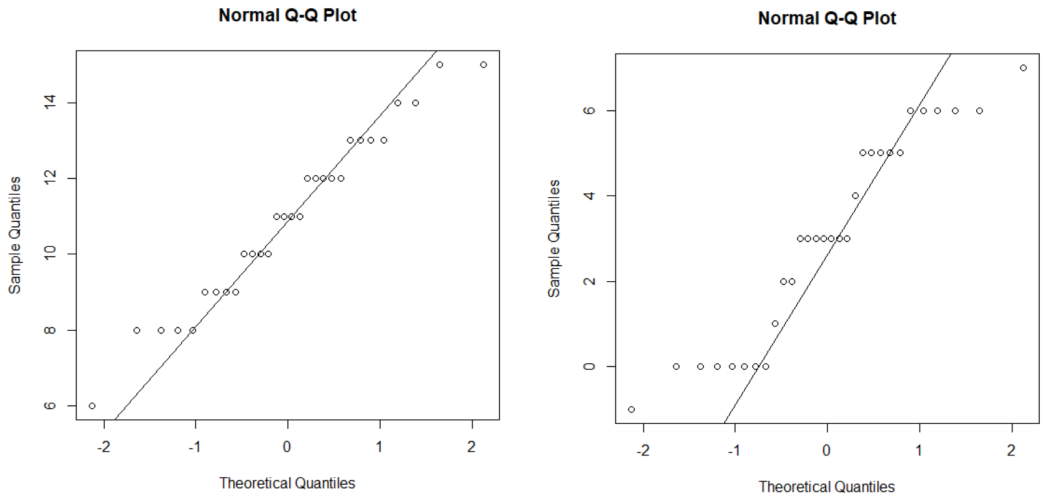
This question aims to determine if the mean sugar content differs significantly between the lowest and highest rated cereals. I want to be able to determine consumer preferences regarding sugar content in cereals.

4.4.1 Method Used

Method: T-Test

A T-Test is selected since the interest lies in comparing the means of two independent groups, specifically the sugar content in the top 30 versus the bottom 30 rated cereals. T-Test is used when comparing the means of two independent groups to determine if they are significantly different.

In order to check the validity of these assumptions, I first checked the normality of the t-test:



(On the left is the last 30 qqnorm plot, on the right is the top 30 qqnorm plot)

The residuals closely align with the straight line in each plot with a slight skew on both ends, suggesting normality overall.

```
> t_test_sugars

        Welch Two Sample t-test

data:  last30$sugars and top30$sugars
t = 13.05, df = 57.87, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.688159 9.111841
sample estimates:
mean of x mean of y
10.933333  3.033333
```

4.4.2 Statistical Results:

- A Welch Two Sample t-test was conducted to compare the mean sugar content between the lowest 30 cereals and the top 30 cereals based on their ratings.
- The t-test resulted in a t-value of 13.05, degrees of freedom (df) of 57.87, and a very small p-value ($<$ 2.2e-16).
- The 95 percent confidence interval for the difference in means is [6.688159, 9.111841].

4.4.3 Interpretation:

- The small p-value (< 0.05) indicates that there is a significant difference in mean sugar content between the lowest and top 30 cereals.
- The positive t-value suggests that the mean sugar content of the lowest 30 cereals is significantly higher than that of the top 30 cereals.

4.4.4 Contextual Insight:

- The significantly higher mean sugar content in the lowest 30 cereals suggests that cereals with lower ratings tend to have higher sugar content.
- Consumers who prefer cereals with lower sugar content might be inclined to choose cereals with higher ratings.

4.5 Question 5

Is the mean rating of the shelf 1 cereals significantly different from the mean ratings of the shelf 3 cereals?

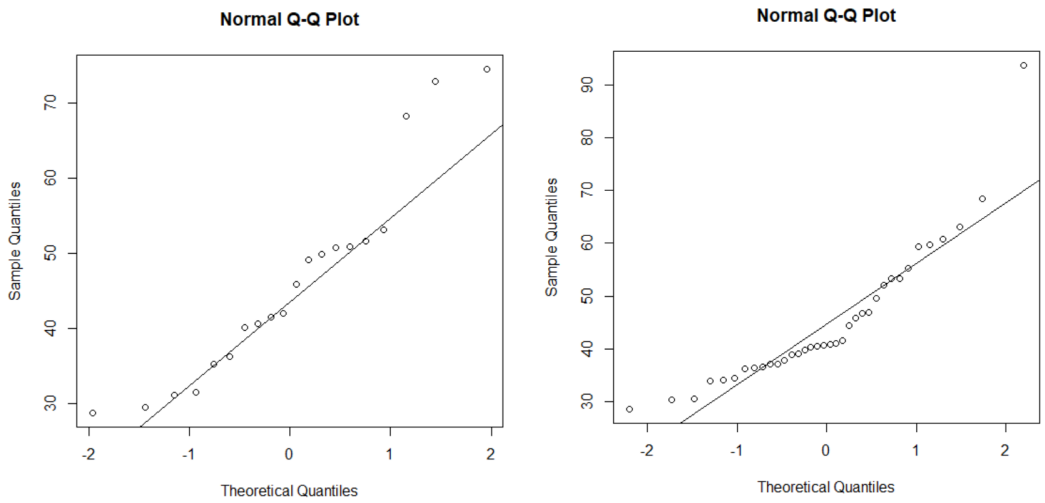
This question explores whether cereals on different shelves receive significantly different mean consumer ratings. Through this, I want to uncover potential connections between shelf placement and consumer preferences.

4.5.1 Method Used

T-Test

Similar to Question 4, a T-Test is appropriate as it involves comparing the means of two independent groups, in this case, the mean ratings of cereals on shelf 1 versus shelf 3. T-Test is used when comparing the means of two independent groups to determine if they are significantly different.

In order to check the validity of these assumptions, I first checked the normality of the t-test:



(On the left is the shelf 1 qqnorm plot, on the right is the shelf 3 qqnorm plot)

The residuals closely align with the straight line in the plot suggesting normality.

```
> t_test_ratings

        Welch Two Sample t-test

data:  shelf1 and shelf3
t = 0.24851, df = 37.917, p-value = 0.8051
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.613582  8.464396
sample estimates:
mean of x mean of y
 46.14544  45.22003
```

4.5.2 Statistical Results:

- A Welch Two Sample t-test was conducted to compare the mean ratings between cereals from shelf 1 and shelf 3.
- The t-test resulted in a t-value of 0.24851, degrees of freedom (df) of 37.917, and a p-value of 0.8051.
- The 95 percent confidence interval for the difference in means is [-6.613582, 8.464396].

4.5.3 Interpretation:

- The p-value of 0.8051 is greater than the common significance level of 0.05. Therefore, we do not have enough evidence to reject the null hypothesis.
- The confidence interval includes zero, further supporting the lack of a significant difference in mean ratings between cereals from shelf 1 and shelf 3.

4.5.4 Contextual Insight:

- The lack of a significant difference in mean ratings suggests that, on average, cereals from shelf 1 and shelf 3 receive similar ratings from consumers.

4.6 Question 6

Is there a linear relationship between the amount of sugars in cereals and their calories content?

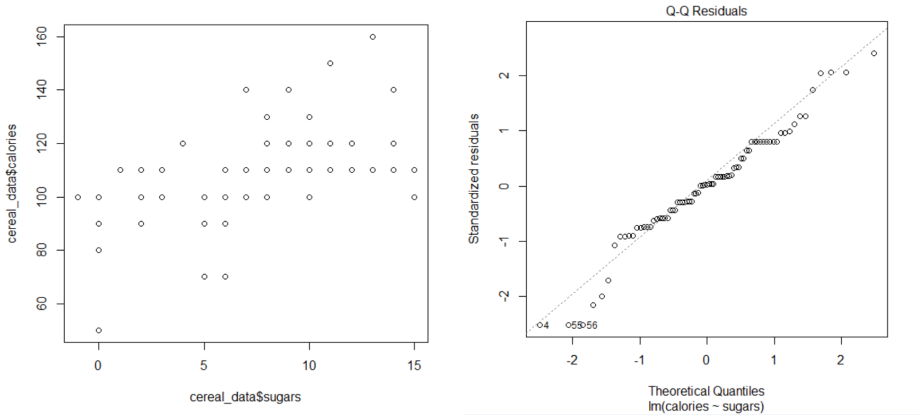
This question looks at the potential linear relationship between sugar and calorie content in cereals. This relationship has many benefits such as making dietary recommendations and assisting manufacturers in creating cereals with certain nutritional content.

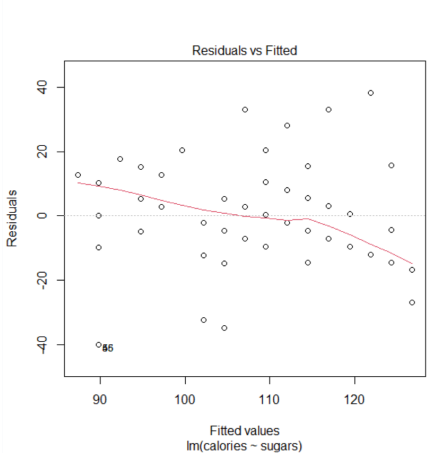
4.6.1 Method Used

Method: Simple Linear Regression

Using Simple Linear Regression is fitting as it helps assess the linear relationship between two continuous variables. It showcases the strength and direction of the correlation between sugar and calorie content in cereals. The goal of simple linear regression is to find the best-fitting straight line (linear equation) that describes the relationship between these two variables.

In order to check the validity of these assumptions, I first checked homogeneity of variances, normality, and linearity of the regression model:





There seems to be a linear relationship within the scatter plot. The residuals closely align with the straight line in the plot suggesting normality. The spread also seems to be roughly similar across groups suggesting homoscedasticity.

```
> summary(lm_model_sugars_calories)

Call:
lm(formula = calories ~ sugars, data = cereal_data)

Residuals:
   Min     1Q Median     3Q    Max
-39.82  -9.40   0.46  12.64  38.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.8201     3.4366  26.137  < 2e-16 ***
sugars        2.4650     0.4185   5.889 1.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.22 on 75 degrees of freedom
Multiple R-squared:  0.3162,    Adjusted R-squared:  0.3071
F-statistic: 34.69 on 1 and 75 DF,  p-value: 1.025e-07
```

4.6.2 Statistical Results:

- A simple linear regression model was used to assess the relationship between calories and sugars in cereals.
- The regression equation is given by: Calories $= 89.8201 + 2.4650 * Sugars$.
- The coefficient for sugars is statistically significant with a p-value of $1.02 * 10^{-7}$.
- The multiple R-squared is 0.3162, indicating that approximately 31.62% of the variation in calories can be explained by the variation in sugars.

4.6.3 Interpretation:

- The coefficient for sugars (2.4650) suggests that, on average, for each additional unit of sugars, there is an increase of 2.4650 units in calories.
- The intercept (89.8201) represents the estimated calories when the sugar content is zero. However, in the context of cereals, this might not have a practical interpretation.
- The p-value of $1.02 * 10^{-7}$ is less than 0.05, indicating that the relationship between sugars and calories is statistically significant.

4.6.4 Contextual Insight:

- The positive coefficient implies a positive linear relationship between sugars and calories. Cereals with higher sugar content tend to have higher caloric content, based on the model.

5 Results

In this analysis of the Cereal Data, various aspects of the dataset were explored to extract valuable insights and understand relationships between different variables.

5.1 Initial Exploration

The initial exploration involved a thorough examination of the dataset to familiarize ourselves with its structure, identifying key variables, and understanding the distribution of data points. Descriptive statistics and visualizations were employed to provide an overview, enabling us to see patterns and trends within the dataset.

5.2 Additional Exploration

Building upon the initial exploration, further analyses were conducted to delve deeper into specific relationships and trends. Correlation analyses, scatter plots, and distribution visualizations were employed to uncover associations between different nutritional components of cereals. This stage of exploration aimed to identify potential factors influencing the nutritional content of cereals and any patterns.

5.3 Research Questions

To address the research questions posed at the beginning of the analysis, statistical methods such as ANOVA, T-Tests, and regression analyses were applied. Each research question was carefully examined, and the results were interpreted to draw possible conclusions regarding the factors influencing shelf placement, nutritional content variations, and potential relationships between key variables. The detailed findings are presented in the subsequent sections, shedding light on the relationships within the Cereal Data.

6 Conclusion:

In conclusion, this examination of the Cereal Data data set has provided valuable insights into the nutritional content and distribution patterns of cereals. Through exploration and statistical analyses of this dataset, I uncovered notable trends among manufacturers and shelves, which could be valuable information for both consumers and producers. The visual representation of nutritional content, along with statistical assessments, supports informed decision-making in cereal product development and marketing. These findings enhance our understanding of cereal choices, which could help you find your own cereal product with your own preference and nutritional consideration.

7 References

Chris Crawford. 2017. 80 Cereals, 8.53. Retrieved December 1 2023 from https://www.kaggle.com/datasets/crawford/80-cereals