DS Project: The Healthiest Cereal

By Sobia Zahid & Aqsa Noreen

About the Case Study

- Data set from Kaggle, titled “80 Cereals”
- Goal: determine the healthiest cereals from this data set
- Methods: comparing nutrition facts and visualizing them using tools in RStudio

Setting Things Up

```
cereal = read.csv(file.choose(),header = T)
library(tidyverse)
library(ggplot2)

colnames(cereal) <- c("Name", "Manufacturer", "Type", "Calories", "Protein",
                      "Fat", "Sodium", "Fiber", "Carbohydrates", "Sugar",
                      "Potassium", "Vitamins", "Shelf", "Weight", "Cups", "Rating")

cereal$Type <- factor(cereal$Type)
cereal$Shelf <- factor(cereal$Shelf)
cereal$Manufacturer <- factor(cereal$Manufacturer)

cereal$Carbohydrates[cereal$Carbohydrates < 0] <- NA
cereal$Sugar[cereal$Sugar < 0] <- NA
cereal$Potassium[cereal$Potassium < 0] <- NA
```

- Data set is made up of the nutrition facts of 77 different kinds of cereals -
- There are 16 columns: Name, Manufacturer, Type (if it's a hot or cold cereal), Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, Potassium, Vitamins, Shelf, Weight, Cups, and Rating
- We made sure that everything had been converted into a factor, and then we changed the incorrect/implausible values to NA as to not skew the results.

Overview

```
summary(cereal)
```

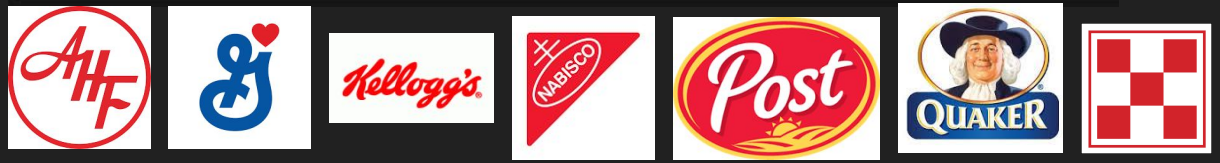
```
    Name           Manufacturer Type      Calories        Protein          Fat        
Length:77          A: 1         C:74   Min.   : 50.0   Min.   :1.000   Min.   :0.000  
Class :character   G:22         H: 3   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000  
Mode  :character   K:23                Median :110.0   Median :3.000   Median :1.000  
                   N: 6                Mean   :106.9   Mean   :2.545   Mean   :1.013  
                   P: 9                3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000  
                   Q: 8                Max.   :160.0   Max.   :6.000   Max.   :5.000  
                   R: 8                                                               
     Sodium          Fiber         Carbohydrates       Sugar         Potassium       Vitamins     
 Min.   :  0.0   Min.   : 0.000   Min.   : 5.0   Min.   : 0.000   Min.   : 15.00   Min.   :  0.00  
 1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 42.50   1st Qu.: 25.00  
 Median :180.0   Median : 2.000   Median :14.5   Median : 7.000   Median : 90.00   Median : 25.00  
 Mean   :159.7   Mean   : 2.152   Mean   :14.8   Mean   : 7.026   Mean   : 98.67   Mean   : 28.25  
 3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00  
 Max.   :320.0   Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00  
                                  NA's   :1      NA's   :1        NA's   :2                        
     Shelf       Weight          Cups            Rating      
 1:20   Min.   :0.50   Min.   :0.250   Min.   :18.04  
 2:21   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17  
 3:36   Median :1.00   Median :0.750   Median :40.40  
        Mean   :1.03   Mean   :0.821   Mean   :42.67  
        3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83  
        Max.   :1.50   Max.   :1.500   Max.   :93.70  
```

This summary of the overall data shows the important values of each column which will help us determine which cereal has higher nutritious value. We can use this information to find the cereals which have higher amounts of good nutrients, and lower amounts of unhealthy ones.

Manufacturers

```
cereal$Manufacturer_Name <- cereal$Manufacturer
cereal$Type <- gsub("H", "Hot", x = cereal$Type)
cereal$Type <- gsub("C", "Cold", x = cereal$Type)
Manufacturers <- cereal %>%
  select(Manufacturer_Name, Type) %>%
  group_by(Manufacturer_Name, Type) %>%
  summarise(Total = n()) %>%
  spread(key = Type, value = Total) %>%
  replace_na(replace = list(Manufacturer_Name = 0, Cold = 0, Hot = 0)) %>%
  mutate(Total = Cold + Hot) %>%
  arrange(desc(Total))
Manufacturers
```

Manufacturer_Name <fctr>	Cold <dbl>	Hot <dbl>	Total <dbl>
K	23	0	23
G	22	0	22
P	9	0	9
Q	7	1	8
R	8	0	8
N	5	1	6
A	0	1	1

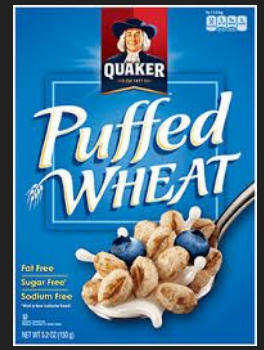


- Manufacturers in this data: American Home Food Products, General Mills, Kellogg's, Nabisco, Post, Quaker Oats, and Ralston Purina
- 74 of the cereals are meant to be eaten cold, while only 3 are the "hot" type
- Kellogg's and General Mills are the two manufacturers with the most amount of products in this data set

Calories


```
min(cereal$calories)
MinCal = subset(cereal, cereal$calories==50)
MinCal
```

	Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
4	All-Bran with Extra Fiber	K	Cold	50
55	Puffed Rice	Q	Cold	50
56	Puffed Wheat	Q	Cold	50



The cereals with the least amount of calories in this data set are All-Bran with Extra Fiber, Puffed Rice, and Puffed Wheat with only 50 per serving.

Calories

```
max(cereal$Calories)
MaxCal = subset(cereal, cereal$Calories==160)
MaxCal
```

	Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
47	Mueslix Crispy Blend	K	Cold	160

The highest calorie cereal is Mueslix Crispy Blend at a whopping 160.

Calories

```
mean=mean(cereal$calories)
mean
```

```
106.8831
```

The overall average calories per serving from this data set is 106.88.

Protein


```
max(cereal$Protein)
MaxProtein = subset(cereal, cereal$Protein==6)
MaxProtein
```

	Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>	Protein <int>
12	Cheerios	G	Cold	110	6
68	Special K	K	Cold	110	6

Cheerios and Special K have the highest protein with 6 grams.

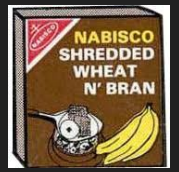
Name	Manufacturer
All-Bran with Extra Fiber	K
Apple Jacks	K
Bran Flakes	P
Corn Chex	R
Corn Flakes	K
Corn Pops	K
Cream of Wheat (Quick)	N
Crispix	K
Double Chex	R
Frosted Flakes	K
Frosted Mini-Wheats	K
Fruitful Bran	K
Golden Crisp	P
Grape-Nuts	P
Honey-comb	P
Nutri-grain Wheat	K
Product 19	K
Puffed Rice	Q
Puffed Wheat	Q
Raisin Squares	K
Rice Chex	R
Rice Krispies	K
Shredded Wheat	N
Shredded Wheat 'n'Bran	N
Shredded Wheat spoon size	N
Special K	K
Strawberry Fruit Wheats	N

Fat

```
min(cereal$Fat)
MinFat = subset(cereal, cereal$Fat==0)
MinFat
```

27 kinds of cereals have 0 grams of fat, which shows that there is an effort being made by manufacturers to reduce this from the products.

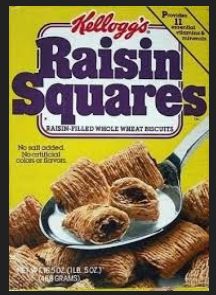
Sodium




```
min(cereal$sodium)
MinSodium = subset(cereal, cereal$sodium==0)
MinSodium
```

Name <chr>	Manufacturer <fctr>	Type <chr>
Frosted Mini-Wheats	K	Cold
Maypo	A	Hot
Puffed Rice	Q	Cold
Puffed Wheat	Q	Cold
Quaker Oatmeal	Q	Hot
Raisin Squares	K	Cold
Shredded Wheat	N	Cold
Shredded Wheat 'n'Bran	N	Cold
Shredded Wheat spoon size	N	Cold

There are 9 kinds of cereals with 0 milligrams of sodium; good for people who are trying to have a lower intake.



Fiber


```
max(cereal$Fiber)
MaxFiber = subset(cereal, cereal$Fiber==14)
MaxFiber
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
All-Bran with Extra Fiber	K	Cold	50

All-Bran with Extra Fiber has 14 grams of fiber, the highest amount in this data.

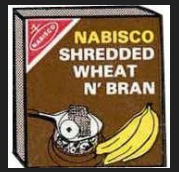
Carbohydrates


```
MinCarbs = subset(cereal, cereal$carbohydrates==5)
MinCarbs
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
100% Bran	N	Cold	70

100% Bran has the lowest amount of carbohydrates with 5 grams.

Sugar




```
MinSugar = subset(cereal, cereal$Sugar==0)
MinSugar
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>	Protein <int>	Fat <int>
All-Bran with Extra Fiber	K	Cold	50	4	0
Cream of Wheat (Quick)	N	Hot	100	3	0
Puffed Rice	Q	Cold	50	1	0
Puffed Wheat	Q	Cold	50	2	0
Shredded Wheat	N	Cold	80	2	0
Shredded Wheat 'n'Bran	N	Cold	90	3	0
Shredded Wheat spoon size	N	Cold	90	3	0

7 cereals have 0 grams of sugar, and all of them also have 0 grams of fat.

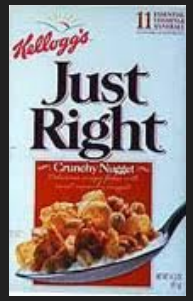
Potassium

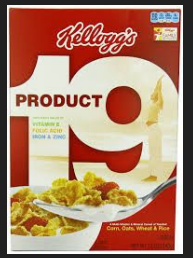
```
MaxPotassium = subset(cereal, cereal$Potassium==330)
MaxPotassium
```

Name <chr>	Manufacturer <fctr>
All-Bran with Extra Fiber	K

The cereal with the highest amount of potassium is All-Bran with Extra Fiber at 330 milligrams per serving.

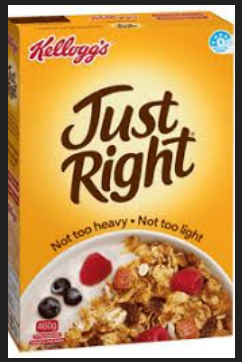
Vitamins

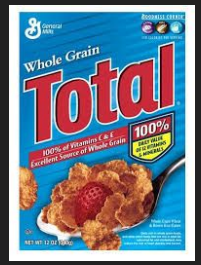


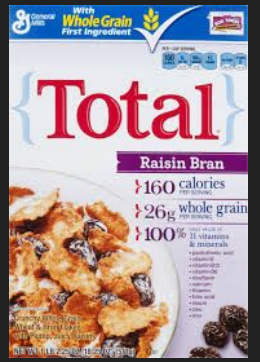



```
max(cereal$vitamins)
MaxVitamins = subset(cereal, cereal$vitamins==100)
MaxVitamins
```

Name <chr>	Manufacturer <fctr>
Just Right Crunchy Nuggets	K
Just Right Fruit & Nut	K
Product 19	K
Total Corn Flakes	G
Total Raisin Bran	G
Total Whole Grain	G







6 cereals have a measure of 100 vitamins and minerals, indicating the typical percentage of FDA recommended, with half of them coming from Kellogg's and the other half from General Mills.

Cups

```
max(cereal$Cups)
MaxCups=subset(cereal, cereal$Cups==1.5)
MaxCups
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
Kix	G	Cold	110

The cereal with the highest amount of cups per serving is Kix with 1.5 cups.

Cups

```
min(cereal$Cups)
MinCups=subset(cereal, cereal$Cups==0.25)
MinCups
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
Grape-Nuts	P	Cold	110

The cereal with the lowest amount of cups per serving is Grape-Nuts at only a quarter of a cup. This is probably because that cereal in particular is very dense.

Ratings

```
max(cereal$Rating)
MaxRating=subset(cereal, cereal$Rating>93)
MaxRating
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
All-Bran with Extra Fiber	K	Cold	50

The highest rated cereal is All-Bran with Extra Fiber, coming in at 93.70491.

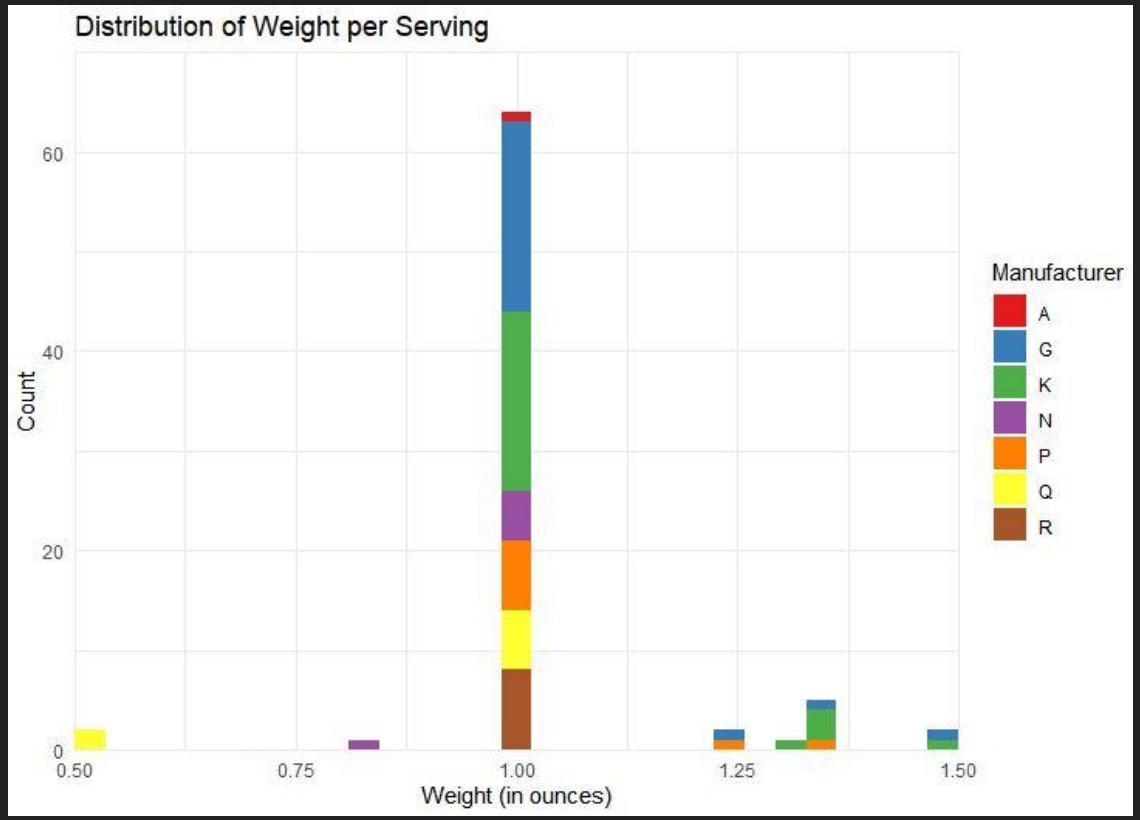
Ratings


```
min(cereal$Rating)
MinRating=subset(cereal, cereal$Rating<18.1)
MinRating
```

Name <chr>	Manufacturer <fctr>	Type <chr>	Calories <int>
Cap'n'Crunch	Q	Cold	120

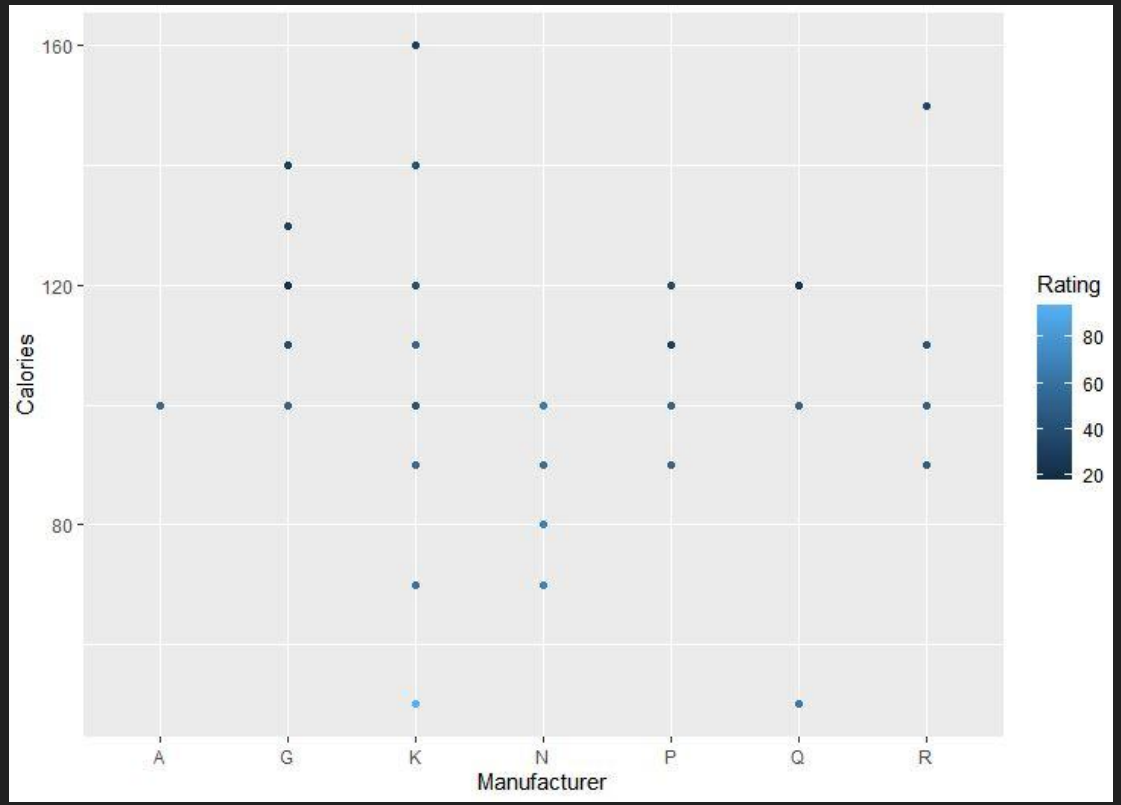
Cap'n Crunch has the lowest rating of only 18.04285, which is probably because it has no fiber, less vitamins, low potassium and protein, but lots of carbs and sugar.

```r
cereal %>%
  ggplot(aes(x = Weight, fill = Manufacturer)) +
  geom_histogram() +
  scale_fill_brewer(palette = "Set1") +
  scale_x_continuous(name = "Weight (in ounces)", expand = c(0,0)) +
  scale_y_continuous(name = "Count", expand = c(0,0), limits = c(0, 70)) +
  labs(fill = "Manufacturer", title = "Distribution of Weight per Serving") +
  theme_minimal()
```



This histogram very clearly shows that the general weight of one serving of cereal for most brands is one ounce. There are some outliers, however, at 0.5 and 1.5 ounces.

```
cal=ggplot(data=cereal)
cal+
  geom_point(mapping = aes(x=Manufacturer,y=Calories, color=Rating))
```

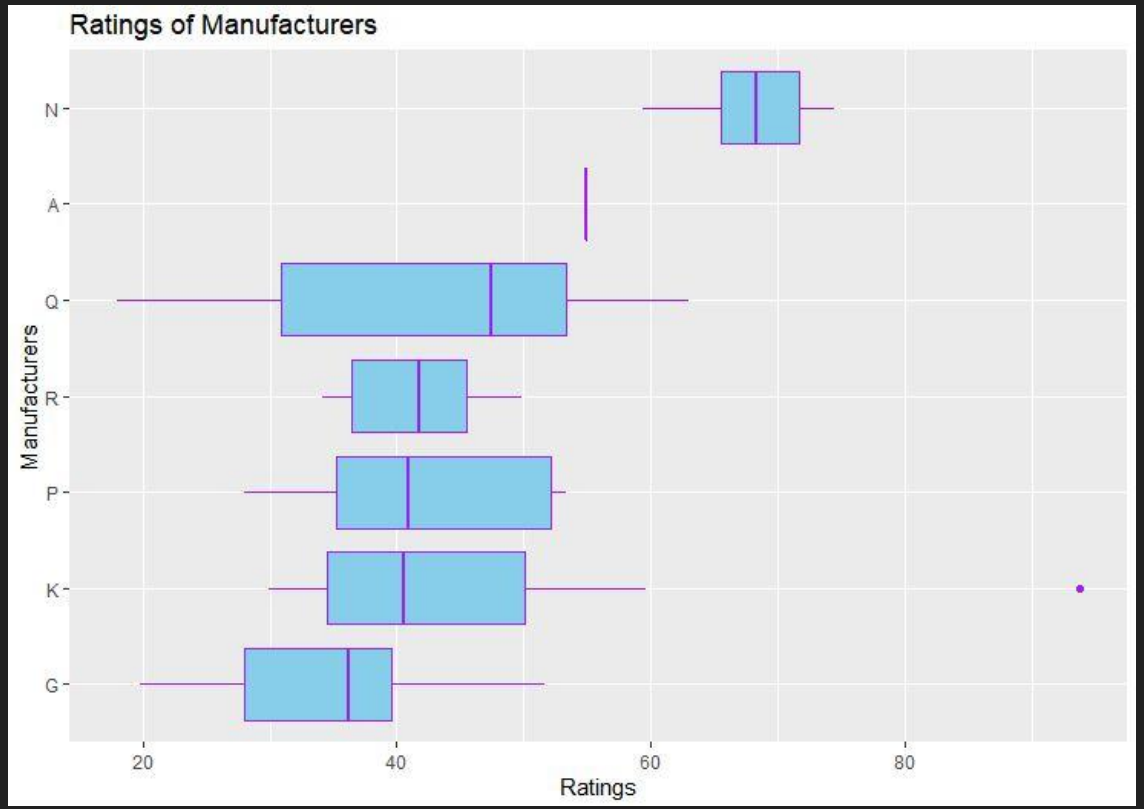


This scatter plot shows the amount of calories the products of each manufacturer contain.

Kellogg's seems to have the biggest range, with both the lowest and highest calorie cereals.

General Mills products are pretty consistent in their calorie count, as well as Nabisco and Post.

```r
ManRating=ggplot(data=cereal)+
  geom_boxplot(aes(x=reorder(Manufacturer,Rating, FUN=median), y=Rating),
               color="purple", fill="skyblue")+
  coord_flip()
ManRating+ ggtitle("Ratings of Manufacturers")+
  xlab("Manufacturers")+ylab("Ratings")
```

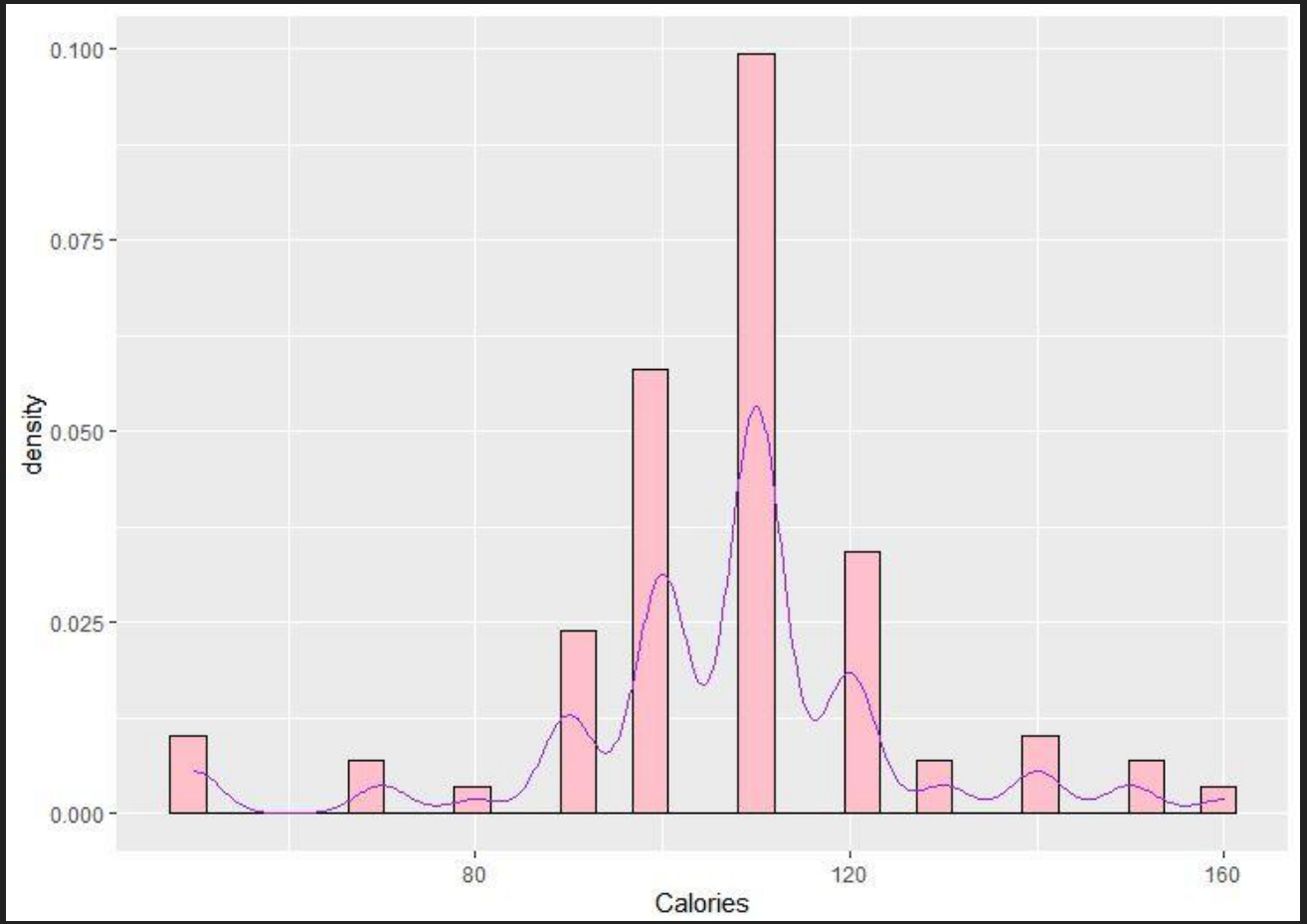


This box plot illustrates the ratings of each manufacturer out of 100. Quaker Oats has the largest range, and also the lowest.

Kellogg's has the highest, although it seems like somewhat of an outlier/unusual value.

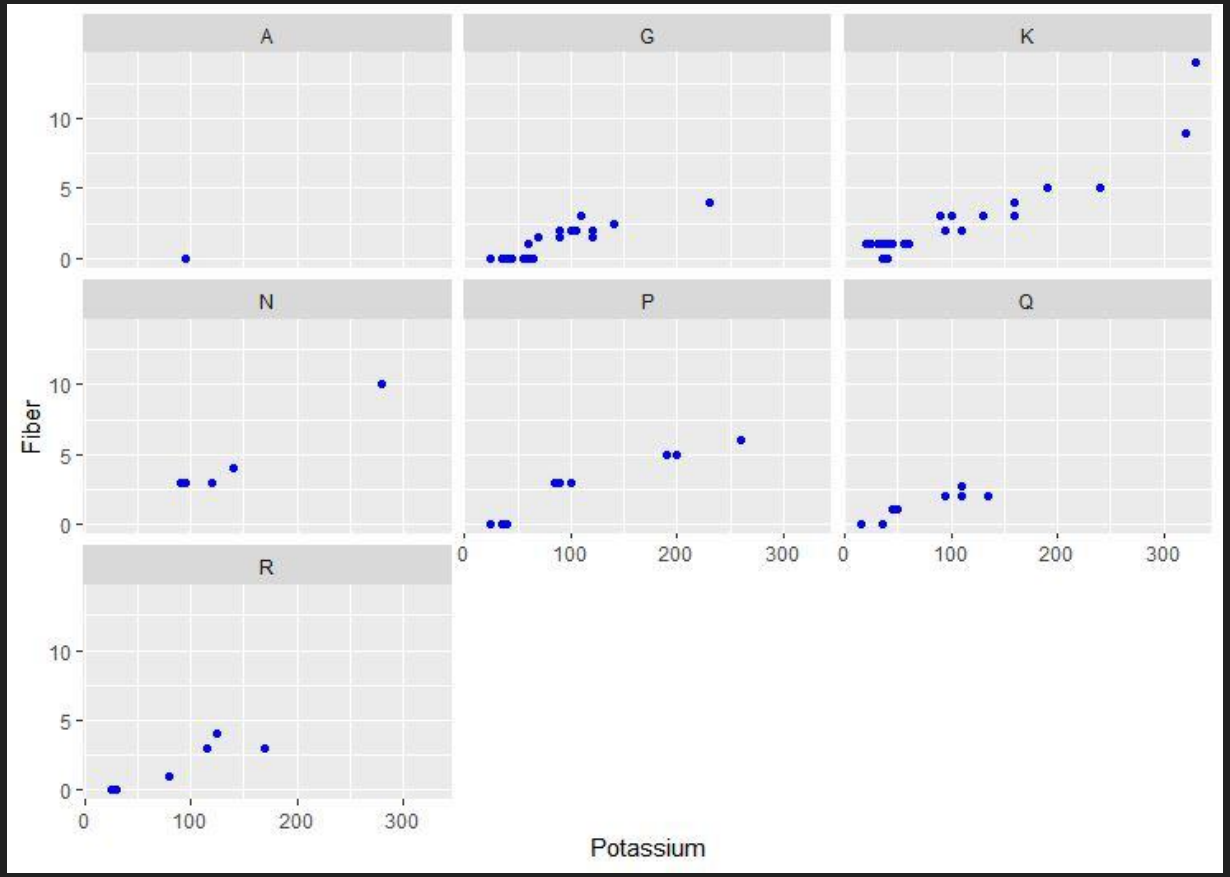
The source of this data did not specify how these ratings were determined, so it should not be the only characteristic taken into account.

```
Cals =ggplot(data=cereal, aes(Calories))
Cals+geom_histogram(aes(y=..density..),color="black",fill="pink")+
  geom_density(color="purple")
```

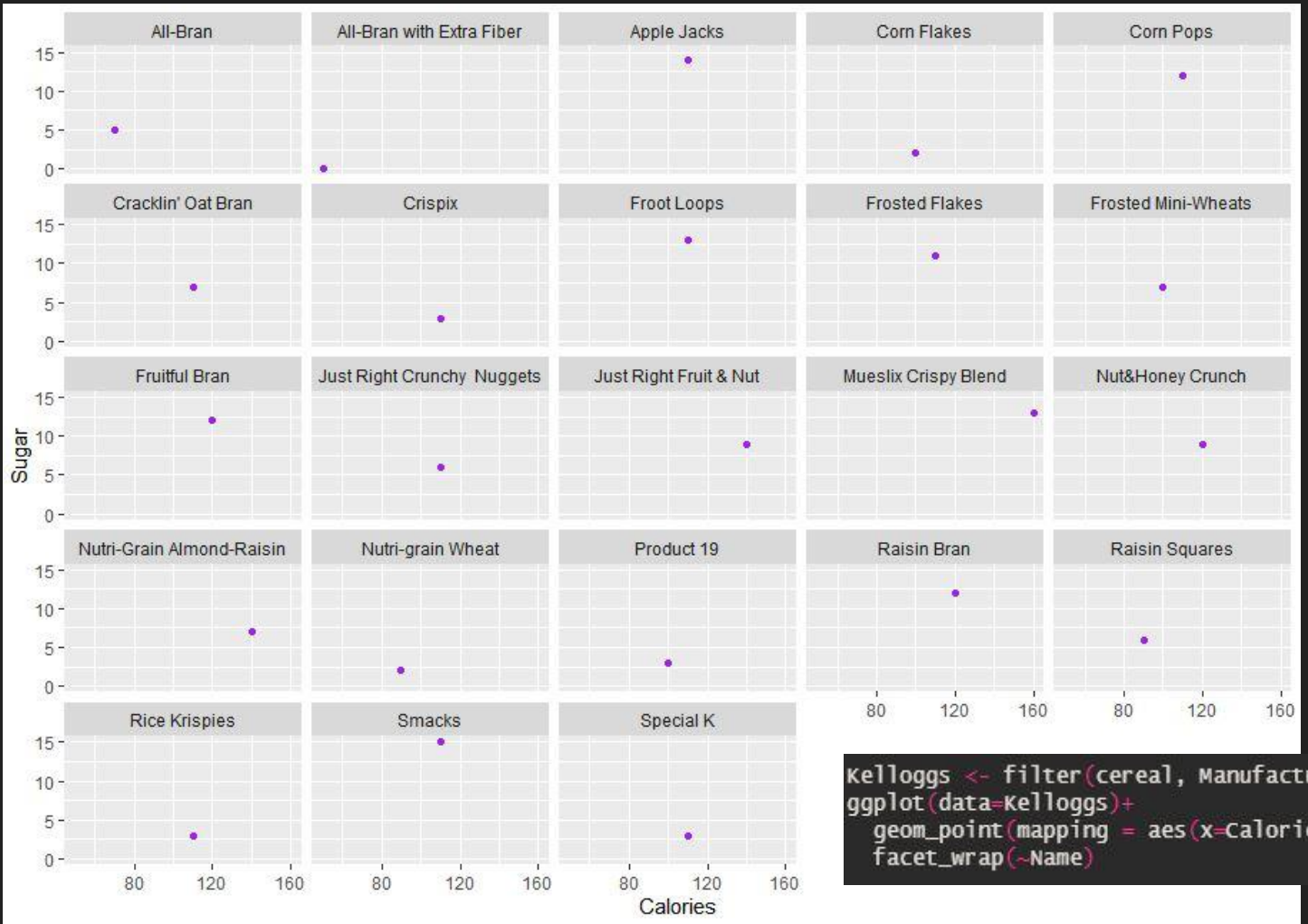


This is a histogram representing the calories of the overall data set. From this, we can see that the majority of these products have a caloric total of about 90 to 120.

```
ggplot(data=cereal)+
  geom_point(mapping=aes(x=Potassium, y=Fiber), color="blue")+
  facet_wrap(~Manufacturer, nrow=3)
```



These plots are separated by manufacturer, and they show the potassium and fiber amounts in each cereal. Points in the top right of each plot represent products with high amount of both things, such as those of Kellogg's. These two factors are an important part of how we determined the healthiest cereal, as it is something consumers look for in particular.




```
Kelloggs <- filter(cereal, Manufacturer=="K")
ggplot(data=Kelloggs)+
  geom_point(mapping = aes(x=Calories, y=Sugar), color="purple")+
  facet_wrap(~Name)
```

Conclusions

- Kellogg's products are some of the highest rated, and they have lots of the needed nutrients and a lot less of th unhealthy ones in their products.
- Make up the majority of this data set, so we decided that the healthiest cereal would reasonably come from Kellogg's.
- Previous plots show the amount of calories and sugar of each cereal made by Kellogg's.
- The healthiest one would be lower calorie and have little to no sugar.
- Any cereal with a point at the top of the very right side of the plot is eliminated.
- From this, we deducted that All-Bran with Extra Fiber is the healthiest cereal out of the 77 products in this data set because of its calories per serving, its low content of unhealthy ingredients, its high nutritious value, and high rating.
- Of course, this study did not take into account taste, texture, and other factors that would affect individual preferences. Since there are many more products out in the market now, this study could be redone with updated data. From an objective standpoint, Kellogg’s is the cereal powerhouse through its variety of offerings that are able to accommodate all kinds of dietary needs.

The Winner:


Kellogg's
Nutrition at a Glance
ALL BRAN
Extra Fiber
Natural Wheat Bran
High Fiber Cereal
TRY THE
10 DAY
CHALLENGE
NET WT 11.4 OZ. (323g)