Cereal EDA

Taylor Anderson

4/6/2020

EDA

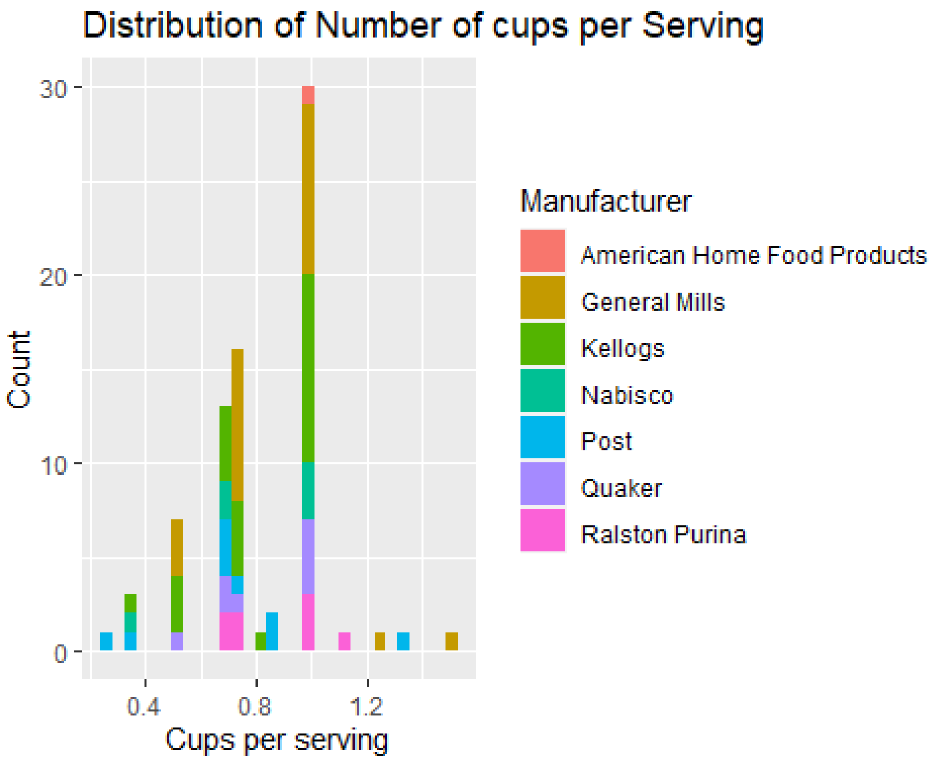
```
summary(dat)
```

```
##        X          name           manufacturer           type
##  Min.   : 0   Length:77          Length:77          Length:77
##  1st Qu.:19   Class :character   Class :character   Class :character
##  Median :38   Mode  :character   Mode  :character   Mode  :character
##  Mean   :38
##  3rd Qu.:57
##  Max.   :76
##     calories         protein            fat             sodium
##  Min.   : 50.0   Min.   :1.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:0.1300
##  Median :110.0   Median :3.000   Median :1.000   Median :0.1800
##  Mean   :106.9   Mean   :2.545   Mean   :1.013   Mean   :0.1597
##  3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:0.2100
##  Max.   :160.0   Max.   :6.000   Max.   :5.000   Max.   :0.3200
##      fiber            carbo           sugars           potass
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 1.000   1st Qu.:12.00   1st Qu.: 3.000   1st Qu.:0.0400
##  Median : 2.000   Median :14.00   Median : 7.000   Median :0.0900
##  Mean   : 2.152   Mean   :14.61   Mean   : 6.935   Mean   :0.0961
##  3rd Qu.: 3.000   3rd Qu.:17.00   3rd Qu.:11.000   3rd Qu.:0.1200
##  Max.   :14.000   Max.   :23.00   Max.   :15.000   Max.   :0.3300
##     vitamins          shelf           weight          cups
##  Min.   :  0.00   Min.   :1.000   Min.   :0.50   Min.   :0.250
##  1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670
##  Median : 25.00   Median :2.000   Median :1.00   Median :0.750
##  Mean   : 28.25   Mean   :2.208   Mean   :1.03   Mean   :0.821
##  3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000
##  Max.   :100.00   Max.   :3.000   Max.   :1.50   Max.   :1.500
##      rating
##  Min.   :18.04
##  1st Qu.:33.17
##  Median :40.40
##  Mean   :42.67
##  3rd Qu.:50.83
##  Max.   :93.70
```

Distribution of Number of cups per Serving

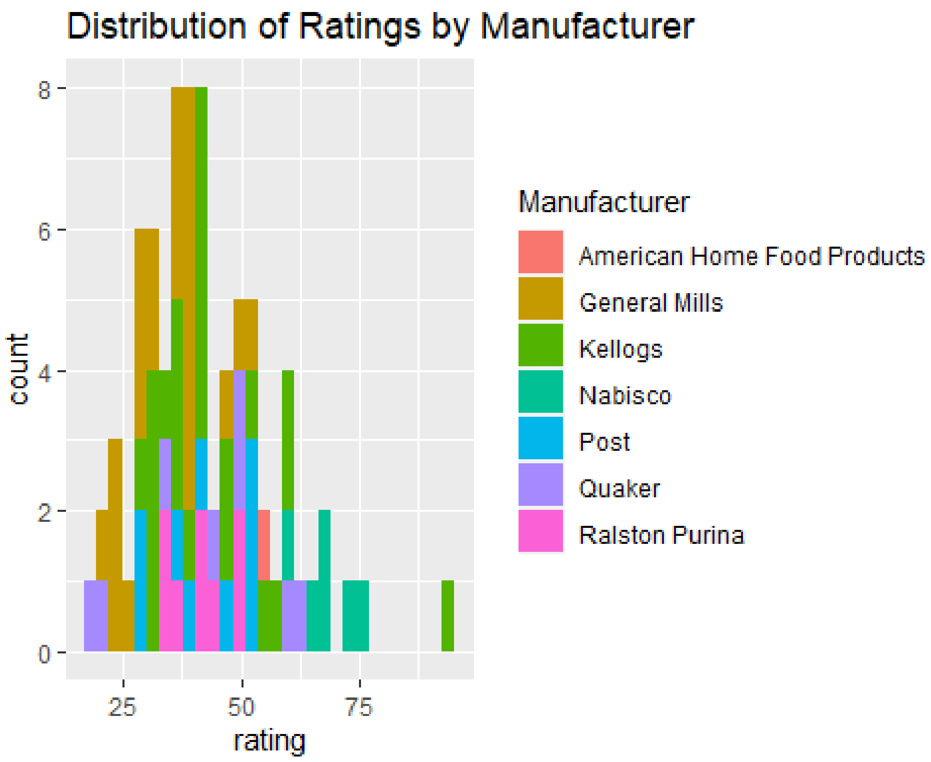
The first thing that caught out attention was the number of cups of cereal per serving. We thought that the cereal’s rating might be affected based on the weight of the cereal being used.

```
library(ggplot2)

ggplot(dat) +
  geom_histogram(aes(x = cups, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Number of cups per
Serving", x = "Cups per serving", y = "Count")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



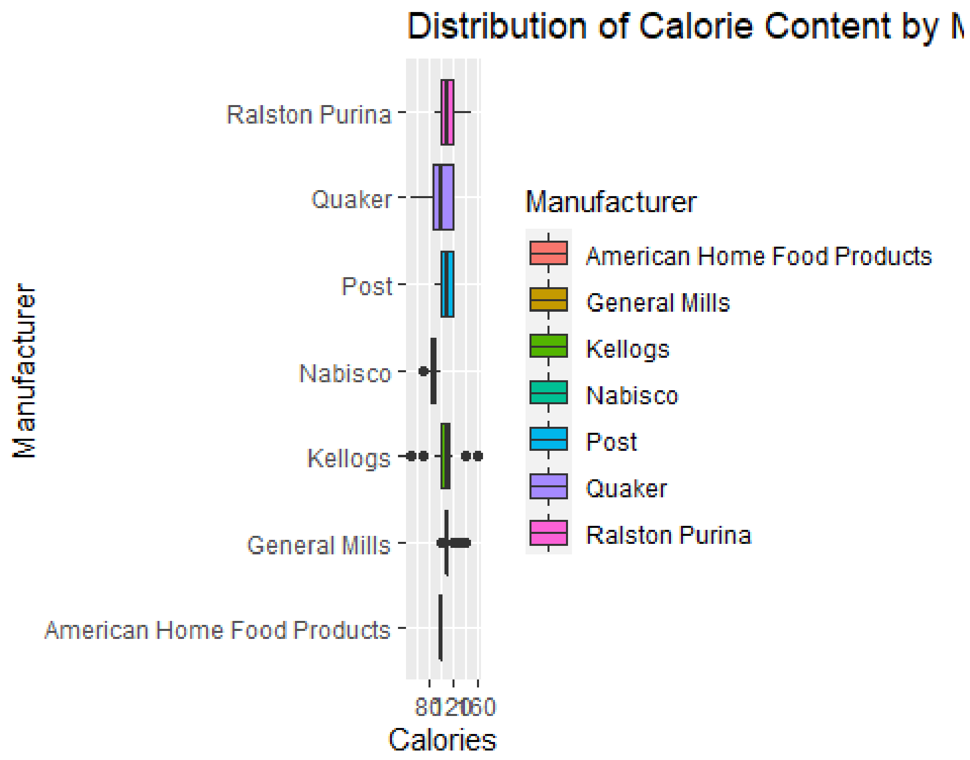
# Distribution of Ratings by Manufacturer

```
ggplot(dat) +
  geom_histogram(aes(x = rating, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Ratings by
Manufacturer")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



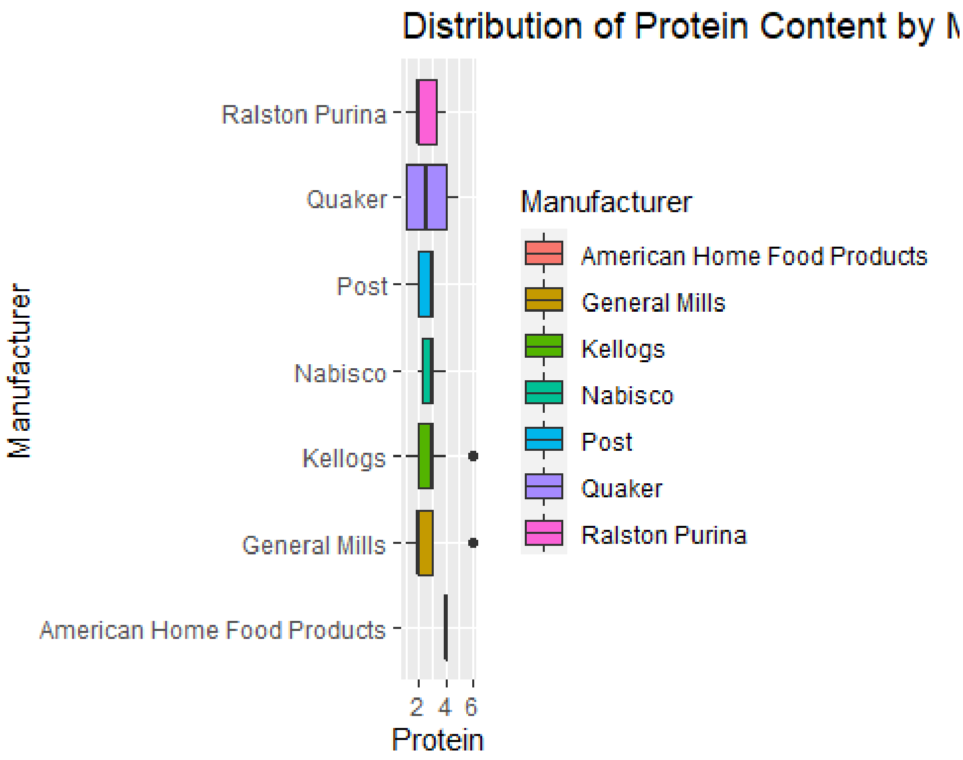
Distribution of Calorie Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = calories, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Calorie Content by
Manufacturer", x = "Calories", y = "Manufacturer")
```



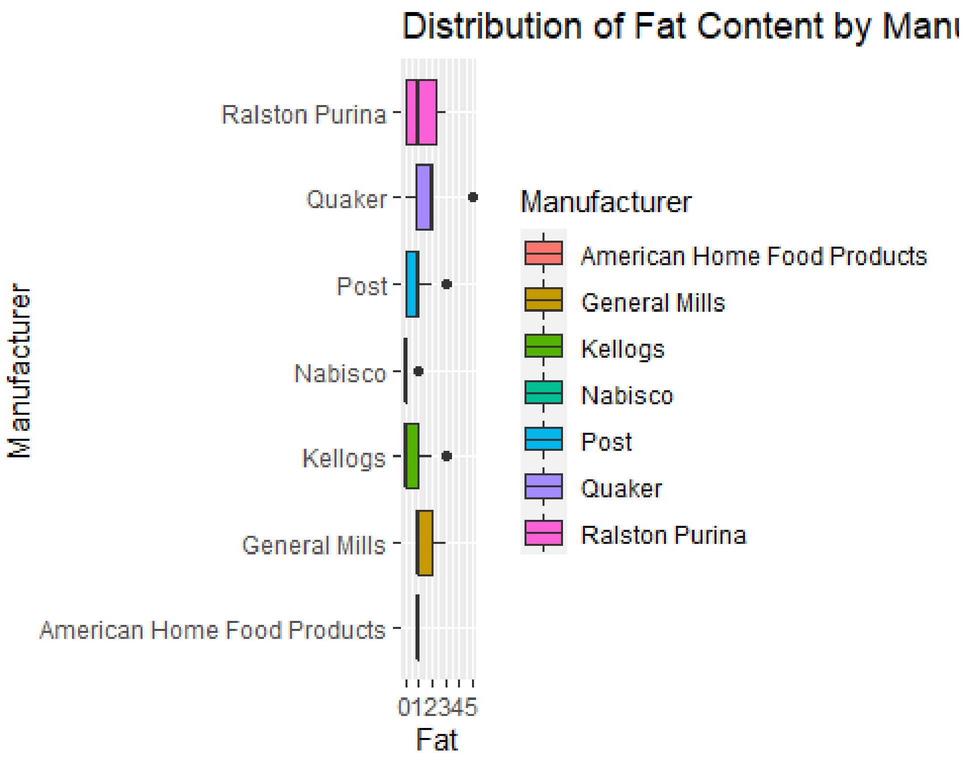
# Distribution of Protein Content by Manufacturer

```r
ggplot(dat) +
  geom_boxplot(aes(x = protein, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Protein Content by 
Manufacturer", x = "Protein", y = "Manufacturer")
```



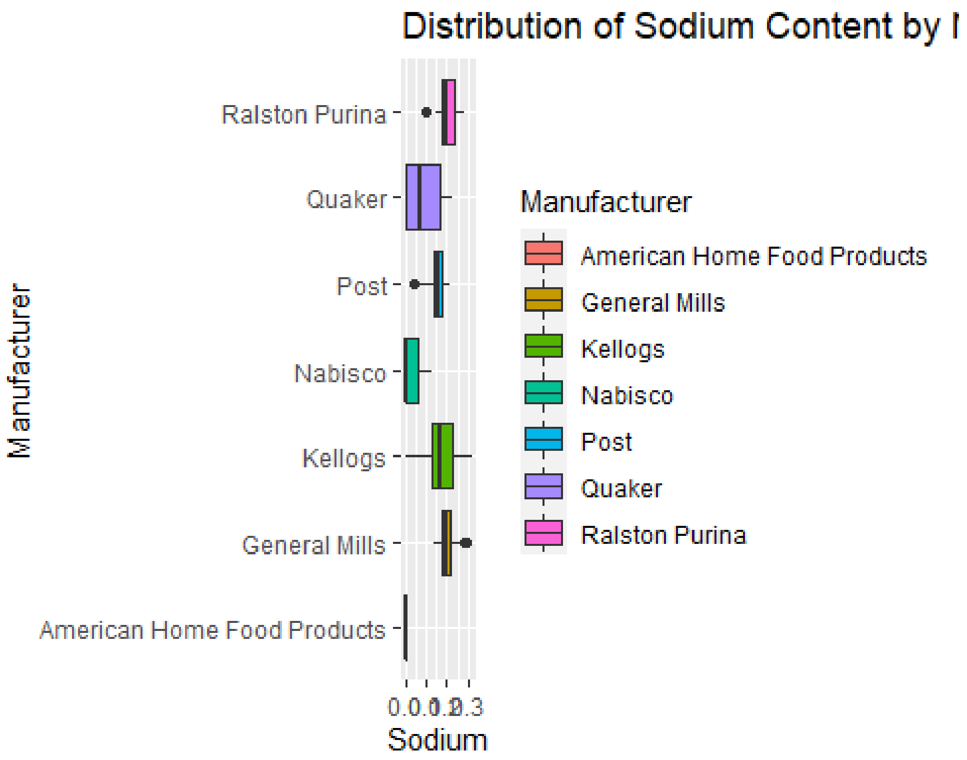
# Distribution of Fat Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = fat, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Fat Content by
Manufacturer", x = "Fat", y = "Manufacturer")
```



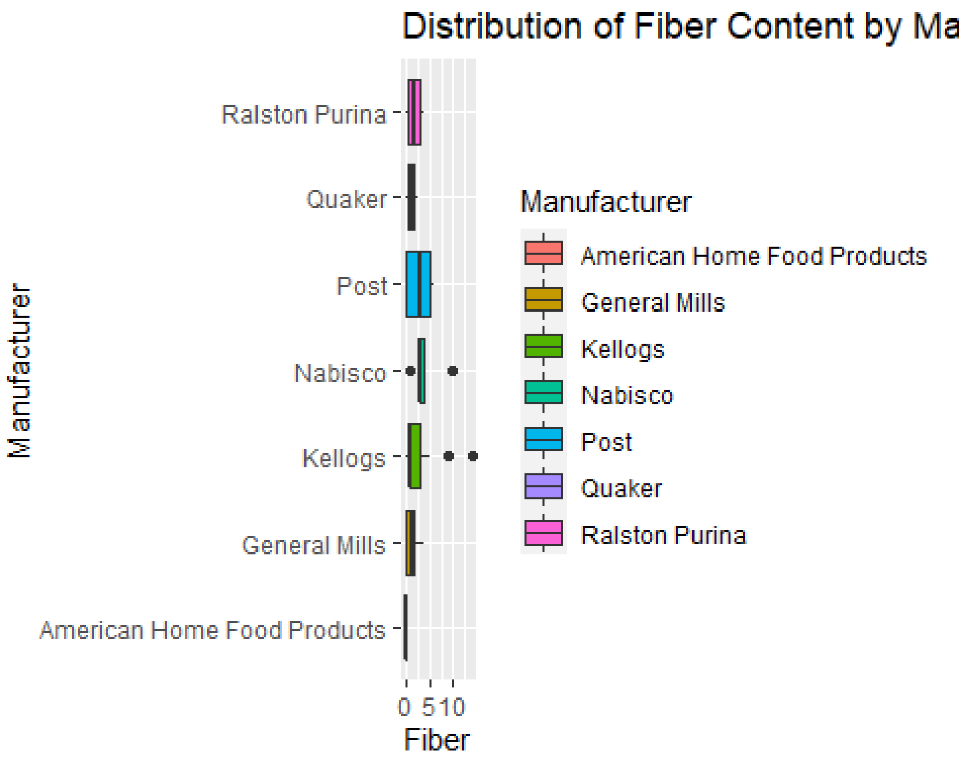
# Distribution of Sodium Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = sodium, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Sodium Content by
Manufacturer", x = "Sodium", y = "Manufacturer")
```



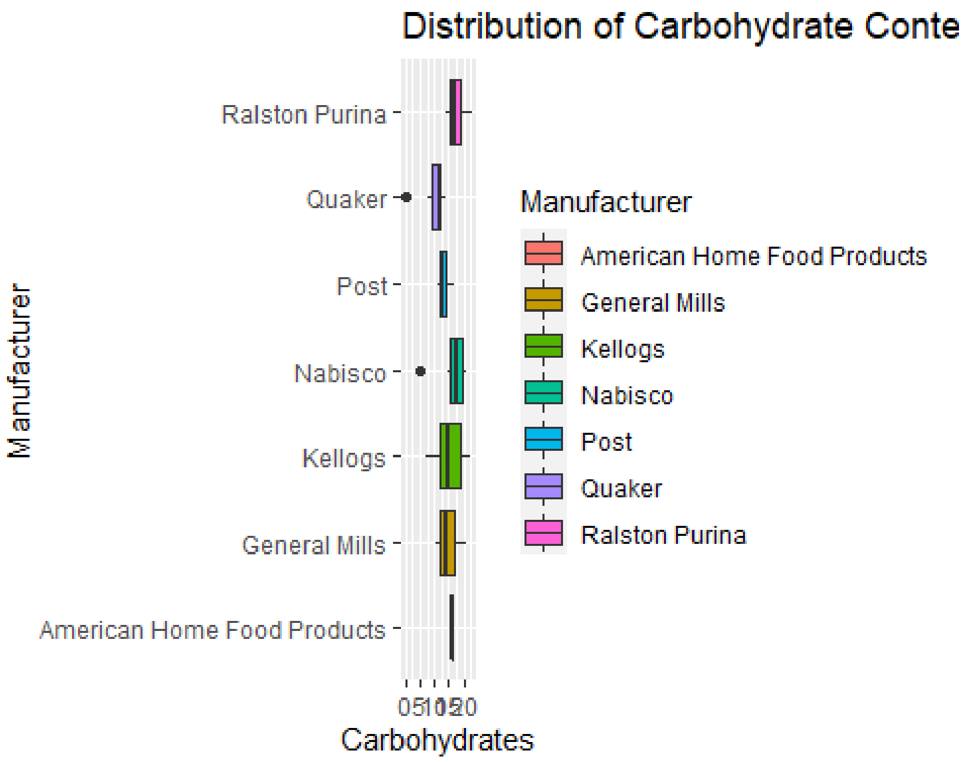
Distribution of Fiber Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = fiber, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Fiber Content by
Manufacturer", x = "Fiber", y = "Manufacturer")
```



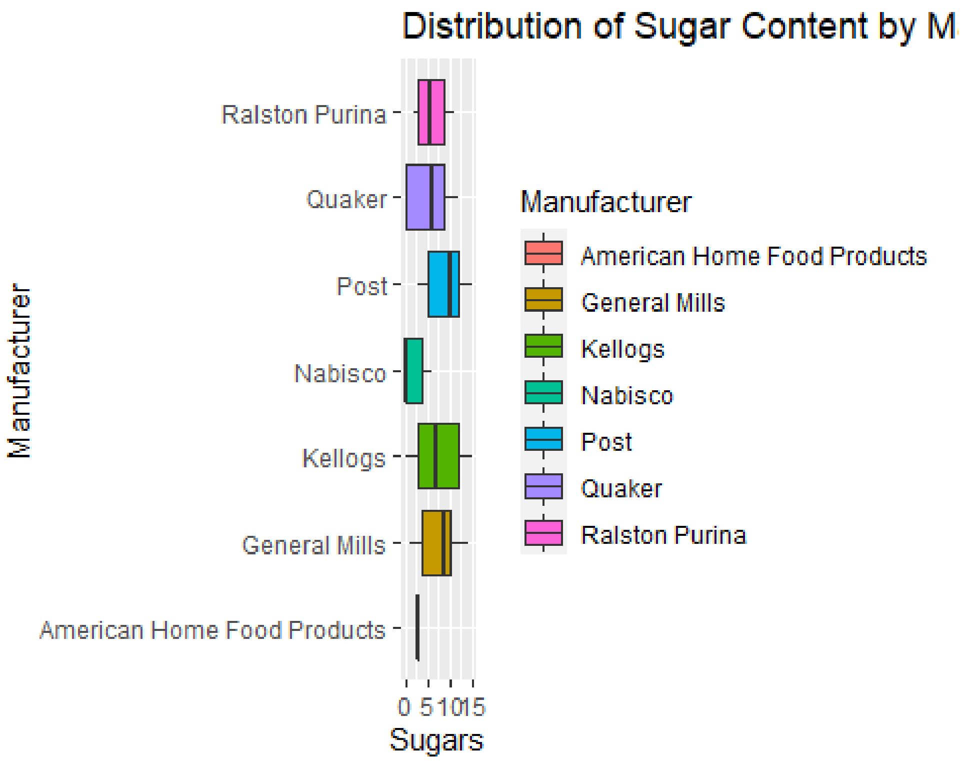
Distribution of Carbohydrates Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = carbo, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Carbohydrate Content
by Manufacturer", x = "Carbohydrates", y = "Manufacturer")
```



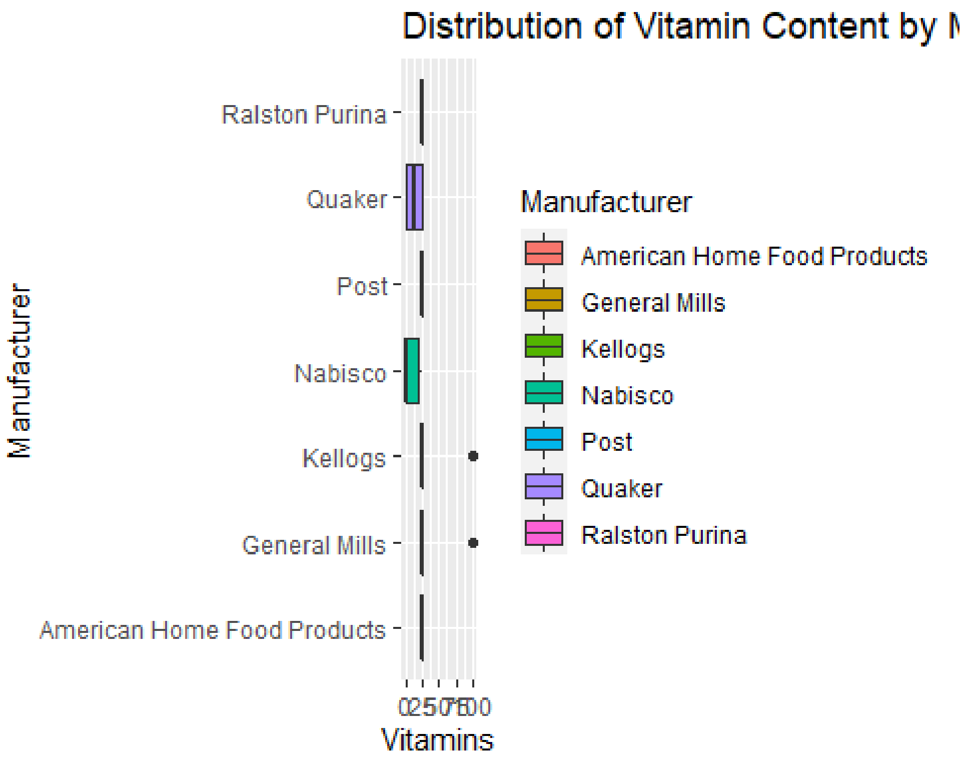
Distribution of Sugars Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = sugars, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Sugar Content by
Manufacturer", x = "Sugars", y = "Manufacturer")
```



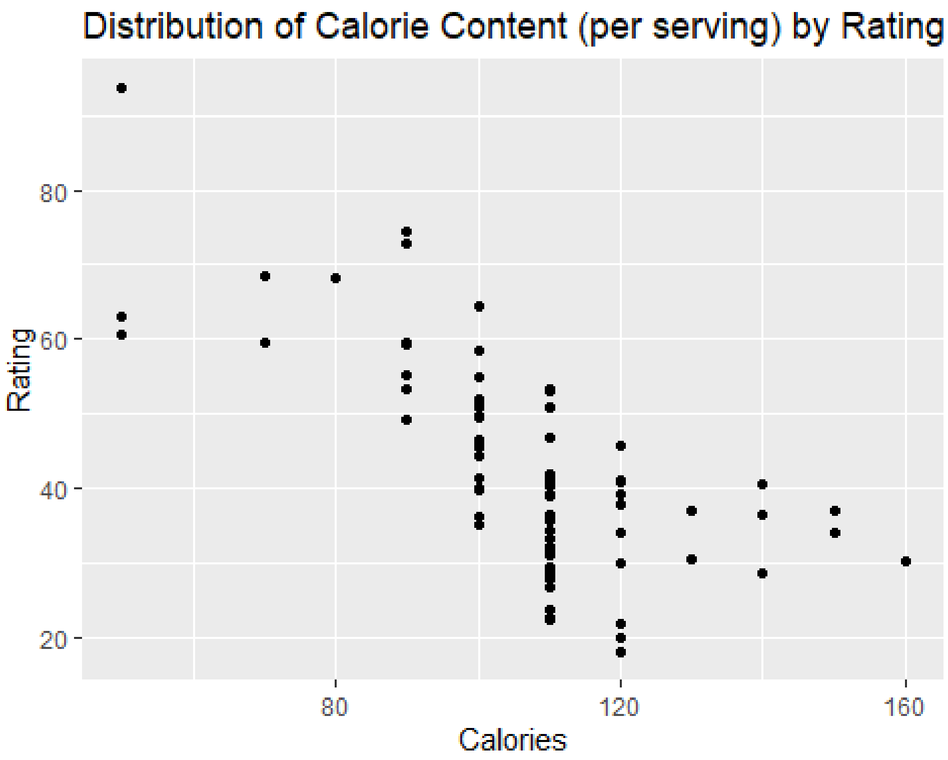
Distribution of Vitamin Content by Manufacturer

```
ggplot(dat) +
  geom_boxplot(aes(x = vitamins, y = manufacturer, fill = manufacturer)) +
  labs(fill = "Manufacturer", title = "Distribution of Vitamin Content by
Manufacturer", x = "Vitamins", y = "Manufacturer")
```



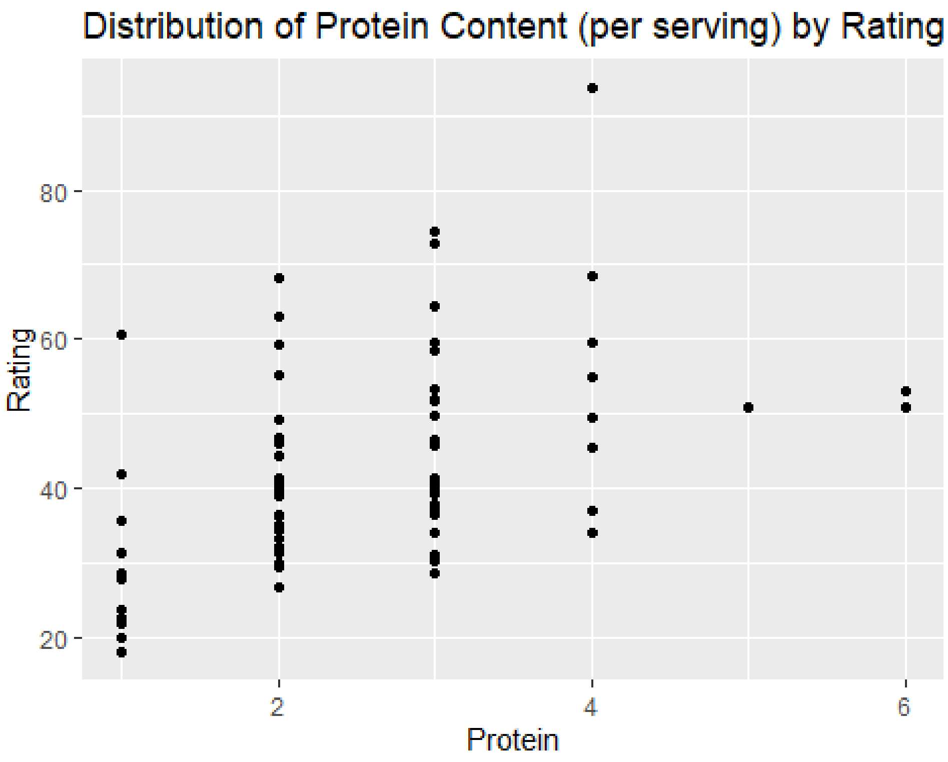
Distribution of Calories Content by Rating

```
ggplot(dat, aes(x = calories, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Calorie Content (per serving) by Rating", x = "Calories", y = "Rating")
```



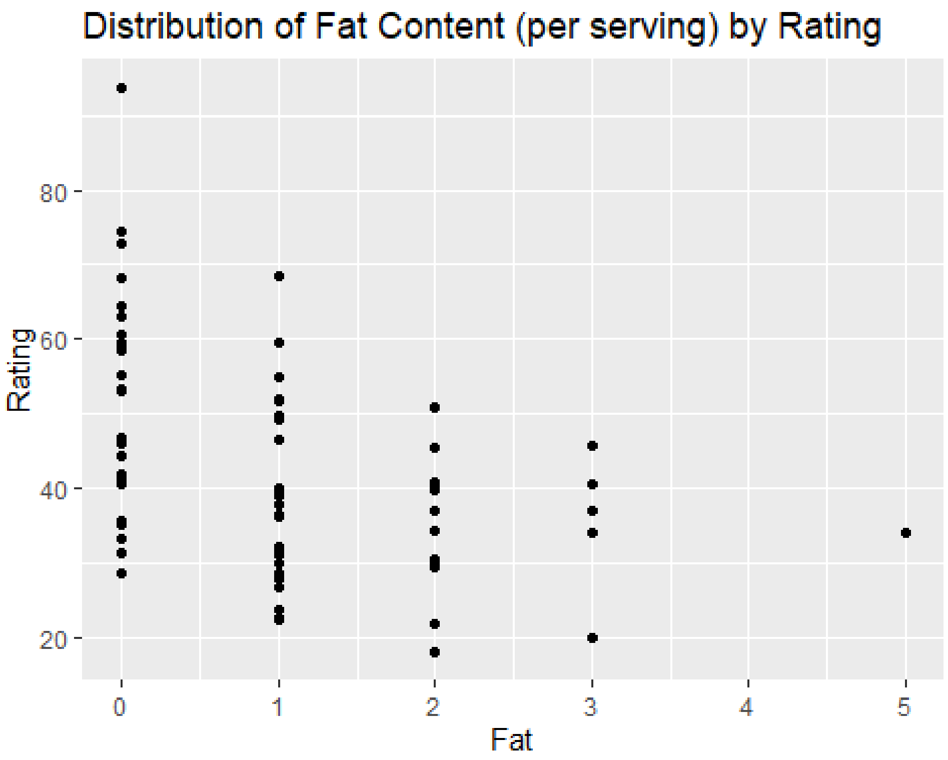
Distribution of Protein Content by Rating

```
ggplot(dat, aes(x = protein, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Protein Content (per serving) by Rating", x =
"Protein", y = "Rating")
```



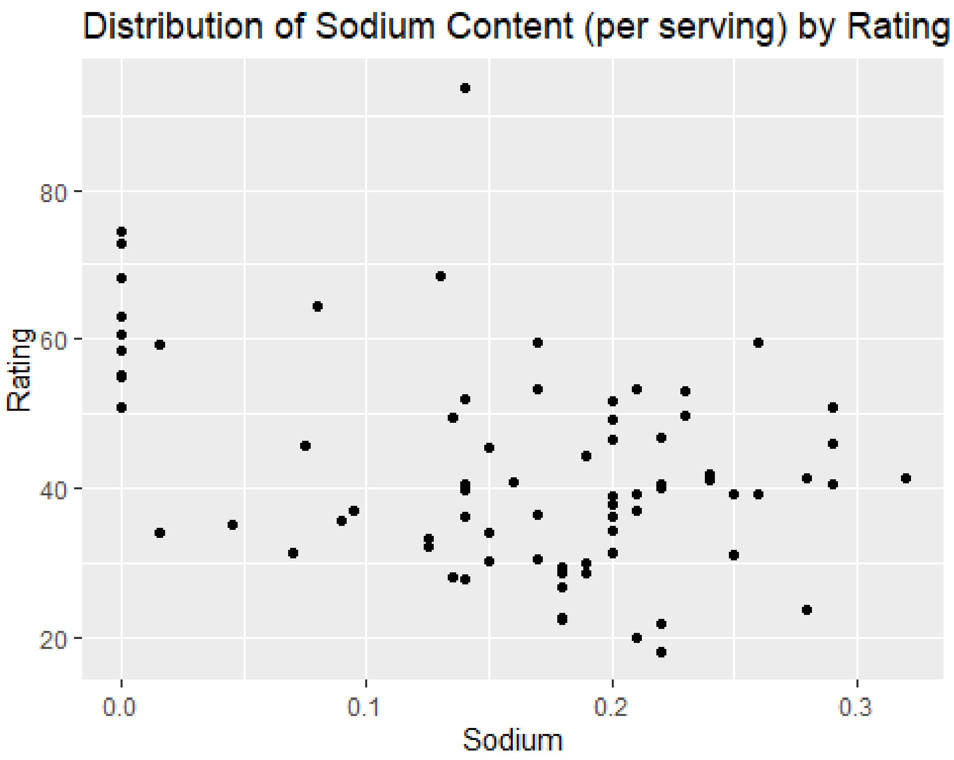
Distribution of Fat Content by Rating

```
ggplot(dat, aes(x = fat, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Fat Content (per serving) by Rating", x =
"Fat", y = "Rating")
```



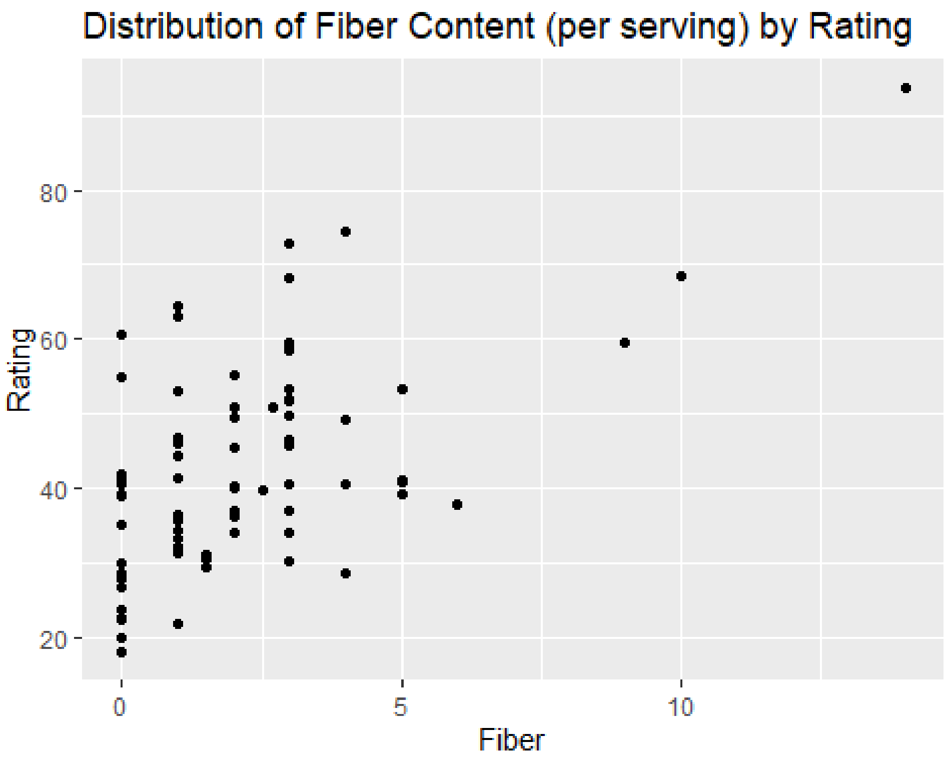
Distribution of Sodium Content by Rating

```
ggplot(dat, aes(x = sodium, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Sodium Content (per serving) by Rating", x =
"Sodium", y = "Rating")
```



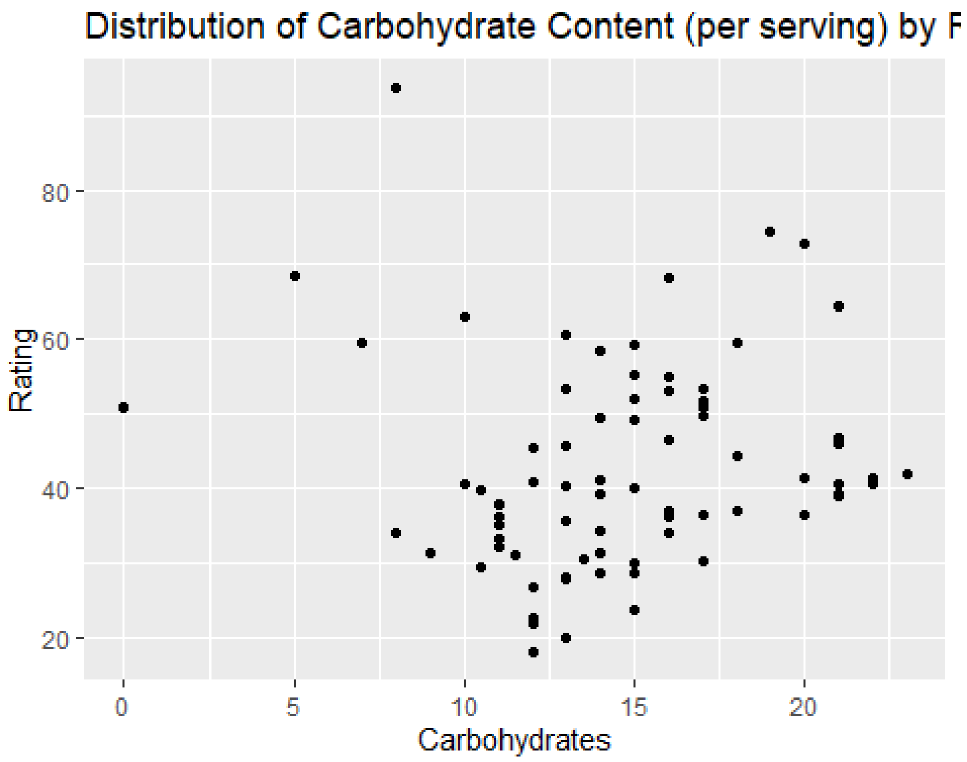
Distribution of Fiber Content by Rating

```
ggplot(dat, aes(x = fiber, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Fiber Content (per serving) by Rating", x =
"Fiber", y = "Rating")
```



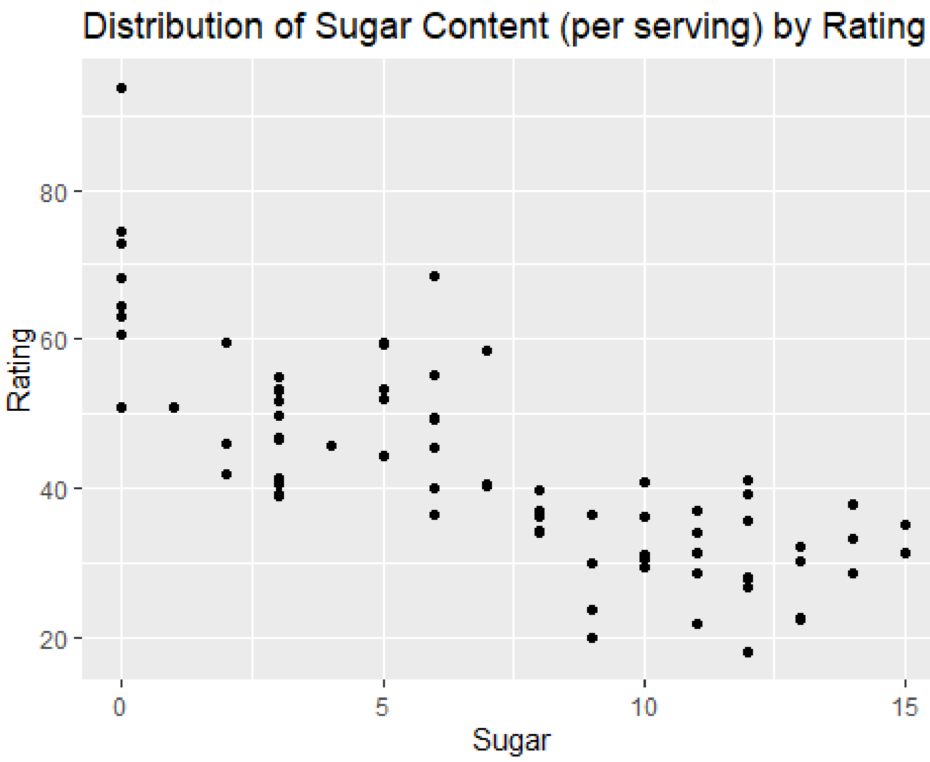
Distribution of Carbohydrate Content by Rating

```
ggplot(dat, aes(x = carbo, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Carbohydrate Content (per serving) by
Rating", x = "Carbohydrates", y = "Rating")
```



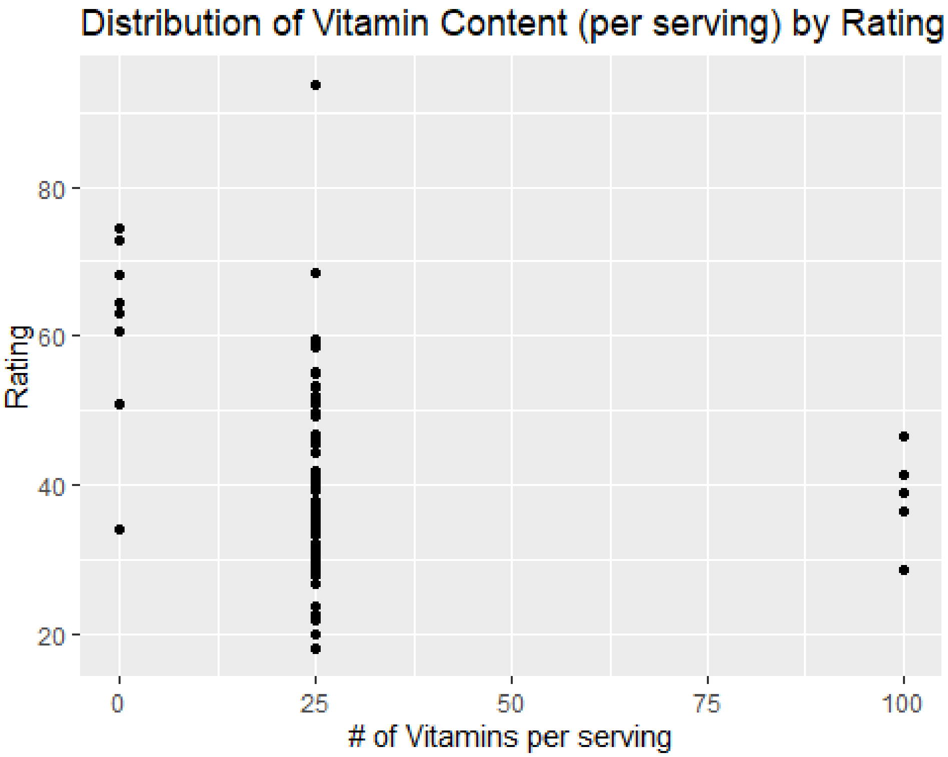
Distribution of Sugar Content by Rating

```
ggplot(dat, aes(x = sugars, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Sugar Content (per serving) by Rating", x =
"Sugar", y = "Rating")
```



Distribution of Vitamin Content by Rating

```
ggplot(dat, aes(x = vitamins, y = rating)) +
  geom_point() +
  labs(title = "Distribution of Vitamin Content (per serving) by Rating", x =
"# of Vitamins per serving", y = "Rating")
```


Distribution of Vitamin Content (per serving) by Rating
Rating
80
60
40
20
0
25
50
75
100
# of Vitamins per serving