Cereals – Exploratory Analysis (EDA)

Made by majafoi

4/25/2021

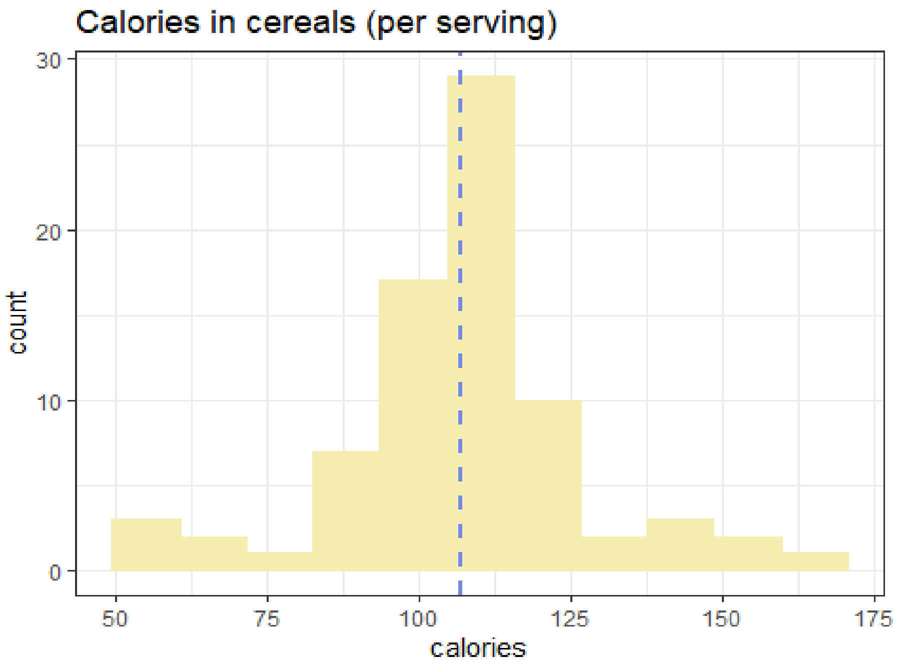
```
library(ggplot2)

cereals <- read.csv("cereal.csv") # to load our dataset called Cereals
attach(cereals) # to omit the need of dataset-name$variable-name

str(cereals) # for checking if the variables are numeric or character

## 'data.frame':   77 obs. of  16 variables:
##  $ name    : chr  "100% Bran" "100% Natural Bran" "All-Bran" "All-Bran with Extra Fiber" ...
##  $ mfr     : chr  "N" "Q" "K" "K" ...
##  $ type    : chr  "C" "C" "C" "C" ...
##  $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
##  $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
##  $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass  : int  280 135 320 330 1 70 30 100 125 190 ...
##  $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
##  $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
##  $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
##  $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
##  $ rating  : num  68.4 34 59.4 93.7 34.4 ...
```

EDA - Exploratory Analysis of the variables in the dataset + Graphics (using ggplot2 package

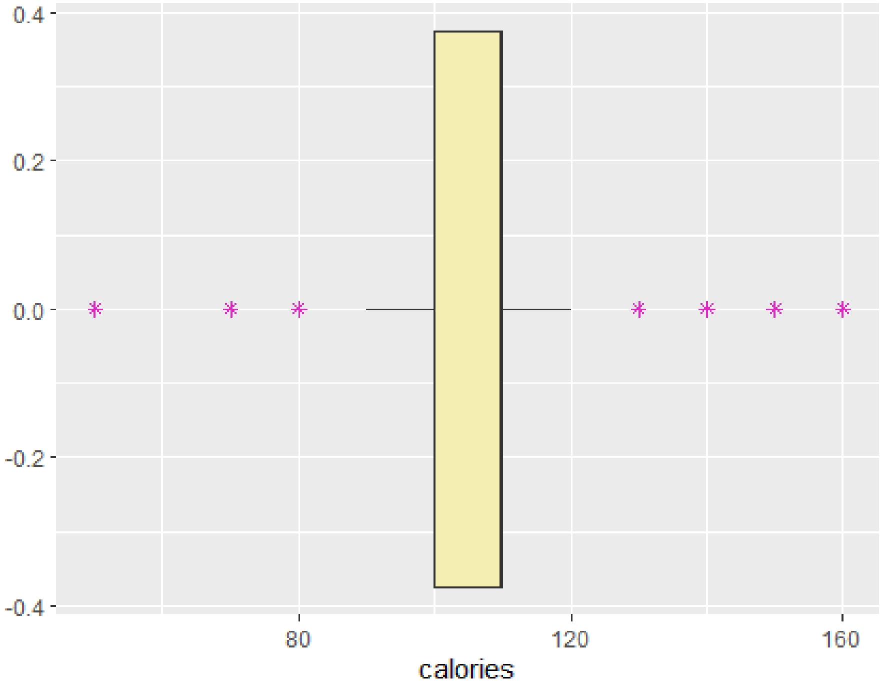
```
summary (type)

##   Length     Class      Mode 
##       77 character character
```

- 74 of the cereals are cold type, and only 3 of them are hot type.

```
summary (calories)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    50.0   100.0   110.0   106.9   110.0   160.0

sd (calories)

## [1] 19.48412

IQR (calories)

## [1] 10

Bin = diff (range(cereals$calories))/10
ggplot (cereals, aes (x=calories)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Calories in cereals (per serving)') + geom_vline (aes (xintercept = mean (calories)), colour = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```



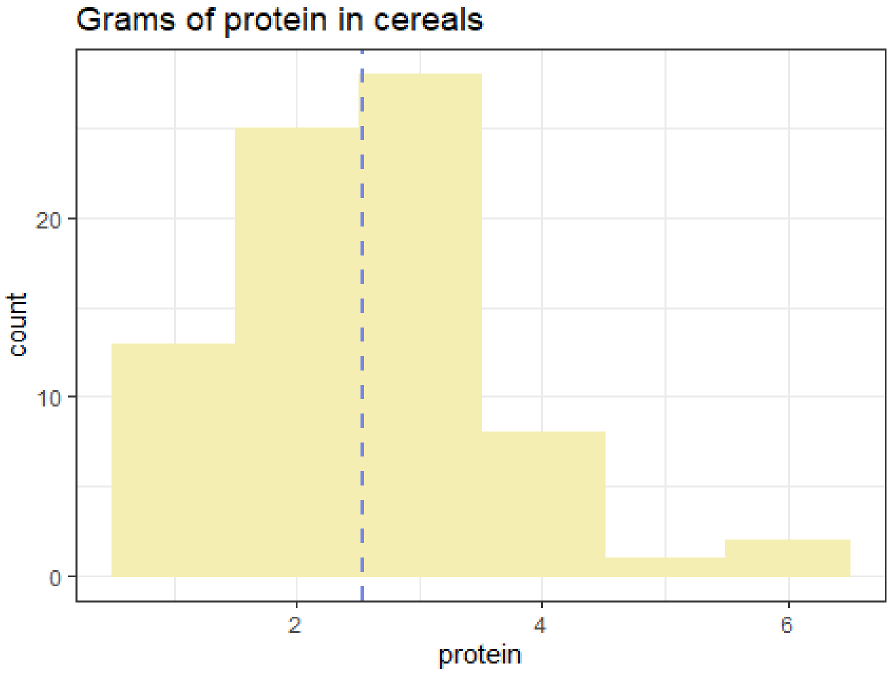
```
ggplot (cereals, aes (x=calories)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape
=8)
```



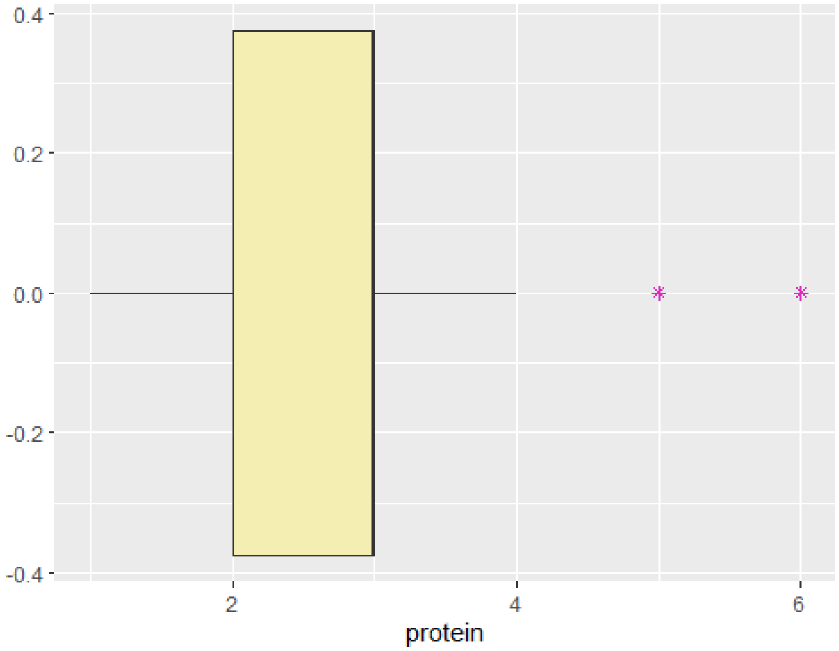
- The average value of calories in the cereals (per serving) is 106.90, with the 50% of the calories value being between 100 and 110 calories per serving (first quartile and the third quartile). Maximum calories per serving is 160, whereas the minimum is 50 calories. Standard deviation (deviation from the mean) is 19.48, which isn't too high.

- On the graph (histogram), you can see a dashed line - that is the pointer for mean value. From histogram perspective, I would say this distribution is close to normal (bell-like shape), but the boxplot shows that we have outliers between 50-80 range and range of 130-160 calories per serving. We can also see that the median is overlapping with the third quartile. IQR (difference between the first and third quartile, which also can show the spread or variability of the data in the middle 50% of the data) is 10. Since the whole variability (maximum - minimum) is 110, there is little variability or spread of data in the middle 50%, which is also shown on the histogram.

```
summary (protein)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.545   3.000   6.000

sd (protein)

## [1] 1.09479

IQR (protein)

## [1] 1

bin = diff (range (cereals$protein))/5
ggplot (cereals, aes(x=protein)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+  t
heme_bw()+ ggtitle ('Grams of protein in cereals') + geom_vline (aes (xintercept = mean (protein)), color="#
7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

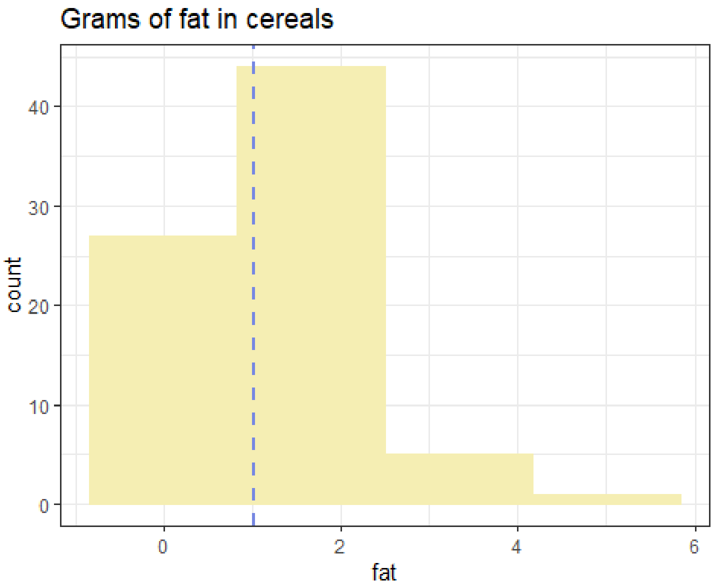


```
ggplot (cereals, aes (x=protein)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape
=8)
```

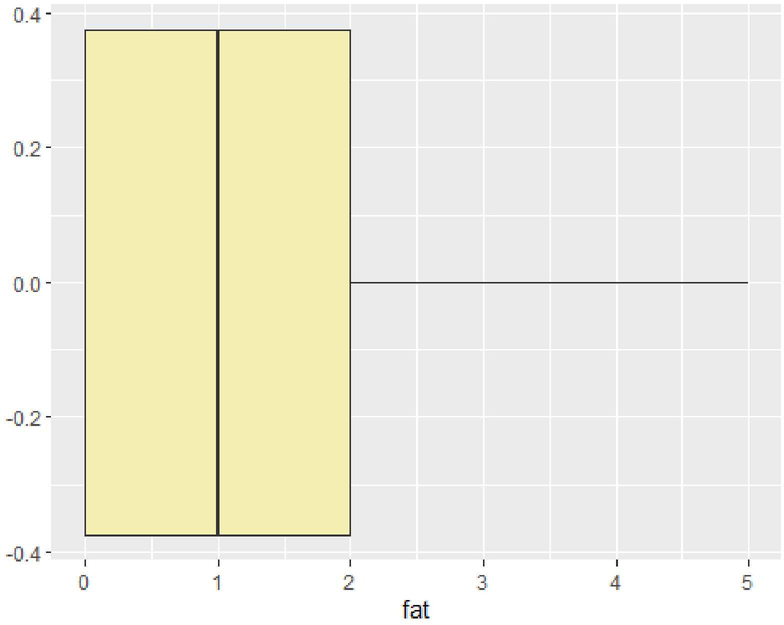


- The minimum grams of protein in the cereals is 1 g, and the maximum 6 g. The average amount of grams of protein is 2.545. 50% of the cereals have between 2 g and 3 g of protein. Standard deviation is 1.09, with IQR of 1 g. Since the variability of this variable is 5 g, the variability in the middle 50% of the distribution is only 1 g which isn't that high. When looking at the distribution on histogram, I can tell that the distribution has a positive skew (right tail), and boxplot confirms that. Also, cereals with 5 g and 6 g of protein in it are outliers, according to boxplot.

```
summary (fat)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.013   2.000   5.000

sd (fat)

## [1] 1.006473

IQR (fat)

## [1] 2

bin = diff (range (cereals$fat))/3
ggplot (cereals, aes (x=fat)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ them
e_bw()+ ggtitle ('Grams of fat in cereals') + geom_vline (aes (xintercept = mean(fat)), color = "#7687E2", fill="
#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

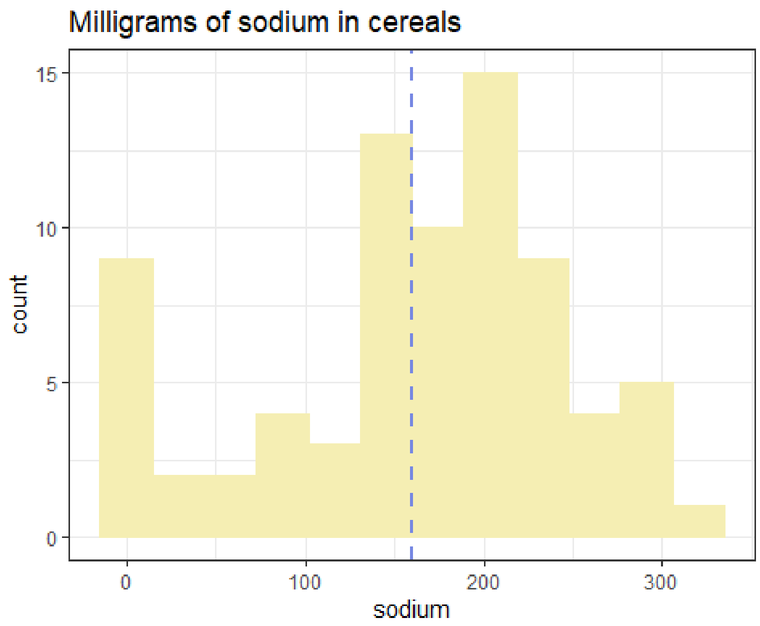


```
ggplot (cereals, aes (x=fat)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8)
```

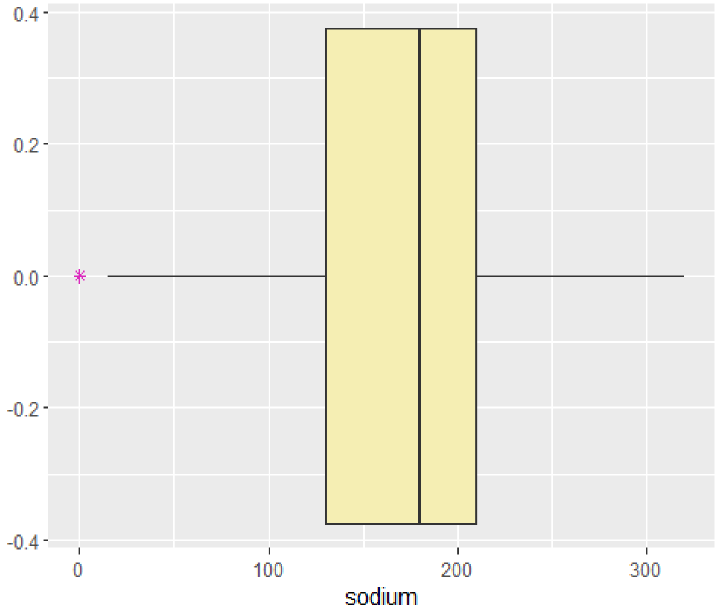


- The minimum grams of fat in cereals is 0g, whereas the maximum value is 5 g. The middle 50% of the distribution has between 0 g and 2 g of fat in it. Standard deviation is 1.006 g, with IQR of 2 g, which is lower than the overall variability in the variable. On the histogram, I can see that the distribution is positively skewed, and the boxplot confirms that. Also, there are no visible outliers in the boxplot.

```
summary (sodium)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   130.0   180.0   159.7   210.0   320.0

sd (sodium)

## [1] 83.8323

IQR (sodium)

## [1] 80

bin = diff (range (cereals$sodium))/11
ggplot (cereals, aes (x=sodium)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw() + ggtitle ('Milligrams of sodium in cereals') + geom_vline (aes (xintercept = mean (sodium)), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

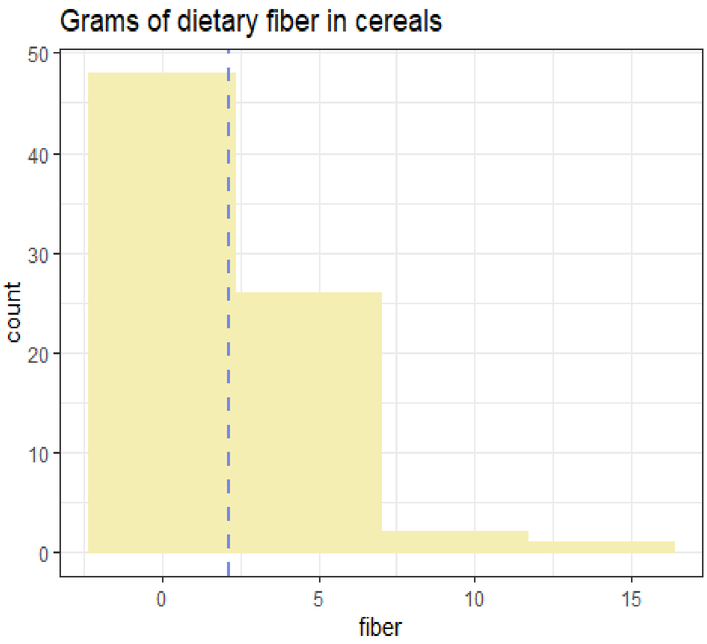


```
ggplot (cereals, aes (x=sodium)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape
=8)
```

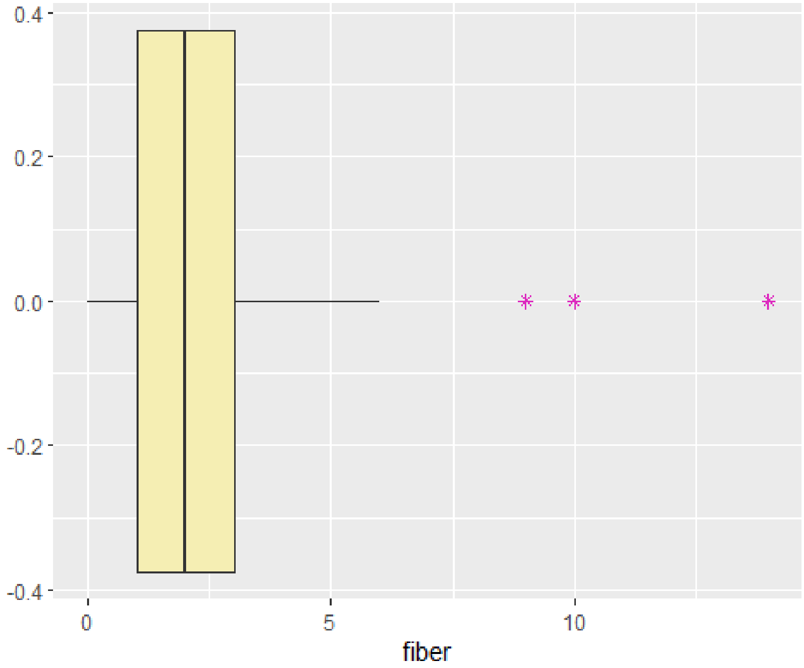


- Minimum amount of milligrams of sodium in the cereals is 0 mg, and the maximum is 320 mg. That immediately shows us that there is a huge variability in the variable distribution. The average value is 159.7 mg, with the middle 50% of the data being in between 130 mg and 210 mg. Standard deviation is a bit higher now - 83.83 mg, with the IQR being 80 mg, which is still lower than the overall variability in this variable's distribution. When looking at the histogram, I can't see a clear conclusion about the skewness, but the boxplot shows that this distribution has negative skewness, and that the outliers are those cereals with 0mg of sodium in it.

```
summary (fiber)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.152   3.000  14.000

sd (fiber)

## [1] 2.383364

IQR (fiber)

## [1] 2

bin = diff (range (cereals$fiber))/3
ggplot (cereals, aes (x=fiber)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ the
me_bw() + ggtitle ('Grams of dietary fiber in cereals') + geom_vline (aes (xintercept = mean(fiber)), color = "
#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

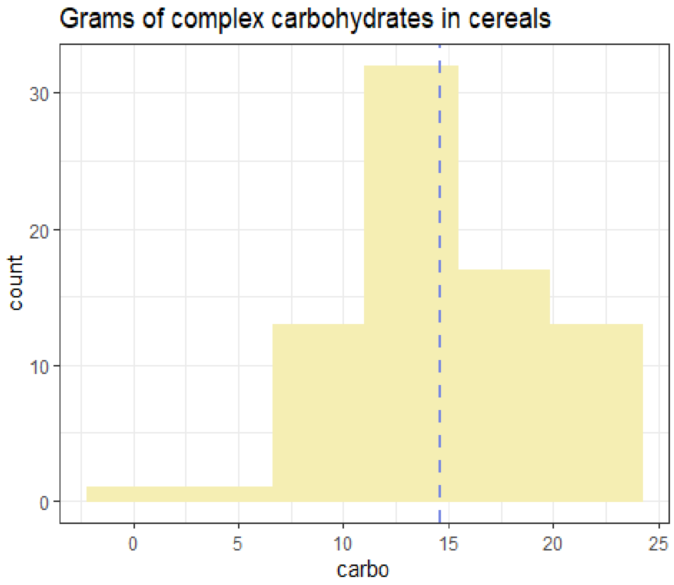


```
ggplot(cereals, aes(x=fiber)) + geom_boxplot(fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8)
```

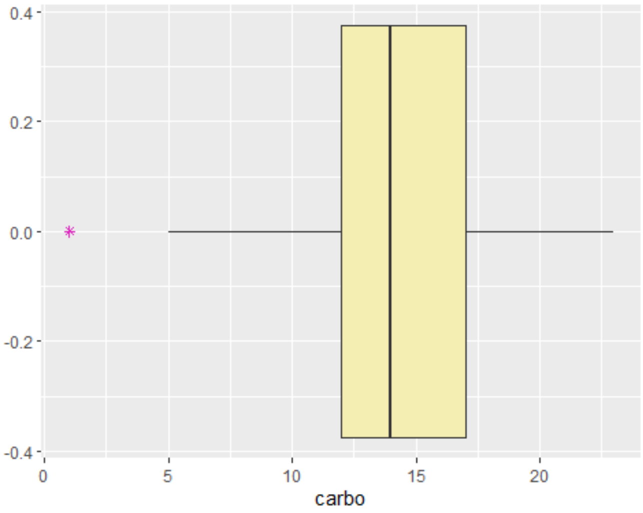


- The minimum amount of grams of dietary fiber in the cereals is 0 g, and the maximum 14 g. The middle 50% of the variable is between 1 g and 3 g, with the average value being 2.152 g. Both standard deviation and IQR are being close to 2 g, but that is still lower than the overall variability of the variable. I can assume we have outliers here because of that. Histogram shows strong positive skewness of the distribution, and on boxplot we can see we have outliers.

```
summary (carbo)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   14.00   14.62   17.00   23.00

sd (carbo)

## [1] 4.188138

IQR (carbo)

## [1] 5

bin = diff (range (cereals$carbo))/5
ggplot (cereals, aes (x=carbo)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Grams of complex carbohydrates in cereals') + geom_vline (aes (xintercept = mean (carbo)), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

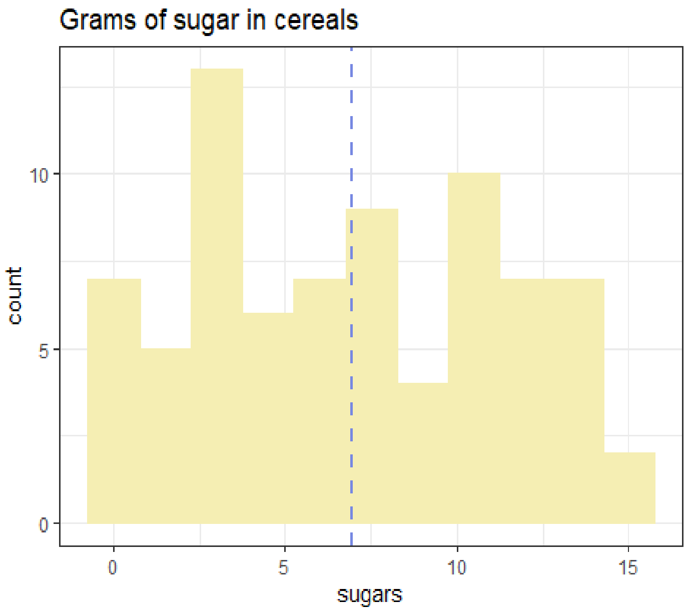


```
ggplot(cereals, aes(x=carbo)) + geom_boxplot(fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8
)
```

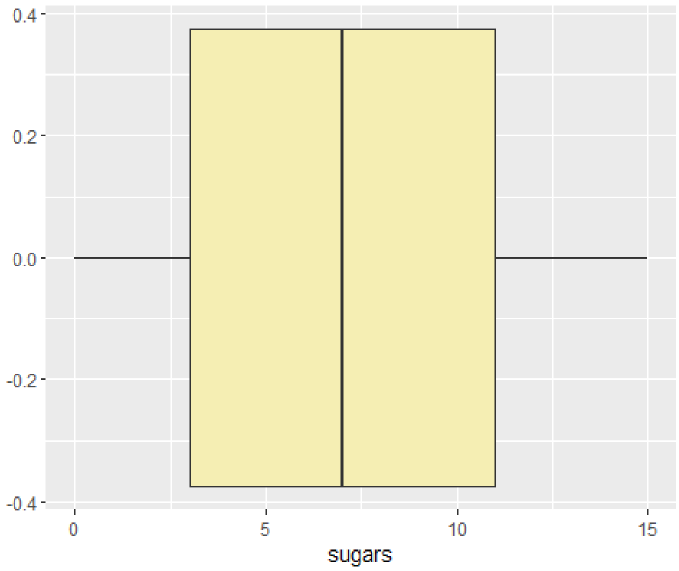


- As for the grams of complex carbohydrates in the cereals, at the beginning one cereal had -1 for this variable, so I have just assumed they mean 1 g and changed it like that. Now we have minimum amount of grams of complex carbohydrates in cereals equal to 1 g, and the maximum amount is 23 g. That could point to a general high variability/spread of the data, but the spread of the middle 50% of distribution is 5 g, which is lower than 22 g. The histogram shows that the distribution is negatively skewed, with one outlier being that cereal with 1 g of complex carbohydrates in it.

```
summary(sugars)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   7.000   6.948  11.000  15.000

sd(sugars)

## [1] 4.403635

IQR(sugars)

## [1] 8

bin = diff(range(cereals$sugars))/10
ggplot(cereals, aes(x=sugars)) + geom_histogram(colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ th
eme_bw()+ ggtitle('Grams of sugar in cereals') + geom_vline(aes(xintercept = mean(sugars)), color = "#768
7E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

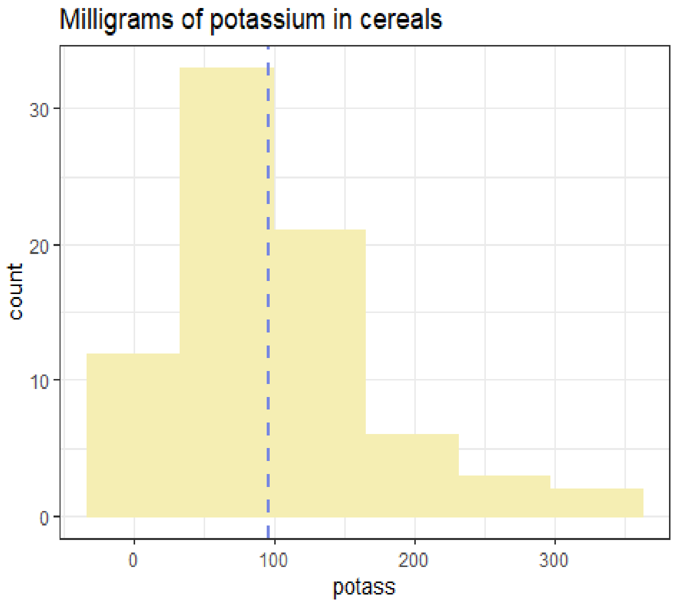


```
ggplot (cereals, aes (x=sugars)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=
8)
```

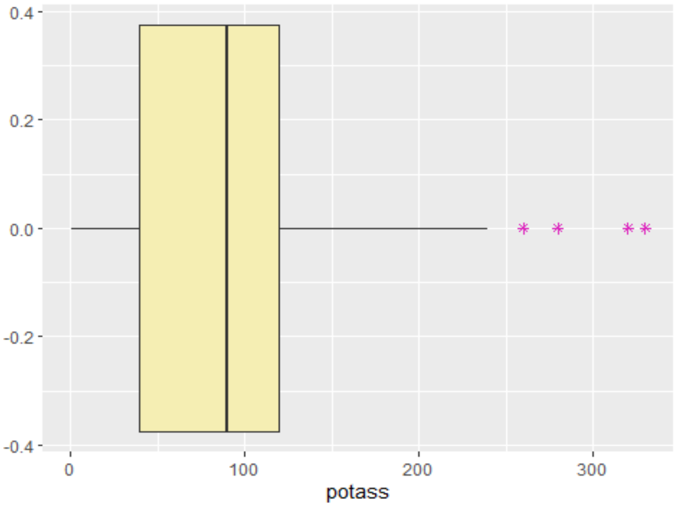


- As the previous case, there was one cereals with -1 g of sugar in it, I have changed it into 1 g. The minimum value of grams of sugar in the cereals is 0 g, whereas the highest amount of sugars is 15 g. That points out to a big variability in the distribution of this variable. The middle 50% of the data is between 3 g and 11 g (variability of 8 g - lower than the general one). The average value is 6.948 g of sugar. Standard deviation is 4.4 g. The histogram doesn't show me anything meaningful about skewness, but I can see on the boxplot that there are no outliers, and that the right whisker is longer than the left one (could point out to a positive skewness).

```
summary (potass)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   90.00   96.13  120.00  330.00

sd (potass)

## [1] 71.21582

IQR (potass)

## [1] 80

bin = diff (range (cereals$potass))/5
ggplot (cereals, aes (x=potass)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Milligrams of potassium in cereals') + geom_vline (aes (xintercept = mean (potass)), colour = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

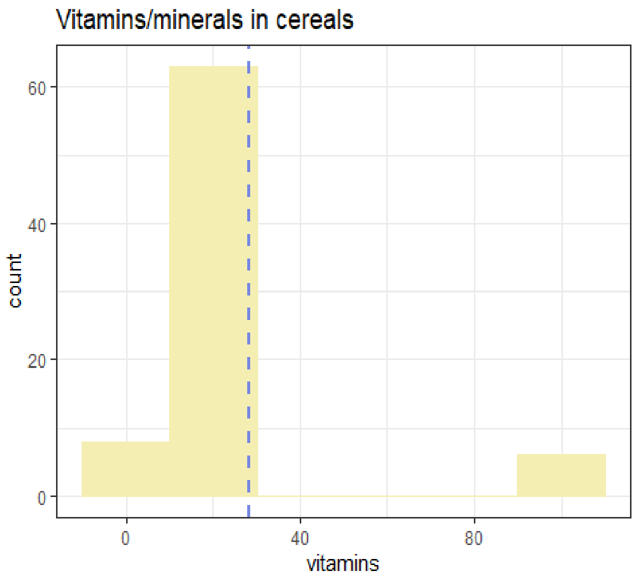


```
ggplot (cereals, aes (x=potass)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=
8)
```

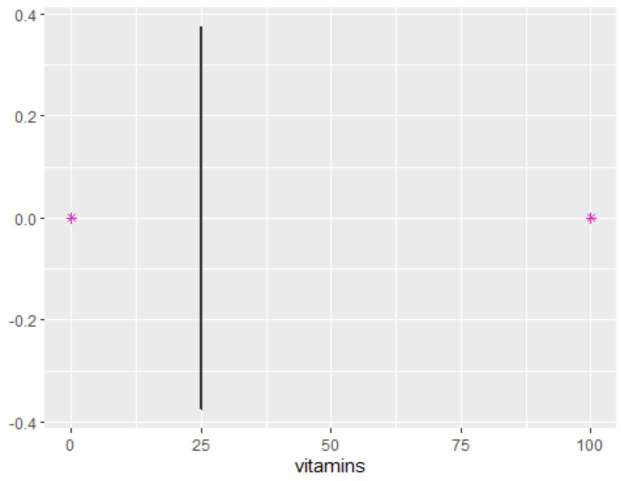


- As for the potassium variable, there were 2 cereals with -1 mg, and I have changed it into 1 mg. The minimum value now is 1 mg, and the maximum value is 330mg. That is the biggest variability/spread of a variable until now. The middle 50% of the data has between 40 mg and 120 mg of potassium in the cereals, which is 80 mg of variability. The average value is 96.13 mg. All this is pointing to probable existence of outliers. The histogram shows that the distribution has a positive skew, with boxplot showing multiple outliers above 250 mg of potassium.

```
summary (vitamins)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   25.00   25.00   28.25   25.00  100.00

bin = diff (range (cereals$vitamins))/5
ggplot (cereals, aes (x=vitamins)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+
theme_bw()+ ggtitle ('Vitamins/minerals in cereals') + geom_vline (aes (xintercept = mean (vitamins)), color=
"#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

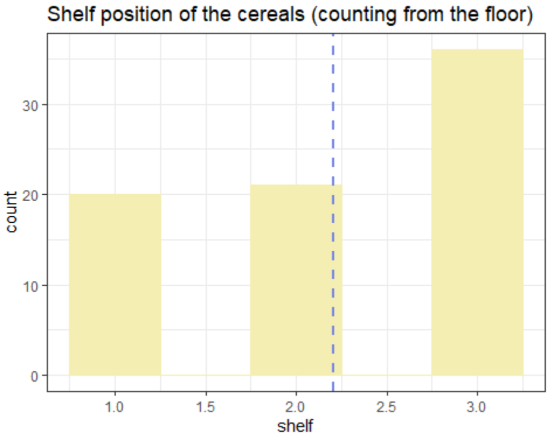


```
ggplot (cereals, aes (x=vitamins)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape
          =8)
```

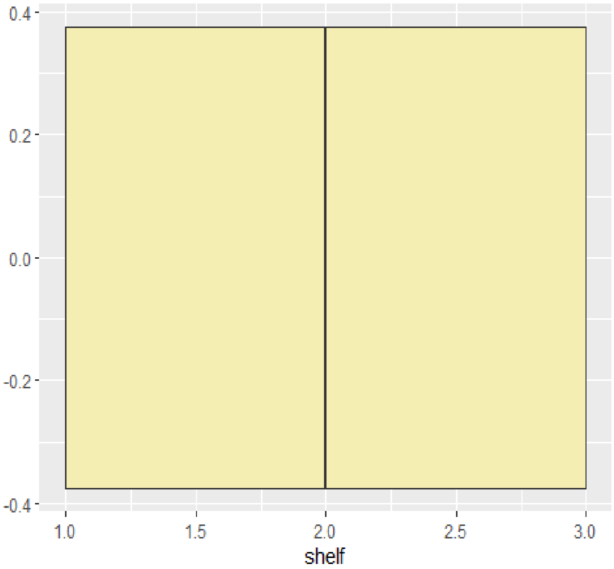


- As for the vitamins, it says in the dataset, it can be 0, 25 or 100, indicating the typical percentage of FDA recommended. The minimum value is 0, and maximum 100, so the whole range is covered. The first and third quartile is 25, with the average value being 28.25. The histogram shows big positive skewness, with outliers being those cereals with vitamins with value of 0 and 100.

```
summary (shelf)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   2.000   2.208   3.000   3.000

sd (shelf)

## [1] 0.8325241

IQR (shelf)

## [1] 2

bin = diff (range (cereals$shelf))/4
ggplot (cereals, aes (x=shelf)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Shelf position of the cereals (counting from the floor)') + geom_vline (aes (xintercept=mean (shelf)), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

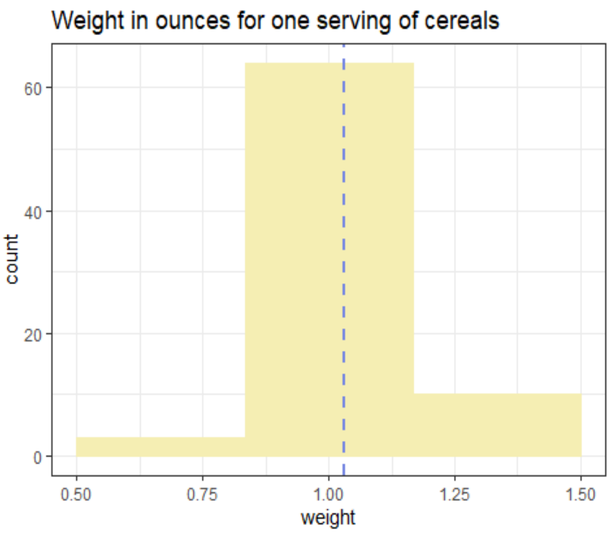


```
ggplot (cereals, aes (x=shelf)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8)
```

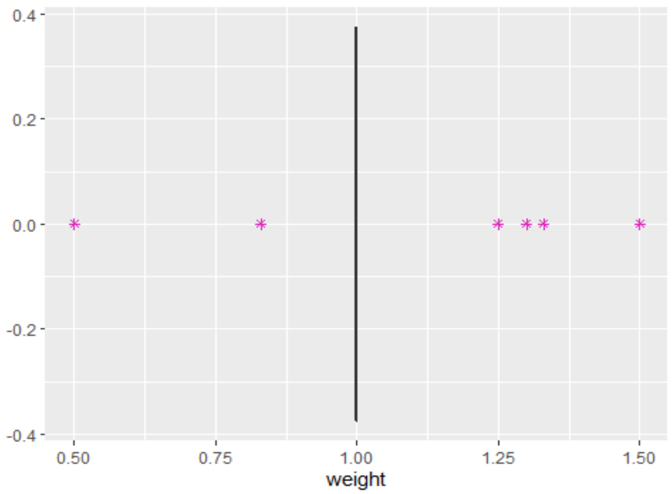


- Shelf variable represents where the cereals are on the shelves, giving them the place of 1st, 2nd or 3rd shelf, counting from the floor. All shelves are represented here, and the average shelf where the cereals are is the 2nd shelf, with the middle 50% being between first and the third shelf. The histogram could point out to negative skew, but the boxplot shows there is no skewness.

```
summary (weight)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50    1.00    1.00    1.03    1.00    1.50

sd (weight)

## [1] 0.1504768

IQR (weight)

## [1] 0

bin = diff (range (cereals$weight))/3
ggplot (cereals, aes (x=weight)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Weight in ounces for one serving of cereals') + geom_vline (aes (xintercept = mean(weight)), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

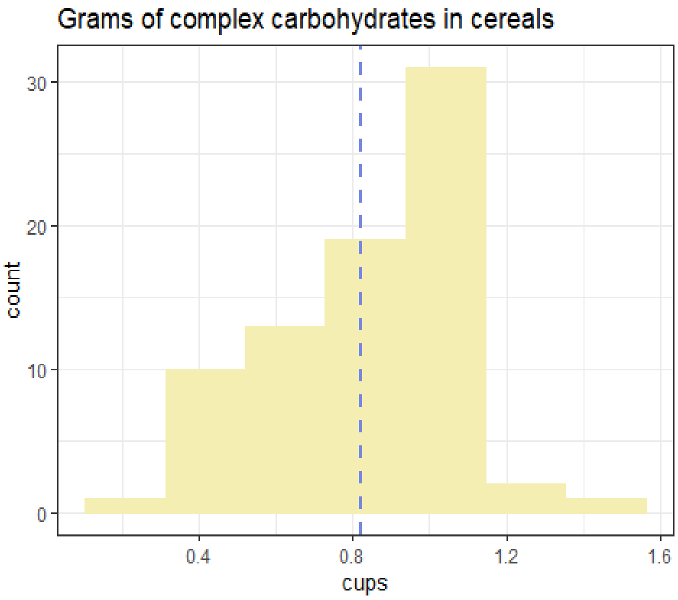


```
ggplot (cereals, aes (x=weight)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape= 8)
```

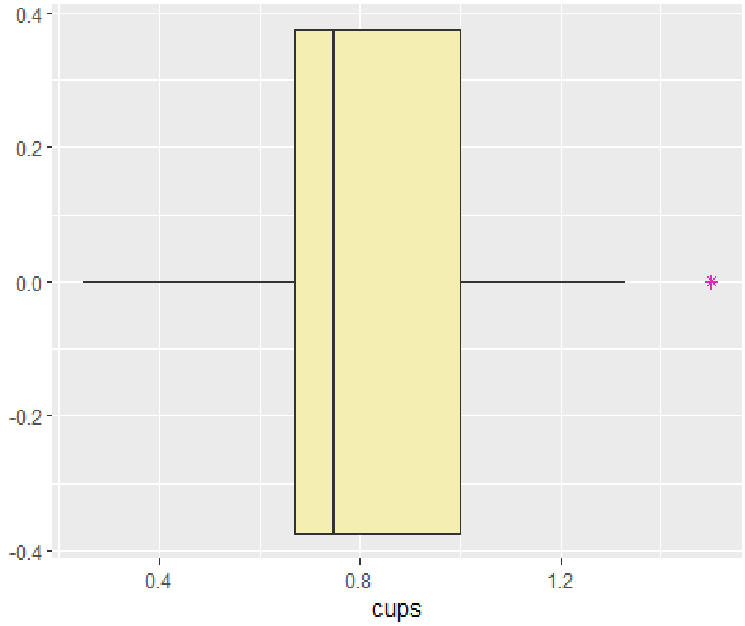


- Weight variable represents the weight in ounces of one serving. The spread is between 0.5 ounces and 1.5 ounces, with the average being 1.03 ounces. The middle 50% of the cereals have 1 ounce of weight. For me, histogram doesn't show any particular skewness but symmetrical, but the boxplot shows that there are outliers on both side, more of them being on the right side.

```
summary (cups)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.250   0.670   0.750   0.821   1.000   1.500

sd (cups)

## [1] 0.2327161

IQR (cups)

## [1] 0.33

bin = diff (range (cereals$cups))/6
ggplot (cereals, aes (x=cups)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ the
me_bw()+ ggtitle ('Grams of complex carbohydrates in cereals') + geom_vline (aes (xintercept = mean (cups)
), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```

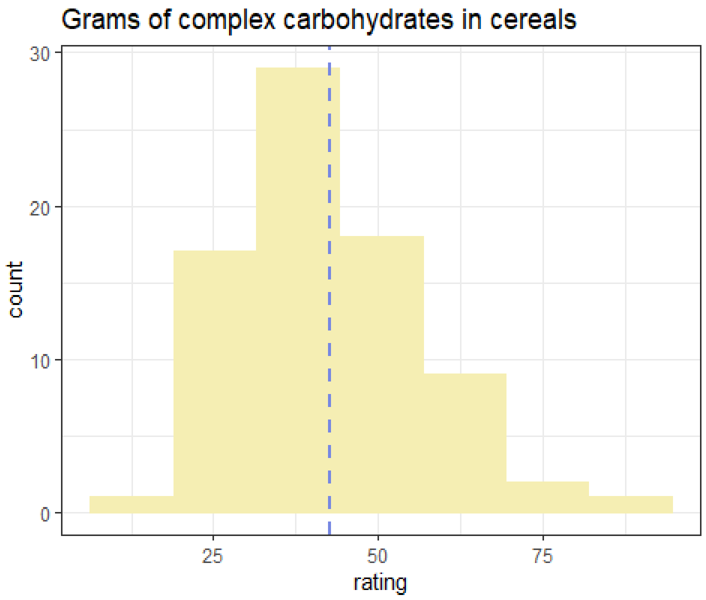


```
ggplot (cereals, aes (x=cups)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8)
```

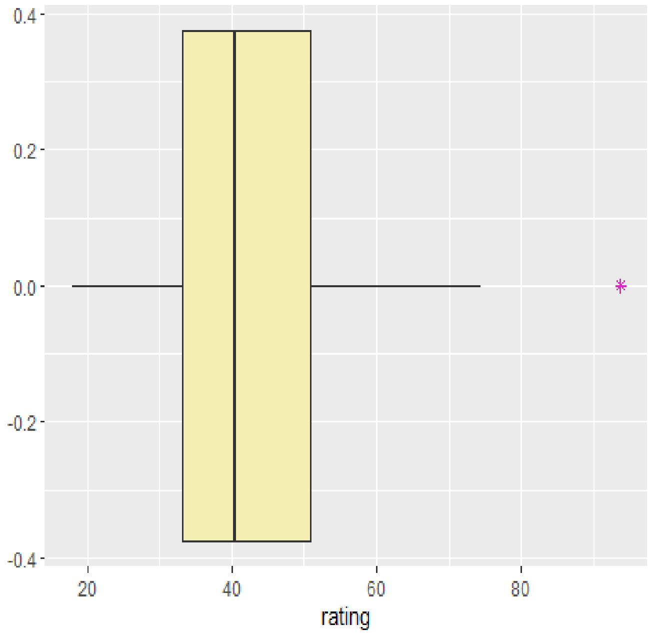


- The cups variable represents the number of cups in one serving. The spread goes between 0.25 cups and 1.5 cups, with the average value being 0.821 cups (not even a full cup). The variability between the middle 50% of the data is lower than the general spread. The histogram shows negative skewness, but boxplot doesn't when it comes to outliers. But, when it comes to the overall look of the boxplot, the distribution of this variable has a negative skewness (+ the left whisker is longer than the right one).

```
summary (rating)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.04   33.17   40.40   42.67   50.83   93.70

sd (rating)

## [1] 14.0473

IQR (rating)

## [1] 17.66

bin = diff (range (cereals$rating))/6
ggplot (cereals, aes (x=rating)) + geom_histogram (colour = "#F5EEB3", fill = "#F5EEB3", binwidth = bin)+ theme_bw()+ ggtitle ('Grams of complex carbohydrates in cereals') + geom_vline (aes (xintercept = mean(rating)), color = "#7687E2", fill = "#7687E2", linetype = "dashed", size=1)

## Warning: Ignoring unknown parameters: fill
```



```
ggplot (cereals, aes (x=rating)) + geom_boxplot (fill = "#F5EEB3", outlier.color = "#DD29BF", outlier.shape=8
)
```



- The last variable is the rating of the cereals (possible from consumer reports?). The spread is between 18.04 and 93.70 (not even 100). The average rating is 42.67, with the standard deviation of 14.04. The spread of data between the middle 50% of the data is lower than the overall spread. The histogram shows a negative skewness of the distribution, which means that there are a few cereals that have a higher rating than the rest of the cereals. The boxplot confirms the negative skewness, with one outlier on the high side.