R Workshop - Practice Exercises - 1.2

Setup

```
library(tidyverse)
```

The Data

```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   mfr = col_character(),
##   type = col_character(),
##   calories = col_double(),
##   protein = col_double(),
##   fat = col_double(),
##   sodium = col_double(),
##   fiber = col_double(),
##   carbo = col_double(),
##   sugars = col_double(),
##   potass = col_double(),
##   vitamins = col_double(),
##   shelf = col_double(),
##   weight = col_double(),
##   cups = col_double(),
##   rating = col_double()
## )
```

```
cereals <- read_csv("/data/cereals.csv")
```

Documentation for dataset: https://www.kaggle.com/crawford/80-cereals/version/2

```
#### Edit this code to examine your dataset
str(cereals)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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: num  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein : num  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat     : num  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium  : num  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  : num  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass  : num  280 135 320 330 -1 70 30 100 125 190 ...
##  $ vitamins: num  25 0 25 25 25 25 25 25 25 25 ...
##  $ shelf   : num  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 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   name = col_character(),
##   ..   mfr = col_character(),
##   ..   type = col_character(),
##   ..   calories = col_double(),
##   ..   protein = col_double(),
##   ..   fat = col_double(),
##   ..   sodium = col_double(),
##   ..   fiber = col_double(),
##   ..   carbo = col_double(),
##   ..   sugars = col_double(),
##   ..   potass = col_double(),
##   ..   vitamins = col_double(),
##   ..   shelf = col_double(),
##   ..   weight = col_double(),
##   ..   cups = col_double(),
##   ..   rating = col_double()
##   .. )
```

```
summary(cereals)
```

```
##      name               mfr                type              calories
##  Length:77          Length:77          Length:77          Min.   : 50.0
##  Class :character   Class :character   Class :character   1st Qu.:100.0
##  Mode  :character   Mode  :character   Mode  :character   Median :110.0
##                                                           Mean   :106.9
##                                                           3rd Qu.:110.0
##                                                           Max.   :160.0
##     protein            fat            sodium          fiber
##  Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000
##  Median :3.000   Median :1.000   Median :180.0   Median : 2.000
##  Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000
##      carbo           sugars           potass          vitamins
##  Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00
##  Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00
##  Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25
##  3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
##  Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00
##      shelf           weight          cups           rating
##  Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
##  Median :2.000   Median :1.00   Median :0.750   Median :40.40
##  Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67
##  3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83
##  Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70
```

[Briefly summarize the dataset here.]

[CHECKPOINT: Knit your Markdown file!]

Data Cleaning and Transformation

```
cereals %>%
  summarize_at(vars(calories, sugars),
               funs(mean, median))
```

```
## # A tibble: 1 x 4
##   calories_mean sugars_mean calories_median sugars_median
##           <dbl>       <dbl>           <dbl>         <dbl>
## 1          107.        6.92             110             7
```

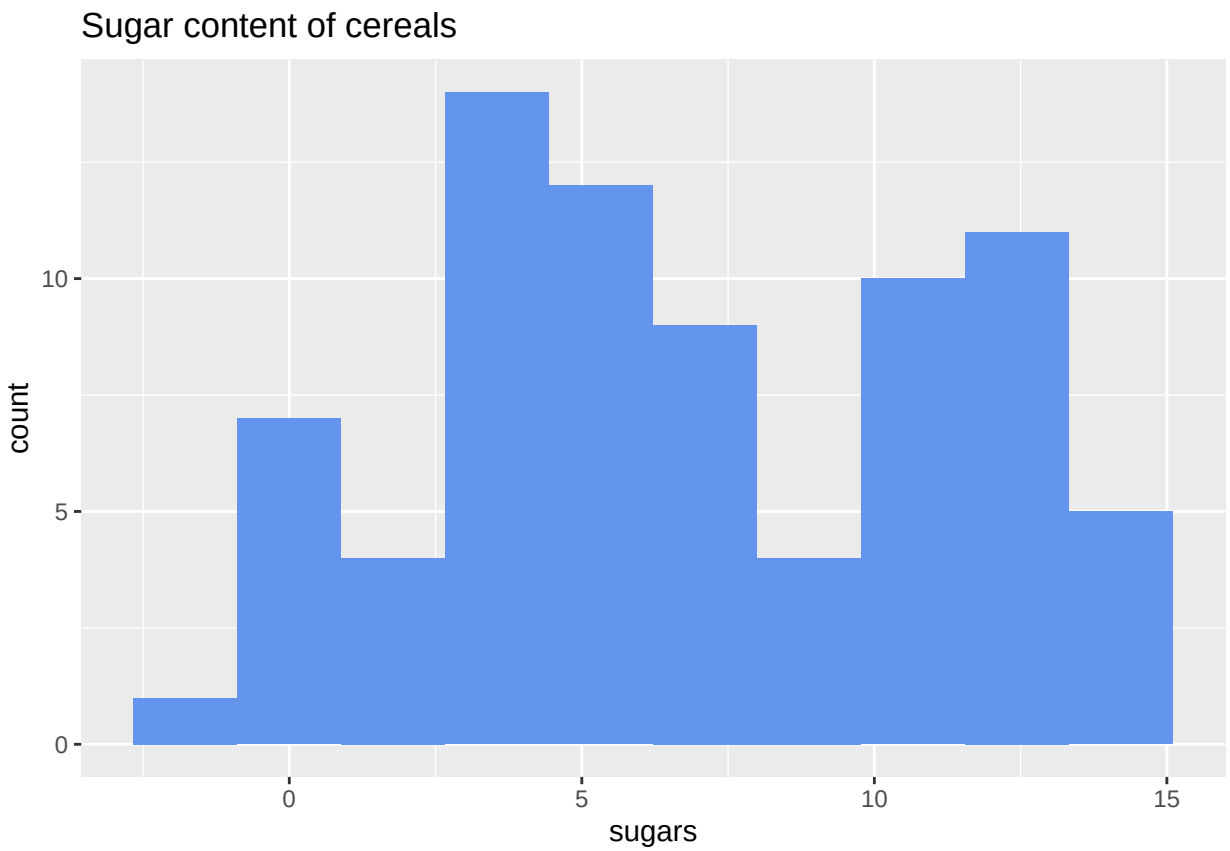
```
cereals_2 <- cereals %>%
  mutate(
    cal_per_cup = calories/cups
  ) %>%
  select(name, mfr, cal_per_cup) %>%
  filter(mfr == "Kellogg") %>%
  arrange(cal_per_cup)
```

[CHECKPOINT: Knit your document!]

Visualizing the data

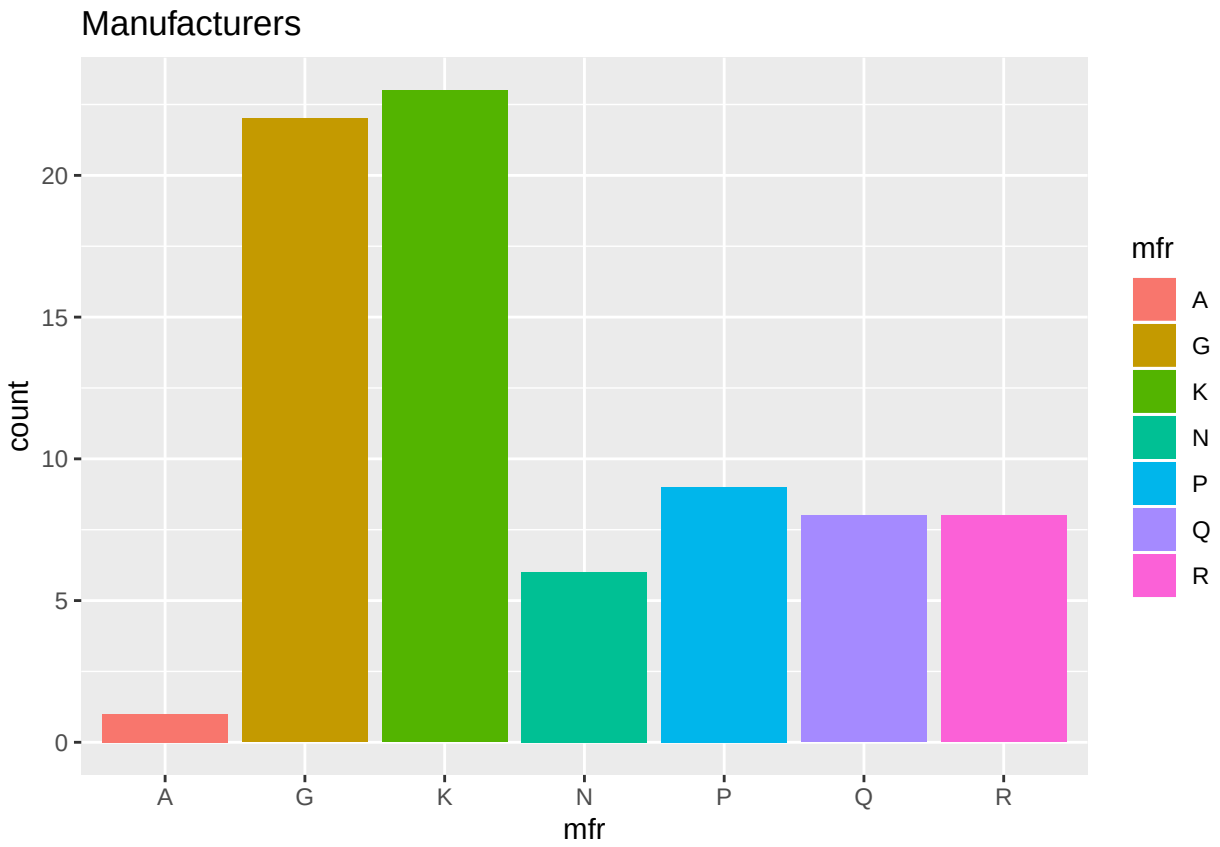
Histogram

```
ggplot(cereals, aes(x = sugars)) +
  geom_histogram(bins = 10, fill = "cornflowerblue") +
  ggtitle("Sugar content of cereals")
```



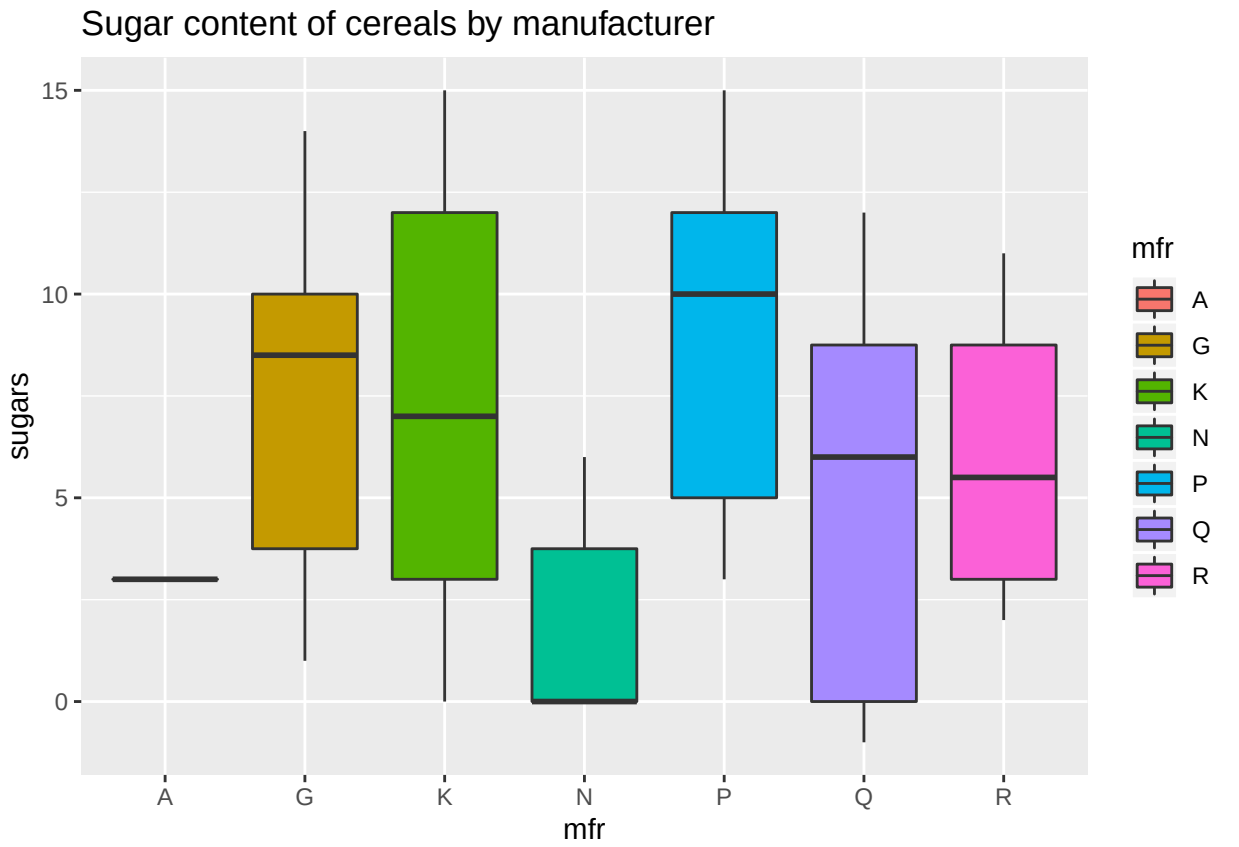
Bar Plot

```
ggplot(cereals, aes(x = mfr, fill = mfr)) +
  geom_bar() +
  ggtitle("Manufacturers")
```



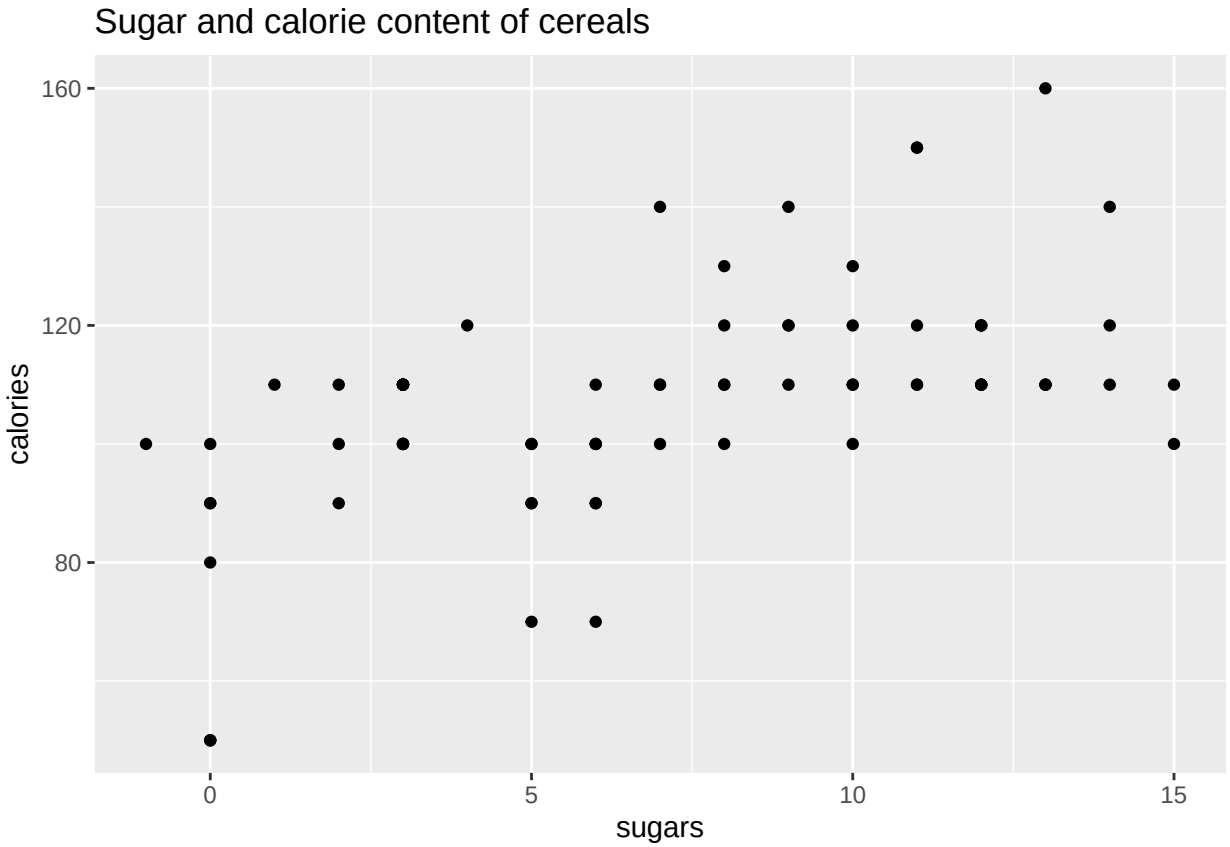
Box Plot

```
ggplot(cereals, aes(x = mfr, y = sugars, fill = mfr)) +
  geom_boxplot() +
  ggtitle("Sugar content of cereals by manufacturer")
```



Scatter Plot

```
ggplot(cereals, aes(x = sugars, y = calories)) +
  geom_point() +
  ggtitle("Sugar and calorie content of cereals")
```



Conclusion

What did you learn about cereals? Write a few sentences summarizing your findings, knit your document, and admire your handiwork!