R4 Exercises: Graphing data relations

Karol Topór

2023-12-14

Table of contents

Packages used in this notebook:

```
library(tidyverse)
library(mosaicData)
```

1 Visualise `RailTrail` data

Use the `mosaicData::RailTrail` data set from the `{mosaicData}` package. See for a description of the data. According to Cambridge.org is a rail-trail a path for walking or bicycle riding, created from a railway that is no longer used by trains.

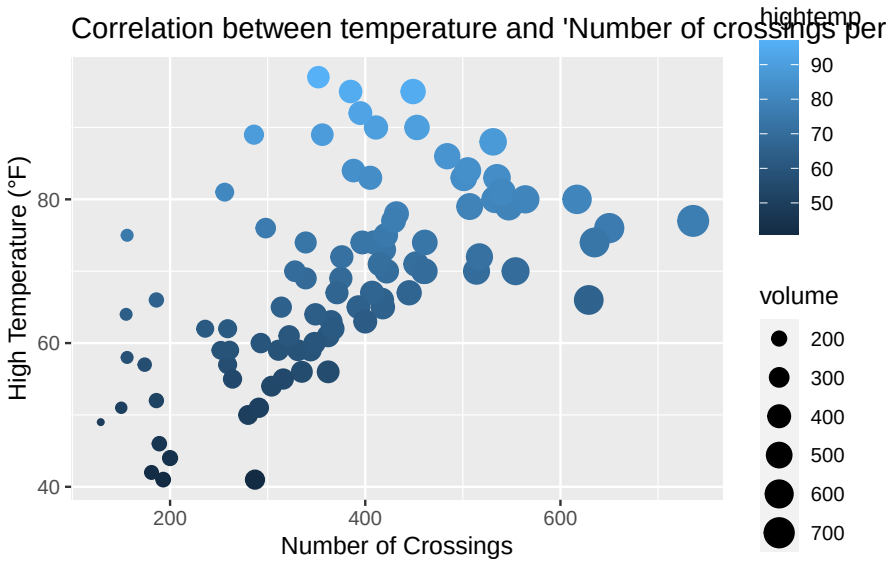
```
railtrail <- as.tibble(RailTrail)
glimpse(railtrail)
```

```
Rows: 90
Columns: 11
$ hightemp   <int> 83, 73, 74, 95, 44, 69, 66, 66, 80, 79, 78, 65, 41, 59, 50,~
$ lowtemp    <int> 50, 49, 52, 61, 52, 54, 39, 38, 55, 45, 55, 48, 49, 35, 35,~
$ avgtemp    <dbl> 66.5, 61.0, 63.0, 78.0, 48.0, 61.5, 52.5, 52.0, 67.5, 62.0,~
$ spring     <int> 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0,~
$ summer     <int> 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
$ fall       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1,~
$ cloudcover <dbl> 7.6, 6.3, 7.5, 2.6, 10.0, 6.6, 2.4, 0.0, 3.8, 4.1, 8.5, 7.2~
$ precip     <dbl> 0.00, 0.29, 0.32, 0.00, 0.14, 0.02, 0.00, 0.00, 0.00, 0.00,~
$ volume     <int> 501, 419, 397, 385, 200, 375, 417, 629, 533, 547, 432, 418,~
$ weekday    <lgl> TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, FALSE, FALSE, TR~
$ dayType    <chr> "weekday", "weekday", "weekday", "weekend", "weekday", "wee~
```

1.1 Scatterplot (2P)

Create a scatterplot of the number of crossings per day, `volume`, against the high temperature, `hightemp`, of that day. Use a header and legends appropriately and interpret the resulting graph.

```
p <-  ggplot(railtrail, aes(x = volume, y = hightemp)) +
      geom_point(aes(col = hightemp, size = volume)) +
      ggtitle("Correlation between temperature and 'Number of crossings per day'") +
      xlab("Number of Crossings") +
      ylab("High Temperature (°F)")
p
```

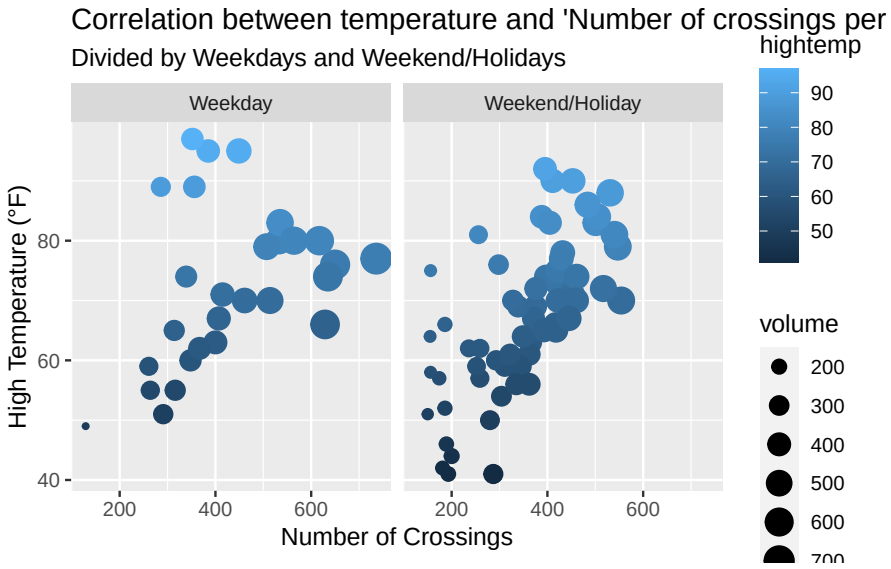


A: The graph shows a correlation between temperature and 'Number of crossing per day'. Overall there are more 'Number of crossings per day' when the temperature is above 60F°.

1.2 Separate scatterplots during week and weekend (2P)

Separate the above scatter plot into facets by weekday. Use a header and legends appropriately and interpret the resulting graph.

```
p <- p +  facet_wrap(~weekday,
                     labeller = labeller(
                       weekday =
                         c(`TRUE` = "Weekend/Holiday", `FALSE` = "Weekday"))) +
          labs(subtitle = "Divided by Weekdays and Weekend/Holidays")
p
```



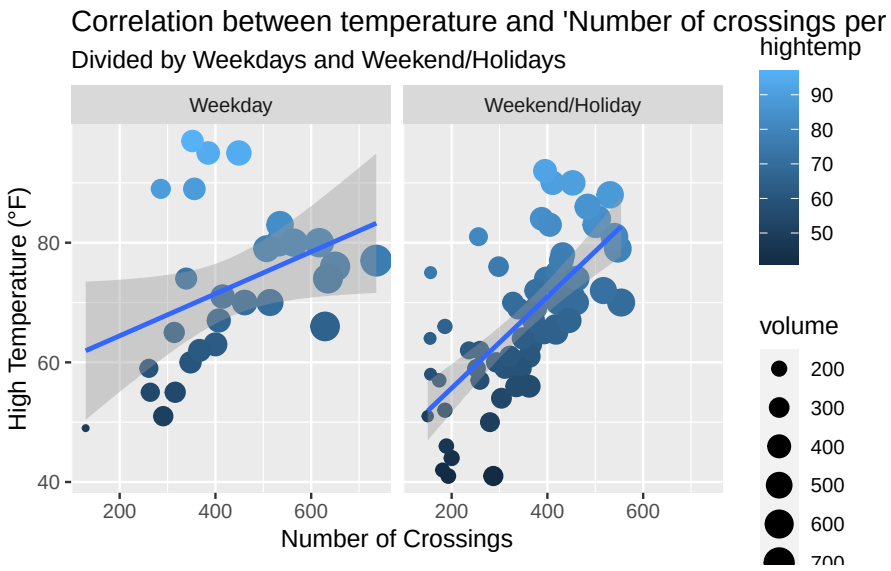
A: Overall there are more occurences on Weekends/Holidays. The graph also shows that there are more 'Number of crossings per day' even in temperatures beneath 60F°.

1.3 Add regression lines (2P)

Add regression lines to the two facets. Show the results for linear and nonlinear regression lines. Use a header and legends appropriately and interpret the resulting graphs.

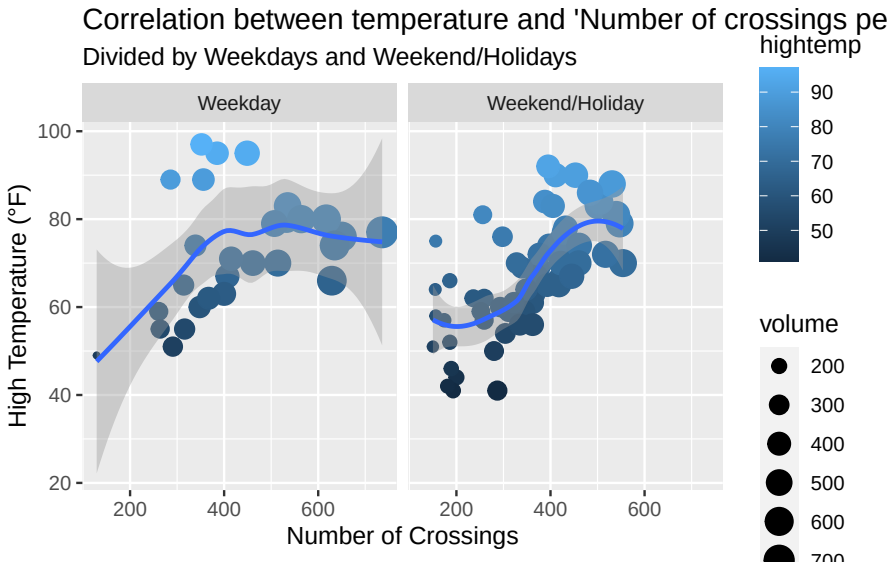
1.3.1 Linear Regression Line

```
p +  geom_smooth(method = "lm")
```



1.3.2 Nonlinear Regression Line

```
p +  geom_smooth(method = "loess")
```



A: The regression lines show that there is indeed a preference to go for a walk/bicycling when the temperature is higher.

2 Visualise `mtcars` data

Use the `mtcars` data. A description of the data is available.

First transform the variables in the `mtcars` data into European measures and then graph the data as requested in the following.

2.1 Transform `mtcars` into European measures (3P)

For the transformation into European measures consider that

- `1 mile = 1.609 km`
- `1 gallon = 3.785 liter`
- `1 liter = 61.0237 cu.in`
- `1 kg = 2.20462 lbs`

Create the following variables:

- transform `mpg` into `verbrauch` as measure for the fuel consumption in `liter` per `100 km`.
- rename `cyl` into `zylinder` as number of cylinders.
- transform `disp` into `hubraum` as the size of the engine measured in liters.
- transform `qseq` into `beschleunigung` to measure the seconds it takes to accelerate to `100 km/h`. `qseq` are the seconds it takes to travel `1/4` mile. For the transformation assume a **constant acceleration** (**constant increase of the speed**) until the car reaches `1/4` mile. The speed starts at zero and the final speed at `1/4` mile is twice the average speed (as measured by traveling a `1/4` mile in `qsec`).
- rename `drat` into `drehmoment`
- transform `wt` (1000 lbs) into `gewicht` (1000 kg).
- transform `am` into `schaltung` as `Automatik` for `am = 0` and `Manuell` for `am = 1`
- transform `vs` into `motor` as `V-Motor` for `vs = 0` and `Reihenmotor` for `vs = 1`

```
mtcars$car_names <- rownames(mtcars)
mtcars$verbrauch <- (100 / (mtcars$mpg * 1.609)) * 3.785
mtcars$zylinder <- mtcars$cyl
mtcars$hubraum <- mtcars$disp / 61.0237

distance_km <- 0.25 * 1.609
mtcars$beschleunigung <- (100 / ((2 * distance_km) / (mtcars$qsec / 3600)))

mtcars$drehmoment <- mtcars$drat
mtcars$gewicht <- mtcars$wt / 2.20462
```

```r
mtcars$schaltung <- ifelse(mtcars$am == 0, "Automatik", "Manuell")
mtcars$motor <- ifelse(mtcars$vs == 0, "V-Motor", "Reihenmotor")
glimpse(mtcars)
```

```
Rows: 32
Columns: 20
$ mpg            <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 1~
$ cyl            <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4~
$ disp           <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7,~
$ hp             <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180~
$ drat           <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3~
$ wt             <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190,~
$ qsec           <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00,~
$ vs             <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1~
$ am             <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1~
$ gear           <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4~
$ carb           <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2~
$ car_names      <chr> "Mazda RX4", "Mazda RX4 Wag", "Datsun 710", "Hornet 4 D~
$ verbrauch      <dbl> 11.201870, 11.201870, 10.317512, 10.992490, 12.579641, ~
$ zylinder       <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4~
$ hubraum        <dbl> 2.621932, 2.621932, 1.769804, 4.227866, 5.899347, 3.687~
$ beschleunigung <dbl> 0.5683309, 0.5876666, 0.6425661, 0.6712244, 0.5876666, ~
$ drehmoment     <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3~
$ gewicht        <dbl> 1.1884134, 1.3040796, 1.0523355, 1.4583012, 1.5603596, ~
$ schaltung      <chr> "Manuell", "Manuell", "Manuell", "Automatik", "Automati~
$ motor          <chr> "V-Motor", "V-Motor", "Reihenmotor", "Reihenmotor", "V-~
```

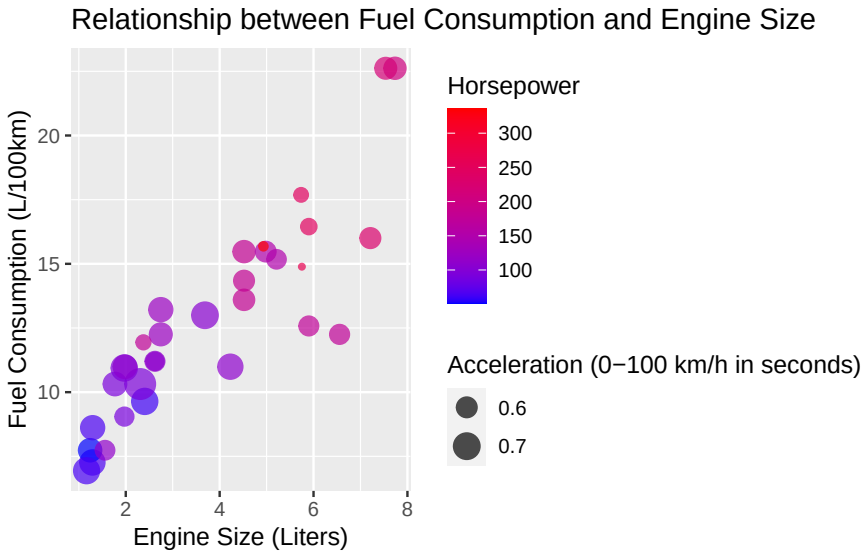
2.2 Graph `verbrauch` as determined by `hubraum` (2P)

Explain the variable **verbrauch** through **hubraum** in a scatter plot. Include **hp** and **beschleunigung** as determinants of the variable **verbrauch** through the color and size of the data points in the scatter plot, respectively. The resulting graph should look like this:

```r
p2 <- ggplot(mtcars, aes(x = hubraum, y = verbrauch, color = hp, size = beschleunigung)) +
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(
    title = "Relationship between Fuel Consumption and Engine Size",
    x = "Engine Size (Liters)",
    y = "Fuel Consumption (L/100km)",
```

```
    color = "Horsepower",
    size = "Acceleration (0-100 km/h in seconds)"
  )


p2
```




```
ggsave("Exercise_cars_verbrauch_hubraum.png")
```

2.3 Interpret the graph (2P)

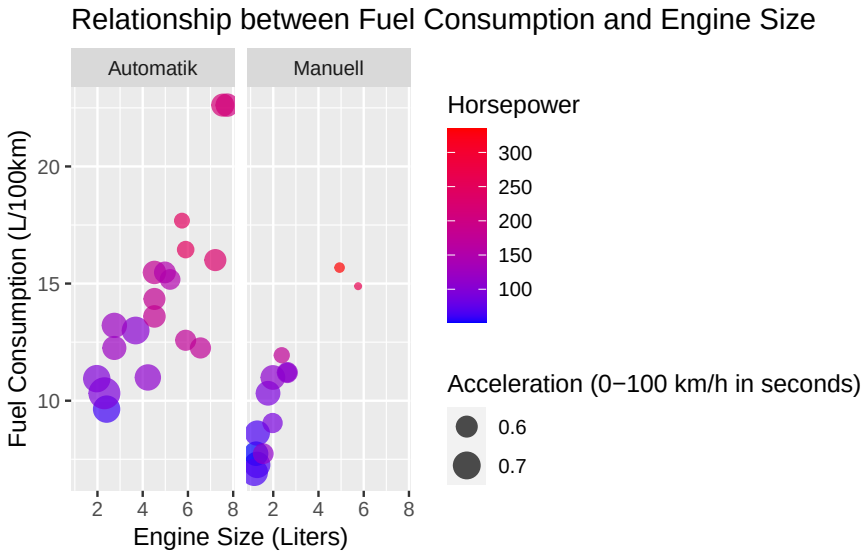
What does the graph reveal about the relation between the different variables in the data set?

```
The graph shows:
A positive correlation between engine size and Fuel consumation.
Cars with bigger engines use more fuel per 100km.
Also, bigger engines tend to have more horsepower but this
does not necessary mean a faster acceleration.
```

2.4 Automatic versus Manual (2P)

Show the above graph for cars with automatic and manual gear shifting and interpret the resulting graph.

```
p2 + facet_wrap(~schaltung)
```




```
#g
ggsave("Exercise_cars_facet_am.png")
```

```
Interpretation of the graph:
Interestingly the face wrap reveals that cars with a manual gear switching mechanism
use less fuel then automatic gear shifting and also tent to have bigger engines.
Which explains the higher fuel consumtiom as we have seen before
--> bigger engin --> more fuel consumtion
```

2.5 Show cars with very high fuel consumption (1P)

Show the names of the cars (together with `verbrauch`, `hubraum`, `beschleunigung`, `hp` and `gewicht`) that have more than 20 liters of fuel consumption per 100 km (in the graph at the top right corner) or an engine size of more than 7 liters.

```r
mtcars[
  (mtcars$verbrauch > 20 | mtcars$hubraum > 7),
  c("verbrauch", "hubraum", "beschleunigung", "hp", "gewicht")
]
```

```
                   verbrauch  hubraum beschleunigung  hp  gewicht
Cadillac Fleetwood   22.61916 7.734700      0.6208135 205 2.381363
Lincoln Continental  22.61916 7.538055      0.6152890 215 2.460288
Chrysler Imperial    16.00267 7.210313      0.6014778 230 2.424454
```

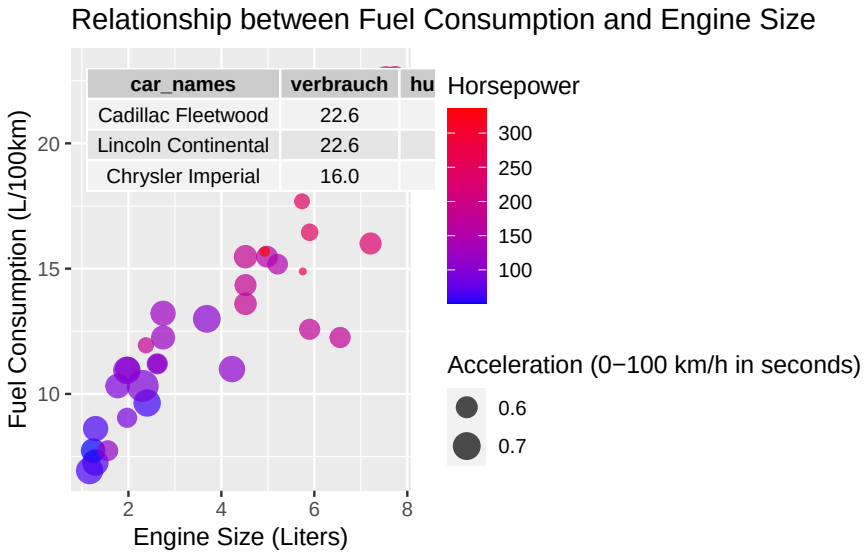
2.6 Show cars with very high acceleration (1P)

Show the names of the cars (together with `verbrauch`, `hubraum`, `beschleunigung`, `hp` and `gewicht`) that have a `beschleunigung` of less than 7.5 seconds.

```r
head(mtcars[
  (mtcars$beschleunigung < 7.5),
  c("verbrauch", "hubraum", "beschleunigung", "hp", "gewicht")
])
```

```
                 verbrauch  hubraum beschleunigung  hp  gewicht
Mazda RX4          11.20187 2.621932      0.5683309 110 1.188413
Mazda RX4 Wag      11.20187 2.621932      0.5876666 110 1.304080
Datsun 710         10.31751 1.769804      0.6425661  93 1.052336
Hornet 4 Drive     10.99249 4.227866      0.6712244 110 1.458301
Hornet Sportabout  12.57964 5.899347      0.5876666 175 1.560360
Valiant            12.99665 3.687092      0.6981562 105 1.569431
```

```r
## Insert a table into a graph
library(ggpmisc)
df <- mtcars %>% filter(hubraum >= 7 | verbrauch >= 20 ) %>%
  select(c('car_names','verbrauch','hubraum')) %>%
  mutate(across(2:3, round, 1))
p2 + geom_table(aes(x=1.1, y=23, label=list(df)))
```



3 Visualise dietary data

Use the code `data("UScereal", package = "MASS")` for the `UScereal` data from the `MASS` package. See https://www.rdocumentation.org/packages/MASS/versions/7.3-53/topics/UScereal for details. Adjust the Manufacturer in `mfr` (represented by its first initial): G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina and the display shelf in `shelf` (1, 2, or 3, counting from the floor) into `bottom-shelf`, `middle-shelf` and `top-shelf`.

```
#library(MASS)  # überschreibt select() in dplyr package ---> do not use library(MASS)
data("UScereal", package = "MASS") # überschreibt select NICHT
UScereal <- UScereal %>%
  mutate(
    mfr = case_when(
      mfr == "G" ~ "General Mills",
      mfr == "K" ~ "Kelloggs",
      mfr == "N" ~ "Nabisco",
      mfr == "P" ~ "Post",
      mfr == "Q" ~ "Quaker Oats",
      mfr == "R" ~ "Ralston Purina",
      TRUE ~ mfr
    ),
    shelf = case_when(
      shelf == 1 ~ "bottom-shelf",
```

```
      shelf == 2 ~ "middle-shelf",
      shelf == 3 ~ "top-shelf",
      TRUE ~ as.character(shelf)
    )
  )
head(UScereal)
```

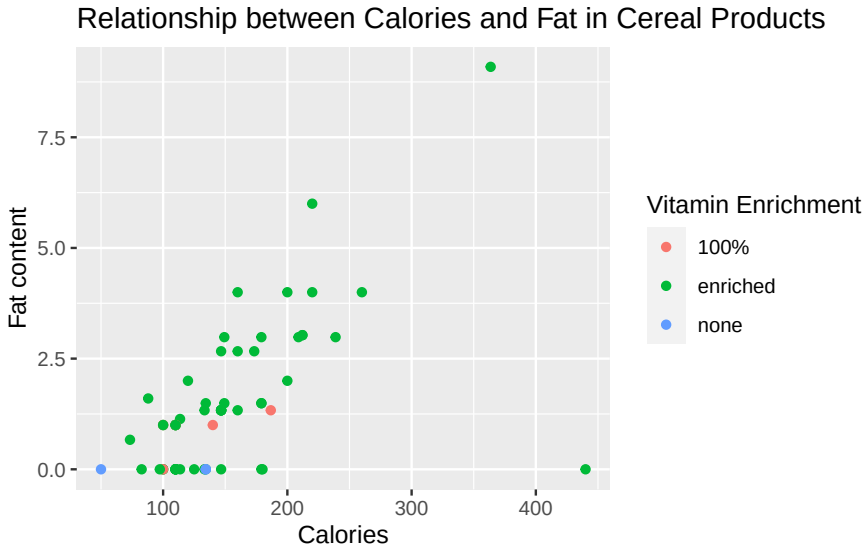
```
                                mfr calories   protein      fat   sodium
100% Bran                    Nabisco 212.1212 12.121212 3.030303 393.9394
All-Bran                    Kelloggs 212.1212 12.121212 3.030303 787.8788
All-Bran with Extra Fiber   Kelloggs 100.0000  8.000000 0.000000 280.0000
Apple Cinnamon Cheerios  General Mills 146.6667  2.666667 2.666667 240.0000
Apple Jacks                 Kelloggs 110.0000  2.000000 0.000000 125.0000
Basic 4                  General Mills 173.3333  4.000000 2.666667 280.0000
                             fibre    carbo   sugars        shelf potassium
100% Bran                30.303030 15.15152 18.18182    top-shelf 848.48485
All-Bran                 27.272727 21.21212 15.15151    top-shelf 969.69697
All-Bran with Extra Fiber 28.000000 16.00000  0.00000    top-shelf 660.00000
Apple Cinnamon Cheerios   2.000000 14.00000 13.33333 bottom-shelf  93.33333
Apple Jacks               1.000000 11.00000 14.00000 middle-shelf  30.00000
Basic 4                   2.666667 24.00000 10.66667    top-shelf 133.33333
                         vitamins
100% Bran                enriched
All-Bran                 enriched
All-Bran with Extra Fiber enriched
Apple Cinnamon Cheerios  enriched
Apple Jacks              enriched
Basic 4                  enriched
```

3.1 Graph the relationship between calories and fat (2P)

Visualize the relationship of calories with fat. Additionally, highlight whether the product has been enriched with vitamins.

```
p3 <- ggplot(UScereal, aes(x = calories, y = fat, color = vitamins)) +
  geom_point() +
  labs(
    title = "Relationship between Calories and Fat in Cereal Products",
    x = "Calories",
    y = "Fat content",
    color = "Vitamin Enrichment"
```

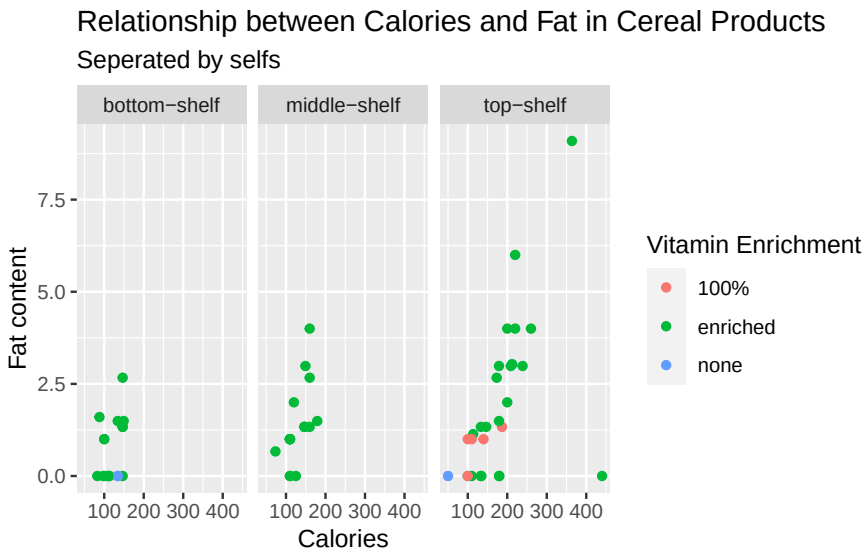
```
  )
p3
```



3.2 Graph the information for different shelves (2P)

As an extension to the previous plot, create plots differentiating between shelves.

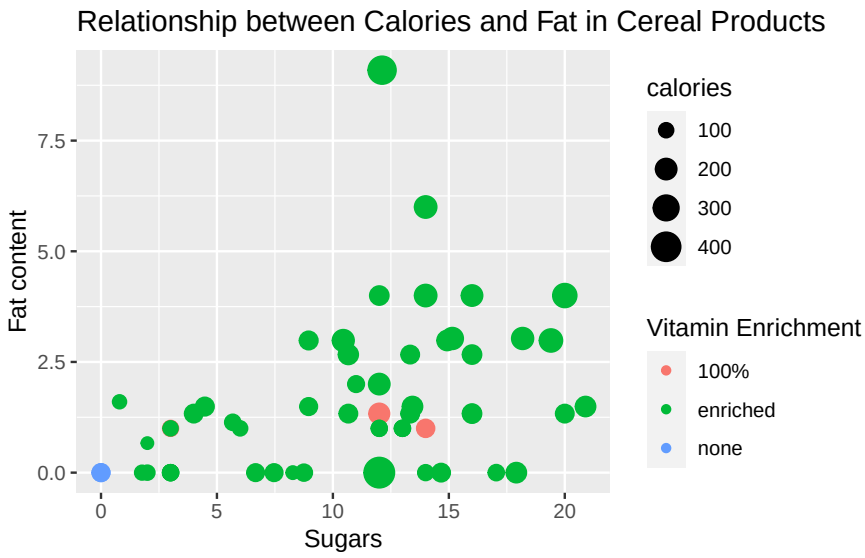
```
p3 + facet_wrap(~shelf) +
     labs(subtitle = "Seperated by selfs")
```



3.3 Graph the relationship between sugar and fat (2P)

Visualize the relationship of sugar and fat. Additionally, highlight whether the product has been enriched with vitamins. Also show the calories.

```
p4 <- ggplot(UScereal, aes(x = sugars, y = fat, color = vitamins, size = calories)) +
  geom_point() +
  labs(
    title = "Relationship between Calories and Fat in Cereal Products",
    x = "Sugars",
    y = "Fat content",
    color = "Vitamin Enrichment"
  )
p4
```



3.4 Graph the information separately for each manufacturer (2P)

As a first extension to the previous plot, show the information separately for each manufacturer, using facets.

```
p4 + facet_wrap(~mfr)
```