R Basics

Blake Pappas

2023-12-17

Data cleaning and visualization in R

We will use the "coffee.csv" dataset to practice.

Load the dplyr package:

```
# install.packages("dplyr")

library(dplyr)
```

Import the "coffee.csv" dataset:

```
coffee = read.csv("coffee.csv", sep = ";")
```

P1: Arrange the data

What is the ProductID of the most profitable coffee?

Answer: The ProductID of the most profitable coffee is 2.

Write the code below:

```
most_profitable = coffee %>% arrange(desc(Profit))

head(most_profitable)
```

```
##   Profit Margin Sales COGS Total.Expenses Marketing Inventory Budget.Profit
## 1    778    613   659   52             46        17     -1493           560
## 2    777    612   658   52             46        17     -2033           560
```

```
## 3    755    595   643   54             45        17     -1006           530
## 4    690    516   614   60             51        19     -2572           460
## 5    646    526   815  239             91        66      1197           450
## 6    599    487   546   64             49        21      -663           430
##   Budget.COGS Budget.Margin Budget.Sales Area.Code ProductId DateTableau
## 1          40           590          630       978         2  07/01/2011
## 2          40           590          630       617         2  08/01/2011
## 3          50           560          610       774         2  06/01/2011
## 4          50           490          540       774         2  09/01/2011
## 5         210           510          720       212         7  10/01/2011
## 6          60           460          520       351         2  05/01/2011
```

P2: Subset the data

Find the subset of data with "Sales" larger than 200.

Write the code below:

```
sales200 = coffee %>% filter(Sales > 200)

head(sales200)
```

```
##   Profit Margin Sales COGS Total.Expenses Marketing Inventory Budget.Profit
## 1     94    130   219   89             36        24       777           100
## 2    101    139   234   95             38        26       821           110
## 3     99    171   341  170             72        47      1091           110
## 4    111    201   345  144             90        47       862           130
## 5     87    139   234   95             52        30       608           100
## 6    203    312   546  234            109        77      1310           260
##   Budget.COGS Budget.Margin Budget.Sales Area.Code ProductId DateTableau
## 1          90           130          220       719         1  01/01/2010
## 2         100           140          240       970         3  01/01/2010
## 3         140           160          300       970         8  01/01/2010
## 4         150           210          360       217         2  01/01/2010
## 5         100           140          240       309         3  01/01/2010
## 6         270           370          640       309         5  01/01/2010
```

P3: Group and summarize the data

Find out the total profit and average inventory level of each product (identified by a unique "ProductID").

```
data_group = coffee %>% group_by(ProductId) %>%
  summarise(total_profit = sum(Profit),
            average_inventory_level = mean(Inventory))
```

```
data_group
```

```
## # A tibble: 13 x 3
##    ProductId total_profit average_inventory_level
##        <int>        <int>                   <dbl>
##  1         1         4890                    741.
##  2         2        55804                    708.
##  3         3        13989                    839.
##  4         4        11375                    256.
##  5         5        17678                    756.
##  6         6        29502                    755.
##  7         7        10065                    880.
##  8         8        27231                    713.
##  9         9        29869                    719.
## 10        10         6154                   1096.
## 11        11        29053                    738.
## 12        12        24164                    757.
## 13        13         -231                    900.
```

P4: Create a new variable, "ProfitRatio", which equals "Profit"/"Sales"

```
data_ProfitRatio = coffee %>% mutate(ProfitRatio = Profit / Sales)

head(data_ProfitRatio)
```

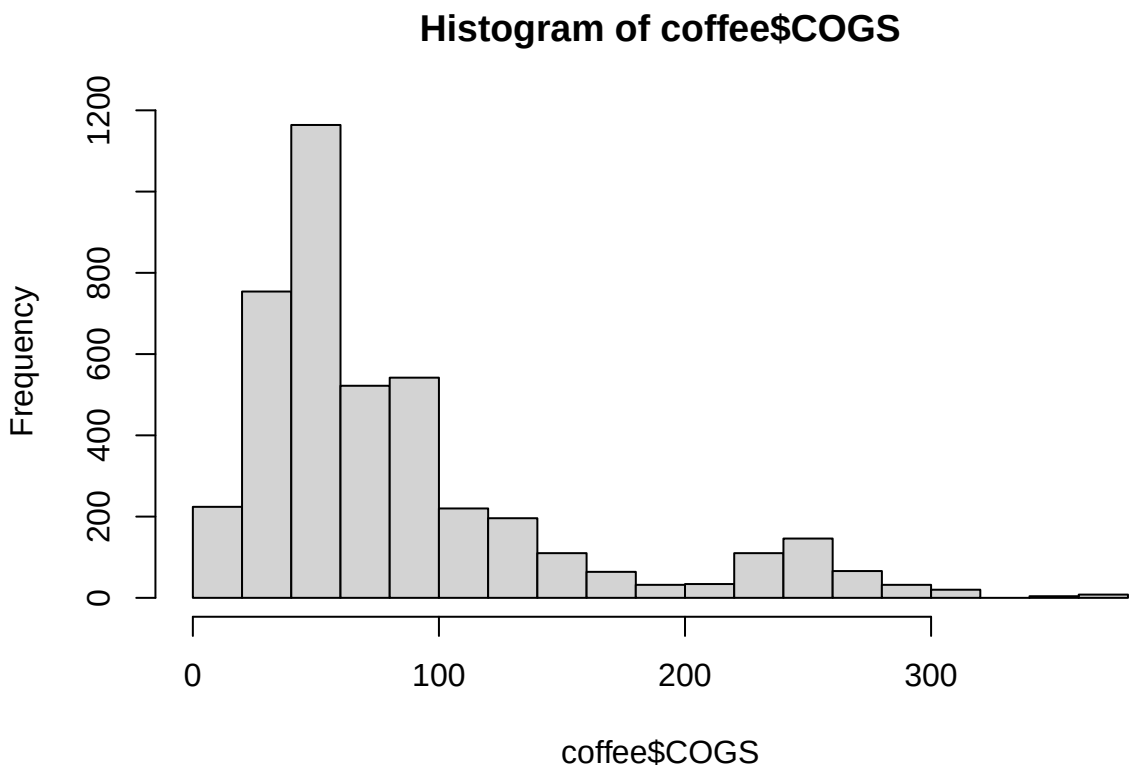
```
##   Profit Margin Sales COGS Total.Expenses Marketing Inventory Budget.Profit
## 1     94    130   219   89             36        24       777            100
## 2     68    107   190   83             39        27       623             80
## 3    101    139   234   95             38        26       821            110
## 4     30     56   100   44             26        14       623             30
## 5     54     80   134   54             26        15       456             70
## 6     53    108   180   72             55        23       558             80
##   Budget.COGS Budget.Margin Budget.Sales Area.Code ProductId DateTableau
## 1          90           130          220       719         1  01/01/2010
## 2          80           110          190       970         2  01/01/2010
## 3         100           140          240       970         3  01/01/2010
## 4          30            50           80       303        13  01/01/2010
## 5          60            90          150       303         5  01/01/2010
## 6          80           130          210       720         6  01/01/2010
##   ProfitRatio
## 1   0.4292237
## 2   0.3578947
## 3   0.4316239
## 4   0.3000000
## 5   0.4029851
## 6   0.2944444
```

P5: Make a histogram

Plot the distribution of "COGS".

Write the code below:

```
hist(coffee$COGS)
```

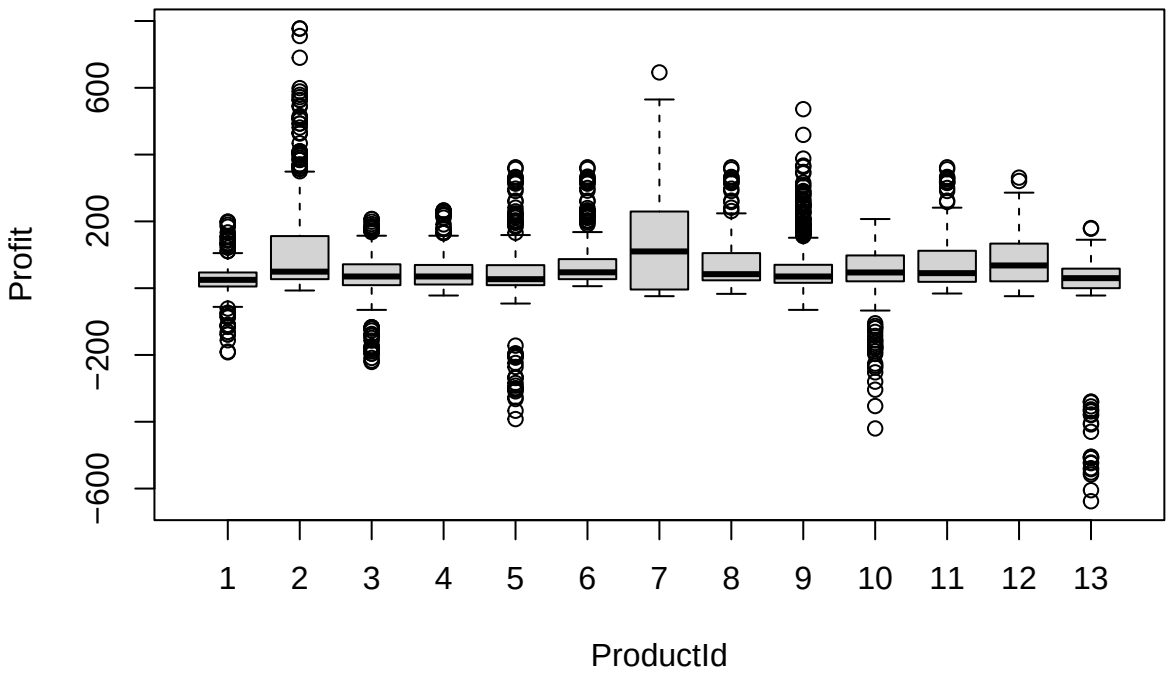


P6: Make a boxplot

Plot the distribution of "Profit" across different "ProductIDs".

Write the code below:

```
boxplot(Profit ~ ProductId, data = coffee)
```

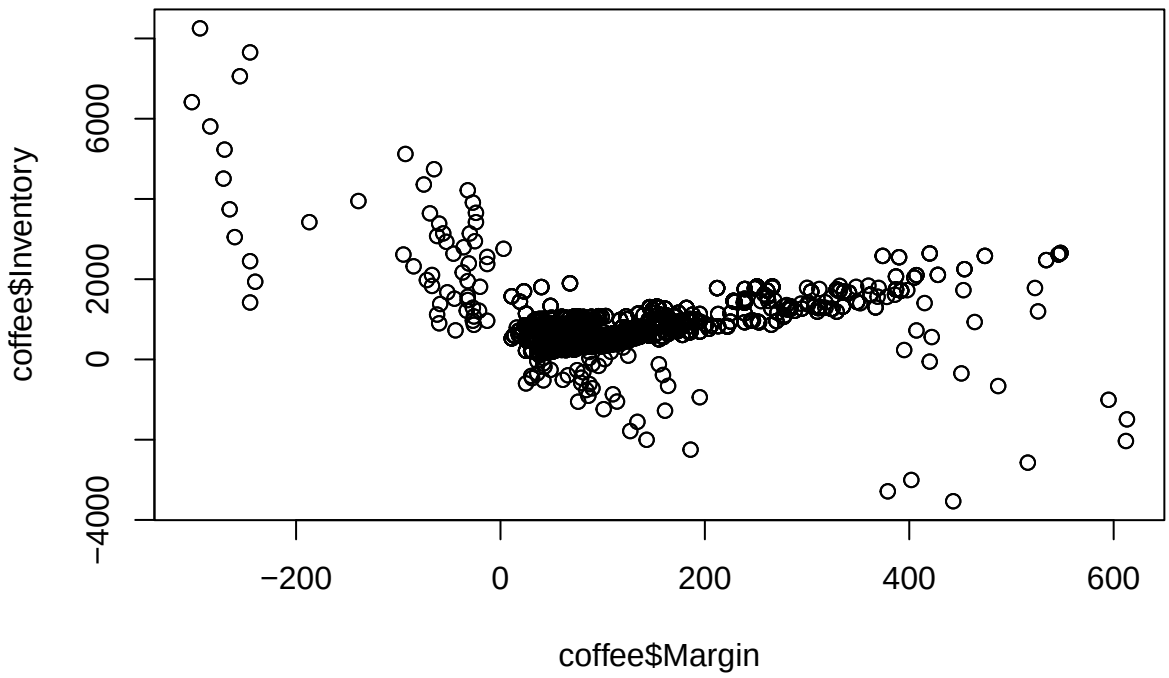


P7: Make a scatterplot

Plot the relationship between "Margin" and "Inventory".

Write the code below:

```
plot(coffee$Margin, coffee$Inventory)
```


coffee$Inventory
6000
2000
0
−4000
−200
0
200
400
600
coffee$Margin