S20 STA 100 A05 Discussion 02

Yishan Huang

2020/04/07

Discussion Time: Tuesday 12:10 – 1:00 pm.

Email Address: yishuang@ucdavis.edu

Zoom: https://ucdstats.zoom.us/j/516752243?pwd=WnhZY3E2SFNxZWhiRHRPcWFrV2hZZz09

Office Hour: Thursday 12:00 – 1:00 pm.

Main points in this notes

- Install packages
- Example of dataset, build-in dataset in R
- Dataframe
- Choose rows and columns from a dataframe, subsampling
- Histograms
- Put multiple plots in a single figure
- Add a histogram on an existing plot
- Set range of x and y coordinate in plots
- Main title, labels, legends
- Draw vertical or horizontal lines, or lines with given slopes and intercepts
- Frequency table for 2 categorical variables
- Mosaic plots

How to install a package

```
# install.packages("ggplot2")
library(ggplot2)
```

Data

Today we are going to talk about data manipulation process, and deal with some real dataset. We will use one of the most popular build-in dataset 'mtcars' as an example. Let's first have a look on the dataset:

```
mtcars_data <- mtcars
mtcars_data
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
```

```
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
# ?mtcars
```

Subset the data

The type of this data is 'dataframe'. Typically, most of the dataset in R are storaged in the format of dataframe. So it is essential to know fundamental usages of dataframe. For example if we need to take some rows or columns from the whole dataset, we could do the following:

```
mtcars_data[1, ]  # the first row
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
```

```
mtcars_data[, c(2, 3)]  # the second and third column
```

```
##                     cyl  disp
## Mazda RX4             6 160.0
## Mazda RX4 Wag         6 160.0
## Datsun 710            4 108.0
## Hornet 4 Drive        6 258.0
## Hornet Sportabout     8 360.0
## Valiant               6 225.0
## Duster 360            8 360.0
## Merc 240D             4 146.7
## Merc 230              4 140.8
## Merc 280              6 167.6
## Merc 280C             6 167.6
## Merc 450SE            8 275.8
## Merc 450SL            8 275.8
## Merc 450SLC           8 275.8
## Cadillac Fleetwood    8 472.0
```

```
## Lincoln Continental   8 460.0
## Chrysler Imperial     8 440.0
## Fiat 128              4  78.7
## Honda Civic           4  75.7
## Toyota Corolla        4  71.1
## Toyota Corona         4 120.1
## Dodge Challenger      8 318.0
## AMC Javelin           8 304.0
## Camaro Z28            8 350.0
## Pontiac Firebird      8 400.0
## Fiat X1-9             4  79.0
## Porsche 914-2         4 120.3
## Lotus Europa          4  95.1
## Ford Pantera L        8 351.0
## Ferrari Dino          6 145.0
## Maserati Bora         8 301.0
## Volvo 142E            4 121.0
```

```
mtcars_data[, c("cyl", "disp")]
```

```
##                     cyl  disp
## Mazda RX4             6 160.0
## Mazda RX4 Wag         6 160.0
## Datsun 710            4 108.0
## Hornet 4 Drive        6 258.0
## Hornet Sportabout     8 360.0
## Valiant               6 225.0
## Duster 360            8 360.0
## Merc 240D             4 146.7
## Merc 230              4 140.8
## Merc 280              6 167.6
## Merc 280C             6 167.6
## Merc 450SE            8 275.8
## Merc 450SL            8 275.8
## Merc 450SLC           8 275.8
## Cadillac Fleetwood    8 472.0
## Lincoln Continental   8 460.0
## Chrysler Imperial     8 440.0
## Fiat 128              4  78.7
## Honda Civic           4  75.7
## Toyota Corolla        4  71.1
## Toyota Corona         4 120.1
## Dodge Challenger      8 318.0
## AMC Javelin           8 304.0
## Camaro Z28            8 350.0
## Pontiac Firebird      8 400.0
## Fiat X1-9             4  79.0
## Porsche 914-2         4 120.3
## Lotus Europa          4  95.1
## Ford Pantera L        8 351.0
## Ferrari Dino          6 145.0
## Maserati Bora         8 301.0
## Volvo 142E            4 121.0
```

```
mtcars_data$mpg
```

```
##  [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2
## [15] 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4
## [29] 15.8 19.7 15.0 21.4
```

```
mtcars_data[3, 5]  # one specific value
```

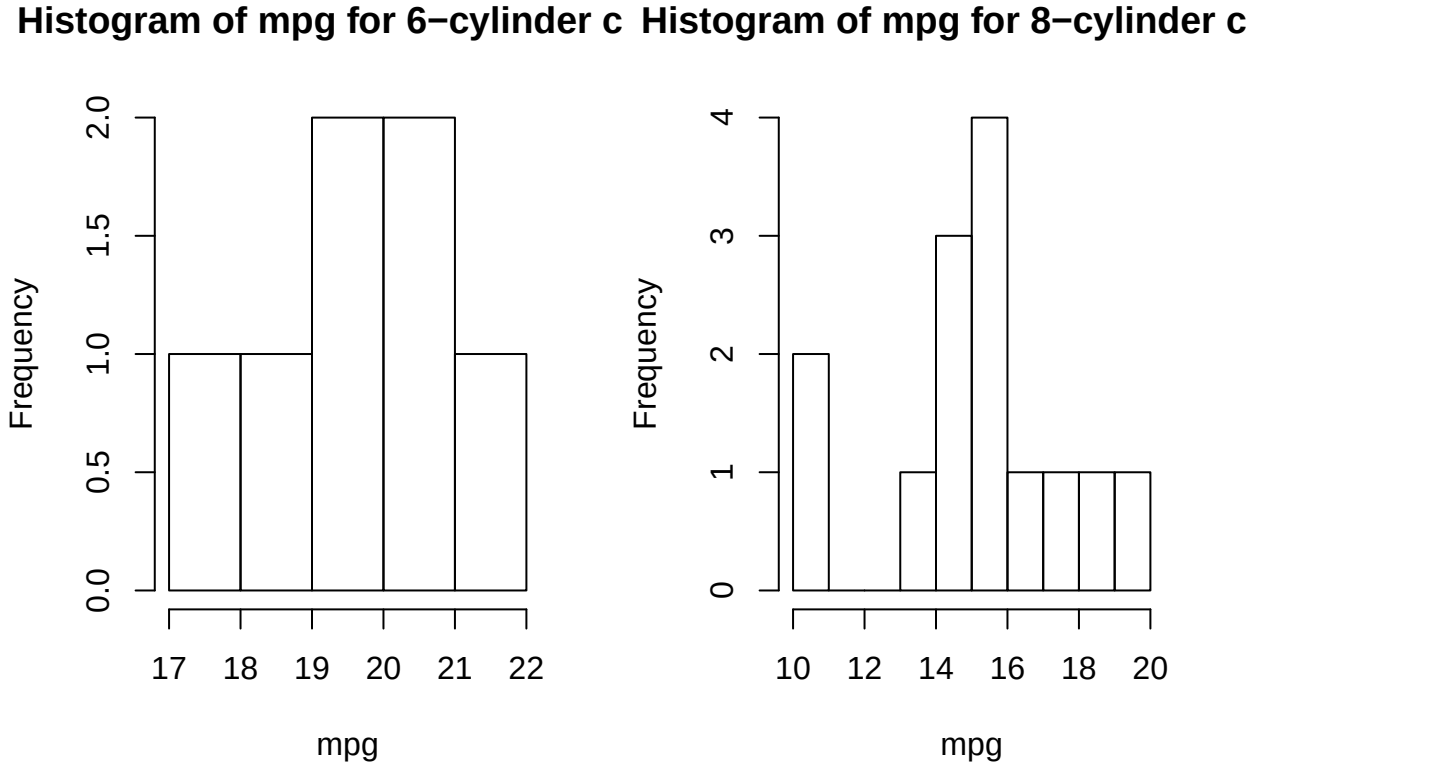
```
## [1] 3.85
```

Let's do it fancier:

```
mtcars_sub1 <- mtcars_data[mtcars_data$cyl == 6, ]
mtcars_sub2 <- mtcars_data[mtcars_data$cyl == 8, ]
```

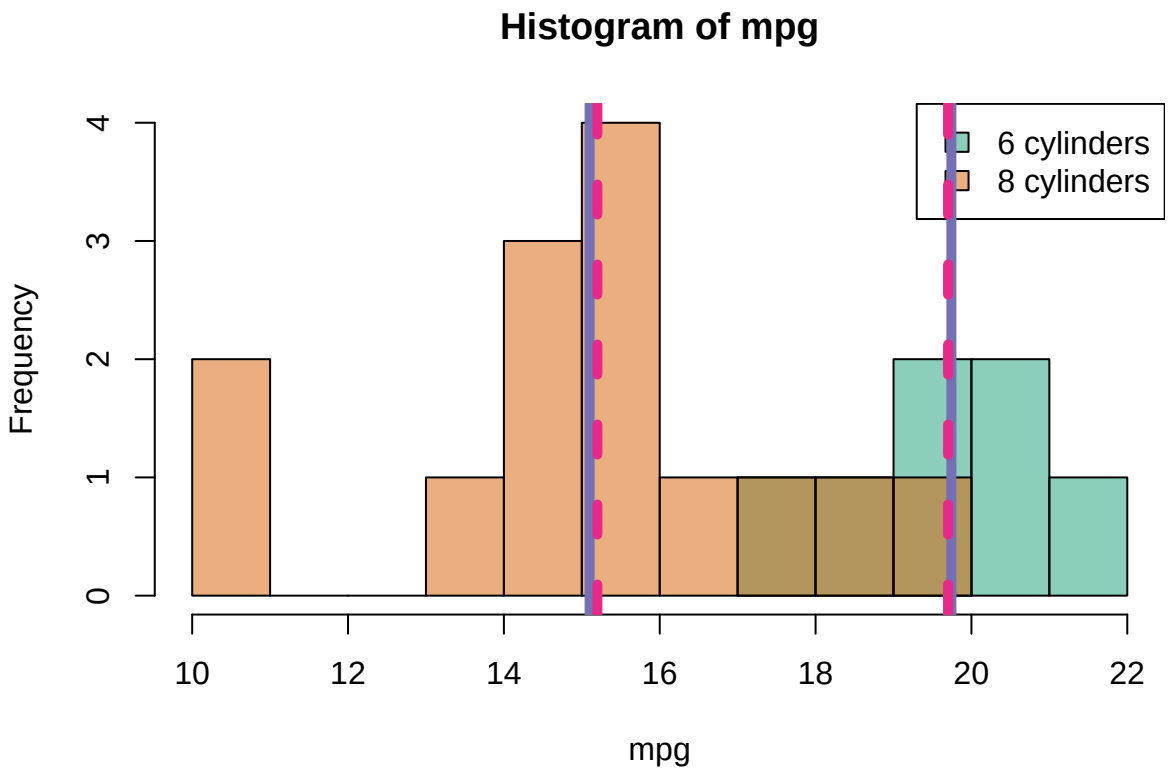
Have a look at the histogram

```
par(mfrow = c(1, 2))
hist1 <- hist(mtcars_sub1$mpg, breaks = 17:22, xlab = "mpg", main = "Histogram of mpg for 6-cylinder ca
hist2 <- hist(mtcars_sub2$mpg, breaks = 10:20, xlab = "mpg", main = "Histogram of mpg for 8-cylinder ca
```




```
library(scales)
library(RColorBrewer)
pal <- brewer.pal(8, 'Dark2')
par(mfrow = c(1, 1))
plot(hist1, col = alpha(pal[1], 0.5), xlim = c(10, 22), ylim = c(0, 4),
     xlab = "mpg", main = "Histogram of mpg")  # transparency = 0.5
plot(hist2, col = alpha(pal[2], 0.5), add = T)
legend("topright", legend = c("6 cylinders", "8 cylinders"), fill = alpha(pal[1:2], 0.5))
```

```
abline(v = c(mean(mtcars_sub1$mpg), mean(mtcars_sub2$mpg)), col = pal[3], lwd = 5)
abline(v = c(median(mtcars_sub1$mpg), median(mtcars_sub2$mpg)), col = pal[4], lty = 2, lwd = 5)
```




```
c(mean = mean(mtcars_sub1$mpg), median = median(mtcars_sub1$mpg))
```

```
##     mean   median
## 19.74286 19.70000
```

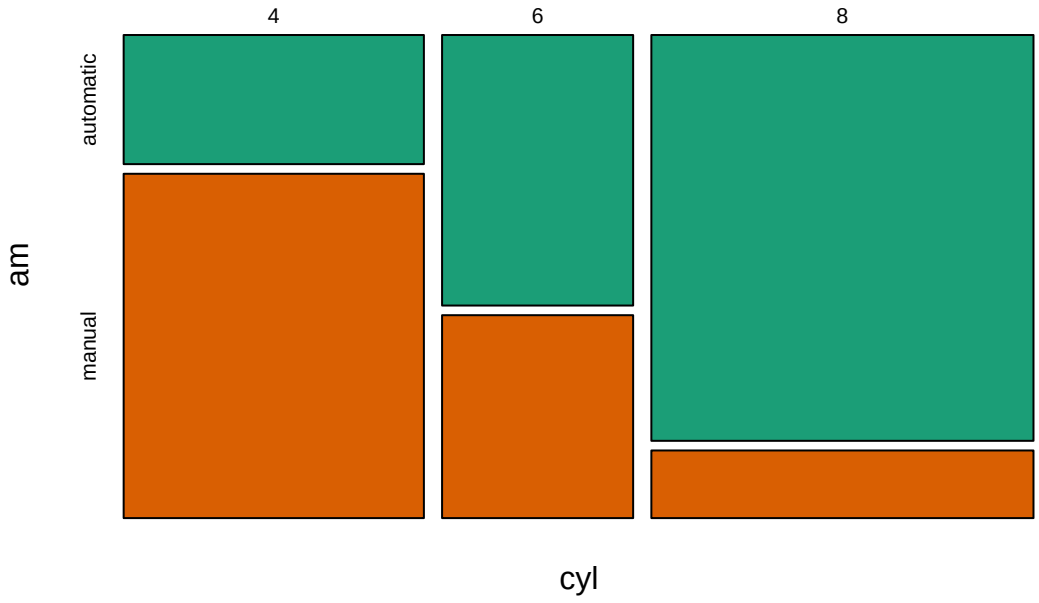
```
c(mean = mean(mtcars_sub2$mpg), median = median(mtcars_sub2$mpg))
```

```
##   mean median
##   15.1   15.2
```

Create mosaic plots

For dataset with two categorical variables, we can create the mosaic plot:

```
cyl_am_table <- table(mtcars_data[, c("cyl", "am")])
colnames(cyl_am_table) <- c("automatic", "manual")
mosaicplot(cyl_am_table, main = "Mosaic plot for cylindar and transmission", col = pal[1:2])
```


Mosaic plot for cylindar and transmission
4
6
8
automatic
manual
am
cyl