Assignment 2

Yue Ma

2018/10/6

1. Matrix Algebra

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
A = matrix(c(5, 6, 1, 2, 2, 3), nrow = 2, ncol = 3)
B = matrix(c(3, -2, 4, -3, 5, 6), nrow = 2, ncol = 3)
C = matrix(c(1, -5, -3, 2, 3, 1), nrow = 3, ncol = 2)
D = matrix(c(2, 4, 1, 3), nrow = 2, ncol = 2)
```

a) A + C Error, A and C can not make the additive operation because they do not have the same number of rows and columns.

```
A + C
```

```
## Error in A + C: non-conformable arrays
```

b) A - B

```
A - B
```

```
##      [,1] [,2] [,3]
## [1,]    2   -3   -3
## [2,]    8    5   -3
```

c) A + 5B

```
A + 5 * B
```

```
##      [,1] [,2] [,3]
## [1,]   20   21   27
## [2,]   -4  -13   33
```

d) 3A

```
3 * A
```

```
##      [,1] [,2] [,3]
## [1,]   15    3    6
## [2,]   18    6    9
```

e) 2B - 5A

```r
2 * B - 5 * A
```

```
##      [,1] [,2] [,3]
## [1,]  -19    3    0
## [2,]  -34  -16   -3
```

f) B^T - C

```r
t(B) - C
```

```
##      [,1] [,2]
## [1,]    2   -4
## [2,]    9   -6
## [3,]    8    5
```

g) BA

Note: B %% *A does not work, so I put* BA here.

```r
B %*% A
```

```
## Error in B %*% A: non-conformable arguments
```

```r
B * A
```

```
##      [,1] [,2] [,3]
## [1,]   15    4   10
## [2,]  -12   -6   18
```

h) DA

```r
D %*% A
```

```
##      [,1] [,2] [,3]
## [1,]   16    4    7
## [2,]   38   10   17
```

i) AD Note: Error here, because the number of column in A is not equal to the number of row in D

```r
A %*% D
```

```
## Error in A %*% D: non-conformable arguments
```

```r
A * D
```

```
## Error in A * D: non-conformable arrays
```

j) CD

```r
C %*% D
```

```
##      [,1] [,2]
## [1,]   10    7
## [2,]    2    4
## [3,]   -2    0
```

k) BC

```r
B %*% C
```

```
##      [,1] [,2]
## [1,]  -32   23
## [2,]   -5   -7
```

l) CB

```
C %*% B
```

```
##      [,1] [,2] [,3]
## [1,]   -1   -2   17
## [2,]  -21  -29   -7
## [3,]  -11  -15   -9
```

2. Inverses of Matrices

It is true and W^(-1) exists. The reason is as below:

$$(W^{-1} - W^{-1}xy^\top W^{-1})/(1 + y^\top W^{-1}x))(W + xy^\top)$$

$$= E + W^{-1}xy^\top - W^{-1}xy^\top(E + W^{-1}xy^\top)/(1 + y^\top W^{-1}x)$$

$$= E + (W^{-1}xy^\top + W^{-1}xy^\top y^\top W^{-1}x - W^{-1}xy^\top(E + W^{-1}xy^\top))/(1 + y^\top W^{-1}x)$$

$$= E + (W^{-1}xy^\top(E + y^\top W^{-1}x) - W^{-1}xy^\top(E + W^{-1}xy^\top))/(1 + y^\top W^{-1}x)$$

$$= E$$

3. Stratifying

```
cdc <- read.csv("https://www.openintro.org/stat/data/cdc.csv")
library(dplyr)
weight_subgroup <- cdc %>%
  group_by(gender, hlthplan) %>%
   summarize(df_mean = mean(weight - wtdesire, na.rm = TRUE),
             df_median = median(weight - wtdesire, na.rm = TRUE))
weight_subgroup
```

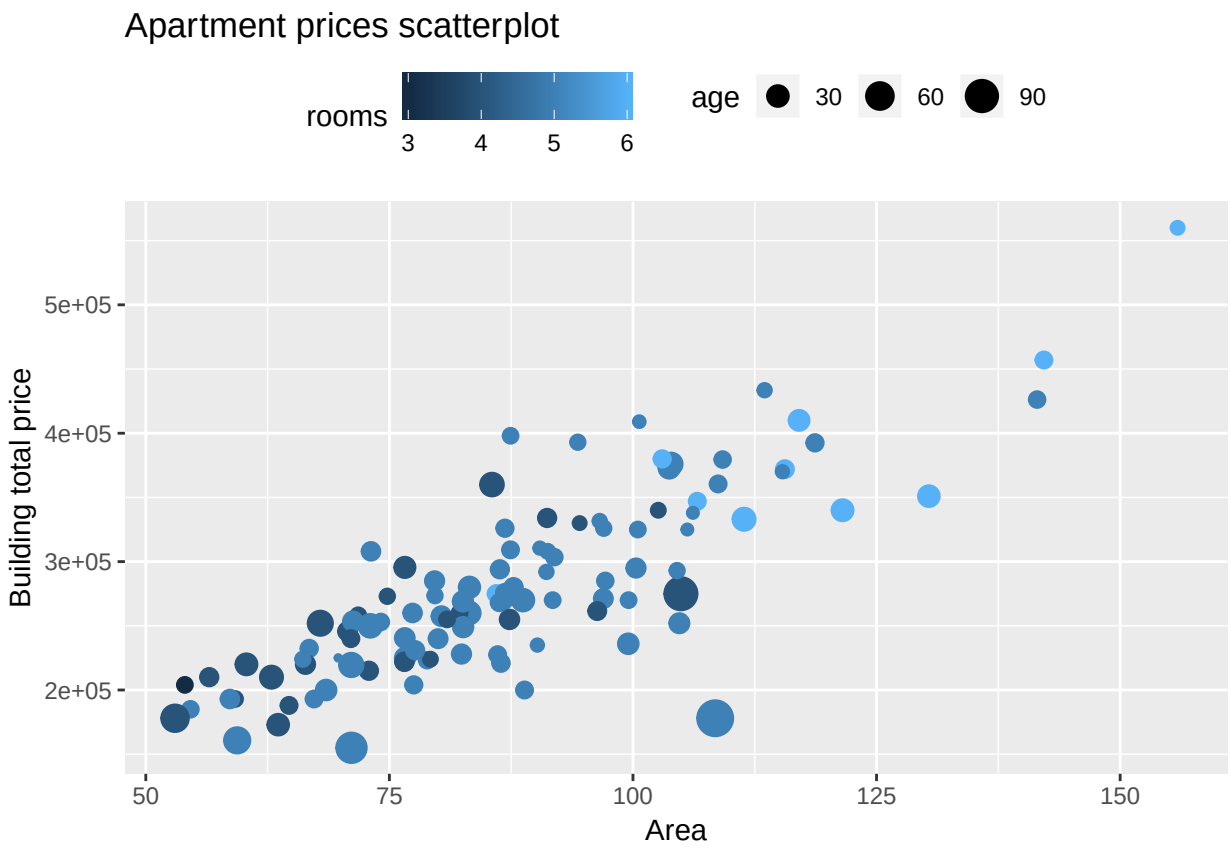
```
## # A tibble: 4 x 4
## # Groups:   gender [?]
##   gender hlthplan df_mean df_median
##   <fct>     <int>   <dbl>     <dbl>
## 1 f             0   20.0         15
## 2 f             1   17.9         10
## 3 m             0    7.98         0
## 4 m             1   11.1          5
```

From the results, we can draw 2 conclusions. The first one is that for male, those with some form of health coverage has larger difference between desire weight and actual weight in average. The median is one measure of the "center" of a distribution of data. The median of difference is also larger for those with health coverage. The second one is the opposite. For female, those without some form of health coverage has larger difference between desire weight and actual weight in average. The median of difference is also larger for those without health coverage. As a result, the health coverage has different incfluence on the weight and desire weight for different genders.

4. Apartment Prices

```
apts <- readRDS(url('https://courseworks.columbia.edu/x/pJdP39'))
library(ggplot2)
apts$age = as.numeric(as.character(apts$age))
apts$rooms = as.numeric(as.character(apts$rooms))
```

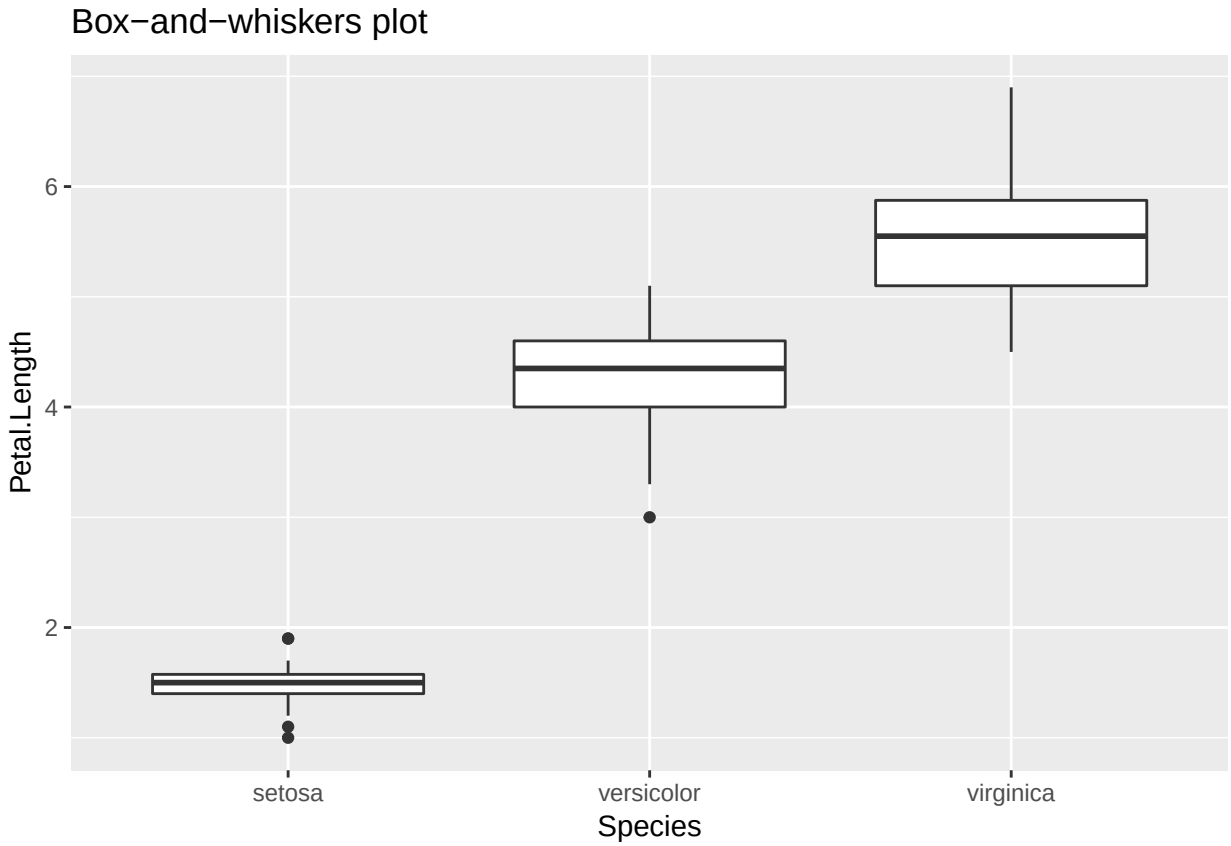
```
gg <- ggplot(apts)
gg <- gg + geom_point(mapping = aes(x = area, y = totalprice,
                                    color = rooms,
                                    size = age)) +
  ggtitle("Apartment prices scatterplot") +
  ylab("Building total price") +
  xlab("Area")
gg <- gg + theme(legend.position = "top")
gg
```



In the scatterplot, there is a positive correlation between the building area and the building total price. I vary the color and size of the points to illustrate the relationsip between the totalprice and rooms and age. Also, there is a positive correlation between the building rooms and the buidling total price. However, there is a negative correlation between the building ages and the building totalprice.

5. Making plots

```
library(ggplot2)
help("iris")
gg1 <- ggplot(iris)
gg1 <- gg1 + geom_boxplot(mapping = aes(x = Species,
                                        y = Petal.Length)) +
  ggtitle("Box-and-whiskers plot")
gg1
```



From the box-and-whiskers plot, we can draw the following conclusions: 1. The thick line within the box is the median, and the median is the center of the data. For the three species, the petal length of Virginica is the largest and that of Setosa is the shortest. 2. The box indicates the 25th percentile and the 75th percentile. The range of the petal length of Viginica is the largest and that of Setosa is the smallest. 3. Points are drawn outside the whiskers, which are in some sense "outliers". There is no "outliers" of Virginica.

6. Histograms

```
library(MASS)
```

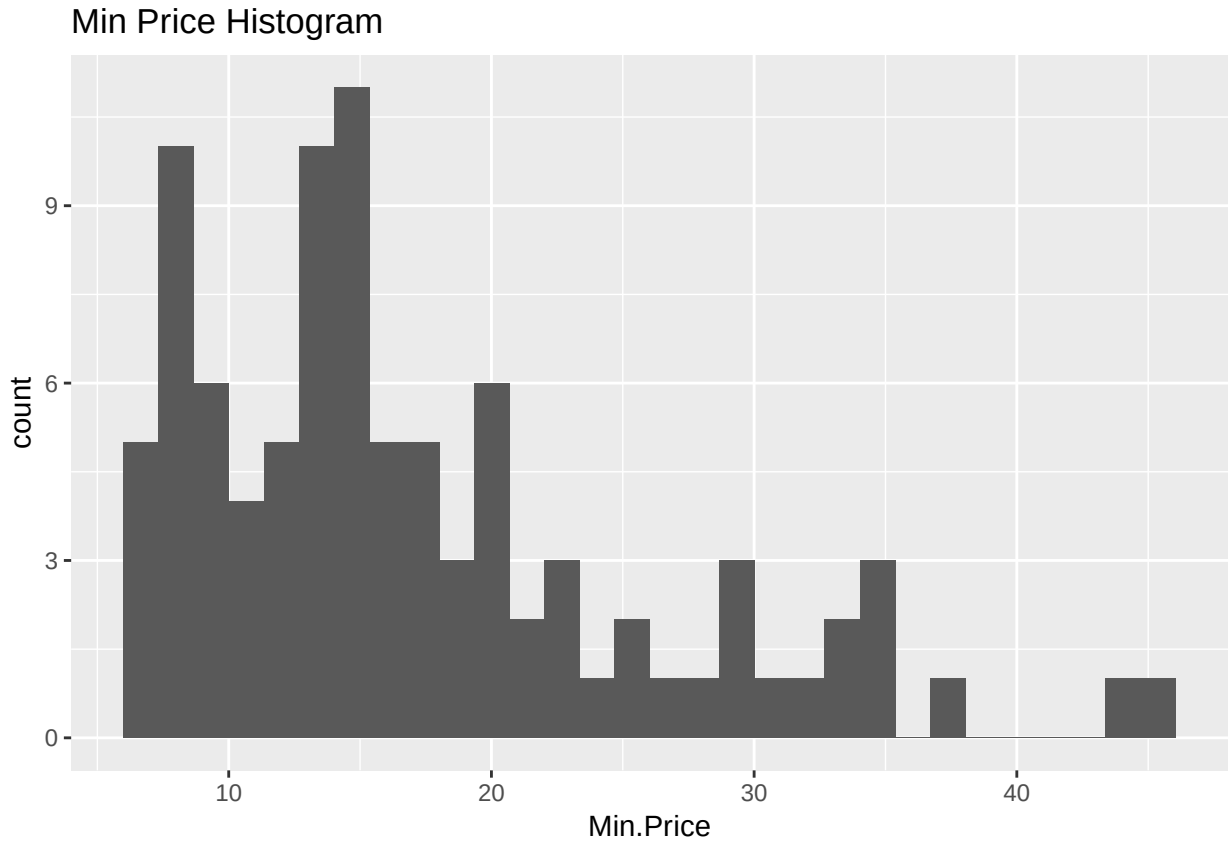
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggplot2)
gg2 <- ggplot(Cars93)
gg_min_price <- gg2 + geom_histogram(mapping = aes(Min.Price)) +
  ggtitle("Min Price Histogram")
gg_max_price <- gg2 + geom_histogram(mapping = aes(Max.Price)) +
  ggtitle("Max Price Histogram")
gg_weight <- gg2 + geom_histogram(mapping = aes(Weight)) +
  ggtitle("Weight Histogram")
gg_length <- gg2 + geom_histogram(mapping = aes(Length)) +
```

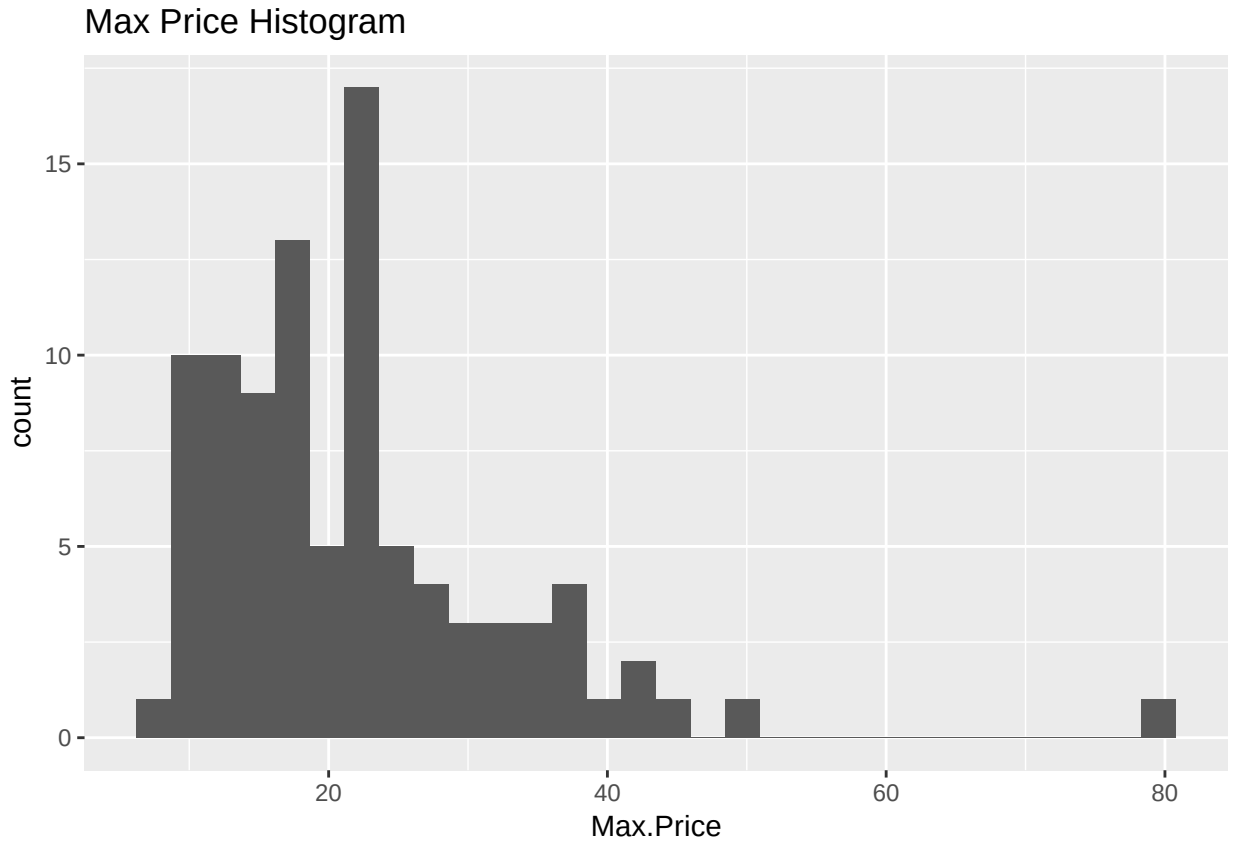
```
  ggtitle("Length Histogram")
gg_min_price
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



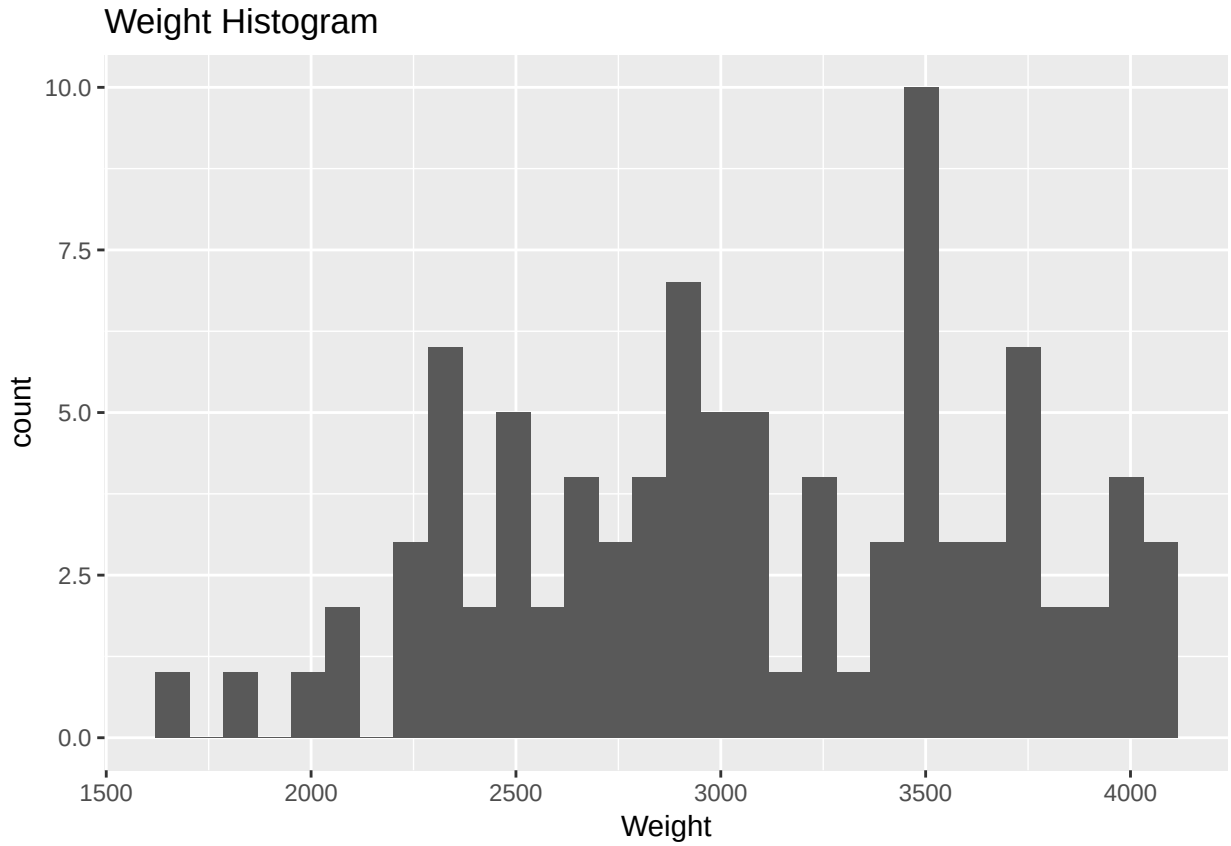

```
gg_max_price
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



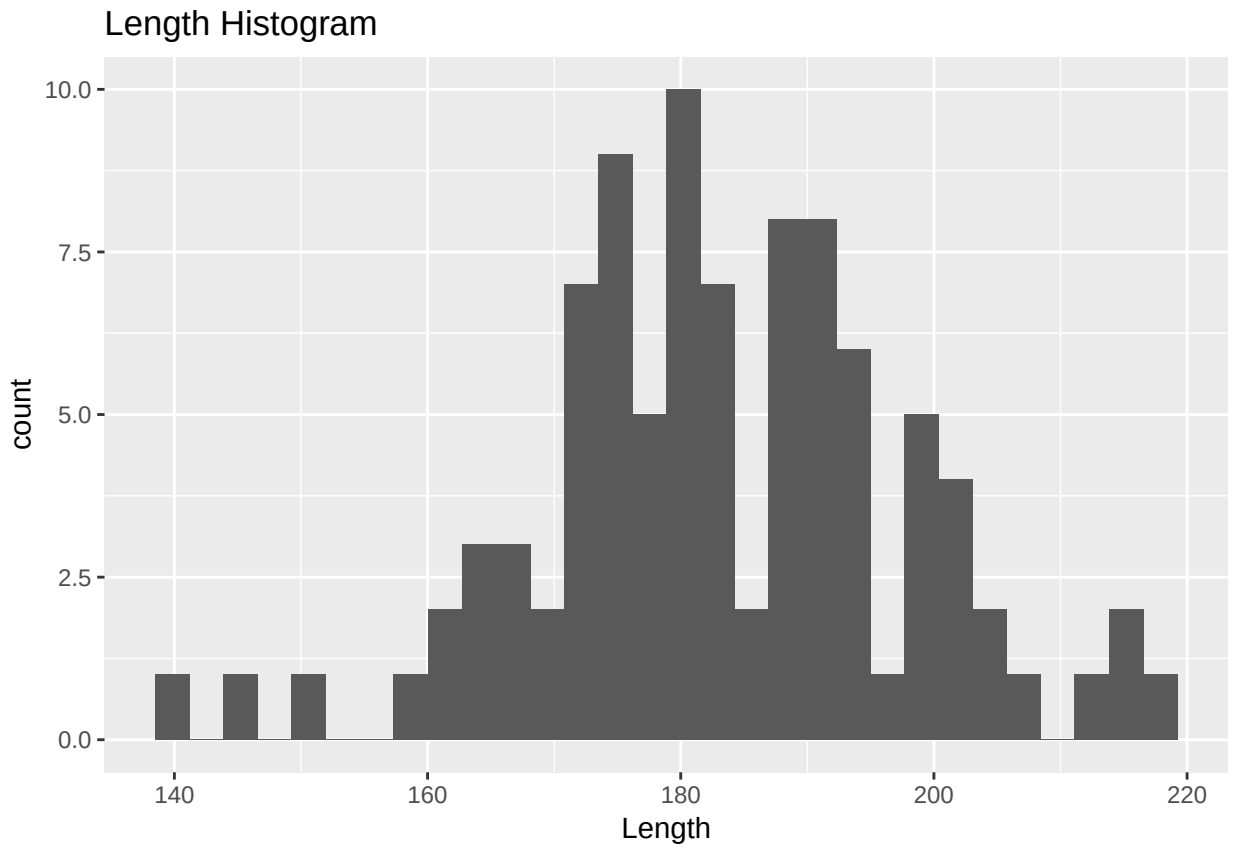

```
gg_weight
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



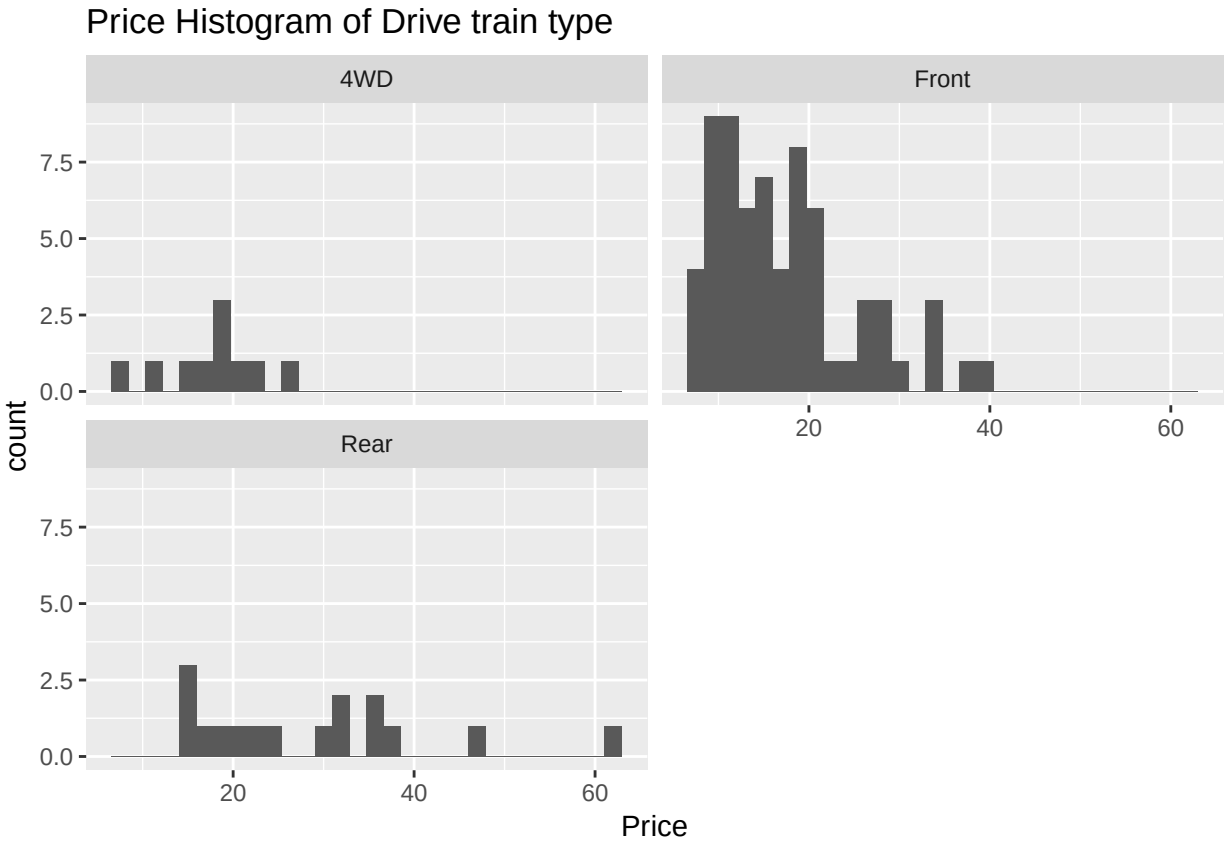

```
gg_length
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
gg_price <- gg2 + geom_histogram(mapping = aes(Price)) +
  facet_wrap(~DriveTrain, nrow = 2) +
  ggtitle("Price Histogram of Drive train type")
gg_price
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the Min Price Histogram, we can see that most minimum price of the cars is below $35,000. From the Max Price Histogram, we can see that most maximum price of the cars is below $45,000. From the Weight Histogram, we can see that most weight of the cars is between 2250 and 4000 pounds. From the Length Histogram, we can see that most length of cars is between 160 and 210 inches. From the Price Histogram of Drive train type, we can see that the midrange price of 4WD and Front wheel is almost all less than 40,000 dollars, and most midrange price of Rear wheel is less than 40,000 dollars. As a result, for Front wheel, the price range is small, and the its minimum price and maximum price are more close compared to other two types of wheels.