hw1.R

asahikuroki222

2021-11-04

```
# Part 1
# a)
my_data = read.csv("LaptopSales (1).csv")
head(my_data)
```

```
##                  Date Configuration Customer.Postcode Store.Postcode
## 1 2008/01/01 00:01:19           163          EC4V 5BH        SE1 2BN
## 2 2008/01/01 00:02:52           320           SW4 0JL       SW12 9HD
## 3 2008/01/01 00:04:18            23          EC3V 1LR         E2 0RY
## 4 2008/01/01 00:04:40           169          SW1P 3AU        SE1 2BN
## 5 2008/01/01 00:06:04           365          EC4V 4EG       SW1V 4QQ
## 6 2008/01/01 00:12:26           309           W1B 5PX       SW1V 4QQ
##   Retail.Price Screen.Size..Inches. Battery.Life..Hours. RAM..GB.
## 1          455                   15                    5        1
## 2          545                   15                    6        1
## 3          515                   15                    4        1
## 4          395                   15                    5        1
## 5          585                   15                    6        2
## 6          555                   15                    6        1
##   Processor.Speeds..GHz. Integrated.Wireless. HD.Size..GB.
## 1                      2                  Yes           80
## 2                      2                   No          300
## 3                      2                  Yes          300
## 4                      2                   No           40
## 5                      2                   No          120
## 6                      2                  Yes          120
##   Bundled.Applications. OS.X.Customer OS.Y.Customer OS.X.Store OS.Y.Store
## 1                   Yes        532041        180995     534057     179682
## 2                    No        529240        175537     528739     173080
## 3                   Yes        533095        181047     535652     182961
## 4                   Yes        529902        179641     534057     179682
## 5                   Yes        531684        180948     528924     178440
## 6                   Yes        529207        180969     528924     178440
##   CustomerStoreDistance
## 1              2405.873
## 2              2507.559
## 3              3194.001
## 4              4155.202
## 5              3729.298
## 6              2544.785
```

```
# b) missing data at OS.X.Store, OS.Y.store, and CustomerStoreDistance
summary(my_data)
```

```
##      Date           Configuration   Customer.Postcode  Store.Postcode
##  Length:2514        Min.   :  1.0   Length:2514        Length:2514
##  Class :character   1st Qu.: 78.0   Class :character   Class :character
##  Mode  :character   Median :212.0   Mode  :character   Mode  :character
##                     Mean   :209.9
##                     3rd Qu.:315.8
##                     Max.   :368.0
##
##   Retail.Price   Screen.Size..Inches. Battery.Life..Hours.    RAM..GB.
##  Min.   :300.0   Min.   :15           Min.   :4.00         Min.   :1.000
##  1st Qu.:455.0   1st Qu.:15           1st Qu.:4.00         1st Qu.:1.000
##  Median :490.0   Median :15           Median :5.00         Median :2.000
##  Mean   :489.8   Mean   :15           Mean   :5.16         Mean   :1.538
##  3rd Qu.:530.0   3rd Qu.:15           3rd Qu.:6.00         3rd Qu.:2.000
##  Max.   :665.0   Max.   :15           Max.   :6.00         Max.   :2.000
##
##  Processor.Speeds..GHz. Integrated.Wireless.  HD.Size..GB.
##  Min.   :1.500          Length:2514          Min.   : 40.0
##  1st Qu.:1.500          Class :character     1st Qu.: 80.0
##  Median :2.000          Mode  :character     Median :120.0
##  Mean   :1.757                               Mean   :150.9
##  3rd Qu.:2.000                               3rd Qu.:300.0
##  Max.   :2.000                               Max.   :300.0
##
##  Bundled.Applications. OS.X.Customer    OS.Y.Customer      OS.X.Store
##  Length:2514           Min.   :512253   Min.   :164886   Min.   :517917
##  Class :character      1st Qu.:529281   1st Qu.:178695   1st Qu.:528924
##  Mode  :character      Median :531190   Median :181082   Median :529902
##                        Mean   :530926   Mean   :179837   Mean   :530821
##                        3rd Qu.:533237   3rd Qu.:182049   3rd Qu.:534057
##                        Max.   :549065   Max.   :199846   Max.   :541428
##                                                          NA's   :4
##    OS.Y.Store     CustomerStoreDistance
##  Min.   :168302   Min.   :    0
##  1st Qu.:178440   1st Qu.: 2385
##  Median :179641   Median : 3368
##  Mean   :179827   Mean   : 3680
##  3rd Qu.:182961   3rd Qu.: 4331
##  Max.   :190628   Max.   :19892
##  NA's   :4        NA's   :4
```

```
which(is.na(my_data$OS.X.Store))
```

```
## [1] 1675 1774 1969 2203
```

```
which(is.na(my_data$OS.Y.Store))
```

```
## [1] 1675 1774 1969 2203
```

```
which(is.na(my_data$CustomerStoreDistance))
```

```
## [1] 1675 1774 1969 2203
```

```
### missing values in row 1675 1774 1969 2203

# c) mean: 489.8, median 490
# d)
data_integrated_wireless <- subset(my_data, Integrated.Wireless. ==
                                    "Yes")
data_non_intergrated_wireless <- subset(my_data, Integrated.Wireless. != "Yes")
summary(data_integrated_wireless)
```

```
##      Date           Configuration   Customer.Postcode  Store.Postcode
##  Length:1301        Min.   :  1.0   Length:1301        Length:1301
##  Class :character   1st Qu.: 71.0   Class :character   Class :character
##  Mode  :character   Median :210.0   Mode  :character   Mode  :character
##                     Mean   :202.6
##                     3rd Qu.:308.0
##                     Max.   :360.0
##
##   Retail.Price   Screen.Size..Inches. Battery.Life..Hours.    RAM..GB.
##  Min.   :320.0   Min.   :15           Min.   :4.00         Min.   :1.000
##  1st Qu.:460.0   1st Qu.:15           1st Qu.:4.00         1st Qu.:1.000
##  Median :495.0   Median :15           Median :5.00         Median :2.000
##  Mean   :495.9   Mean   :15           Mean   :5.14         Mean   :1.533
##  3rd Qu.:535.0   3rd Qu.:15           3rd Qu.:6.00         3rd Qu.:2.000
##  Max.   :665.0   Max.   :15           Max.   :6.00         Max.   :2.000
##
##  Processor.Speeds..GHz. Integrated.Wireless.  HD.Size..GB.
##  Min.   :1.500         Length:1301          Min.   : 40.0
##  1st Qu.:1.500         Class :character     1st Qu.: 80.0
##  Median :2.000         Mode  :character     Median :120.0
##  Mean   :1.752                              Mean   :147.7
##  3rd Qu.:2.000                              3rd Qu.:300.0
##  Max.   :2.000                              Max.   :300.0
##
##  Bundled.Applications. OS.X.Customer    OS.Y.Customer      OS.X.Store
##  Length:1301           Min.   :512253   Min.   :164886   Min.   :517917
##  Class :character      1st Qu.:529174   1st Qu.:178524   1st Qu.:528924
##  Mode  :character      Median :531065   Median :181063   Median :529902
##                        Mean   :530869   Mean   :179822   Mean   :530883
##                        3rd Qu.:533246   3rd Qu.:182055   3rd Qu.:534057
##                        Max.   :549065   Max.   :199846   Max.   :541428
##                                                          NA's   :1
##    OS.Y.Store     CustomerStoreDistance
##  Min.   :168302   Min.   :    0
##  1st Qu.:178440   1st Qu.: 2424
##  Median :179641   Median : 3418
##  Mean   :179787   Mean   : 3774
##  3rd Qu.:182961   3rd Qu.: 4406
##  Max.   :190628   Max.   :19892
##  NA's   :1        NA's   :1
```

```
summary(data_non_intergrated_wireless)
```

```
##      Date           Configuration   Customer.Postcode  Store.Postcode
##  Length:1213        Min.   :  9.0   Length:1213        Length:1213
##  Class :character   1st Qu.: 80.0   Class :character   Class :character
##  Mode  :character   Median :219.0   Mode  :character   Mode  :character
##                     Mean   :217.7
##                      3rd Qu.:318.0
##                      Max.   :368.0
##
##   Retail.Price   Screen.Size..Inches. Battery.Life..Hours.    RAM..GB.
##  Min.   :300.0   Min.   :15           Min.   :4.000        Min.   :1.000
##  1st Qu.:455.0   1st Qu.:15           1st Qu.:4.000        1st Qu.:1.000
##  Median :485.0   Median :15           Median :5.000        Median :2.000
##  Mean   :483.3   Mean   :15           Mean   :5.182        Mean   :1.544
##  3rd Qu.:520.0   3rd Qu.:15           3rd Qu.:6.000        3rd Qu.:2.000
##  Max.   :645.0   Max.   :15           Max.   :6.000        Max.   :2.000
##
##  Processor.Speeds..GHz. Integrated.Wireless.  HD.Size..GB.
##  Min.   :1.500          Length:1213          Min.   : 40.0
##  1st Qu.:1.500          Class  :character    1st Qu.: 80.0
##  Median :2.000          Mode   :character    Median :120.0
##  Mean   :1.763                               Mean   :154.3
##  3rd Qu.:2.000                               3rd Qu.:300.0
##  Max.   :2.000                               Max.   :300.0
##
##  Bundled.Applications. OS.X.Customer    OS.Y.Customer      OS.X.Store
##  Length:1213           Min.   :512253   Min.   :165028   Min.   :517917
##  Class  :character     1st Qu.:529342   1st Qu.:178835   1st Qu.:528924
##  Mode   :character     Median :531255   Median :181083   Median :529902
##                        Mean   :530987   Mean   :179853   Mean   :530753
##                        3rd Qu.:533180   3rd Qu.:182019   3rd Qu.:534057
##                        Max.   :549065   Max.   :193894   Max.   :541428
##                                                          NA's   :3
##    OS.Y.Store     CustomerStoreDistance
##  Min.   :168302   Min.   :    0
##  1st Qu.:178440   1st Qu.: 2322
##  Median :179641   Median : 3258
##  Mean   :179871   Mean   : 3579
##  3rd Qu.:182961   3rd Qu.: 4228
##  Max.   :190628   Max.   :13530
##  NA's   :3        NA's   :3
```

```
### Average price of a laptop with Integrated Wireless $495.9
### Average price of a laptop without Integrated Wireless $483.3

# e)
my_data_sorted <- my_data[order(my_data$Retail.Price, decreasing = TRUE),]
my_data_sorted[1, ]
```

```
##                    Date Configuration Customer.Postcode Store.Postcode
## 12 2008/01/01 01:03:25           359           W1T 1DG        NW5 2QH
##    Retail.Price Screen.Size..Inches. Battery.Life..Hours. RAM..GB.
## 12          665                   15                    6        2
##    Processor.Speeds..GHz. Integrated.Wireless. HD.Size..GB.
## 12                      2                  Yes          300
##    Bundled.Applications. OS.X.Customer OS.Y.Customer OS.X.Store OS.Y.Store
## 12                   Yes        529584        181554     529248     185213
##    CustomerStoreDistance
## 12              3674.395
```

```r
### Configuration type with the highest price is 359

# f)
sum(my_data$HD.Size..GB. < 150)
```

```
## [1] 1749
```

```r
### 1749

# g)
sum(my_data$Retail.Price)
```

```
## [1] 1231470
```

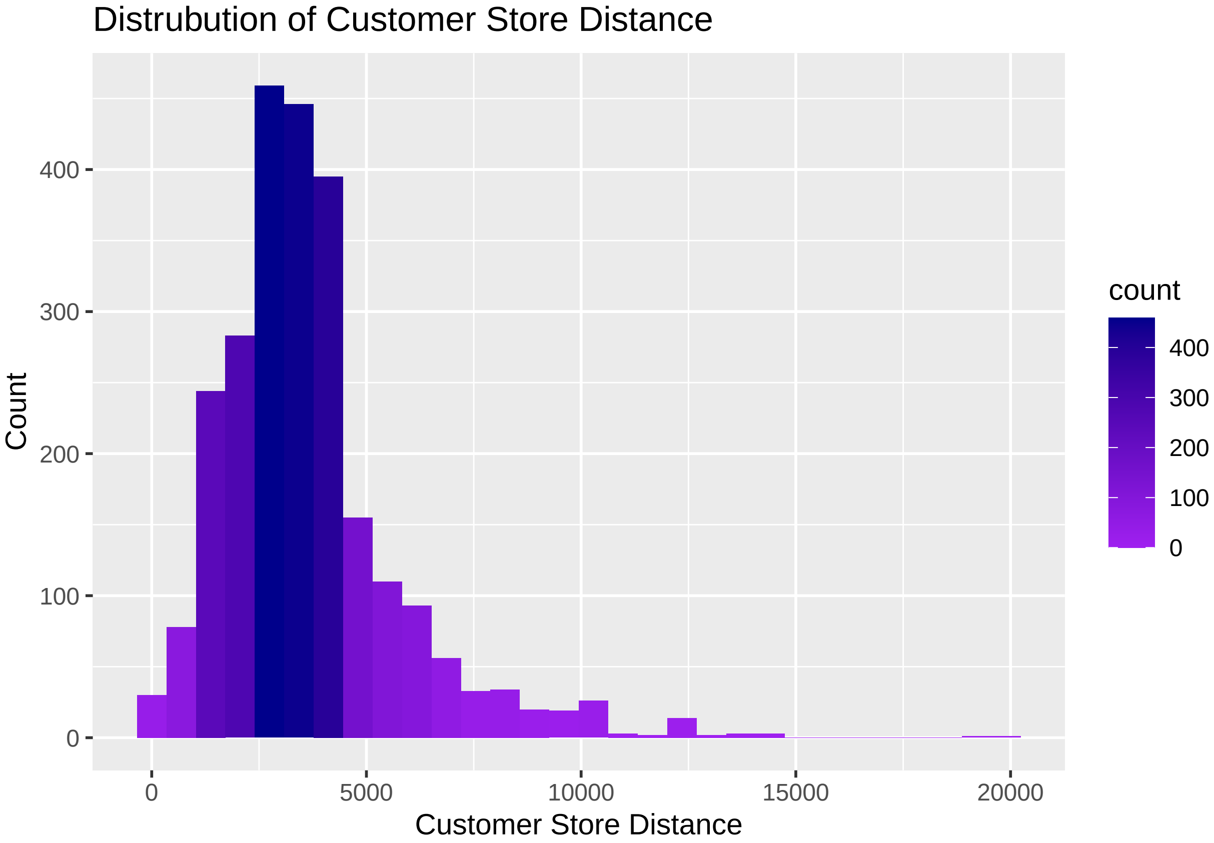
```r
### Total price = $ 1231470


### Part2
library(ggplot2)
# a)
summary(my_data)
```

```
##      Date           Configuration   Customer.Postcode  Store.Postcode
##  Length:2514        Min.   :  1.0   Length:2514        Length:2514
##  Class :character   1st Qu.: 78.0   Class :character   Class :character
##  Mode  :character   Median :212.0   Mode  :character   Mode  :character
##                     Mean   :209.9
##                     3rd Qu.:315.8
##                     Max.   :368.0
##
##   Retail.Price   Screen.Size..Inches. Battery.Life..Hours.    RAM..GB.
##  Min.   :300.0   Min.   :15           Min.   :4.00         Min.   :1.000
##  1st Qu.:455.0   1st Qu.:15           1st Qu.:4.00         1st Qu.:1.000
##  Median :490.0   Median :15           Median :5.00         Median :2.000
##  Mean   :489.8   Mean   :15           Mean   :5.16         Mean   :1.538
##  3rd Qu.:530.0   3rd Qu.:15           3rd Qu.:6.00         3rd Qu.:2.000
##  Max.   :665.0   Max.   :15           Max.   :6.00         Max.   :2.000
##
##  Processor.Speeds..GHz. Integrated.Wireless.  HD.Size..GB.
##  Min.   :1.500          Length:2514          Min.   : 40.0
##  1st Qu.:1.500          Class :character     1st Qu.: 80.0
##  Median :2.000          Mode  :character     Median :120.0
##  Mean   :1.757                               Mean   :150.9
##  3rd Qu.:2.000                               3rd Qu.:300.0
##  Max.   :2.000                               Max.   :300.0
##
##  Bundled.Applications. OS.X.Customer    OS.Y.Customer      OS.X.Store
##  Length:2514           Min.   :512253   Min.   :164886   Min.   :517917
##  Class :character      1st Qu.:529281   1st Qu.:178695   1st Qu.:528924
##  Mode  :character      Median :531190   Median :181082   Median :529902
##                        Mean   :530926   Mean   :179837   Mean   :530821
##                        3rd Qu.:533237   3rd Qu.:182049   3rd Qu.:534057
##                        Max.   :549065   Max.   :199846   Max.   :541428
##                                                          NA's   :4
##    OS.Y.Store     CustomerStoreDistance
##  Min.   :168302   Min.   :    0
##  1st Qu.:178440   1st Qu.: 2385
##  Median :179641   Median : 3368
##  Mean   :179827   Mean   : 3680
##  3rd Qu.:182961   3rd Qu.: 4331
##  Max.   :190628   Max.   :19892
##  NA's   :4        NA's   :4
```

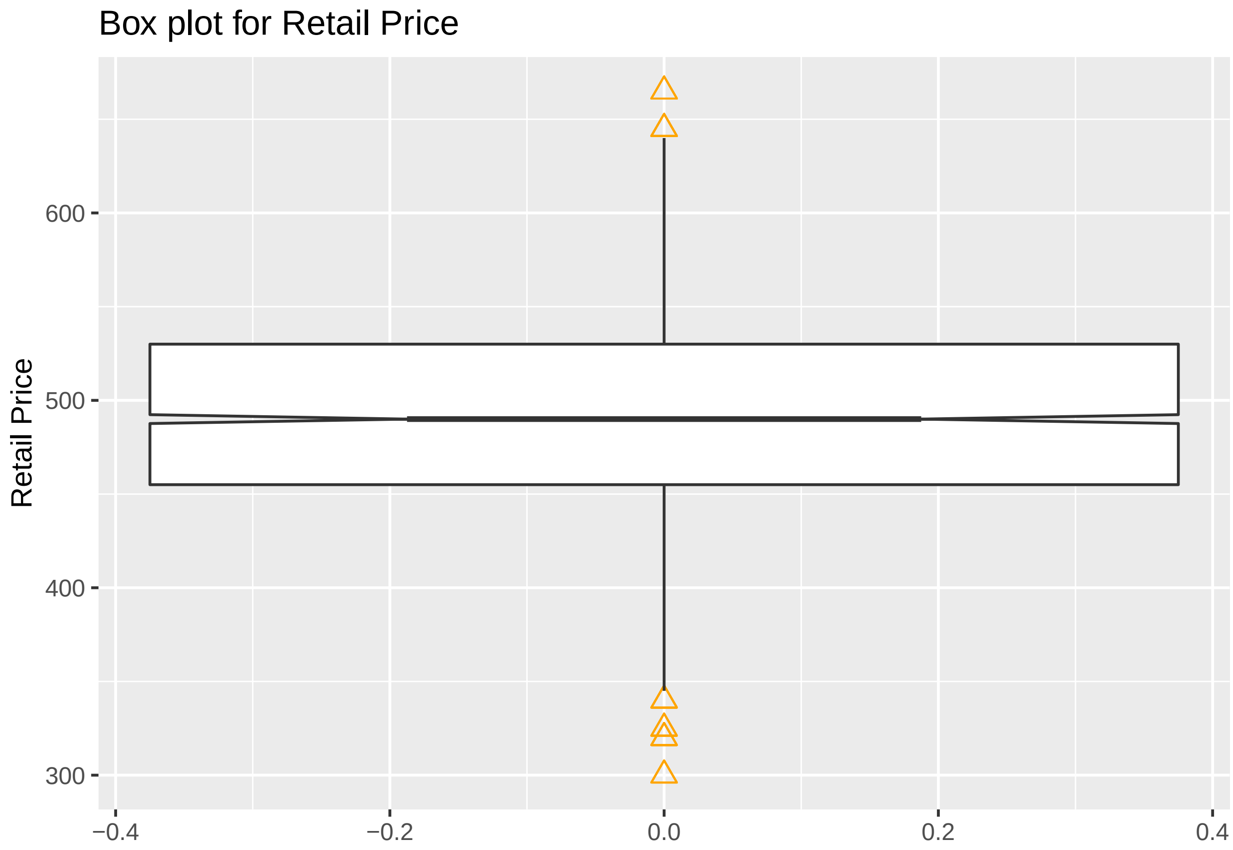
```
ggplot(data= my_data, aes(x = CustomerStoreDistance, fill = ..count..)) +
  geom_histogram(alpha=1) +
  scale_fill_gradient(low="purple", high="darkblue") +
  ggtitle("Distrubution of Customer Store Distance") +
  labs(x = "Customer Store Distance", y = "Count")
```

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

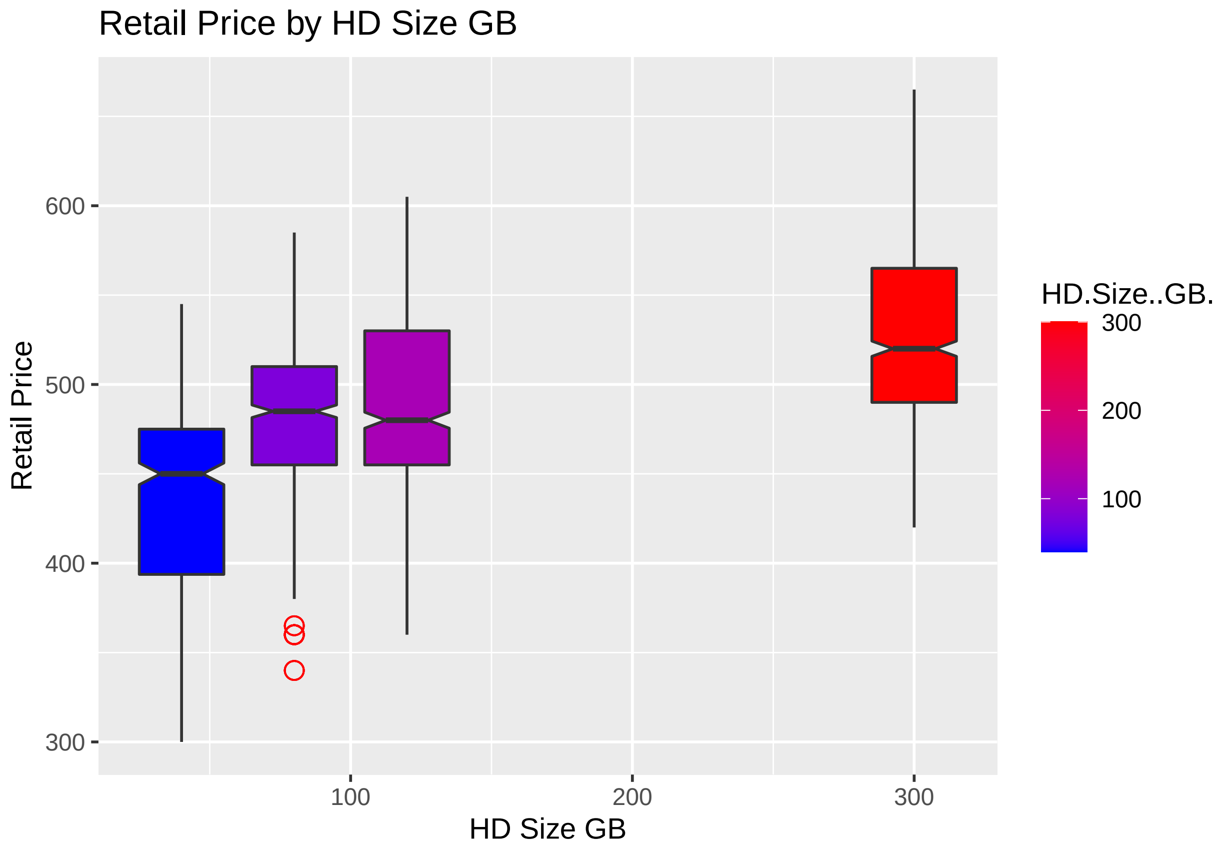
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



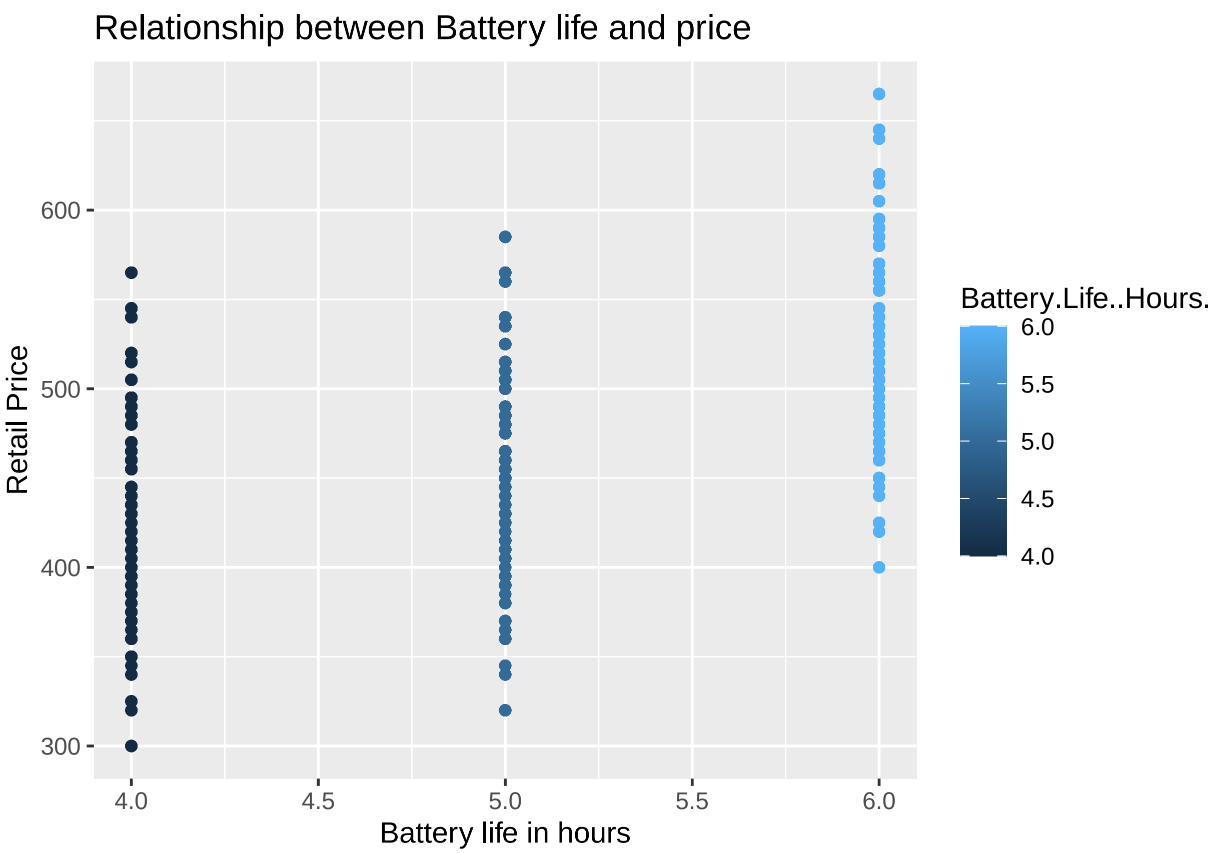
```
### Insights: We could see from the data that as the distance grow, counts decreases.
# Therefore, distance is one of an important aspect of shopping for customers.
# Also, we could see that counts are the highest between 2500 - approximately 4000

# b)
ggplot(data = my_data, aes(y=Retail.Price)) +
  geom_boxplot(notch = TRUE, outlier.colour="orange", outlier.shape=2, outlier.size=3) +
  ggtitle("Box plot for Retail Price") +
  labs(y = "Retail Price")
```



```
### Insights: We could see from the boxplot that most of the Retail Prices are
# in the range from 455 - 530.
# Also, there are more outliers in the minimum side than the maximum side.


# c)
ggplot(data <- my_data, aes(x = HD.Size..GB., y=Retail.Price, group = HD.Size..GB.
, fill = HD.Size..GB.)) +
  geom_boxplot(notch = TRUE, outlier.colour="red", outlier.shape=1, outlier.size=3) +
  scale_fill_gradient(low="blue", high="red") +
  ggtitle("Retail Price by HD Size GB") +
  labs(x = "HD Size GB", y = "Retail Price")
```

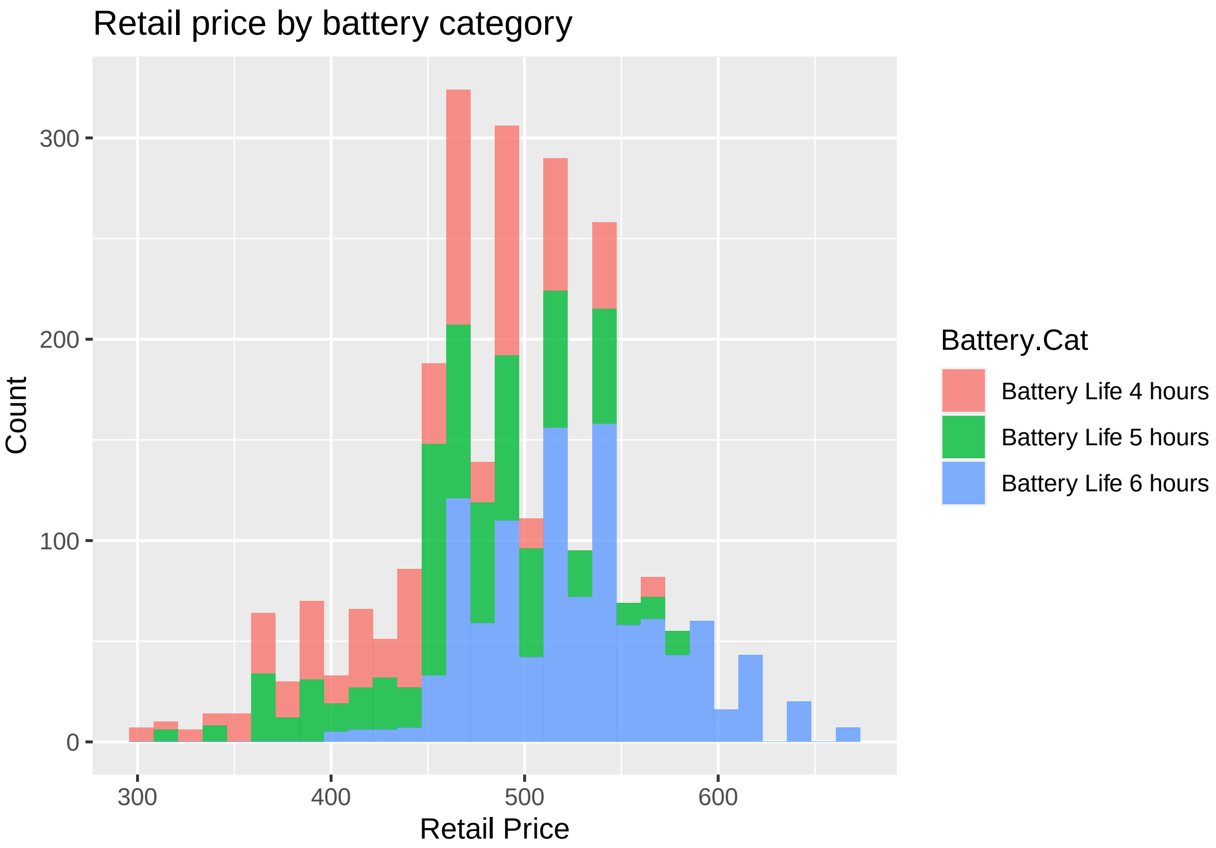


```
### Insights: The four box plot shows that as the HD Size GB become larger the the price will
# increase too. Also, there are couple of outliers in the purple boxplot which is the size 80.

# d)
# part a)
ggplot(data <- my_data, aes(x= Battery.Life..Hours., y = Retail.Price, color = Battery.Life..Hours. )) +
  geom_point() +
  ggtitle("Relationship between Battery life and price") +
  labs(x = "Battery life in hours", y = "Retail Price")
```



```
### Insights: We could see from the graph that as battery life hours increase, so do the price.
# The highest price with 6 hour Battery life is way higher than the highest price with
# 4 hour battery life


# part b)
my_data$Battery.Cat[my_data$Battery.Life..Hours. == 4] <- "Battery Life 4 hours"
my_data$Battery.Cat[my_data$Battery.Life..Hours. == 5] <- "Battery Life 5 hours"
my_data$Battery.Cat[my_data$Battery.Life..Hours. == 6] <- "Battery Life 6 hours"
ggplot(data= my_data, aes(x = Retail.Price, fill = Battery.Cat)) +
  geom_histogram(alpha=0.8) +
  ggtitle("Retail price by battery category") +
  labs(x = "Retail Price", y = "Count" )
```

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



```
### Insights: We could see from the graph that battery life with 6 hours appears more
# in the right side of the graph than 4 hours and 5 hours. This means battery life with
# 6 hours are priced higher than them. Histogram makes the comaprison easier than the scatter plot
```