hw1-Deep-Bhalodia

Loading required packages

```
library(ggplot2)
library(dplyr)
library(rlang)
```

Prob-1

Function to subset a dataset in columns

```
selectCols <- function(data, ...) {
cols <- list(...)
# convert indices to column names
cols <- lapply(cols, function(var) {
  if ( is.numeric(var) ) {
    names(data)[var]
    }
  else {
    var
    }
  })

cols <- unlist(cols) # unlist the list into a character vector
cols <- unique(cols) # unique() removes repeated columns; optional
data[,cols,drop=FALSE]
}

selectCols(mpg, "model", "year")
```

```
## # A tibble: 234 x 2
##    model       year
##    <chr>      <int>
##  1 a4          1999
##  2 a4          1999
##  3 a4          2008
##  4 a4          2008
##  5 a4          1999
##  6 a4          1999
##  7 a4          2008
##  8 a4 quattro  1999
##  9 a4 quattro  1999
## 10 a4 quattro  2008
## # ... with 224 more rows
```

```
selectCols(mpg, 1, 2:3)
```

```
## # A tibble: 234 x 3
##    manufacturer model      displ
##    <chr>        <chr>      <dbl>
```

```
##  1 audi         a4           1.8
##  2 audi         a4           1.8
##  3 audi         a4           2
##  4 audi         a4           2
##  5 audi         a4           2.8
##  6 audi         a4           2.8
##  7 audi         a4           3.1
##  8 audi         a4 quattro   1.8
##  9 audi         a4 quattro   1.8
## 10 audi         a4 quattro   2
## # ... with 224 more rows
```

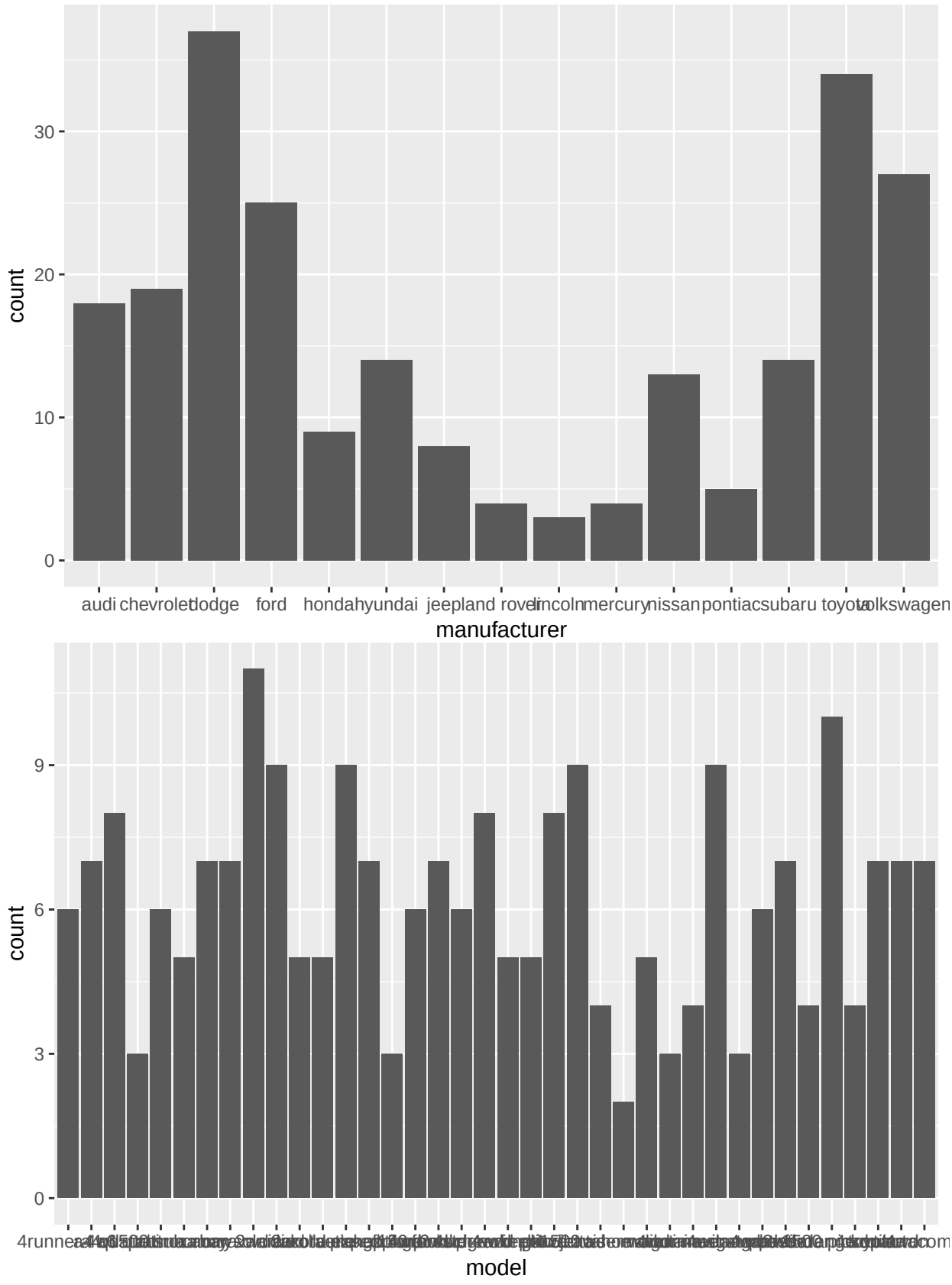
```
selectCols(mpg, 2, "cty", "hwy")
```

```
## # A tibble: 234 x 3
##    model        cty   hwy
##    <chr>      <int> <int>
##  1 a4            18    29
##  2 a4            21    29
##  3 a4            20    31
##  4 a4            21    30
##  5 a4            16    26
##  6 a4            18    26
##  7 a4            18    27
##  8 a4 quattro    18    26
##  9 a4 quattro    16    25
## 10 a4 quattro    20    28
## # ... with 224 more rows
```

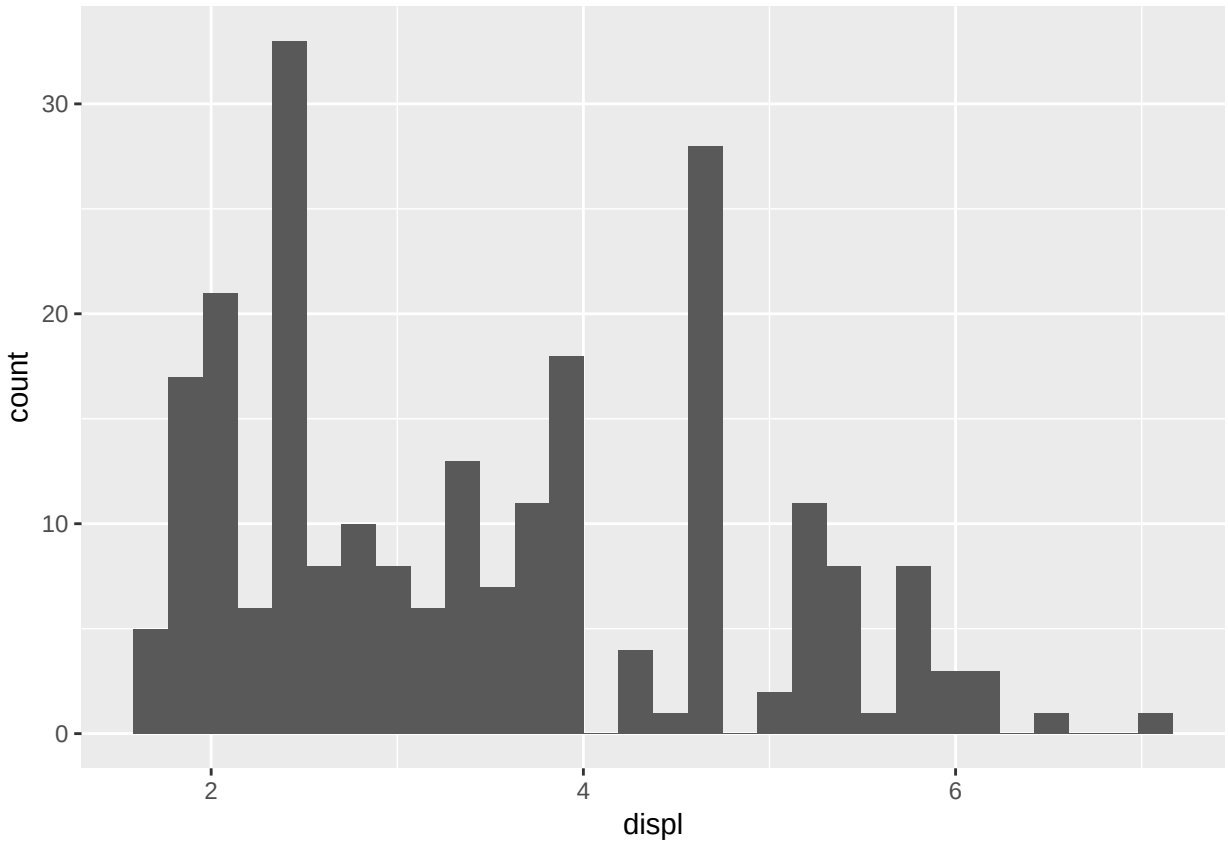
Part A - Problem 2

Write a function of the following form: plotCols(data)

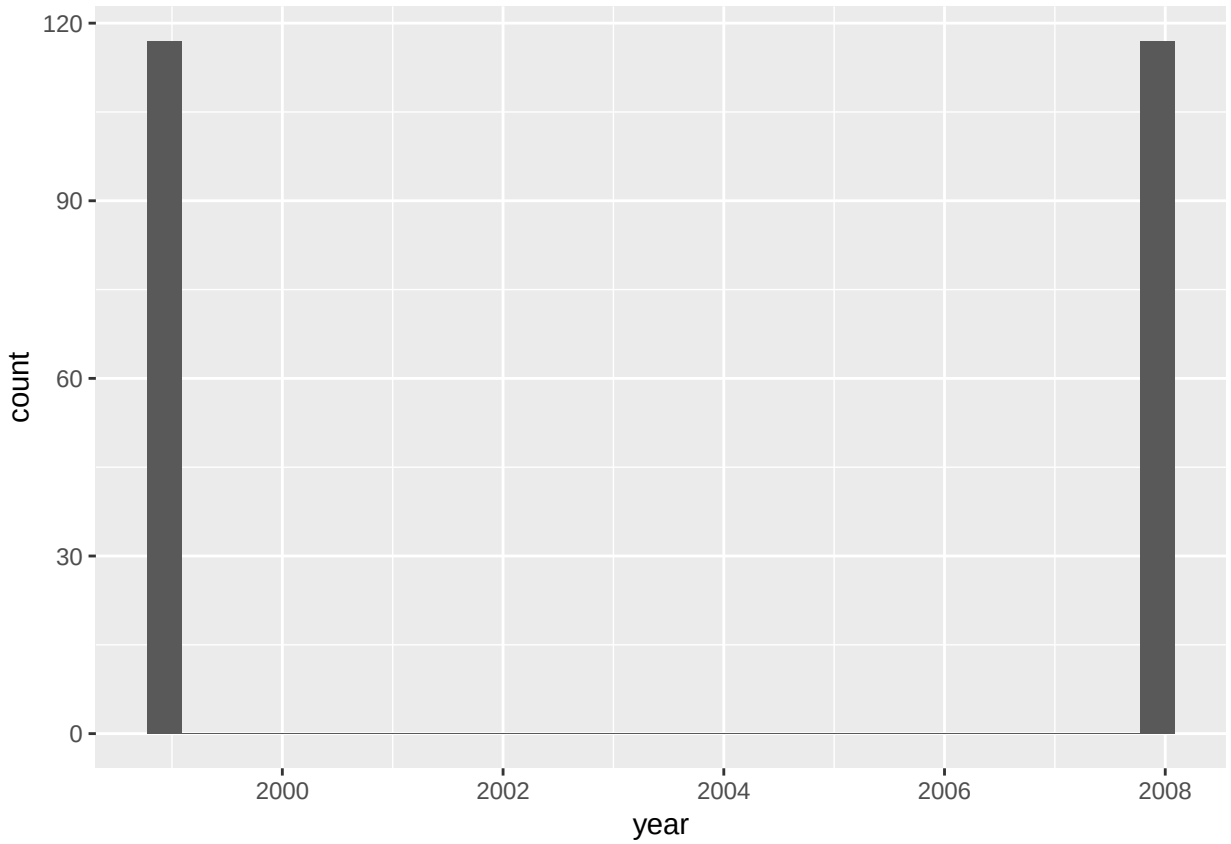
```
plotCols <- function(data) {
  for ( var in names(data) ) {
    if ( is.numeric(data[[var]]) ) {
      print(ggplot(data, aes_string(x=var)) + geom_histogram())
      }
    else {
      print(ggplot(data, aes_string(x=var)) + geom_bar())
    }
  }
}

plotCols(mpg)
```



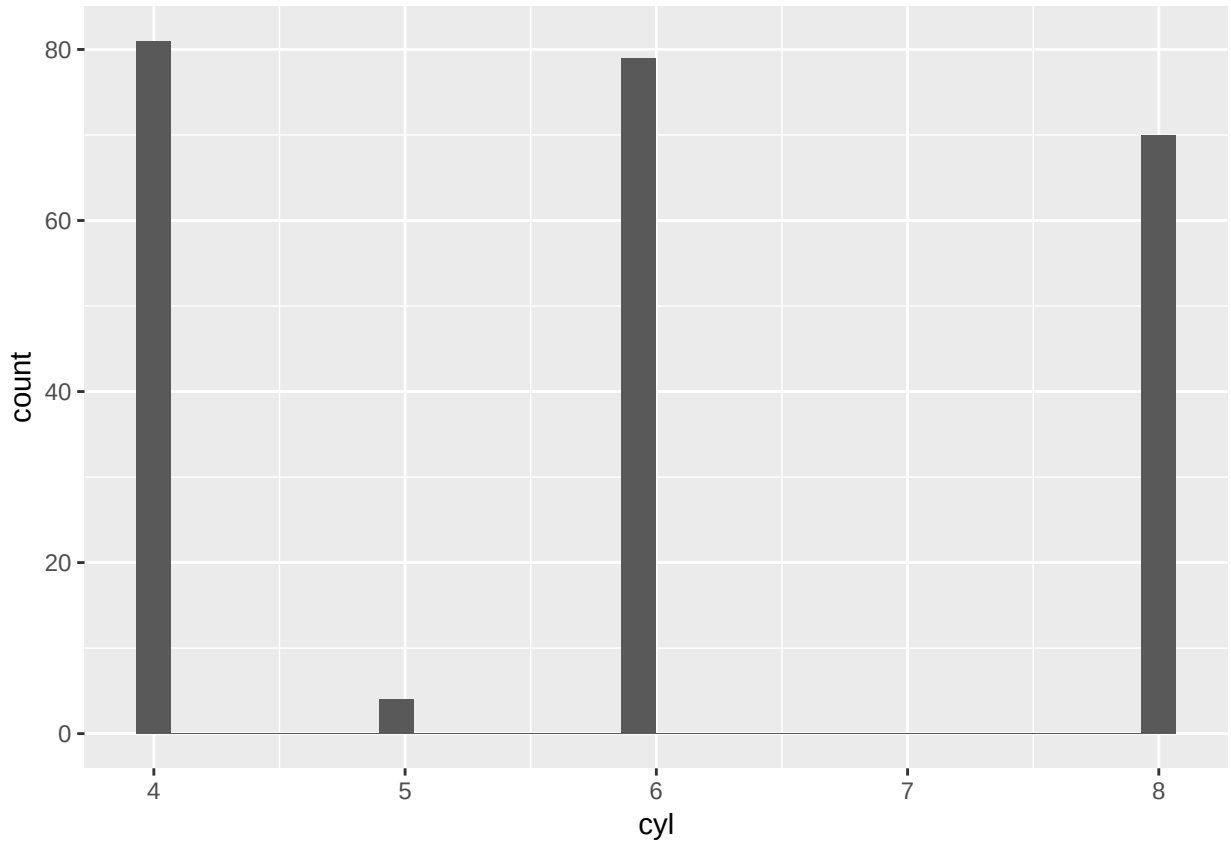

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

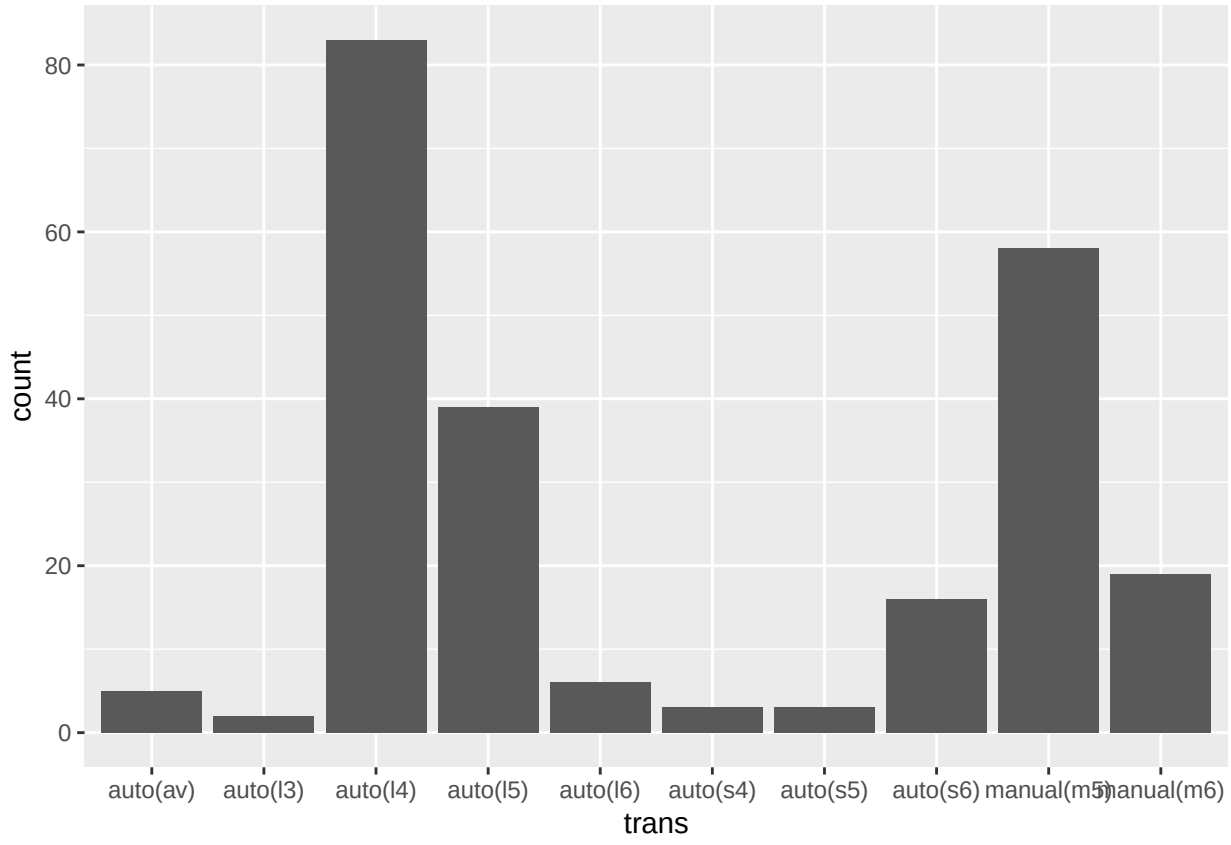


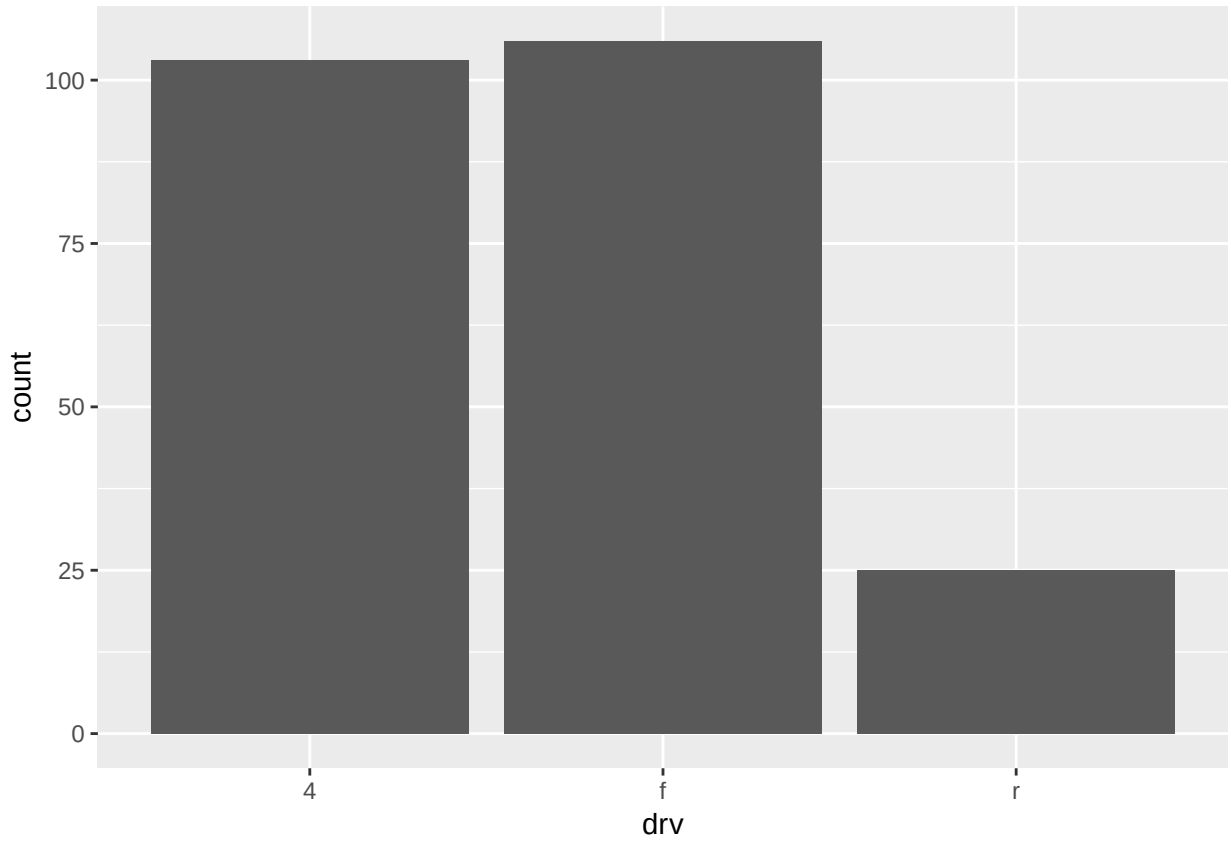

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



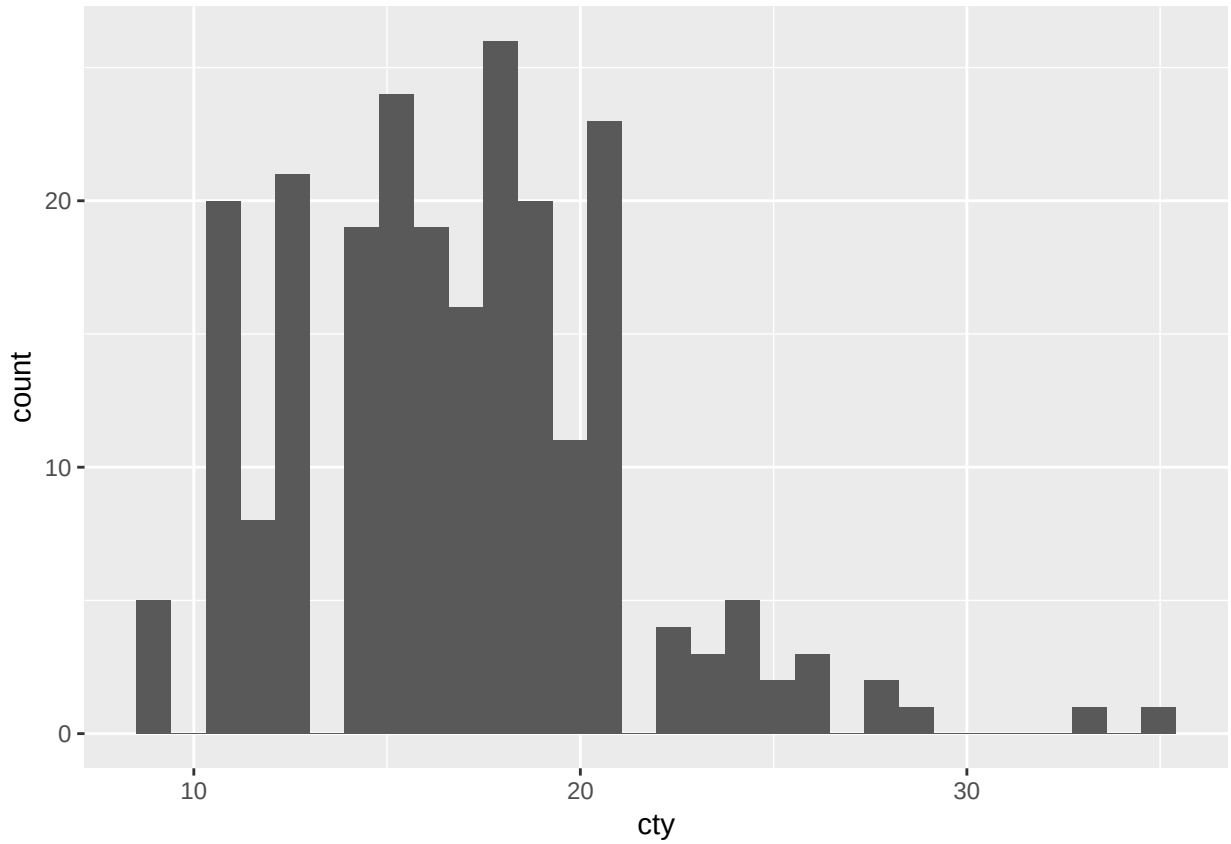

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


80
60
40
20
0
count
4
5
6
7
8
cyl

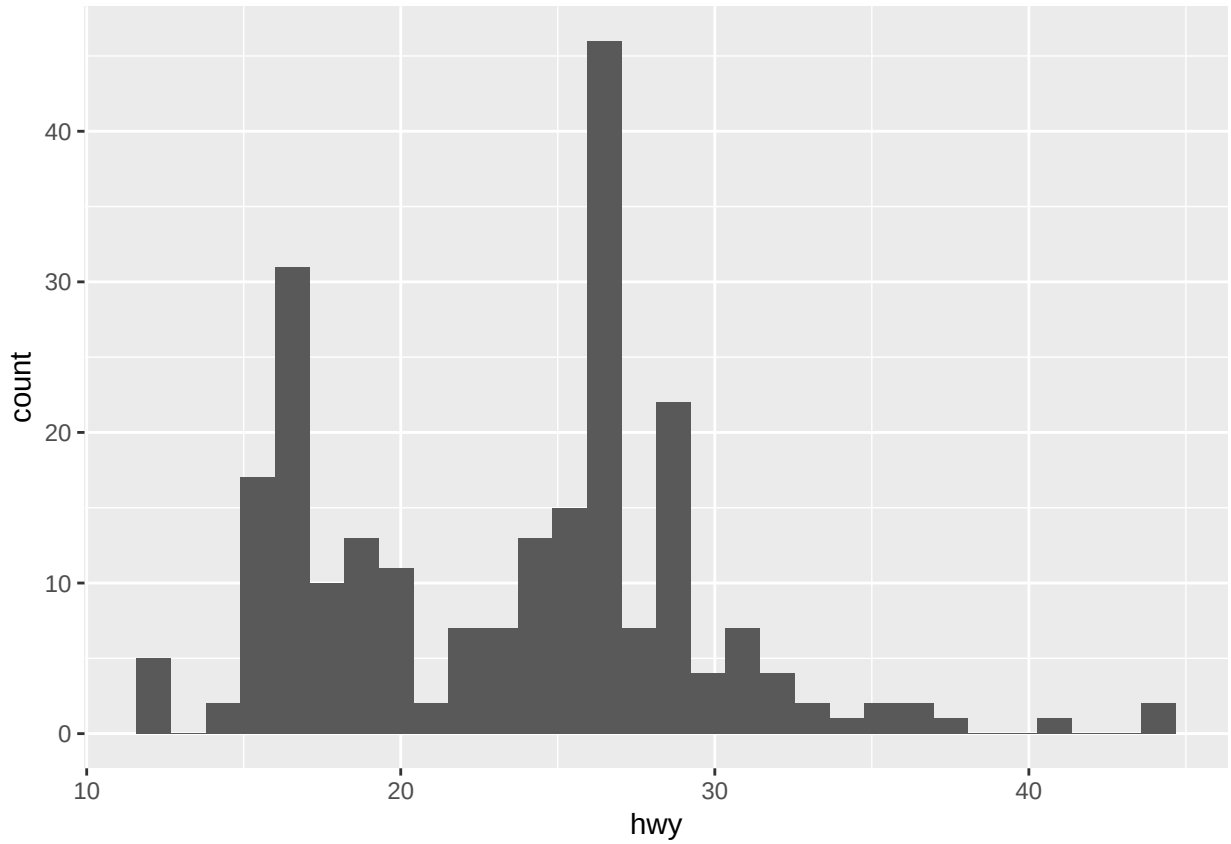

80
60
40
20
0
count
auto(av)
auto(l3)
auto(l4)
auto(l5)
auto(l6)
auto(s4)
auto(s5)
auto(s6)
manual(m5)
manual(m6)
trans

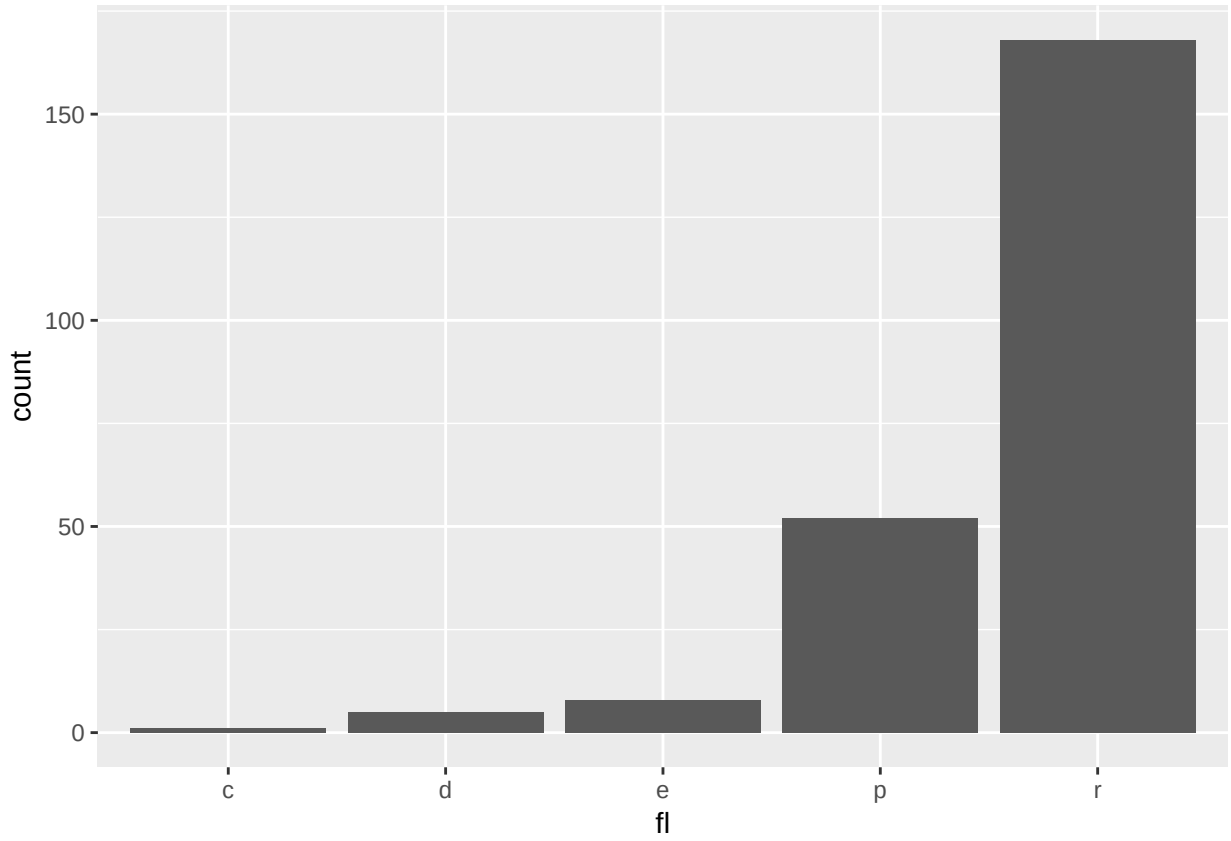


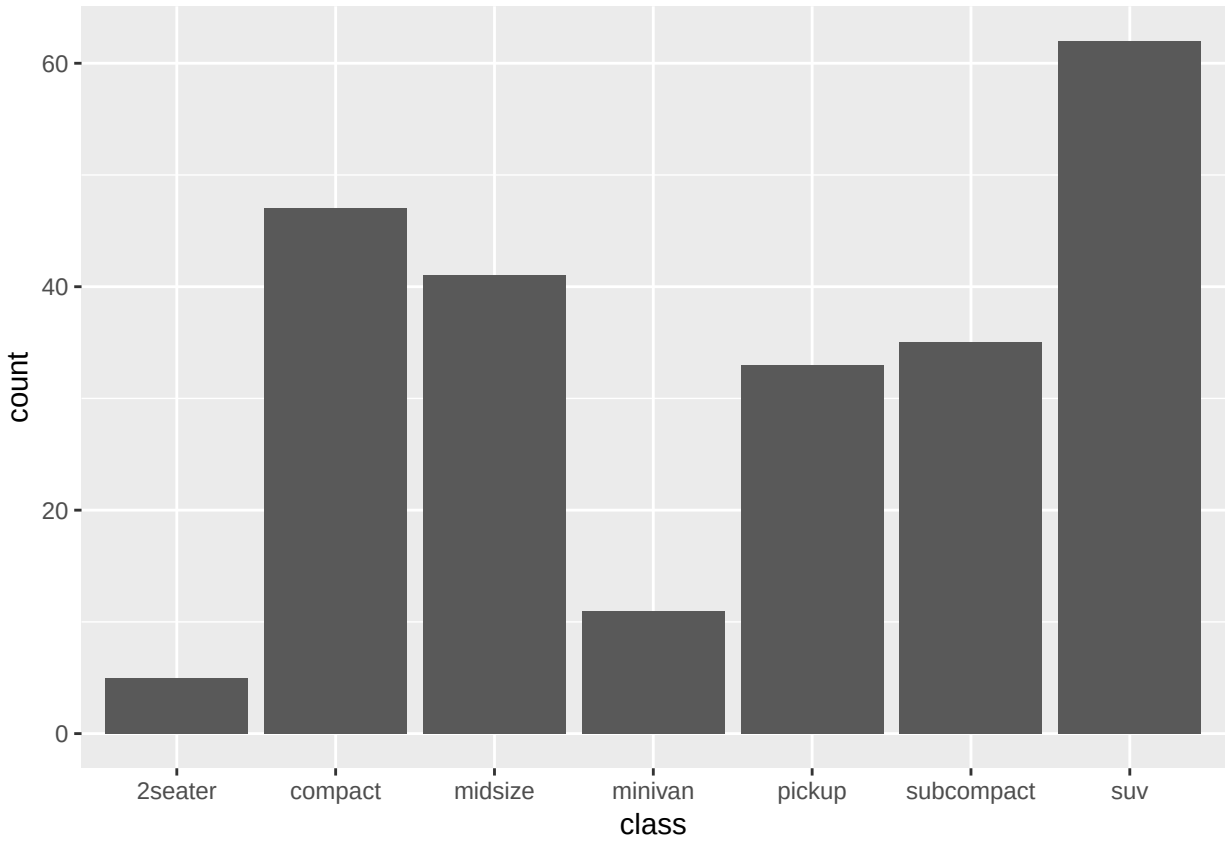

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

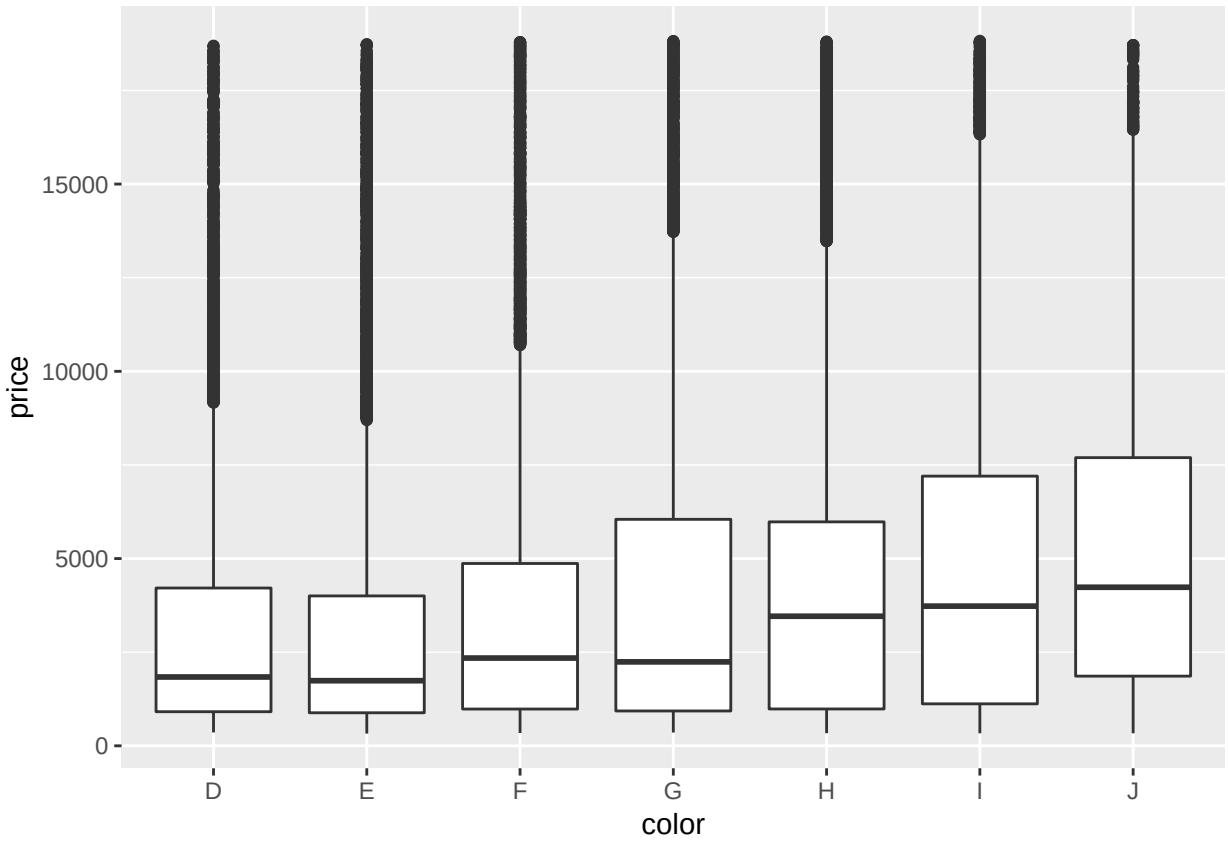

40
30
20
10
0
count
10
20
30
40
hwy


150
100
50
0
count
c
d
e
p
r
fl



Part B - Problem 3

```
#boxplot
ggplot(diamonds, aes(x=color, y=price)) + geom_boxplot()
```



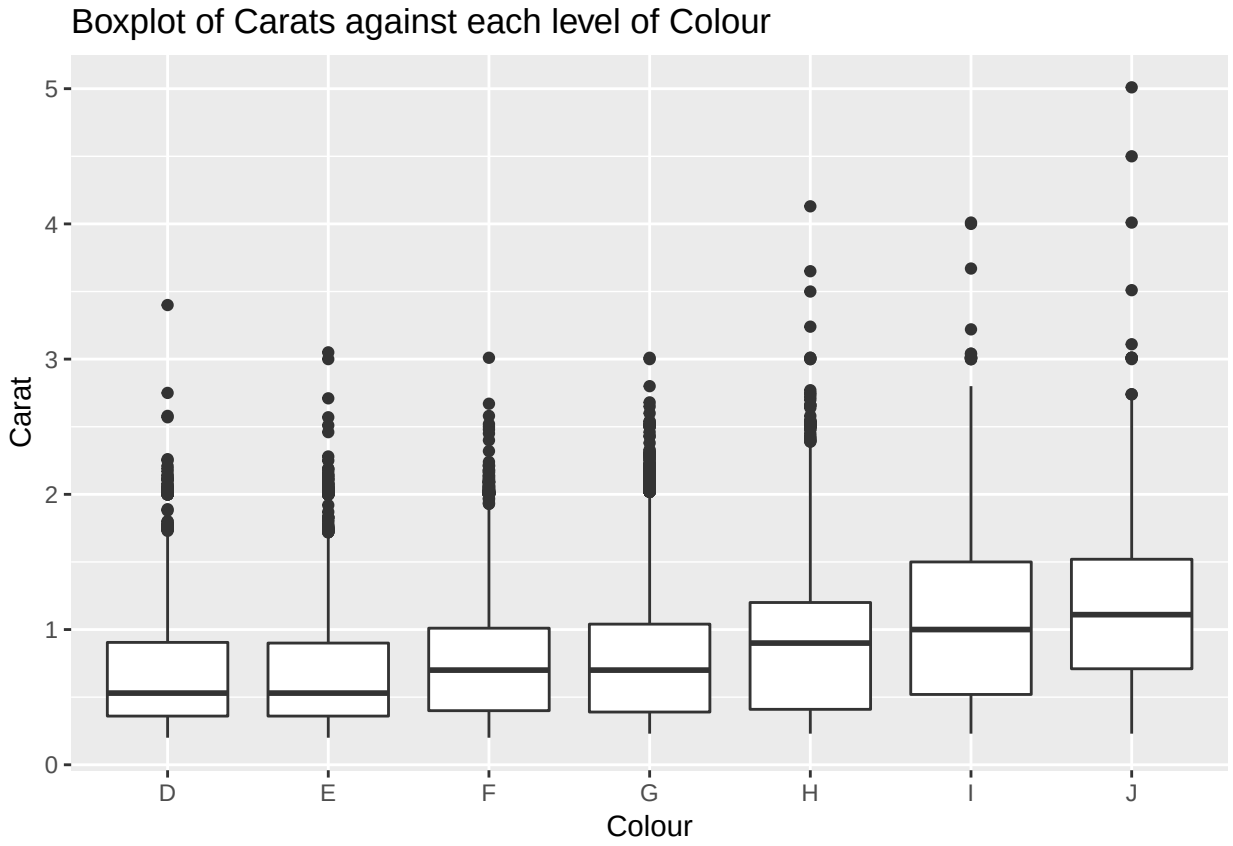
Comments:

The best color is "D" and the worst color is "J", so the worse color diamonds appear to have higher prices on average. This doesn't make sense, as we would expect that better color diamonds should have higher prices.

Part B - Problem 4

Use side-by-side boxplots to visualize the distribution of carat for each level of color. What do you notice about the relationship between carat and color? Could this help make sense of the previous plot?

```
#boxplot
ggplot(diamonds , aes(x=color, y=carat)) + ylab("Carat") + xlab("Colour") + ggtitle("Boxplot of Carats a
```



Comments:

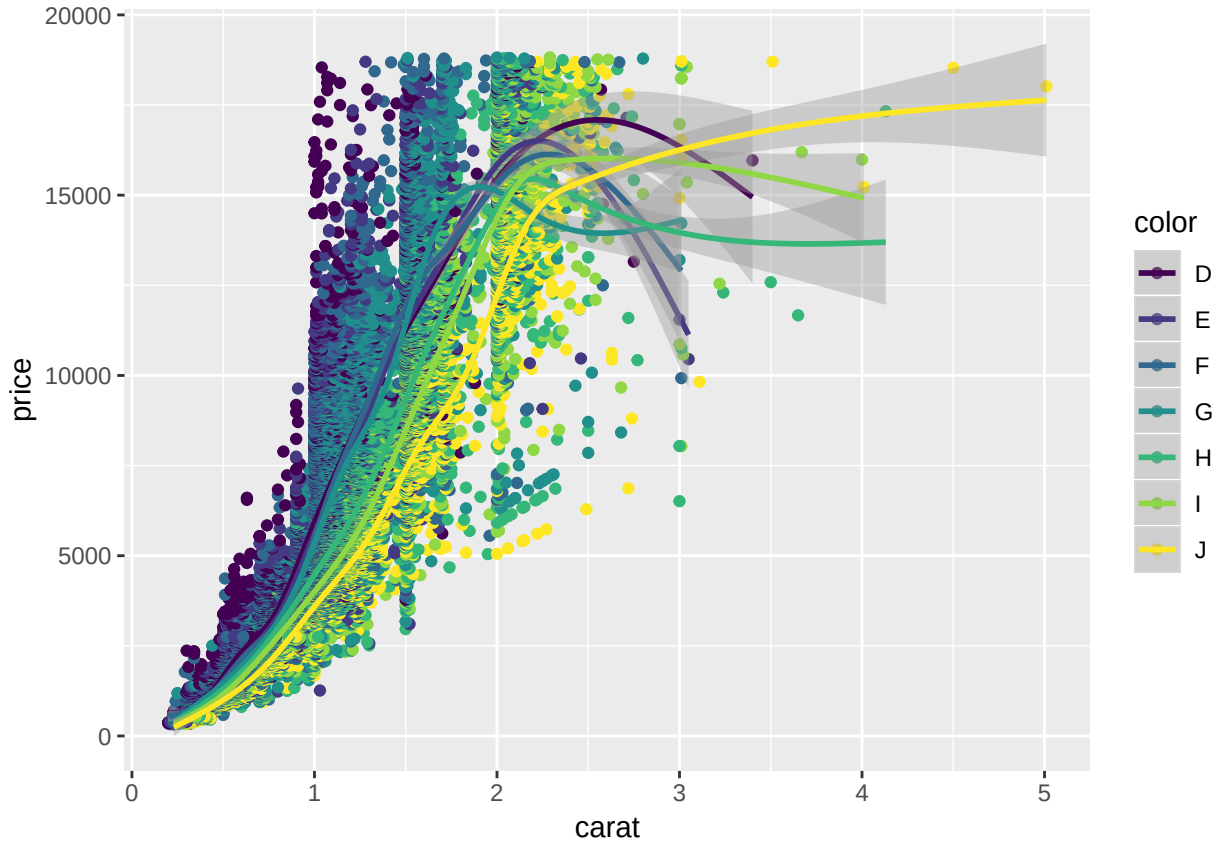
This plot shows that worse color diamonds tend to have larger carat sizes than better color diamonds. If worse color diamonds are larger on average (and we would expect that larger diamonds are more expensive), then it may help explain why worse color diamonds tend to have a higher average price.

Part B - Problem #5

Create a scatter plot of carat versus price, using either an additional aesthetic or faceting to visualize the relationship between carat and price for each level of color. Overlay smooth lines for each level of color. Comment on what you notice about the relationship between carat, price, and color.

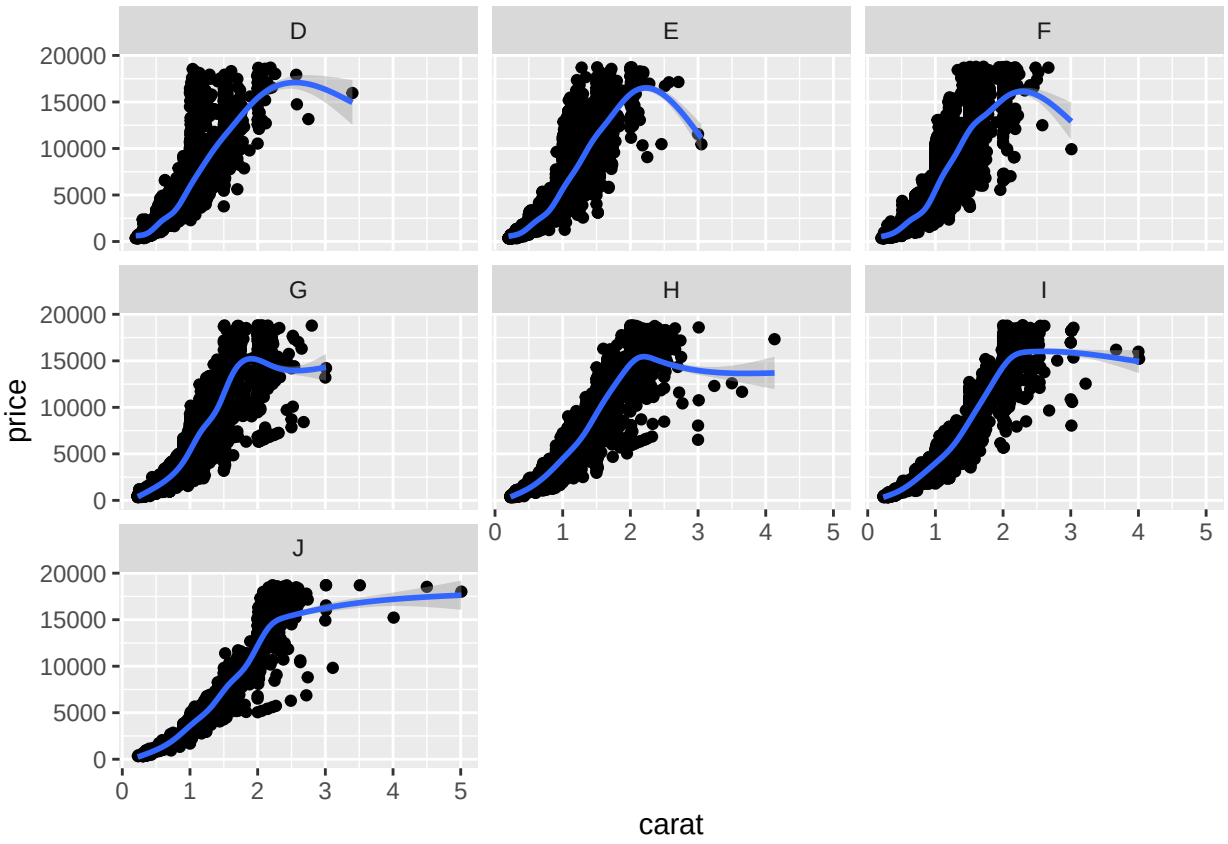
```
ggplot(diamonds, aes(x=carat, y=price, color=color)) +
geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
#Scatter plot
ggplot(diamonds, aes(x=carat, y=price)) +
geom_point() + geom_smooth() + facet_wrap(~color)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Comments:

These plots show a positive relationship between carat and price. Larger diamonds demand higher prices. In addition, the colors (or facets) show that among diamonds of similar size, the better color diamonds tend to be more expensive. However, worse color diamonds tend to be larger, as shown by the separate fitted lines.