Exploratory Data Analysis Using R

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# Load the Motor Trend Car Road Tests (mtcars) dataset
carData = read.csv('mtcars.csv')
```

```
# Create a vector of categorical columns
categorical_cols = c('vs', 'am')

# Convert the columns to factor type
carData[categorical_cols] = lapply(carData[categorical_cols], as.factor)

# Print the structure of the dataframe
str(carData)
```

```
## 'data.frame':    32 obs. of  12 variables:
##  $ X   : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

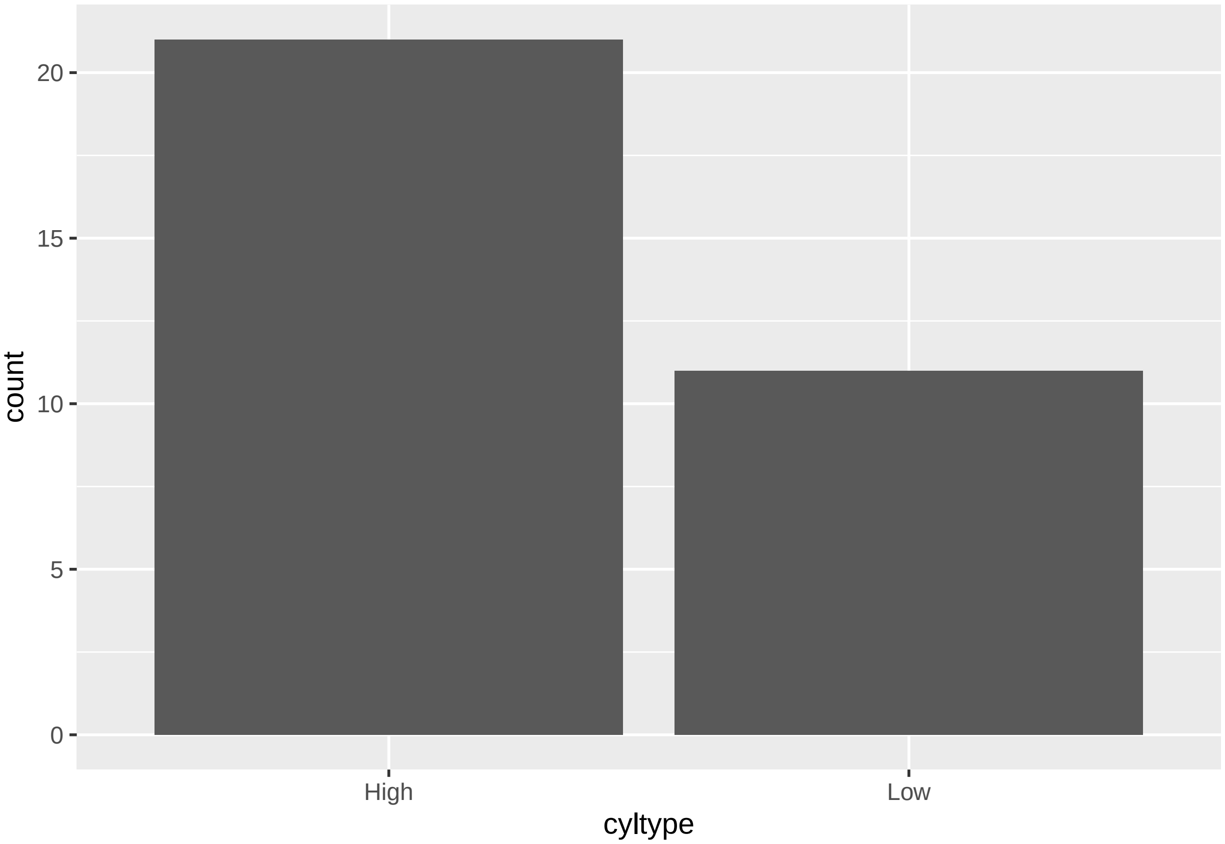
```
# Add a new column called cyltype with value High
# is cyl is greater than 4 and Low otherwise
carData = carData %>% mutate(cyltype = ifelse(cyl > 4, 'High', 'Low'))
head(carData)
```

```
##                   X  mpg cyl disp  hp drat    wt  qsec vs am gear carb cyltype
## 1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4    High
## 2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4    High
## 3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1     Low
## 4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1    High
## 5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2    High
## 6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1    High
```

```r
# Summarize the features
summary(carData)
```

```
##       X                  mpg             cyl             disp
##  Length:32          Min.   :10.40   Min.   :4.000   Min.   : 71.1
##  Class :character   1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8
##  Mode  :character   Median :19.20   Median :6.000   Median :196.3
##                     Mean   :20.09   Mean   :6.188   Mean   :230.7
##                     3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0
##                     Max.   :33.90   Max.   :8.000   Max.   :472.0
##        hp             drat              wt             qsec          vs     am
##  Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   0:19
##  1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1:13
##  Median :123.0   Median :3.695   Median :3.325   Median :17.71
##  Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85
##  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90
##       gear            carb          cyltype
##  Min.   :3.000   Min.   :1.000   Length:32
##  1st Qu.:3.000   1st Qu.:2.000   Class :character
##  Median :4.000   Median :2.000   Mode  :character
##  Mean   :3.688   Mean   :2.812
##  3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :8.000
```

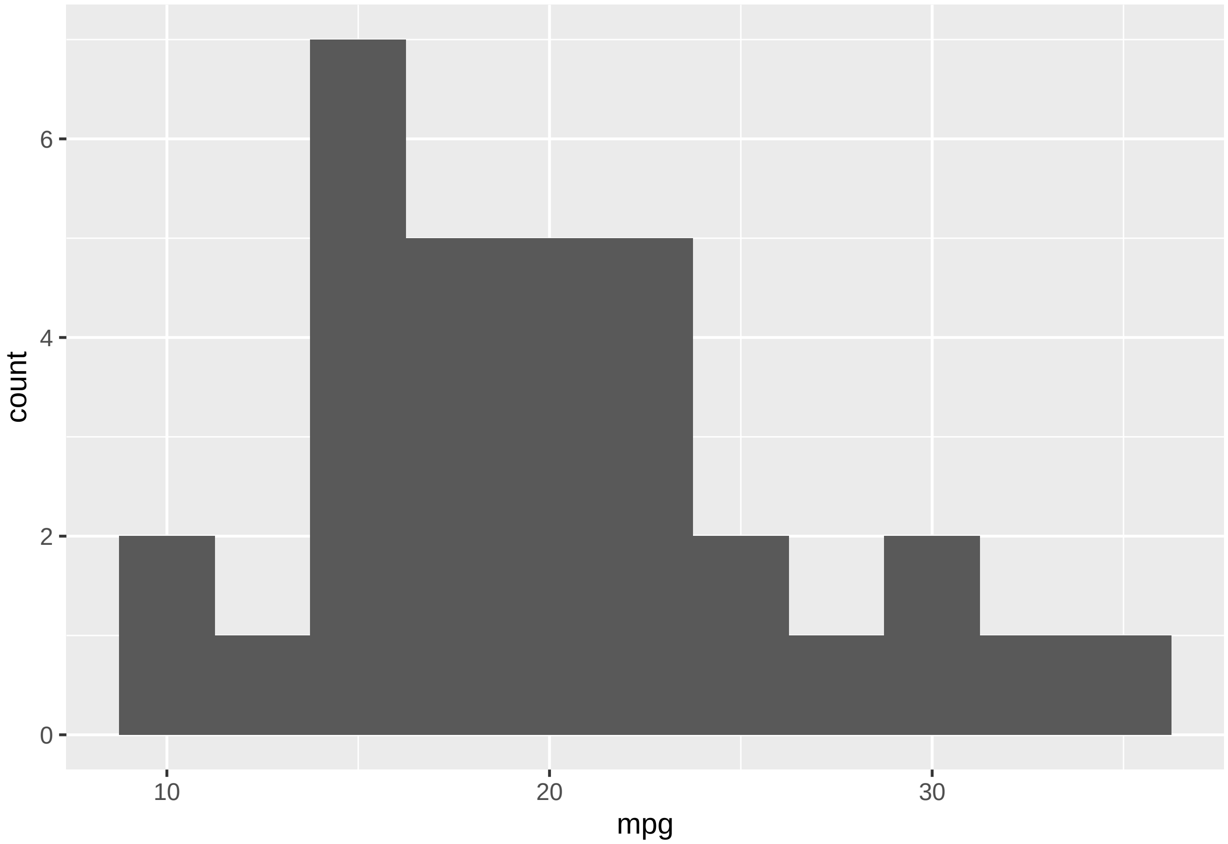
```r
# Visualize distribution of a categorical
# variable using bar chart
ggplot(data = carData) +
  geom_bar(aes(x = cyltype))
```



```
# Count the number of observations in each category
carData %>% count(cyltype)
```

```
##   cyltype  n
## 1    High 21
## 2     Low 11
```

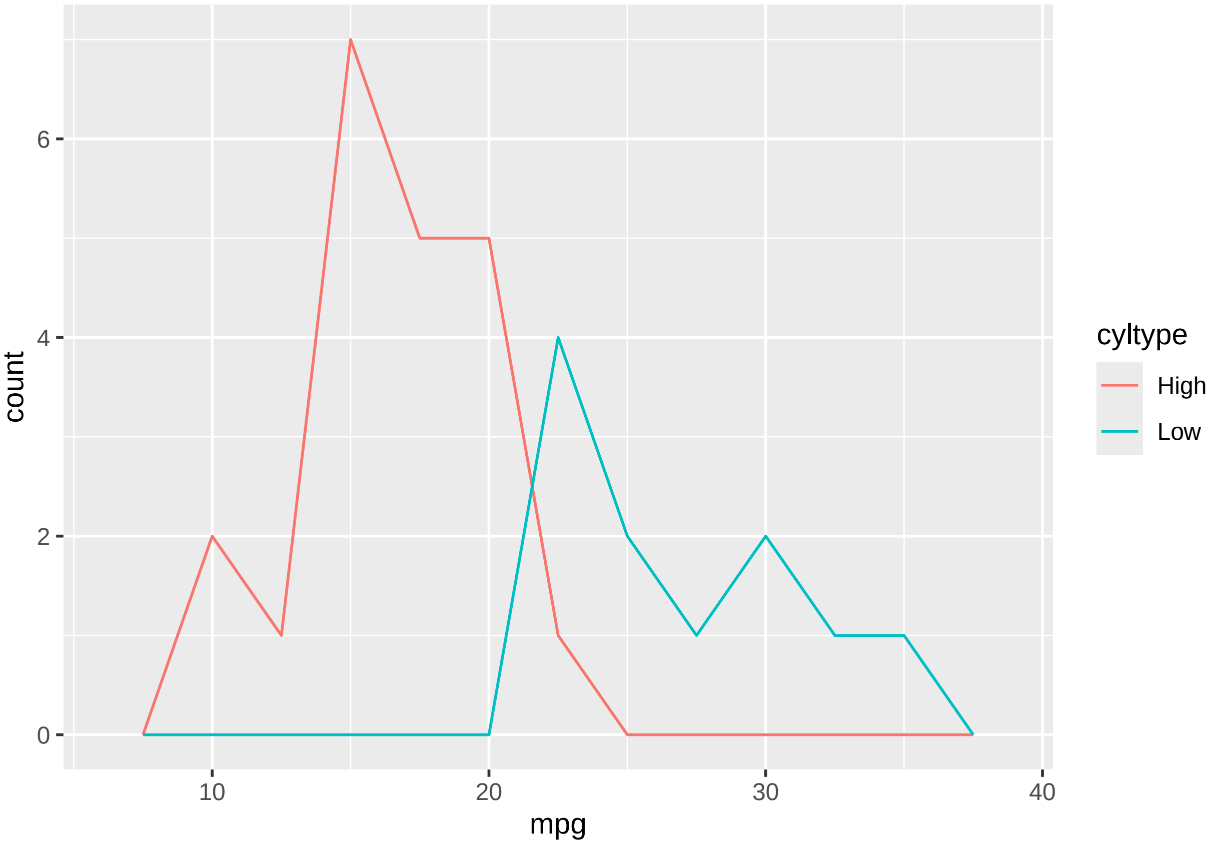
```
# Visualize distribution of a continuous
# variable using histogram
ggplot(data = carData) +
  geom_histogram(aes(x = mpg), binwidth = 2.5)
```



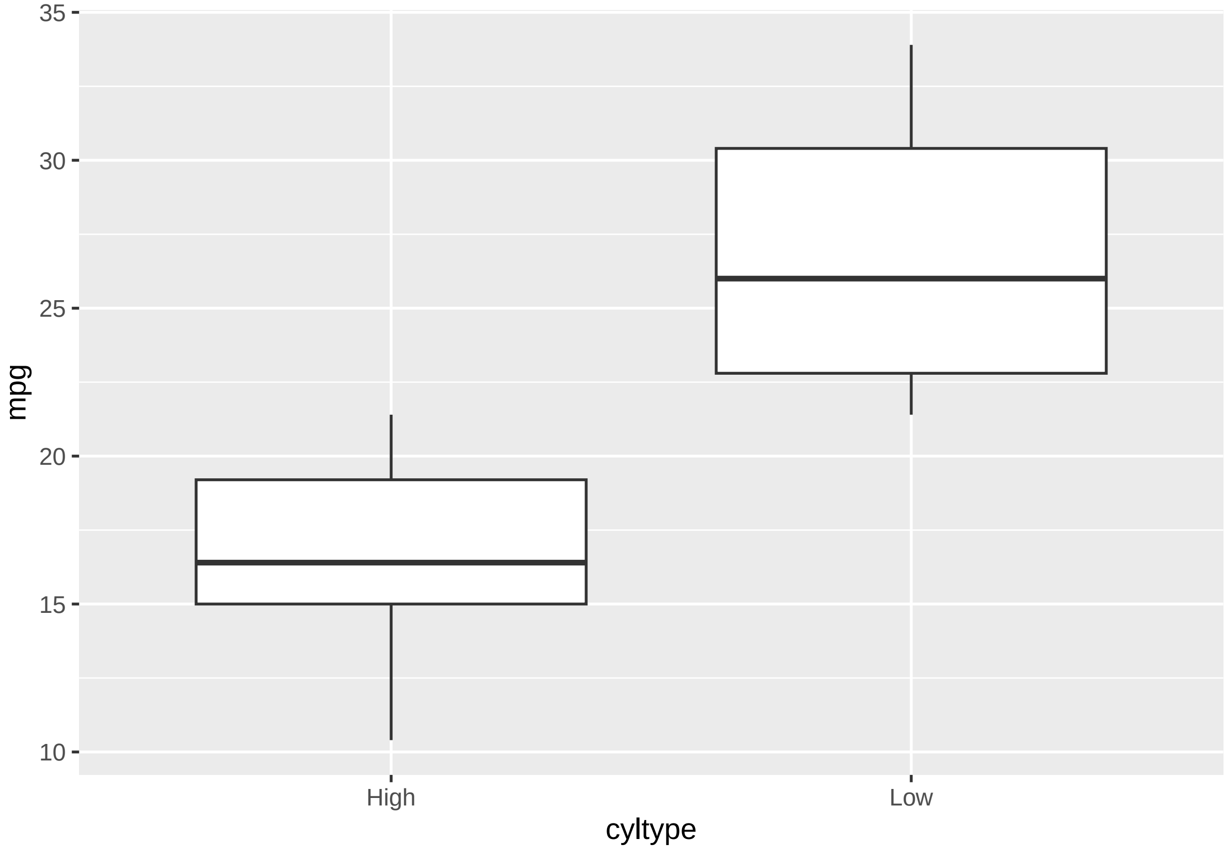
```
# Visualzing the histogram using counts
carData %>%
  count(cut_width(mpg, 2.5))
```

```
##    cut_width(mpg, 2.5) n
## 1          [8.75,11.2] 2
## 2          (11.2,13.8] 1
## 3          (13.8,16.2] 7
## 4          (16.2,18.8] 5
## 5          (18.8,21.2] 5
## 6          (21.2,23.8] 5
## 7          (23.8,26.2] 2
## 8          (26.2,28.8] 1
## 9          (28.8,31.2] 2
## 10         (31.2,33.8] 1
## 11         (33.8,36.2] 1
```

```
# Visualizing multiple histograms
ggplot(data = carData, mapping = aes(x = mpg)) +
  geom_freqpoly(binwidth = 2.5, mapping = aes(colour = cyltype))
```



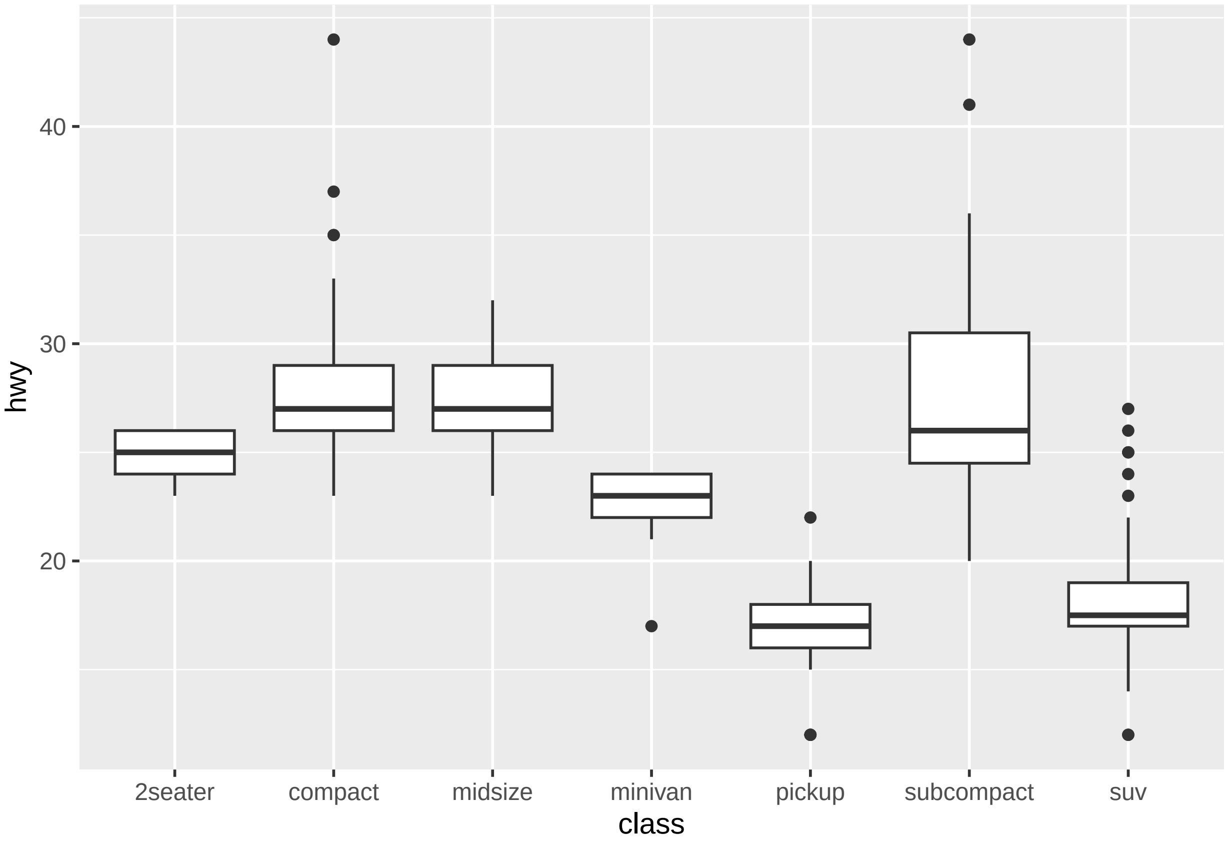
```
# Boxplot to visualize the covariance
# between a continuous and categorical
# feature
ggplot(data = carData, mapping = aes(x = cyltype, y = mpg)) +
  geom_boxplot()
```



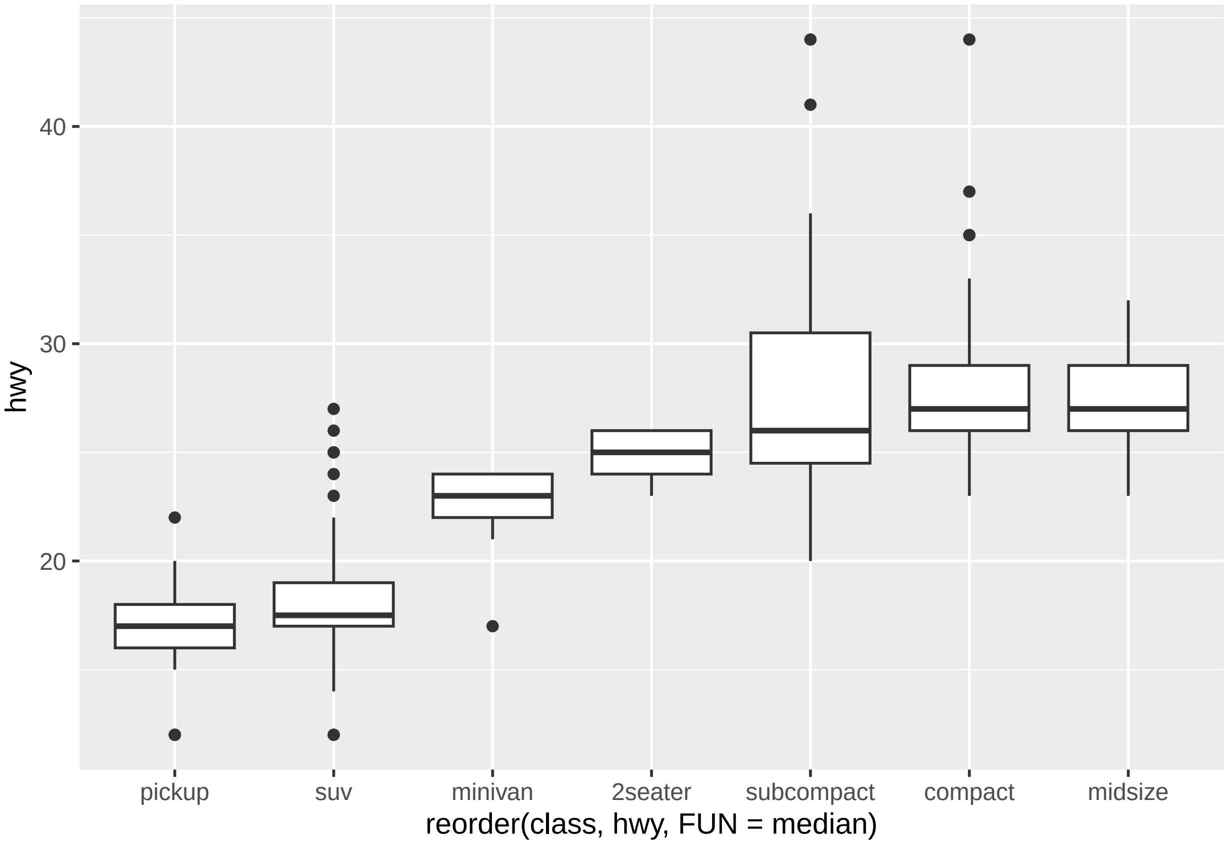
```r
# Load the mpg dataset
data('mpg')
mpgData = mpg
head(mpgData)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

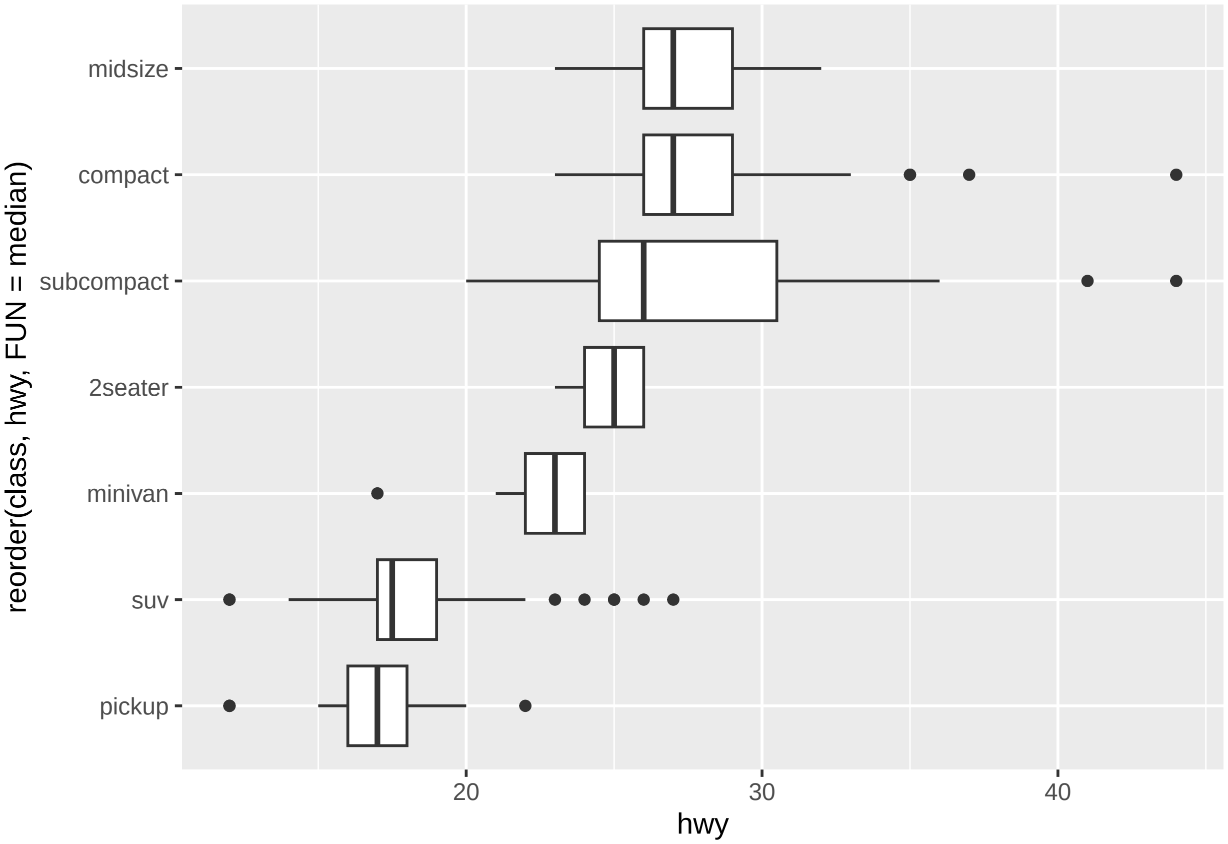
```r
# Boxplot to visualize highway mpg according to
# car type
ggplot(data = mpgData, mapping = aes(x = class, y = hwy)) +
  geom_boxplot()
```



```
# Reorder boxplot according to median
# to visualize the trend
ggplot(data = mpgData, mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy)) +
  geom_boxplot()
```



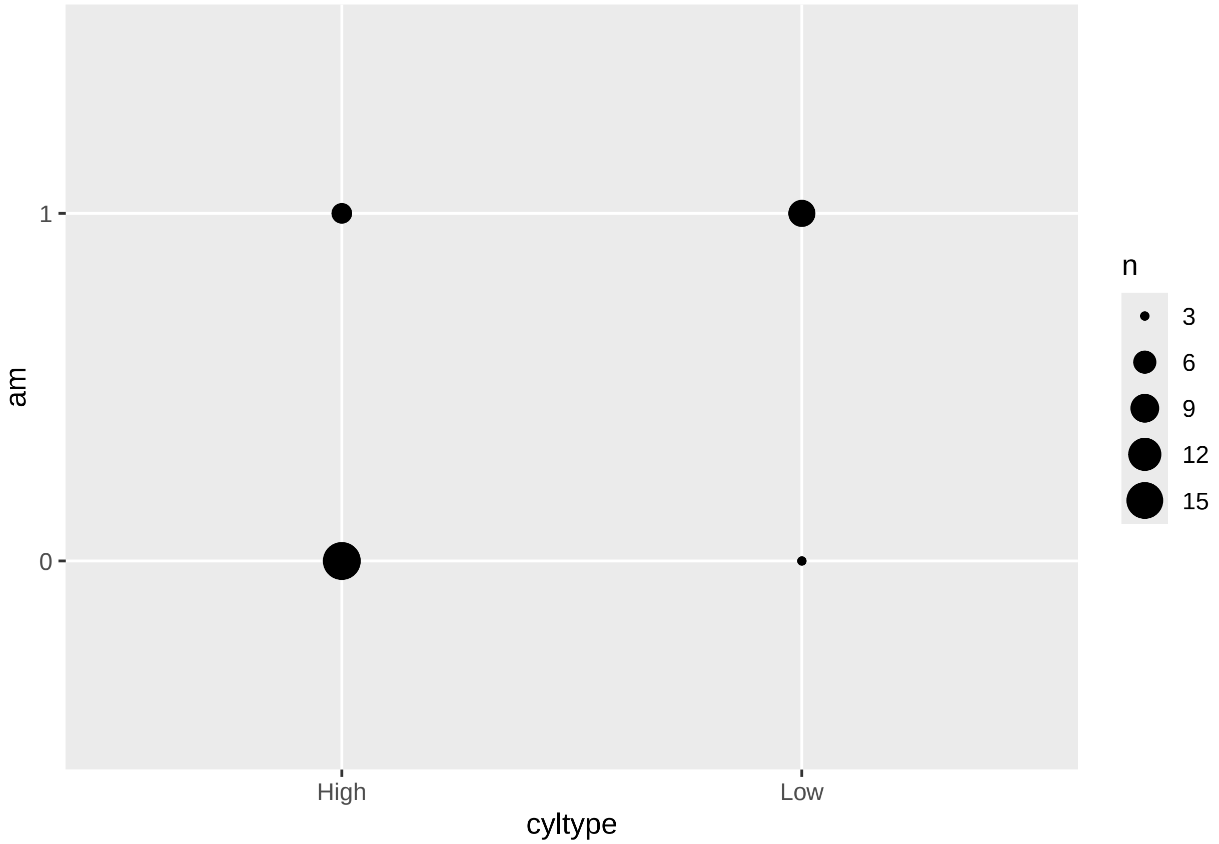
```
# Flip the boxplot for better visualization
ggplot(data = mpgData) +
  geom_boxplot(mapping = aes(x = reorder(class, hwy, FUN = median), y = hwy)) +
  coord_flip()
```



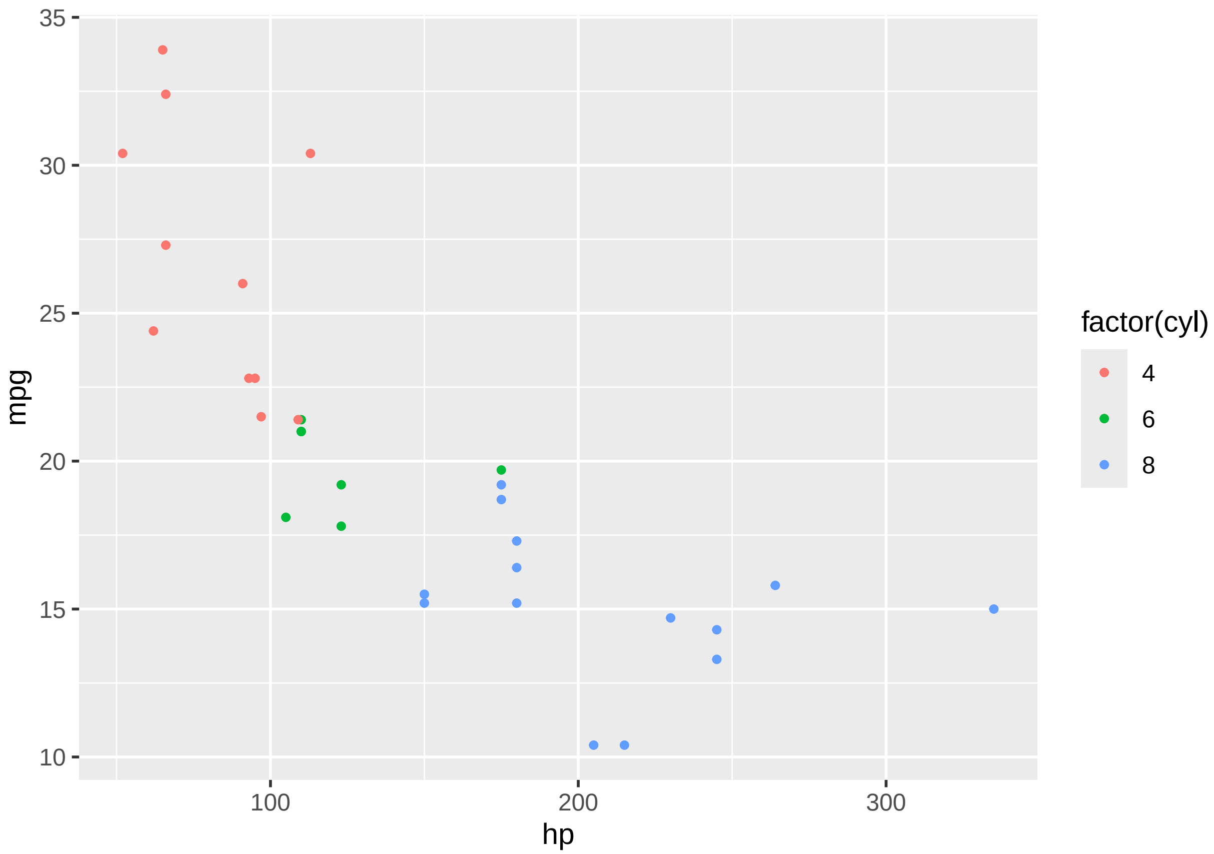
```
# Visualize covariance between two
# categorical features
carData %>% count(cyltype, am)
```

```
##   cyltype am  n
## 1    High  0 16
## 2    High  1  5
## 3     Low  0  3
## 4     Low  1  8
```

```
ggplot(data = carData) +
  geom_count(mapping = aes(x = cyltype, y = am))
```



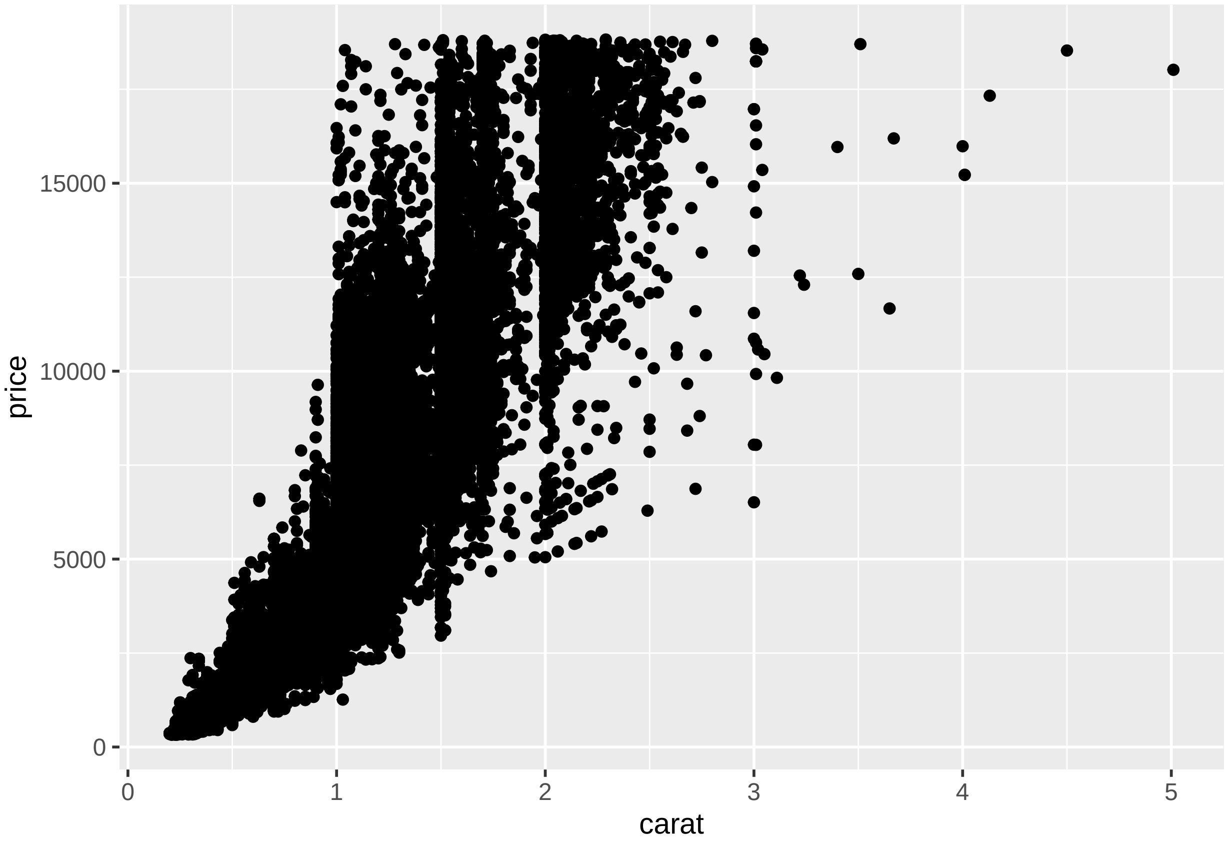
```
# Visualize covariance between two
# continuous features - create a
# scatter plot of mpg vs. HP
ggplot(data = carData, aes(x = hp, y = mpg, color = factor(cyl))) +
  geom_point(size = 1)
```



```
# Load the diamonds dataset
data(diamonds)
head(diamonds)
```

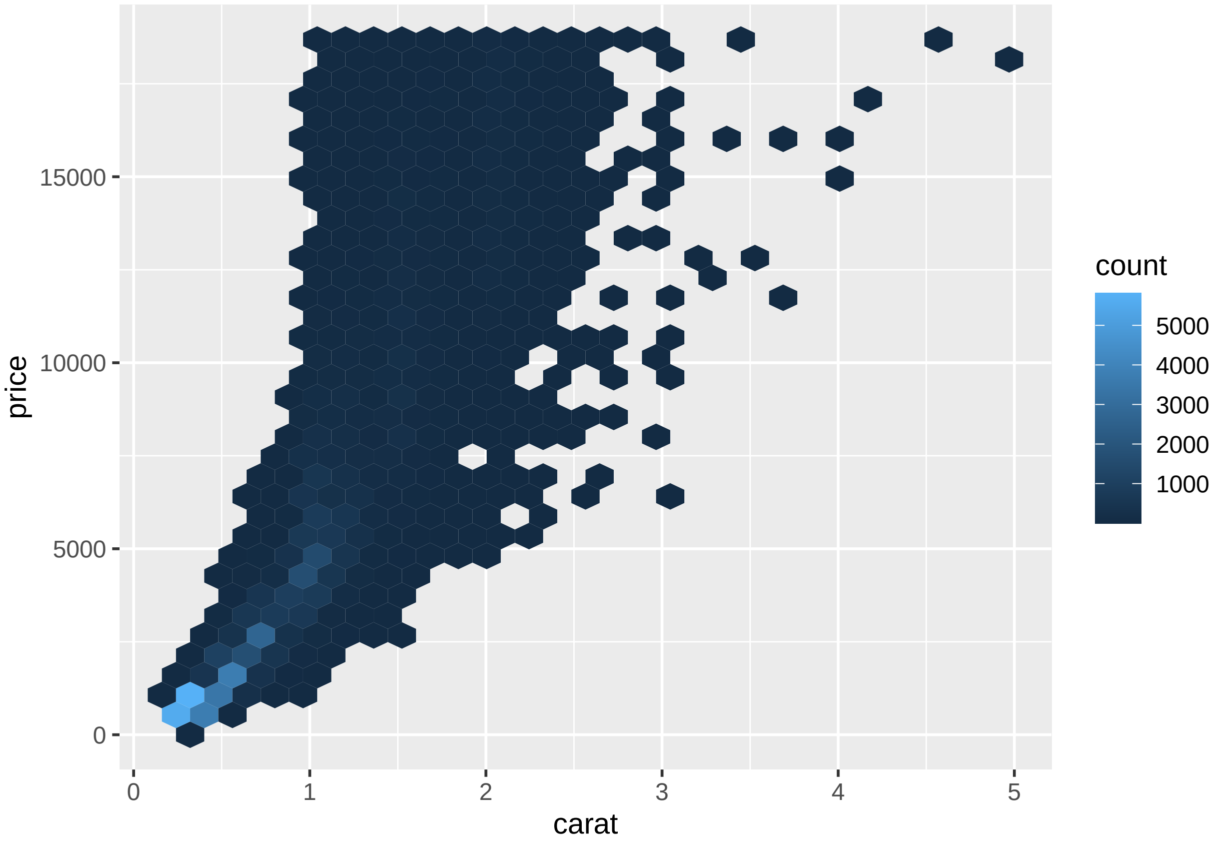
```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

```
# Visualize covariance between two
# continuous features - create a
# scatter plot of carat vs. price
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price))
```



```
# Load the hexbin package
library(hexbin)
```

```
ggplot(data = diamonds) +
  geom_hex(mapping = aes(x = carat, y = price))
```


price
15000
10000
5000
0
0
1
2
3
4
5
carat
count
5000
4000
3000
2000
1000