

Eric Pitman Summer Workshop in Computational Science



5. Visualizing Data



CENTER FOR **COMPUTATIONAL RESEARCH**

UB University at Buffalo
The State University of New York

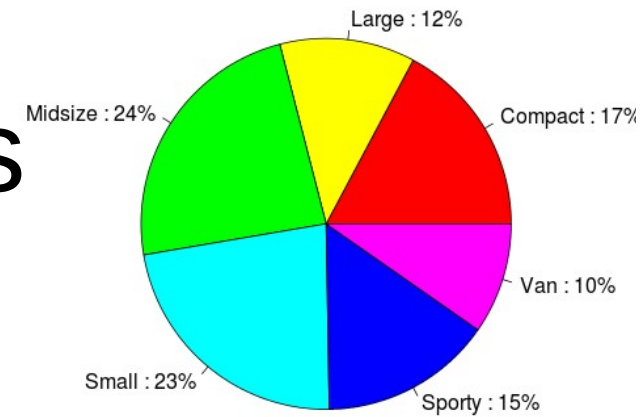
Plotting Data



Plotting is another way to explore a dataset, visually:

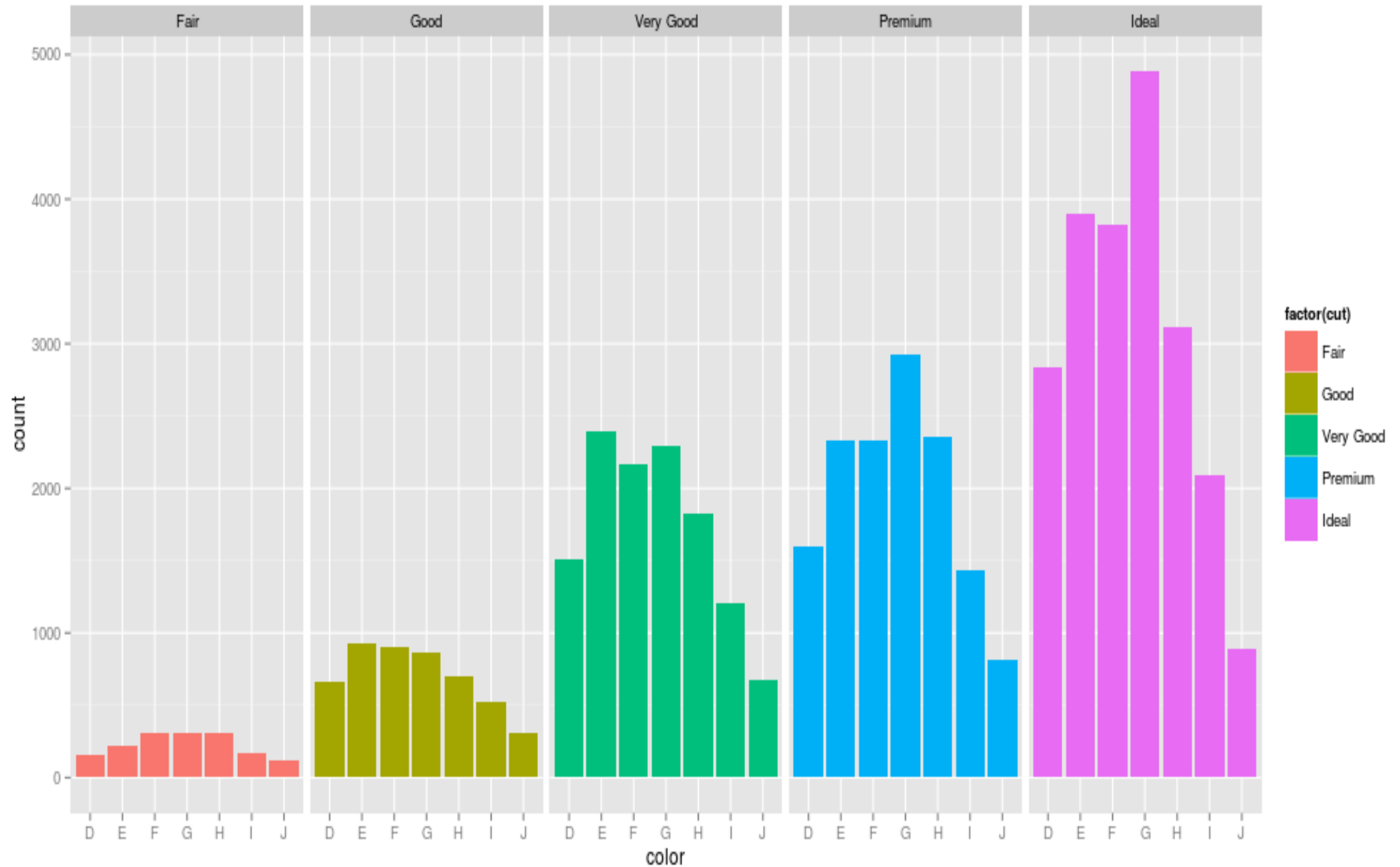
- What's in the dataset?
- What does it mean?
- What if there's *a lot* of it?

Some Plot Types

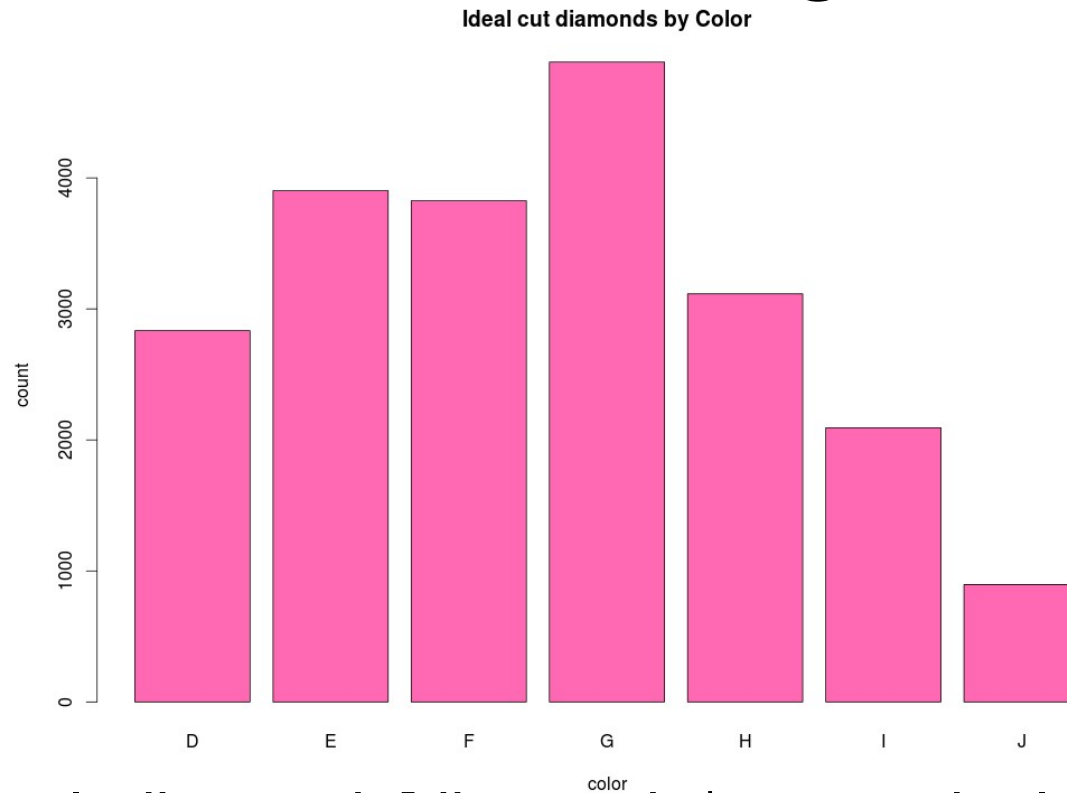


- Pie Chart
 - Display proportions of different values for a variable
- Bar Plot
 - Display counts of values for a categorical variable
- Histogram, Density Plot
 - Display counts of values for a binned, numeric variable
- Scatter Plot
 - y vs. x
- Box Plot
 - Display distributions over different values of a variable

Barplot: Counts of Categorical Values



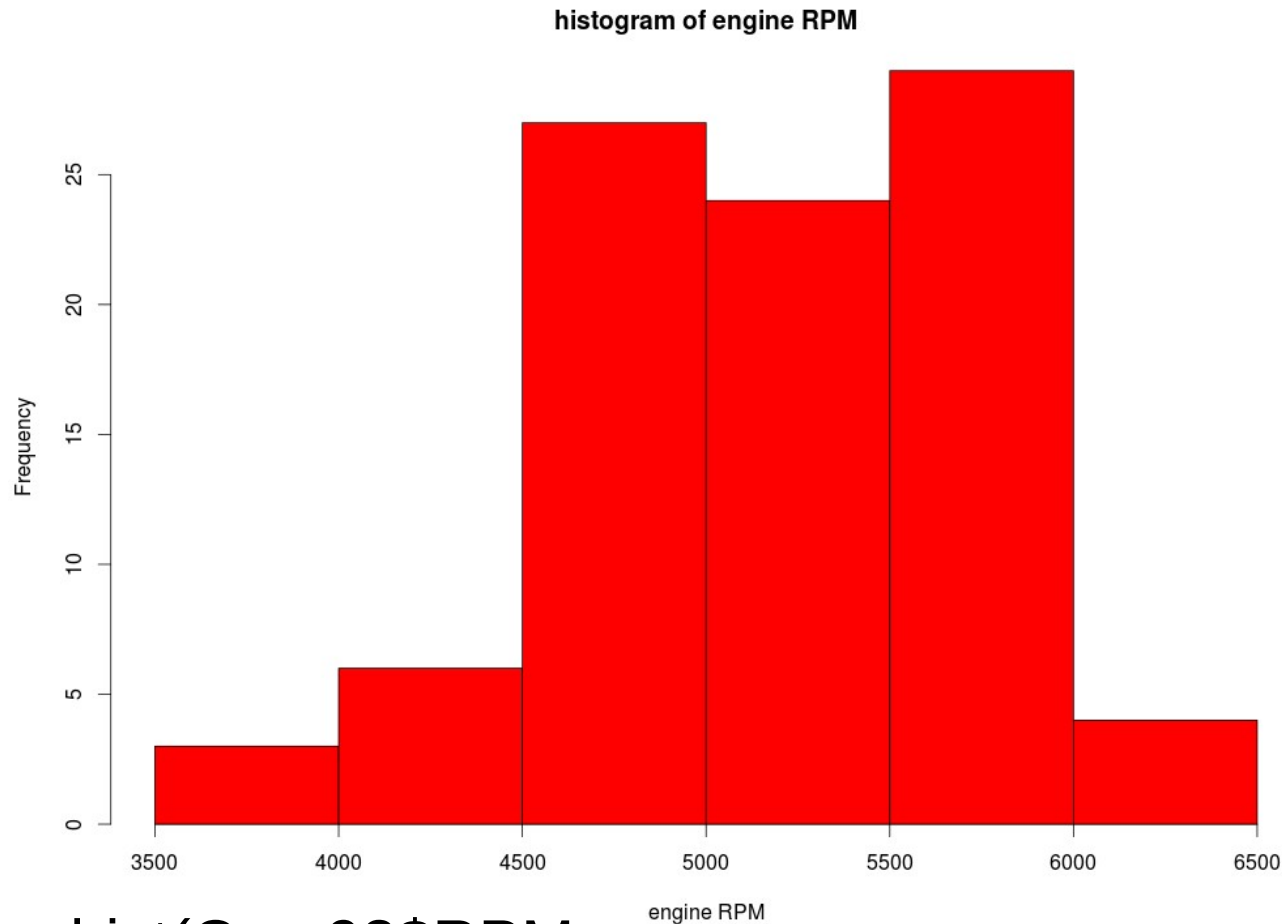
Barplot: Counts of Categorical Values



```
ideal=diamonds[diamonds$cut=="Ideal","color"]
```

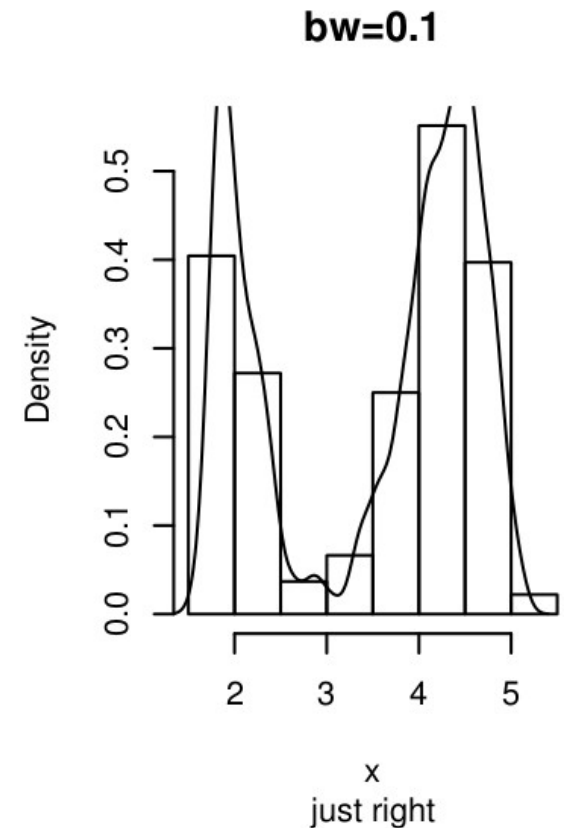
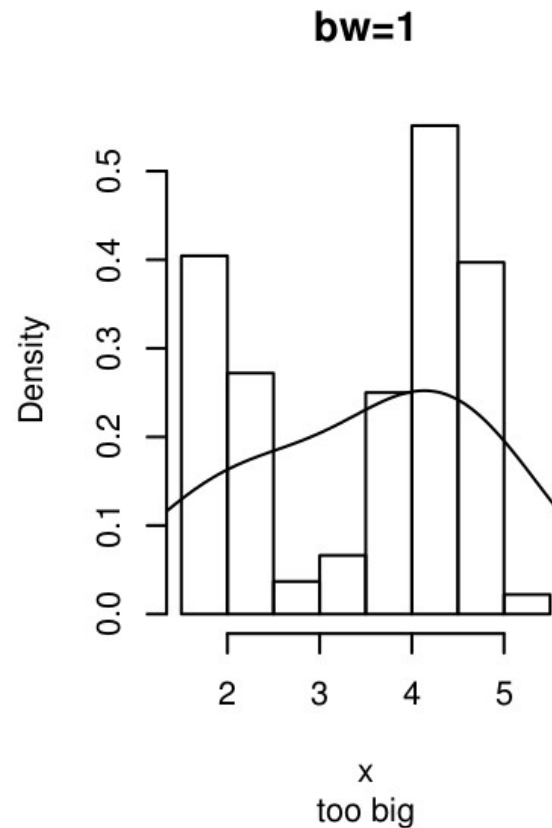
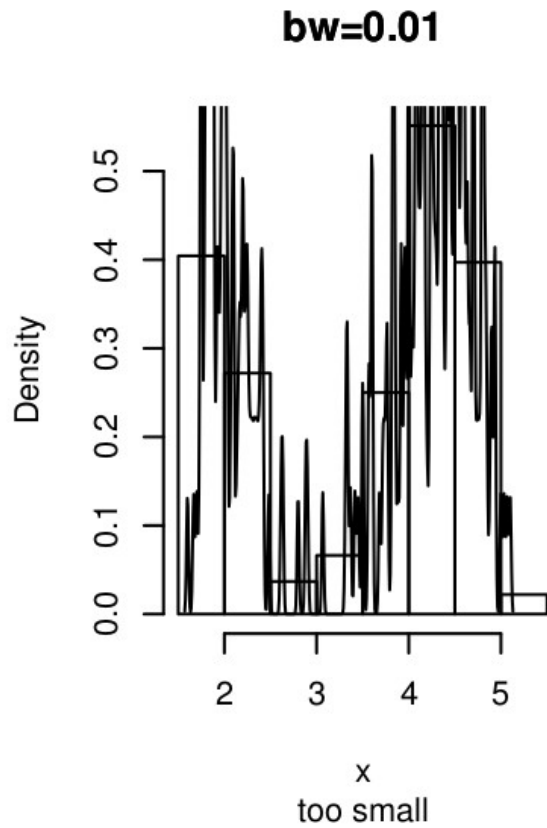
```
barplot(table(ideal),  
        xlab="color",  
        ylab="count",  
        main="Ideal cut diamonds by Color",  
        col="hotpink")
```

Histogram: Frequencies of Numeric Values

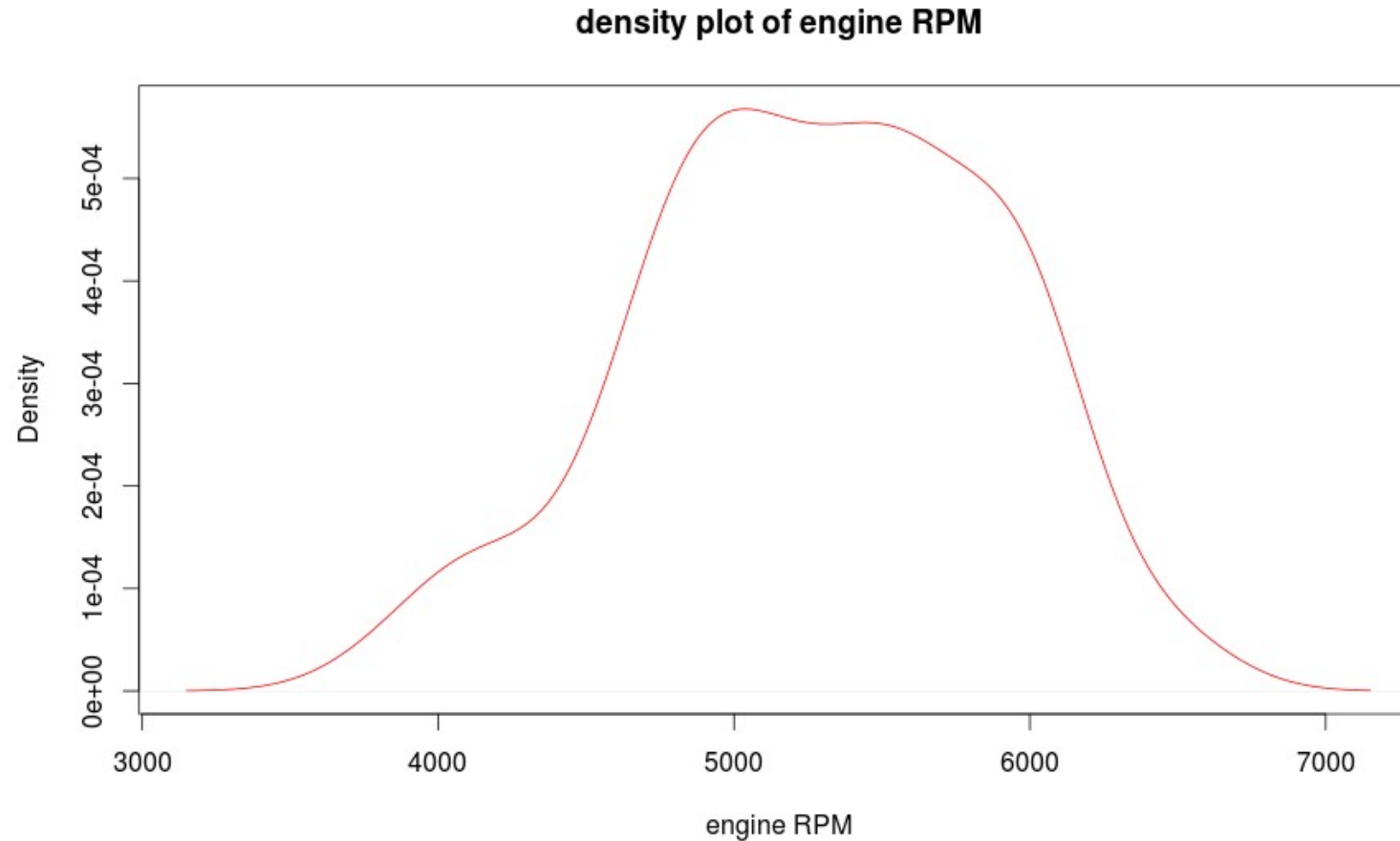


```
hist(Cars93$RPM,  
      xlab="engine RPM",  
      main="histogram of engine RPM",  
      col="red")
```

Histogram and Density Binning



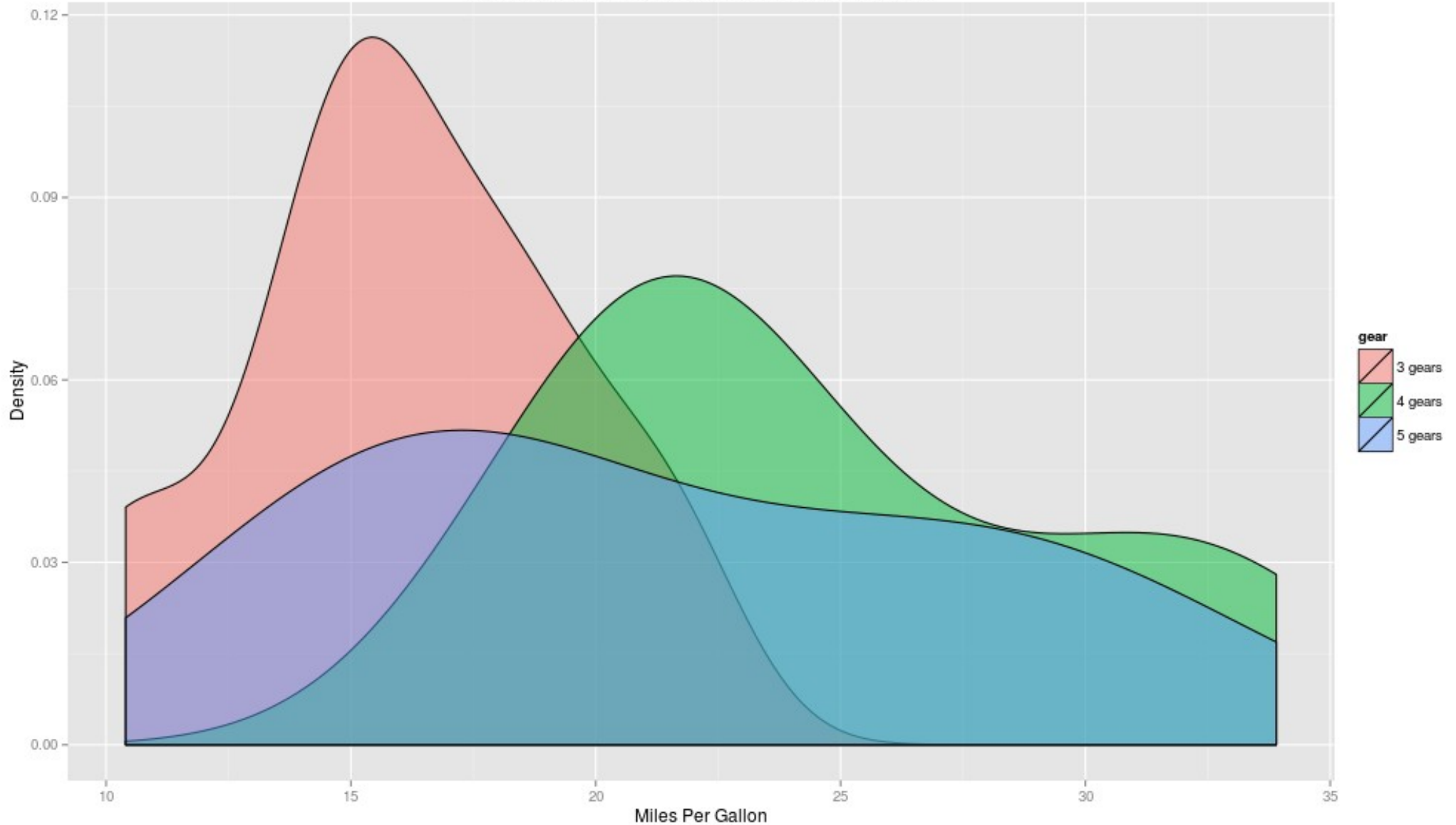
Kernel Density Plot



```
plot(density(Cars93$RPM),  
     xlab="engine RPM",  
     main="density plot of engine RPM",  
     col="red")
```


Density Plot

Distribution of Gas Mileage with Number of Gears



Scatterplot: Numeric Data

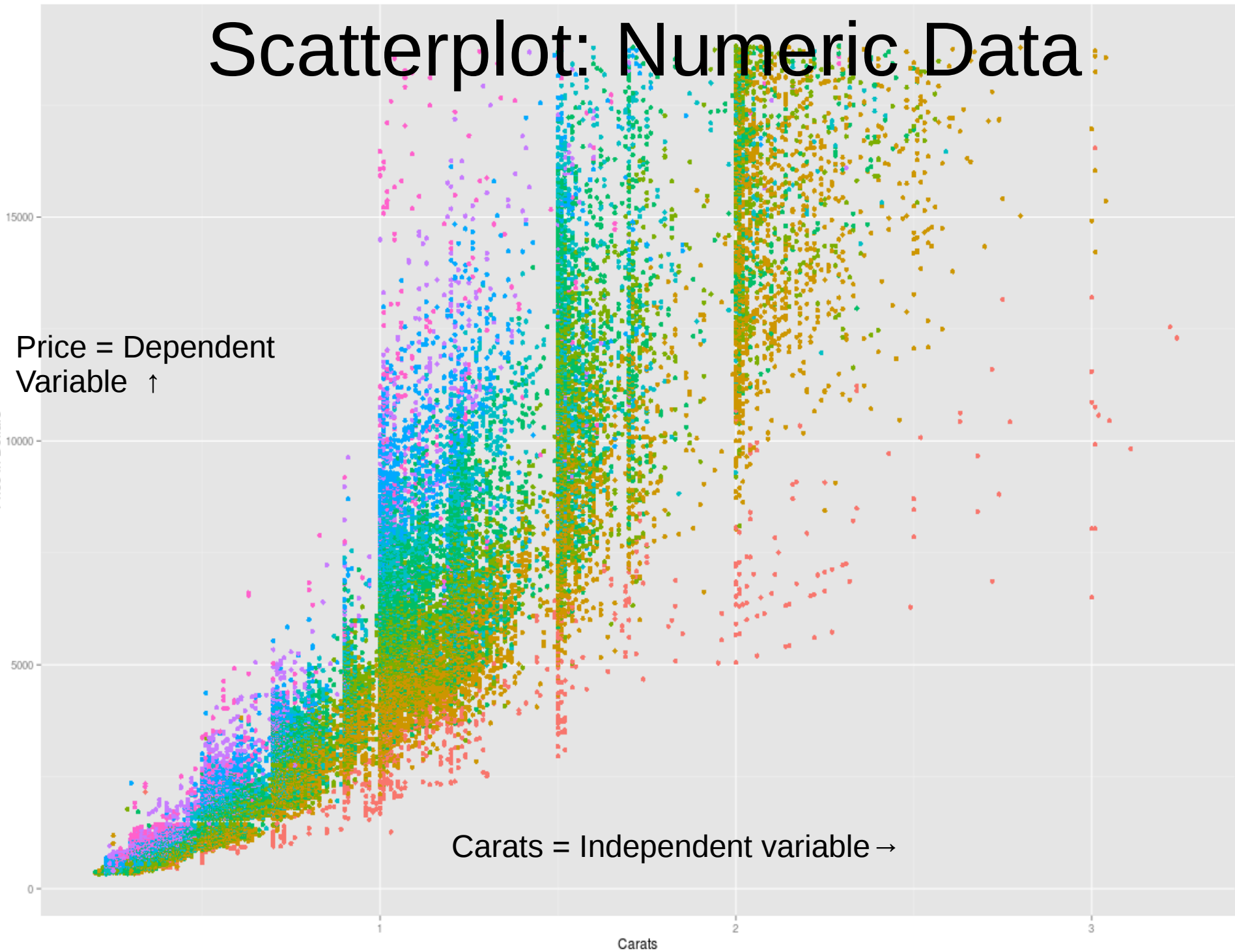
Price in Dollars

Price = Dependent Variable ↑

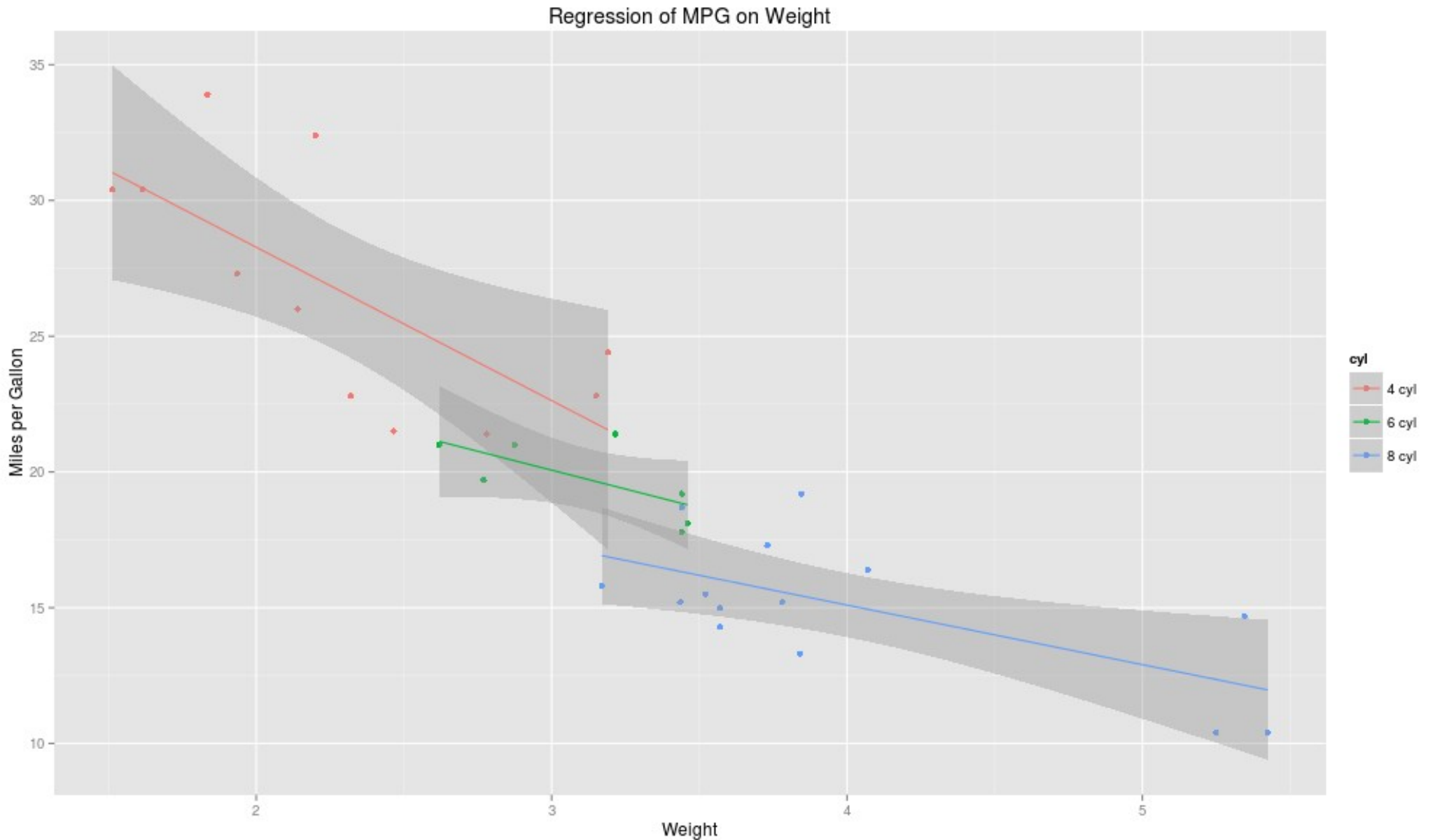
Carats = Independent variable →

clarity

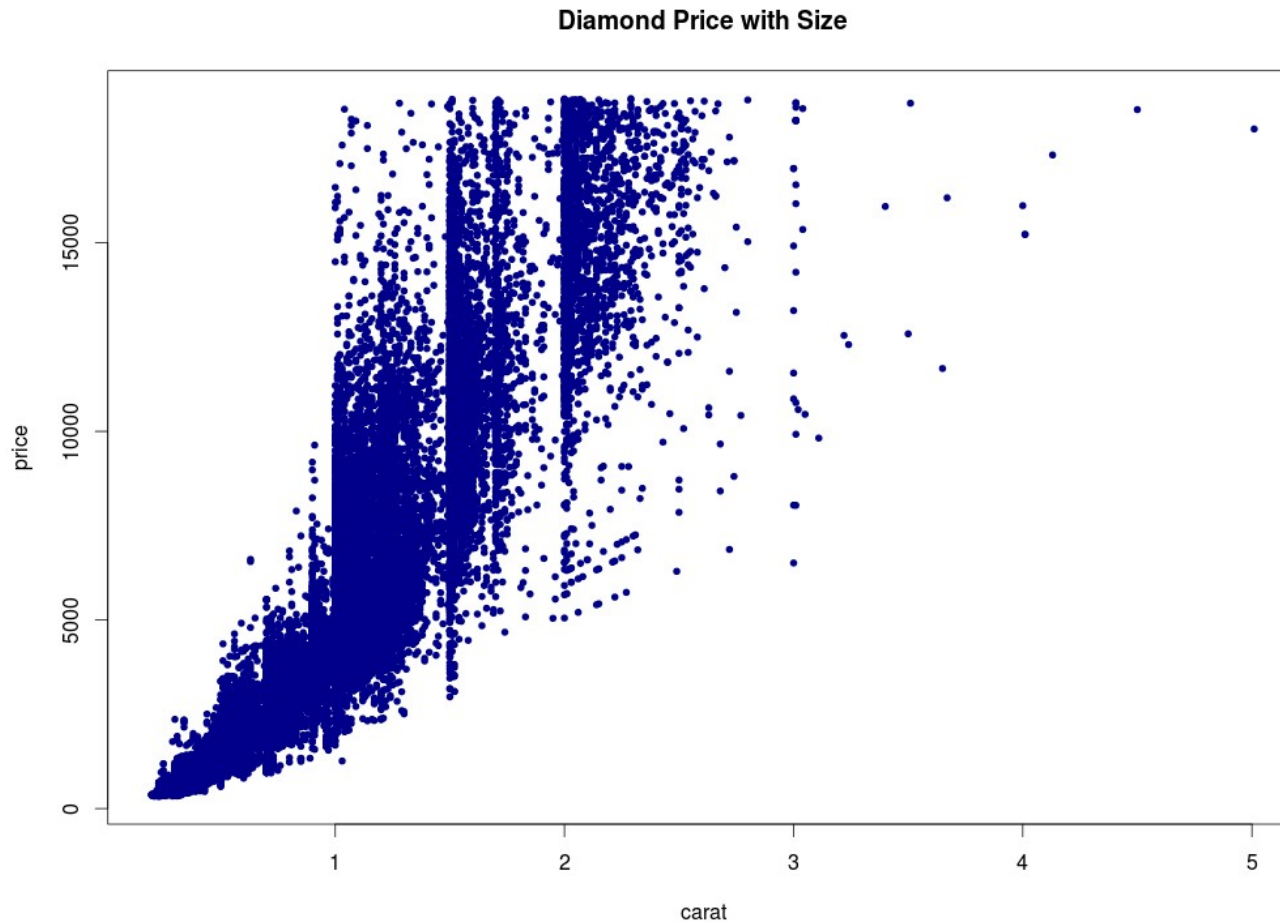
- I1
- SI2
- SI1
- VS2
- VS1
- WS2
- WS1
- IF



Scatterplot with Regression Lines

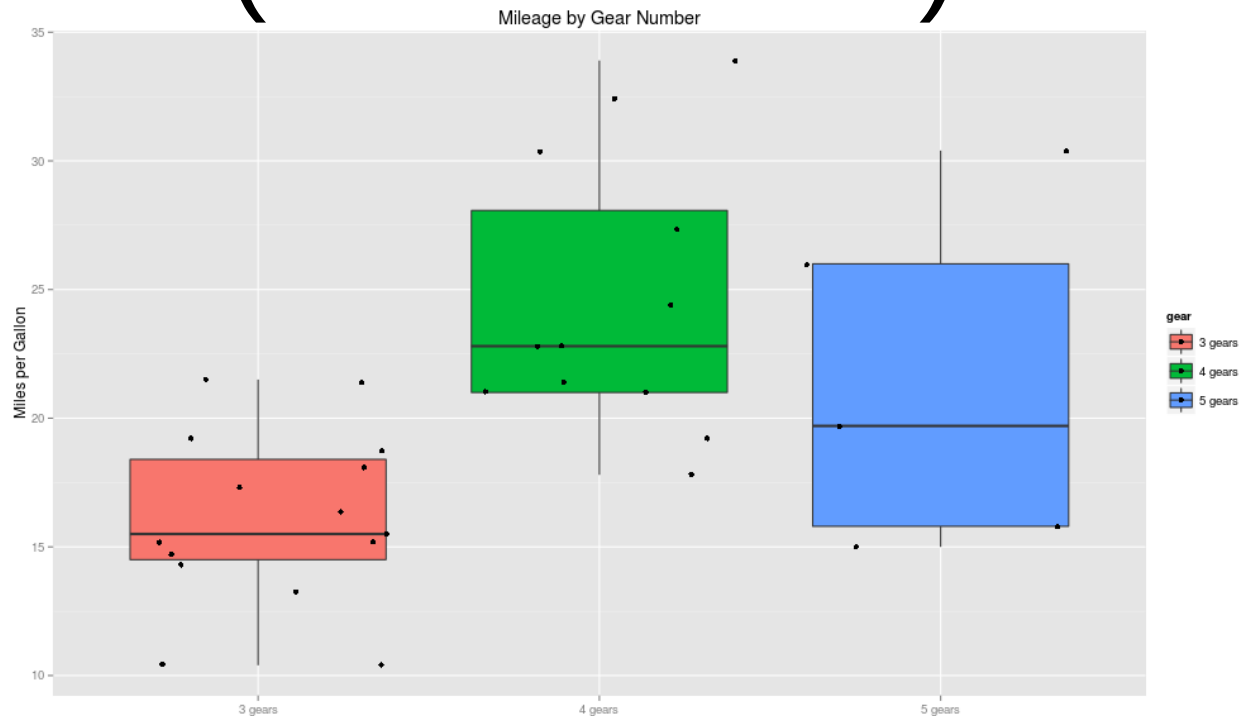


Scatterplot: Numeric Data, y vs. x



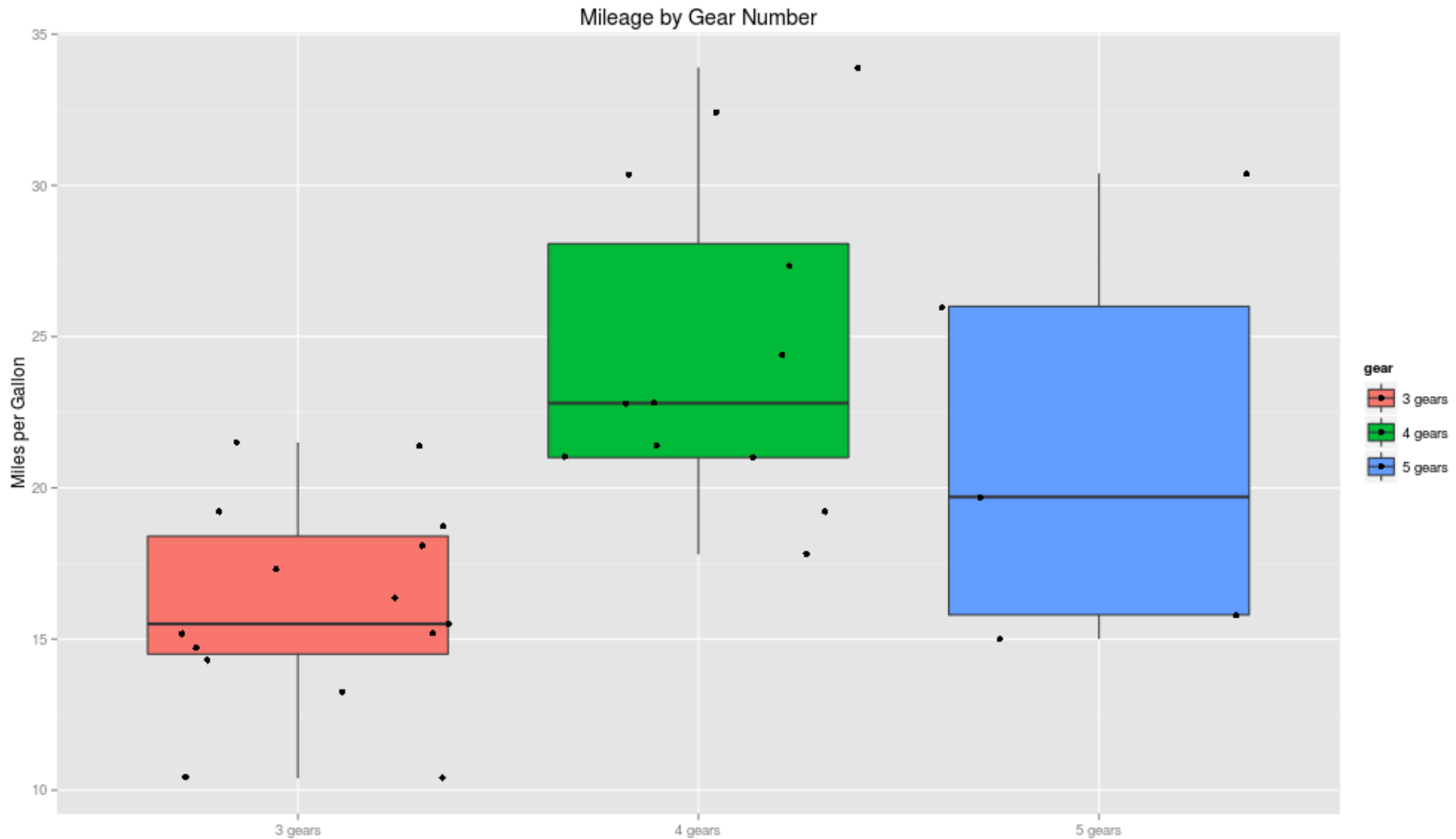
```
plot(formula=price~carat,  
     data=diamonds,  
     col="darkblue",  
     pch=20,  
     main="Diamond Price with Size")
```

Box (and Whisker) Plot

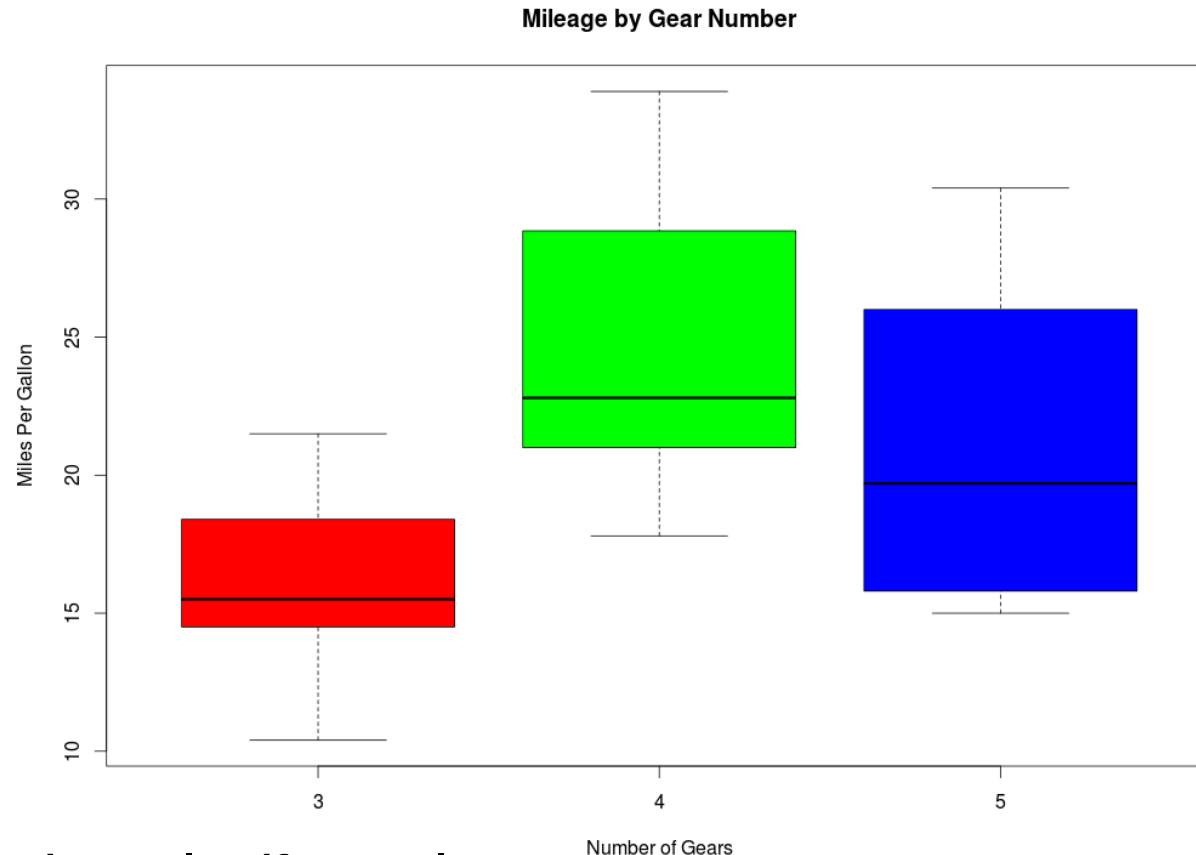


- The *box* extends from Q1 to Q3
- The *median*, Q2, is marked inside the box
- The *whiskers* extend to the min and max
 - Whiskers: required to lie within $1.5 \times (\text{IQR})$
 - *Outliers*: beyond $1.5 \times (\text{IQR})$

Boxplot: Data Symmetry?



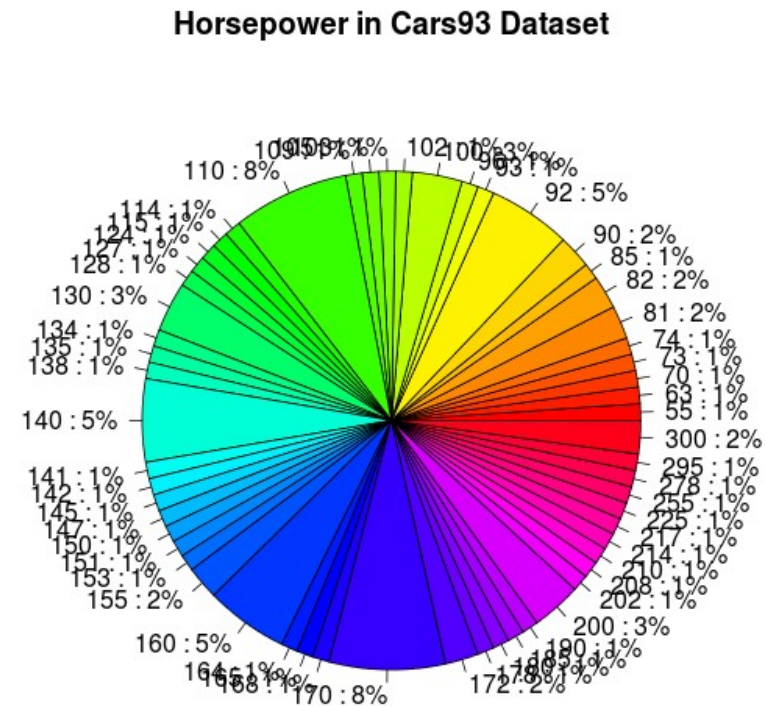
Box (and Whisker) Plot



```
boxplot(formula=mpg~gear,  
        data=mtcars,  
        main="Mileage by Gear Number",  
        xlab="Number of Gears",  
        ylab="Miles Per Gallon",  
        col=c("red","green","blue"))
```


Approach to Plotting

- Remember, you're getting to know your data.
- Don't be afraid to tinker and play.
- Sometimes the outcomes are silly (make sure you learn something!)



```
pie(table(Cars93$Horsepower))
```

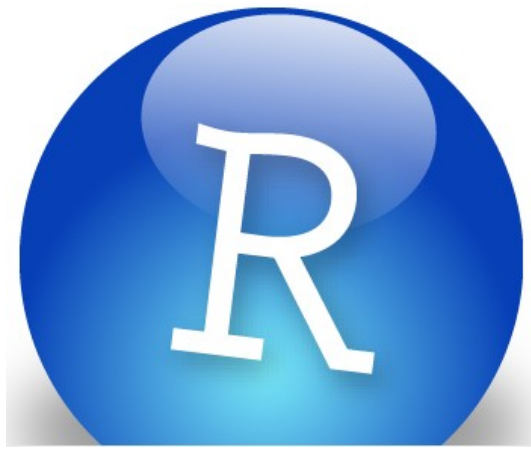
Interlude

Complete plotting exercises.



Open in the RStudio source editor:

```
<workshop>/exercises/exercises-plotting-basic.R
```



...is free

If you want to experiment further with R and RStudio, you can install them on your favorite operating system at home.

First, install R:

<http://cran.r-project.org/>

Then, install the Rstudio IDE:

<http://www.rstudio.com/ide/>