Intro to R - Plots Part II

Andrea Lane (adapted from Steve Pittard)

April 8, 2019

Recall - 4 Major Graphics Packages in R

- Base Graphics (covered last week)
- lattice (will cover today)
- ggplot (will cover today)
- grid - used by all other graphics packages

What makes a good plot?

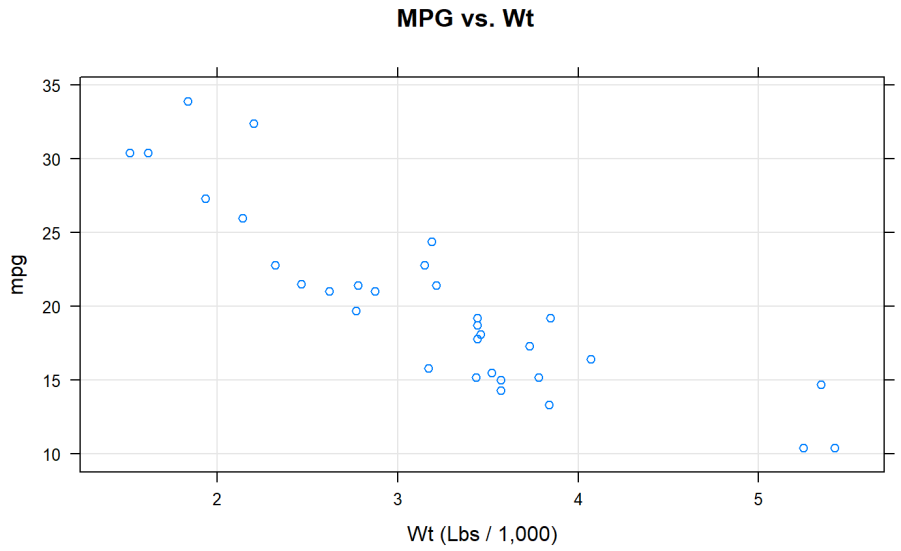
Lattice Graphics

Lattice was written to provide grouping and paneling

- Consistent look and feel
- Great for multivariate data
- Takes care of lots of things for you
- Has a formula interface
- Lots of examples and support on Google
- See http://lmdvr.r-forge.r-project.org/figures/figures.html/
- Picks useful defaults for you

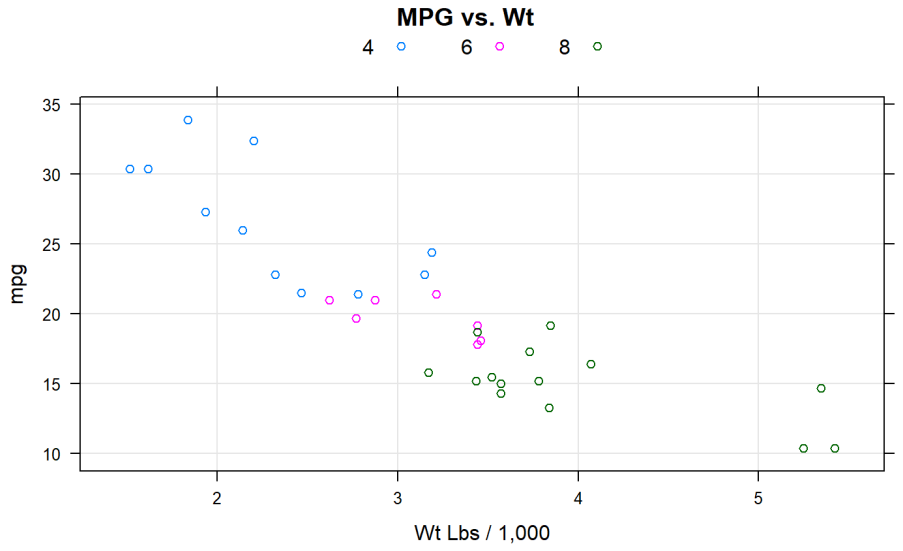
Lattice Graphics

```
library(lattice)
xlab <- "Wt (Lbs / 1,000)"
main <- "MPG vs. Wt"
xyplot(mpg~wt,data=mtcars,main=main,xlab=xlab,type=c("p","g"))
```



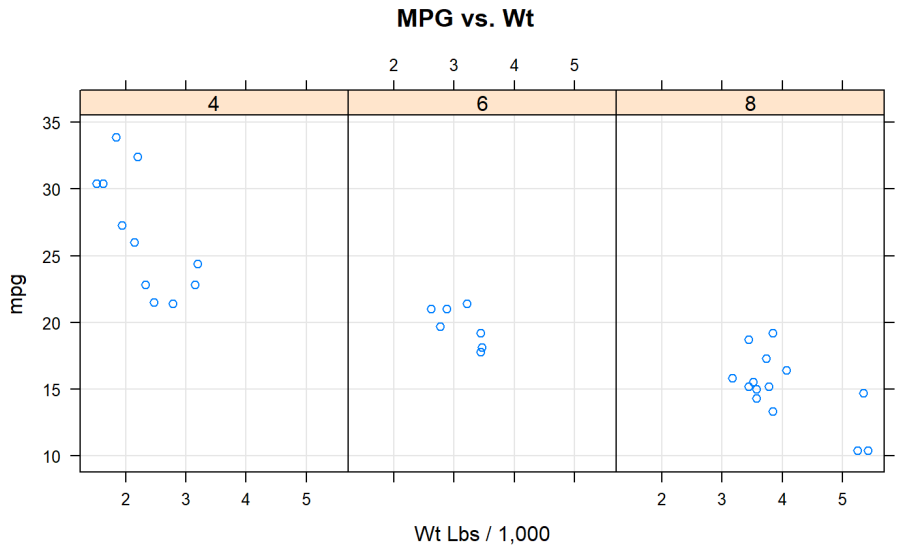
Lattice Graphics - Grouping

```
library(lattice)
xlab <- "Wt Lbs / 1,000"
main <- "MPG vs. Wt"
xyplot(mpg~wt,groups=factor(cyl),data=mtcars,main=main,xlab=xlab,
       type=c("p","g"),auto.key=list(columns=3))
```



Lattice Graphics - Panels

```
library(lattice)
xlab <- "Wt Lbs / 1,000"; main <- "MPG vs. Wt"
xyplot(mpg~wt|factor(cyl),data=mtcars,main=main,xlab=xlab,
       type=c("p","g"),layout=c(3,1))
```



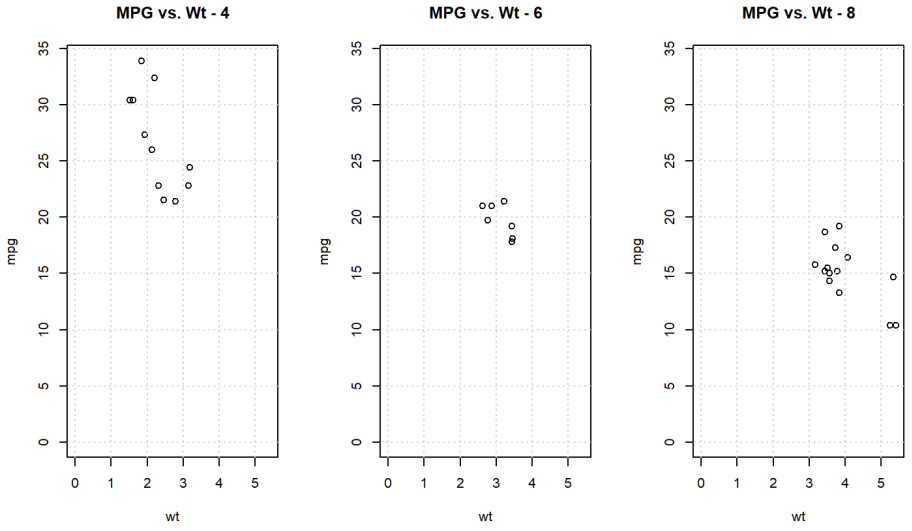
Panels in Base R

A manual process: creating 3 side-by-side plots with a loop

```
xlab <- "Wt Lbs / 1,000"; main <- "MPG vs. Wt"
par(mfrow=c(1,3))
maxmpg <- max(mtcars$mpg)
maxwt <- max(mtcars$wt)
mydf <- split(mtcars,mtcars$cyl)
for (ii in 1:length(mydf)) {
  tmpdf <- mydf[[ii]]
  main <- paste("MPG vs. Wt",names(mydf)[ii],sep=" - ")
  plot(mpg~wt,data=tmpdf,main=main,
  xlim=c(0,maxwt),
  ylim=c(0,maxmpg))
  grid()
}
```

Panels in Base R

A manual process: creating 3 side-by-side plots with a loop



ggplot2

- Rapidly becoming the default R graphics package
- Attempts to leverage the good parts of lattice and Base graphics
- Written according to a "Grammar of Graphics" (Wilkinson, 2005)
    - "I find myself still thinking about the book and its ideas, several weeks after I finished reading it. I love that kind of book"
    - "a richly rewarding work, an outstanding achievement by one of the leaders of statistical graphics"
    - "a pleasure to read, whether a novice or an expert in graphics"

ggplot2 resources

- home page for ggplot: http://ggplot2.org/
- Presentation: http://ggplot2.org/resources/2007-past-present-future.pdf
- Book: ggplot2: Elegant Graphics for Data Analysis (check Amazon)
- Vanderbilt Workshop: http://ggplot2.org/resources/2007-vanderbilt.pdf
- Documentation: http://ggplot2.tidyverse.org/reference
- R for Data Science Online Book - http://r4ds.had.co.nz/
- R Graphics Cookbook: http://www.cookbook-r.com/Graphs/index.html

ggplot2 resources

The cheat sheet is very useful!

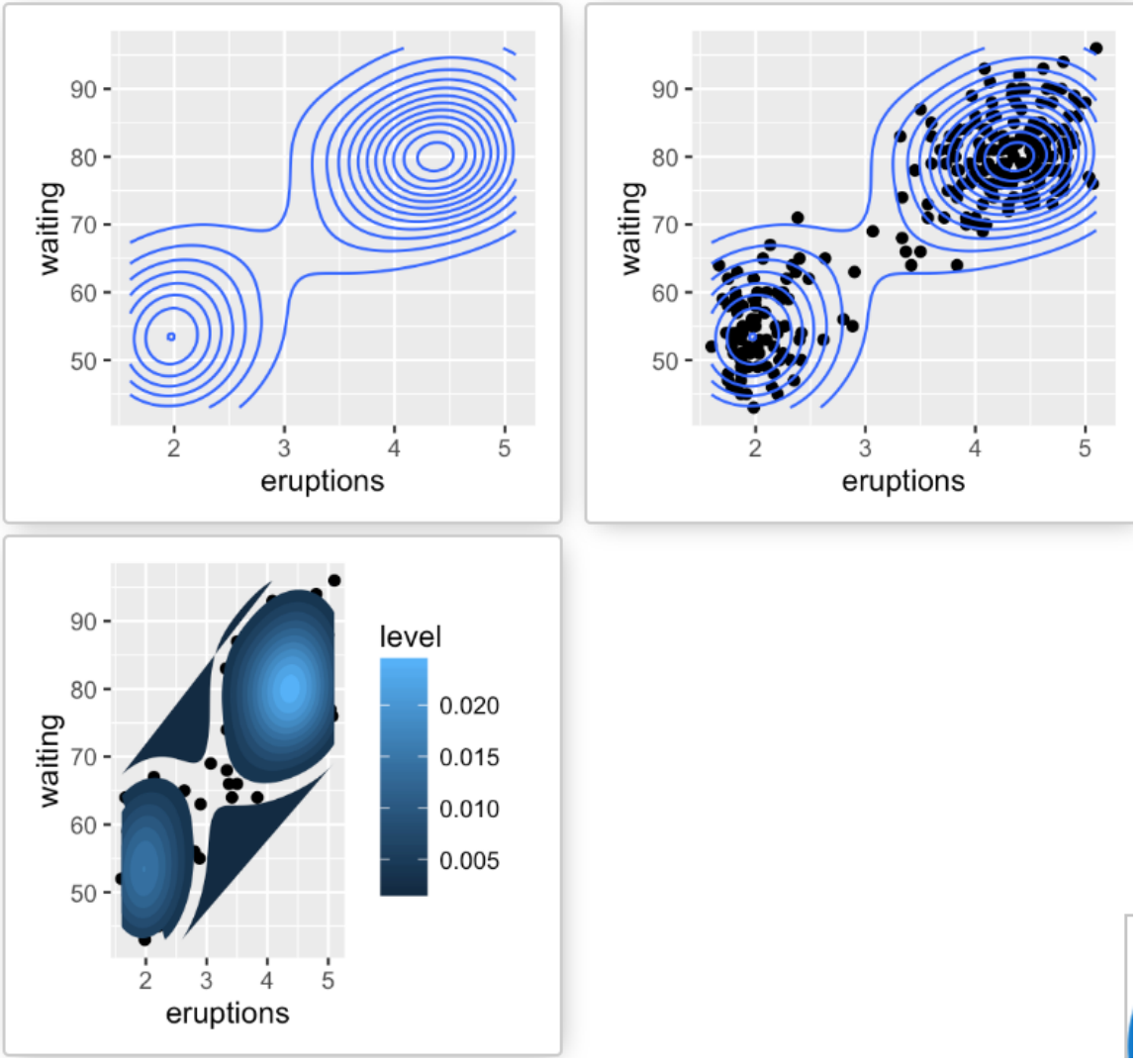
http://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

tidyverse

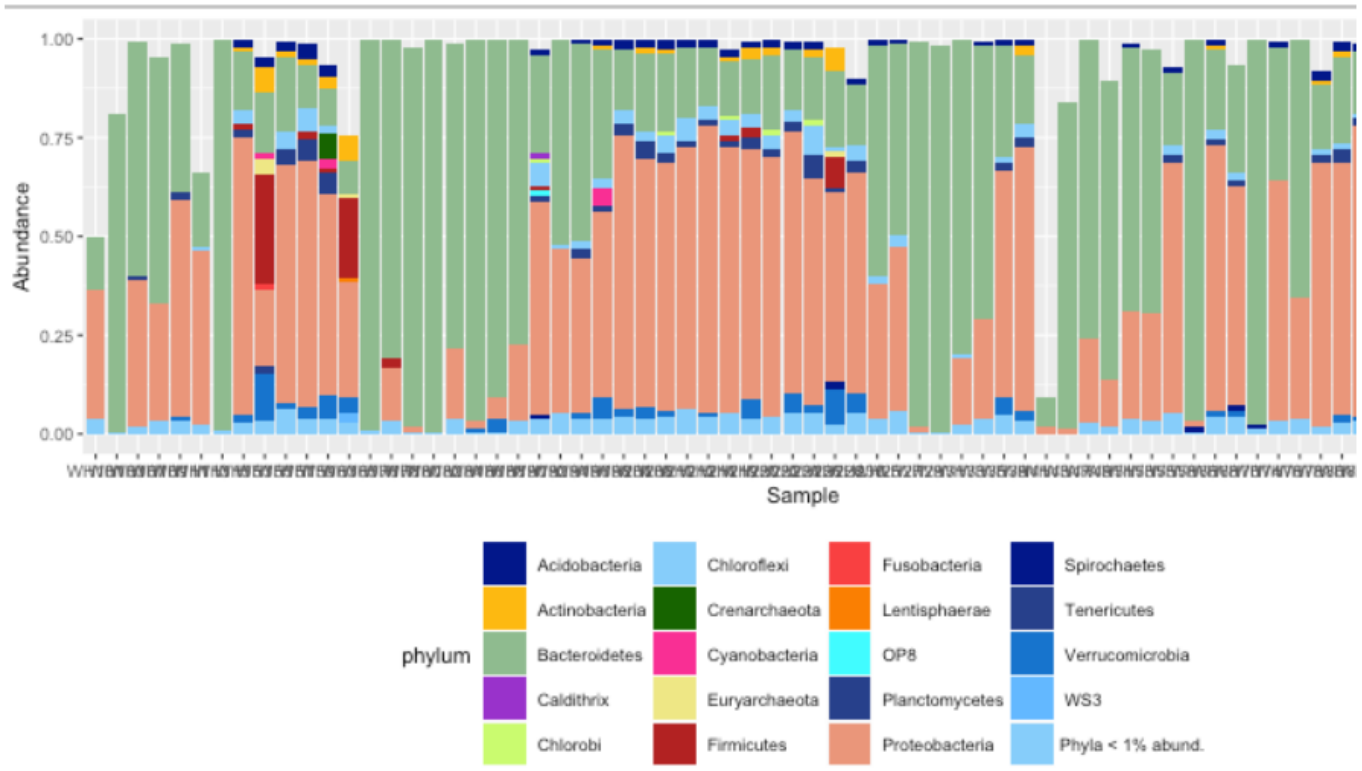
ggplot is part of the "tidyverse"

- A collection of R packages that share common philosophies to work well together
- Home page for project is at http://tidyverse.org/
- Main packages are: ggplot2, tibblr, tidyr, readr, purrr, dplyr
- Can install from within R Studio just like any other package
- The name of the package is simply tidyverse

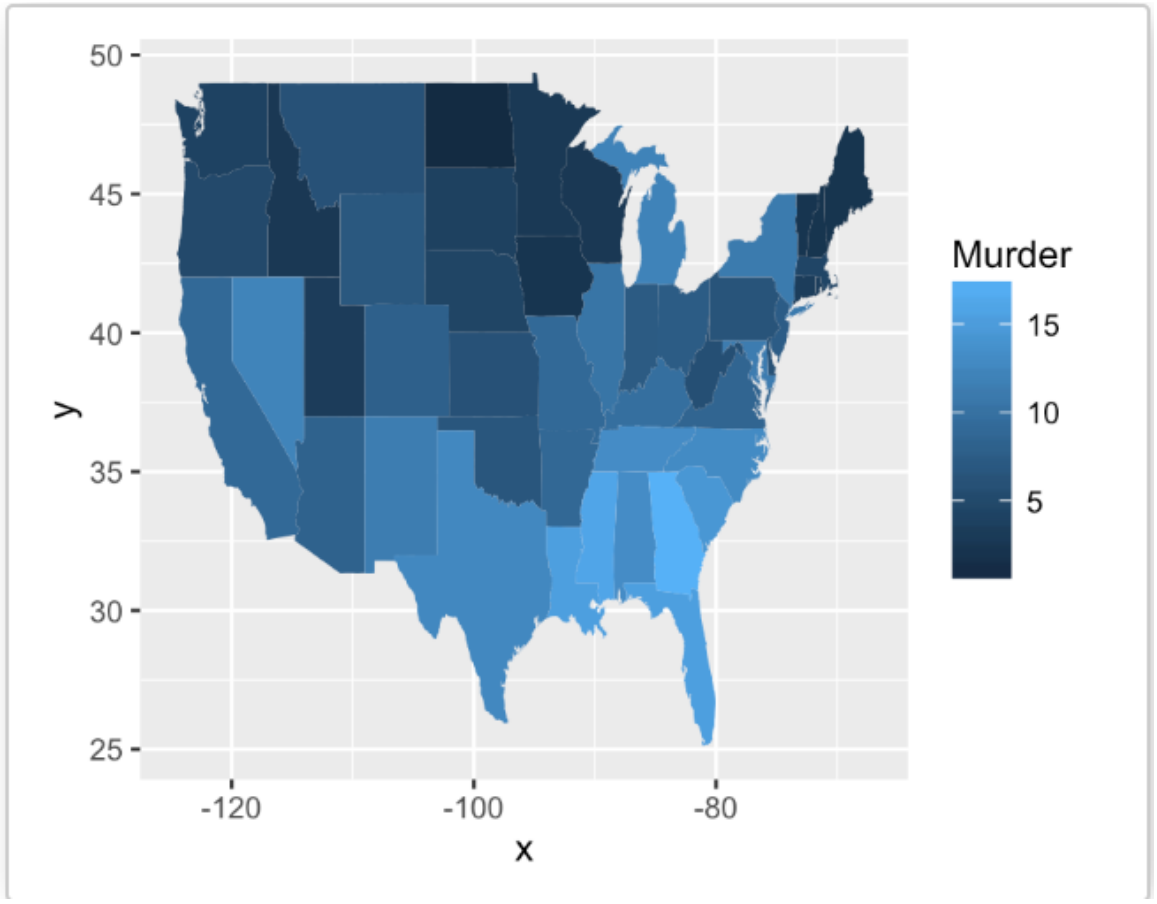
ggplot2 examples



ggplot2 examples



ggplot2 examples

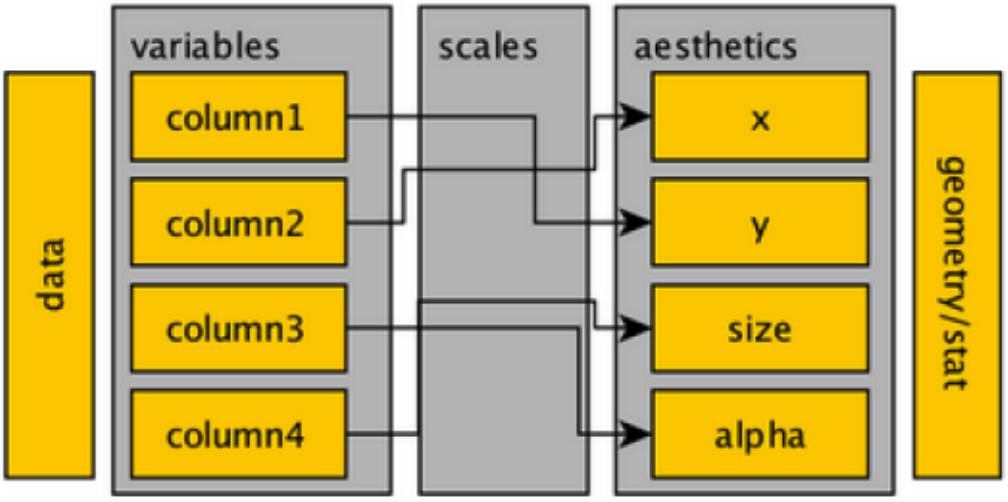


ggplot2 - Key Ideas

These ideas come from the Grammar of Graphics

- Understanding these ideas will help you define a plot in general terms that can be implemented using ggplot commands
- Data: the actual data frame under consideration
- Aesthetics: visual elements mapped to the data (axis, lines, colors, bars, etc)
- Scales: Transformations you might want to apply (e.g. logarithm, polar coordinates)
- Geometries: The shape mapped to the aesthetic(s)

ggplot2 - Key Ideas



Visualizing the data, aesthetics, scales, and geometries

Aesthetics

Here are some of the aesthetics that help make a plot:

- x and y position
- size of the elements
- shape
- color

We use geometries to view the data:

- lines and variations (dashed, segments, etc)
- bars, histograms
- text labels
- points
- http://ggplot2.tidyverse.org/reference/#section-layer-geoms

Examples using mtcars

Let's use the simple data set mtcars to create some plots with ggplot2

A data frame with 32 observations on 11 (numeric) variables.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (1000 lbs)

[, 7] qsec 1/4 mile time

[, 8] vs Engine (0 = V-shaped, 1 = straight)

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

Examples using mtcars

Let's use the simple data set mtcars to create some plots with ggplot2. What are the categorical variables in this data? What are the continuous variables?

Note: We often want to compare continuous quantities across groups (categorical variables)

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Plotting is all about exploring relationships!

Let's say we have the following 4 questions of interest:

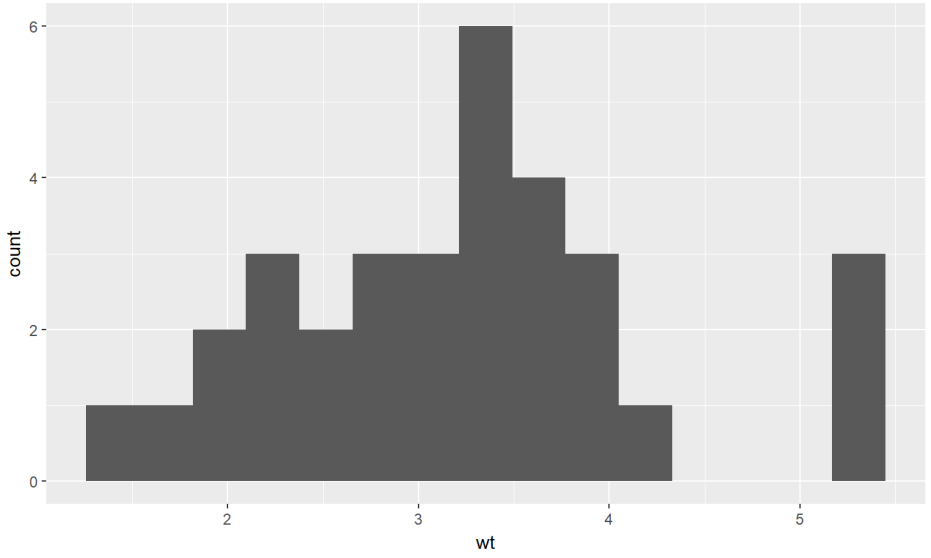
- What does the distribution of wt values look like?
- Is there a relationship between mpg and wt?
- Does mpg appear to be different over individual cylinder groups?
- What are the counts of transmission types and cylinder groups?

1. What does the distribution of wt values look like?

- What kind of variable is wt?
- Which type of plot corresponds to that variable type and the question we want to answer?

1. What does the distribution of wt values look like?

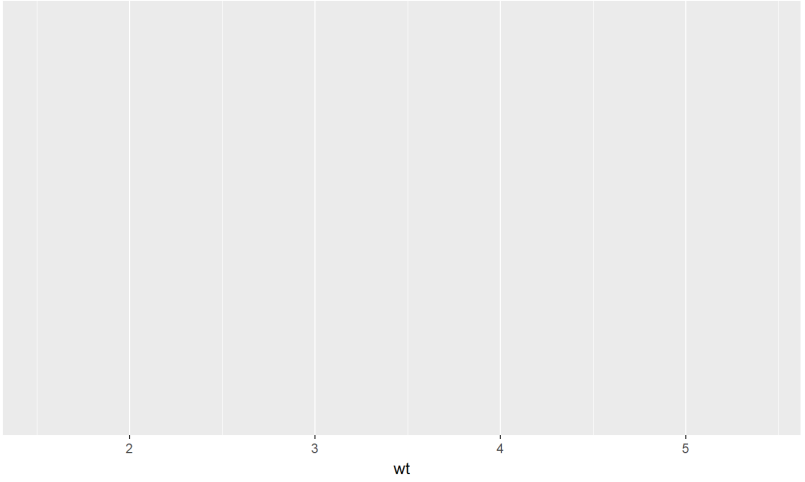
```
library(ggplot2)
ggplot(mtcars,aes(x=wt)) + geom_histogram(bins=15)
```



1. What does the distribution of wt values look like?

The geometry is crucial!

```
library(ggplot2)
ggplot(mtcars,aes(x=wt))
```

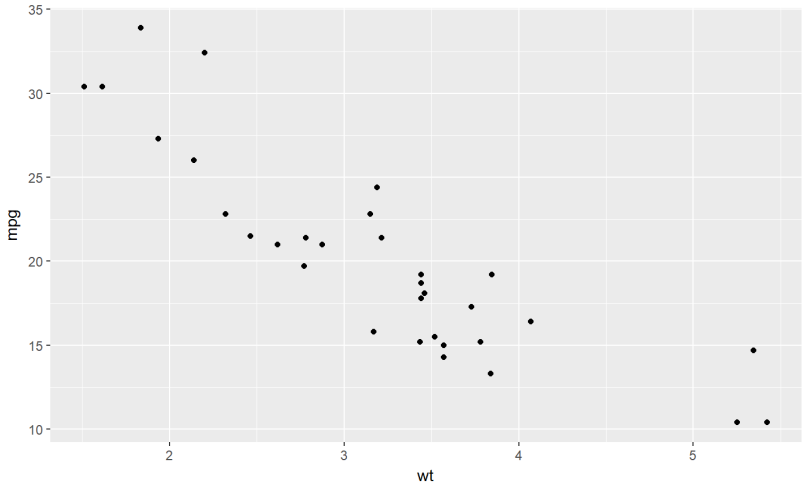


2. Is there a relationship between mpg and wt?

- What kinds of variables do we have?
- Which type of plot corresponds to these variable types and the question we want to answer?

2. Is there a relationship between mpg and wt?

```
ggplot(mtcars,aes(x=wt)) + geom_point(aes(y=mpg))
```

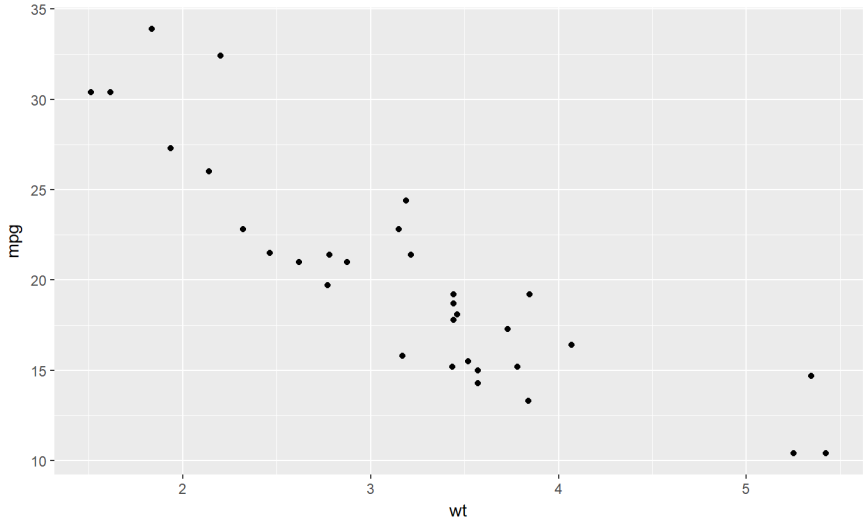


Note how we added a new geometry on an existing aesthetic mapping then added another aesthetic mapping - we mapped the y-axis to the mpg variable

2. Is there a relationship between mpg and wt?

We could also do this:

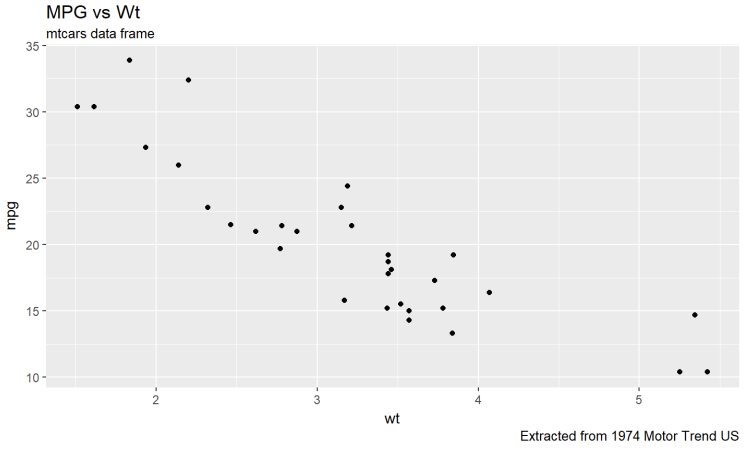
```
ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point()
```



2. Is there a relationship between mpg and wt?

Adding titles, labels, captions:

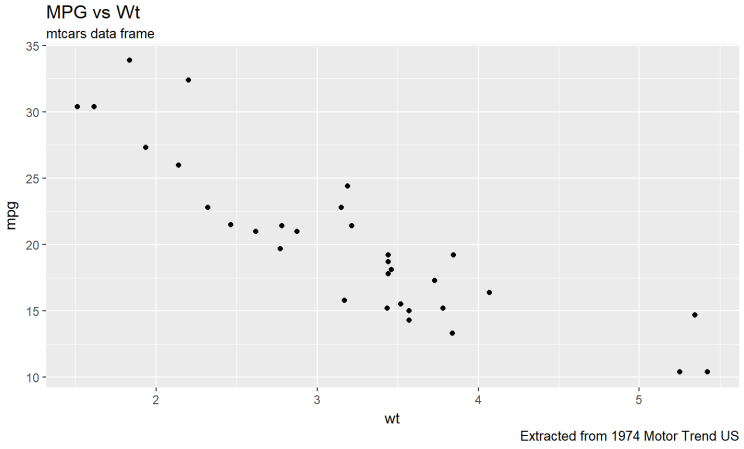
```
ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point() +
  ggtitle("MPG vs Wt","mtcars data frame") +
  labs(caption="Extracted from 1974 Motor Trend US")
```



2. Is there a relationship between mpg and wt?

Can also specify title in labs():

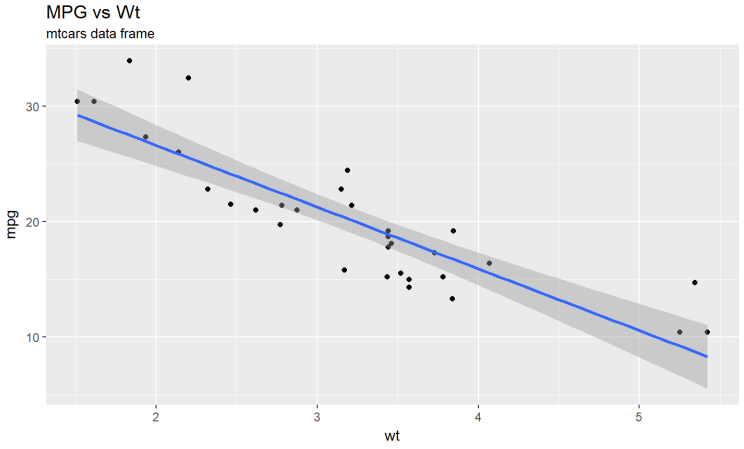
```
ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point() +
  labs(title="MPG vs Wt",subtitle="mtcars data frame",
       caption="Extracted from 1974 Motor Trend US")
```



2. Is there a relationship between mpg and wt?

Adding more geometry:

```
ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point() +
  ggtitle("MPG vs Wt","mtcars data frame") +
  geom_smooth(method="lm")
```



3. Does mpg appear to be different over individual cylinder groups?

- What kinds of variables do we have?
- Which type of plot corresponds to these variable types and the question we want to answer?

3. Does mpg appear to be different over individual cylinder groups?

- We can use color, shapes, and size to see how unique values of a factor or category impact the plot (this is called "grouping")
- Note that the cyl variable assumes 3 unique values:

```
unique(mtcars$cyl)
```

```
## [1] 6 4 8
```

```
#Let's make cyl an "official" factor:
mtcars$cyl <- factor(mtcars$cyl)
```

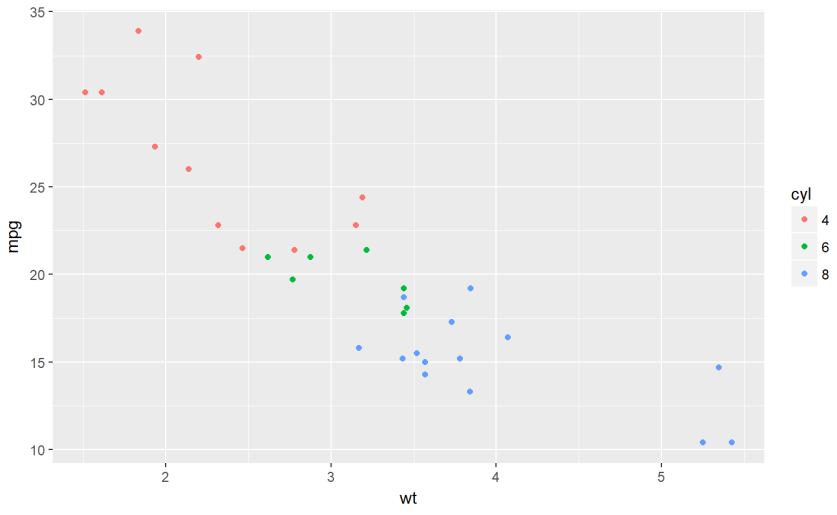
3. Does mpg appear to be different over individual cylinder groups?

Do you think specifying this grouping variable would be an aesthetics command or a geometry command?

3. Does mpg appear to be different over individual cylinder groups?

- In ggplot we use an "aesthetic mapping" to specify a grouping variable

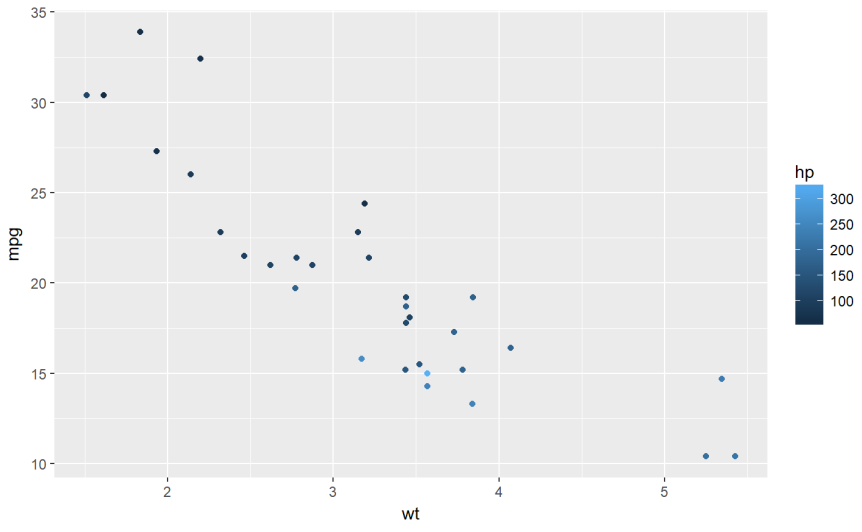
```
ggplot(mtcars,aes(x=wt,y=mpg,color=cyl)) + geom_point()
```



3. Does mpg appear to be different over individual cylinder groups?

What if we use a continuous quantity as a color aesthetic?

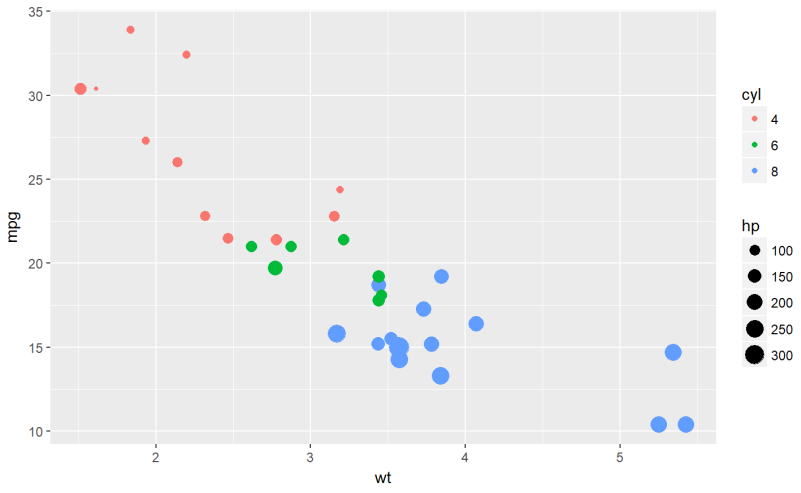
```
ggplot(mtcars,aes(x=wt,y=mpg,color=hp)) + geom_point()
```



3. Does mpg appear to be different over individual cylinder groups?

We can use multiple layers for grouping. (The aesthetic command can also go inside the geometry!)

```
ggplot(mtcars,aes(x=wt,y=mpg,color=cyl,size=hp)) + geom_point()
```

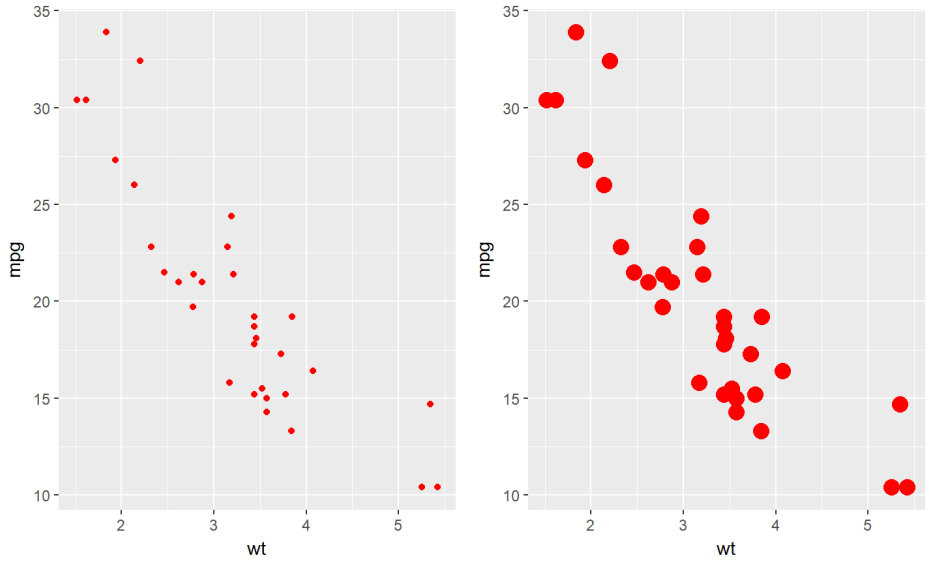


Note the difference between mappings and settings

- Mappings are usually functions of some variable in the data
- Settings alter appearance in a "fixed" way
- Previously we used "size" as a mapping. Here it is used as a setting.

```
library(gridExtra)
p1 <- ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point(color="red")
p2 <- ggplot(mtcars,aes(x=wt,y=mpg)) + geom_point(color="red",size=4)
grid.arrange(p1, p2, nrow=1, ncol=2)
```

Note the difference between mappings and settings



4. What are the counts of transmission types and cylinder groups?

To answer this, let's discuss how to handle counts and tabular data in ggplot

Counts and Tabular Data

- Let's say we are given data in a 2x2 table. We would first need to convert the table into a data frame

```
(ctab <- table(carb=mtcars$carb))
```

```
## carb
##  1  2  3  4  6  8
##  7 10  3 10  1  1
```

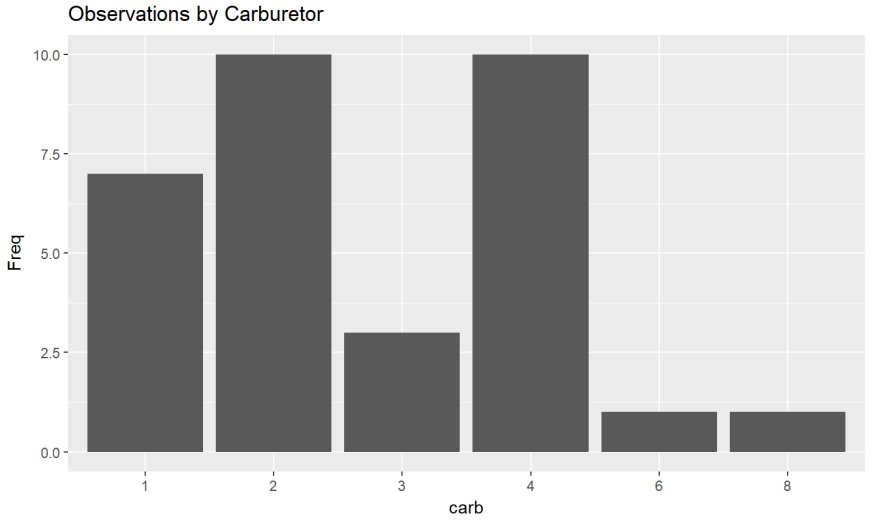
```
(df <- as.data.frame(ctab))
```

```
##   carb Freq
## 1    1    7
## 2    2   10
## 3    3    3
## 4    4   10
## 5    6    1
## 6    8    1
```

Counts and Tabular Data

Then we can use this data frame to create a plot. (Note stat="identity" because geom_bar uses stat_count by default)

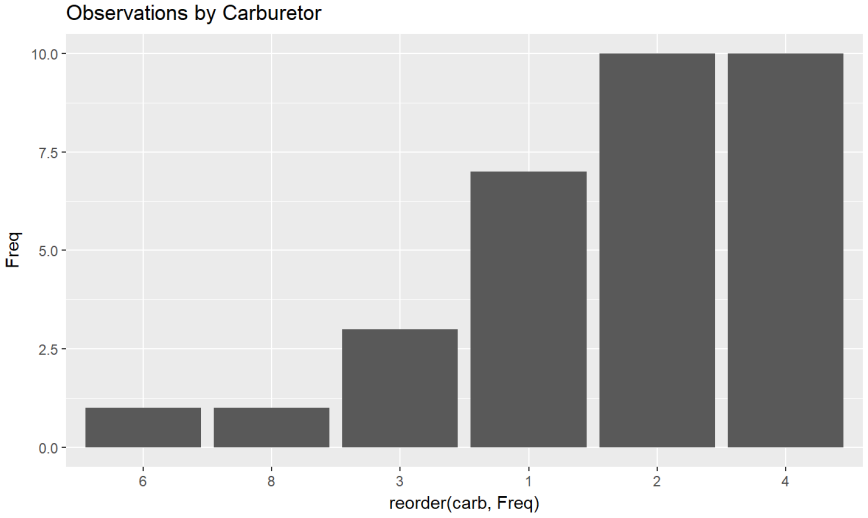
```
ggplot(df,aes(x=carb,y=Freq)) + geom_bar(stat="identity") +
  ggtitle("Observations by Carburetor")
```



Counts and Tabular Data

- What if we want to rearrange the bars?
- Here we want the variable "carb" in order of "Freq"

```
ggplot(df,aes(x=reorder(carb,Freq),y=Freq)) + geom_bar(stat="identity")
  ggtitle("Observations by Carburetor")
```

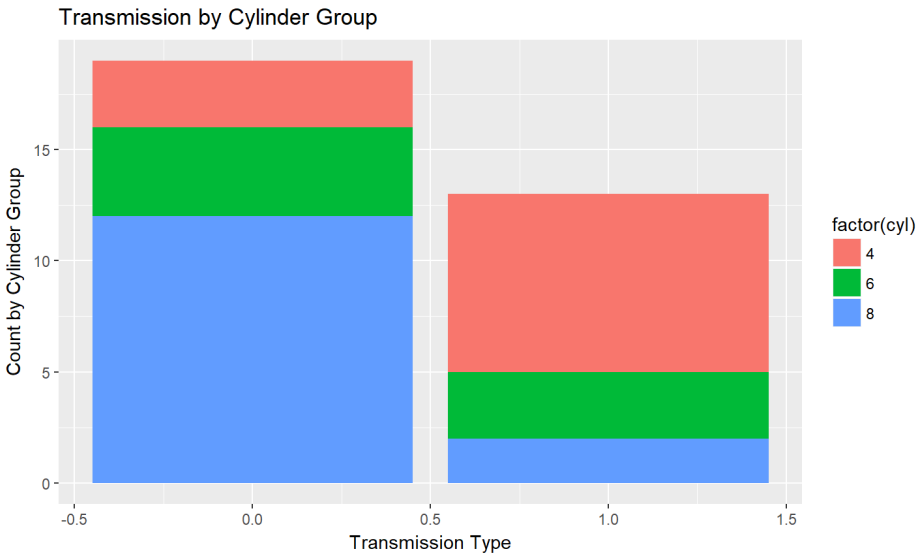


4. What are the counts of transmission types and cylinder groups?

- In this case, we already have the data frame. We can use the "fill" aesthetic for the factor variable "cyl"
- Similar to the "grouping" we did earlier, but over a factor variable

```
ggplot(mtcars,aes(x=am)) + geom_bar(aes(fill=factor(cyl))) +
  ggtitle("Transmission by Cylinder Group") +
  xlab("Transmission Type") +
  ylab("Count by Cylinder Group")
```

4. What are the counts of transmission types and cylinder groups?

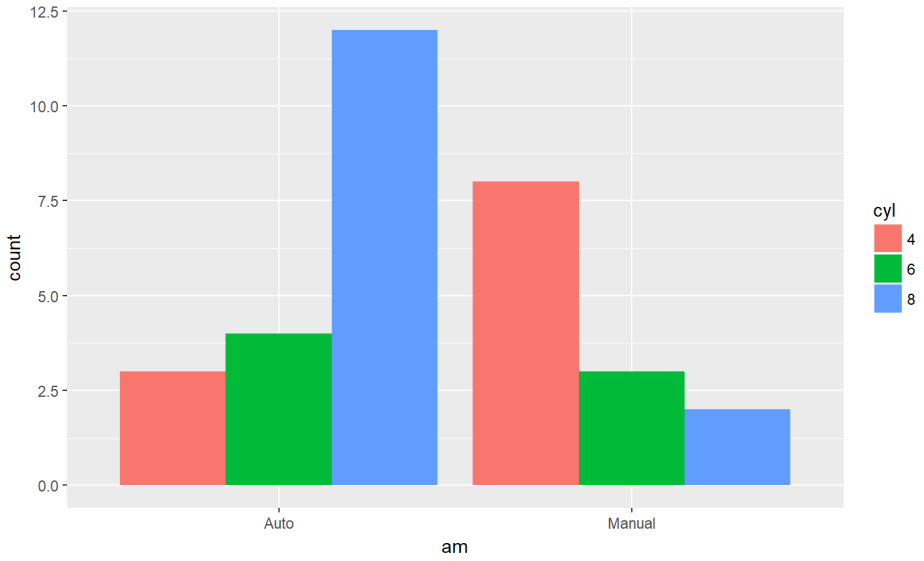


4. What are the counts of transmission types and cylinder groups?

- To have the bars side by side, create factor variables
- position="dodge" adjusts the horizontal positioning of the bars (compare to position="dodge2")

```
mtcars <- transform(mtcars,am=factor(am,labels=c("Auto","Manual"
    )),cyl=factor(cyl))
    ggplot(mtcars,aes(x=am)) +
    geom_bar(aes(fill=cyl),position="dodge")
```

4. What are the counts of transmission types and cylinder groups?



A few more fun things

- Facets (panels for ggplot)
- density and color fill
- boxplots

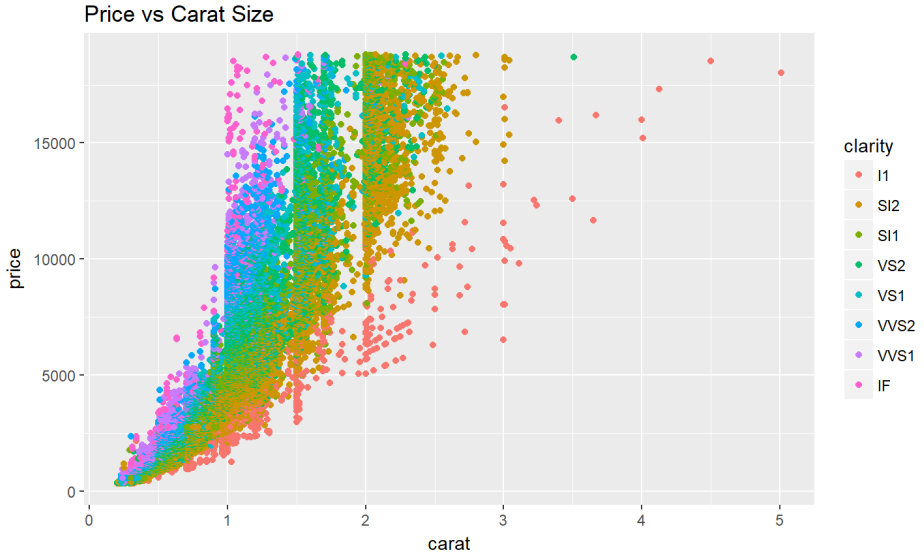
Facets

- Equivalent to paneling in lattice

Consider the following example with the diamonds data set (without facets):

```
ggplot(diamonds,aes(x=carat,y=price)) +
geom_point(aes(color=clarity)) +
ggtitle("Price vs Carat Size")
```

Facets



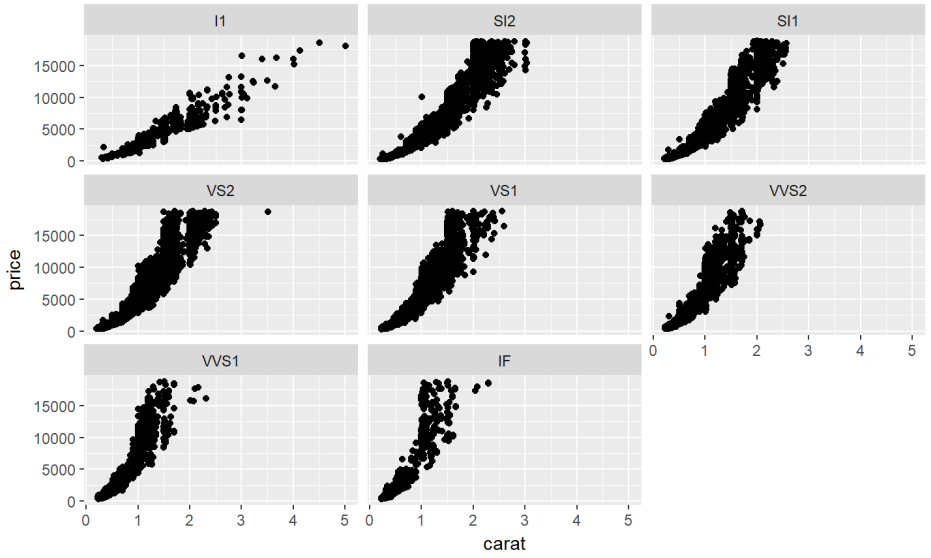
Hard to read and interpret!

Facets

- The facet_wrap() is used to break down a large plot into multiple small plots for individual categories. It takes a formula as the main argument. The items to the left of ~ forms the rows while those to the right form the columns.

```
ggplot(diamonds,aes(x=carat,y=price)) +
geom_point() +
facet_wrap(~clarity)
```

Facets

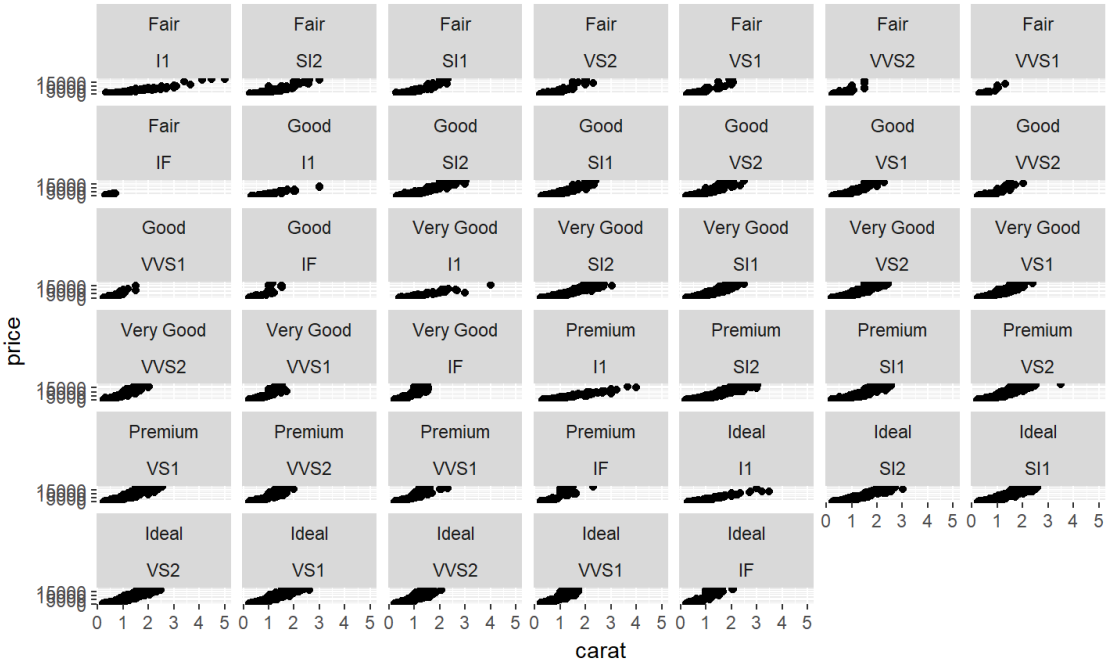

I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
price
carat
0
5000
10000
15000
0
1
2
3
4
5

Facets

- The facet_wrap() is used to break down a large plot into multiple small plots for individual categories. It takes a formula as the main argument. The items to the left of ~ forms the rows while those to the right form the columns.

```
ggplot(diamonds,aes(x=carat,y=price)) +
geom_point() +
facet_wrap(cut~clarity)
```

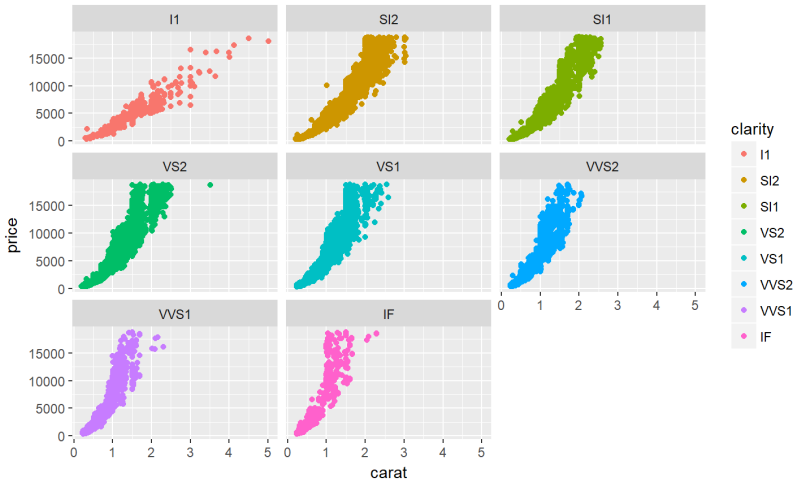
Facets


Fair I1
Fair SI2
Fair SI1
Fair VS2
Fair VS1
Fair VVS2
Fair VVS1
Fair IF
Good I1
Good SI2
Good SI1
Good VS2
Good VS1
Good VVS2
Good VVS1
Good IF
Very Good I1
Very Good SI2
Very Good SI1
Very Good VS2
Very Good VS1
Very Good VVS2
Very Good VVS1
Very Good IF
Premium I1
Premium SI2
Premium SI1
Premium VS2
Premium VS1
Premium VVS2
Premium VVS1
Premium IF
Ideal I1
Ideal SI2
Ideal SI1
Ideal VS2
Ideal VS1
Ideal VVS2
Ideal VVS1
Ideal IF
price
carat
0 1 2 3 4 5

Facets

But let's not leave out the pretty colors…

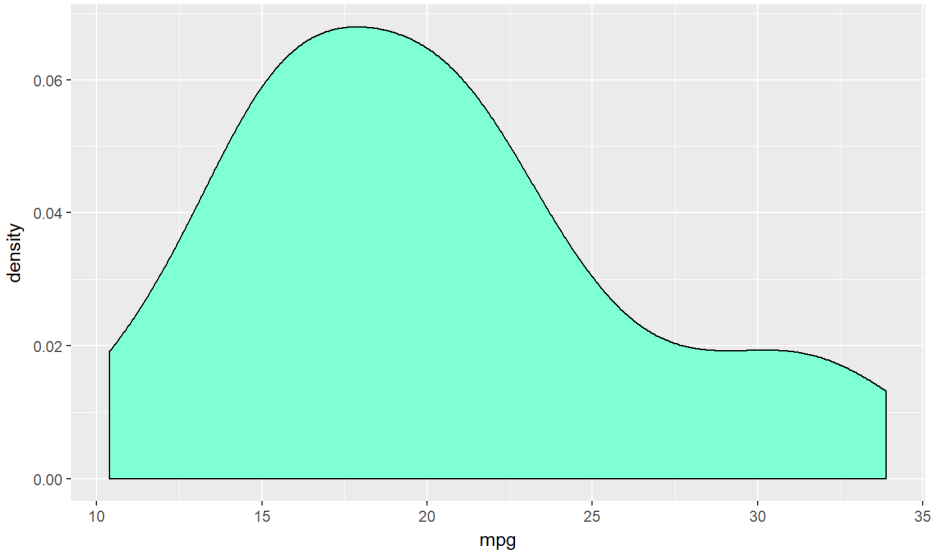
```
ggplot(diamonds,aes(x=carat,y=price)) +
geom_point(aes(color=clarity)) +
facet_wrap(~clarity)
```



Density

In general anything you wish to set to a static value should be set outside of the aes function

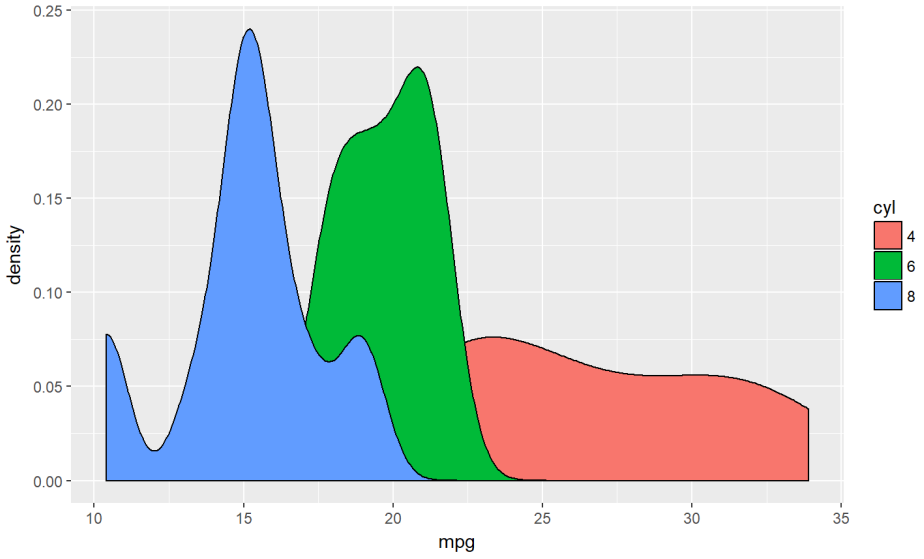
```
ggplot(mtcars) + geom_density(aes(x=mpg),fill="aquamarine")
```



Density

We can group the density (fill moves inside the aes function):

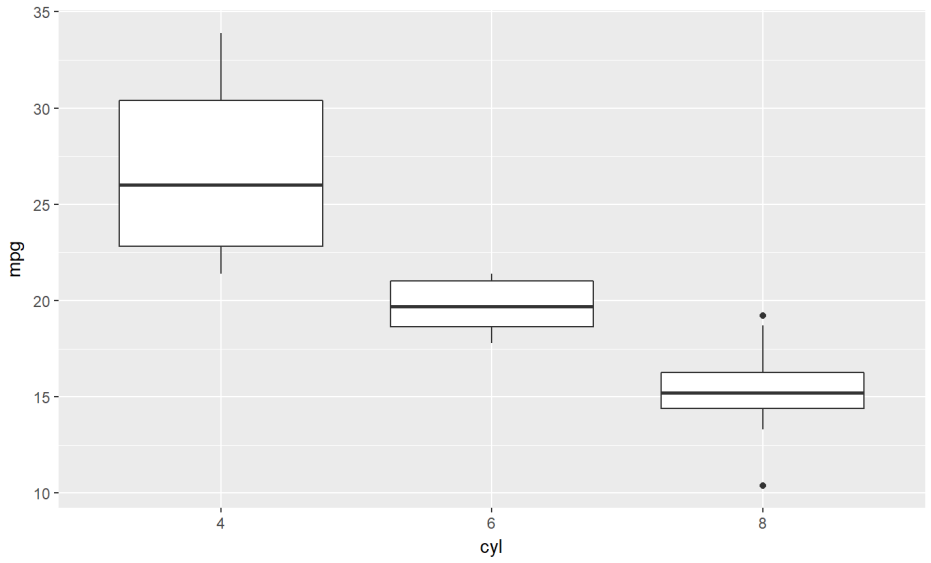
```
ggplot(mtcars) + geom_density(aes(x=mpg,fill=cyl))
```



Boxplot

A boxplot of mpg across cylinder groups:

```
ggplot(mtcars) + geom_boxplot(aes(x=cyl,y=mpg))
```



Summary of ggplot geometry

- One continuous variable
  - geom_density
  - geom_histogram
- Two continuous variables
  - geom_point
  - geom_smooth
- One discrete variable + one continuous variable
  - geom_boxplot
  - geom_bar
- See the ggplot2 cheat sheet for many other geometry options!

References

- http://r-statistics.co/Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html#5.%20Faceting:%20Draw%20multiple%20plo
- http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-visualization/
- https://ggplot2.tidyverse.org/reference/position_dodge.h

Happy Plotting!