Empirical Methods of Data Science

PROFESSOR MICHELLE LEVINE

WEEK 10: DATA ANALYSIS

4/3/19

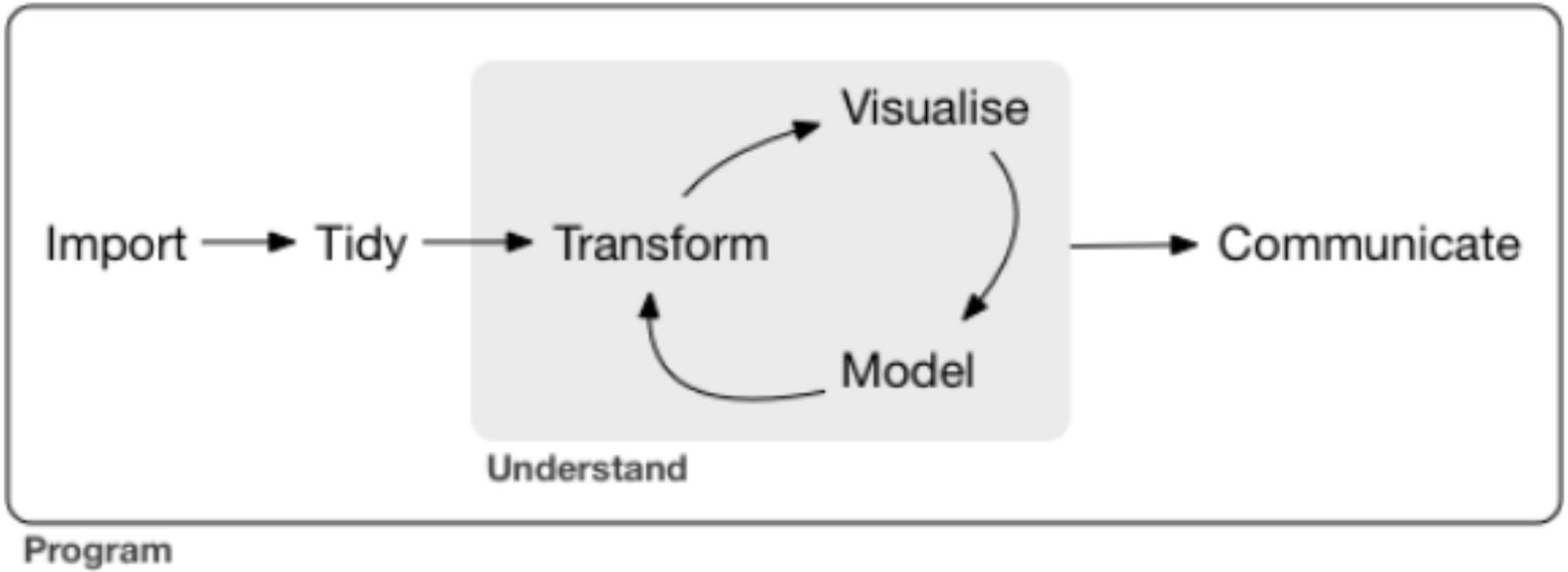
Today

Data analysis

Using R

- Online guides:
  - R for Data Science – all online at https://r4ds.had.co.nz
  - Quick-R: http://www.statmethods.net
  - Cookbook for R: http://www.cookbook-r.com

Tools needed for data science research



Unless noted, graphs in presentation are From R for Data Science

Visualizing your data

One of the first steps you want to do is explore and get a visual sense of your data.

This will tell you a lot about general patterns and guide your analyses.

Visually look at mean, median, mode, standard deviation

Graphs

Boxplot

Histogram

Barchart

Scatterplot

Boxplot

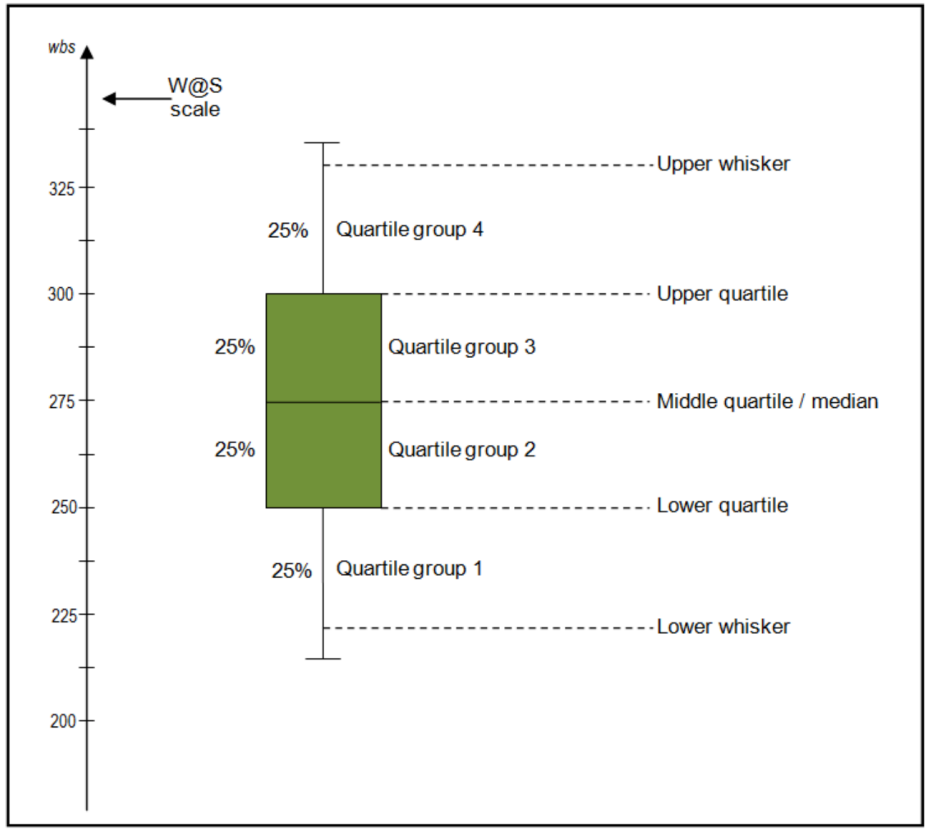
Displays the distribution of a dataset that includes both categorical and continuous data.

- median, range, quartiles, whiskers

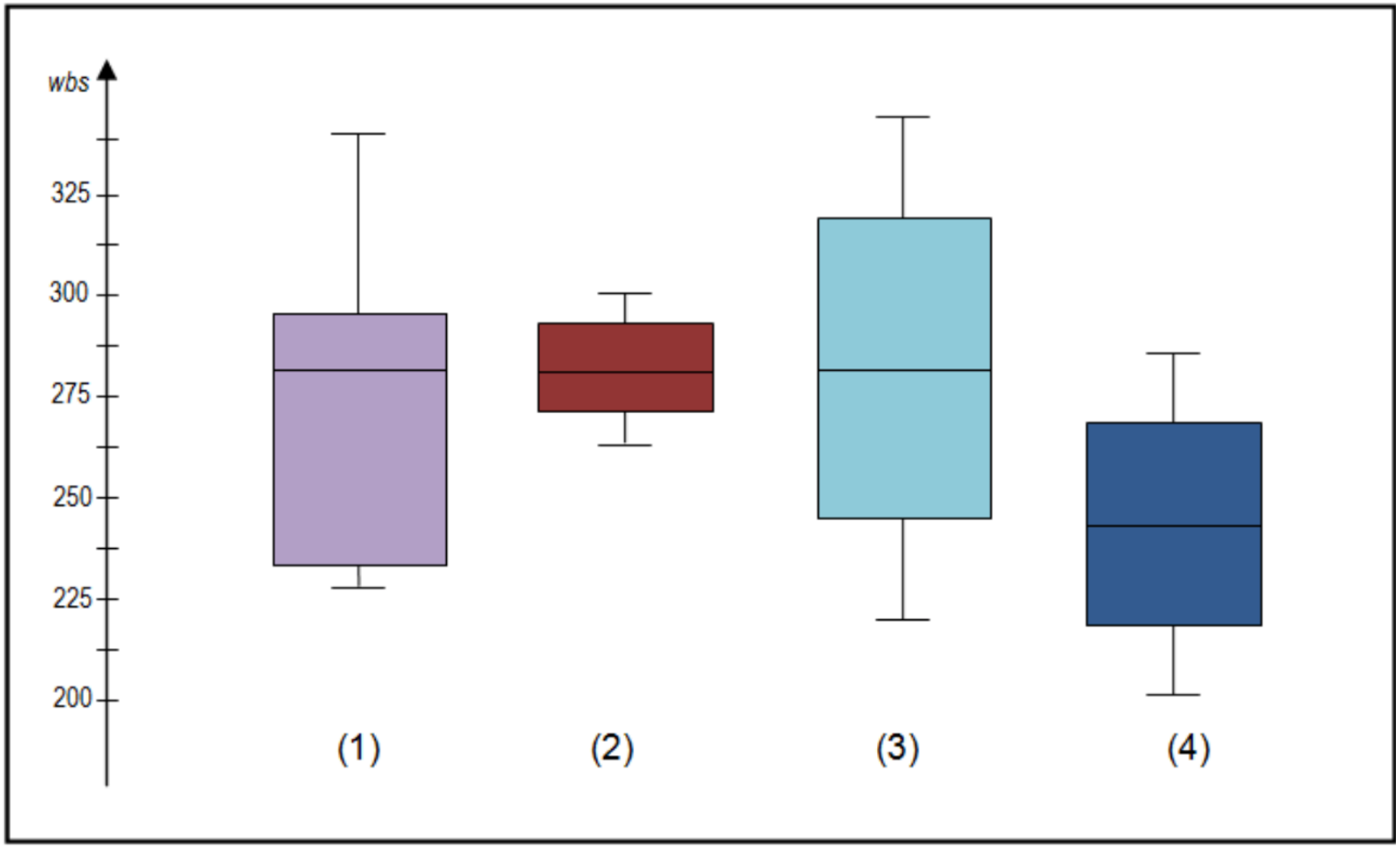
Good for clearly seeing outliers.

Good for comparing different variables or different sets of data as more than one can be plotted per graph (horizontally or vertically).

Boxplot terminology



Examples from: https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots


wbs
325
300
275
250
225
200
(1)
(2)
(3)
(4)

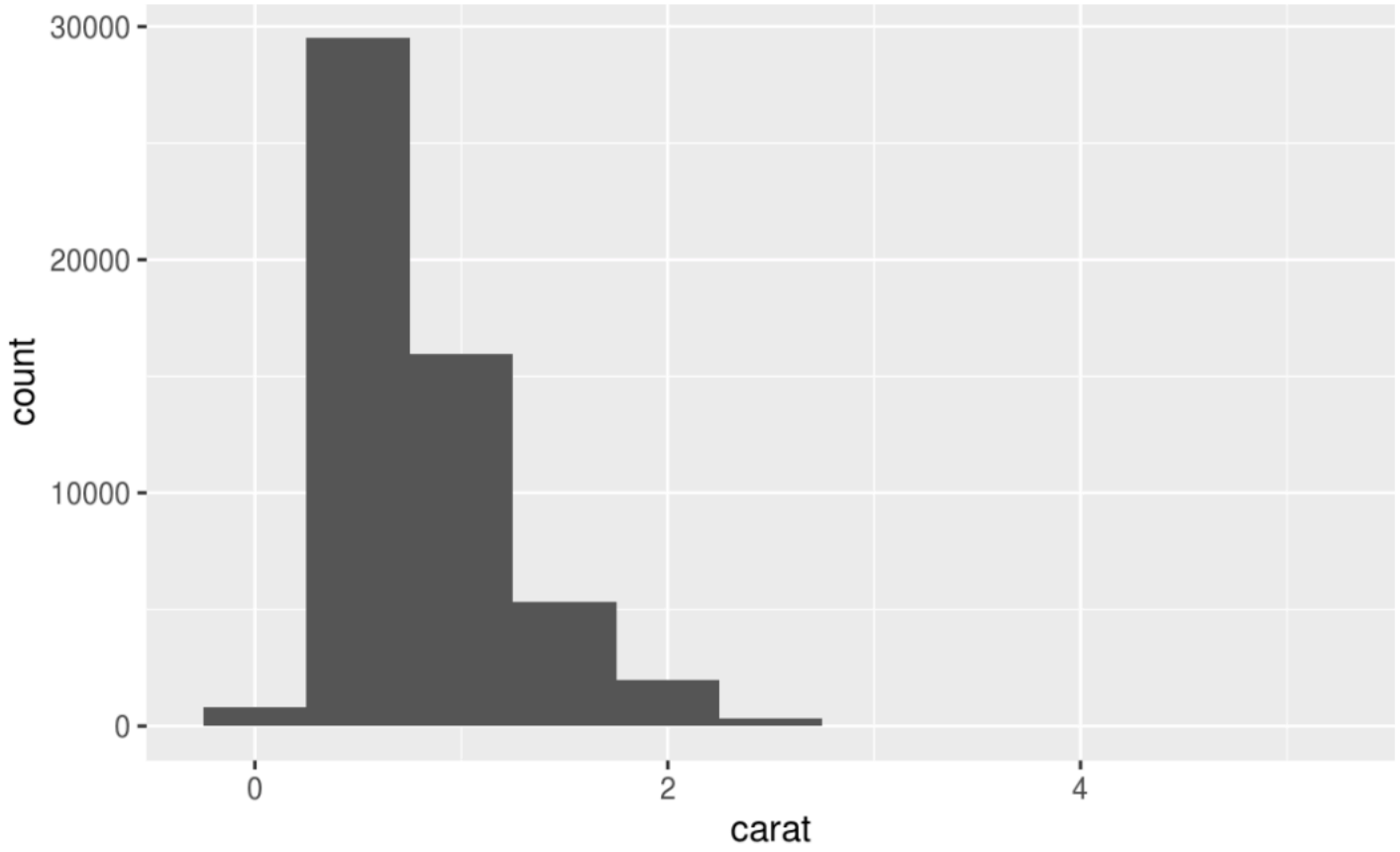
Histogram

Used with continuous data.

Similar use to a boxplot in that it shows you the general distribution of the data.

But histograms also shows frequencies for each response.

In a traditional histogram, each graph can only represent one variable (however more recent methods can illustrate multiple variables).


30000
20000
10000
0
count
0
2
4
carat

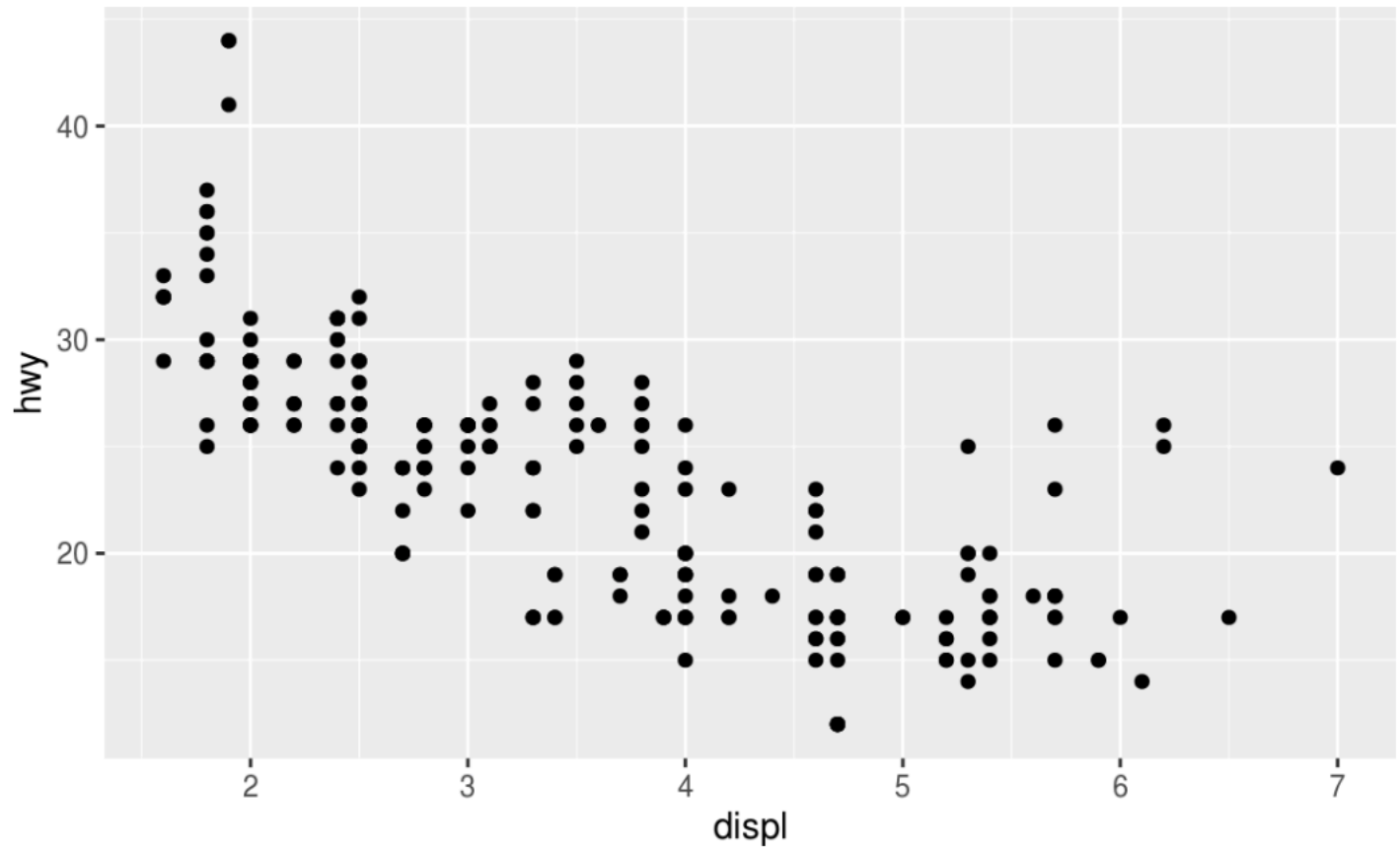
Scatterplot

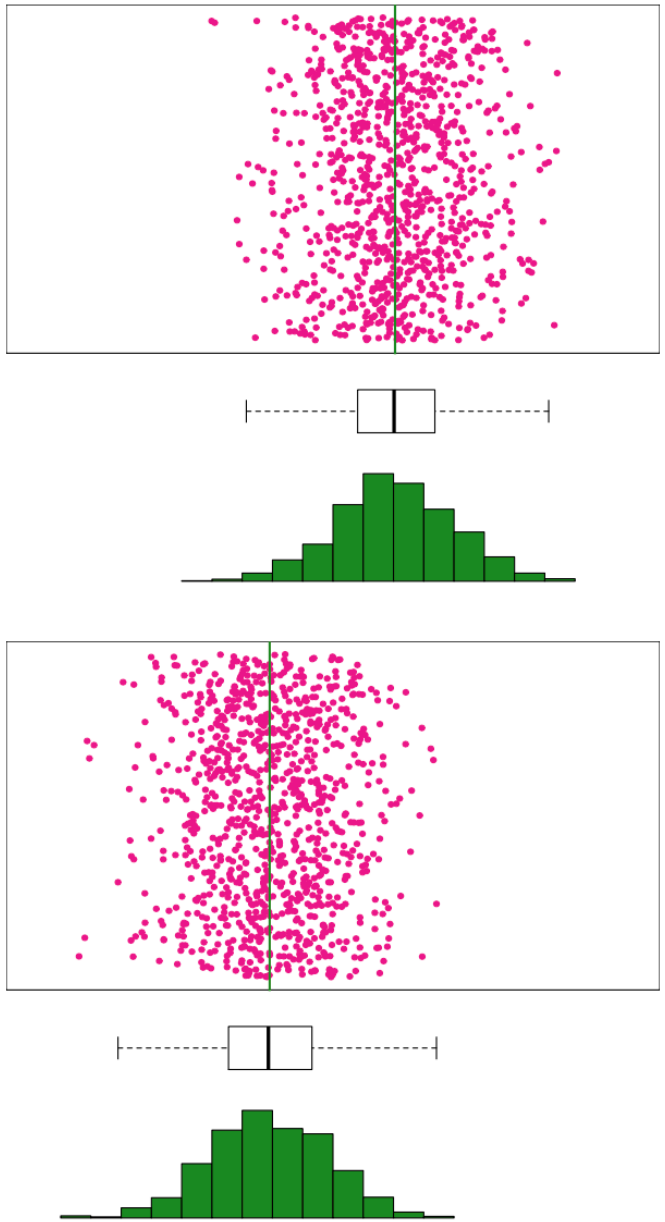
Visualizes the covariation of two continuous variables.

Best with small or medium sized data sets (although there are ways to present large data sets in scatterplots).

Most often used with:

- Paired numerical data
- Dependent variable with multiple independent variables
- Two variables that may be related or have a causal relationship.


hwy
40
30
20
2
3
4
5
6
7
displ

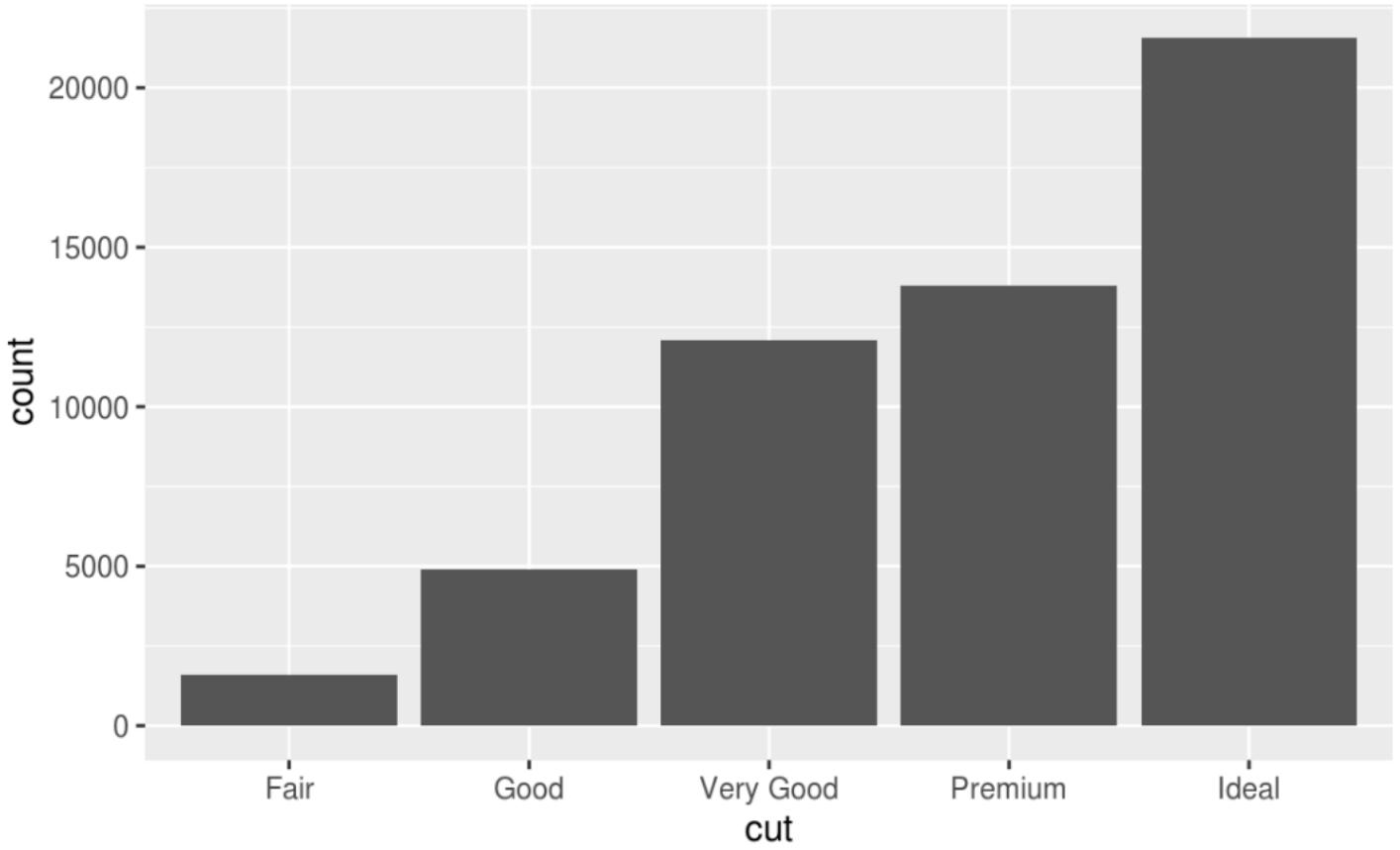


Bar chart

Unlike previous graphs, bar charts calculate new values to show.

Used for a dataset that includes both categorical and continuous data.

If you are creating a bar chart after *running an analysis*, make sure to include error bars.


20000
15000
10000
5000
0
count
Fair
Good
Very Good
Premium
Ideal
cut

Modeling your data

What is actually taking place?

Is your hypothesis supported or refuted?

- Make sure to explain in words and to report the statistical values

Keeping track of your statistics

Create a text document

- Write out each hypothesis in the order you are testing them.
- Write out the statistics you are running to test that hypothesis.
- Put the findings, including statistical values, in the document.

Save syntax (if you are running a program where that is possible)

RStudio

Download at https://www.rstudio.com

R: programming language and environment for statistical analyses

RStudio: user interface for R that includes code editor, debugging & visualization tools

R Studio

4 parts in R Studio opening screen:

- Editor
- Console
- global environment
- output

R Studio

Console:

- Type command
- View output
- Once a file is opened, the console splits into two parts with the editor on the top

Global environment:

- Environment: shows you your data variables & stores everything you create in R
- History**:** stores all commands used in current session

Output:

- File: default workspace
- Plots: shows the graphs created
- Packages: lists and installs option add-ons
- Help: search user manual
- Viewer: displays local web content

RStudio

Go to file/function | Addins | Project: (None)

Untitled1

Source on Save | Run | Source

```
1
```

1:1 (Top Level) R Script

Console Terminal

```
87          ford        f150 pickup 4wd   4.6 1999    8   auto(l4)   4  13  16  r      pickup
88          ford        f150 pickup 4wd   4.6 2008    8   auto(l4)   4  13  17  r      pickup
89          ford        f150 pickup 4wd   5.4 1999    8   auto(l4)   4  11  15  r      pickup
90          ford        f150 pickup 4wd   5.4 2008    8   auto(l4)   4  13  17  r      pickup
 [ reached getOption("max.print") -- omitted 144 rows ]
> ?mpg
> ggplot(data=mpg) + geom_point(mapping = aes(x=displ, y = hwy))
> ggplot(data=mpg) + geom_point(mapping = aes(x=displ, y = hwy, color = class)
+
+
```

Environment History Connections

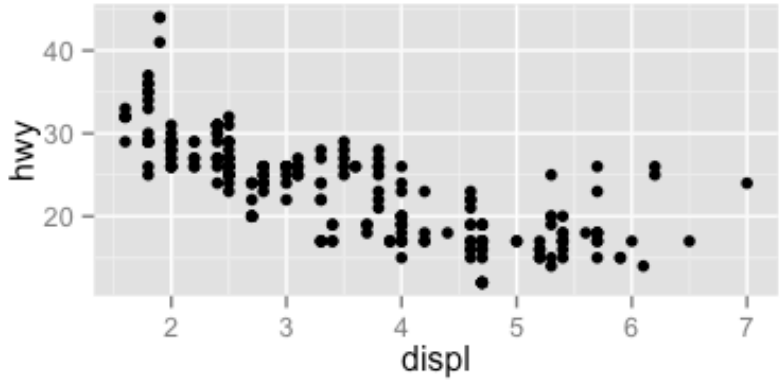
Import Dataset | List

Global Environment

Environment is empty

Files Plots Packages Help Viewer

Zoom | Export | Publish



Some useful hints

Alt-shift-k – pulls up keyboard shortcut quick reference

Capitalization and punctuation matters but spacing does not

For more information on a command, type ? followed by command name

Esc – exits out of a command not running

Loading data

Can load from any source

Can extract a variable from a dataset, attach a dataset, or detach a dataset

Can create a subset of a dataset

Can take a random sample of a specified size

Set/check working directory and load data

setwd()

- Set working directory
- E.g. setwd("~Desktop/assignment 2 files")

getwd()

- Prints your current location

read.csv()

- Read in a csv data file
- E.g. classData.file<-read.csv('classData.csv')
  - *classData.csv is your data file*
  - *classData.file is the name you will refer to it in R*

View()

- See data displayed in top left panel
- E.g. View(classData.file)

Visualizing your data

One option for graphing: Ggplot2

Dplyr (data manipulation)

- Filter()
- Arrange()
- Select()
- Mutate()
- Summarize()

use with group_by()

R in-class demo

In the demo I ran, I used the following code for the correlation:

- cor.test(mpg$displ,mpg$hwy)
  - mpg is the name of the data set. You have the option to specify whether you want to set which data set you are using or you can specify it before each variable name as I did.
  - Also note that there are different types of correlation tests that can be run in R. Make sure you choose the correct one for your purposes. For example, if you are running many different correlations on one dataset, then you would want to run a function that accounts for that.

Class demo, correlation output

```
data:  mpg$displ and mpg$hwy
t = -18.1508, df = 232, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8142727 -0.7072539
sample estimates:
      cor
-0.76602
```

For reporting this finding: r(232) = -0.77, $p < 0.001$; Can also report R-squared (optional).

Class exercise / Take home assignment

1) Download R Studio

2) Choose a research area to look at from the National Center for Health Statistics website:

- https://www.cdc.gov/nchs/hus/contents2017.htm

3) You may want to explore the data a bit. Then,

- Come up with an hypothesis and:
  - Create a graph
  - Decide which analysis is the best to run
  - Optional: run the analysis and potentially modify the graph based on the findings
- Write up the findings. You may include actual findings or projected findings.
  - *The goal is to learn how to properly word your findings.