

Data visualization and a grammar of graphics



Overview

Quick review of dplyr

The grammar of graphics

ggplot



Midterm exam

Homework 5 has been and a practice midterm exam has been posted

Midterm exam is on Thursday October 10th **in person** during regular class time

- The exam is on paper
- If you have accommodations, please schedule to take your exam with SAS and let me know
- **Please post an exam practice question to Canvas by 3pm on Friday**

You are allowed an exam “cheat sheet” that you will turn in with your exam

- See last class video and announcement for guidelines



Very quick dplyr review

The **tidyverse** is a set of packages that makes it easy to process data frames

dplyr is a package that has a set of verbs for transformations data

- All these function **take a data frame** and other arguments and **return a data frame**

1. `filter()`
2. `select()`
3. `mutate()`
4. `arrange()`
5. `summarize()`
6. `group_by()`

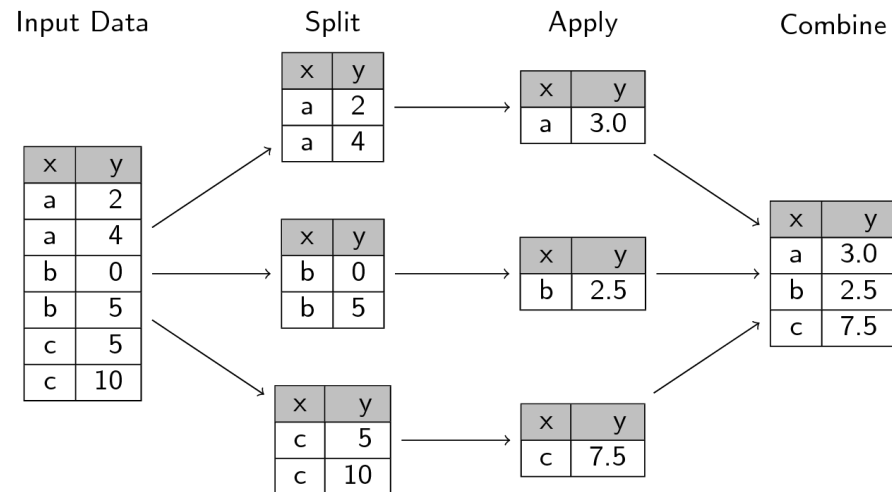


```
film_results <- movies |>
  filter(title_type == "Feature Film") |>
  select(critics_score, audience_score, genre) |>
  mutate(audience_prefs =
    audience_score - critics_score) |>
  group_by(genre) |>
  summarize(mean_audience_prefs =
    mean(audience_prefs)) |>
  arrange(desc(mean_audience_prefs))

head(film_results )
```

Very quick dplyr review: group_by

group_by: split, apply, combine



group_by multiple items:

```
group_by(genre, mpaa_rating) |>  
summarize(ms = mean(critics_score))
```

```
film_results <- movies |>
```

```
filter(title_type == "Feature Film") |>  
select(critics_score, audience_score, genre) |>  
mutate(audience_prefers =  
  audience_score - critics_score) |>  
group_by(genre) |>  
summarize(mean_audience_prefers =  
  mean(audience_prefers)) |>  
arrange(desc(mean_audience_prefers))
```

```
head(film_results )
```

Very quick dplyr review: summarize

One can summarize multiple variables:

```
summarize(mean_pref = mean(audience_prefers),  
           med_pref = median(audience_prefers))
```

One can use the `n()` function to count how many items are in each group

```
group_by(genre) |>  
summarize(num_genre = n())
```

```
film_results <- movies |>  
  filter(title_type == "Feature Film") |>  
  select(critics_score, audience_score, genre) |>  
  mutate(audience_prefers =  
          audience_score - critics_score) |>  
  group_by(genre) |>  
  summarize(mean_audience_prefers =  
             mean(audience_prefers)) |>  
  arrange(desc(mean_audience_prefers))  
  
head(film_results )
```


Homework 5, part 2: weather predictions

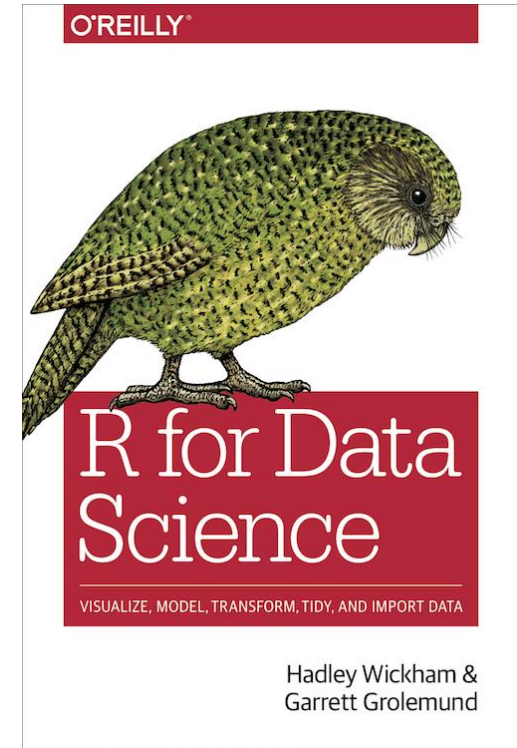
Assessing the accuracy and visualizing weather predictions



Steps:

1. What result do I want?
2. What steps can I take to get the result?
3. How can I implement these steps using dplyr?

Questions about dplyr?



Data visualization

Q: What are some reasons we visualize data rather than just reporting statistics?

*Statistical projections which **speak to the senses without fatiguing the mind**, possess the advantage of fixing the attention on a great number of important facts.*

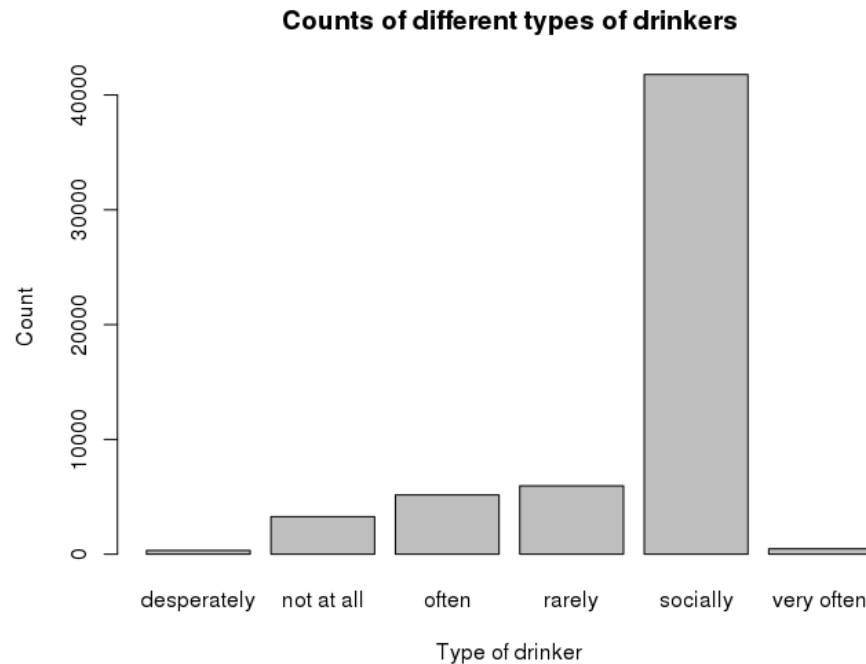
—Alexander von Humboldt, 1811



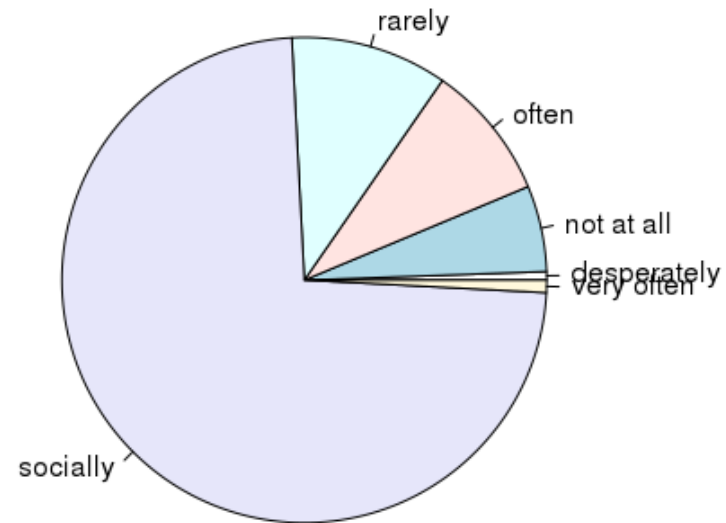
A grammar of graphics and ggplot

How have we plotted a single categorical variable?

Bar plot

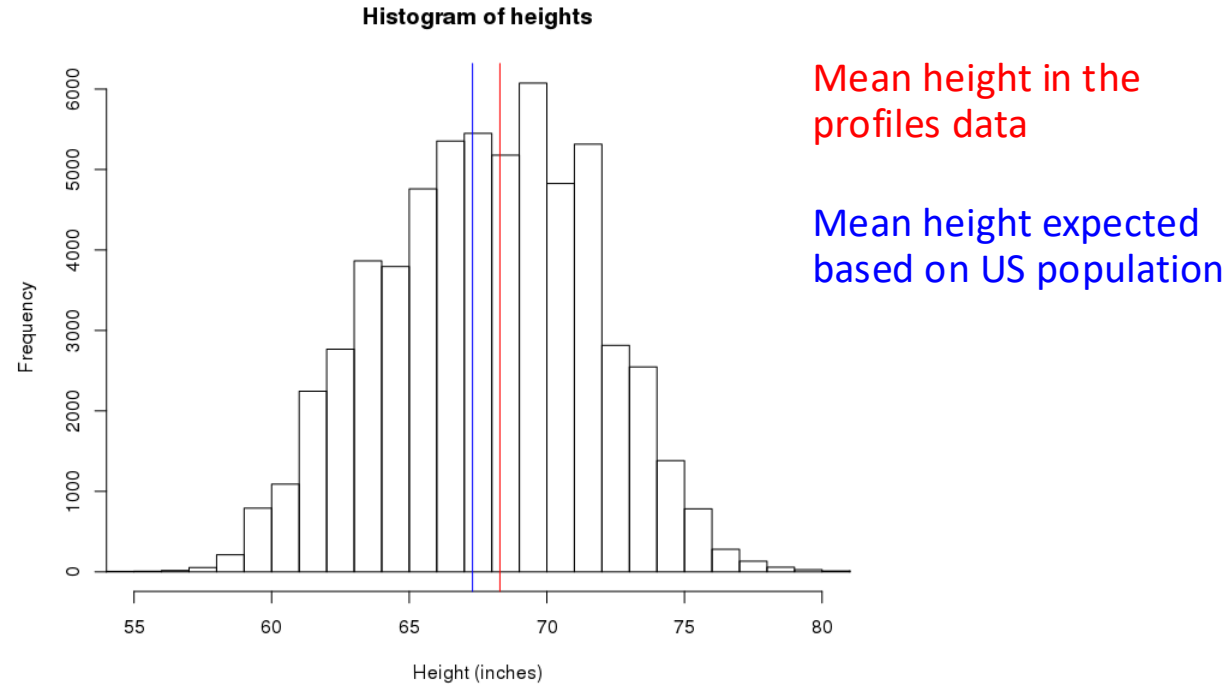


Pie chart



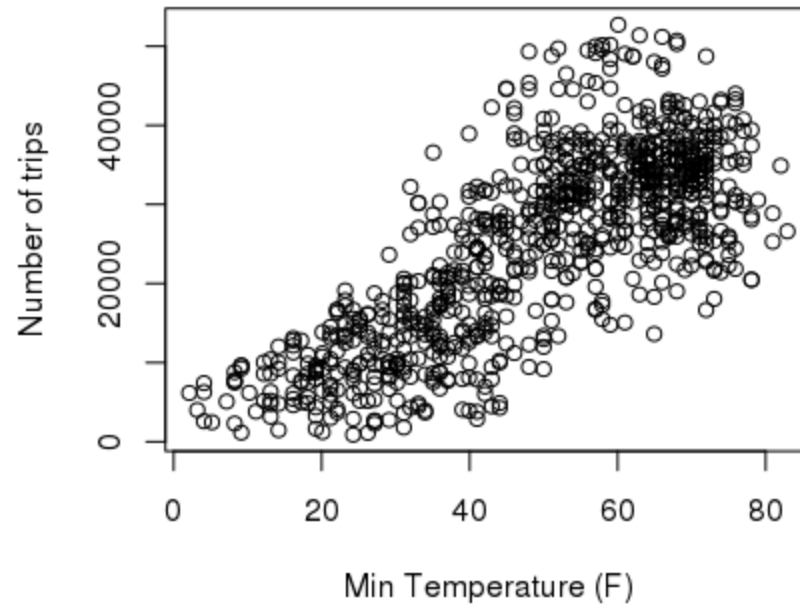
How have we plotted a single quantitative variable?

Histograms

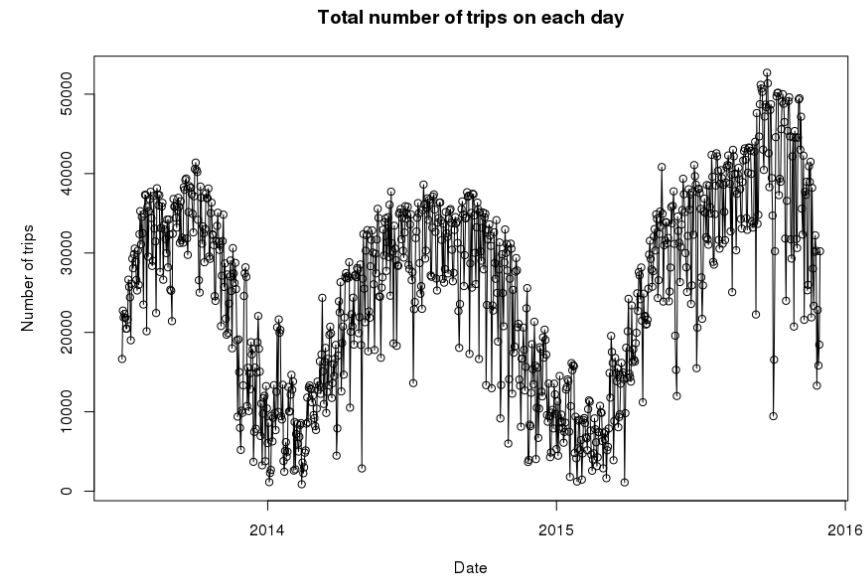


How have we plotted a two quantitative variables?

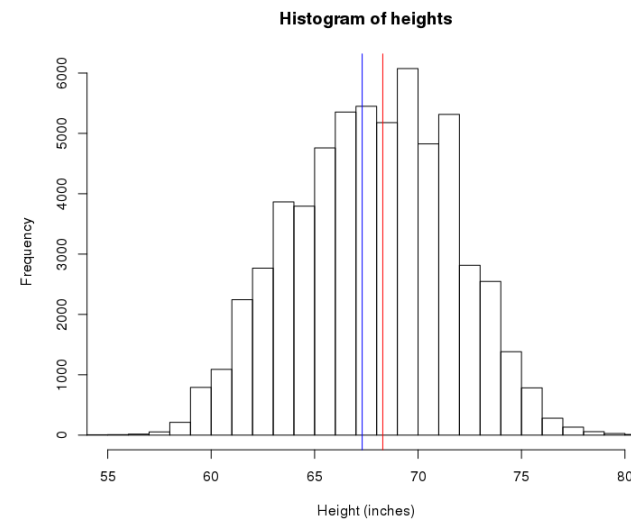
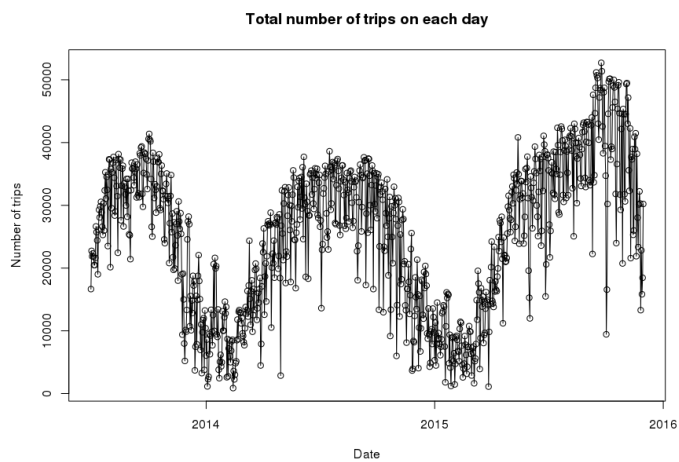
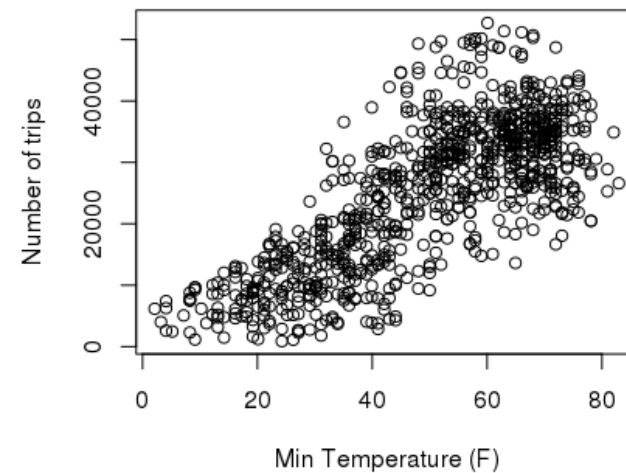
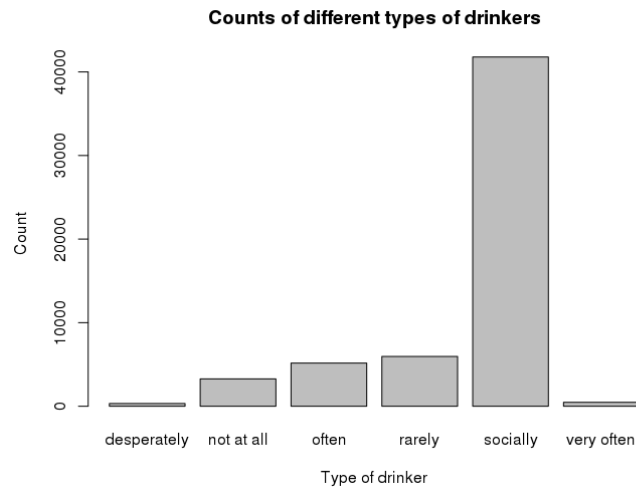
Scatter plots



Line chart



What are some similarities between these graphs?

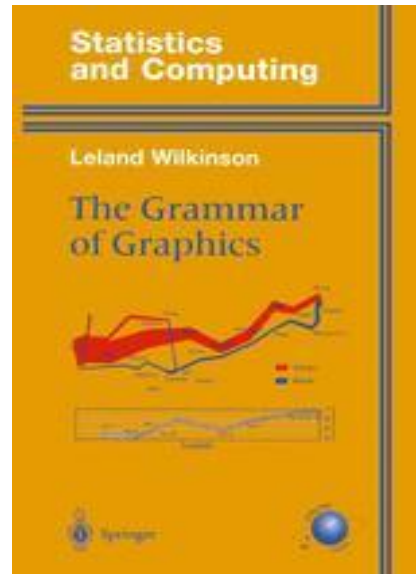


The grammar of graphics

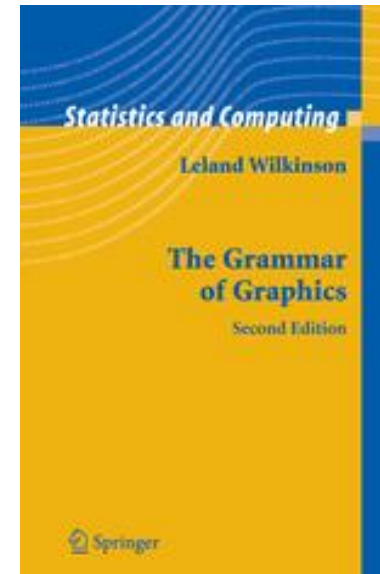
Leland Wilkinson noticed similarities between many graphs and tried to generate a 'grammar' that could be used to express a graph

- i.e., a list elements that can be combined together to create a graph

First edition



Second edition



Graphs are composed of...

A Frame: Coordinate system on which data is placed

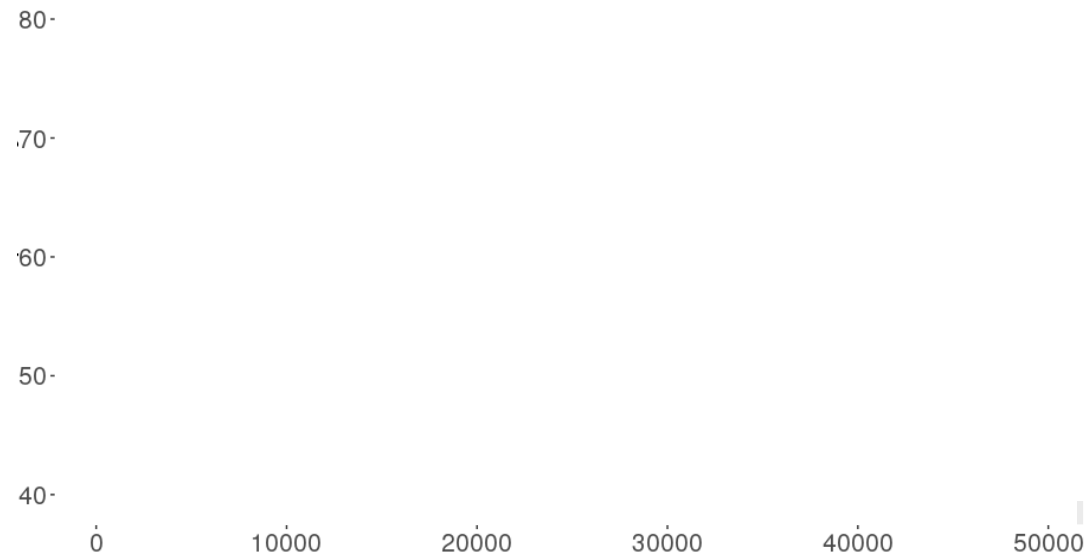
- E.g., Cartesian coordinate system, polar coordinates, etc.

Glyphs: basic graphic unit representing cases or statistics

- Contains visual properties (aesthetics) such as: shape, color, size, etc.
- Need to specify how properties of the data are **mapped** onto these aesthetics

Scales and guides: shows how to interpret axes and other properties of the glyphs

- i.e., tells us how the data values are mapped into glyph properties



Plots can also contain...

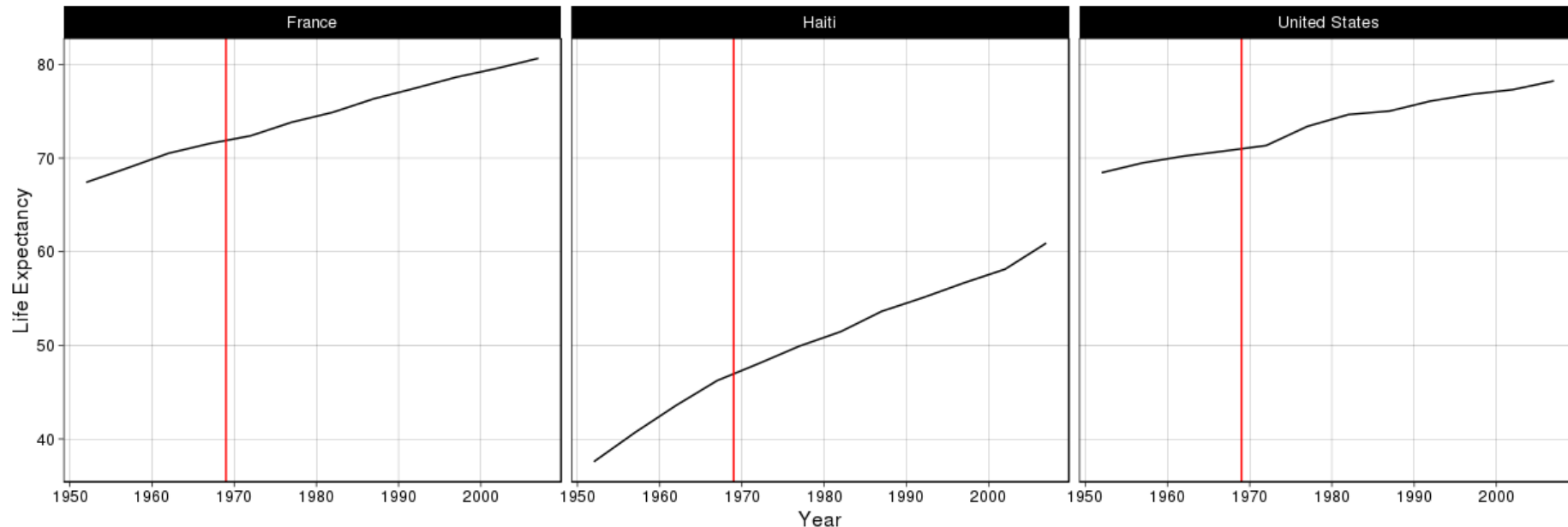
Facets: allows for multiple side-by-side graphs based on a categorical variable

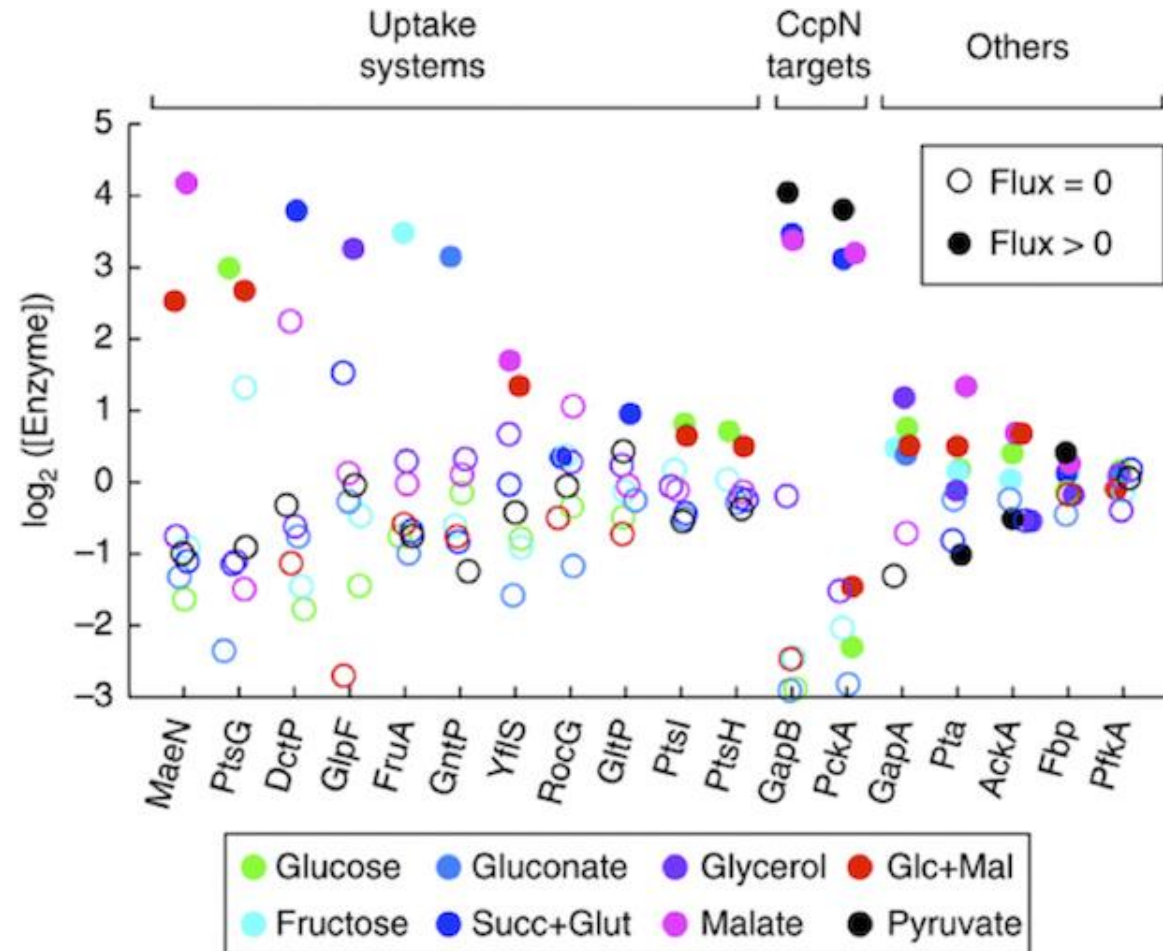
- Makes it easier to compare different conditions

Layers: allows for more than one types of data to be mapped onto the same figure

Theme: contains finer points of display

- E.g., font size, background color, etc.





The variables are:

1. Log enzyme concentration
 - -3 to 5
2. Gene
 - MaeN, PtsG, ...
3. Target
 - CcpN, Uptake,...
4. Flux
 - Zero or positive
5. Molecule:
 - Glucose, Fructose, ...

What are the mappings between each variable and visual attribute?

ggplot

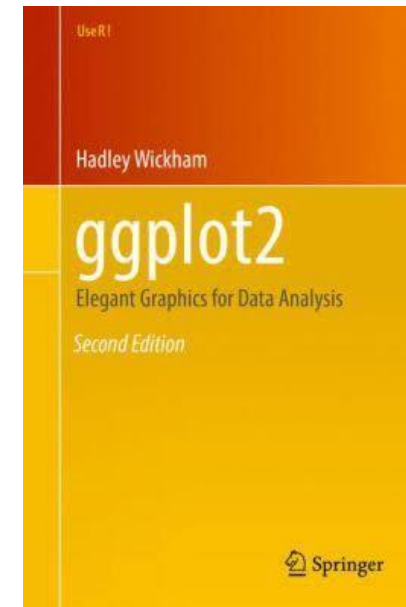
ggplot2 is an R package that implements the grammar of graphics

- It builds up graphics by starting with a frame, adding glyphs, etc.

```
# load the ggplot2 library
```

```
> library('ggplot2')
```

[Get the book on GitHub](#)



Example data: mtcars

ROAD TEST By Jim Brokaw

THE LUXURY CARS

Imperial Palace
Fortress Fleetwood
Castle Continental

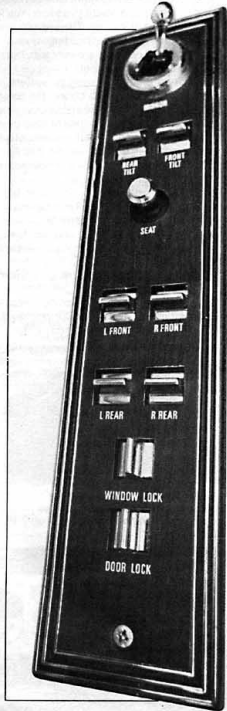
Crippled by the fuel flap and en-
gined at lewdly by those smug
Mercedes owners, today it seems that
these great mastadons are dismissed as
symbols of an ancient aristocracy whose
strata was marked by expanse of wheel-
base, and the heavenly quantity of gross

cubic mass able to be shouldered by
four beleaguered tires. As the sole sur-
viving heirs of princely Packards, dy-
nosaurean Duesenbergs and the Brob-
dingnagian Bugattis, these marvels of
grand proportion should be headed for
the Smithsonian Institute by way of the
mucky La Brea tar pits.

But are they?
Are these behemoths simply vestiges
of a defunct class of affluence and in-
fluence, or are they firmly entrenched—
perhaps now more than ever—in their
positions atop the very peak of slippery

nipulation and partisan hatcheteering in
government. They leave us little to be-
lieve in, and less to trust.

The very qualities that appear to con-
demn these sail-less luxury liners will
very likely ensure their perpetuation. In
days past, the block-long Cadillac and
shiny Lincoln with paint jobs three feet
deep flaunted a socio-economic posi-
tion we could never hope to achieve.



Lincoln's placement of seat controls on
arm rest panel is less desirable than
lower side-of-seat location of Cad
and Imperial.

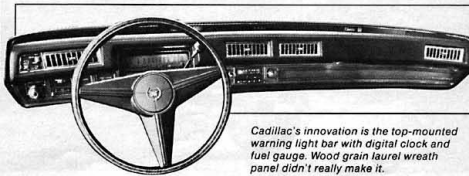


Imperial's seats featured amazingly
soft kid-glove leather, but lacked support.

Mount Status, whose crags we scale
daily, whether we wish to acknowledge
that fact or not?

Ah, but welcome to the current state of
American affairs: boot prices manipu-
lated by withholding the animals from
market; endless rounds of strikes by
unions whose members are engaged in
serving the public; gasoline supplies
that rise and fall in mysterious coinci-
dence with rising prices; dry rot, ma-

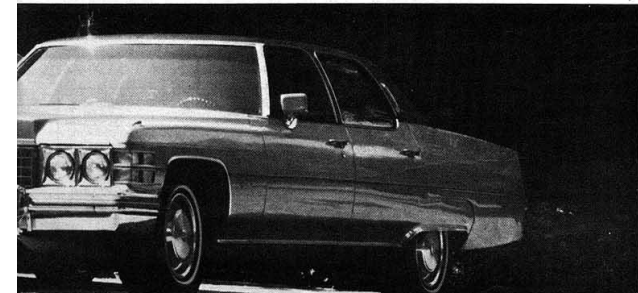
They did, however, constantly remind us
that such positions, such wealth, such
power, did, in fact, exist. The gliding
specter of the shiny Imperial eagle
stirred within a few heretical souls the
bold idea that if such positions of power
and wealth existed, there must be some
means of attainment. More than a few of
the haughty, distant drivers of the velvet
tanks clawed their way up from the very
pavement they now whisper over to



Cadillac's innovation is the top-mounted
warning light bar with digital clock and
fuel gauge. Wood grain laurel wreath
panel didn't really make it.

JUNE 1974 39

PERFORMANCE	CADILLAC	LINCOLN	IMPERIAL
Acceleration			
0-30 mph	4.30	3.97	4.2
0-50 mph	8.49	8.00	9.15
0-60 mph	12.00	9.50	12.1
Standing Start 1/4-mile Mph	77.05	77.65	80.28
Elapsed time	17.98	17.82	17.42
Passing speeds			
40-60 mph	6.58	5.9	7.1
50-70 mph	7.00	6.8	6.8
Stopping distance			
From 30 mph	32'1"	31'4"	27'5"
From 60 mph	182'7"	153'10"	129'3"
Gas mileage range	10.43	10.42	14.7
Width—in.	79.8	80.0	79.7
Front Track—in.	63.5	64.3	64
Rear Track—in.	63.3	64.3	63.7
Wheelbase—in.	133.0	127.0	124.0
Overall length—in.	233.7	232.6	231.1
Height—in.	55.6	55.4	54.7
Curb Weight—lbs.	5,250	5,425	5,345
Fuel Capacity—gals.	27	22.5	25
Oil Capacity—qts.	4 (1)	4 (1)	4 (1)
Storage Capacity—cu. ft.	19.27	20.9	20+
Base Price	\$9,312	\$7,637	\$7,062
Price as tested	\$11,435	\$9,452	\$8,737
Engine:	OHV V-8	OHV V-8	OHV V-8
Bore & Stroke—in.	4.3x4.06	4.36x3.85	4.32x3.75
Displacement—cu. in.	472	460	440
HP @ RPM	205 @ 3600	215 @ 4000	230 @ 4000
Torque: lbs.-ft. @ rpm	365 @ 2000	350 @ 2600	350 @ 3200
Compression Ratio	8.25:1	NA	8.2:1
Carburetion	4V	4V	4V
Transmission	Auto. Turbo Hydra-Matic	Auto. Select Shift	Auto. Torqueflite
Final Drive Ratio	2.93	3.00	3.23 (?)
Steering Type	Recirculating Ball & Nut Power	Recirculating Ball & Nut With Integral Power Unit	Recirculating Ball Power
Steering Ratio	17.8-9.0	21.6 To 1	18.9:1
Turning Diameter (curb-to-curb-ft.)	(Wall To Wall) 24.54'	46.7"	44.69'
Wheel Turns (lock-to-lock)	2.83	3.99	3.5
Tire Size	LR78X15 Steel Belted Radials	LR78X15 Steel Belted Radials	LR78X15 Steel Belted Radial Ply
Brakes	Power Disc/Drum	Power Disc/Drum	Power Disc/Disc
Front Suspension	Coils/Shocks Front Diagonal Tie Struts Stabilizer	Coils/Shocks Axial Strut Stabilizer	Torsion Bar Shocks Stabilizer
Rear Suspension	4 Link, Coils/ Shocks	Three Link, Rubber Cushioned Pivots Coils/Shocks	Leaf Springs Shocks
Body/Frame Construction	Perimeter Frame	Body On Perimeter Frame	Unitized Construction



mtcars data frame

How can you determine what variables are in a data frame?

```
> View(mtcars)    # only works in Rstudio, not in Markdown  
> glimpse(mtcars)  
> ? mtcars       # this data frame as a code book
```

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 4]	hp	Gross horsepower
[, 6]	wt	Weight (1000 lbs)
[, 9]	am	Transmission (0 = automatic, 1 = manual)

Do cars that weigh more use more fuel?

Question: do cars that weigh more use more fuel?

What variables in the mtcars data frame are of interest?

- mpg
- wt

We can create a scatter plot using base graphics...

```
> plot(mtcars$wt, mtcars$mpg)
```

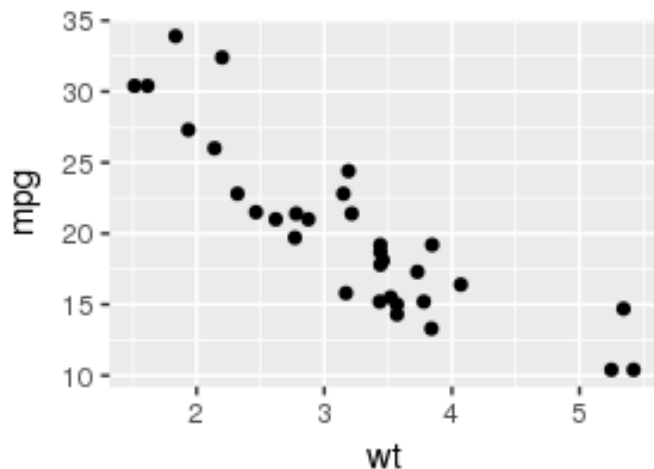
Creating a scatter plot in ggplot

Data frame to be used

Aesthetic mapping

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

Adds a layer with glyphs



	wt	cyl	hp	mpg	disp
Mazda RX4	2.620	6	110	21.0	160.0
Mazda RX4 Wag	2.875	6	110	21.0	160.0
Datsun 710	2.320	4	93	22.8	108.0
Hornet 4 Drive	3.215	6	110	21.4	258.0
Hornet Sportabout	3.440	8	175	18.7	360.0

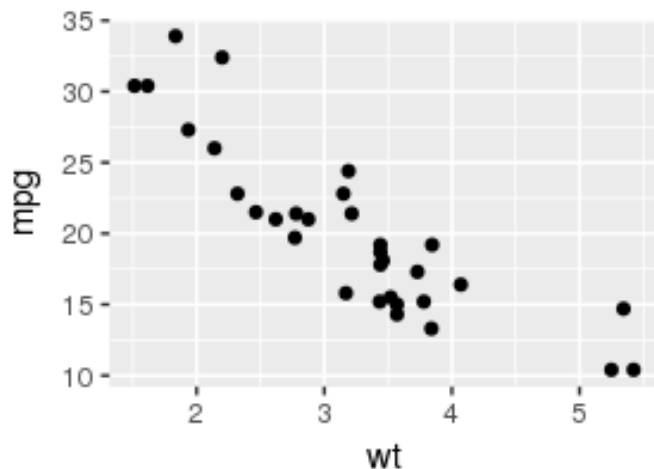
Creating a scatter plot in ggplot

Data frame to be used

Aesthetic mapping

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

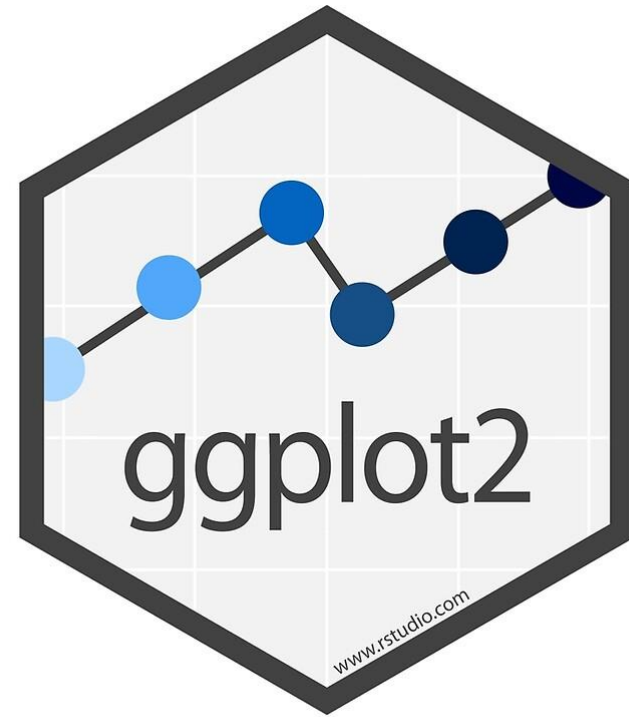
Adds a layer with glyphs



	wt	cyl	hp	mpg	disp
Mazda RX4	2.620	6	110	21.0	160.0
Mazda RX4 Wag	2.875	6	110	21.0	160.0
Datsun 710	2.320	4	93	22.8	108.0
Hornet 4 Drive	3.215	6	110	21.4	258.0
Hornet Sportabout	3.440	8	175	18.7	360.0

A lot more that ggplot can do!

- More aesthetic mapping
- Multiple glyphs/layers
- Axis labels
- Facets
- Visual themes
- Different coordinate systems
- Etc.



[The R Graph Gallery](http://www.rstudio.com)

Let's try the rest in R!

If life gives you questions, Google
gives you answers.

