Visualise!

Hadley Wickham

@hadleywickham

Chief Scientist, RStudio

July 2015

Outline

- Warmups
- Getting started
- Basic plots
- RStudio tips
- Where next

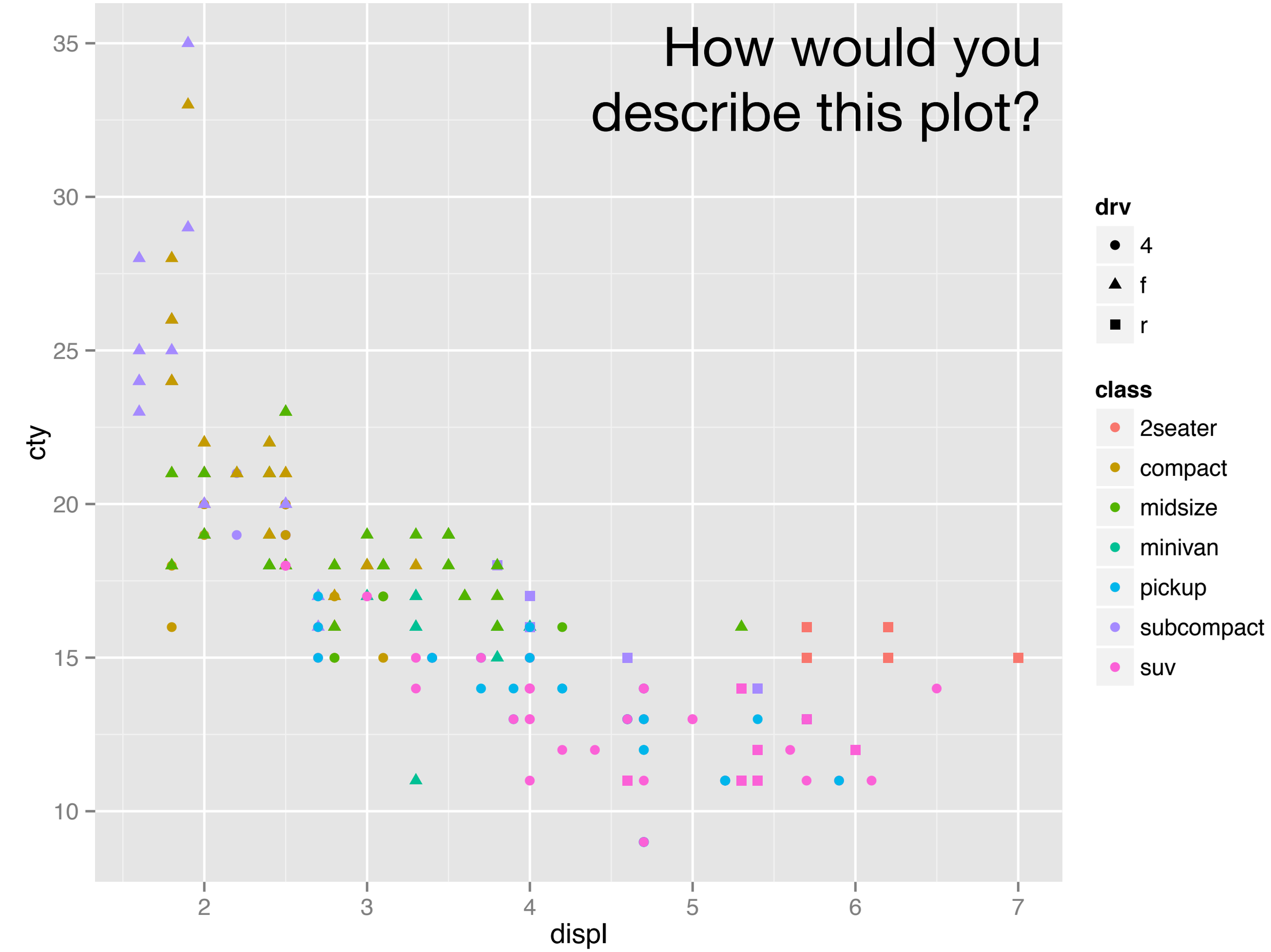
Warmups

Your turn

What is a plot?

What are the three most important types of plot?

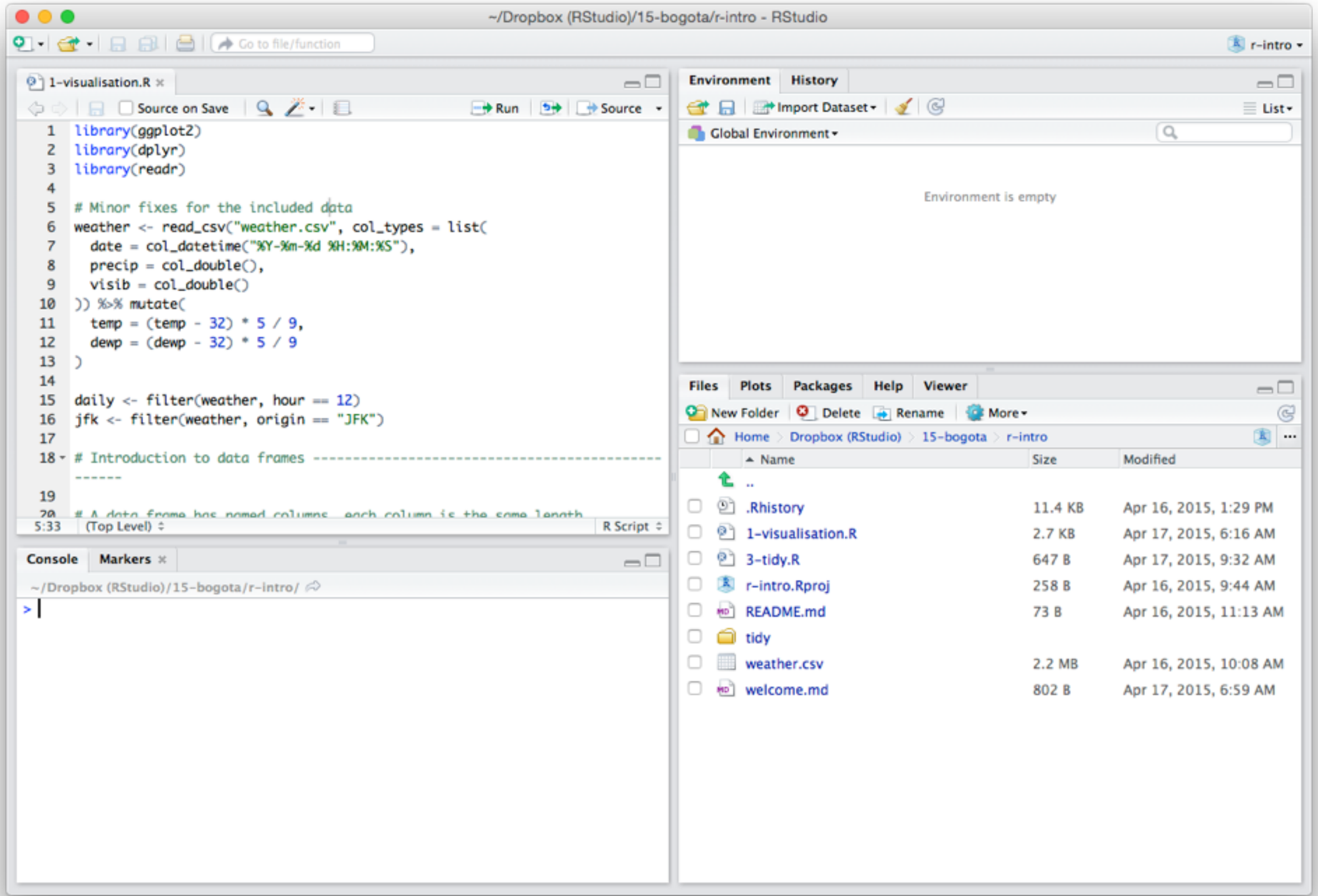
How would you describe this plot?



Getting started

Open

code-data.Rproj


~/Dropbox (RStudio)/15-bogota/r-intro - RStudio
Go to file/function
r-intro
1-visualisation.R
Source on Save
Run
Source
1 library(ggplot2)
2 library(dplyr)
3 library(readr)
4
5 # Minor fixes for the included data
6 weather <- read_csv("weather.csv", col_types = list(
7 date = col_datetime("%Y-%m-%d %H:%M:%S"),
8 precip = col_double(),
9 visib = col_double()
10)) %>% mutate(
11 temp = (temp - 32) * 5 / 9,
12 dewp = (dewp - 32) * 5 / 9
13)
14
15 daily <- filter(weather, hour == 12)
16 jfk <- filter(weather, origin == "JFK")
17
18 # Introduction to data frames ----------
19
5:33
(Top Level)
R Script
Console
Markers
~/Dropbox (RStudio)/15-bogota/r-intro/
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Dropbox (RStudio)
15-bogota
r-intro
Name
Size
Modified
..
.Rhistory 11.4 KB Apr 16, 2015, 1:29 PM
1-visualisation.R 2.7 KB Apr 17, 2015, 6:16 AM
3-tidy.R 647 B Apr 17, 2015, 9:32 AM
r-intro.Rproj 258 B Apr 16, 2015, 9:44 AM
README.md 73 B Apr 16, 2015, 11:13 AM
tidy
weather.csv 2.2 MB Apr 16, 2015, 10:08 AM
welcome.md 802 B Apr 17, 2015, 6:59 AM

~/Dropbox (RStudio)/15-bogota/r-intro - RStudio

Go to file/function

r-intro

1-visualisation.R

Source on Save | Run | Source

```
1  library(ggplot2)
2  library(dplyr)
3  library(readr)
4
5  # Minor fixes for the included data
6  weather <- read_csv("weather.csv", col_types = list(
7    date = col_datetime("%Y-%m-%d %H:%M:%S"),
8    precip = col_double(),
9    visib = col_double()
10 )) %>% mutate(
11   temp = (temp - 32) * 5 / 9,
12   dewp = (dewp - 32) * 5 / 9
13 )
14
15 daily <- filter(weather, hour == 12)
16 jfk <- filter(weather, origin == "JFK")
17
18 # Introduction to data frames ----------------------------------------
   ------
19
20 # A data frame has named columns, each column is the same length
```

5:33 (Top Level) R Script

Console | Markers

~/Dropbox (RStudio)/15-bogota/r-intro/

```
>
```

Environment | History

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Dropbox (RStudio) > 15-bogota > r-intro

	Name	Size	Modified
	..		
	.Rhistory	11.4 KB	Apr 16, 2015, 1:29 PM
	1-visualisation.R		Apr 17, 2015, 6:16 AM
	3-tidy.R	647 B	Apr 17, 2015, 9:32 AM
	r-intro.Rproj	258 B	Apr 16, 2015, 9:44 AM
	README.md	73 B	Apr 16, 2015, 11:13 AM
	tidy		
	weather.csv	2.2 MB	Apr 16, 2015, 10:08 AM
	welcome.md	802 B	Apr 17, 2015, 6:59 AM

Advice

Most of the answers to quizzes are in the slides.

Use them if you get stuck, but always try first.

You learn much better if you have tried and failed first.

Your turn

How do you load a package?

How do you install a package?

What's the difference?

```
# library() makes a package available for use

library(readr)
library(ggplot2)

# If you're missing any of these packages you can
# install with:
install.packages(c("ggplot2", "readr"))

# You only need to install once per computer.
# You need to library() every time you start R.
```

```
# Load some data

weather <- read_csv("weather.csv", col_types = list(
  date = col_datetime("%Y-%m-%d %H:%M:%S"),
  precip = col_double(),
  visib = col_double()
)) %>% mutate(
  temp = (temp - 32) * 5 / 9,
  dewp = (dewp - 32) * 5 / 9
)
```

Your turn

How can you see what's is this **data frame**?

Explore RStudio, and if you're familiar with R, share your expertise!

~/Dropbox (RStudio)/15-bogota/r-intro - RStudio

Go to file/function

r-intro

1-visualisation.R

Source on Save | Run | Source

```
1  library(ggplot2)
2  library(dplyr)
3  library(readr)
4
5  # Minor fixes for the included data
6  weather <- read_csv("weather.csv", col_types = list(
7    date = col_datetime("%Y-%m-%d %H:%M:%S"),
8    precip = col_double(),
9    visib = col_double()
10 )) %>% mutate(
11   temp = (temp - 32) * 5 / 9,
12   dewp = (dewp - 32) * 5 / 9
13 )
14
15 daily <- filter(weather, hour == 12)
16 jfk <- filter(weather, origin == "JFK")
17
18 # Introduction to data frames ----------------------------------------
   ------
19
20 # A data frame has named columns, each column is the same length
```

10:15 (Top Level) R Script

Console Markers

~/Dropbox (RStudio)/15-bogota/r-intro/

```
    intersect, setdiff, setequal, union

> library(readr)
>
> # Minor fixes for the included data
> weather <- read_csv("weather.csv", col_types = list(
+   date = col_datetime("%Y-%m-%d %H:%M:%S"),
+   precip = col_double(),
+   visib = col_double()
+ )) %>% mutate(
+   temp = (temp - 32) * 5 / 9,
+   dewp = (dewp - 32) * 5 / 9
+ )
>
```

Environment History

Import Dataset | List

Global Environment

Data

weather	26130 obs. of 15 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > Dropbox (RStudio) > 15-bogota > r-intro

Name	Size	Modified
..		
.Rhistory	11.4 KB	Apr 16, 2015, 1:29 PM
1-visualisation.R	2.7 KB	Apr 17, 2015, 6:16 AM
3-tidy.R	647 B	Apr 17, 2015, 9:32 AM
r-intro.Rproj	258 B	Apr 16, 2015, 9:44 AM
README.md	73 B	Apr 16, 2015, 11:13 AM
tidy		
weather.csv	2.2 MB	Apr 16, 2015, 10:08 AM
welcome.md	802 B	Apr 17, 2015, 6:59 AM

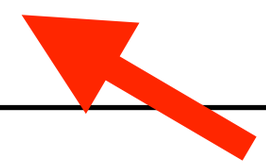
```
# Or in code

weather
View(weather)
str(weather)

# Hourly weather data for three New York City
# airports (we'll see them again later).
```

```
View(weather)
str(weather)
```

Your turn

What do `chr`, `int`, `num` and `POSIXct` stand for?

Abbreviation	Meaning
int	Integer
num	Numeric (real number)
chr	Character
POSIXct	Date/time (don't ask!)
Factor	Categorical (fixed set of values)



We'll avoid

```
weather <- mutate(weather, origin = factor(origin))
str(weather)
```

```
# If you have factors, almost always best to
# convert to character. We'll talk more about
# this later
weather <- mutate(daily,
  origin = as.character(origin))
```

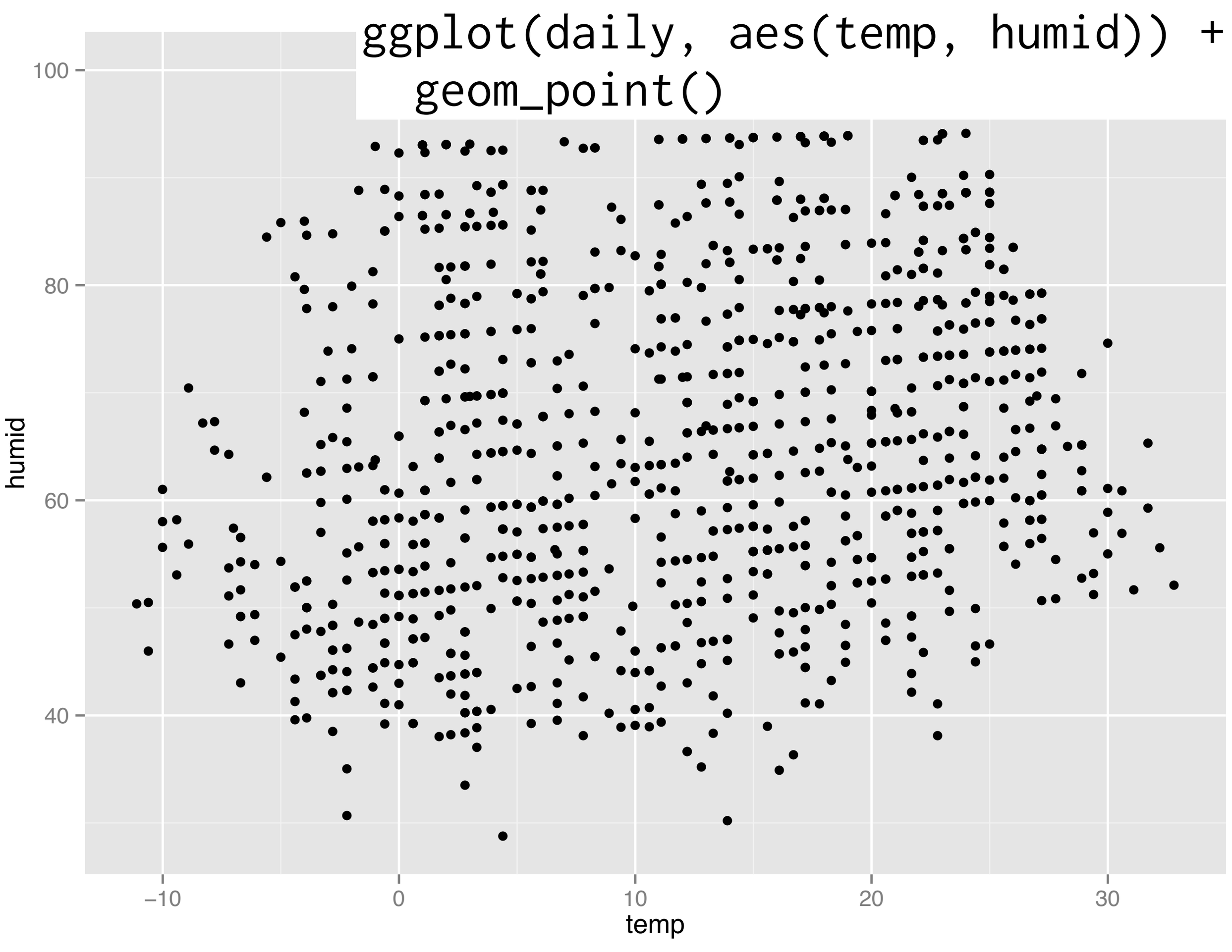
Basic plots

Most important plots

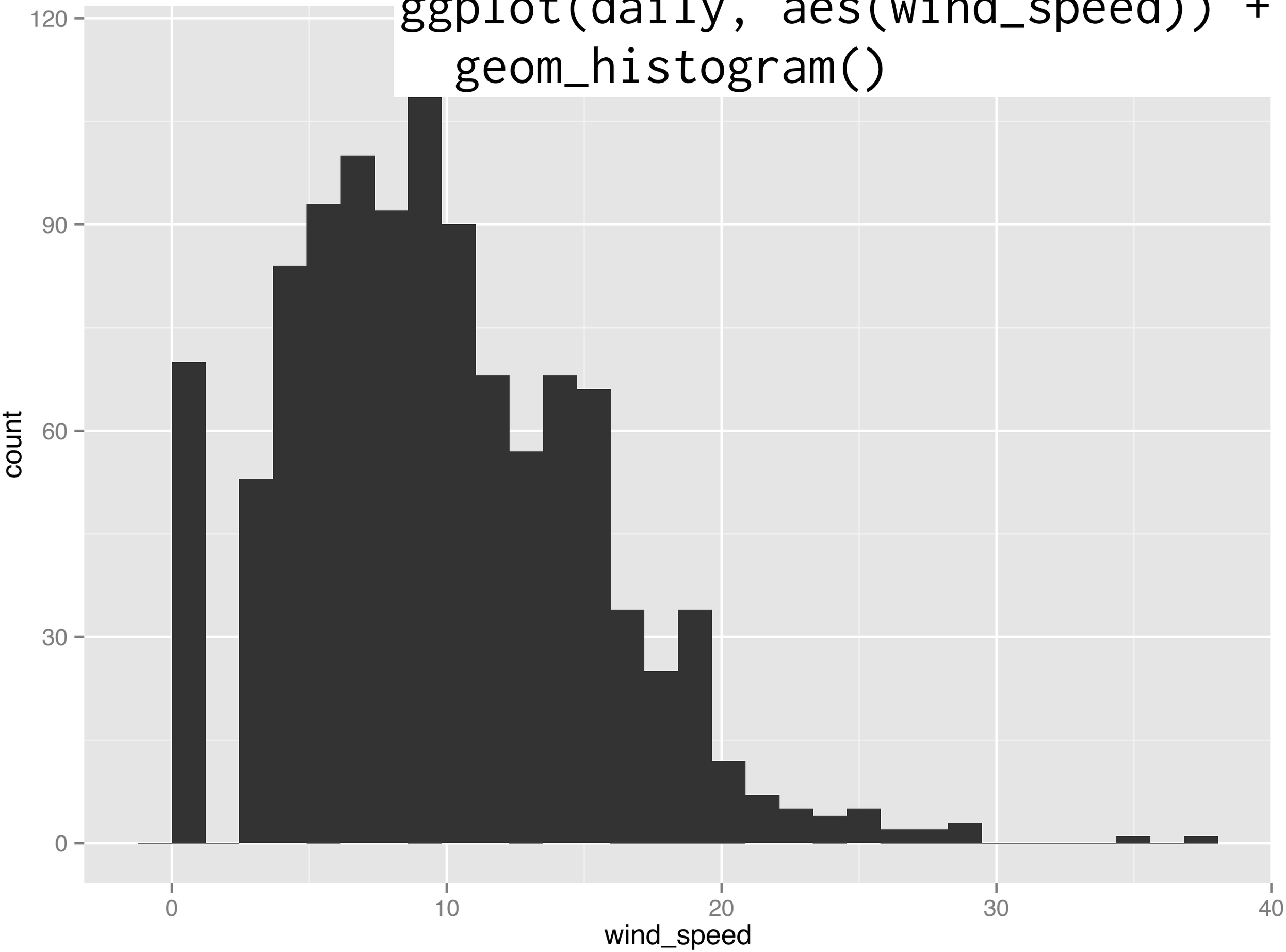
- Scatterplot: two continuous variables
- Histogram: one continuous variable
- Line chart: one time, one continuous
- Obviously many options missing! But these are a good place to start

```
# Two new variants: you'll learn how
# this code works tomorrow

# Just one measurement a day
daily <- filter(weather, hour == 12)

# Just one airport
jfk <- filter(weather, origin == "JFK")
```

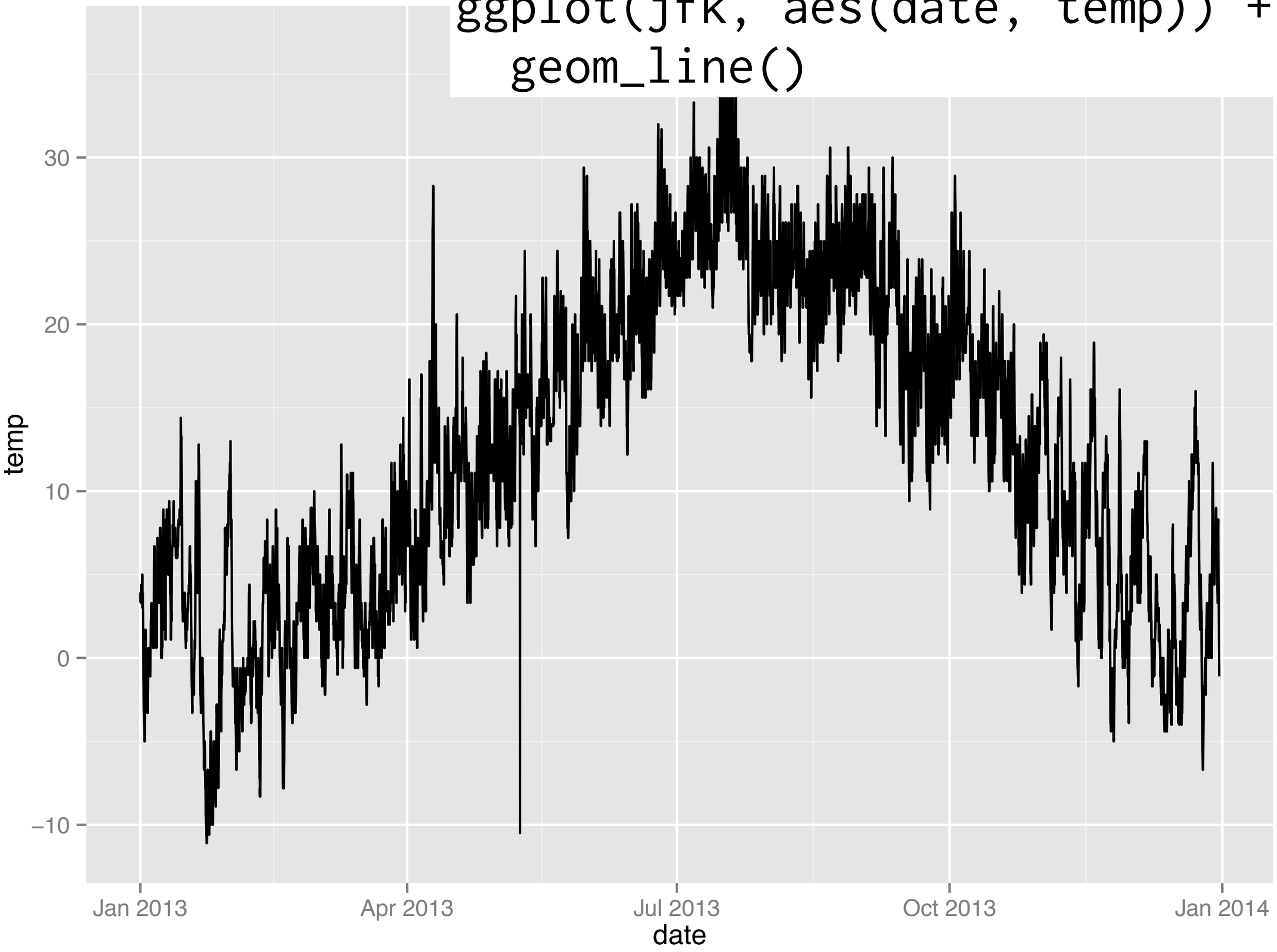
```
ggplot(daily, aes(temp, humid)) +
  geom_point()
```



```
ggplot(daily, aes(wind_speed)) +
  geom_histogram()
```



```
ggplot(jfk, aes(date, temp)) +
  geom_line()
```



Your turn

Compare and contrast the inputs and output of these three function calls.

What does the first argument to `ggplot()` do? What does the second argument do?

What does the geom function do?

Scatterplots

```
ggplot(daily, aes(temp, humid)) + geom_point()
# Add extra properties with shape or colour
ggplot(daily, aes(temp, humid, shape = origin)) +
  geom_point()
ggplot(daily, aes(temp, humid, colour = origin)) +
  geom_point()

# Your turn: create some more scatterplots
# What happens if you map a continuous variable
# or colour? Or shape? Or size?
# Do you find any interesting patterns?
```

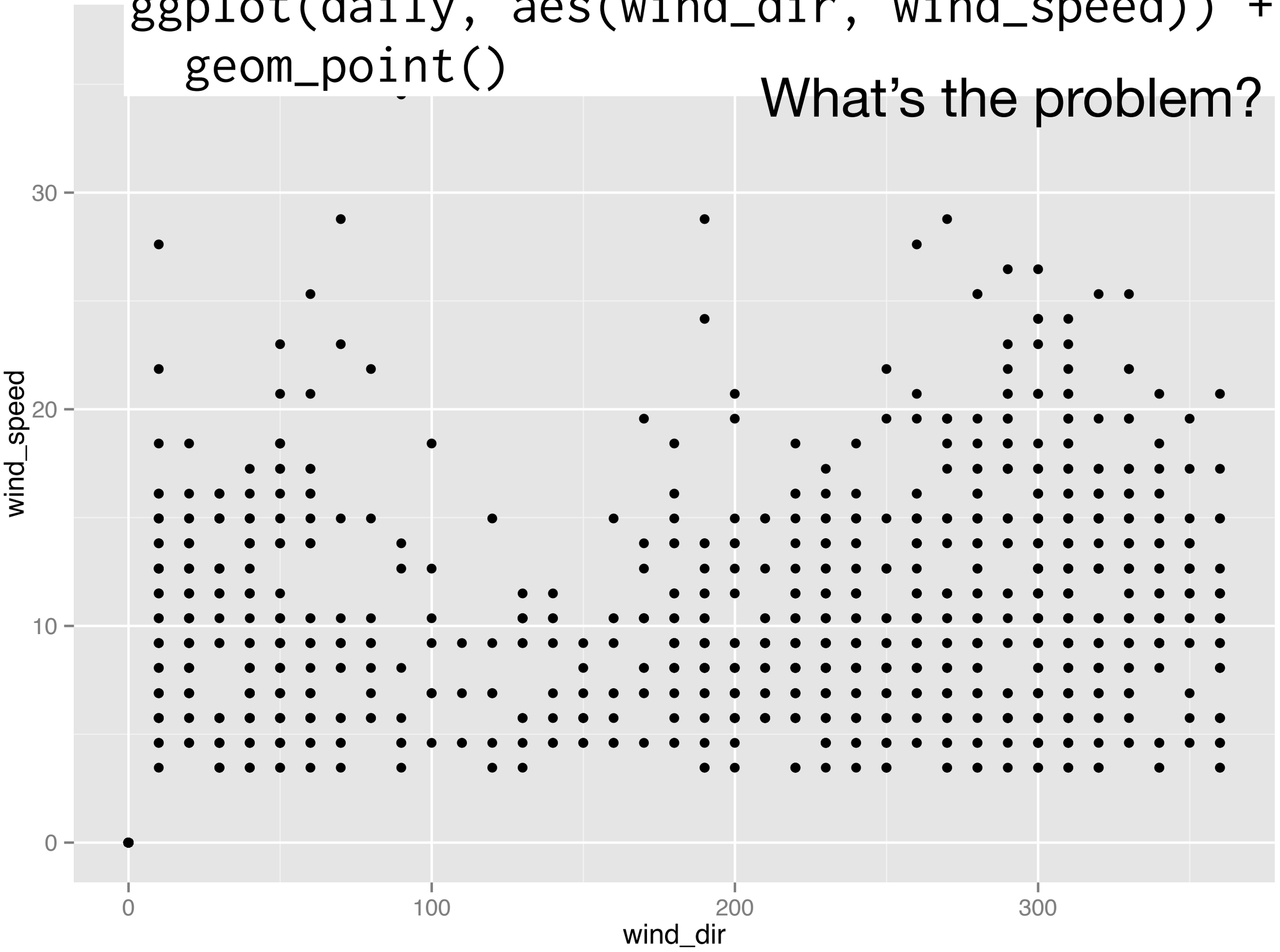
Learning a new language is hard!

When you get an error

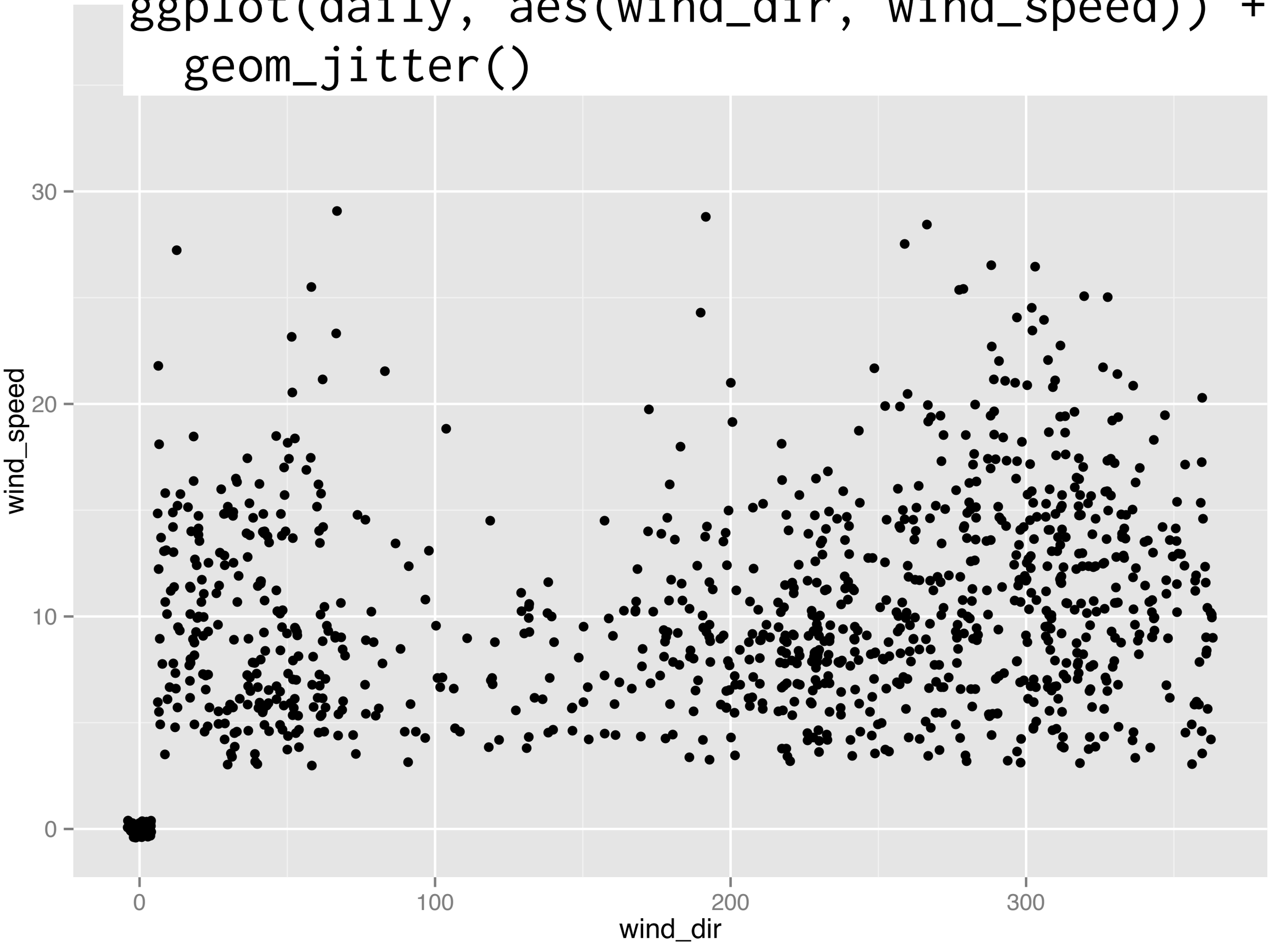
- Check that commas and spaces are in the right places
- Check that every " has a matching ", and every (has a matching)
- Read the error message!
- Still stuck? Google the error.

```
ggplot(daily, aes(wind_dir, wind_speed)) +
  geom_point()
```

What's the problem?



```
ggplot(daily, aes(wind_dir, wind_speed)) +
  geom_jitter()
```

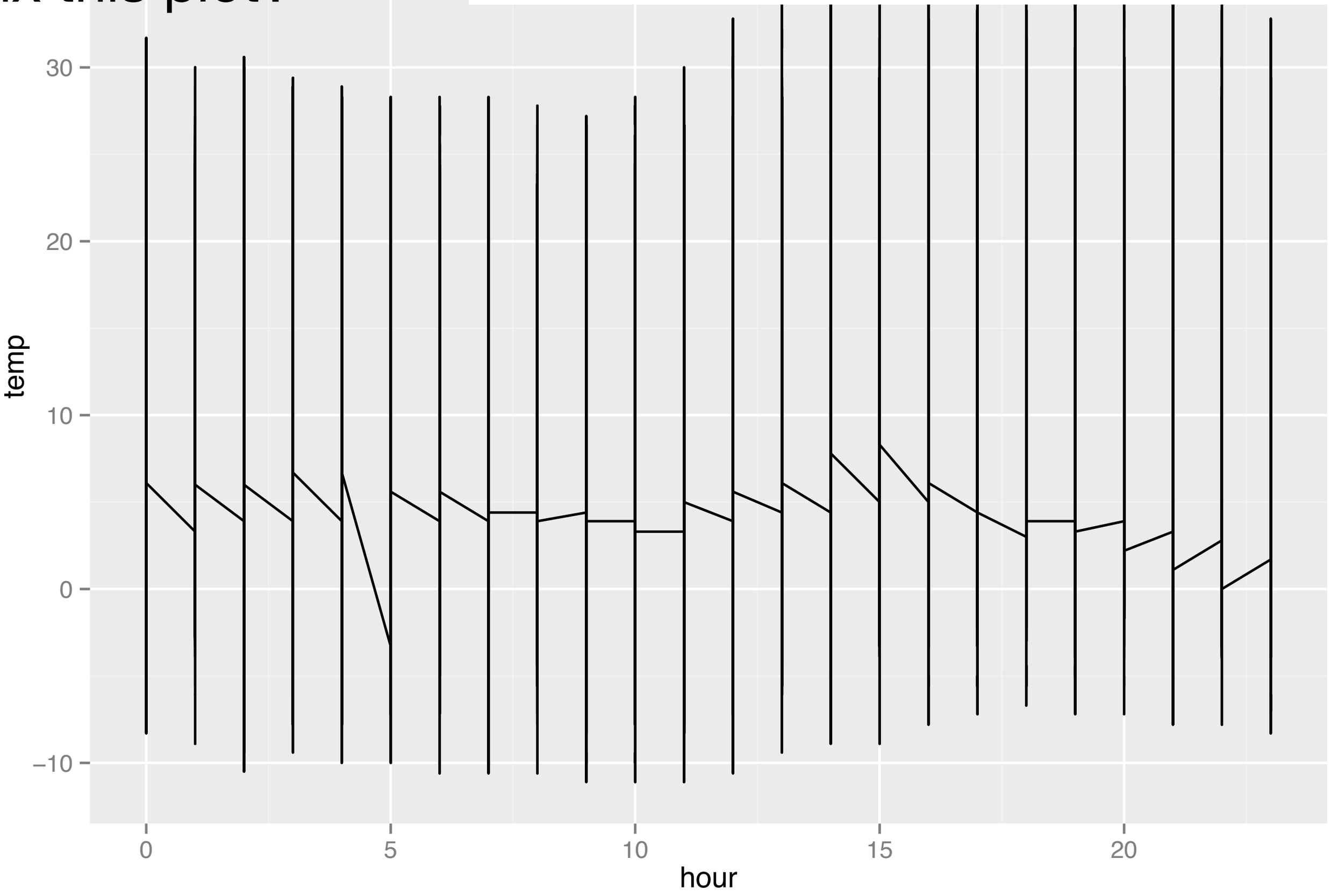


Line plots

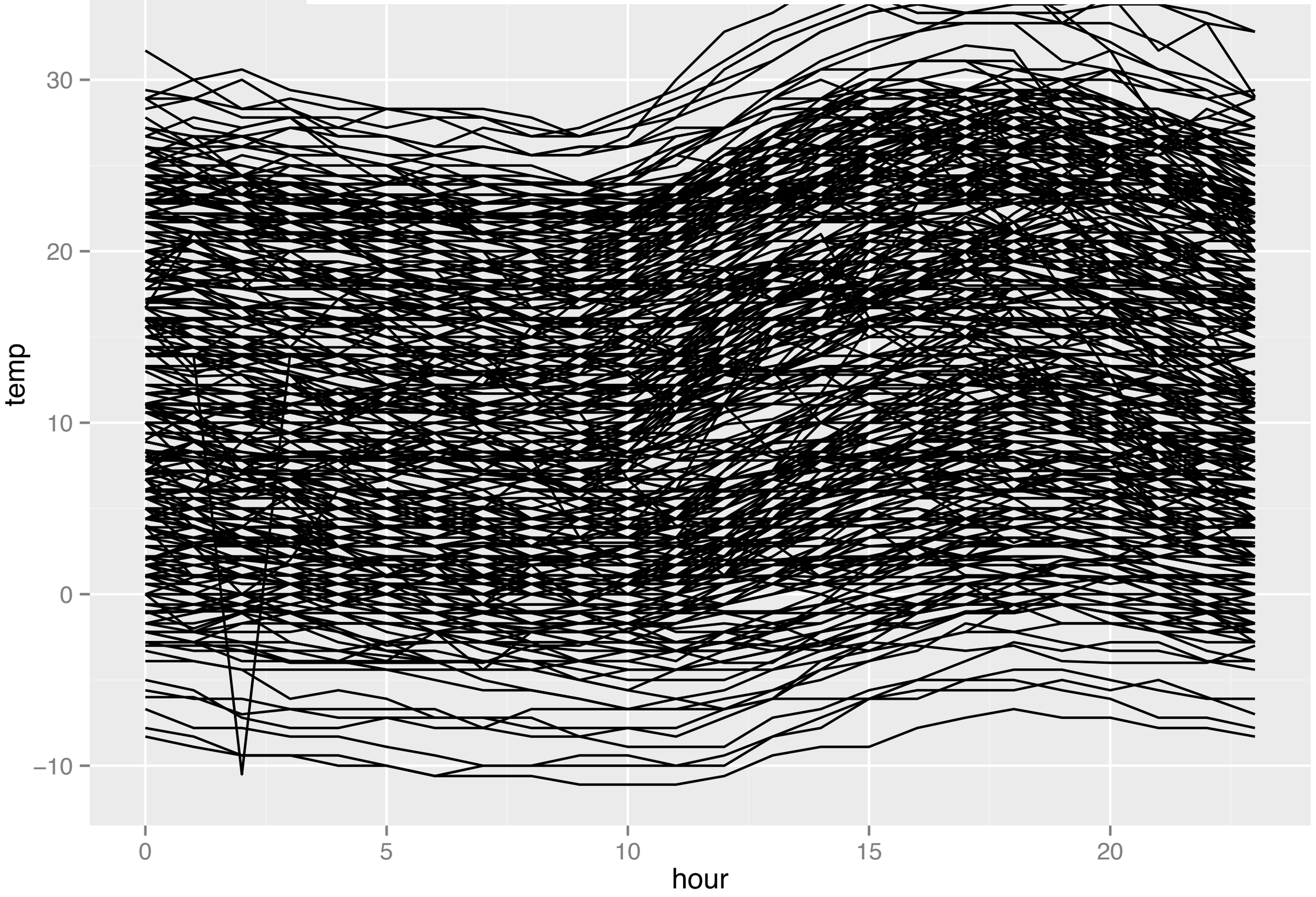
```
# Useful when you have a time series
ggplot(daily, aes(date, temp)) +
  geom_point()
ggplot(daily, aes(date, temp)) +
  geom_line()
ggplot(daily, aes(date, temp, group = origin)) +
  geom_line()
ggplot(daily, aes(date, temp, colour = origin)) +
  geom_line()

# Alternative to aesthetics is to use _facetting_
ggplot(daily, aes(date, temp)) +
  geom_line() +
  facet_wrap(~origin)
```

How could we fix this plot?

```
ggplot(jfk, aes(hour, temp)) +
  geom_line()
```



```
ggplot(jfk, aes(hour, temp)) +
  geom_line(aes(group = interaction(month, day)))
```



Your turn

How does humidity vary over the course of the day? How does it vary over the course of the year?

Histogram

```
ggplot(daily, aes(wind_speed)) + geom_histogram()

ggplot(daily, aes(wind_dir)) + geom_histogram()
ggplot(daily, aes(wind_dir)) +
   geom_histogram(binwidth = 10)
```

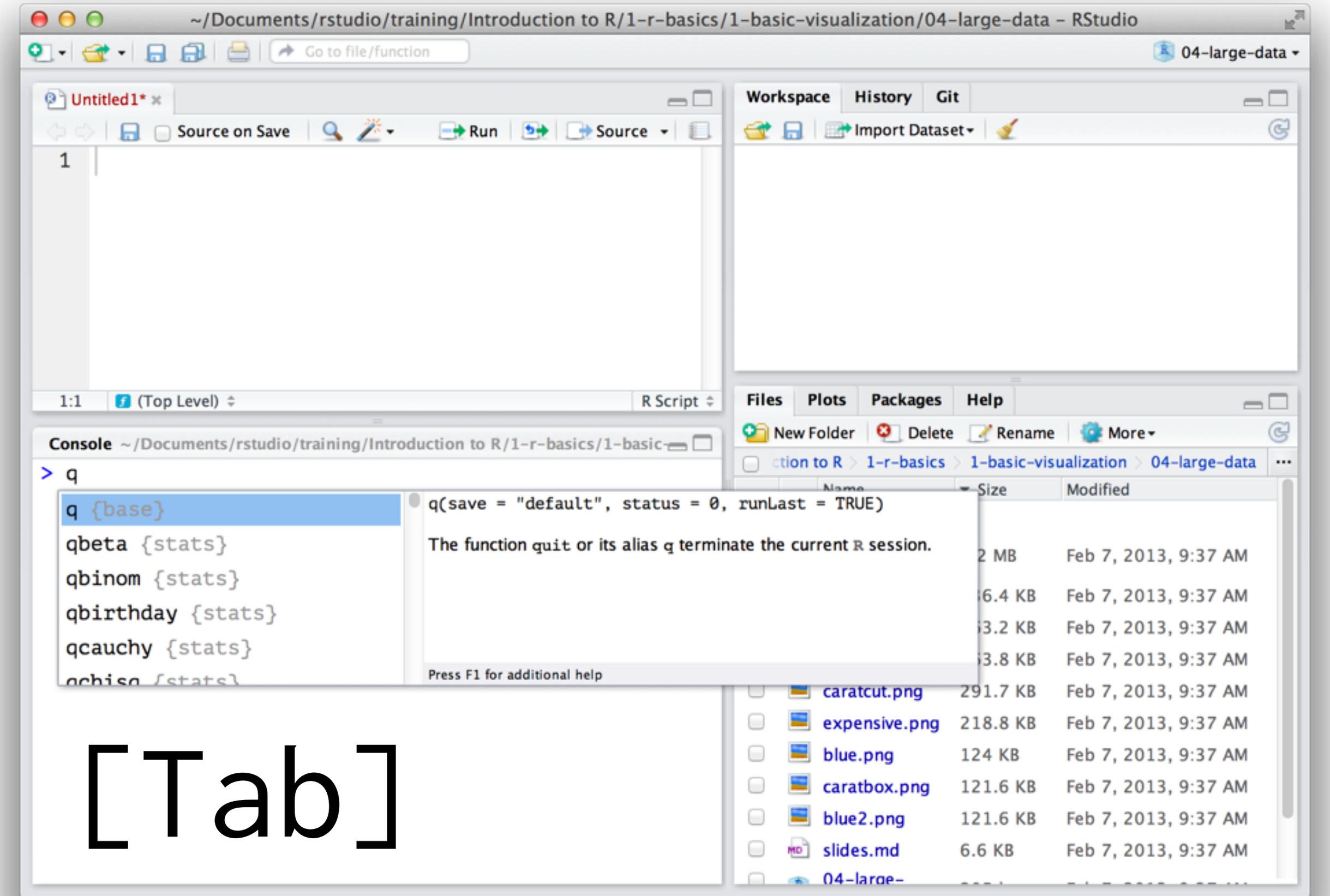
```
ggplot(daily, aes(precip)) + geom_histogram()
ggplot(daily, aes(precip)) +
   geom_histogram(binwidth = 0.01)

# ALWAYS EXPERIMENT WITH THE BIN WIDTH!
```

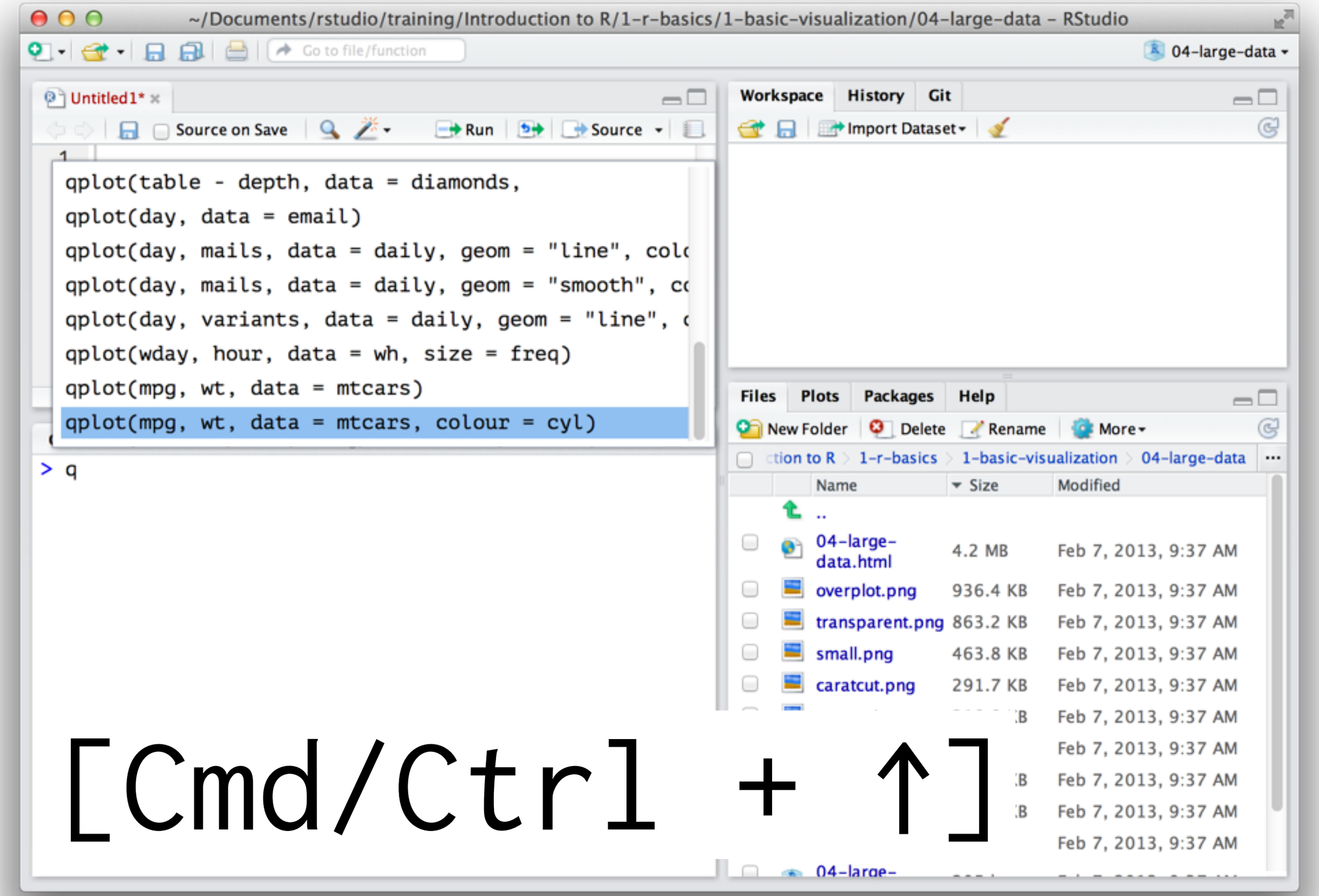
Your turn

What's strange about the distribution of wind speed? Why?

(Hint: the `resolution()` function might help)

Rstudio tips


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
> q
q {base}
qbeta {stats}
qbinom {stats}
qbirthday {stats}
qcauchy {stats}
q(save = "default", status = 0, runLast = TRUE)
The function quit or its alias q terminate the current R session.
Press F1 for additional help
2 MB
Feb 7, 2013, 9:37 AM
6.4 KB
Feb 7, 2013, 9:37 AM
3.2 KB
Feb 7, 2013, 9:37 AM
3.8 KB
Feb 7, 2013, 9:37 AM
caratcut.png
291.7 KB
Feb 7, 2013, 9:37 AM
expensive.png
218.8 KB
Feb 7, 2013, 9:37 AM
blue.png
124 KB
Feb 7, 2013, 9:37 AM
caratbox.png
121.6 KB
Feb 7, 2013, 9:37 AM
blue2.png
121.6 KB
Feb 7, 2013, 9:37 AM
slides.md
6.6 KB
Feb 7, 2013, 9:37 AM
[Tab]


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
qplot(table - depth, data = diamonds,
qplot(day, data = email)
qplot(day, mails, data = daily, geom = "line", colo
qplot(day, mails, data = daily, geom = "smooth", co
qplot(day, variants, data = daily, geom = "line", c
qplot(wday, hour, data = wh, size = freq)
qplot(mpg, wt, data = mtcars)
qplot(mpg, wt, data = mtcars, colour = cyl)
> q
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
tion to R
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
..
04-large-data.html
4.2 MB
Feb 7, 2013, 9:37 AM
overplot.png
936.4 KB
Feb 7, 2013, 9:37 AM
transparent.png
863.2 KB
Feb 7, 2013, 9:37 AM
small.png
463.8 KB
Feb 7, 2013, 9:37 AM
caratcut.png
291.7 KB
Feb 7, 2013, 9:37 AM
[Cmd/Ctrl + ↑]

~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio

Go to file/function

04-large-data

Untitled1*

Source on Save

Run

Source

```
1  library(ggplot2)
```

1:17 (Top Level) R Script

Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-

```
>
```

Workspace History Git

Import Dataset

Files Plots Packages Help

New Folder Delete Rename More

tion to R 1-r-basics 1-basic-visualization 04-large-data

	Name	Size	Modified
	..		
	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
	blue.png	124 KB	Feb 7, 2013, 9:37 AM
	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
	04-large-		

~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio

Go to file/function

04-large-data

Untitled1*

Source on Save | Run | Source

```
1  library(ggplot2)
```

[Cmd/ctrl + enter]

1:17 (Top Level) R Script

Workspace | History | Git

Import Dataset

Files | Plots | Packages | Help

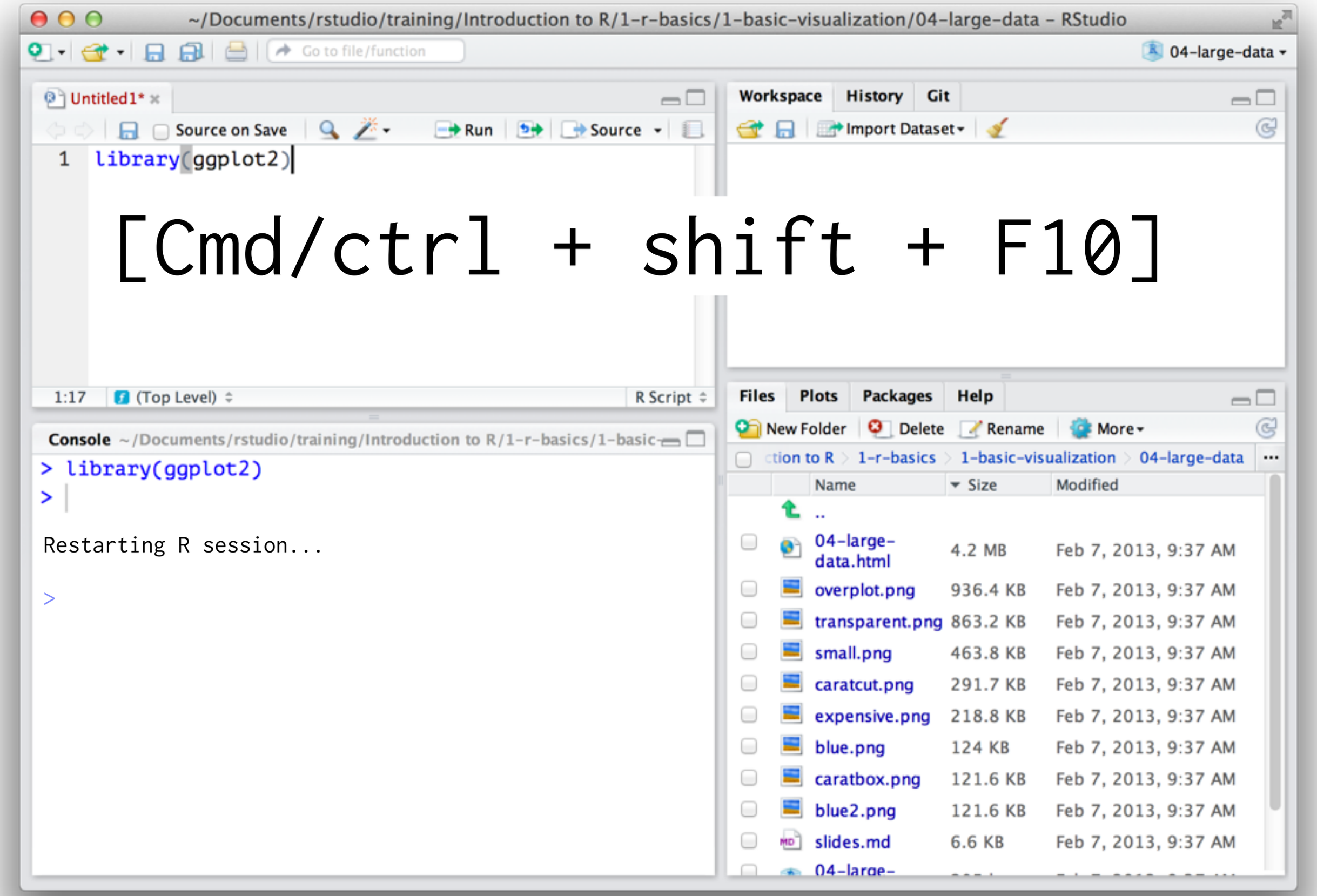
New Folder | Delete | Rename | More

tion to R › 1-r-basics › 1-basic-visualization › 04-large-data

Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-

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

	Name	Size	Modified
	..		
	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
	blue.png	124 KB	Feb 7, 2013, 9:37 AM
	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
	04-large-		


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1 library(ggplot2)
[Cmd/ctrl + shift + F10]
1:17
(Top Level)
R Script
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
> library(ggplot2)
>
Restarting R session...
>
..
04-large-data.html 4.2 MB Feb 7, 2013, 9:37 AM
overplot.png 936.4 KB Feb 7, 2013, 9:37 AM
transparent.png 863.2 KB Feb 7, 2013, 9:37 AM
small.png 463.8 KB Feb 7, 2013, 9:37 AM
caratcut.png 291.7 KB Feb 7, 2013, 9:37 AM
expensive.png 218.8 KB Feb 7, 2013, 9:37 AM
blue.png 124 KB Feb 7, 2013, 9:37 AM
caratbox.png 121.6 KB Feb 7, 2013, 9:37 AM
blue2.png 121.6 KB Feb 7, 2013, 9:37 AM
slides.md 6.6 KB Feb 7, 2013, 9:37 AM

https://twitter.com/rstudiotips

Where to go from here

Index. ggplot2 0.9.2.1

docs.ggplot2.org/current/

ggplot2 0.9.2.1 Index

Help topics

Geoms

Geoms, short for geometric objects, describe the type of plot you will produce.

- `geom_abline`
  Line specified by slope and intercept.
- `geom_area`
  Area plot.
- `geom_bar`
  Bars, rectangles with bases on x-axis
- `geom_bin2d`
  Add heatmap of 2d bin counts.
- `geom_blank`
  Blank, draws nothing.
- `geom_boxplot`
  Box and whiskers plot.
- `geom_contour`
  Display contours of a 3d surface in 2d.
- `geom_crossbar`
  Hollow bar with middle indicated by horizontal line.
- `geom_density`
  Display a smooth density estimate.
- `geom_density2d`
  Contours from a 2d density estimate.
- `geom_dotplot`
  Dot plot

Dependencies

- **Depends**: stats, methods
- **Imports**: plyr, digest, grid, gtable, reshape2, scales, memoise, proto, MASS
- **Suggests**: quantreg, Hmisc, mapproj, maps, hexbin, maptools, multcomp, nlme, testthat
- **Extends**: sp

Learning ggplot2

R graphics cookbook
http://amzn.com/1449316956

ggplot2 book
http://amzn.com/0387981403

ggplot2 mailing list
http://groups.google.com/group/ggplot2

stackoverflow
http://stackoverflow.com/tags/ggplot2