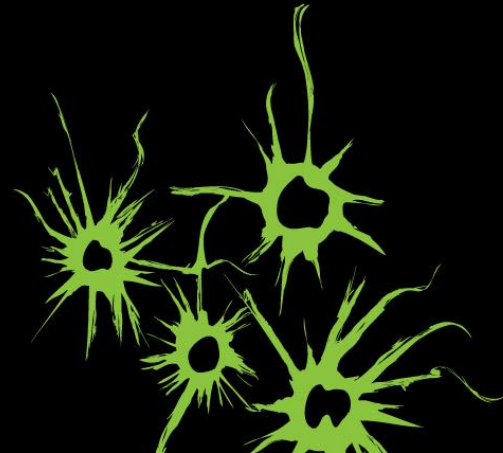
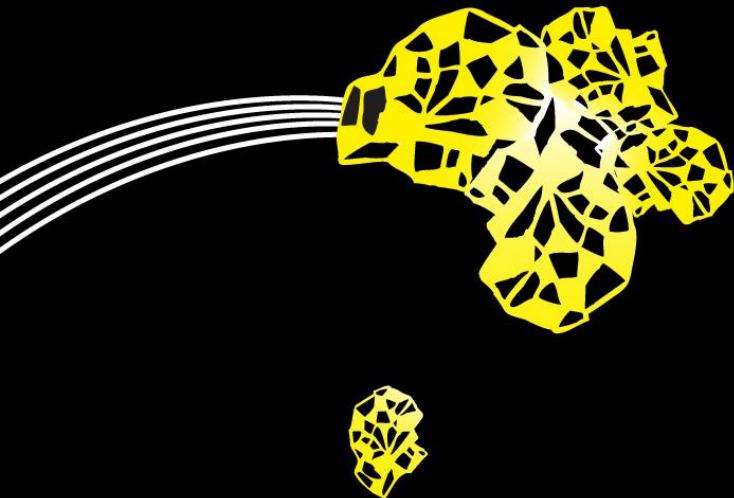


INTRODUCTION INTO WORKING WITH R

SESSION 1 – VERSION 17/11/2019

BENJAMIN ZIEPERT



INTRODUCTION INTO WORKING WITH R

SESSION 1

Lecturers: Benjamin Ziepert

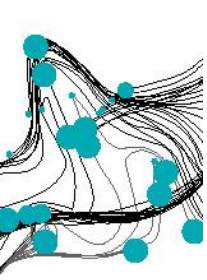
Authors: Benjamin Ziepert & Dr. Elze G. Ufkes

The course will:

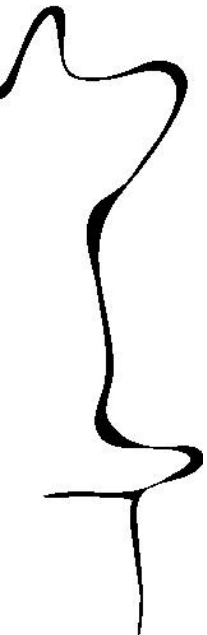
- Teach you the basics of R
- Practice an advanced data-analyses that can't be done with SPSS
- Enable you to further study R on your own

The course will not:

- Enable you to do all statistical analysis in R



WHY R?

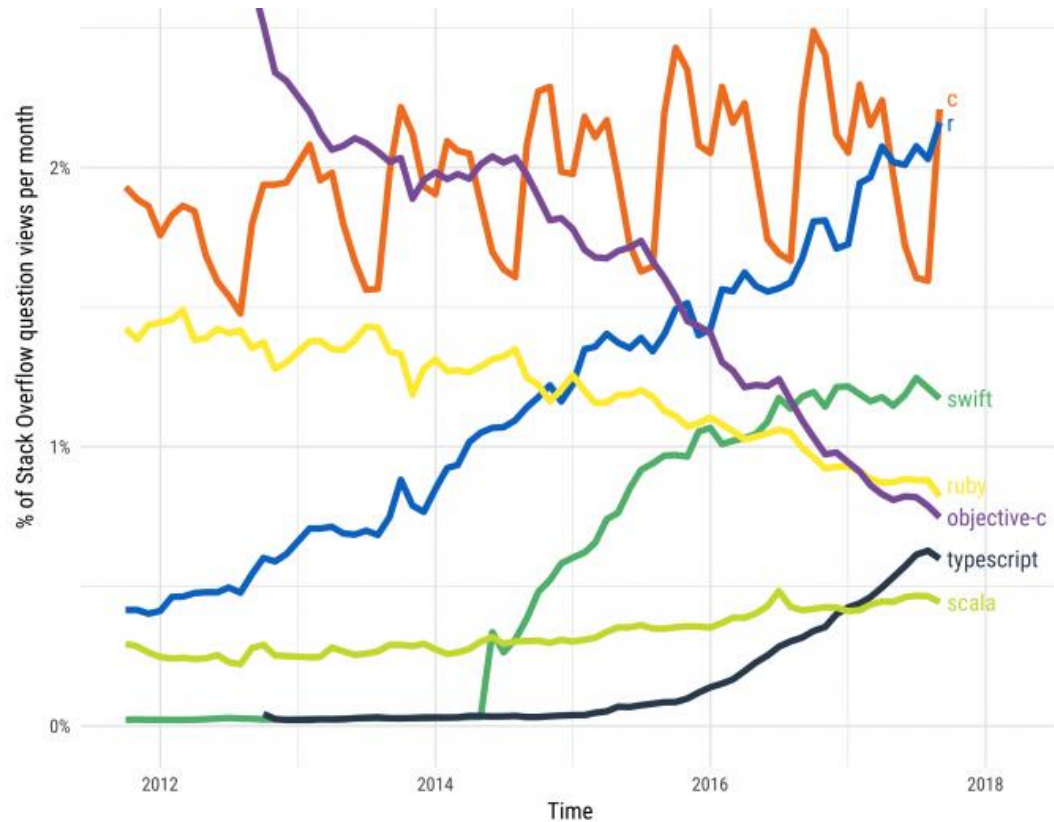
- 
- A large, abstract black line graphic on the left side of the slide, resembling a stylized, elongated shape with a jagged, irregular outline, possibly representing a map or a data visualization.
- Open Source
 - Powerful and flexible
 - The standard for data science

Programming becomes more important in the workplace and as teachers we want to prepare you for that reality.



WHY R?

R GROWTH



Source: stackoverflow.blog

WHY R?

COMPANIES USING R



McKinsey&Company



The New York Times

UBER

Source: listendata.com

HOW TO DEAL WITH CODE?



HOW TO DEAL WITH CODE?

MAKE AN INVESTMENT

“Learning to code is empowering and can hugely improve a researcher’s career prospects. But it does require an investment”

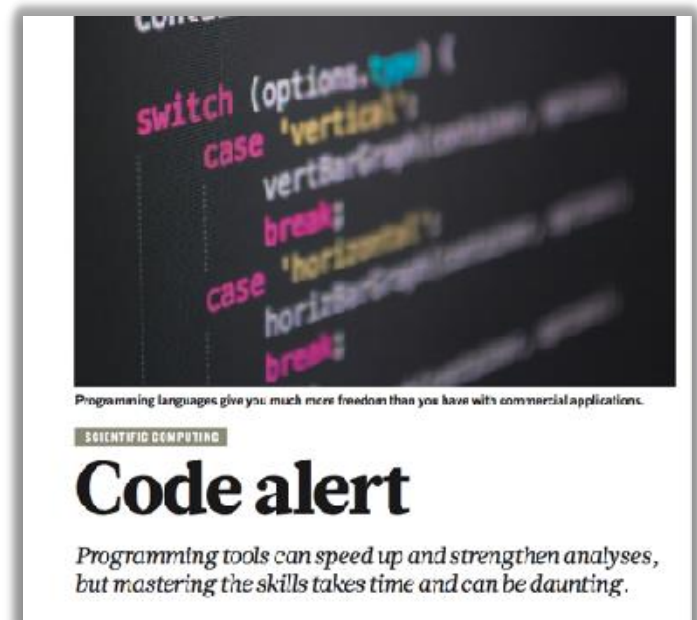


Baker, M. (2017). Scientific computing:
Code alert. *Nature*, 541(7638), 563–565.
doi:10.1038/nj7638-563a

HOW TO DEAL WITH CODE?

ANTICIPATE HURDLES IN THE BEGINNING

“Typos, for example, bring work to a standstill, she says. They didn’t put a space and the script won’t run; they put two dashes and the script won’t run.”

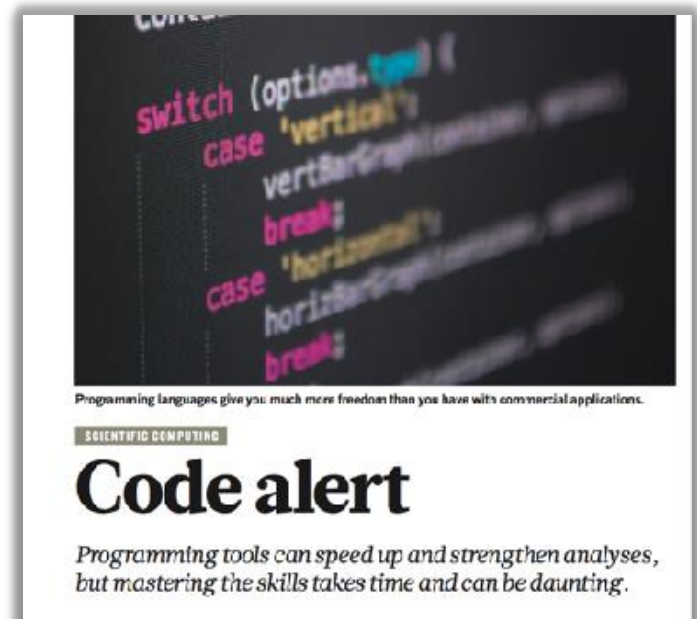


Baker, M. (2017). Scientific computing:
Code alert. *Nature*, 541(7638), 563–565.
doi:10.1038/nj7638-563a

HOW TO DEAL WITH CODE?

PLAN CODING TIME WITH PEERS

“... people [should] pick a language that’s popular with their colleagues and work initially in four-hour blocks, which he says provide enough time to work through hurdles and get a sense of progress.”



Baker, M. (2017). Scientific computing:
Code alert. *Nature*, 541(7638), 563–565.
doi:10.1038/nj7638-563a

HOW TO DEAL WITH CODE?

SEEK HELP FROM THE START

Perhaps the biggest barrier is insecurity ... “Many people think, I’ll just figure it out on my own first. I’m not good enough yet to ask questions’,” she says. Instead, they should seek help from others to gain more skills.



Baker, M. (2017). Scientific computing:
Code alert. *Nature*, 541(7638), 563–565.
doi:10.1038/nj7638-563a



PLANNING



1. Learn the benefits
2. Getting up to speed with the basics of R
 - Create figures
 - Run analysis
 - Basic R coding knowledge
3. Getting introduced to the extensive possibilities of R
 - Completing a R-project wherein you challenge yourself



PLANNING

OVERVIEW

3 Lectures

- Introduction into R
- Statistical analysis
- Analyzing social media content

2 Self-study assignment's using DataCamp

Reading

- R is for Revolution (Culpepper & Aguinis, 2010)
- Scientific computing: Code alert (Baker, 2017)

PLANNING

OVERVIEW

Passing requirements

- Attendance of all sessions
- Complete DataCamp assignments with at least 8000 XP (Self-study)
- Complete R script assignment with statistical analysis (Session 2)
- Complete Twitter analysis and present results (Session 3)

PLANNING

TODAY

- Introduction in R
- Graphics
- Statistical analysis
- Preparing next lecture



R BASICS

SOFTWARE

R

- Core software
- <https://cloud.r-project.org>



RStudio

- Integrated development environment (IDE) for R
- <https://www.rstudio.com>

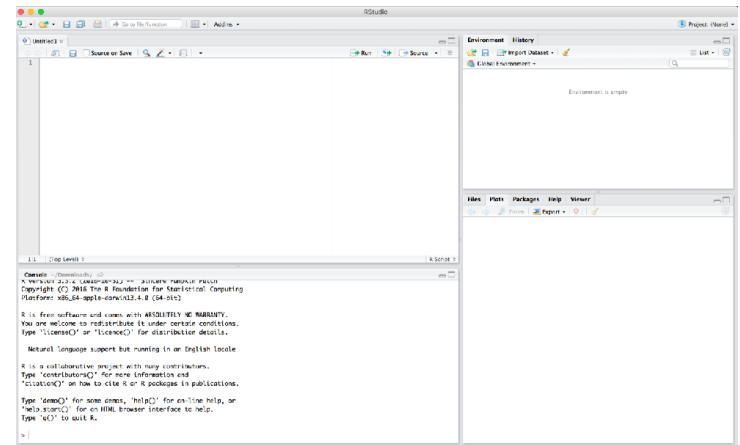
R BASICS

RSTUDIO



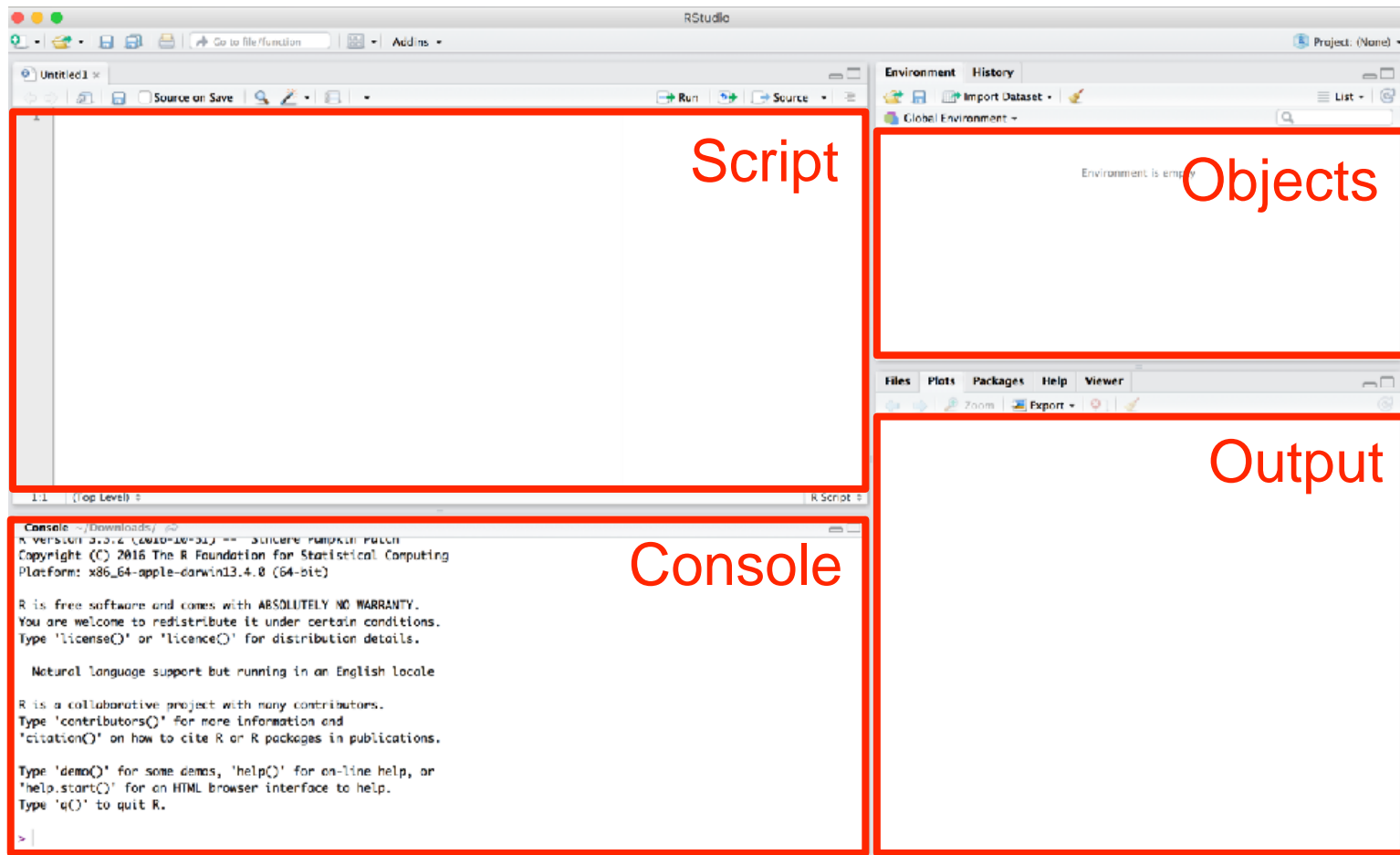
Let's have a look at the software.

✓ Please open RStudio now.



R BASICS

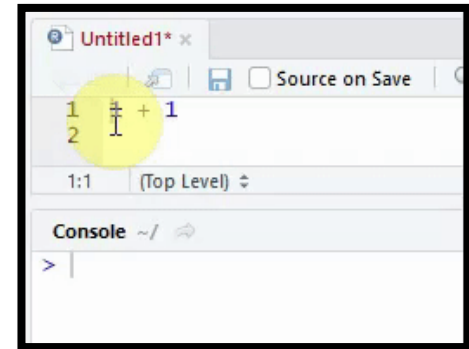
RSTUDIO



R BASICS

RUNNING CODE

- Run line or selection: [Cmd] / [Ctrl] + [Enter]
 - Code will be transferred to the console and runs there
- Document your code well with comments
 - Characters that come after # are skipped
- Be precise, punctuation and capitalization is important
 - DataBase ≠ database



R BASICS

OPEN HANDOUT

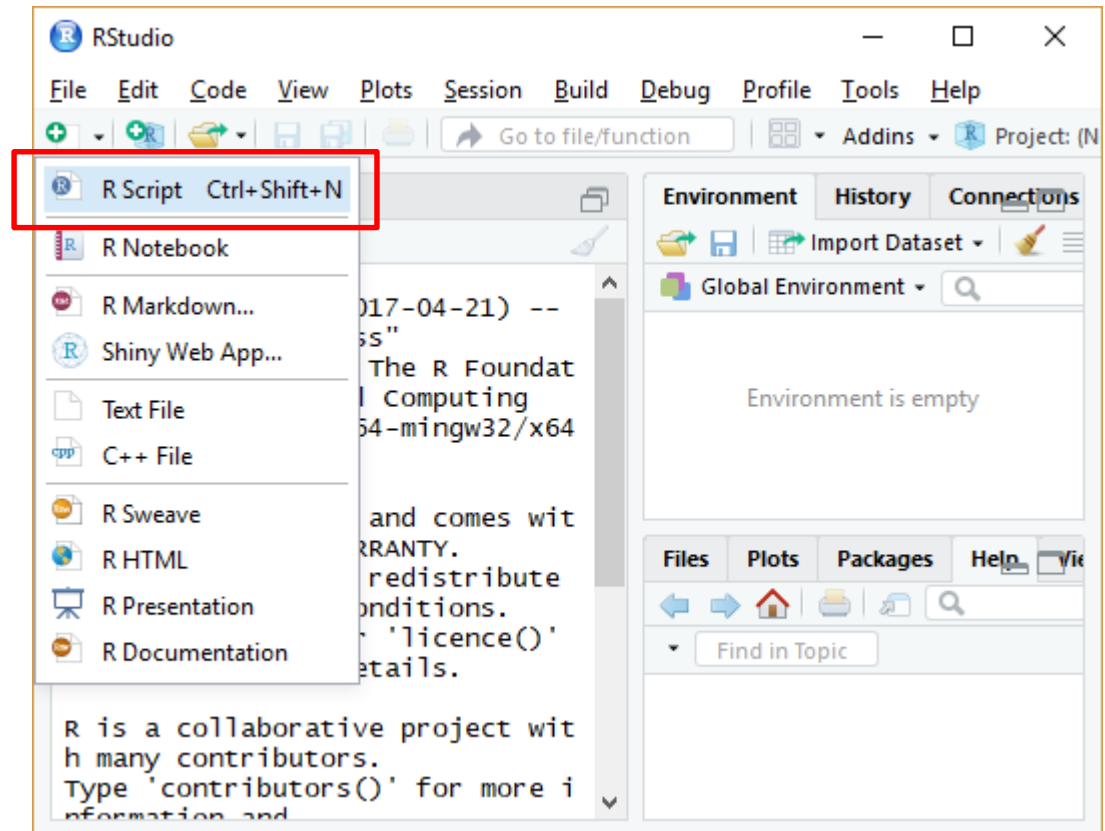
- ✓ Go now to benjaminziepert.com/teaching
- ✓ Download all files and save them in one folder
- ✓ Open Session 1 → Handout R basics: statistical graphs and analysis

R BASICS

CREATE SCRIPT FILE

- ✓ Open R Studio
- ✓ Create R Script
- ✓ Save R Script

Tip: save all files in one location



R BASICS

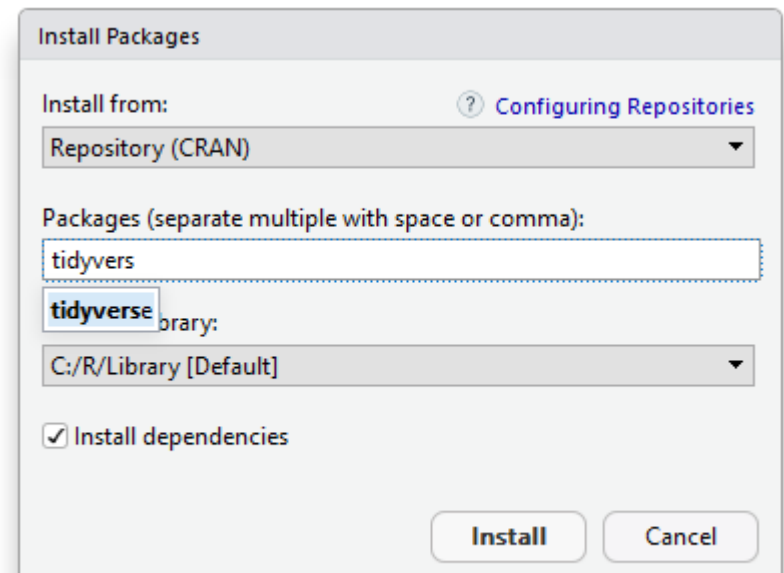
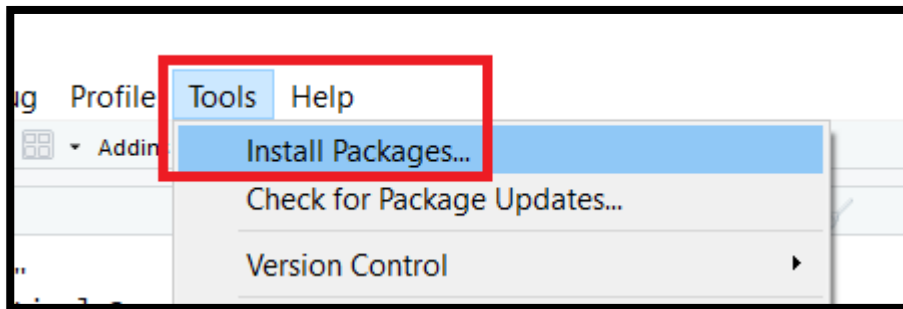
1 INSTALLING AND ACTIVATING PACKAGES

- Packages add functionality to R
 - Use `install.packages()`
 - For instance: `install.packages("tidyverse")`
 - You only have to install the package once
 - When asked, decline to install from source package or to compile a package.
 - Installation doesn't work? Check the FAQ.
- ✓ Copy the text from the gray box in the handout to your R file and then run the line with [Cmd] / [Ctrl] + [Enter].

R BASICS

1 INSTALLING AND ACTIVATING PACKAGES

RStudio Menu alternative



R BASICS

1 INSTALLING AND ACTIVATING PACKAGES

Activate the package using `library()`

You have to do this every time / session you want to use the package

GRAPHICS

2.1 OPEN THE DATA FRAME MPG

- ✓ Run `library("ggplot2")`
- ✓ Run `mpg` to open the data frame

mpg is a data set for the fuel economy data from 1999 and 2008 for 38 popular car models

```
# # A tibble: 234 x 11
#   manufacturer model      displ  year   cyl trans      drv    cty   hwy fl    class
#   <chr>         <chr>    <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
# 1 audi         a4         1.8    1999     4 auto(l5)  f       18    29 p    compact
# 2 audi         a4         1.8    1999     4 manual(m5) f       21    29 p    compact
# 3 audi         a4         2.0    2008     4 manual(m6) f       20    31 p    compact
# 4 audi         a4         2.0    2008     4 auto(av)   f       21    30 p    compact
# 5 audi         a4         2.8    1999     6 auto(l5)  f       16    26 p    compact
# 6 audi         a4         2.8    1999     6 manual(m5) f       18    26 p    compact
# 7 audi         a4         3.1    2008     6 auto(av)   f       18    27 p    compact
# 8 audi         a4 quattro 1.8    1999     4 manual(m5) 4       18    26 p    compact
# 9 audi         a4 quattro 1.8    1999     4 auto(l5)   4       16    25 p    compact
# 10 audi        a4 quattro 2.0    2008     4 manual(m6) 4       20    28 p    compact
# # ... with 224 more rows
```

GRAPHICS

HOW TO CREATE A VISUALIZATION?

How can we visualize this data?

- For instance, what is the frequency of engine sizes?

→ We use the graphics package ggplot2

ggplot2 was installed with tidy verse packages and is used for graphics.

GRAPHICS

2.2 HISTOGRAM

Creates coordinate system
based on a data frame

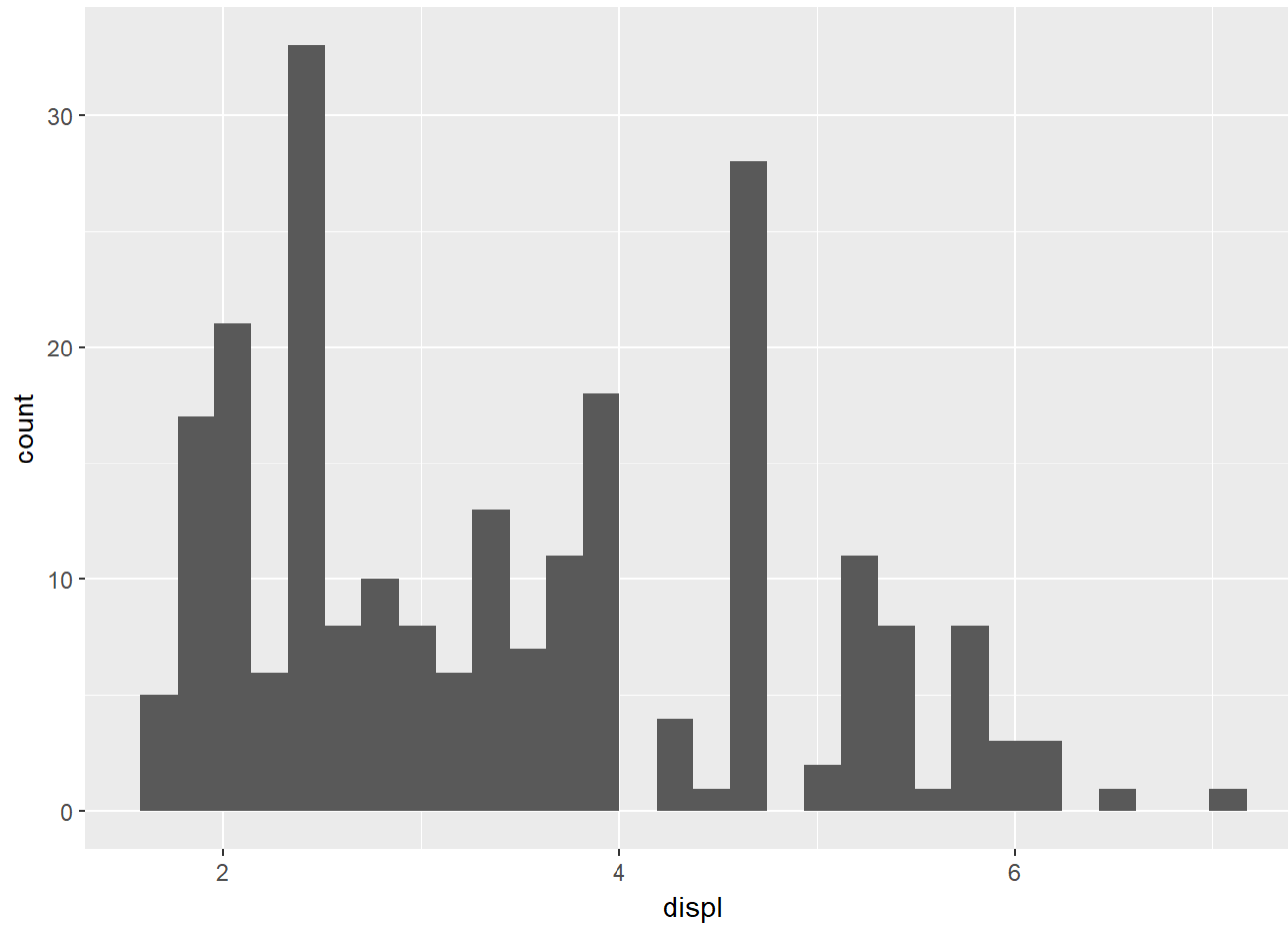
Adds a layer of some
geometric object

```
ggplot(data = mpg) + aes(displ) + geom_histogram()
```

Specifies mapping of
variables in the data frame onto
aesthetic attributes

GRAPHICS

2.2 HISTOGRAM

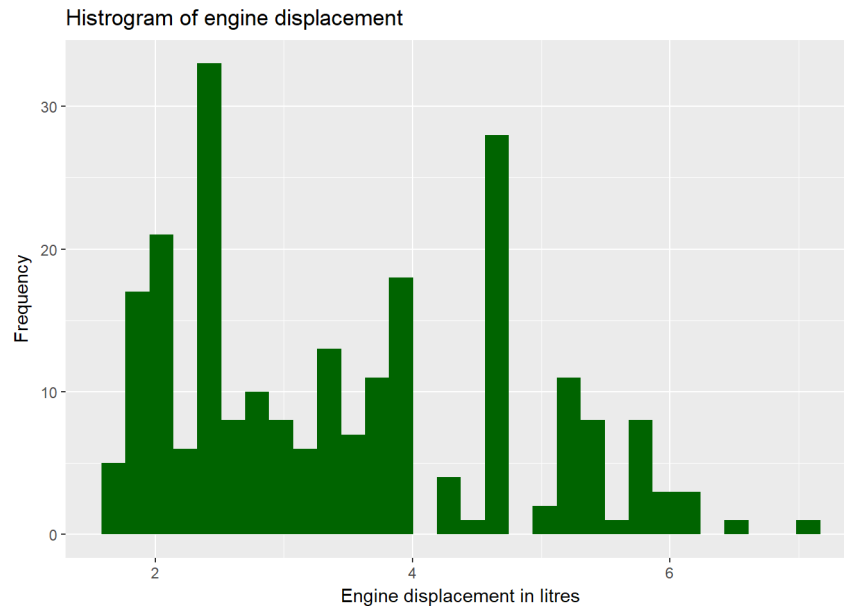


GRAPHICS

2.3 UPDATE LABELS AND COLOR

```
ggplot(data = mpg) +  
  aes(x = displ) +  
  geom_histogram(fill = 'darkgreen') +  
  labs(title = "Histogram of engine displacement",  
       x = "Engine displacement in litres",  
       y = "Frequency")
```

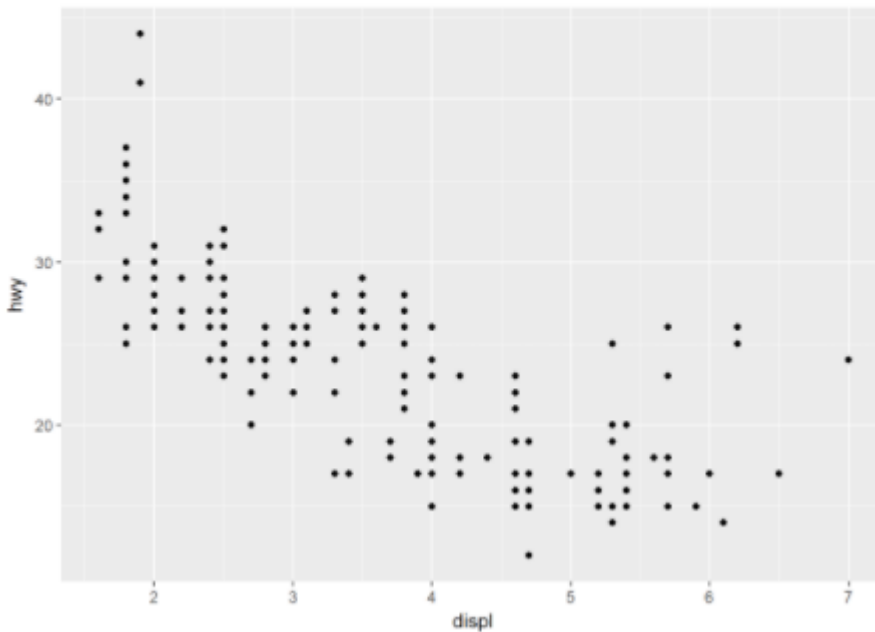
`geom_histogram()` is now filled with an color and labels (`labs()`) are added.



GRAPHICS

2.4 CREATE A SCATTER DOT

```
ggplot(data = mpg) +  
  aes(x = displ, y = hwy) +  
  geom_point()
```



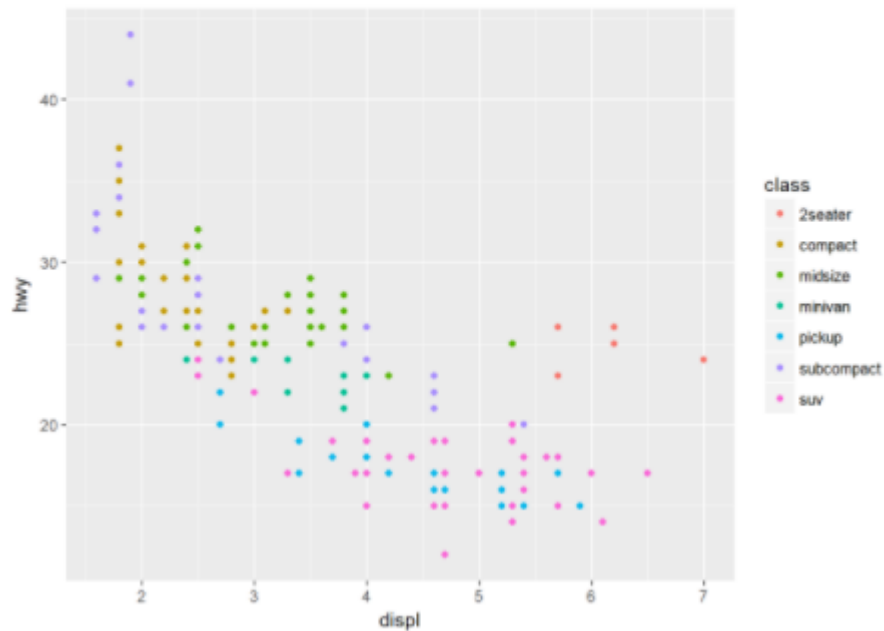
`geom_histogram()` is now replaced with `geom_point()` and we added `hwy` to the variables.

GRAPHICS

2.5 ADDING MORE AESTHETIC MAPPINGS

```
ggplot(data = mpg) +  
  aes(x = displ, y = hwy, color = class) +  
  geom_point()
```

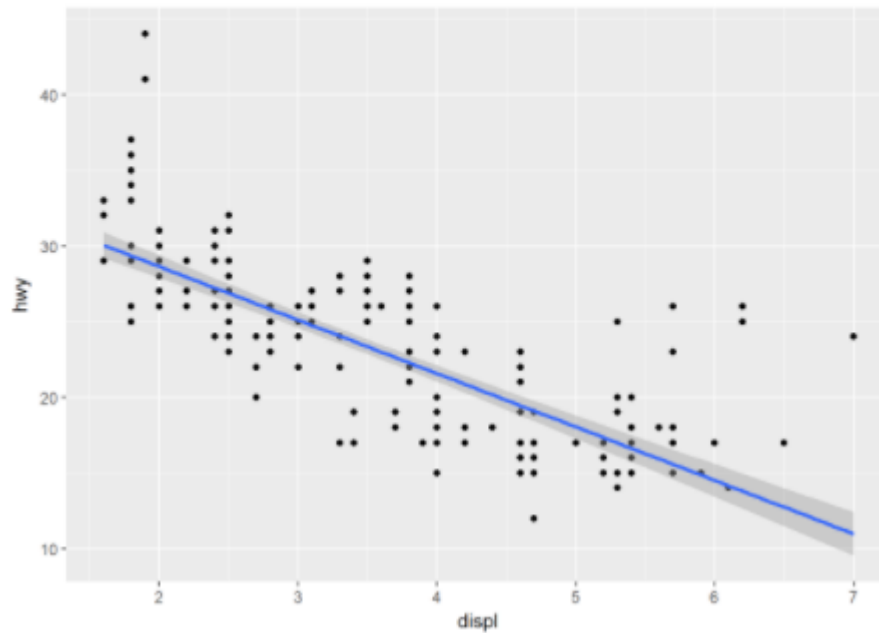
Colours per car class



GRAPHICS

2.6 ADDING REGRESSION LINE

```
ggplot(data = mpg) +  
  aes(x = displ, y = hwy) +  
  geom_point() +  
  geom_smooth(method=lm)
```



`geom_smooth(method=lm)`

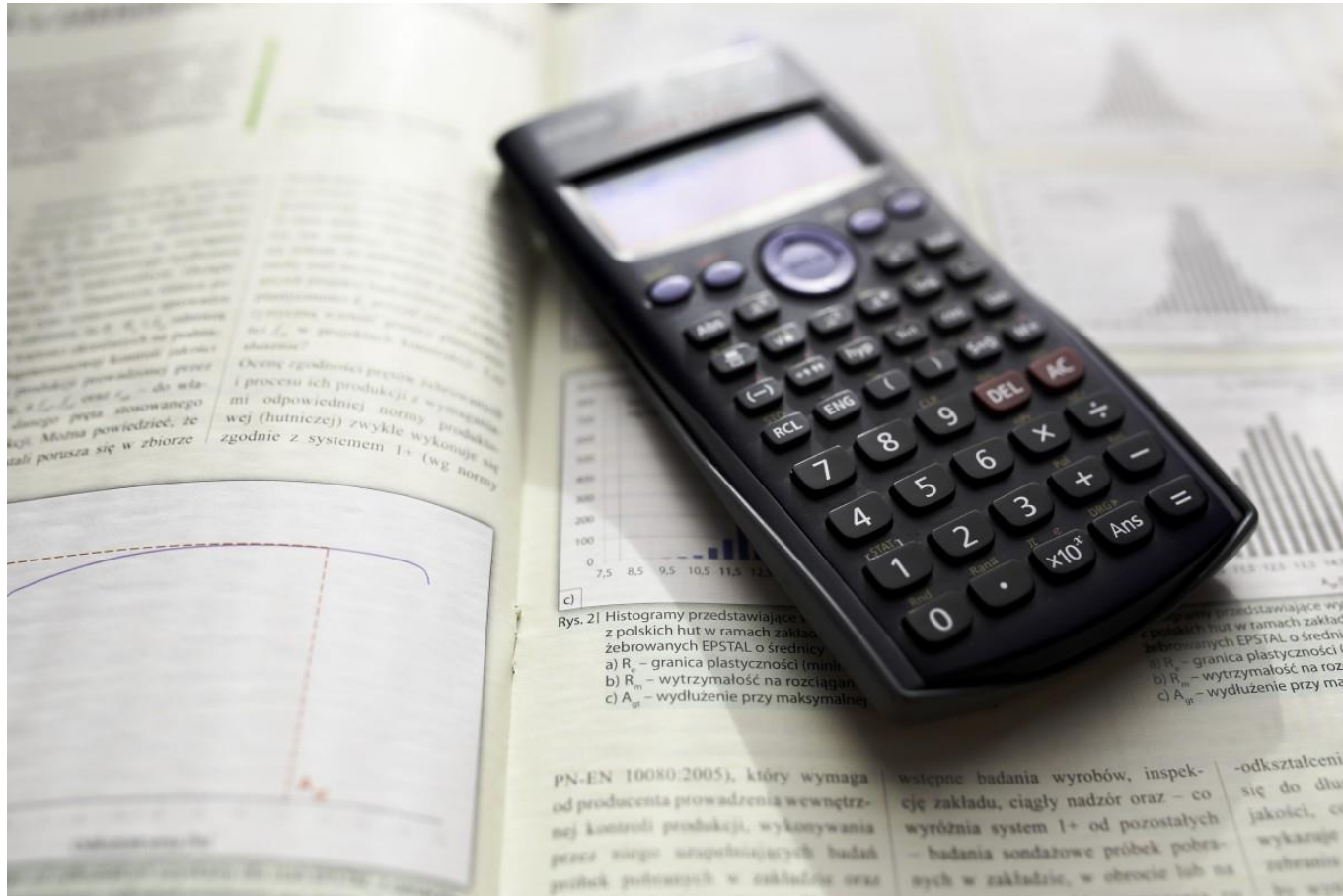
What does this graph tell us?

You can find more info about graphics at

- <http://www.sthda.com/english/wiki/ggplot2-essentials>
- <http://www.r-graph-gallery.com>

STATISTICS

DESCRIPTIVE, CORRELATION & LINEAR



STATISTICS

3.1 DESCRIPTIVE STATISTICS

```
summary(mpg)
```

```
## manufacturer      model      displ      year
## Length:234      Length:234      Min.   :1.600      Min.   :1999
## Class :character  Class :character  1st Qu.:2.400      1st Qu.:1999
## Mode  :character  Mode  :character  Median :3.300      Median :2004
##                                     Mean  :3.472      Mean  :2004
##                                     3rd Qu.:4.600      3rd Qu.:2008
##                                     Max.   :7.000      Max.   :2008
##      cyl      trans      drv      cty
## Min.   :4.000      Length:234      Length:234      Min.   : 9.00
## 1st Qu.:4.000      Class :character  Class :character  1st Qu.:14.00
## Median :6.000      Mode  :character  Mode  :character  Median :17.00
## Mean   :5.889                                     Mean  :16.86
## 3rd Qu.:8.000                                     3rd Qu.:19.00
## Max.   :8.000                                     Max.   :35.00
##      hwy      fl      class
## Min.   :12.00      Length:234      Length:234
## 1st Qu.:18.00      Class :character  Class :character
## Median :24.00      Mode  :character  Mode  :character
## Mean   :23.44
## 3rd Qu.:27.00
## Max.   :44.00
```

STATISTICS

LINEAR STATISTICS

3.3 Independent T-Test

- `t.test(x, y)`

3.4 One Way Anova

- `aov(y ~ x, data = mydata)`

3.5 Multiple Linear regression

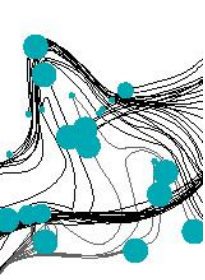
- `lm(y ~ x1 + x2 + x3, data = mydata)`

Formula

- $y = x_1 + \dots + x_k$
- $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$

More statistics:

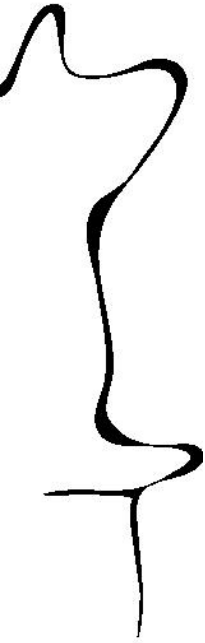
- <https://www.statmethods.net/stats/index.html>
- Discovering Statistics Using R by Andy Field.



NEXT LECTURE

PLANNING

- Preparation
 - At home: DataCamp assignment
 - Now: Check R and RStudio installation



NEXT LECTURE

SELF-STUDY ASSIGNMENTS



Complete the 3 assignments before the day of the next lecture:

1. Introduction to R (4 hours)
 - Whole course
2. Importing data (2 hours)
 - Only do the chapter "Importing data from statistical software packages" in the course "Importing Data in R (Part 2)"
3. Bring at least one question for the Q&A next lecture

To pass the DataCamp assignments your XP must stay above 7000.

- Therefore, try to understand what you do before clicking on hint or show solution.

NEXT LECTURE

PREPARING AND CHECKING INSTALLATION

- ✓ Make sure R, RStudio and Rtools (windows only) are up to date.
- ✓ Please install or update the following packages: "tidyverse", "ggplot2", "Hmisc", "twitteR", "tm", "wordcloud", "psych" , "devtools" and "gplots".
- ✓ Update all packages
- ✓ Open "S01F03 Test Package Installation.R" and call me.

ADDITIONAL INFORMATION

Check the R Studio Cheat sheets: [Base R](#), [R Studio](#) & [more](#) ...

Statistics

- <https://www.statmethods.net/stats/index.html>
- Discovering Statistics Using R by Andy Field.

Graphics

- <http://www.sthda.com/english/wiki/ggplot2-essentials>
- <http://www.r-graph-gallery.com>