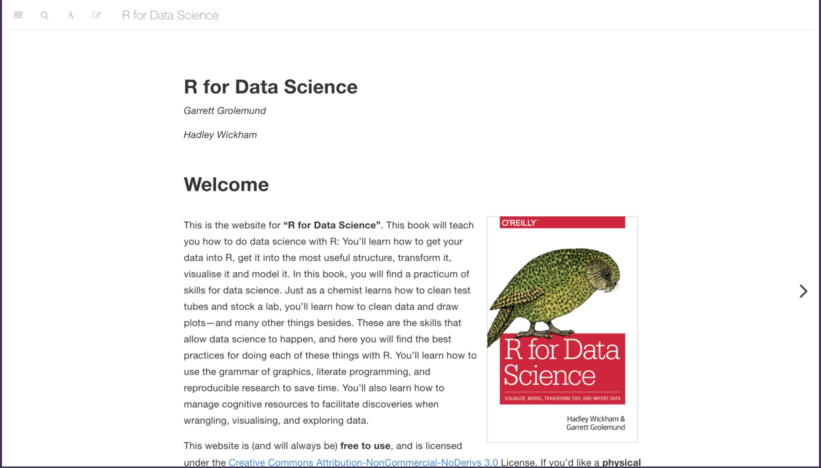
–

COMP360: Research Dissertation

11: Visualising Data in R

The Book



R for Data Science

Garrett Grolemund

Hadley Wickham

Welcome

This is the website for **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License. If you'd like a **physical**

Figure 1: Link to free book: R for Data Science

What is R?

Load Tidyverse

```
> library(tidyverse)

## -- Attaching packages -------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ----------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data Frame

```
> mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model displ  year   cyl trans drv     cty   hwy fl    cla~
##    <chr>        <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <ch>
##  1 audi         a4      1.8  1999     4 auto~ f        18    29 p     com~
##  2 audi         a4      1.8  1999     4 manu~ f        21    29 p     com~
##  3 audi         a4      2    2008     4 manu~ f        20    31 p     com~
##  4 audi         a4      2    2008     4 auto~ f        21    30 p     com~
##  5 audi         a4      2.8  1999     6 auto~ f        16    26 p     com~
##  6 audi         a4      2.8  1999     6 manu~ f        18    26 p     com~
##  7 audi         a4      3.1  2008     6 auto~ f        18    27 p     com~
##  8 audi         a4 q~   1.8  1999     4 manu~ 4        18    26 p     com~
##  9 audi         a4 q~   1.8  1999     4 auto~ 4        16    25 p     com~
## 10 audi         a4 q~   2    2008     4 manu~ 4        20    28 p     com~
## # ... with 224 more rows
```

Loading data

```
library(readr)
dat <- read_csv('assets/obfuscated_data.csv')
```

```
## # A tibble: 159 x 14
##    GENDER BIRTH_YEAR PRIOR_EXP DEPTH PROCAST APTITUDE CONCEPT ANXIETY
##     <int>      <int>     <int> <dbl>   <dbl>    <dbl>   <dbl>   <dbl>
##  1      1       1988         1  64.3   44.4     52.4   -9.02   55.9
##  2      1       1997         2  57.9    6.73    17.0   16.3    12.8
##  3      1       1997         3  69.8   -5.96     1.16  21.7     0.889
##  4      1       1997         1  95.0    8.43     3.05  27.5     1.14
##  5      1       1998         2  42.8    8.58     2.99  14.8    12.8
##  6      1       1997         3  58.3   -5.36     1.46  22.9     2.78
##  7      1       1997         1  60.2   22.7     16.4   -0.998  58.0
##  8      1       1998         3  42.0    9.89    13.0    6.42   42.3
##  9      1       1996         2  63.5   15.2     44.3   20.6     7.50
## 10      1       1996         1  71.1    1.08    23.3   14.3    44.5
## # ... with 149 more rows, and 6 more variables: INTEREST <dbl>,
## #   EXTRAVERSION <int>, AGREEABLENESS <int>, CONSCIENTIOUSNESS <int>,
## #   NEUROTICISM <int>, OPENNESS <int>
```

Summary

```
summary(dat)
```

```
##      GENDER        BIRTH_YEAR     PRIOR_EXP         DEPTH
##  Min.   :1.000   Min.   :1976   Min.   :1.000   Min.   :-170.12
##  1st Qu.:1.000   1st Qu.:1996   1st Qu.:1.000   1st Qu.:  46.56
##  Median :1.000   Median :1997   Median :1.000   Median :  59.51
##  Mean   :1.119   Mean   :1996   Mean   :1.704   Mean   :  57.39
##  3rd Qu.:1.000   3rd Qu.:1998   3rd Qu.:3.000   3rd Qu.:  71.48
##  Max.   :2.000   Max.   :1999   Max.   :3.000   Max.   :  96.68
##     PROCAST          APTITUDE          CONCEPT            ANXIETY
##  Min.   :-5.964   Min.   :-170.23   Min.   :-35.6758   Min.   : 0.8036
##  1st Qu.:12.083   1st Qu.:  17.58   1st Qu.:-10.1707   1st Qu.:19.3574
##  Median :26.658   Median :  25.24   Median :  1.7783   Median :37.0401
##  Mean   :26.062   Mean   :  29.24   Mean   :  0.4195   Mean   :37.6957
##  3rd Qu.:39.166   3rd Qu.:  44.76   3rd Qu.: 11.8288   3rd Qu.:53.8863
##  Max.   :69.619   Max.   :  95.28   Max.   : 28.7092   Max.   :92.0369
##     INTEREST      EXTRAVERSION    AGREEABLENESS   CONSCIENTIOUSNESS
##  Min.   :12.53   Min.   :  0.00   Min.   : 20.0   Min.   : 20.0
##  1st Qu.:60.50   1st Qu.: 62.50   1st Qu.:100.0   1st Qu.: 91.0
##  Median :71.46   Median :100.00   Median :120.0   Median :120.0
##  Mean   :68.71   Mean   : 94.44   Mean   :121.4   Mean   :115.6
##  3rd Qu.:83.57   3rd Qu.:120.00   3rd Qu.:140.5   3rd Qu.:140.0
##  Max.   :91.06   Max.   :188.00   Max.   :200.0   Max.   :200.0
##   NEUROTICISM       OPENNESS
##  Min.   :  0.0   Min.   : 24.0
##  1st Qu.: 60.0   1st Qu.:100.0
##  Median : 98.0   Median :120.0
##  Mean   : 94.4   Mean   :117.8
##  3rd Qu.:126.5   3rd Qu.:140.0
##  Max.   :200.0   Max.   :200.0
```

ggplot

```
ggplot(data = dat) +  geom_point(mapping = aes(x =
BIRTH_YEAR, y = ANXIETY))
```

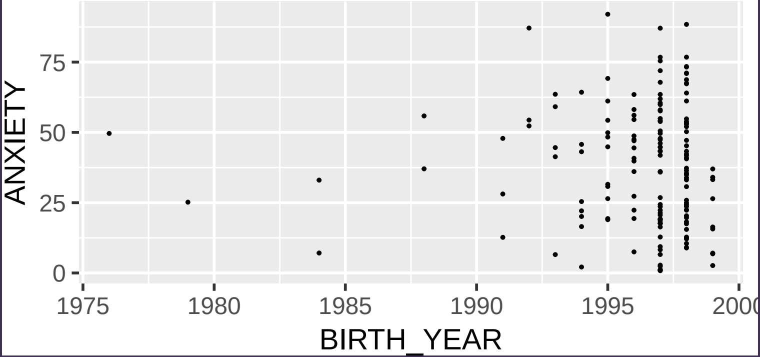
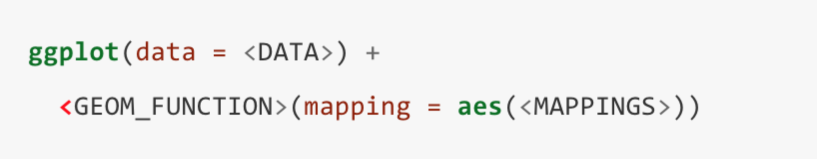


Figure 2: ggplot point graph

ggplot



```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Figure 3: The anatompy of a ggplot command

Correlation

```
> cor(x, y)
> cor.test(x, y, method)

## [1] 0.6157466
```

T-Test

```
##
##  Welch Two Sample t-test
##
## data:  dat$ANXIETY by dat$GENDER
## t = -0.97505, df = 26.122, p-value = 0.3385
## alternative hypothesis: true difference in means
## 95 percent confidence interval:
##  -13.655452   4.867173
## sample estimates:
## mean in group 1 mean in group 2
##        37.17062        41.56476
```

Further Reading

- Official Docs
- Stat Methods
- Harvard Tutorial Series
- R Studio Docs
- R Markdown Docs