HSS8005 Stream C: Introduction to Quantitative Linguistics

Dr Danielle Turton: danielle.turton@newcastle.ac.uk

Overview

This course is:

- an introduction to analysing large-scale linguistic data sets in R
- aimed at linguists working with quantitative data
- mainly going to focus on phonetic, sociolinguistic and reaction time datasets, but the principles can be applied to any subfield of any kind

What this course is *not*

- An introduction to Natural Language Processing
- An introduction to corpus linguistics
- Discourse analysis, sentiment analysis etc.

Schedule

- Today: Intro to course and to R
- Thursday: Data visualisation
- Next Tuesday: Basic statistical tests
- Next Thursday: Data wrangling, advanced visualisation, statistical tests and requests!

Class website

You can follow the class website here:
https://danielleturton.github.io/quantling

All of the materials we'll be using can be found here for each session.

What is your research?

- Who are you and what do you want to analyse?
- What is your dependent variable?
- What about independent variables?

Introduction to R

Why use R?

- R is the statistics software paradigm of our day
- It's free!
- It's platform independent
- Packages for everything (constantly being updated)
- All the cool kids use it

This lesson's goals

- Work with an R notebook (simpler than working with R proper for now)
- Read in and manipulate data
- Make some figures

R can be used as a calculator

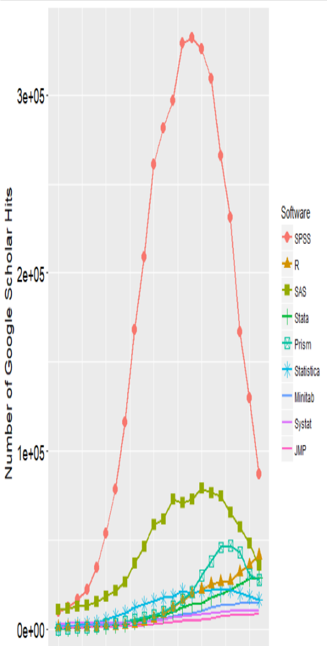
The difference between R and RStudio. . .

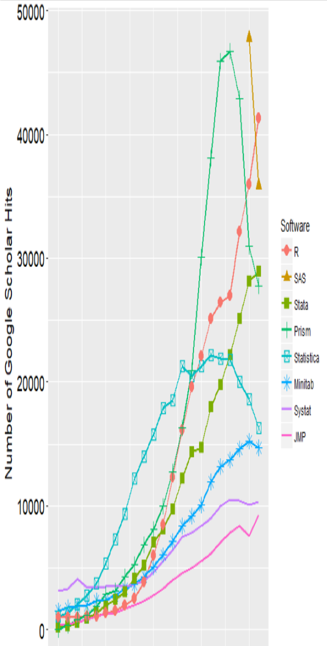
“RStudio is like an Instagram filter over R, to make your R user experience better.” - Joe Fruehwald (R course)

- Check out Joe’s Workshop at this year’s Newcastle Postgraduate Conference in Linguistics

Figure 1: caption

Why are we using R?





Data files

- Need to be `.csv` or `.txt` NOT Excel
- Need to have one token per row

Bad data formatting for R

- Never organise your data like this:

Demonstration: importing and data basics

Importing data

- If you are importing `.txt` files, use `read.delim()`
- If you are importing `.csv` files, use `read.csv()`
- Also possible:
- Double slashes for PC, single forward slashes for Mac

General tips

- R is case sensitive (it will treat an `s` as a completely different character to `S`)
- Don't use spaces in your filenames and folders
- Softwrap your code by going to **Tools** > **Global Options** > **Code** and ticking **Soft-wrap R source files**.

Basics: data assignment

```
## [1] 10
```

"There are only two hard things in Computer Science: cache invalidation and naming things." — Phil Karlton

For best practices on naming variables, checkout the tidyverse style guide by Hadley Wickham

Factors and levels: `factors`

Our variables are called **factors** in R terminology.

```
## [1] 6596   34
```

```
##  [1] "sex"            "occupation"     "age"
##  [5] "town"           "postcode_birth" "postcode_now"
##  [9] "furniture"      "clothing"       "evening_meal"
## [13] "foot_strut"     "for_more"       "one_gone"
## [17] "fur_bear"       "sauce_source"   "pour_poor"
## [21] "bangor_banger"  "mute_moot"      "spa_spar"
## [25] "give_it_me"     "I_done_it"      "it_was"
## [29] "beaches_was"    "I_werent"       "they_was"
## [33] "dress_what"     "things_what"
```

Factors and levels: `factor levels`

Our variables are called **factors** in R terminology. Each option for a factor is a **factor level**.

```
## [1] "don't rhyme" "rhyme"
```

Simple functions: head and tail, dim

```
##       sex               occupation age age_group
## 1 female                  Teacher  32     young
## 2 female Government Administrator  47    middle
## 3   male    Management Consultant  61       old
## 4 female                  student  19     young
## 5   male               Accountant  34     young
## 6   male                  Retired  63       old
##                                     town postcode_birth p
## 1 Bishopton, Renfrewshire, Scotland, UK            PA7
## 2                            Dumbarton            G82
## 3                             EDNBURGH           EH15
## 4                                wigan            WA3
## 5               Bellshill\nAtherstone             CV9
## 6                              Bristol            BS8
##   furniture clothing evening_meal   group  foot_strut
## 1     couch trousers       dinner you all don't rhyme do
## 2      sofa trousers       dinner     you don't rhyme do
## 3    settee trousers       dinner     you don't rhyme do
```

Simple functions: dim and colnames

```
## [1] 6596   34
```

```
##  [1] "sex"            "occupation"     "age"
##  [5] "town"           "postcode_birth" "postcode_now"
##  [9] "furniture"      "clothing"       "evening_meal"
## [13] "foot_strut"     "for_more"       "one_gone"
## [17] "fur_bear"       "sauce_source"   "pour_poor"
## [21] "bangor_banger"  "mute_moot"      "spa_spar"
## [25] "give_it_me"     "I_done_it"      "it_was"
## [29] "beaches_was"    "I_werent"       "they_was"
## [33] "dress_what"     "things_what"
```

Download the materials for today's class here

https://danielleturton.github.io/quantling

Packages

- What are packages?
- The first time you use them, you will need to install the packages (you only need to do this once)

```r
install.packages("dplyr")
install.packages("ggplot2")
```

- Load the packages

```r
library(dplyr)
library(ggplot2)
```