An Introduction to R

Tuesday 16th May 2023

Dr Simon Rudkin

University of Manchester

RStudio

MANCHESTER
1824

Preliminary

- Download R from https://cran.r-project.org/
- Download R studio from https://posit.co/download/rstudio-desktop/
- Cluster PCs can add using the software centre (if not already added)

By the end of the session attendees will be able to

Download R from https://cran.r-project.org/

Download R studio from https://posit.co/download/rstudio-desktop/

1. Install R and RStudio on their own machine
2. Be familiar with the RStudio GUI
3. Understand R variables, data types and objects

By the end of the session attendees will be able to

Download R from https://cran.r-project.org/

Download R studio from https://posit.co/download/rstudio-desktop/

1. Understand the use of vectors and Dataframes
2. Understand how to get help and make use of R libraries
3. Read datasets of different formats into the R environment

By the end of the session attendees will be able to

Download R from https://cran.r-project.org/

Download R studio from https://posit.co/download/rstudio-desktop/

1. Perform data cleaning and manipulation using core R and the 'tidyverse' package
2. Perform visualisations of data using the 'ggplot2' package
3. Appreciate how such things as statistical analysis, machine learning and mapping of data can be performed using a variety of R packages which are readily available

Schedule

09:30	Introduction	13:30	Further plotting
10:00	Loading and viewing data	14:15	Summarising data
10:30	Introduction to the tidyverse	14:45	Exercises and own data
11:00	Working in the tidyverse	15:45	Further methods and notes
11:30	Creating variables	16:15	Summary and review
11:50	Introduction to plotting		
12:10	Summary of morning		
Lunch Break 12:30-13:30			

This schedule is indicative and sections may be lengthened / shortened as appropriate

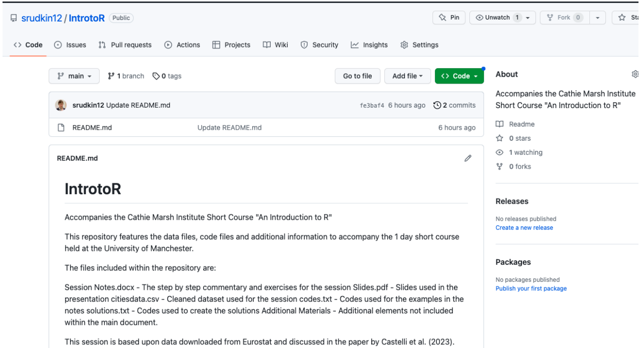
Preliminaries

In order to complete this session you should have:

- A blank word document into which you can paste output from the session
- A blank notepad file into which you can paste any code
- A folder which contains the downloaded material from the GitHub site

Introduction to GitHub next

GitHub https://github.com/srudkin12/IntrotoR



Accompanies the Cathie Marsh Institute Short Course "An Introduction to R"

IntrotoR

Accompanies the Cathie Marsh Institute Short Course "An Introduction to R"

This repository features the data files, code files and additional information to accompany the 1 day short course held at the University of Manchester.

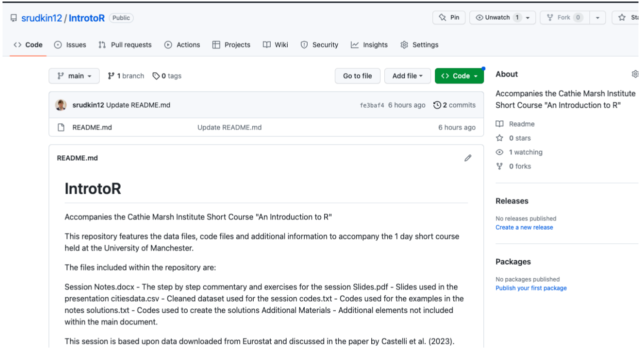
The files included within the repository are:

Session Notes.docx - The step by step commentary and exercises for the session Slides.pdf - Slides used in the presentation citiesdata.csv - Cleaned dataset used for the session codes.txt - Codes used for the examples in the notes solutions.txt - Codes used to create the solutions Additional Materials - Additional elements not included within the main document.

This session is based upon data downloaded from Eurostat and discussed in the paper by Castelli et al. (2023).

- GitHub is a site used by statistics / data science community for projects
- All files relating to this session on GitHub LINK
- Download the files on next slide

GitHub https://github.com/srudkin12/IntrotoR 2



About

Accompanies the Cathie Marsh Institute Short Course "An Introduction to R"

Releases

Packages

IntrotoR

Accompanies the Cathie Marsh Institute Short Course "An Introduction to R"

This repository features the data files, code files and additional information to accompany the 1 day short course held at the University of Manchester.

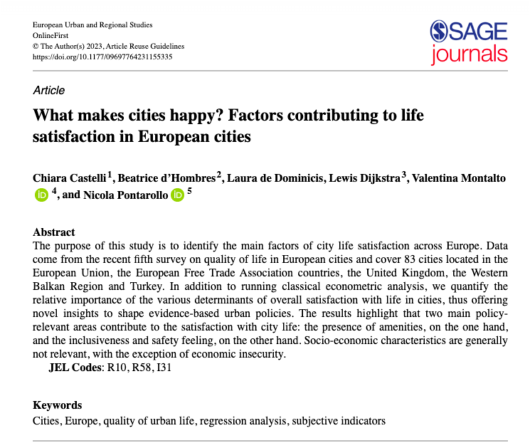
The files included within the repository are:

Session Notes.docx - The step by step commentary and exercises for the session Slides.pdf - Slides used in the presentation citiesdata.csv - Cleaned dataset used for the session codes.txt - Codes used for the examples in the notes solutions.txt - Codes used to create the solutions Additional Materials - Additional elements not included within the main document.

This session is based upon data downloaded from Eurostat and discussed in the paper by Castelli et al. (2023).

- GitHub is a site used by statistics / data science community for projects
- All files relating to this session on GitHub LINK
- UPDATE SO THE GITHUB SHOWS THE FILES

Data



European Urban and Regional Studies
OnlineFirst
© The Author(s) 2023, Article Reuse Guidelines
https://doi.org/10.1177/09697764231155335

Article

What makes cities happy? Factors contributing to life satisfaction in European cities

Chiara Castelli[1], Beatrice d'Hombres[2], Laura de Dominicis, Lewis Dijkstra[3], Valentina Montalto [4], and Nicola Pontarollo [5]

Abstract

The purpose of this study is to identify the main factors of city life satisfaction across Europe. Data come from the recent fifth survey on quality of life in European cities and cover 83 cities located in the European Union, the European Free Trade Association countries, the United Kingdom, the Western Balkan Region and Turkey. In addition to running classical econometric analysis, we quantify the relative importance of the various determinants of overall satisfaction with life in cities, thus offering novel insights to shape evidence-based urban policies. The results highlight that two main policy-relevant areas contribute to the satisfaction with city life: the presence of amenities, on the one hand, and the inclusiveness and safety feeling, on the other hand. Socio-economic characteristics are generally not relevant, with the exception of economic insecurity.

JEL Codes: R10, R58, I31

Keywords
Cities, Europe, quality of urban life, regression analysis, subjective indicators

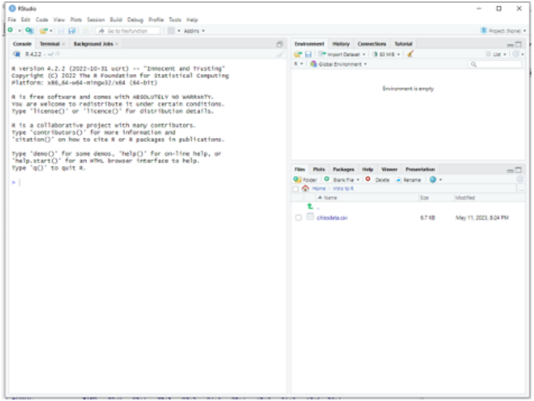
- Session uses paper by Castelli et al. (2023) on the satisfaction of individuals living in European cities
- Consider satisfaction with facilities, safety and area affordability
- Survey data collected by Eurostat
- Data link: CLICK HERE
- Cleaned subset of variables is available on the GitHub

Todays Question

What makes cities happy?

Public Transport	Health Services	Cultural Facilities
Green Spaces	Public Squares	Cleanliness
Trust in Others	Safety	Affordability

RStudio Environment



```
R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment is empty

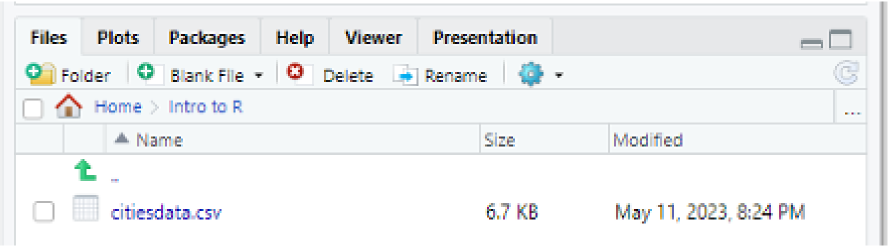
Name	Size	Modified
..		
citiesdata.csv	6.7 KB	May 11, 2023, 8:24 PM

Elements of RStudio:

- Terminal (Left)
- Environment (Top Right)
- Files and plots (Bottom Right)

Those familiar with R will know the terminal as the only window when using R standalone

Setting the Working Directory



	Name	Size	Modified
	..		
□	citiesdata.csv	6.7 KB	May 11, 2023, 8:24 PM

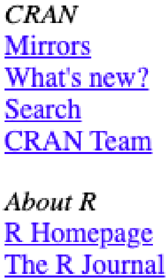
- Navigate to the folder created for todays session
- Screenshot shows the citiesdata.csv file in folder

Loading and Viewing Data

Complete pages 4 to 6 of the accompanying notes

Questions 1 to 4 should also be attempted

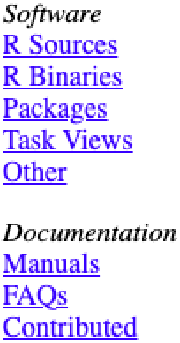
R Packages



CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

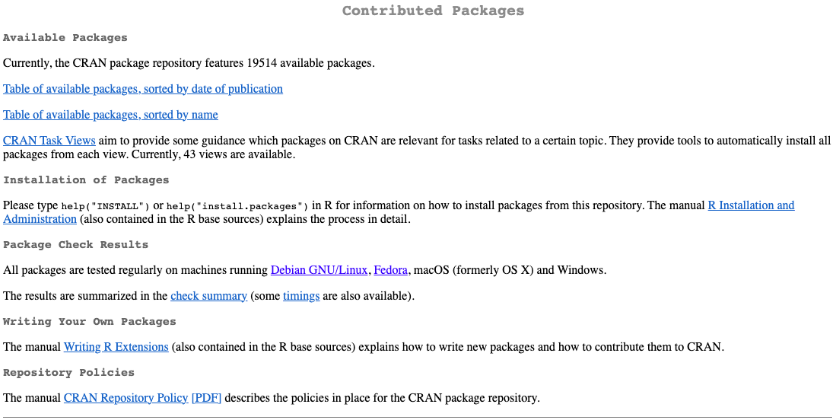
- On the side bar you will see link to packages
- Packages are functions, or groups of functions, written by R users
- Packages are available for most analyses
- We will use the collection of packages `tidyverse` (Wickham and Grolemund, 2016)



Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

R Packages 2



Contributed Packages

Available Packages

Currently, the CRAN package repository features 19514 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

CRAN Task Views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic. They provide tools to automatically install all packages from each view. Currently, 43 views are available.

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux, Fedora, macOS (formerly OS X) and Windows.

The results are summarized in the check summary (some timings are also available).

Writing Your Own Packages

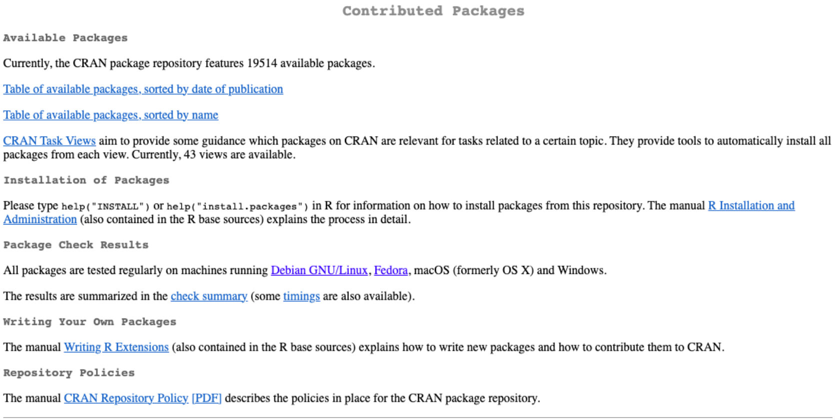
The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

Repository Policies

The manual CRAN Repository Policy [PDF] describes the policies in place for the CRAN package repository.

Packages page has details of how to build packages and links to lists of new packages

R Packages 3



Contributed Packages

Available Packages

Currently, the CRAN package repository features 19514 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

CRAN Task Views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic. They provide tools to automatically install all packages from each view. Currently, 43 views are available.

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux, Fedora, macOS (formerly OS X) and Windows.

The results are summarized in the check summary (some timings are also available).

Writing Your Own Packages

The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

Repository Policies

The manual CRAN Repository Policy [PDF] describes the policies in place for the CRAN package repository.

Task Views are collections of packages which are linked to specific tasks. There is an option to download and install all packages within a particular Task View

R Packages 4: Two commands

```
install.packages(‘‘packagename’’)
```

```
library(packagename)
```

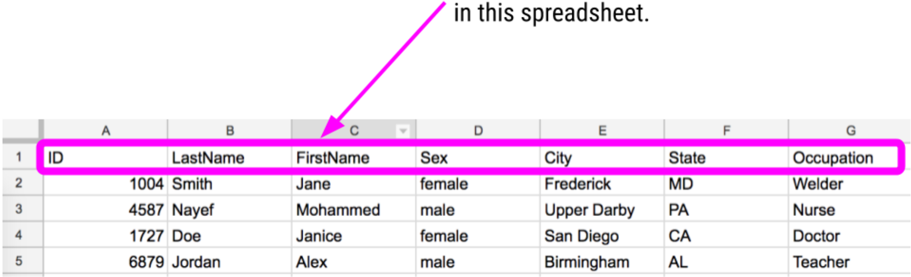
- When installing packages include the name in “ ”
- Only need to install packages once*

The Tidyverse (Wickham and Grolemund, 2016)

an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures

Tidy Data

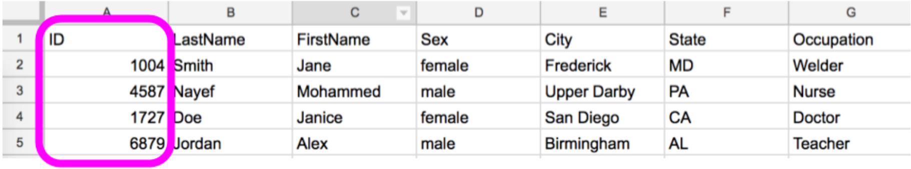
There are 7 different **variables** in this spreadsheet.



	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Tidy Data

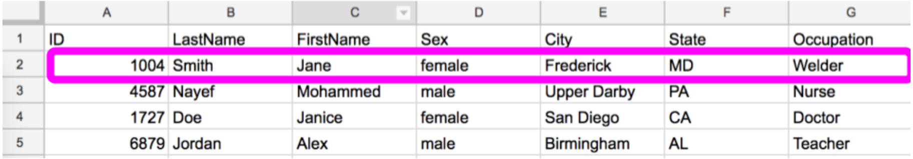
1. Each variable you measure should be in a **single column**



	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

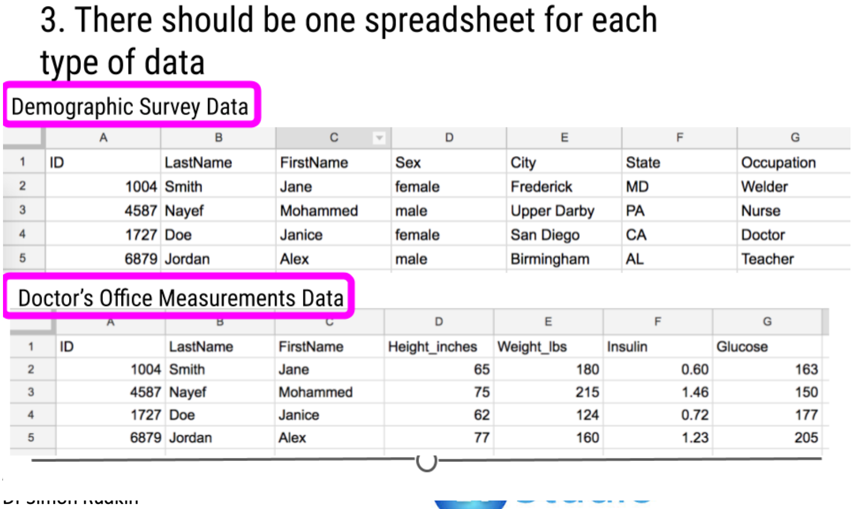
Tidy Data

2. Every observation of a variable should be in a **different row**



	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Tidy Data



3. There should be one spreadsheet for each type of data

Demographic Survey Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

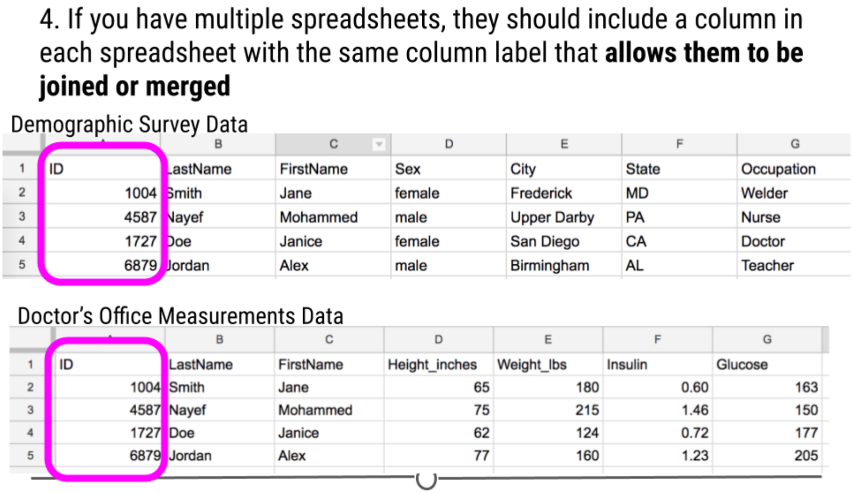
Doctor's Office Measurements Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

Note that in this session all merging has been done

https://github.com/srudkin12/IntrotoR

Tidy Data



4. If you have multiple spreadsheets, they should include a column in each spreadsheet with the same column label that **allows them to be joined or merged**

Demographic Survey Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Doctor's Office Measurements Data

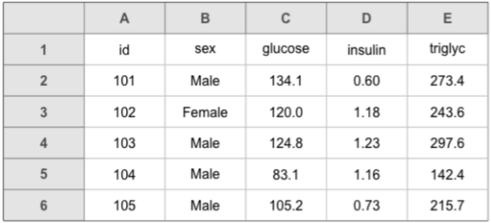
	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

Note that in this session all merging has been done

Tidy Data

Tidy data = rectangular data

A



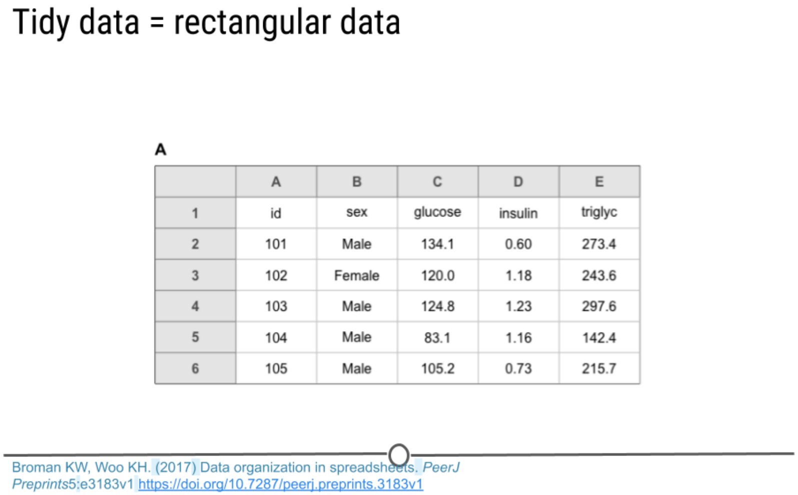
	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

A spreadsheet may also be thought of as **dataframe**

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints*5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

Dr Simon Rudkin
https://github.com/srudkin12/IntrotoR

Tidy Data



Tidy data = rectangular data

A

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints*5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

A spreadsheet may also be thought of as **dataframe**

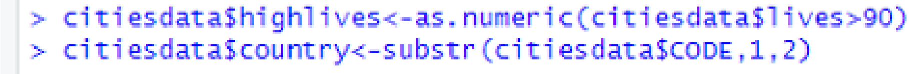
R as an Object Based Language

- Objects have no interpretation until we assign
- R allows users to assign value to objects
- Objects may be single numbers, variables or tables
- Objects may also be collections of results from models
- Many packages have further object types
- Begin with R as a calculator on Page 9

R as an Object Based Language 2

Complete page 9 of the accompanying notes

- *Questions 5 to 8 should also be attempted*
- Produce two variables for the next stages:



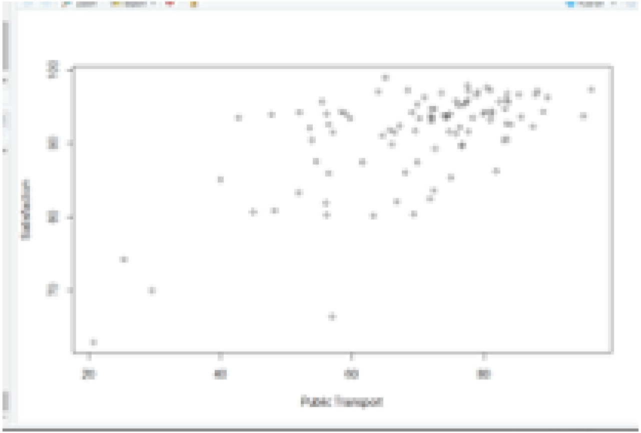
```
> citiesdata$highlives<-as.numeric(citiesdata$lives>90)
> citiesdata$country<-substr(citiesdata$CODE,1,2)
```

Plotting in R

`plot(` *horizontal axis*
vertical axis
options
)

- Axis labels (`xlab = “xlabel”` and `ylab = “ylabel”`
- Point colour, style and size (`color =` and `pch =` and `cex =`)
- Plot type default scatter can change to line (`type = "l"`)

Plotting in R 2



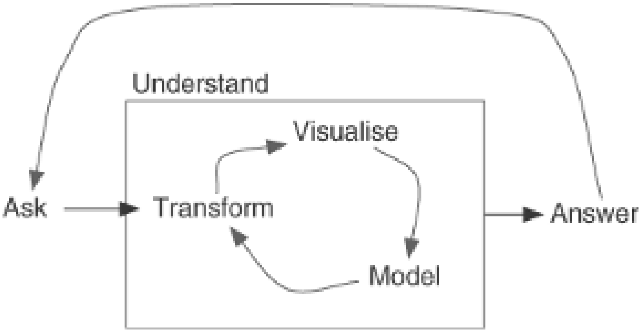
```
plot(citiesdata$ptrans,
citiesdata$lives,
xlab="Public Transport",
ylab="Satisfaction")
```

Plotting in R 3

Complete page 10 of the accompanying notes

- *Questions 9 to 12 should also be attempted*
- There are many further plotting options in base R
- Further types include boxplots, bar charts, pie charts, histograms, density plots etc.
- Many guides available including STHDA

First Half Summary



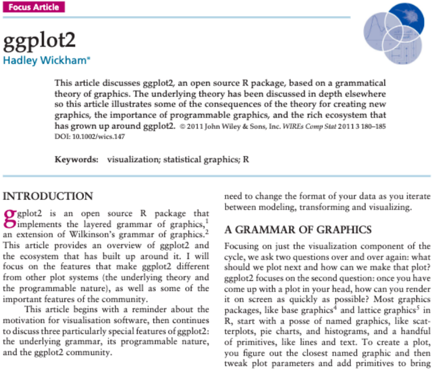
- Project flow from Wickham and Grolemund (2016)
- Visualisation is very important Anscombe (1973)
- Introduced R studio
- See R as an object based language
- Read in data and simple manipulations
- Plots in base R

Schedule

09:30	Introduction	13:30	Further plotting
10:00	Loading and viewing data	14:15	Summarising data
10:30	Introduction to the tidyverse	14:45	Exercises and own data
11:00	Working in the tidyverse	15:45	Further methods and notes
11:30	Creating variables	16:15	Summary and review
11:50	Introduction to plotting		
12:10	Summary of morning		
Lunch Break 12:30-13:30			

This schedule is indicative and sections may be lengthened / shortened as appropriate

ggplot package



Focus Article

ggplot2

Hadley Wickham*

This article discusses ggplot2, an open source R package, based on a grammatical theory of graphics. The underlying theory has been discussed in depth elsewhere so this article illustrates some of the consequences of the theory for creating new graphics, the importance of programmable graphics, and the rich ecosystem that has grown up around ggplot2. © 2011 John Wiley & Sons, Inc. *WIREs Comp Stat* 2011 3 180–185
DOI: 10.1002/wics.147

Keywords: visualization; statistical graphics; R

INTRODUCTION

ggplot2 is an open source R package that implements the layered grammar of graphics,[1] an extension of Wilkinson's grammar of graphics.[2] This article provides an overview of ggplot2 and the ecosystem that has built up around it. I will focus on the features that make ggplot2 different from other plot systems (the underlying theory and the programmable nature), as well as some of the important features of the community.

This article begins with a reminder about the motivation for visualisation software, then continues to discuss three particularly special features of ggplot2: the underlying grammar, its programmable nature, and the ggplot2 community.

need to change the format of your data as you iterate between modeling, transforming and visualizing.

A GRAMMAR OF GRAPHICS

Focusing on just the visualization component of the cycle, we ask two questions over and over again: what should we plot next and how can we make that plot? ggplot2 focuses on the second question: once you have come up with a plot in your head, how can you render it on screen as quickly as possible? Most graphics packages, like base graphics[4] and lattice graphics[5] in R, start with a posse of named graphics, like scatterplots, pie charts, and histograms, and a handful of primitives, like lines and text. To create a plot, you figure out the closest named graphic and then tweak plot parameters and add primitives to bring

- `ggplot2` is the current name for graphics package in tidyverse
- Installs and loads as part of tidyverse
- Plotting commands are different as we shall see
- Many options beyond what we cover here

Plotting with ggplot

```
ggplot(data=citiesdata) +
  geom_point(mapping=aes(x = ptrans, y = lives, size =
  highlives))
  labs(x = "Public Transport", y = "Life Satisfaction")
```

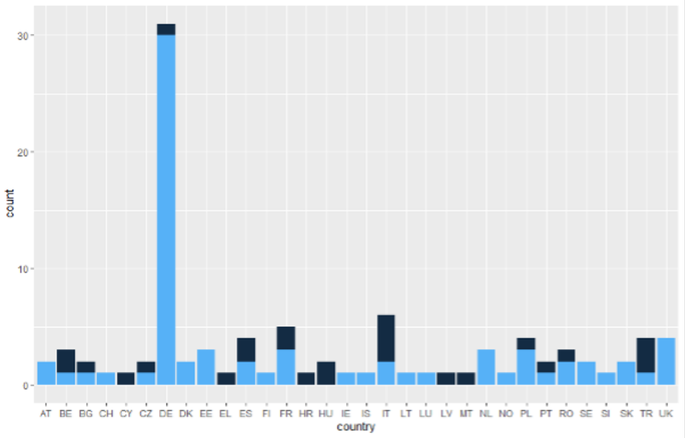
- ggplot uses the + notation to say add something else
- First argument informs about the data
- Second argument are about what to plot
- Third line is about the axis labels

Plotting with ggplot2

Complete pages 13 to 15 of the accompanying notes

- *Questions 13 to 18 should also be attempted*
- Here we have seen how the scatter plot can be enhanced in ggplot
- Aim to gain inference on our overall question of the session

Plotting with ggplot2: Bar Charts



- Code will build up to producing this plot
- Should there be a legend?
- There are many more options to enhance

Plotting with ggplot2: Bar Charts 2

Complete pages 16, 17 and the top half of page 18 of the accompanying notes

- *Questions 19 and 20 should also be attempted*
- Colouration within the bar chart helps visualise with low categories
- Which other categories could we create?
- Consult further guides to plotting including Wickham (2011)

Arranging Data

```
arrange(citiesdata,lives)
```

```
arrange(citiesdata,desc(lives))
```

- Assign the arrange outcome to an object to store
- Can also use the column names in the top left window to sort

Grouping and Summarising Data

```
by_country<-group_by(citiesdata,country)
```

```
clives<-summarise(by_country,mlives=mean(lives,na.rm=TRUE))
```

- Define the grouping process using group_by
- Assign the summary to a new object
- Watch what happens in the top right Environment tab

Arranging and Summarising Data

Complete the remaining pages of the accompanying notes

- *Questions 21 to 28 should also be attempted*
- There are many standard summary functions `mean()`, `sd()`, `min()`, `max()`
- Can also obtain quantiles `quantile(<variable>, quantile=0.10)`
- Other functions are available

Exercise

How can we use the data within citiesdata.csv to understand the factors which link to satisfaction with life in the cities?

Further Topics

```
install.packages(datasauRus)
```

- Install the `datasauRus` package from Matejka and Fitzmaurice (2017)
- Work through the Vignette at https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html
- Lesson is that we should always look at data and not just summary statistics - this session has shown all elements

Summary

- R is an object based language for statistical analysis
- RStudio provides a GUI for using R in an intuitive way
- The `tidyverse` offers a well used suite of packages with data philosophy
- R has plotting functionality within base R - `ggplot2` adds functionality
- R supports statistical analysis with a wealth of specialist functions
- This session just introduced R...

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Castelli, C., d'Hombres, B., Dominicis, L. d., Dijkstra, L., Montalto, V., and Pontarollo, N. (2023). What makes cities happy? factors contributing to life satisfaction in european cities. *European Urban and Regional Studies*, page 09697764231155335.

Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1290–1294.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2):180–185.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".