Introduction to Data Science

Jan Kirenz

2021-01-29

Contents

Welcome

This course contains materials to learn data acquisition and wrangling, exploratory data analysis, data visualization, inference, modelling, and effective communication of results using the programming language R.

The goal of this course is to enable students to work on a fully reproducible data science project analysing a dataset of their choice and answering questions they care about.

License

Acknowledements

This introductory data science course is based on Mine Çetinkaya-Rundel's excellent "Data Science in a Box" materials.

The website is built with bookdown.

Part I

Hello

Chapter 1

Overview

Hello!

The core content of this introductory data science course focuses on

- data acquisition and wrangling,
- exploratory data analysis,
- data visualization,
- inference,
- modelling, and
- effective communication of results.

A heavy emphasis is placed on a consistent syntax (with tools from the tidyverse), reproducibility (with R Markdown), and version control and collaboration (with Git and GitHub).

The RStudio Cloud workspace for your course is provided in your Moodle-Course. You have to join the workspace to use the sample application exercises.

Part II

Course content

Chapter 2

Hello world

In this first application, we conduct our first data visualization.

The RStudio Cloud workspace for your course is provided in your Moodle-Course. You have to join the workspace to use the sample application exercises.

2.1 Slides and application exercises

Unit 1 - Deck 1: Welcome

Slides

First dataviz

> **Option 1 - UN Votes**
>
> Source
>
> **Option 2 - COVID-19**
>
> Source

Unit 1 - Deck 2: Meet the toolkit - Programming

Slides

Source

- R4DS :: Chp 2 - Introduction
- IMS :: Sec 1.1 & 1.2 - Case study & Data basics

Bechdel + R Markdown

Source

Unit 1 - Deck 3: Meet the toolkit - Version control and collaboration

Slides

Source

Chapter 3

Exploring data

This unit focuses on data visualization and data wrangling. Specifically we cover fundamentals of data and data visualization, confounding variables, and Simpson's paradox as well as the concept of tidy data, data import, data cleaning, and data curation.

We end the unit with web scraping and introduce the idea of iteration in preparation for the next unit.

Also in this unit we introduce the toolkit: R, RStudio, R Markdown, Git, and GitHub.

3.1 Slides, videos, and application exercises

3.1.1 Visualising data

Unit 2 - Deck 1: Data and visualisation

Slides

Source

Unit 2 - Deck 2: Visualising data with ggplot2

Slides

Source

R4DS :: Chp 3 - Data visualization

Unit 2 - Deck 3: Visualising numerical data

Slides

Source

IMS :: Sec 2.1 - Exploring numerical data

Unit 2 - Deck 4: Visualising categorical data

Slides

Source

IMS :: Sec 2.2 - Exploring categorical data

StarWars + Dataviz

Source

3.1.2 Wrangling and tidying data

Unit 2 - Deck 5: Tidy data

Slides

Source

JSS :: Tidy data

Unit 2 - Deck 6: Grammar of data wrangling

Slides

Source

Unit 2 - Deck 7: Working with a single data frame

Slides

Source

R4DS :: Chp 5 - Data transformation

Unit 2 - Deck 8: Working with multiple data frames

Slides

Source

R4DS :: Chp 13 - Relational data

Unit 2 - Deck 9: Tidying data

Slides

Source

R4DS :: Chp 12 - Tidy data

Hotels + Data wrangling

Source

3.1.3 Importing and recoding data

Unit 2 - Deck 10: Data types

Slides

Source

Unit 2 - Deck 11: Data classes

Slides

Source

R4DS :: Chp 15 - Factors

Unit 2 - Deck 12: Importing data

Slides

Source

R4DS :: Chp 11 - Data import

Unit 2 - Deck 13: Recoding data

Slides

R4DS :: Sec 16.1 - 16.3 - Dates and times

Hotels + Data types

Source

Source

Nobels + Sales + Data import

Source

Source

3.1.4 Communicating data science results effectively

Unit 2 - Deck 14: Tips for effective data visualization

Slides

Source

IMS :: Sec 2.3 - Effective data visualisation

Brexit + Telling stories with dataviz

Source

Unit 2 - Deck 15: Scientific studies and confounding

Slides

Source

- IMS :: Sec 1.3 - Sampling principles and strategies
- IMS :: Sec 1.4 - Experiments

Unit 2 - Deck 16: Simpson's paradox

Slides

Source

Unit 2 - Deck 17: Doing data science

Slides

Source

R4DS :: Chp 7 - Exploratory data analysis

3.1.5 Web scraping and programming

Unit 2 - Deck 18: Web scraping

Slides

Source

Unit 2 - Deck 19: Scraping top 250 movies on IMDB

Slides

Source

Unit 2 - Deck 20: Web scraping considerations

Slides

Source

IMDB + Web scraping

Source

Unit 2 - Deck 21: Functions

Slides

Source

R4DS :: Chp 19 - Functions

Unit 2 - Deck 22: Iteration

Slides

Source

R4DS :: Chp 20 - Iteration

3.2 Labs

Lab 1: Hello R

Introduction to R, R Markdown, Git, and GitHub

Instructions

Source

Starter

Lab 2: Plastic waste

Introduction to working with data in R with the tidyverse

Instructions

Source

Starter

Lab 3: Nobel laureates

Data wrangling and tidying

Instructions

Source

Starter

Lab 4: La Quinta is Spanish for 'next to Denny's,' Pt. 1

Visualizing spatial data

Instructions

Source

Starter

Lab 5: La Quinta is Spanish for 'next to Denny's,' Pt. 2

Wrangling spatial data

Instructions

Source

Starter

Lab 6: Sad plots

Critiquing and improving data visualisations

Instructions

Source

Starter

Lab 7: Simpson's paradox

Data visualisation, confounding, multivariable relationships

Instructions

Source

Starter

Lab 8: University of Edinburgh Art Collection

Web scraping, function, iteration

Instructions

Source

Starter

3.3 Homework assignments

HW 1: Pet names

Introduction to working with data in R with the tidyverse

Instructions

Source

Starter

HW 2: Edinburgh Airbnb rentals

Data visualisation with the tidyverse

Instructions

Source

Starter

HW 3: Road traffic accidents

Data wrangling, tidying, and visualization

Instructions

Source

Starter

HW 4: What should I major in?

More data wrangling, summarizing, and visualization

Instructions

Source

Starter

HW 5: Legos

More data wrangling, summarizing, and visualization

Instructions

Source

Starter

HW 6: Money in politics

Web scraping, functions, and iteration

Instructions

Source

Starter

Chapter 4

Making rigorous conclusions

In this part we introduce modelling and statistical inference for making data-based conclusions.

We discuss building, interpreting, and selecting models, visualizing interaction effects, and prediction and model validation.

Statistical inference is introduced from a simulation based perspective, and the Central Limit Theorem is discussed very briefly to lay the foundation for future coursework in statistics.

4.1 Slides, videos, and application exercises

4.1.1 Modelling data

Unit 4 - Deck 1: The language of models

Slides

Source

Unit 4 - Deck 2: Fitting and interpreting models

Slides

Source

IMS :: Chp 3 - Introduction to linear models

Unit 4 - Deck 3: Modelling nonlinear relationships

Slides

Source

Unit 4 - Deck 4: Models with multiple predictors

Slides

Source

IMS :: Sec 4.1 - Regression with multiple predictors

Unit 4 - Deck 5: More models with multiple predictors

Slides

Source

4.1.2 Classification and model building

Unit 4 - Deck 6: Logistic regression

Slides

Source

IMS :: Sec 4.5 - Logistic regression

Unit 4 - Deck 7: Prediction and overfitting

Slides

Source

tidymodels :: Build a model

Unit 4 - Deck 8: Feature engineering

Slides

Source

tidymodels :: Preprocess your data with recipes

4.1.3 Model validation

Unit 4 - Deck 9: Cross validation

Slides

Source

tidymodels :: Evaluate your model with resampling

The Office + Feature engineering, Pt. 1

Source

The Office + Cross validation, Pt. 2

Source

4.1.4 Uncertainty quantification

Unit 4 - Deck 10: Quantifying uncertainty

Slides

Source

Unit 4 - Deck 11: Bootstrapping

Slides

Source

IMS :: Sec 5.2 - Bootstrap confidence intervals

Unit 4 - Deck 12: Hypothesis testing

Slides

Source

IMS :: Sec 5.1 - Randomization tests

Unit 4 - Deck 13: Inference overview

Slides

Source

4.2 Labs

Lab 10: Grading the professor, Pt. 1

Fitting and interpreting simple linear regression models

Instructions

Source

Starter

Lab 11: Grading the professor, Pt. 2

Fitting and interpreting multiple linear regression models

Instructions

Source

Starter

Lab 12: Smoking while pregnant

Constructing confidence intervals, conducting hypothesis tests, and interpreting results in context of the data

Instructions

Source

Starter

4.3 Homework assignments

HW 7: Bike rentals in DC

Exploratory data analysis and fitting and interpreting models

Instructions

Source

Starter

HW 8: Exploring the GSS

Fitting and interpreting models

Instructions

Source

Starter

HW 9: Modelling the GSS

Model validation and inference

Instructions

Source

Starter

Chapter 5

Interactive tutorials

The following interactive tutorials have been built with **learnr** and **gradethis**.

They're available on shinyapps.io (linked) as well as distributed with the **dsbox** package.[1] With the dsbox package installed, you can also run these tutorials in the Tutorials pane of your RStudio window.

Tutorial 1: Airbnb listings in Edinburgh

The goal of this tutorial is not to conduct a thorough analysis of Airbnb listings in Edinburgh, but instead to give you a chance to practice your data visualisation and interpretation skills.

[Tutorial] [Source]

Tutorial 2: Road Traffic Accidents

- Continue practising data visualization skills with ggplot2.
- Filter data for certain attributes with `filter()`.
- Create new variables based on existing variables in the data with `mutate()`.

[Tutorial] [Source]

Tutorial 3: What should I major in?

- Continue practising data tidying and visualisation.
- Calculate summary statistics with `summarise()`.
- Arrange output of dplyr chains with `arrange()`.

[1] The dsbox package is not yet on CRAN, until then you will need to install from GitHub with `devtools::install_github("rstudio-education/dsbox")`.

[Tutorial] [Source]

Tutorial 4: Lego sales

- Practice the analysis skills you have learned so far.
- Develop a question you can answer with the data.
- Deepen your understanding of building and interpreting visualisations.

[Tutorial] [Source]

Tutorial 5: Money in US politics

- Get started with scraping data from the web.
- Continue to build on your data cleaning and visualisation skills.

[Tutorial] [Source]

Tutorial 6: Bike Rentals in D.C.

- Continue to hone your data wrangling skills.
- Practice modelling and interpreting model results and performance.
- Conduct backwards selection for finding the "best" model.

[Tutorial] [Source]

Tutorial 7: Exploring the General Social Survey

- Work on your data manipulation skills.
- Fit linear models with multiple predictors.
- Interpret regression output.

[Tutorial] [Source]

Tutorial 8: Bootstrapping the General Social Survey

- Continue to hone your data wrangling skills.
- Use bootstrapping to construct confidence intervals.
- Interpret of confidence intervals in context of the data.

[Tutorial] [Source]