HOMEWORK 1 | TASK 2

CSCI-548

Sarvesh Parab
sparab@usc.edu

Scrapper questions – answers:

Q1. *What are the data science keywords you used for this task? What is the website you are extracting data from (give the base URI and a one-line description about the website)? Name and describe each of the fields (at least 8) you are extracting, provide a representative screenshot of the website and annotate (on the screenshot) the field values that you are extracting. In order to describe the fields, use a table with two columns. The first column will be the name of the field and the second its description.*

A1. The website I have used is : https://www.datacamp.com/. DataCamp.com provides a wide variety of data science courses for free (9 courses per month). They also have skills and career tracks groomed to master specific areas in data science.
The fields I am extracting :

1. URL: The web URL for the course content *[Mandatory Field]*
2. Title: The title of the course *[Mandatory Field]*
3. Author:
    a. Name: The name of the author *[Mandatory Field]*
    b. Organization: The organization/affiliation the author/speaker has *[Mandatory Field]*
    c. URL: Link to a page for more information about he author and lists more courses related to data science covered by the same author *[Extra Field]*
4. Description: A short brief about the course and its objectives *[Mandatory Field]*
5. Duration: The length of the course in hours *[Extra Field]*
6. Participants: Number of people who have enrolled in this course in the past *[Extra Field]*
7. Videos: The number of videos in the course *[Extra Field]*
8. Exercises: The number of hands-on exercises included in the course *[Extra Field]*
9. Chapters: The list of topics/chapters covered in the course *[Extra Field]*
10. Datasets:
    a. Name: Of the dataset used/leveraged in the course *[Extra Field]*
    b. URL: Link to the dataset files (.csv, .zip, .dat, etc.) *[Extra Field]*

Q2. *In 1-3 sentences write the name of the tool you used to scrape the data and describe why you decided to use that tool.*

A2. I have used 'BeautifulSoup4' in Python 3 to scrape the data from DataCamp.com. Since the website has a very well-structured HTML and CSS layout, I could leverage the BS4's API to crawl through and parse the HTML structure of the website.

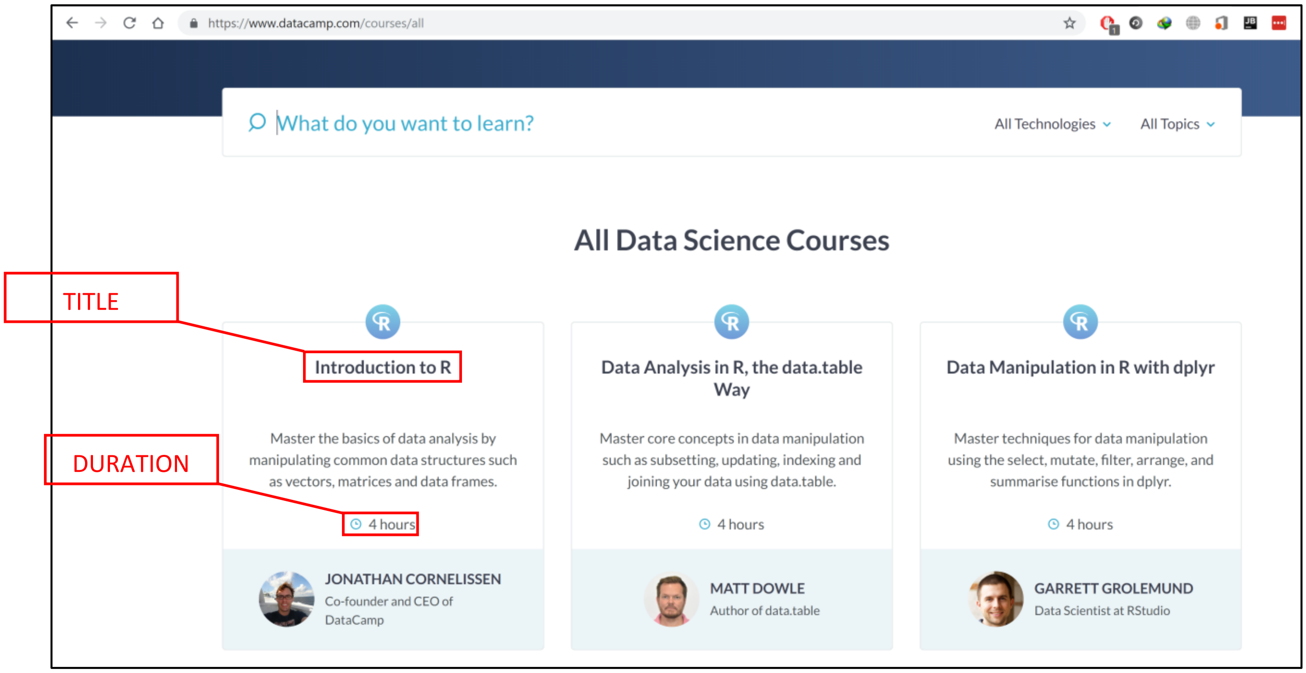
Q3. *In a short paragraph, describe your wrapper. Try to be as specific as possible. We will be looking for details like what kind of wrapper you used (e.g. manual, automatic…?), what is the wrapper model (e.g. HLRT, LR…?), where we can access the wrapper algorithm, if your wrapper is non-manual etc. A good rule of thumb is, can someone familiar with wrappers read your description and be able to (roughly) replicate your wrapper for themselves?*

A3. The wrapper I have built is a manual wrapper. I have gone through the website source code and the challenge was to find and isolate tags and their ids or classes which would uniquely identify a specific element of the webpage, like the fields I was interested in. Once I had the specific identifiers, BS4 provides the API to parse and extract the field values from the page source code.
Then I created a dict in python to hold all the extracted values and make a JSON dump of all the data.
I have used beautifulSoup4 and my python code is well commented for easy maintainability.

Few reference links/materials I used to build the wrapper:

- https://www.dataquest.io/blog/web-scraping-tutorial-python/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Screenshots and annotations of the fields extracted:


https://www.datacamp.com/courses/all
What do you want to learn?
All Technologies
All Topics
TITLE
DURATION

All Data Science Courses

Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

4 hours

JONATHAN CORNELISSEN
Co-founder and CEO of DataCamp

Data Analysis in R, the data.table Way

Master core concepts in data manipulation such as subsetting, updating, indexing and joining your data using data.table.

4 hours

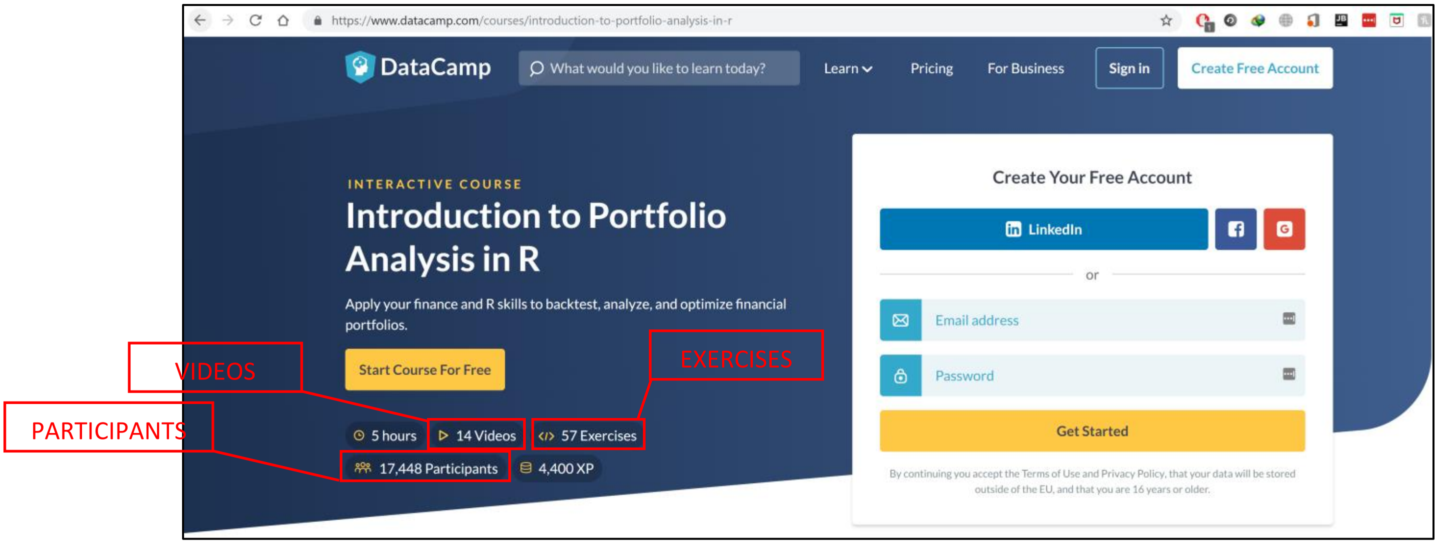
MATT DOWLE
Author of data.table

Data Manipulation in R with dplyr

Master techniques for data manipulation using the select, mutate, filter, arrange, and summarise functions in dplyr.

4 hours

GARRETT GROLEMUND
Data Scientist at RStudio



INTERACTIVE COURSE

Introduction to Portfolio Analysis in R

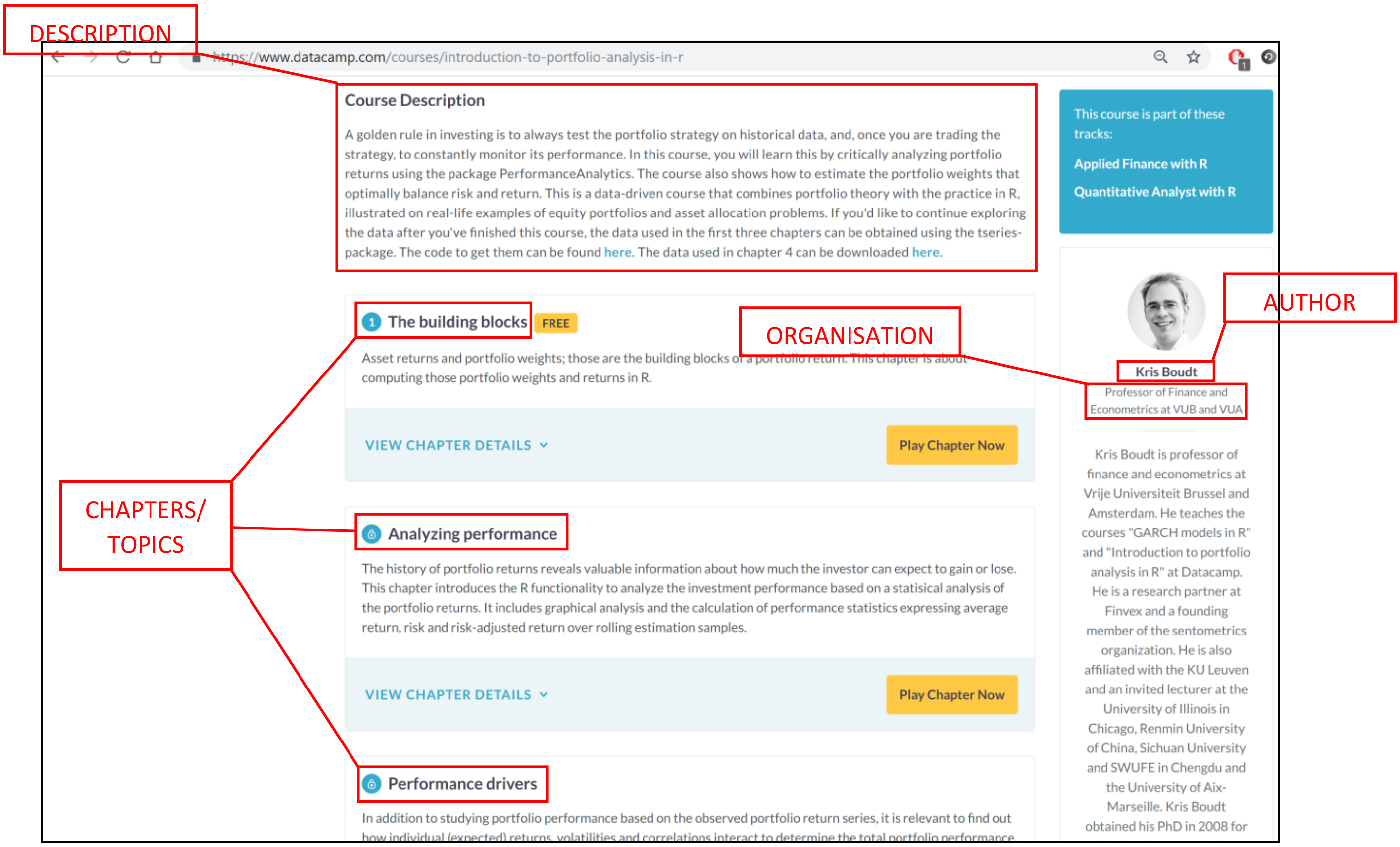
Apply your finance and R skills to backtest, analyze, and optimize financial portfolios.

5 hours | 14 Videos | 57 Exercises

17,448 Participants | 4,400 XP

Create Your Free Account

By continuing you accept the Terms of Use and Privacy Policy, that your data will be stored outside of the EU, and that you are 16 years or older.



Course Description

A golden rule in investing is to always test the portfolio strategy on historical data, and, once you are trading the strategy, to constantly monitor its performance. In this course, you will learn this by critically analyzing portfolio returns using the package PerformanceAnalytics. The course also shows how to estimate the portfolio weights that optimally balance risk and return. This is a data-driven course that combines portfolio theory with the practice in R, illustrated on real-life examples of equity portfolios and asset allocation problems. If you'd like to continue exploring the data after you've finished this course, the data used in the first three chapters can be obtained using the tseries-package. The code to get them can be found here. The data used in chapter 4 can be downloaded here.

This course is part of these tracks:

Applied Finance with R

Quantitative Analyst with R

1 The building blocks

Asset returns and portfolio weights; those are the building blocks of a portfolio return. This chapter is about computing those portfolio weights and returns in R.

Analyzing performance

The history of portfolio returns reveals valuable information about how much the investor can expect to gain or lose. This chapter introduces the R functionality to analyze the investment performance based on a statisical analysis of the portfolio returns. It includes graphical analysis and the calculation of performance statistics expressing average return, risk and risk-adjusted return over rolling estimation samples.

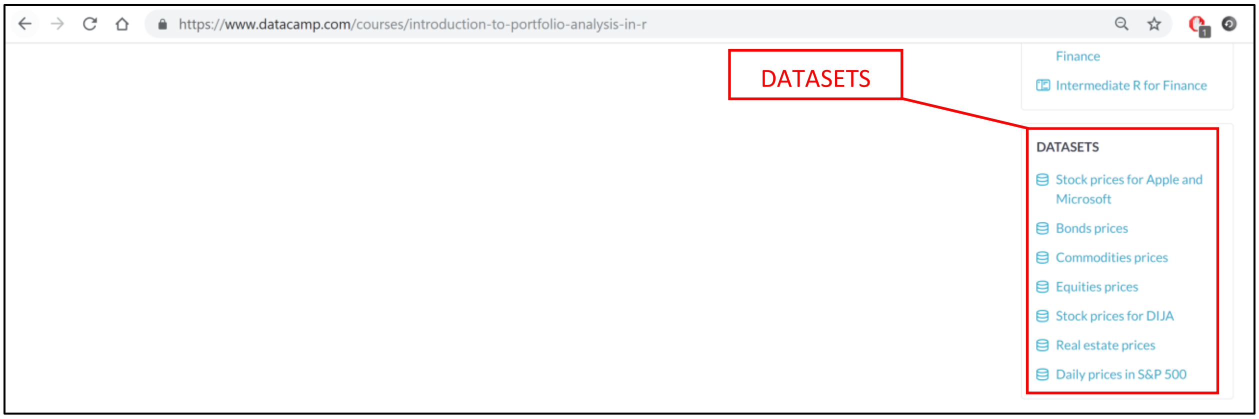
Performance drivers

In addition to studying portfolio performance based on the observed portfolio return series, it is relevant to find out how individual (expected) returns, volatilities and correlations interact to determine the total portfolio performance

Kris Boudt

Professor of Finance and Econometrics at VUB and VUA

Kris Boudt is professor of finance and econometrics at Vrije Universiteit Brussel and Amsterdam. He teaches the courses "GARCH models in R" and "Introduction to portfolio analysis in R" at Datacamp. He is a research partner at Finvex and a founding member of the sentometrics organization. He is also affiliated with the KU Leuven and an invited lecturer at the University of Illinois in Chicago, Renmin University of China, Sichuan University and SWUFE in Chengdu and the University of Aix-Marseille. Kris Boudt obtained his PhD in 2008 for



https://www.datacamp.com/courses/introduction-to-portfolio-analysis-in-r

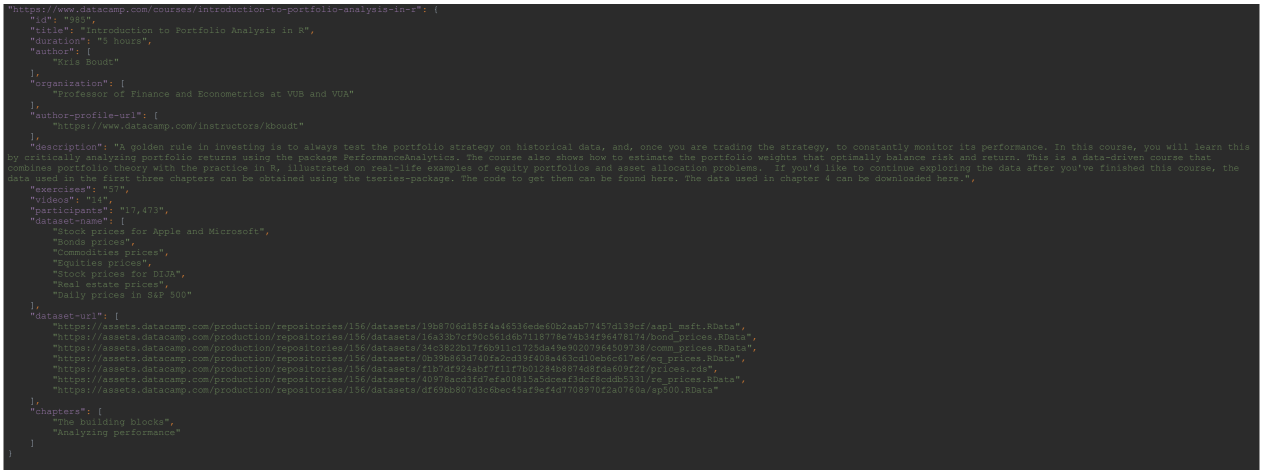
Finance

Intermediate R for Finance

DATASETS

- Stock prices for Apple and Microsoft
- Bonds prices
- Commodities prices
- Equities prices
- Stock prices for DIJA
- Real estate prices
- Daily prices in S&P 500

My JSON Structure: (Sample Course Data)



```
"https://www.datacamp.com/courses/introduction-to-portfolio-analysis-in-r": {
    "id": "985",
    "title": "Introduction to Portfolio Analysis in R",
    "duration": "5 hours",
    "author": [
        "Kris Boudt"
    ],
    "organization": [
        "Professor of Finance and Econometrics at VUB and VUA"
    ],
    "author-profile-url": [
        "https://www.datacamp.com/instructors/kboudt"
    ],
    "description": "A golden rule in investing is to always test the portfolio strategy on historical data, and, once you are trading the strategy, to constantly monitor its performance. In this course, you will learn this
by critically analyzing portfolio returns using the package PerformanceAnalytics. The course also shows how to estimate the portfolio weights that optimally balance risk and return. This is a data-driven course that
combines portfolio theory with the practice in R, illustrated on real-life examples of equity portfolios and asset allocation problems.  If you'd like to continue exploring the data after you've finished this course, the
data used in the first three chapters can be obtained using the tseries-package. The code to get them can be found here. The data used in chapter 4 can be downloaded here.",
    "exercises": "57",
    "videos": "14",
    "participants": "17,473",
    "dataset-name": [
        "Stock prices for Apple and Microsoft",
        "Bonds prices",
        "Commodities prices",
        "Equities prices",
        "Stock prices for DIJA",
        "Real estate prices",
        "Daily prices in S&P 500"
    ],
    "dataset-url": [
        "https://assets.datacamp.com/production/repositories/156/datasets/19b8706d185f4a46536ede60b2aab77457d139cf/aapl_msft.RData",
        "https://assets.datacamp.com/production/repositories/156/datasets/16a33b7cf90c561d6b7118778e74b34f96478174/bond_prices.RData",
        "https://assets.datacamp.com/production/repositories/156/datasets/34c3822b17f6b911c1725da49e90207964509738/comm_prices.RData",
        "https://assets.datacamp.com/production/repositories/156/datasets/0b39b863d740fa2cd39f408a463cd10eb6c617e6/eq_prices.RData",
        "https://assets.datacamp.com/production/repositories/156/datasets/f1b7df924abf7f11f7b01284b8874d8fda609f2f/prices.rds",
        "https://assets.datacamp.com/production/repositories/156/datasets/40978acd3fd7efa00815a5dceaf3dcf8cddb5331/re_prices.RData",
        "https://assets.datacamp.com/production/repositories/156/datasets/df69bb807d3c6bec45af9ef4d7708970f2a0760a/sp500.RData"
    ],
    "chapters": [
        "The building blocks",
        "Analyzing performance"
    ]
}
```