FIN3230 Data Science and Machine Learning for Finance

Workshop 1: Exploratory Data Analysis with R

Sylvia Gottschalk

Outline

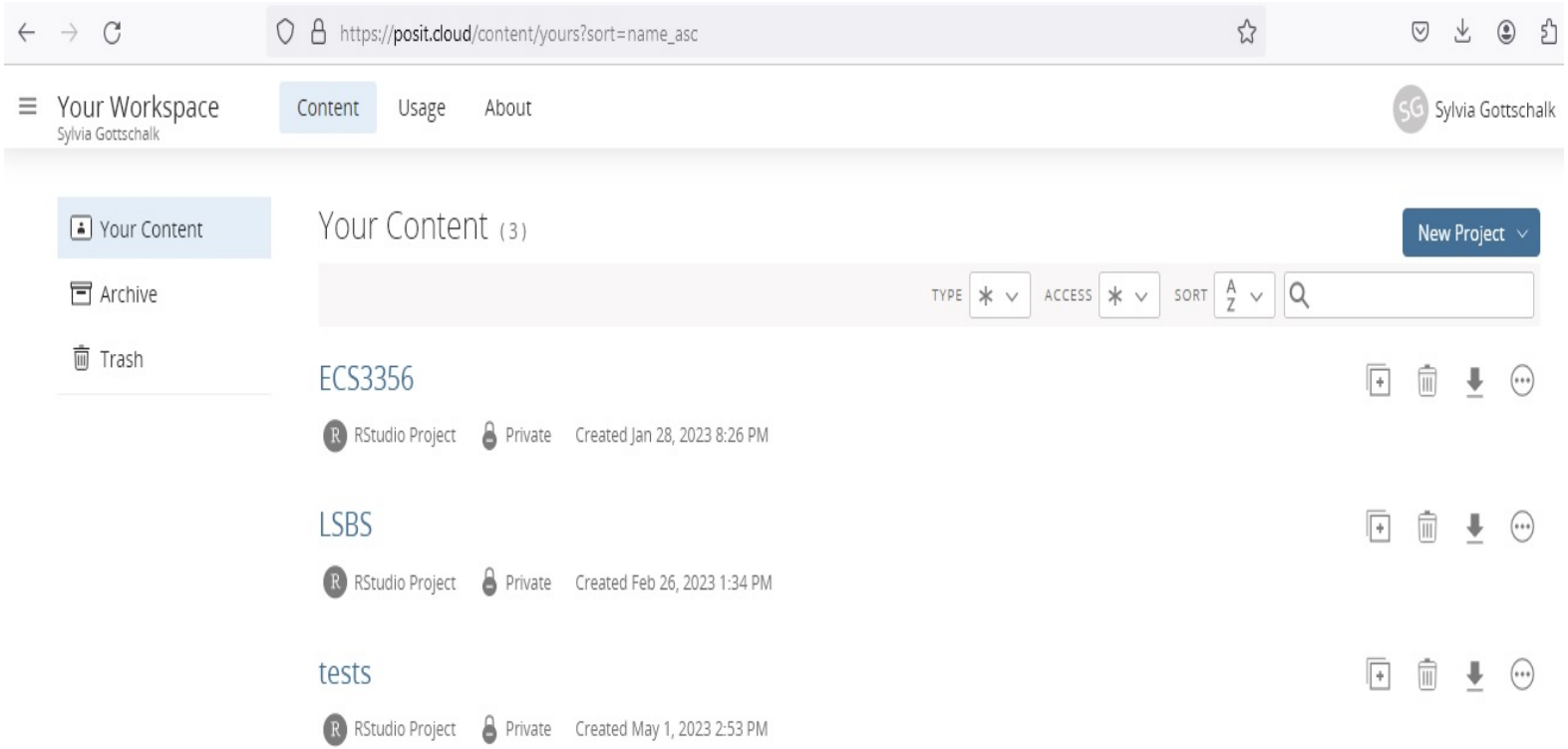
- Posit and R Studio
- Exercises: Football Analytics and Russian Twitter trolls
- Basic R commands

Getting RStudio

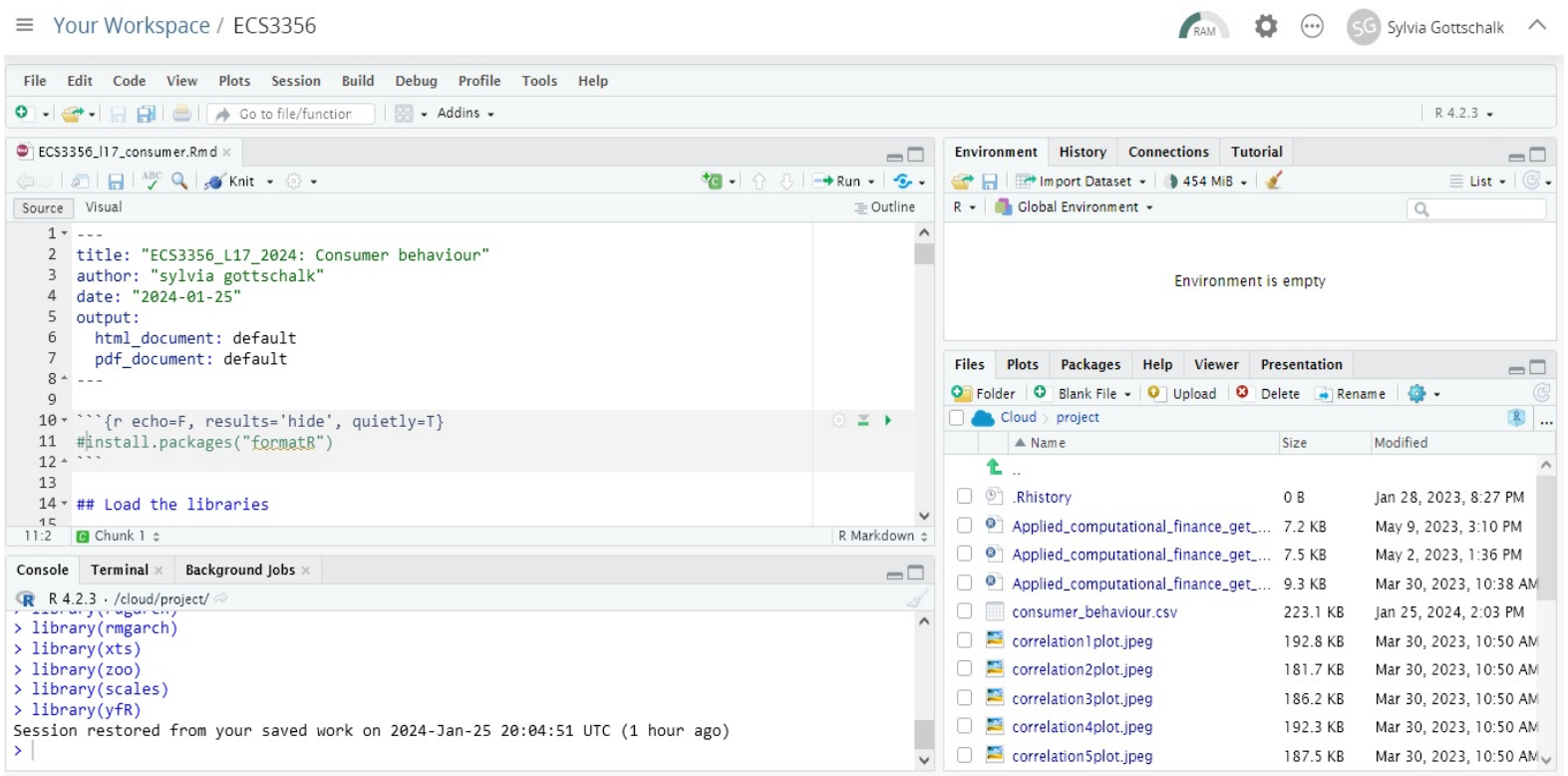
1. If you want to install R on your laptop
   Download the R software from:
   https://www.r-project.org/
   Download RStudio from:
   https://posit.co/
2. If you are using a tablet or are on the go, use R Cloud:
   https://posit.cloud

During the workshops we will use R Cloud, and Google Colab for Python.

Posit Cloud


https://posit.cloud/content/yours?sort=name_asc
Your Workspace
Sylvia Gottschalk
Content
Usage
About
Sylvia Gottschalk
Your Content
Archive
Trash
Your Content (3)
New Project
TYPE
ACCESS
SORT
ECS3356
RStudio Project
Private
Created Jan 28, 2023 8:26 PM
LSBS
RStudio Project
Private
Created Feb 26, 2023 1:34 PM
tests
RStudio Project
Private
Created May 1, 2023 2:53 PM

RStudio in Posit Cloud



packages

Install a package: install.packages("tidyverse")

- Load a package: library(tidyverse)
- Load data from a package: data(iris)

Football infographics

https://fbref.com/en/ is a website that collates statistics about worldwide football leagues:

- leagues
- matches: scores, away/home, ranking
- players: passes, expected passes, position, team

The exercise consists of:

1. uploading data about different leagues
2. create a plot showing passes vs expected passes and a plot showing assists vs expected assists.
3. creating an Rmarkdown file from the R code
4. creating a pdf and an html file from the RMardown code

Exercise: Football Analytics

1. Download from MyLearning Week 1 the files epl_player_summary.csv, laliga_player_summary.csv, ligue1_player_summary.csv, seriea_player_summary.csv; and the code file FIN3230_W1_Football.R
2. Upload all these files into Posit

FIN3230_W1_Football.R creates the plots for EPL. Create the same plots for La Liga, Bundesliga, Serie A, and Ligue 1.

3 million Russian disinformation campaign tweets

A Github directory contains data on nearly 3 million tweets sent from Twitter accounts connected to the Internet Research Agency, the Russian "troll factory" that interfered in the 2016 US election.

The tweets in this database were sent between February 2012 and May 2018, with the vast majority posted from 2015 through 2017.

Source: https://github.com/fivethirtyeight/russian-troll-tweets

Twitter trolls exercise

- Download from MyLearning Week 17 the files: IRAhandle_tweets_1.csv and ECS3356_L17_Russian_Twitter.r
- Upload all these files onto Posit
- Run the R code
- Create an Rmarkdown file from the R code
- Create a pdf and an html file from the RMardown code

Twitter trolls code

The code produces tables answering the questions:

1. What are the most frequent and second-most frequent languages?
2. On average, how many followers did each tweet reach in each region?
3. How many tweets are retweets in each language?
4. How many tweets are not retweets in each language?
5. How often are "Trump" and "Clinton" mentioned in the tweets?

Basics

Calculate: (2 + 5) / 7

- Create a variable:
  var_1← c(1, 4, 5, 7)
- Create another variable:
  var_2 ← c(1, 15, 25, -1)
- Elements of a vector: var_1[1:3]
- Arithmetic with variables:
  var_1 * var_2
  var_1 * 2
- Create a data.frame
- D ← data.frame(var_1, x c(1, 2, 3, 4)) D ← data.frame(var_1, c(1, 2, 3))

Dataframe commands

Variables in dataset D: names(D)

- Index variable within a data.frame:
  D$var1[2:4]
- Create new variable within existing data.frame:
  D$new_var $\leftarrow$ c(100, 200, 300, 400)
- Basic sampling functions:
  sample(1:6, 10)
  rnorm(10, mean = 5, sd = 0.25)

Next sessions

- workshop: data visualisation in finance
- lab: data visualisation in finance