Advanced R

Kálmán Abari

2021-09-27

Contents

Chapter 1

Introduction

Welcome to the second book in R Fundamentals series! This second book takes you through how to do manipulation of tabular data and how to create modern graphics in R. We'll primarily be using capabilities from the set of packages called the tidyverse within the book. The book is aimed at beginners to R who understand the basics (check out the Basic R).

Chapter 2

Warm-up exercise

Loftus, S. C. (2021). Basic Statistics with R: Reaching Decisions with Data. Retrieved from https://books.google.hu/books?id=vTASEAAAQBAJ

2.1 Data structures

2.1.1 Problems

Consider the following set of attributes about the American Film Institute's top-five movies ever from their 2007 list.

1. What code would you use to create a vector named `Movie` with the values `Citizen Kane`, `The Godfather`, `Casablanca`, `Raging Bull`, and `Singing in the Rain`? (Hints: `object <- c()`, Working with character in R)
2. What code would you use to create a vector — giving the year that the movies in Problem 1 were made — named `Year` with the values 1941, 1972, 1942, 1980, and 1952?
3. What code would you use to create a vector — giving the run times in minutes of the movies in Problem 1 — named `RunTime` with the values 119, 177, 102, 129, and 103?
4. What code would you use to find the run times of the movies in hours and save them in a vector called `RunTimeHours`? (Hints: Numeric tranformation)
5. What code would you use to create a data frame named `MovieInfo` containing the vectors created in Problem 1, Problem 2, and Problem 3? (Hints: `data.frame()`)

2.2 Manipulation

2.2.1 Problems

Suppose we have the following data frame named `colleges` (download here):

College	Employees	TopSalary	MedianSalary
William and Mary	2104	425000	56496
Christopher Newport	922	381486	47895
George Mason	4043	536714	63029
James Madison	2833	428400	53080
Longwood	746	328268	52000
Norfolk State	919	295000	49605
Old Dominion	2369	448272	54416
Radford	1273	312080	51000
Mary Washington	721	449865	53045
Virginia	7431	561099	60048
Virginia Commonwealth	5825	503154	55000
Virginia Military Institute	550	364269	44999
Virginia Tech	7303	500000	51656
Virginia State	761	356524	55925

1. What code would you use to select the first, third, tenth, and twelfth entries in the `TopSalary` vector from the `Colleges` data frame? (Hints: Indexing with `[]` operator)
2. What code would you use to select the elements of the `MedianSalary` vector where the `TopSalary` is greater than $400,000? (Hints: `d$MedianSalary[d$TopSalary>400000]`)
3. What code would you use to select the rows of the data frame for colleges with less than or equal to 1000 employees? (Hints: `d[condition, ]`)
4. What code would you use to select a sample of 5 colleges from this data frame (there are 14 rows)? (Hints: `d[sample(x = 1:14, size = 5, replace = F),]`)

Suppose we have the following data frame named Countries (download here):

Nation	Region	Population	PctIncrease	GDPcapita
China	Asia	1409517397	0.4	8582
India	Asia	1339180127	1.1	1852
United States	North America	324459463	0.7	57467
Indonesia	Asia	263991379	1.1	3895
Brazil	South America	209288278	0.8	10309
Pakistan	Asia	197015955	2.0	1629
Nigeria	Africa	190886311	2.6	2640
Bangladesh	Asia	164669751	1.1	1524
Russia	Europe	143989754	0.0	10248
Mexico	North America	129163276	1.3	8562

5. What could would you use to select the rows of the data frame that have GDP per capita less than 10000 and are not in the Asia region?
6. What code would you use to select a sample of three nations from this data frame (There are 10 rows)?
7. What code would you use to select which nations saw a population percent increase greater that 1.5%?

Suppose we have the following data frame named Olympics (download here):

Year	Type	Host	Competitors	Events	Nations	Leader
1992	Summer	Spain	9356	257	169	Unified Team
1992	Winter	France	1801	57	64	Germany
1994	Winter	Norway	1737	61	67	Russia
1996	Summer	United States	10318	271	197	United States
1998	Winter	Japan	2176	68	72	Germany
2000	Summer	Australia	10651	300	199	United States
2002	Winter	United States	2399	78	78	Norway
2004	Summer	Greece	10625	301	201	United States
2006	Winter	Italy	2508	84	80	Germany
2008	Summer	China	10942	302	204	China
2010	Winter	Canada	2566	86	82	Canada
2012	Summer	United Kingdom	10768	302	204	United States
2014	Winter	Russia	2873	98	88	Russia
2016	Summer	Brazil	11238	306	207	United States
2018	Winter	South Korea	2922	102	92	Norway

8. What code would you use to select the rows of the data frame where the host nation was also the medal leader?
9. What code would you use to select the rows of the data frame where the number of competitors per event is greater than 35?
10. What code would you use to select the rows of the data frame where the number of competing nations in the Winter Olympics is at least 80?

2.3 Packages

2.3.1 Problems

1. Install the **Ecdat** package. (Hints: `install.packages()`)
2. Say that we previously installed the **Ecdat** library into R and wanted to call the library to access datasets from it. What code would we use to call the library? (Hints: `library()`)
3. Say that we then wanted to call the dataset `Diamond` from the **Ecdat** library. What code would we use to load this dataset into R? (Hints: `data()`)

2.4 Frequency and numerical exploratory analyses

2.4.1 Problems

Load the `leuk` dataset from the *MASS* library. This dataset is the survival times (`time`), white blood cell count (`wbc`), and the presence of a morphologic characteristic of white blood cells (`ag`).

1. Generate the frequency table for the presence of the morphologic characteristic.
2. Find the median and mean for survival time.
3. Find the range, IQR, variance, and standard deviation for white blood cell count.
4. Find the correlation between white blood cell count and survival time.

Load the `survey` dataset from the *MASS* library. This dataset contains the survey responses of a class of college students.

5. Create the contingency table of whether or not the student smoked (`Smoke`) and the student's exercise regimen (`Exer`). (Hints: `table()`, `DescTools::Desc()`)
6. Find the mean and median of the student's heart rate (`Pulse`). (Hints: `summary()`, `DescTools::Desc()`, `psych::describe()`)
7. Find the range, IQR, variance, and standard deviation for student age (`Age`).
8. Find the correlation between the span of the student's writing hand (`Wr.Hnd`) and nonwriting hand (`NW.Hnd`). (Hints: `cor()`, `DescTools::Desc()`)

Load the `Housing` dataset from the *Ecdat* library. This dataset looks at the variables that affect the sales price of houses.

9. Create the contingency table of whether or not the house has a recreation room (`recroom`) and whether or not the house had a full basement (`fullbase`).
10. Find the mean and median of the house's lot size (`lotsize`).
11. Find the range, IQR, variance, and standard deviation for the sales price (`price`).
12. Find the correlation between the sales price of the house (`price`) and the number of bedrooms (`bedrooms`).

2.5 Graphical exploratory analyses

Load the `Star` dataset from the *Ecdat* library. This dataset looks at the affect on class sizes on student learning.

1. Generate the scatterplot of the student's math score `tmathssk` and reading score `treadssk`. (Hints: `plot()`, `ggplot() + geom_point()`)
2. Generate the histogram of the years of teaching experience `totexpk`. (Hints: `hist()`, `ggplot() + geom_histogram()`)
3. Create a new variable in the `Star` dataset called `totalscore` that is the sum of the student's math score `tmathssk` and reading score `treadssk`. (Hints: tranformation)
4. Generate a boxplot of the student's total score `totalscore` split out by the class size type `classk`. (Hints: `boxplot()`, `ggplot() + geom_boxplot()`)

Load the `survey` dataset from the *MASS* library. This dataset contains the survey responses of a class of college students.

5. Generate the scatterplot of the student's height `Height` and writing hand span `Wr.Hnd`.
6. Generate the histogram of student age `Age`.
7. Generate a boxplot of the student's heart rate `Pulse` split out by the student's exercise regimen `Exer`.

Chapter 3

RMarkdown

RMarkdown is a framework from RStudio for easily combining your code, data, text and interactive charts into both reports and slide decks. RMarkdown is based on Markdown.

3.1 Markdown

Markdown is a markup language. It is an extremely simple markup language, so it is very popular on the Web and in other application. Markdown is used to format text on GitHub, Reddit, Stack Exchange, and Trello, and in RMarkdown.

Markdown was created by John Gruber and Aaron Swartz in 2004. Markup was designed that a human reader could easily parse the content.

- Markdown Cheat Sheet A quick reference to the Markdown syntax.
- Basic Syntax The Markdown elements outlined in John Gruber's design document.
- Extended Syntax Advanced features that build on the basic Markdown syntax.

3.2 RMarkdown

R Markdown understands Pandoc's Markdown, a version of Markdown with more features. This Pandoc guide provides and extensive resource for formatting options.

Happy collaboration with Rmd to docx

Using the flextable R package

3.3 Additional Resources

- R Markdown Cookbook A comprehensive free online book that contains almost everything you need to know about RMarkdown.
- RMarkdown for Scientists
- RStudio Articles for RMarkdown RStudio has published a few in-depth how to articles about using RMarkdown.
- R for Data Science Hadley Wickham provides a great overview of authoring with RMarkdown.
- R Markdown: The Definitive Guide It contains a large number of technical details, it may serve you better as a reference book than a textbook.
- Online lesson from RStudio

Markdown is file format.

https://github.com/citation-style-language/styles

Chapter 4

Advanced data manipulation

This chapter focuses exclusively on advanced data manipulation. I therefore assume a basic level of comfort with data manipulation.

Chapter 5

Modern graphics

Chapter 6

Tidyverse R

Chapter 7

Bioconductor

Chapter 8

RNA-Seq (an example)