

Exploring data using R

Kamarul Imran Musa, Wan Nor Arifin

2017-10-27

Contents

1	Introduction to R	5
1.1	R and RStudio	6
1.2	Working with packages	6
1.3	Working directory	6
1.4	Data management	6
1.5	More about data management	6
1.6	Summary	6
2	Textual	7
2.1	Basic descriptive statistics	8
2.2	Data sets in package ‘datasets’:	8
2.3	AirPassengers Monthly Airline Passenger Numbers 1949-1960	8
2.4	BJsales Sales Data with Leading Indicator	8
2.5	BJsales.lead (BJsales) Sales Data with Leading Indicator	8
2.6	BOD Biochemical Oxygen Demand	8
2.7	CO2 Carbon Dioxide Uptake in Grass Plants	8
2.8	8
2.9	More on descriptive statistics	8
2.10	Summary	8
3	Graphical	9
3.1	Preliminaries	9
3.2	One variable	9
3.3	Two variables	9
3.4	Summary	9
4	Reporting results	11
5	Summary	13
5.1	What we have learned so far	13
5.2	Some important packages	13
6	References	15
7	Preparation	17
7.1	Set the working directory	18
7.2	Read data	18
7.3	Training data	18
7.4	Questions to ask when graphing or plotting data	18
7.5	One variable: A categorical variable	18
7.6	One variable: A numerical variable	18
7.7	Two variables : A numerical with another numerical variable	18
7.8	One variable: A numerical variable	18

7.9	One variable: A categorical variable	18
7.10	Two variables: A numerical and a categorical variable	18
7.11	Two or more categorical variables	18
7.12	Data and package	18
7.13	View data	18
7.14	One numerical variable - Histogram	18
7.15	One numerical and one categorical variable - histogram	18
7.16	Barchart	18
7.17	Scatterplot	18
7.18	Reading dataset from STATA file (.dta)	18
7.19	Graphing or Plotting data	18
7.20	Saving plot	18
8	Data transformation (data munging or data wrangling)	19
8.1	Definition of data wrangling	20
8.2	Package: dplyr	20
8.3	Data wrangling using dplyr	20
8.4	Preparation and data	20
8.5	Function: select and mutate	20
8.6	arrange and filter	20
8.7	group_by	20
8.8	Summary	20
8.9	Categorical variables	20
8.10	forcats	20

Chapter 1

Introduction to R

1.1 R and RStudio

1.1.1 Installation of R

1.1.2 Starting R

1.1.3 Installation of RStudio

1.1.3.1 Starting RStudio

1.1.3.2 Why RStudio?

1.1.3.3 RStudio interface

1.2 Working with packages

1.2.1 About packages

1.2.2 Package installation

1.2.3 Loading packages

1.3 Working directory

1.3.1 Setting a working directory

1.4 Data management

1.4.1 Reading data set

1.4.2 Viewing data set

1.4.3 Exporting data set from R

1.5 More about data management

1.5.1 Subsetting

Chapter 2

Textual

2.1 Basic descriptive statistics

2.2 Data sets in package ‘datasets’:

2.3 AirPassengers Monthly Airline Passenger Numbers 1949-1960

2.4 BJsales Sales Data with Leading Indicator

2.5 BJsales.lead (BJsales) Sales Data with Leading Indicator

2.6 BOD Biochemical Oxygen Demand

2.7 CO2 Carbon Dioxide Uptake in Grass Plants

2.8 ...

2.8.1 Describing a numerical variable

2.8.2 Describing a categorical variable

2.9 More on descriptive statistics

2.9.1 Describing numerical variables

2.9.2 Describing categorical variables

2.9.3 Describing the variables together

2.9.3.1 By groups

2.9.3.2 Simple cross-tabulation

2.10 Summary

Chapter 3

Graphical

3.1 Preliminaries

3.1.1 Reading dataset

3.1.2 Describing data

3.2 One variable

3.2.1 One variable: A categorical or factor variable

3.2.2 One variable: A numerical variable

3.3 Two variables

3.3.1 Two variables : A numerical with another numerical variable

3.3.2 Two variables : A categorical variable with a categorical variable

3.4 Summary

Chapter 4

Reporting results

Chapter 5

Summary

5.1 What we have learned so far

5.2 Some important packages

Chapter 6

References

Chapter 7

Preparation

7.1 Set the working directory

7.2 Read data

7.3 Training data

7.4 Questions to ask when graphing or plotting data

7.5 One variable: A categorical variable

7.6 One variable: A numerical variable

7.6.1 Histogram

7.6.2 Kernel density plot

7.7 Two variables : A numerical with another numerical variable

7.7.1 Scatterplot

7.8 One variable: A numerical variable

7.8.1 Histogram

7.8.2 Density curve

7.8.3 Combining the histogram and the density curve

7.9 One variable: A categorical variable

7.10 Two variables: A numerical and a categorical variable

7.10.1 Overlaying histograms

Chapter 8

Data transformation (data munging or data wrangling)

8.1 Definition of data wrangling

8.2 Package: dplyr

8.3 Data wrangling using dplyr

8.4 Preparation and data

8.4.1 Working directory and data format

8.4.2 Training data

8.5 Function: select and mutate

8.5.1 Select

8.5.2 Mutate

8.6 arrange and filter

8.6.1 arrange

8.6.2 filter

8.7 group_by

8.7.1 Summarize data - summarize

8.8 Summary

8.9 Categorical variables