

Analysing Social Science Data Using R

Class 1: Introduction

Michał Bojanowski¹ mbojan@icm.edu.pl
Zbigniew Karpiński² zkarpinski@ifispan.waw.pl

¹ICM University of Warsaw

²IFiS PAN

February 13, 2012
SNS, Warsaw

Objectives

Main objective:

Educate the participants to use R for a typical data analysis in sociology

Including:

- Loading data
- Basic data manipulation: recoding, aggregating, variable transformation
- Descriptive statistics
- Data visualization
- Regression modeling (linear and logistic)

Work flow

Every meeting will consist of:

- Presentation
- Tutorial exercises
- Practical exercises to be completed in class

You will receive short exercises to complete at home.

Materials

Course materials will consist of:

- Slides
- Data files
- R scripts

All slides, data files, and scripts will be progressively available on-line on course webpage:

<https://sites.google.com/site/r4sns2012>

The course is rather self-contained: no extra obligatory reading materials etc. are planned. We will provide pointers to literature should anybody feel a need to refresh his/her statistical knowledge.

Grading

Final grade will be computed using:

- Attendance (20%)
- Homeworks (40%)
- Final test (40%)

Schedule

Updated syllabus is available on

<https://sites.google.com/site/r4sns2012>

Please note:

- Dates are provisional. It is likely that we will skip 12.03 and 19.03 and resume on 26.03.

What is R?

Program for data analysis For many people **R** is a program of choice for statistical analysis, visualization.

Programming language Data analysis with functions and scripting. Interactive programming language.

Environment for statistical analysis Access to standard models and cutting edge methods.

Open Source project Open source code. Free. Over 15 years of peer review. Integration with other tools/systems.

Community 20 people in R Core. Approximately 2 mln users worldwide: forums, mailing lists, blogs.

CRAN and addon packages

- R's functionality is contained in packages.
 - Base distribution contains 27 packages.
 - Developers and users created the next 3604 packages (as on 2012-02-13) available on the Internet on the CRAN (Comprehensive R Archive Network) servers.
- Other things on CRANs
 - Official documentation
 - User-contributed documentation
 - R Journal

CRAN server in Wrocław <http://r.meteo.uni.wroc.pl/>

RStudio

<http://www.rstudio.org>

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading data, summarizing it, and creating a faceted scatter plot.


```

1 library(ggplot2)
2
3 view(diamonds)
4 summary(diamonds)
5
6 summary(diamonds$price)
7 aveSize <- round(mean(diamonds$carat), 4)
8 clarity <- levels(diamonds$clarity)
9
10 p <- qplot(carat, price,
11            data=diamonds, color=clarity,
12            xlab="carat", ylab="Price",
13            main="Diamond Pricing")
14
      
```
- Console:** Shows the execution output, including summary statistics for 'x', 'y', and 'z' (representing carat, price, and clarity), and the execution of the R code above.


```

14:1 [1] (Top Level)
R Script

Console
x      y      z
Min.  : 0.000   Min.  : 0.000   Min.  : 0.000
1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
Median : 5.700   Median : 5.710   Median : 3.530
Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
Max.   :10.740   Max.   :18.900   Max.   :31.800

> summary(diamonds$price)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 326   950   2401   3933   5324  18820

> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+           data=diamonds, color=clarity,
+           xlab="carat", ylab="Price",
+           main="Diamond Pricing")
>
> format.plot(plot=p, size=23)
>
      
```
- Environment/History:** Shows the 'diamonds' data frame with 53940 observations and 10 variables. It also lists the 'aveSize' variable and the 'p' plot object.
- Plots:** A scatter plot titled "Diamond Pricing" is displayed. The x-axis is labeled "Carat" and the y-axis is labeled "Price". The plot shows a positive correlation between carat weight and price, with points colored by clarity (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF).