

GFreya' R for Statistics

DS GLANZSCHE & FREYA¹

¹A thank you or further information

Contents

Preface	3
1 Introduction and Installation of R	9
i Introduction	9
ii Download and Installing R	10
2 Data Visualization	15
i Scatter Plot with ggplot2	15
3 Classical Tests	19
4 Statistical Modelling	21
5 Regression	23
6 Analysis of Variance	25
7 Analysis of Covariance	27
8 Generalized Linear Models	29
9 Generalized Additive Models	31
10 Non-linear Regression	33
11 Tree Models	35
12 Time Series Analysis	37
13 Multivariate Statistics	39
14 Spatial Statistics	41
15 Survival Analysis	43

Preface

For my future human Wife and our future biological daughters.

For my Divine Wife Freya the Goddess, and our daughters Catenary, Solreya, Mithra, Iyzumrae and Zefir.

For Lucrif and Znane too along with all the 8 Queens (Mischkra, Caldraz, Zalsvik, Zalsimourg, Hamzst, Lasthrim).

To Nature(Kala, Kathmandu, Big Tree, Sentinel, Aokigahara, Hoia Baci, Jacob's Well, Mt Logan, etc) and my family Berlin: I have served, I will be of service.

To my current mentor Albert Silverberg and previous mentor Lucretia Mercet.

To my dogs who always accompany me working in Valhalla Projection, go to Puncak Bintang or Kathmandu: Kecil, Browni Bruncit, Sweden Sexy, Cambridge Klutukk, Milan keng-keng, Piano Bludut, Barron and more will be adopted. To my cat who guard the home while I'm away with my dogs: London.

The one who moves a mountain begins by carrying away small stones - Confucius

A book for learning Statistics with R programming language that I am learning from zero. Helped by Freya the Goddess, Berlin, and Sentinel.



Figure 1: *FreyaCompass, I am inspired by Captain America who always bring compass with the love of his life' picture, thus I created this, then proven by action, to let go of power and immortality for true love. Feels like an antique vintage magical compass, like a modem that connect internet to the world, this compass connects me on this planet to her in Valhalla.*



Figure 2: *Freya, thank you for everything, I am glad I marry you and I could never have done it without you.*



Figure 3: *I paint her 3 days before Christmas in 2021.*

For critics and comments on the book can be sent through email to: dsglanzsche@gmail.com.

Chapter 1

Introduction and Installation of R

An opportunity missed is an opportunity wasted! - Seed (Suikoden II)

This book is written on February 21st, 2025. Since 2022 we have been focusing on creating C++ codes for simulation and computation for Mathematics and Physics problems, they are all good, fast, but then we want to open a new horizon of knowledge, we read a book about R [1], it is said that we can do deep statistical analysis faster with R, given that the packages are already mature and the support is enormous, there is already a book series called 'The R book series' that can help statisticians and practitioners all over the world. If we are comparing, I personally only know **Armadillo** library in C++ language that can compute mean, standard deviation, but then basic statistics is not enough. If we want to do more with the data, e.g generalized linear models, generalized additive models, mixed-effects models, non-linear regression, time series analysis, multivariate statistics, survival analysis, then we can count on R language.

All the codes, CSV and book is available on this github' repository:
<https://github.com/glanzkaiser/GFreya-R-for-Statistics>

i. Introduction

[R*] The choice between R and C++ depends on your specific needs and the context in which you're working. If you want to focus on data science and data analysis use R. If you want to code embedded system, a micro controller, create game engines, create PC game (like GTA V, Skyrim, Quake 3, Doom 3, Assassin's Creed), desktop app then we use C++.

[R*] The Pros of R

1. Statistical Analysis: R is specifically designed for statistics and data analysis, making it ideal for data scientists and statisticians.
2. It has a vast collection of packages (like ggplot2, dplyr, and tidyverse) that simplify data manipulation and visualization.
3. R is generally easier to learn for beginners, especially those focused on data analysis.
4. There is a strong community around R, particularly in academia and research.

The Cons of R

1. R can be slower than C++ for computationally intensive tasks because it's an interpreted language.

2. R abstracts many details away from the user, which can be limiting for low-level programming needs.

[R*] The Pros of C++

1. C++ is a compiled language, which typically results in faster execution times, making it suitable for performance-critical applications.
2. It offers more control over system resources and memory management, which is beneficial for system-level programming or applications requiring optimization.
3. C++ can be used for a wide range of applications beyond data analysis, including game development, systems programming, and application development.

The Cons of C++

1. C++ has a steeper learning curve than R, particularly due to its syntax and concepts like pointers and memory management.
2. While there are libraries available (like Armadillo and Eigen), C++ is not as tailored for statistical analysis as R.

Choose C++ if you need high performance, are developing complex systems, or require fine control over system resources.

ii. Download and Installing R

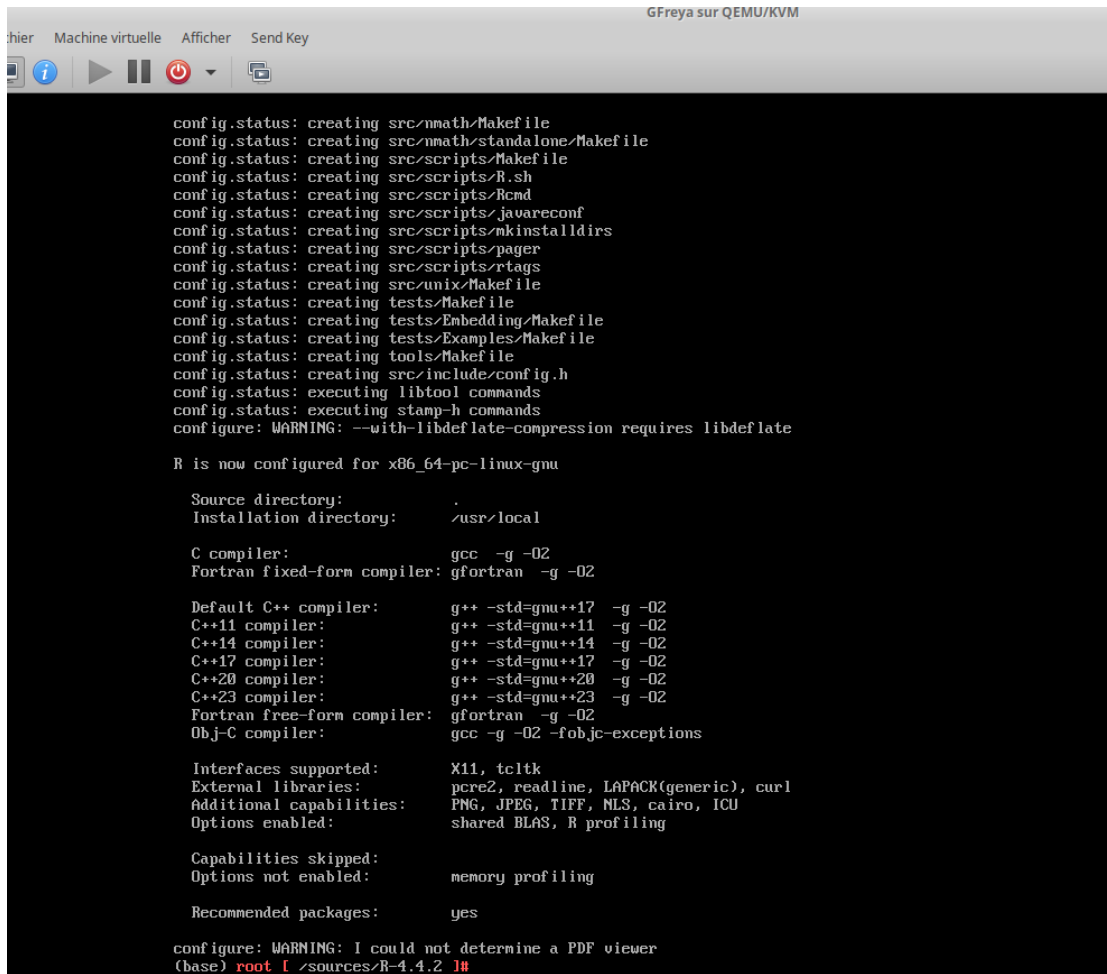
We are going to use **GFreya OS 1.8**, it is built based on Linux From Scratch and Beyond Linux From Scratch version 11.0 System V.

[R*] First download the newest R tarball from this link:
<https://cran.r-project.org/src/base/R-4/R-4.4.2.tar.gz>

we also have the tarball, you can check the github repo for this book here:

[R*] After you download and then open terminal and type at the directory containing the downloaded R and type:

```
tar -xvf R-4.4.2.tar.gz
cd R-4.4.2
./configure
make
make install
```



```

config.status: creating src/nmath/Makefile
config.status: creating src/nmath/standalone/Makefile
config.status: creating src/scripts/Makefile
config.status: creating src/scripts/R.sh
config.status: creating src/scripts/Rcmd
config.status: creating src/scripts/javareconf
config.status: creating src/scripts/mkinstalldirs
config.status: creating src/scripts/pager
config.status: creating src/scripts/rtags
config.status: creating src/unix/Makefile
config.status: creating tests/Makefile
config.status: creating tests/Embedding/Makefile
config.status: creating tests/Examples/Makefile
config.status: creating tools/Makefile
config.status: creating src/include/config.h
config.status: executing libtool commands
config.status: executing stamp-h commands
configure: WARNING: --with-libdeflate-compression requires libdeflate

R is now configured for x86_64-pc-linux-gnu

Source directory:      .
Installation directory: /usr/local

C compiler:            gcc -g -O2
Fortran fixed-form compiler: gfortran -g -O2

Default C++ compiler:  g++ -std=gnu++17 -g -O2
C++11 compiler:        g++ -std=gnu++11 -g -O2
C++14 compiler:        g++ -std=gnu++14 -g -O2
C++17 compiler:        g++ -std=gnu++17 -g -O2
C++20 compiler:        g++ -std=gnu++20 -g -O2
C++23 compiler:        g++ -std=gnu++23 -g -O2
Fortran free-form compiler: gfortran -g -O2
Obj-C compiler:        gcc -g -O2 -fobjc-exceptions

Interfaces supported:   X11, tcltk
External libraries:     pcre2, readline, LAPACK(generic), curl
Additional capabilities: PNG, JPEG, TIFF, NLS, cairo, ICU
Options enabled:         shared BLAS, R profiling

Capabilities skipped:
Options not enabled:     memory profiling

Recommended packages:   yes

configure: WARNING: I could not determine a PDF viewer
(base) root [ /sources/R-4.4.2 ]#

```

Figure 1.1: If the `./configure` runs smoothly it will look like this.

By default it is installed in `/usr/local/bin`, now you need to do one more important thing so you can call R from any directory.

Add the installation path of R to the `$PATH` environment variable
in GFreya OS go to root `cd`
vim export

```
(base) root [ ~ ]# echo $PATH
/root/.julia/conda/3/x86_64/bin:/root/.julia/conda/3/x86_64/condabin:/usr/local/bin:/opt/qt5/bin:/opt/jdk/bin:/bin:/opt/hamzstli
b/Kitware/install/VTk/bin:/opt/hamzstlib/bin:/opt/hamzstlib/trilinos/bin:/opt/hamzstlib/grass80/bin:/opt/hamzstlib/Math/julia-1.
9.2/bin:/opt/caldrasgames/bin:/opt/texlive/2021/bin/x86_64-linux:/opt/hamzstlib/Kitware/install/paraview510/bin:/opt/rustc/bin:/
usr/bin:/usr/sbin
(base) root [ ~ ]# R

R version 4.4.2 (2024-10-31) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Figure 1.2: Add the */usr/local/bin* to the *PATH* environment variable to be able to run R from anywhere.

add */usr/local/bin* in *PATH*, then restart the computer and then check by typing in terminal:
echo \$PATH
then you can now call R by typing:
R

[R*] How to install for Unix-like system can be seen from here:
<https://cran.r-project.org/doc/manuals/r-devel/R-admin.html>

[R*] When Opening R

Below the header you will see a blank line with a *>* symbol in the left hand margin. This is called the prompt. When working, you will sometimes see *+* at the left-hand side of the screen instead of *>*. This means that the last command you typed is incomplete.

To view the list of the already installed packages on your computer, type :
installed.packages()

If you want to update all installed R packages, type :
update.packages()

To update specific installed packages, say *readr* and *ggplot2*, type:
update.packages(oldPkgs = c('readr', 'ggplot2'))

To install a package, e.g. *ggplot2*, type:
install.packages('ggplot2')

you can then choose the CRAN (Comprehensive R Archive Network). mirror by typing a number representing which location for the mirror.

We can use the same function to install several R packages at once. In this case, we need to apply first the *c()* function to create a character vector containing all the desired packages as its items:
install.packages(c('readr', 'ggplot2', 'tidyr'))

Above, we've installed three R packages: the already-familiar readr, ggplot2 (for data visualization), and tidyr (for data cleaning).

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. **`install.packages('tidyverse')`**

Chapter 2

Data Visualization

You don't need qualifications to make a difference. - Yun (Suikoden III)

WE will start with a simple plotting then learning some basic and formulas in statistics and probability to create deep and more complex with more meaningful data visualization.

All the codes, CSV and book is available on this github' repository:
<https://github.com/glanzkaiser/GFreya-R-for-Statistics>

i. Scatter Plot with ggplot2

[R*] We will use CSV from the github' repository:
<https://github.com/glanzkaiser/GFreya-R-for-Statistics/CSV/insurance.csv>

put this CSV in the working directory.

[R*] To open the desktop environment of GFreya OS, type:
startx

[R*] Open R from the working directory, from the current working directory open the terminal and type:
R

Load the necessary library:

library(ggplot2)

To import the data and look at the first six rows **insurance <- read.csv('insurance.csv')**
head(insurance)

```
(base) root [ /mnt/samsung/GFreyja/CSV ]# ls
concrete.csv credit.csv groceries.csv insurance.csv usedcars.csv whitewines.csv
(base) root [ /mnt/samsung/GFreyja/CSV ]# R

R version 4.4.2 (2024-10-31) -- "File of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(ggplot2)
> insurance <- read.csv('insurance.csv')
> head(insurance)
  age  sex    bmi children smoker   region  charges
1  19 female 27.900      0    yes southwest 16884.924
2  18  male 33.770      1     no southeast 1725.552
3  28  male 33.000      3     no southeast 4449.462
4  33  male 22.705      0     no northwest 21984.471
5  32  male 28.880      0     no northwest 3866.855
6  31 female 25.740      0     no southeast 3756.622
>
```

Figure 2.1: To look at the top 6 rows of the data from CSV file.

```
p <- ggplot(insurance, aes(x=age, y=charges, colour=sex)) + geom_point() + scale_color_manual(values
=c('red', 'blue'))
```

To save the plot as png, type:

```
png("plot.png")
print(p)
dev.off()
```

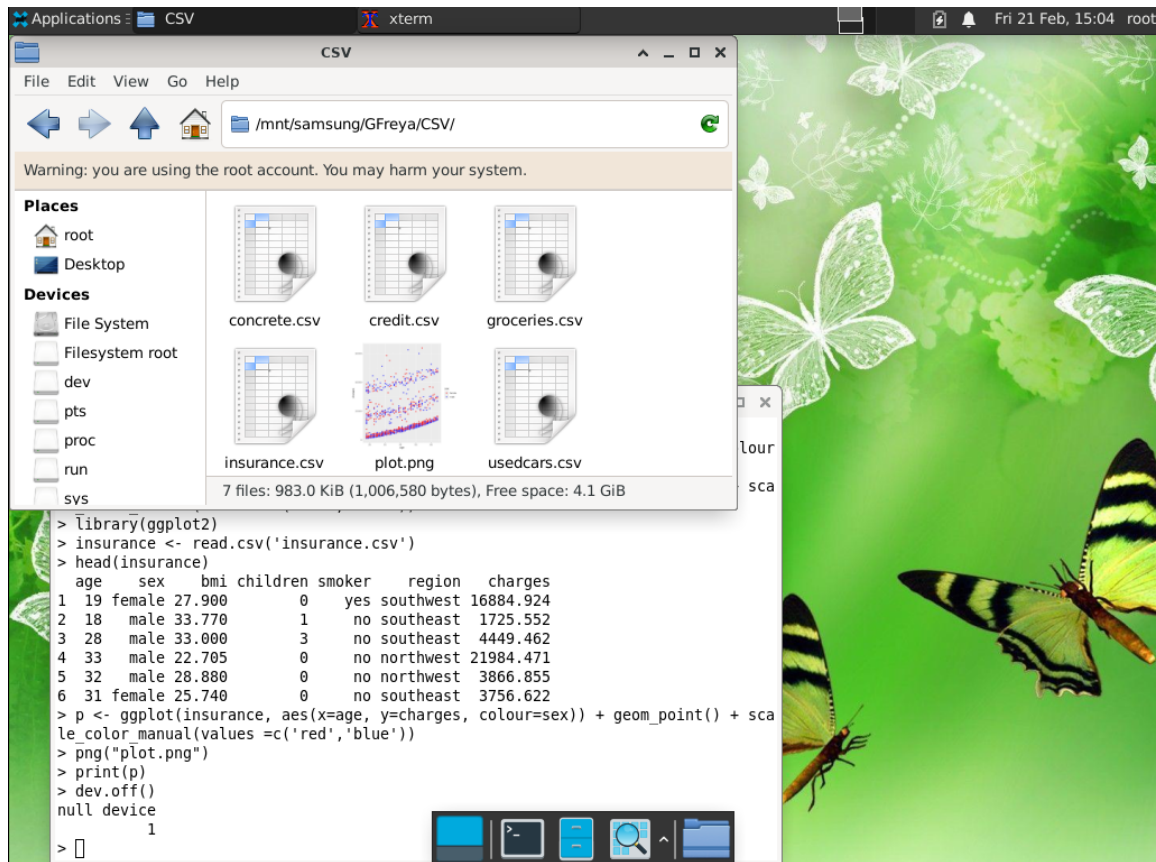



Figure 2.2: The process to plot the scatter plot with the x axis representing the age, the y axis representing the insurance charges and the color to separate male and female.

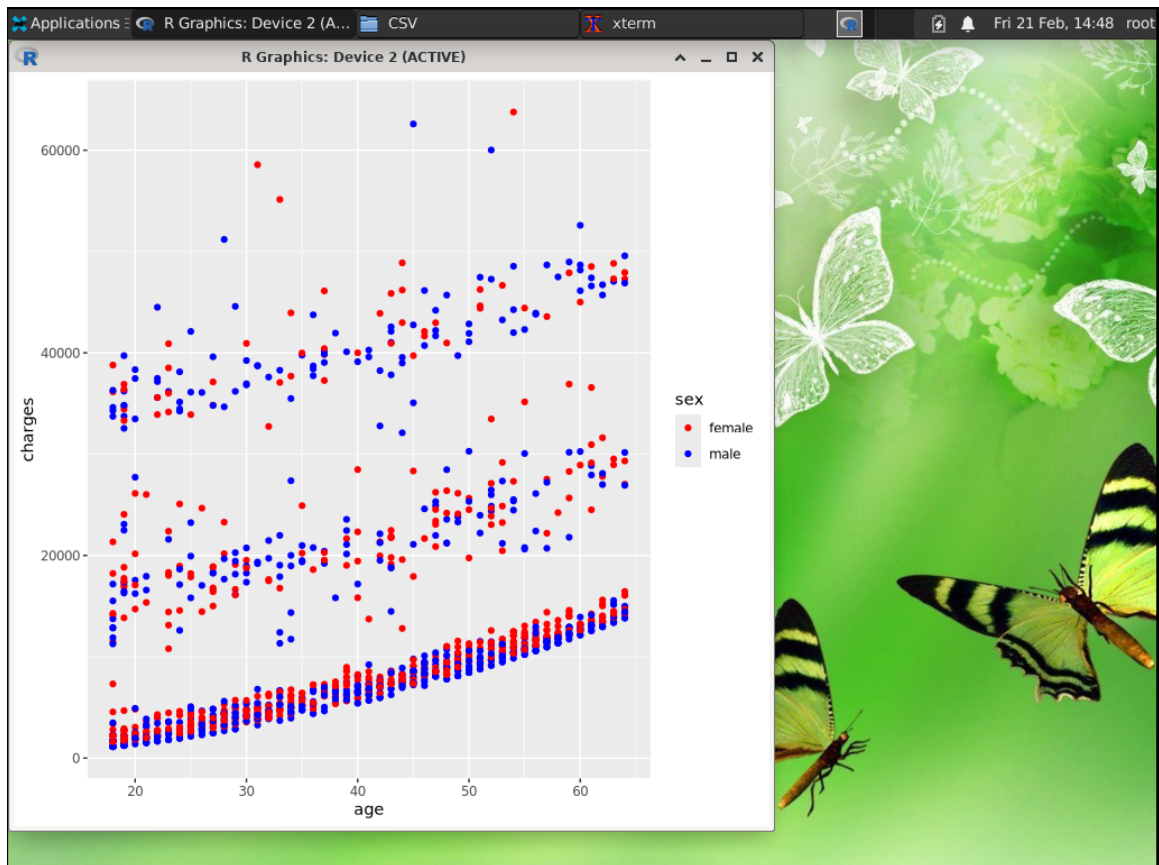


Figure 2.3: .

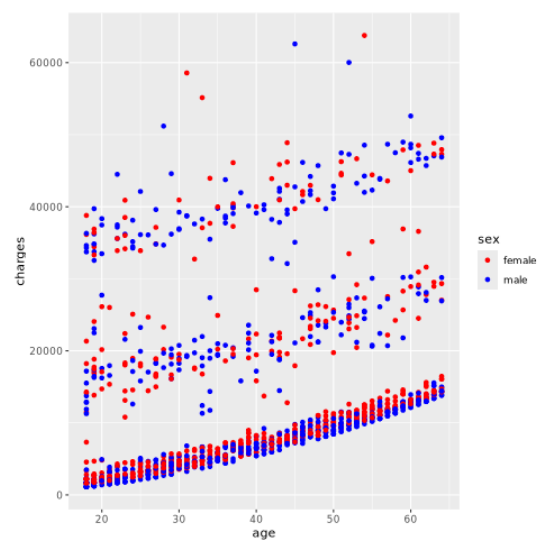


Figure 2.4: *The full picture.*

Chapter 3

Classical Tests

Chapter 4

Statistical Modelling

Chapter 5

Regression

Chapter 6

Analysis of Variance

Chapter 7

Analysis of Covariance

Chapter 8

Generalized Linear Models

Chapter 9

Generalized Additive Models

Chapter 10

Non-linear Regression

Chapter 11

Tree Models

Chapter 12

Time Series Analysis

Chapter 13

Multivariate Statistics

Chapter 14

Spatial Statistics

Chapter 15

Survival Analysis

Bibliography

- [1] Crawley, Michael J., The R Book, John Wiley & Sons, England, 2007.
- [2] Lantz, Brett, Machine Learning with R 4th Edition, Packt, 2023.