UNIVERSITÄT BONN

Nutrition & Planetary Health Research Group

Introduction to R

Zentrum für Entwicklungsforschung (ZEF)

Author: Mahir Bhatt

2025/2026

Difference between R & RStudio

(a) R is a programming language

(b) RStudio is programming environment

Working with RStudio

- Always create a project for specifc anaylsis.
- Set the project directory to the location where your data is stored.
- Do not open R files from multiple directories at the same time (**RStudio could crash!**)

Using Rmarkdown vs RScript

There is nothing wrong with using either of them, however .Rmd has bit of an advantage.

- .Rmd hlps you make your code more dynamic by combining text with code and output.
- You can knit .Rmd files in to html, pdf or word format
- .Rmd ensures that your analysis is reproducible.
- You can include interactive elements like shny apps & widgets with .Rmd
- You can easily make your code accessible on Github with .Rmd

Packages in R

- There are pre-installed packagese in R which can perform basic statistical functions.
- To perfrom specific tasks, you would have to install package the contains the given function
- R has a large repository of packages (**Constantly updated!**)

Packages in R

- There are pre-installed packages in R which can perform basic statistical functions.
- To perfrom specific tasks, you would have to install package the contains the given function
- R has a large repository of packages (**Constantly updated!**)
- Example of a mathematical operation

$$\binom{n}{k} = \frac{n!}{n!(n-k)!}$$

Lists, Vectors & Data Frames

Lists

It is the most basic type of data storage unit in R. It generates a list of inputs stored which can be heterogenous (It is like storing different attributes of a person in a single variable).

Vectors

Vector is a simple one-dimensional array of data. Elements of all vectors are homogeneous (all numeric, all characters). It is more like storing the same attribute of a different person.

Data Frames

They are the building block of data analysis in R. It stores data in tabular format (two-dimensions). Each column of the table acts as an individual vector.

Loading a data frame in R

- R is capable of loading any type of data frame (SAS, STATA, SPSS, csv, excel, *QGIS*, *ArcGIS*)
- Unless you are loading a dataset from another which has an extension of another statistical software, always use **csv!**

Visualizing data in R

R has inbuilt capabilites to graphically present your data. Below are examplesof saome packages that can be used.

- ggplot2 $\rightarrow$ Combine multiple data sources to create graphical summary
- plotly $\rightarrow$ Create interactive plots in 2D and 3D
- Leaflet $\rightarrow$ Build interactive maps
- tmaps $\rightarrow$ Use shapefiles to prodcue descriptve maps
- highcharter $\rightarrow$ Used for dynamic charting
- Lattice $\rightarrow$ Powerful tool for visualizing multivariate data

We will first look at the use of ggplot2 and slowly advance skills with other packages.

Types of ggplots

```
# geom_boxplot()
ggplot(df1, aes(x=, y
    =))+
geom_boxplot()+
theme_bw()+
labs(x ="", y="",
    title="")

# geom_smooth()
ggplot(df1, aes(x=, y
    =))+geom_smooth()+
theme_bw()+labs(x =""
    , y="", title="")
```

```
# geom_line()
ggplot(df1, aes(x=, y
    =))+
geom_line()+
theme_bw()+
labs(x ="", y="",
    title="")

# geom_point()
ggplot(df1, aes(x=, y
    =))+
geom_boxplot()+
theme_bw()+
labs(x ="", y="",
    title="")
```

Types of ggplots

- ggplot allows you to create wide range of plots to graphically show your data, you do make more plots your skills with atuomatically improve.

Some examples

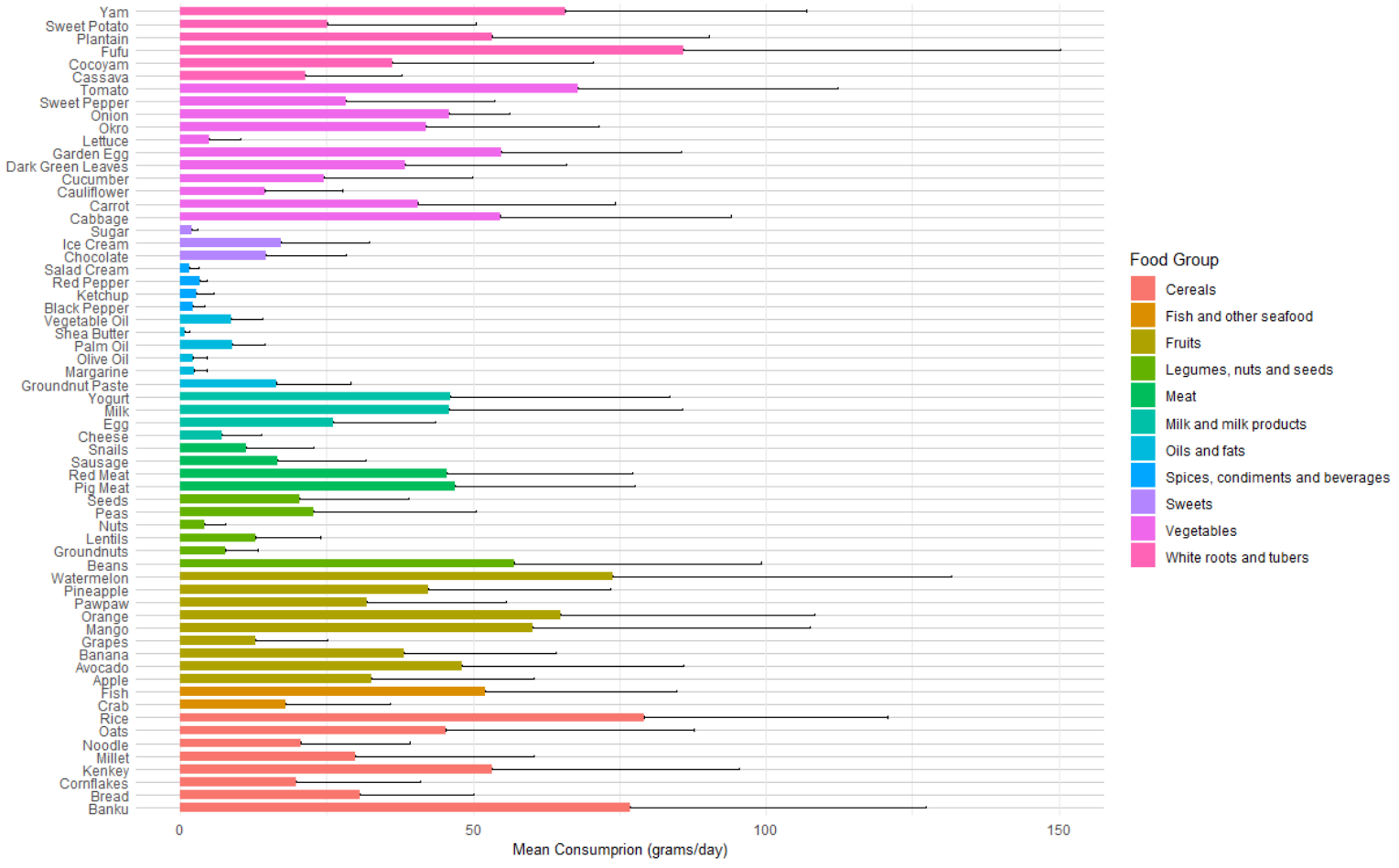
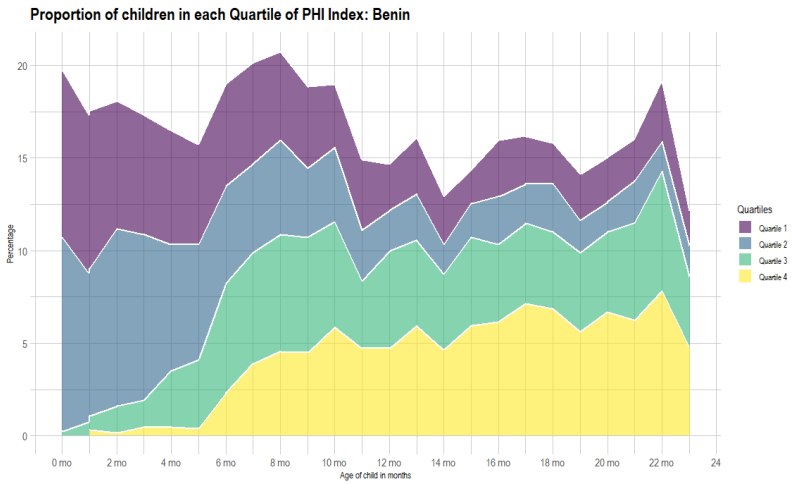
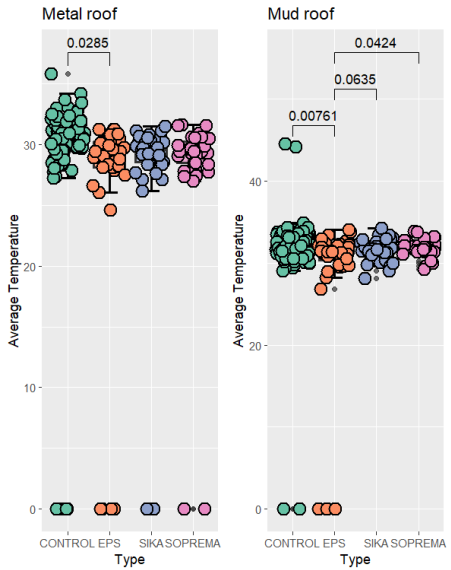


Figure: Mean and standard deviation of all the food items consumed by children in Agogo

More examples



(a) Proportion of children in each quartile of Planetary Health Diet Index by age



(b) Tukey plot of show effect of cool roof paints on indoor temperature in Mud and Metal roof

Linear models in R

Begining with bivariate data exploration

```
#make a plot looking
    at two variables
plot(df1$height, df1$
    weight)

#Cross tabulation can be
    used for categorical
    data
table(df1$rdt, df1$sex,
     useNA="always")
```

```
# variance, correlation
    and covariance
     var(df1$height
     cor(df1age,df1$
         height)
     var(df1$height,
         df1$weight)
```

Fitting a linear model

- Let us consider the simple case of linear relationship between a response variable and predictor variable $y \approx \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots + \delta$ Where
- β_0 is intercept
- $\beta_1, \ldots\ldots\ldots, \beta_n$ are coefficients
- δ is error term

```
1  #simple linear model in R
2  model1<-lm(hemoglobin ~ weight, data=df1)
3  summary(model)
4
5  model2<-lm(hemoglobin~weight+rdt,data=df1)
6  summary(model2)
```

Tables

I think this is enough for today. Thank you for your attention ☺