

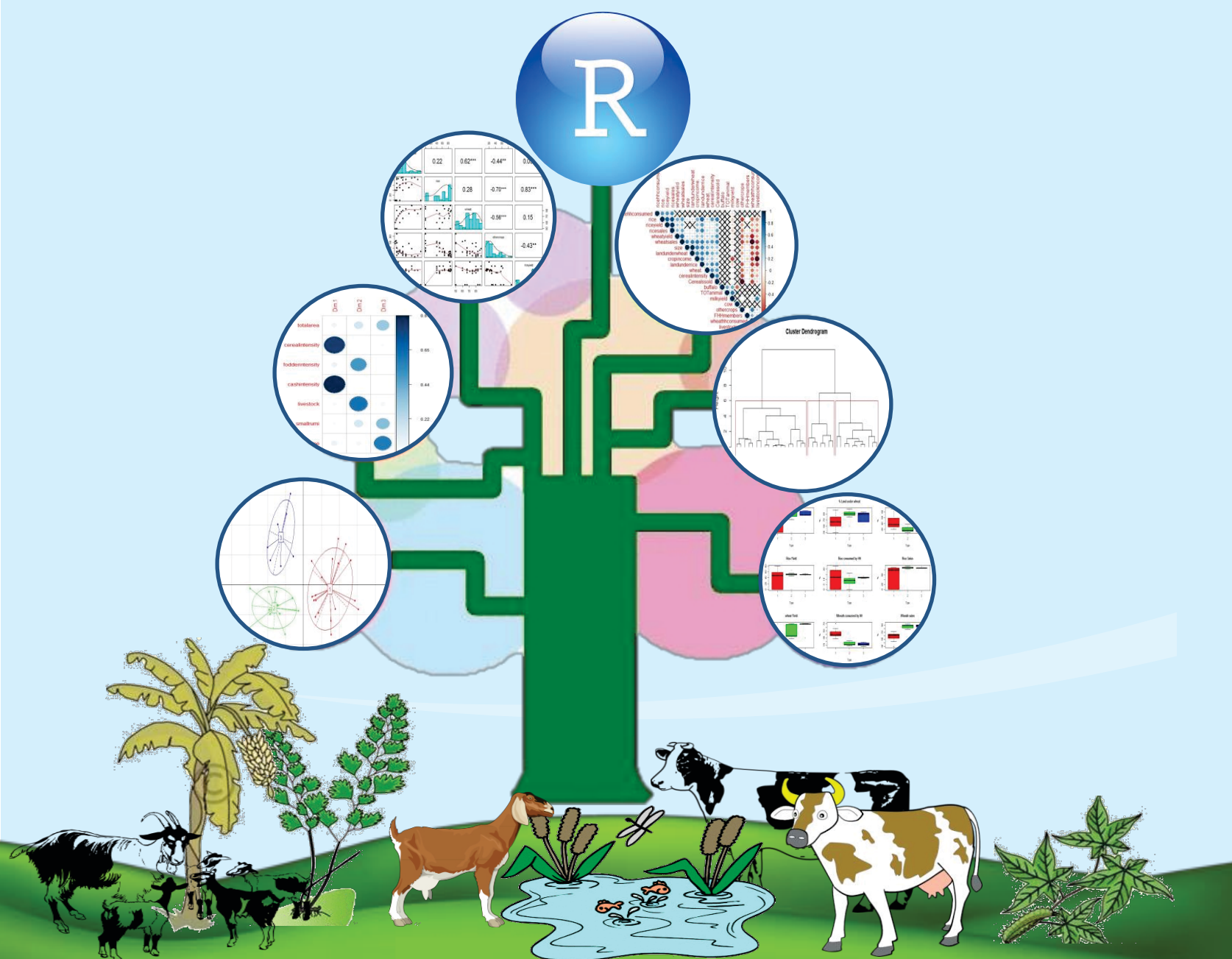


RESEARCH PROGRAM ON
Climate Change,
Agriculture and
Food Security



Researchers' Manual for Quantitative Farming Systems Typologies Applications using the R Statistical Tool

**Luis Barba-Escoto, A.K. Prusty, Santiago Lopez-Ridaura
N. Ravisankar, M. L. Jat, J. P. Tatarwal, A. S. Panwar**



**ICAR- Indian Institute of Farming Systems Research
Modipuram, Meerut – 250110, U. P.**

Researchers' Manual for Quantitative Farming Systems Typologies Applications using the R Statistical Tool

Authors

Luis Barba Escoto
A.K. Prusty
Santiago Lopez Ridaura
N. Ravisankar
M. L. Jat
J. P. Tetarwal
A. S. Panwar



ICAR-Indian Institute of Farming Systems Research
Modipuram, Meerut - 250 110, India



Correct Citation:

Barba-Escoto, L., Prusty, A. K., Lopez-Ridaura, S., Ravisankar, N., Jat, M.L., Tetarwal, J. P. and Panwar, A. S. (2019). Researchers' Manual for Quantitative Farming Systems Typologies Applications using the R Statistical Tool, ICAR-Indian Institute of Farming Systems Research, Modipuram, Meerut and International Maize and Wheat Improvement Centre (CIMMYT), Mexico, pp. 1-72.

Acknowledgements:

The contribution of Chief Agronomist, Agronomist and other scientists working in the AICRP on Integrated Farming Systems scheme is gratefully acknowledged for their involvement in data collection and facilitating in conducting the regional training workshops. We also acknowledge the support of Indian Council of Agricultural Research (ICAR), International Maize and Wheat Improvement Center (CIMMYT), CGIAR Research Programs on Wheat Agri-Food Systems and Climate Change, Agriculture and Food Security (CCAFS).

Published By: **Director**
ICAR-Indian Institute of Farming Systems Research
Modipuram, Meerut-250 110, India

© Director, ICAR-IIFSR

Printed at: Yugantar Prakashan Pvt. Ltd.
WH-23, Mayapuri Industrial Area Phase-I, New Delhi-64
Ph.: 011-28115949, 28116018, 9811349619, 9953134595
E-mail: yugpress01@gmail.com, yugpress@rediffmail.com

മി. എ. ഗുണസുന്ദരം

മി. എ. ഗുണസുന്ദരം

K. Alagusundram

Deputy Director General (Engineering)

&

I/c Deputy Director General (NRM)



ഇന്ത്യൻ കർഷക വ്യവസ്ഥാപനം

ഇന്ത്യൻ കർഷക വ്യവസ്ഥാപനം

ഇന്ത്യൻ കർഷക വ്യവസ്ഥാപനം

Indian Council of Agricultural Research

KRISHIANUSANDHAN BHAWAN-II,

PUSA, NEW DELHI-110 012

29/01/2019



FOREWORD

Integrated Farming Systems is crucial for sustaining the income of marginal and small farms under changing climatic scenario. Scientific studies on integration of various components is essential for which understanding the farming systems typologies practiced by different farmers and also their constraints and potential is critical for developing the Science based integrated farming system models. ICAR-Indian Institute of Farming Systems Research, Modipuram through its schemes such as AICRP on Integrated Farming Systems and All India Network Programme on Organic Farming have developed 45 IFS models and also refined 63 on-farm farming systems through farmer participatory research. These models have been developed based on specific needs of the region. On-Farm Research (OFR) component of AICRP on Integrated Farming Systems was working with large number of marginal and small farmers from 2011 in 31 districts covering 21 states to systematically characterize the existing farming systems, identify the constraints, make collective, compatible and convenient farm interventions and study the changes. Use of advanced tools such as linear programming, R etc is essential to study the farming systems typology and also optimization of components. Capacity building of Scientists in this area is essential and should be given priority.

ICAR-Indian Institute of Farming Systems Research in collaboration with International Maize and Wheat Improvement Centre (CIMMYT) under ICAR-CIMMYT work plan have done work on to capture the variability existing among farming systems through typology construction. The training workshop series have helped in enhancing the analytical capacity of young researchers and resulted in operational manual. It will also act as eye opener for developmental agencies in planning developmental action plans for targeted interventions in farming systems perspective for improving the income of farm family. The procedure for typology analysis using R statistical tool have been compiled as **“Researchers’ Manual for Quantitative Farming Systems Typologies Applications using the R Statistical Tool”** which is user friendly and can serve as a reference document for farming systems typology analysis. I congratulate the authors for bringing out the practical manual for farming systems typology analysis using R statistical tool.

(K. Alagusundaram)

Ph.: 91-11-2584 3415 Fax: 91-11-2584 2660 E-mail: ddgengg@icar.org.in

PREFACE

In India, contribution of small farmers to total farm output exceeds 50%, while they cultivate 44% of land. The holding sizes of marginal farms have decreased from the level of 0.40 ha in 1970-71 to 0.38 ha in 2010-11 and likely to reduce to the level of 0.32 ha with in this decade. By virtue of increased number of operational holdings (mainly due to fragmentation), their size is small but can be made profitable through interventions in farming system approach. In India, crop + livestock is the pre-dominant farming system and around 85 % of farm households practice it. Characterization of existing farming system in the farm household is essential for understanding the constraints and temporal dynamics of the system. On-Farm Research (OFR) component of AICRP on Integrated Farming Systems was working with large number of marginal and small farmers from 2011 in 31 districts covering 20 states to systematically characterize the existing farming systems, identify the constraints, make collective, compatible and convenient farm interventions and study the changes. A practical way of dealing with the complexity of farming systems variability and diversity is constructing typologies for distinction between farming systems.

Quantitative typologies based on multivariate analyses allows to identify significant differences among farm types and use this as the basis for targeting interventions as well as design alternative farming systems for different types of farms. As part of the ICAR-IIFSR-CIMMYT collaboration, four quantitative farming systems analyses and training courses have been carried out for the OFR scientists working under AICRP on IFS in four zones (Eastern zone at ICAR-RC, Patna, Bihar; Western Zone at AU, Kota, Rajasthan; Southern zone at TNAU, Coimbatore, Tamil Nadu and Northern zone at ICAR-IIFSR, Modipuram, Uttar Pradesh) during September, 2018 on “Quantitative farming systems typologies applications with the R statistical computing software”. This manual is the output of the workshop series and the document presented here is a key milestone for providing guidelines for constructing typologies in a step-wise approach to structure this process for its appropriation by young scientists. The editors are very much thankful to Dr T Mohapatra, Secretary DARE & DG ICAR; Dr Alagusundaram, DDG (NRM), Dr S Bhaskar, ADG (AAFCC) and Director, ICAR-IIFSR, Modipuram for their support and encouragement. We are also thankful to all the researchers from OFR centres of AICRP on IFS, CIMMYT, CGIAR Research Programs on Climate Change, Agriculture and Food Security (CCAFS) & Wheat Agri-Food Systems (WHEAT) for collaboration in successful completion of the workshop series and support in bringing out the document.

Authors

CONTENTS

	Page No.
Foreword	iii
Preface	v
1. Background	1
2. Summary of Training course	3
3. Methodology Adopted	5
3.1 Data Cleaning	7
3.1.1 Data Format	9
3.1.2 Working Directory	10
3.1.3 Download and install R and R studio	12
3.1.4 Starting a R studio session	13
3.1.5 Creating a new script	15
3.1.6 Saving script	15
3.1.7 Set working directory	16
3.1.8 Running codes	17
3.1.9 About executing the codes	18
3.1.10 Loading packages	18
3.1.11 Loading data	19
3.1.12 Building histograms	20
3.1.13 Building boxplots	21
3.1.14 Correlations	24
3.1.15 Choosing variables for PCA from the correlation matrix	28
3.1.16 Construct a vector with the names of the selected variables for PCA	30
3.2 Principal Component Analysis (PCA)	31
3.2.1 Run First PCA	33
3.2.2 Run 2 nd PCA	34
3.2.3 Accesses the most determinant variables on the PCs	36
3.3 Clustering	37
3.3.1 Clustering	39
3.3.2 Visualizing dendrogram	39
3.3.3 Cluster number selection	40
3.3.4 Cut the dendrogram with the number of clusters selected	42

3.3.5	Plot the clusters against the PC dimensions	42
3.3.6	Plot the PCs, the Variables and The Clusters	43
3.4	Describing Farm Types	45
3.4.1	Types profiling	47
3.4.2	Boxplots of variables vs types	47
3.4.3	Descriptive Statistics by type and total sample	47
3.4.4	Kruskal-Wallis and post hoc tests	48
4.	References	51
5.	Appendices	52
Appendix I	Format for summary table	52
Appendix II	Zone wise training workshop participant list	53
Appendix III	Ludhiana dataset and codebook	65
Appendix IV	Ludhiana typology R script	67

1. Background

Identification and characterization of farming systems is of utmost importance for simplifying the huge diversity of farm types in complex agro-ecological environment. Farming system and farms heterogeneity is enormous due to the same huge diversity of agro-ecological, socio-economic and resource endowment conditions in which they develop. Transforming, increasing agricultural productivity or in general rural livelihoods improvement, must consider the small holders variability for several reasons. Rural families develop different livelihood strategies driven by the opportunities and constraints derived from such diversity. Agroecology, markets and culture determine different land use patterns, but also within villages one may encounter differences in resources endowment, production orientation and objectives, even ethnicity, education, age, management skills and attitudes towards risks of the farmers, shape the diversity of strategies of natural resources exploitation (Titonell et al., 2010, Giller et al., 2010).

Adoption and scaling of technologies in agricultural systems is of central interest for academics and policymakers, as higher levels of adoption increase a higher return of investment in research and development impacting the economy of rural livelihoods is expected to happen. But numerous examples have proved that technologies with great potential are not adopted because the complexity and heterogeneity of the smallholders is not addressed. Particular farmers may need specific technologies as single “one size fits all” solutions do not exist (Goswami et al., 2014). Even more, reconfiguring farming systems to reach various productive and environmental objectives while meeting farm and policy constraints is challenging because of the large array of farm components and the multitude of interactions among them (Groot et al. 2012). A practical way of dealing with farming system complexity and diversity is to artificially stratify smallholders into subsets or group that are homogenous according to specific criteria. The term typology designates both, the science of type delineation and the system of types resulting from this procedure. Farms typologies are an attempt to capture farming systems heterogeneity and are considered a useful first step in the analysis of farm performance and rural livelihoods. Farm typologies have been used for nearly 20 years now (Kuivanan et al., 2016). Farm typology study recognizes that farmers are not a monolithic group and face differential constraints in their farming decisions depending on the resources available to them and their lifestyle (Soule 2001). Moreover, typology studies are of paramount importance for understanding the factors that explain the adoption and/or rejection of new technologies (Kostrowicki 1977; Mahapatra and Mitchell 2001). The heterogeneity of farming systems is created by a host of biophysical (e.g. climate, soil fertility, slope etc.) and socio-economic (e.g. preferences, prices, production objectives etc.) factors (Ojiem et al. 2006).

Several approaches can be used for developing farm typologies, from participatory workshops where local knowledge and stakeholders perception of the main factors that explain local diversity

is taken into account, to the use of surveys and the statistical multivariate analysis of data for typologies construction (Alvarez et al., 2018). Quantitative typologies based on multivariate analyses allows to identify significant differences among farm types and use this as the basis for targeting interventions as well as design alternative farming systems for different types of farms. As part of the ICAR-IIFSR-CIMMYT collaboration, quantitative farming systems analyses an training courses have been carried out and the course presented here is a key milestone for the appropriation of these approaches by local scientists.

The main objective of these series of workshops is to provide AICRP on IFS and other NARS researchers with the tools to capitalize on the farm level data collected and build a typology to understand the diversity of farming systems in the areas where they work.

2.Summary of Training Course/ Course Design

The course was designed in such way so that participants of these courses will gain knowledge on

- ✓ Why and how to build farming system typologies
- ✓ Overview of the R statistical computing software environment
- ✓ Exploratory data analysis for data curation, variables selection
- ✓ Multivariate methods: principal components analysis (PCA) and hierarchical clustering (HC) for households' ordination
- ✓ Types profiling
- ✓ Reporting a farm/farming system typology

The result of typology construction is the grouping of a large number of farm households into a set of fewer homogenous categories. We mean homogeneous because the members belonging to a certain group should share more or less the same characteristics.

By the end the typology building process participants will be able to prepare a basic report containing:

- Introduction
 - ◆ A brief description of the zone under investigation
 - ◆ A description of the aims of this study and how typologies are needed
- Methods
 - ◆ A brief description of the applied survey and nature of the data
 - ◆ Total number of households
 - ◆ The multivariate methods a brief description (PCA, Clustering)
- Results
 - ◆ Total number of variables.
 - ◆ Variables selected for multivariate ordination with a brief description of the reasons why they were selected.

- ◆ The reasons why some variables might have been excluded from the original database, as some might not provide any additional or all farm households have similar values.
- ◆ The number of PCs selected from PCA, based on which criteria (Screeplot, eigenvalues>1...).
- ◆ The total % of variance explained by the retained PCs.
- ◆ The number of clusters selected based on what criteria such as: by observing the dendrogram and selecting equitable groups, the within sum of squares criterium, or by checking if the partition is meaningful, that is if clusters make the diversity of HH more explicit and meaningful.
- ◆ Boxplots of variables per type to allow comparison.
- ◆ Table of mean and standard deviation of each variable by type and total sample (**Appendix II**).
- ◆ A text describing each type including their most conspicuous characteristics and, when possible, making reference of the whole sample mean or median and in reference to the mean or median for the other types. Actual mean values and in reference to the other types.
- ◆ Each type description must include a synthetic descriptive name for the type.
- ◆ A brief description for each type of constraints and opportunities of each type in terms of possible interventions helps for portraying how interventions should or should not be different for each type.

3. Methodology Adopted

In this manual an example of the steps to perform a typology from data cleaning to Principal Component Analysis (PCA) and a Hierarchical Clustering (HC) with the R software in RStudio is provided, with a case study of Ludhiana, Punjab India.

The methods followed in R consists of:

1. Data cleaning

- a. Data format in .csv files
- b. Detection of Outliers and zero or near zero variance variables through
 - i. Building histograms
 - ii. Building boxplots
- c. Selection of variables for PCA by detecting meaningful correlations
 - i. Significant correlations matrix
 - ii. Rule of thumb: No. of households (HH) should be atleast 5 times of variables selected for PCA

2. PCA

- a. Selection of Principal Components (PCs)
 - i. Screeplot
 - ii. PCs Eigenvalues >1

3. Clustering

- a. Hierarchical Clustering
- b. Selecting the optimum number of cluster (WGSS within groups sum of squares criteria)

4. Describing farm types

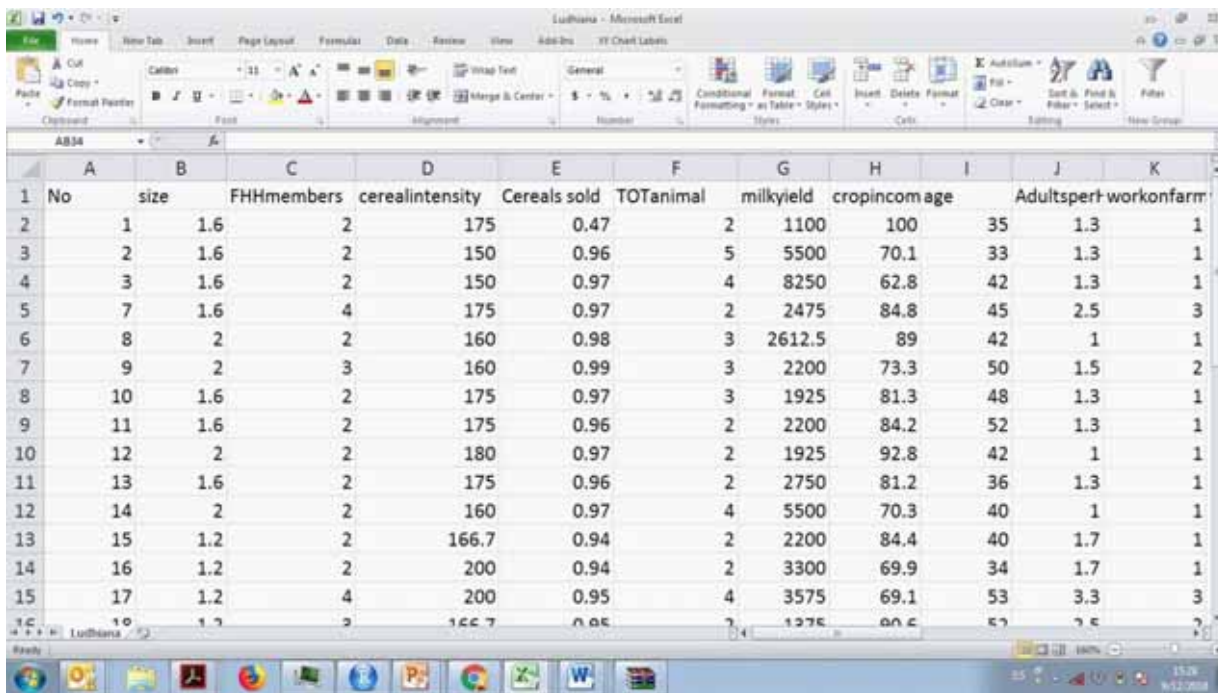
- a. Boxplots variables vs Types(clusters)
- b. Table with descriptive statistics

Sample data set used in this example is available as annexure IV or soft copy of the same may be obtained on request by email from aasiana143@gmail.com (Dr. A. K. Prusty)

3.1 DATA CLEANING

3.1.1 Data format

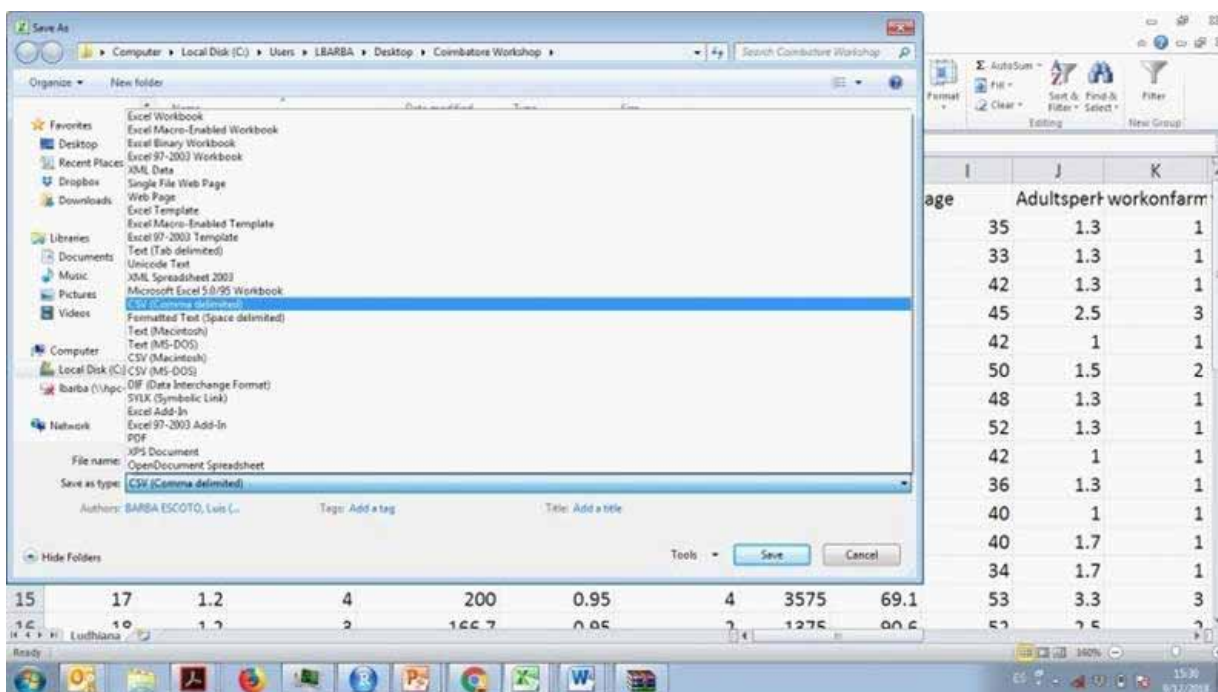
Data must be in a matrix format in excel, that means in rows households in columns variables. For example see the Ludhiana data set (Annexure III), once you copy and paste the data set from this manual to excel:



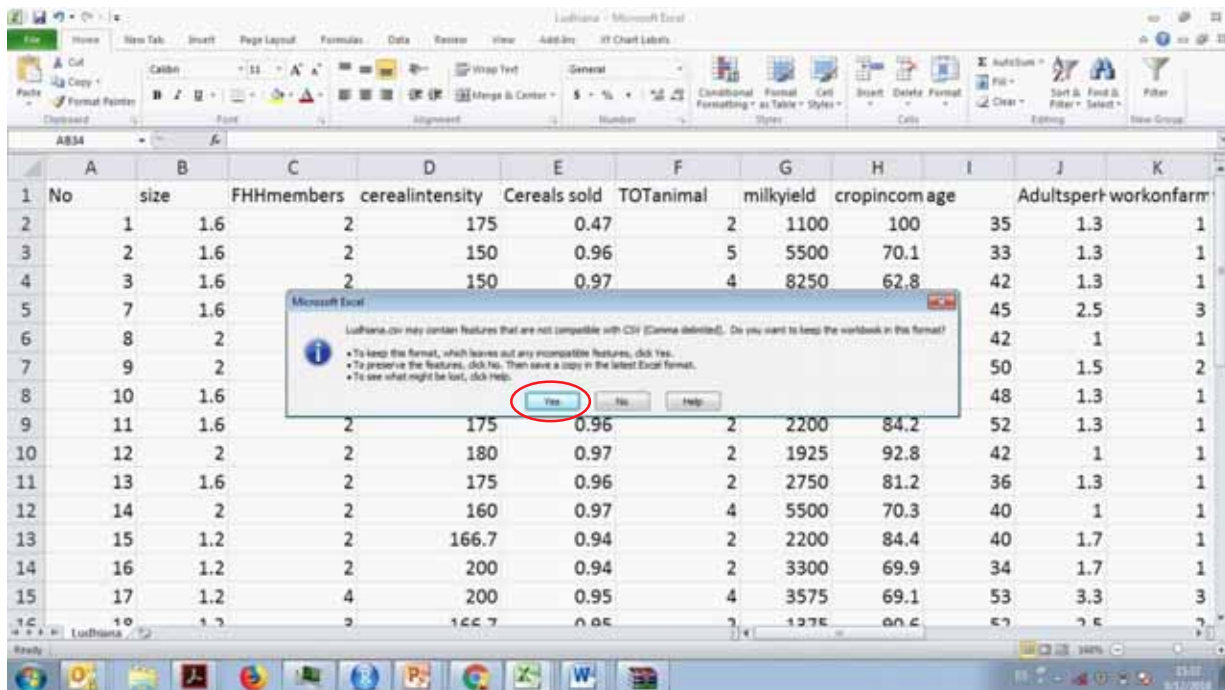
No	size	FHHmembers	cerealintensity	Cereals sold	TOTanimal	milkyield	cropincomage	Adultsper/workonfarm			
1	1	1.6	2	175	0.47	2	1100	100	35	1.3	1
2	2	1.6	2	150	0.96	5	5500	70.1	33	1.3	1
3	3	1.6	2	150	0.97	4	8250	62.8	42	1.3	1
4	7	1.6	4	175	0.97	2	2475	84.8	45	2.5	3
5	8	2	2	160	0.98	3	2612.5	89	42	1	1
6	9	2	3	160	0.99	3	2200	73.3	50	1.5	2
7	10	1.6	2	175	0.97	3	1925	81.3	48	1.3	1
8	11	1.6	2	175	0.96	2	2200	84.2	52	1.3	1
9	12	2	2	180	0.97	2	1925	92.8	42	1	1
10	13	1.6	2	175	0.96	2	2750	81.2	36	1.3	1
11	14	2	2	160	0.97	4	5500	70.3	40	1	1
12	15	1.2	2	166.7	0.94	2	2200	84.4	40	1.7	1
13	16	1.2	2	200	0.94	2	3300	69.9	34	1.7	1
14	17	1.2	4	200	0.95	4	3575	69.1	53	3.3	3

Here, “No” the first column of the excel sheet refers to the household ID.

Data should be saved as a .csv file. This can be done in excel by choosing in menu bar: File>Save as> Save as type> CSV(Coma delimited)



A message will appear



Click: Yes. Now the file is saved as a .csv file.

3.1.2 Working Directory

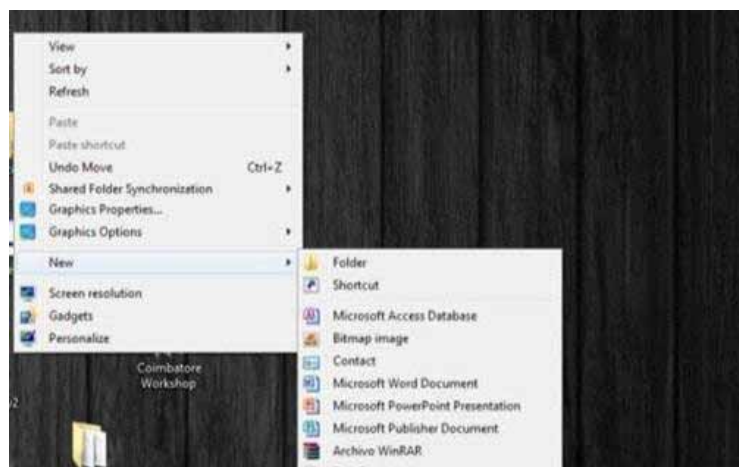
R requires to read the data (our data set) and store data (tables, graphs, etc.) in a specified working directory.

You must create a folder in a convenient place in your hard disk. For ease of this manual instructions create one in your desktop for example:

C:\Users\LBARBA\Desktop\CoimbatoreWorkshop

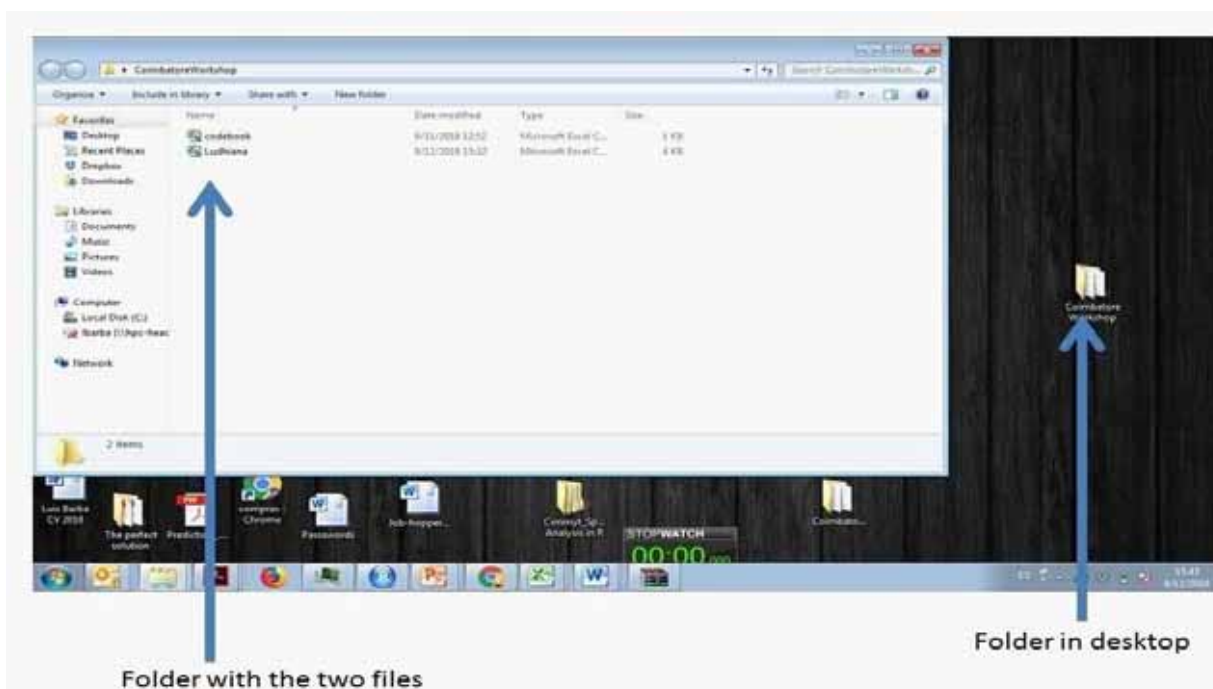
Here, Coimbatore workshop is the Folder name created on your desktop.

In desktop: mouse right click>New>Folder



Now, you must save in this directory the following two files:

1. Your data set, in this case Ludhiana.csv
2. A code book that must be constructed with the variables Acronym, variable explicit name and units, this excel file name should be codebook and saved also as .csv (Here codebook is the excel file having acronym of variables saved as a .csv file).



The codebook should look like this:

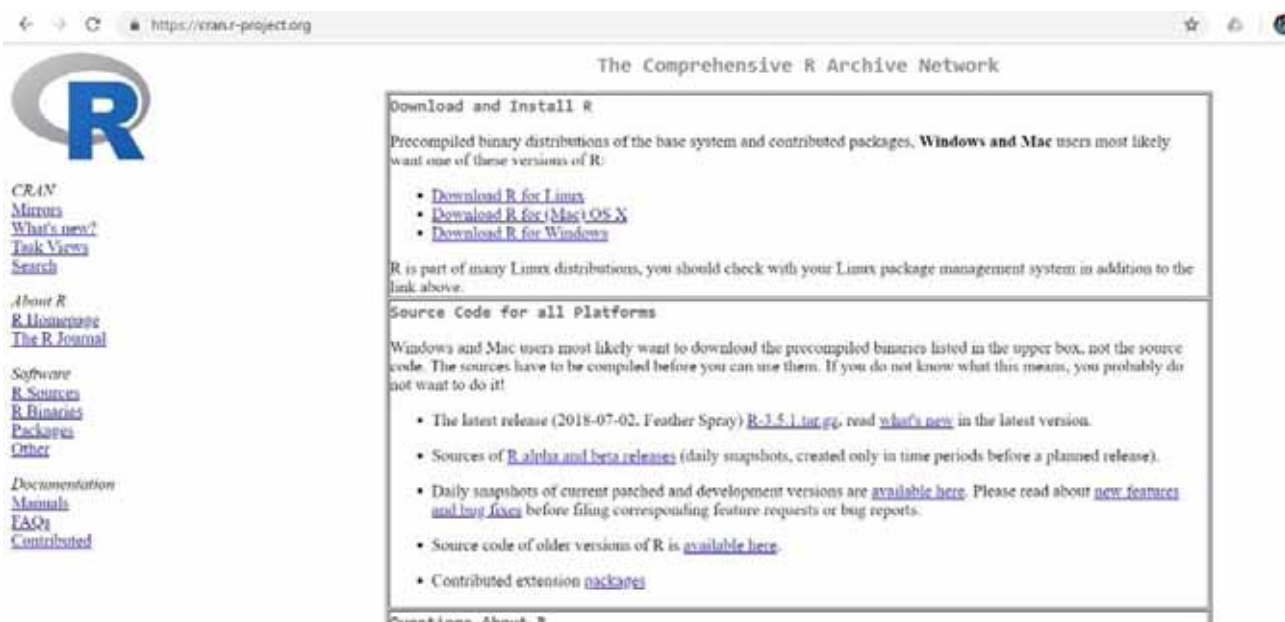
Acronym	Variable	Unit
size	Land Size	ha
FHHmembers	HH Size	Persons
cerealintensity	Cereals Intensity	Ratio
Cereals sold	Cereals Sold	%
TOTanimal	Total Animals	Animals
milkyield	Milk Yield	Liters/Year
cropincome%	Crops Income	%
age	Household Head Ag Years	
AdultsperHA	Adults per Ha	persons/ha
workonfarm	Work on Farm	Number
workofffarm	Work off Farm	Number
children	Children	Number
nonveg	Non-Vegetarians	Number
veg	Vegetarians	Number
landunderrice	Land under Rice	ha
landunderwheat	Land under Wheat	ha
rice%	% Land under rice	%

3.1.3 Download and install R and R studio

Download and install latest version of R and R studio (For R)

First you must install R, depending on your operating systems

<https://cran.r-project.org/> (For R)

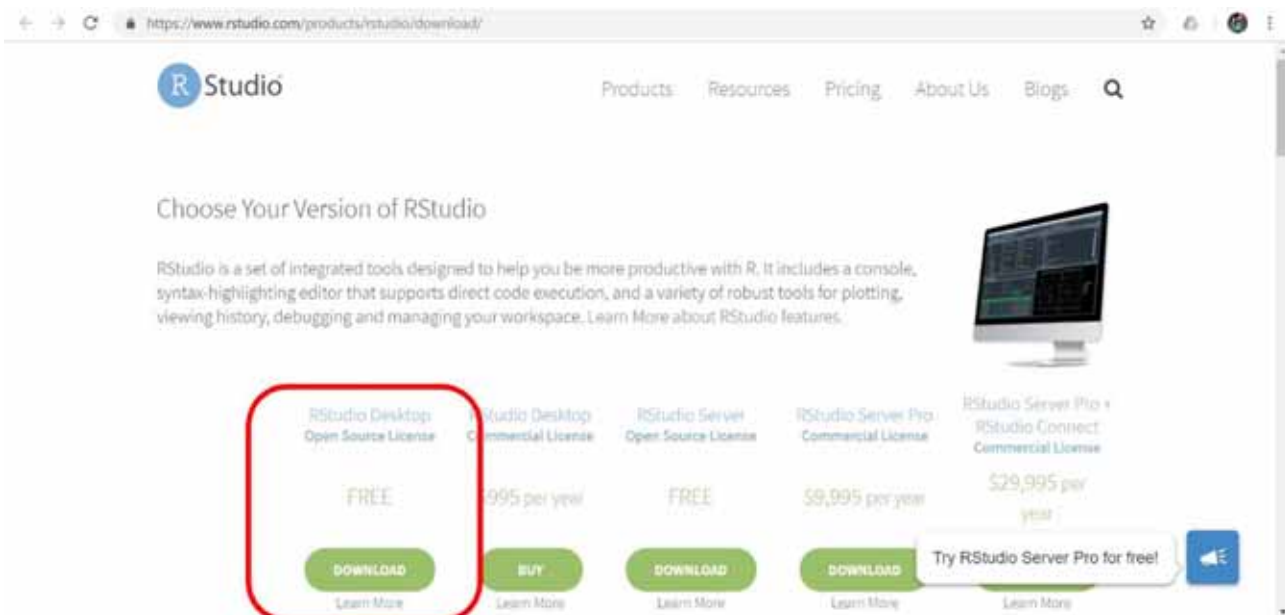


Then R studio should be installed

<https://www.rstudio.com/products/rstudio/download/> (for R studio)

Choose the free version

While installing Rstudio, the installer might ask you if you want to create desktop shortcut, It's up to you to create it in desktop for ease of finding the programme.



3.1.4 Starting a R Studio session

Start R Studio: click on Rstudio icon



Install Packages

If this is the first time you have started RStudio **EVER**

You will need to **INSTALL** three packages

ade4 **psych**

corrplot

agricolae

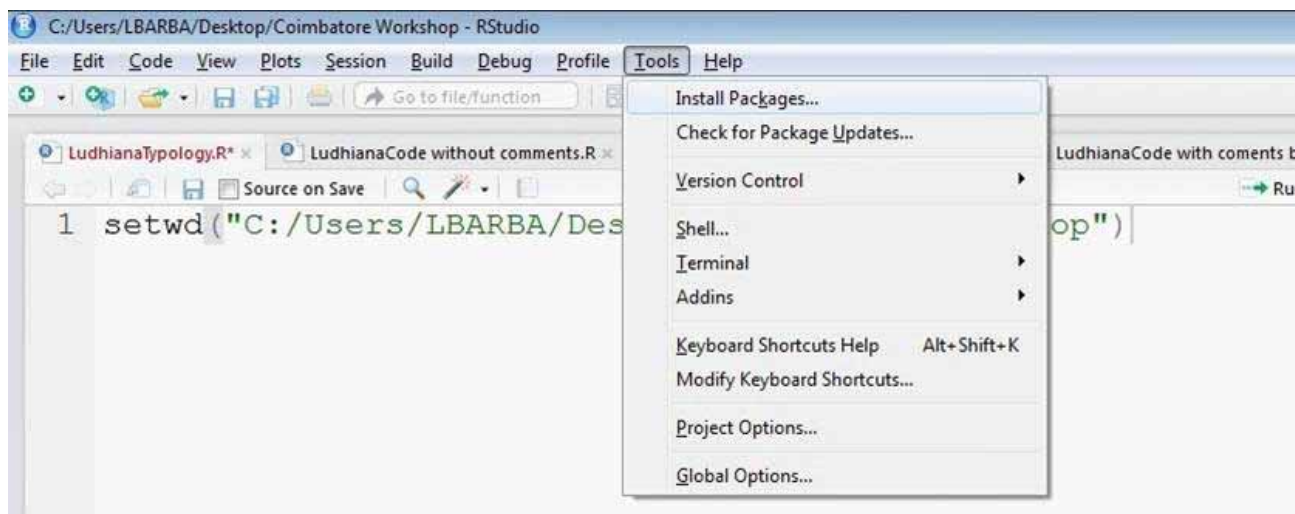
(not mandatory you may install **factoextra**)

Once correctly installed **they don't' need to be installed again**

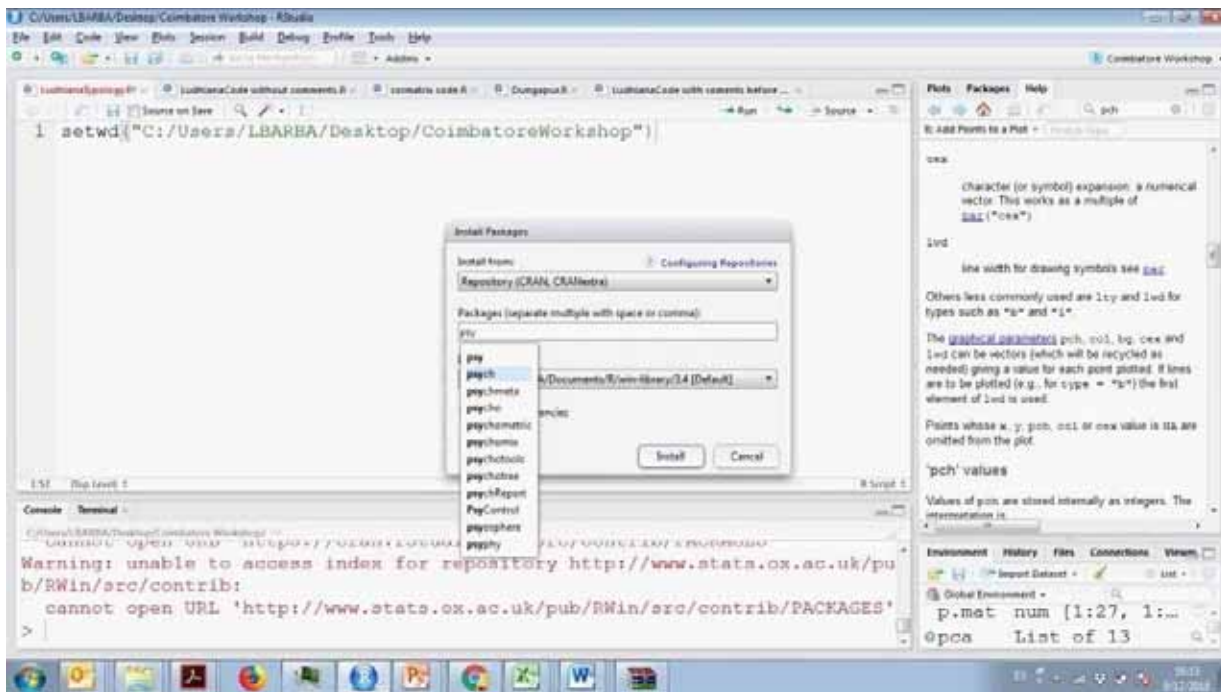
For installing packages **you must have internet connection** as they are retrieved from remote depositories

To install one package you do:

Menu>Tools>Install Packages>Packages (separate multiple with space or comma):

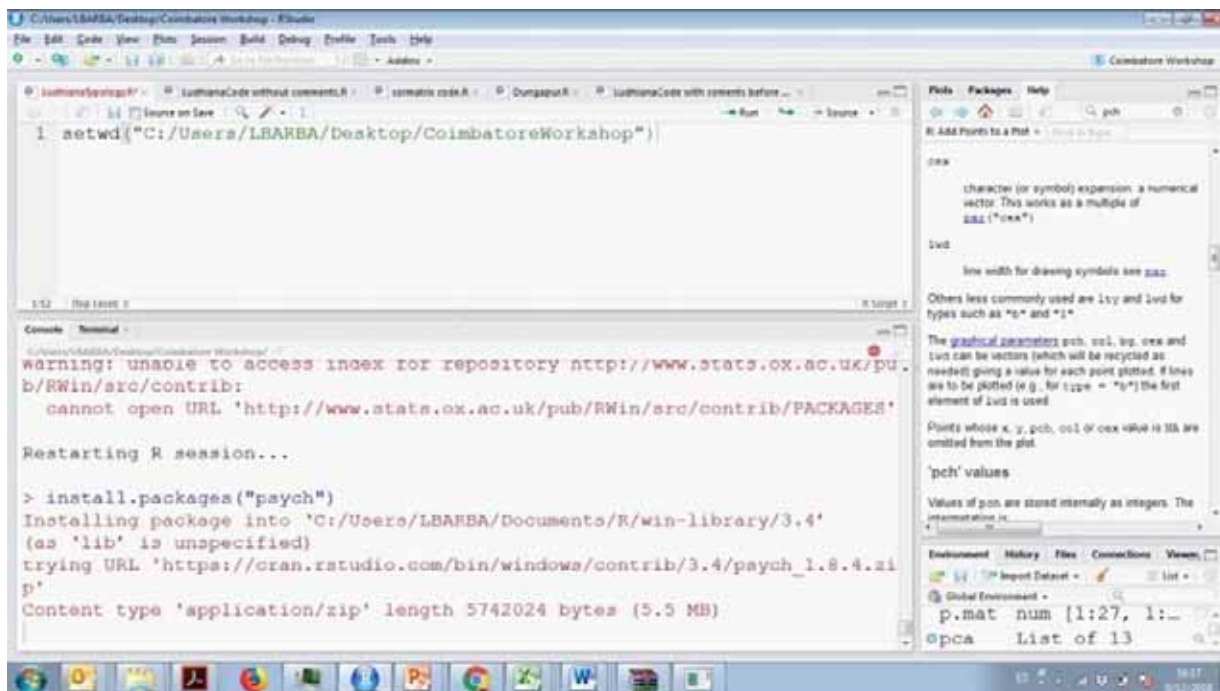


Type the name of the package, if internet connection is OK, just typing 3 letters might display the options, select the correct package and click: Install



Do the same for the other two packages

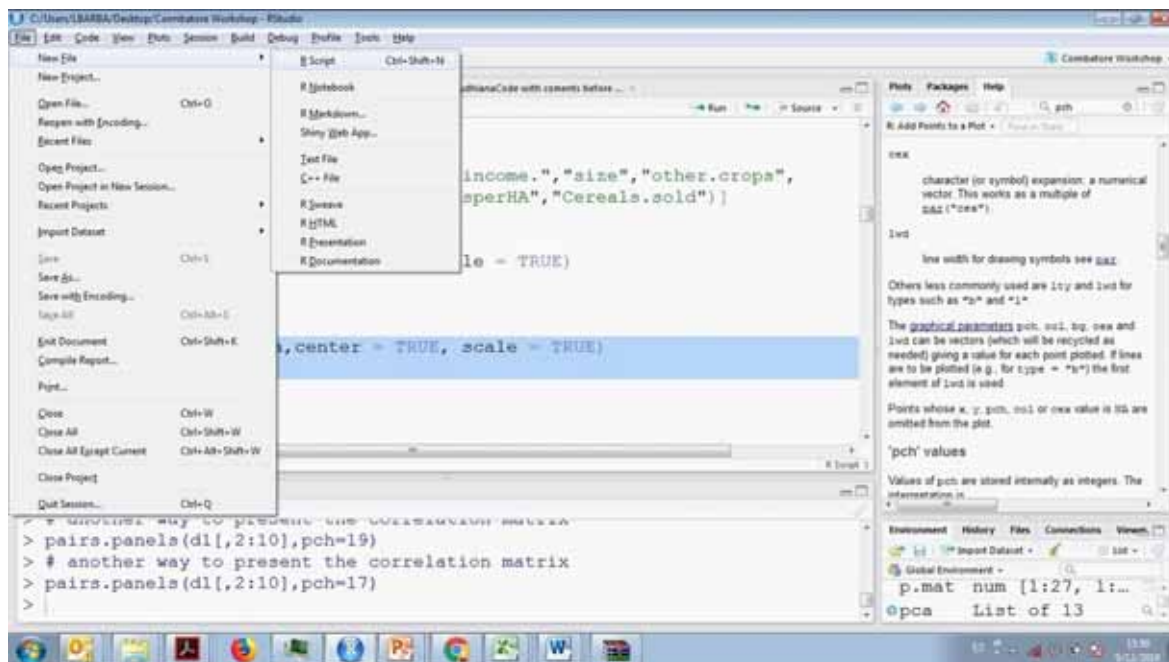
You may be able to see the installation process in the console, left panel below



By the end of the installation process of each package you must get no error messages

3.1.5 Create a new script

File> New File> R script (Ctrl+Shift+N)



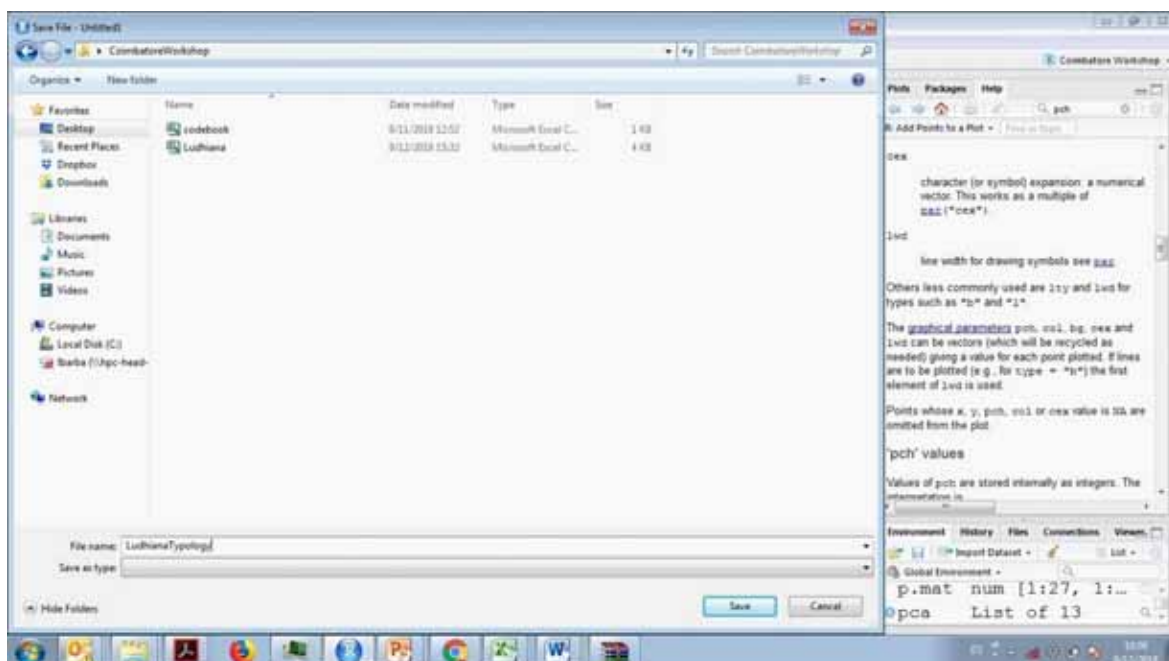
3.1.6 Save the script

Save the code in your working directory

Menu>File>Save as

Save it as: LudhianaTypology

(Note: Save it in Folder created in Desktop, see section 3.1.2)

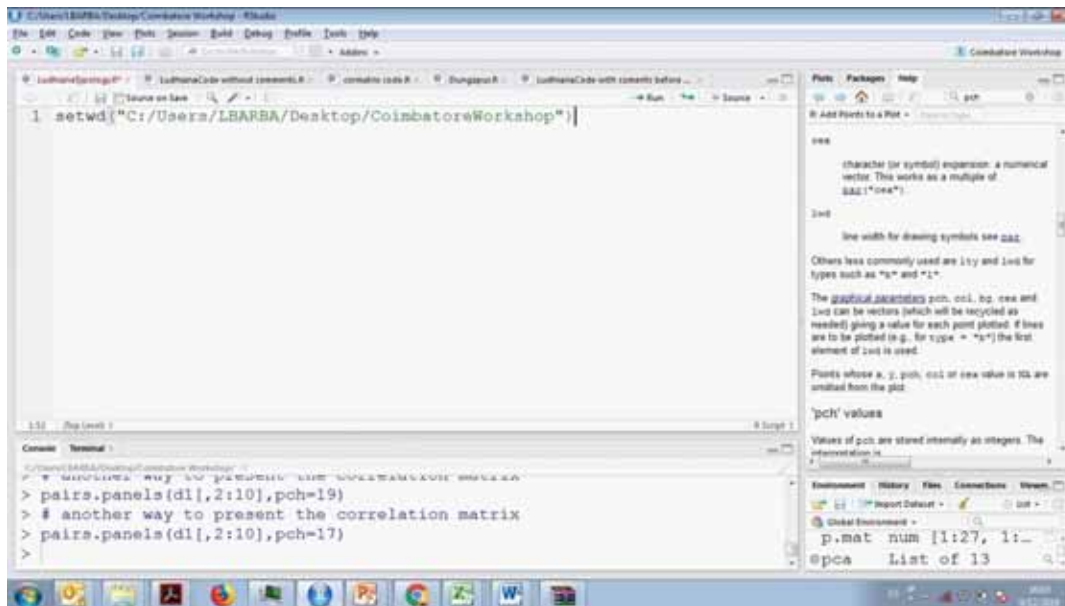


3.1.7 Set working Directory

Write your first line of code:

```
setwd("C:/Users/LBARBA/Desktop/CoimbatoreWorkshop")
```

`setwd()` stands for set working directory



The route inside `setwd()`

“C:/Users/LBARBA/Desktop/CoimbatoreWorkshop”

Makes reference to the directory you created in desktop

It is very important that you set the working directory as R will extract the data base from this directory and also will save the outputs from analysis.

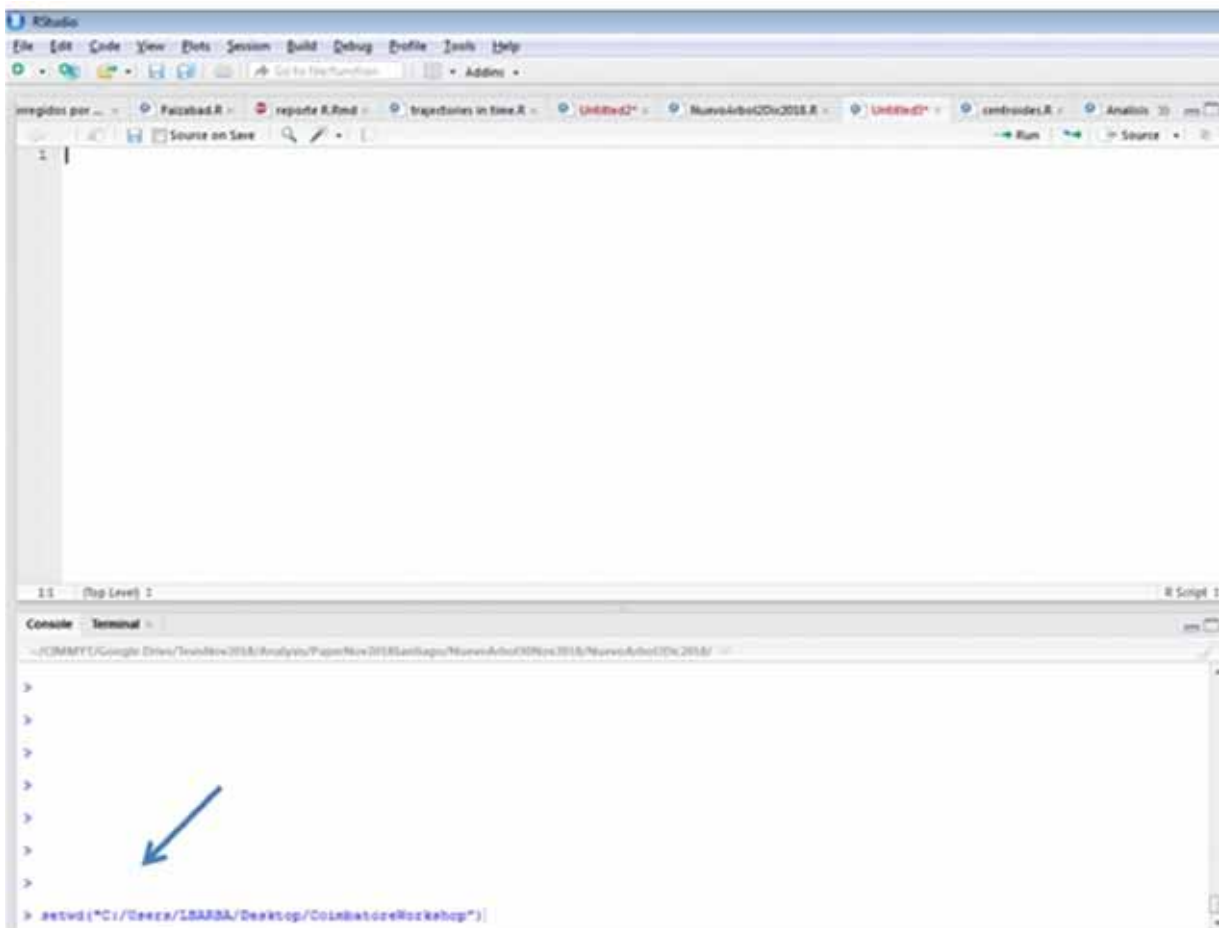
Alternatively, you may choose the working directory following:

Menu>Session>Set Working Directory>Choose Directory>
and choose the folder you want as working directory

Note: [Choose the folder created in desktop as working directory] i.e. Coimbatore Workshop



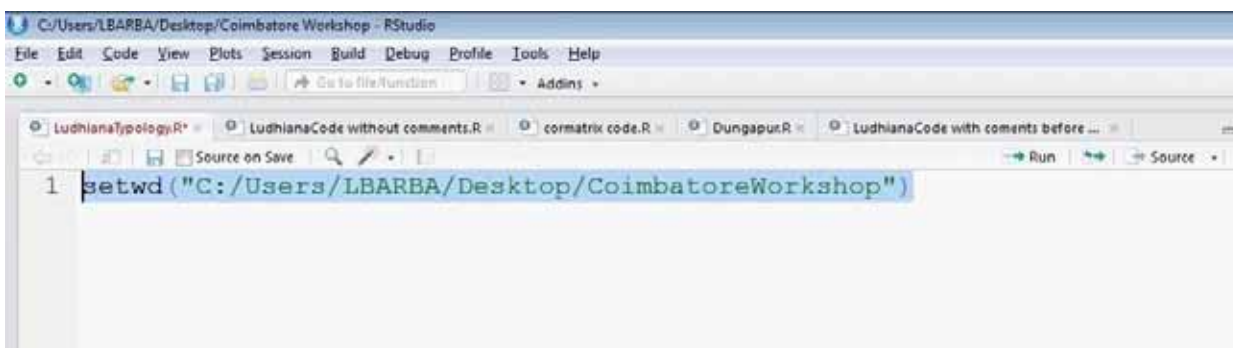
The line of code setting the working directory will be displayed in the console.



You must copy and paste this line without the “>”, as your first line of code, next time you open Rstudio you will just run this line for setting the working directory, saving you some time.

3.1.8 Running lines of code

To run (execute) lines of code select the line to be run



and press Ctrl + Enter

Or place the cursor in the line to be run and



press Ctrl + Enter

3.1.9 About executing code

- Executing lines of code means that a function, in this case `setwd()` is applied over an object, in this case, the working directory route
- Objects are created only if they are executed (if they are selected and Ctrl+Enter is pressed)
- Codes are sequential:
 - ◆ if I call for the object `d`, and it has not been previously created I won't be able to call it.
 - ◆ if for example: line 2 calls for object "d" from line 1, and line 1 is not previously executed, "d" will not exist, object "d" can be called in line 2 only after line 1 has been executed.
- Each time you start Rstudio again, you need to run code from right from line 1 again.
- After the symbol `#` nothing is executed, `#` is for commenting code in R

3.1.10 Load Packages

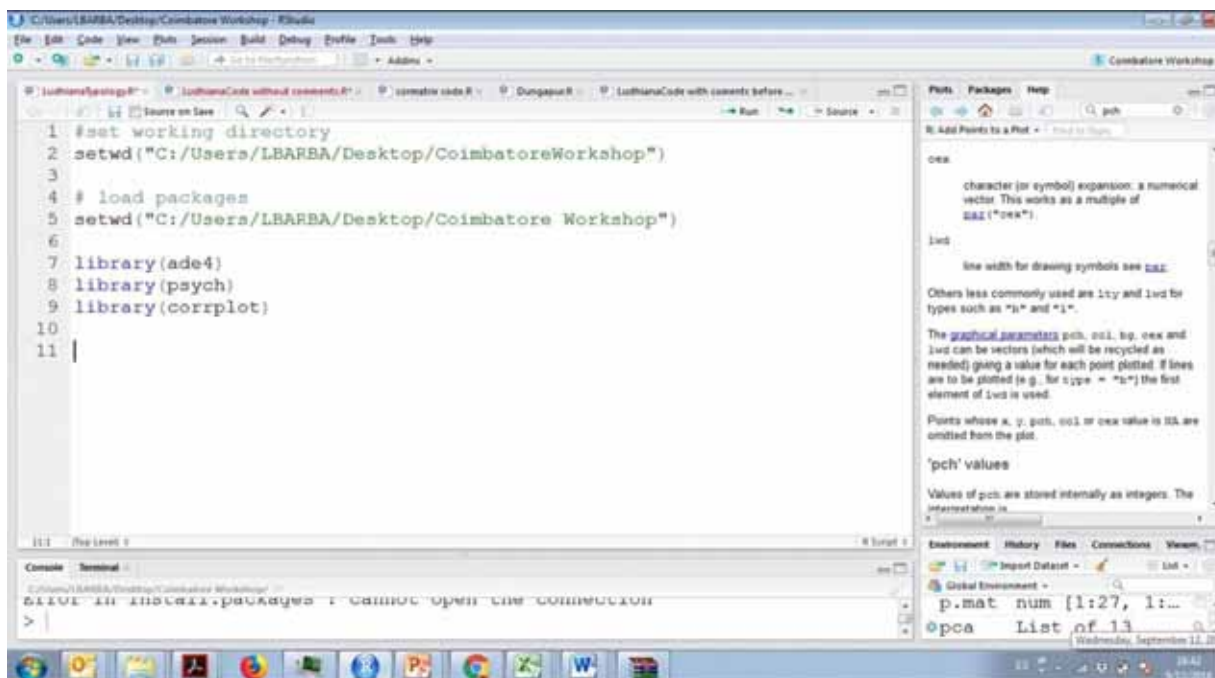
Once packages have been installed they don't need to be installed again but they need to be **loaded** each time we open RStudio again.

Packages contain programs (codes in R) that are full of functions that will do the actual computations in our analysis.

For loading the packages (**ade4**, **psych**, **corrplot** and **agricolae**) type:

library(ade4)
library(psych)
library(corrplot)
library(agricolae)

press Ctrl+Enter after each line to load the packages. (above packages are needed for typology)
Your code should look by now like:



3.1.11 Loading data

We will assign the name “d” to the data set of Ludhiana. For that we will use the function **read.csv()**

d<-read.csv(“Ludhiana.csv”)

Remember the command for objects assignment is:

<-

Load also the codebook file, as we will extract the graphs titles as well as x- and y-axis labels

codebook<-read.csv(“codebook.csv”)

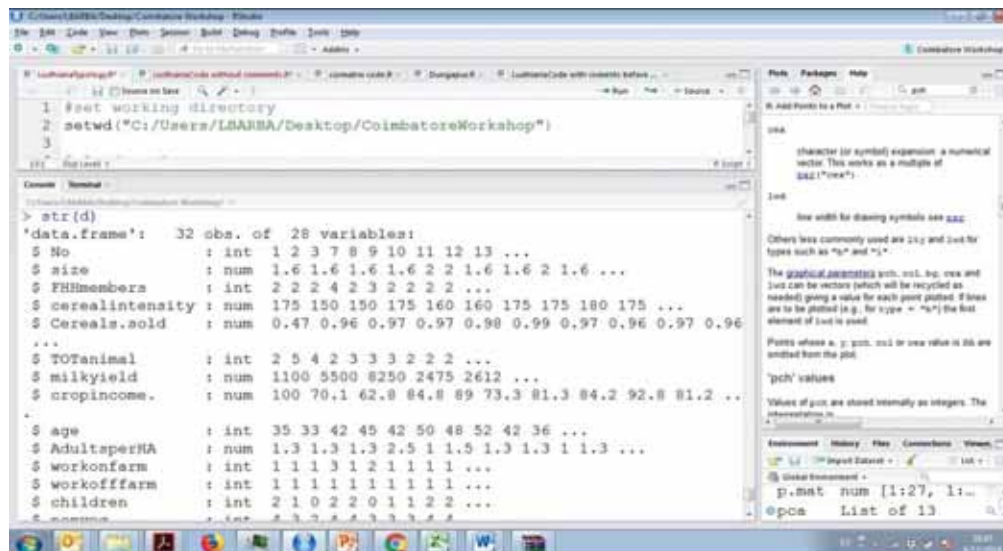
Inspect the number of variables and observations and look for unwanted variables

Check how many variables and observations we have in Ludhiana.csv

str(d)

str(), is a function to check the structure of an object, in this case a dataframe

You should be able to see the output (32 obs. 28 variables) in the console



If you look at the output, we must not take into consideration the variable “No” as it is only the HH IDs, for that we select only columns 2 to 28 of from “d”, and name this new object “d1”

`d1<-d[,2:28]`

Here “d1” is the subset of “d” with only columns 2 to 28 (Total 27 variables)

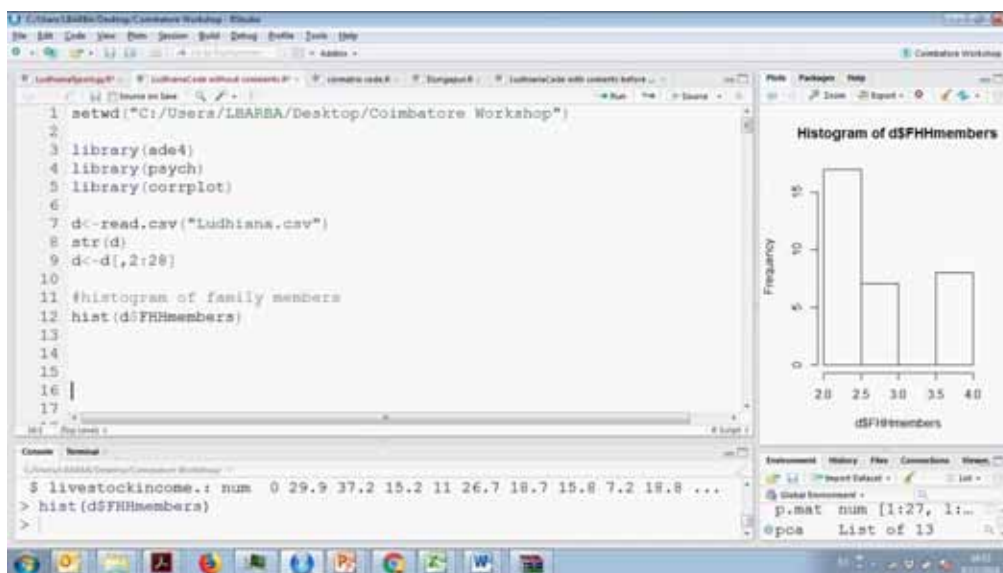
3.1.12 Build some histograms of the 27 variables

To check for zero and non-zero variance variables or possible outliers, histograms and boxplots are useful visualizations of our variables.

You can make a histogram for only one variable with the function `hist()`

`hist(d$FHHmembers)`

you can see the output graph in the plots panel



But we need to inspect 27 variables!!

We will do 27 histograms with a **for** loop

```
par(mfrow=c(3,4))
for (i in 1:27) {
  hist(d1[,i],main=codebook$Variable[i],
  main=codebook$Unit[i]
}
```

par(mfrow=c(3,4)) means we split the plots panel in 3 rows and 4 columns we will be able to see 3X4=12 histograms per page

With the **for** loop we are able to apply the function **hist()** for each i-th element of the sequence 1:27, the function name stands for histogram.

with **d1[,i]** we are indexing each column of the dataset in d1

with **main=codebook\$Variable[i]** we are selecting the title of the histogram from the codebook, column Variable, and in [i], the i-th row

with **main=codebook\$Unit[i]** we are selecting the title of the histogram from the codebook, column Unit, and with [i], the i-th row

The output will show in the plots panel



3.1.13 Build the boxplots

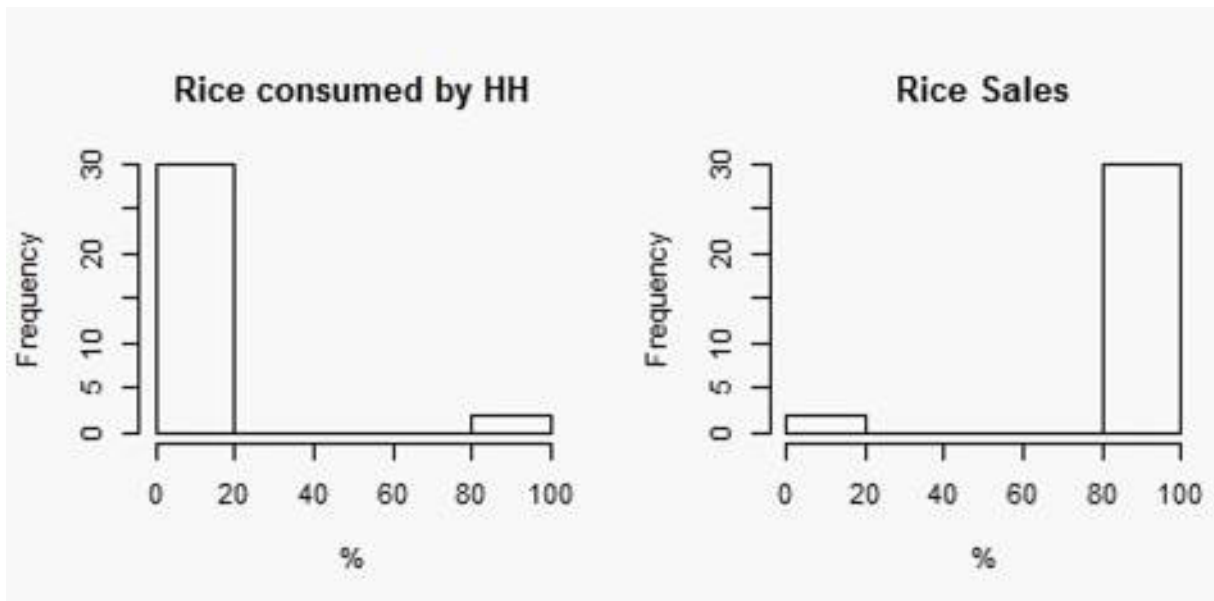
For boxplots we only change **boxplot()** instead of **hist()**. The y-axis now will show the variables units

```

par(mfrow=c(3,4))
for (i in 1:27) {
boxplot(d1[,i],main=codebook$Variable[i],ylab =
codebook$Unit[i])
}

```

How Histograms and Boxplots help for cleaning data and selecting variables for dimension reduction (PCA).

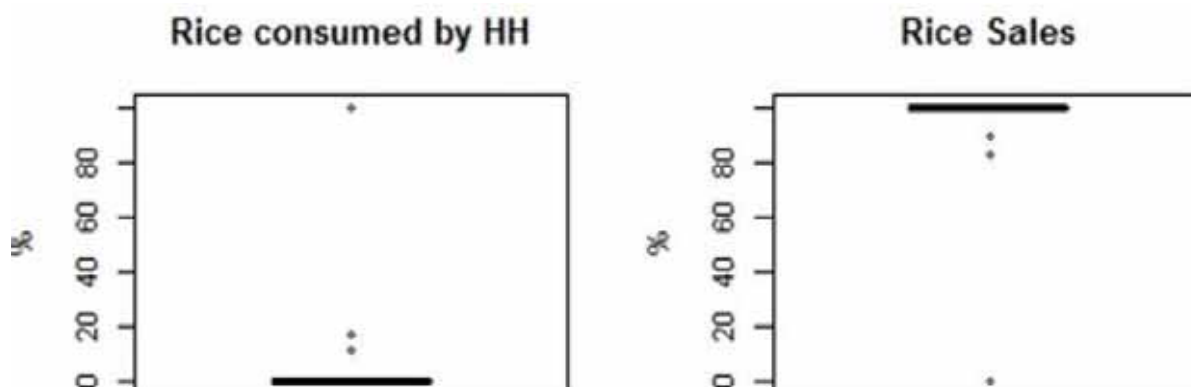


For example, we have found this behavior on Rice consumed by household and Rice sales

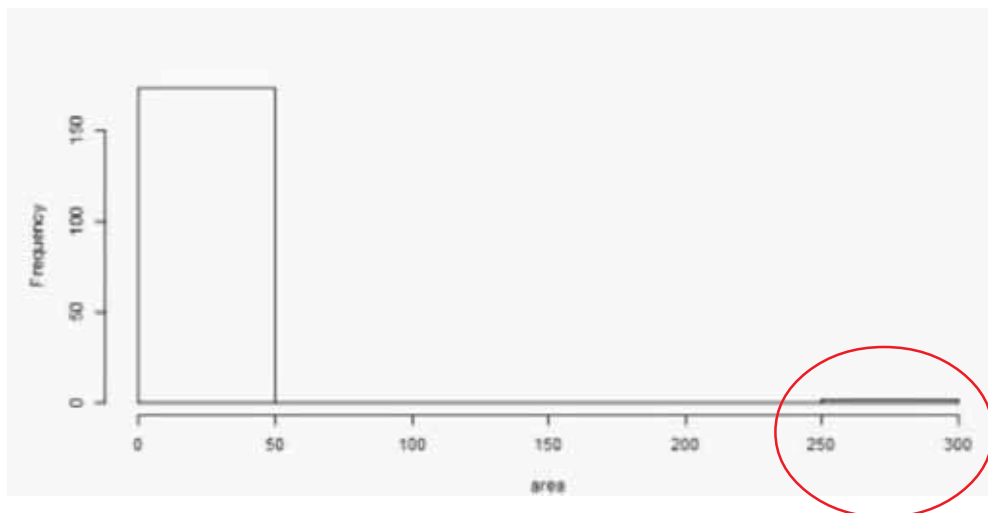
We may conclude that these variables:

- In a certain way give the same information, the less rice is consumed the more is sold
- But most importantly , the variance of this variables is near-zero, that is, they bare little information as almost 30 of 32 HH sale their rice, value of 100%

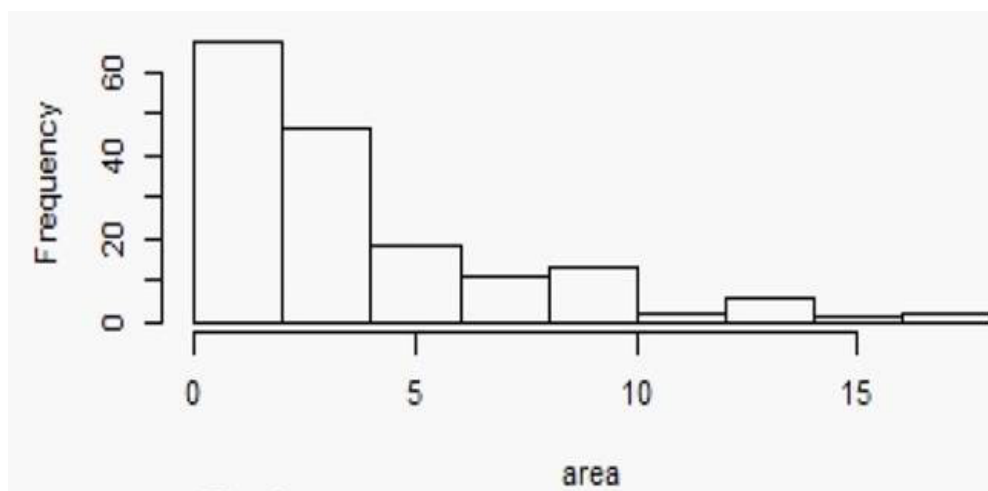
If we take a look at the boxplots the same information could be extracted



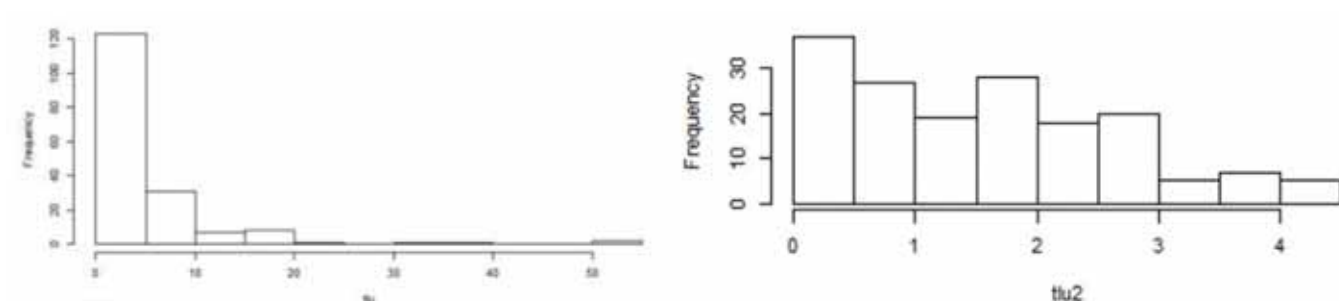
In the case of the following example (Alvarez et al. 2014). When plotting the variable area they found an outlier



And after removing it, we can see the distribution changes a lot



The same for TLU after removing a HH with TLU of 50



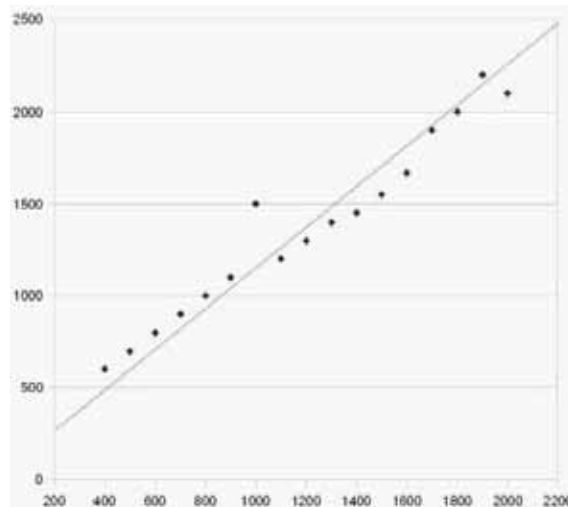
3.1.14 Correlation

Correlation among variables helps for identifying linear dependencies or collinearities. Some variables might be constructs of another, for example

$$x=2y$$

this case shows that x is only a construction of another variable, a case of linear dependency. IF

we plot them they might show a classic 45 degrees line if we plot them in a scatterplot



The correlation among those variables would be one or very close to 1. If we find those correlated variables, we must only take one, of the correlated variables.

Additional important information that correlation analysis throws, is that, only by analyzing the correlation matrix we can begin to have an idea of what the factors (latent variables, PCs...) will be as a result of our factor analysis.

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	1.000					
sentence	.696	1.000				
paragrap	.743	.724	1.000			
lozenges	.369	.335	.326	1.000		
cubes	.184	.179	.211	.492	1.000	
visperc	.230	.367	.343	.492	.483	1.000

We might conclude from the figure above that sentence, paragraph and wordmean might constitute one factor and that cubes, visperc and lozenges constitute another.

So in summary, we are looking for correlated variables below **r=1 and above 0.3**, if we cannot find variables correlated above 0.3 our analysis won't work.

Compute the correlation matrix

`cor(d1)` computes the correlation matrix

We will store the correlation matrix of the data set d1, `cor(d1)` in the object `corMatrix`

```
corMatrix<-cor(d1)
corMatrix
```

Here, `corMatrix` is the name given to the output matrix of correlated function of "d1" data set.

We can save this matrix if we want as a csv file with:

```
write.csv(corMatrix,"corMatrix.csv")
```

This will be stored in our working directory folder.

Compute the significant correlations

The following function `cor.mtest` will help us compute the p-values of all the `corMatrix`

This code below only creates the function, we later will apply it to the object `corMatrix`

```
cor.mtest <- function(mat, ...) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat<- matrix(NA, n, n)
  diag(p.mat) <- 0
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], ...)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
    }
  }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
```

We compute the p.values of `corMatrix`

```
p.mat <- cor.mtest(corMatrix)
```

Here, `p.mat` is the name given to the matrix of p.values of `corMatrix`.

we check the output by looking only at the first 5 variables

```
head(p.mat[, 1:5])
```

we can also save this matrix of p-values of the correlations

```
write.csv(p.mat, "CorpvaluesMatrix.csv")
```

Visualization of correlations

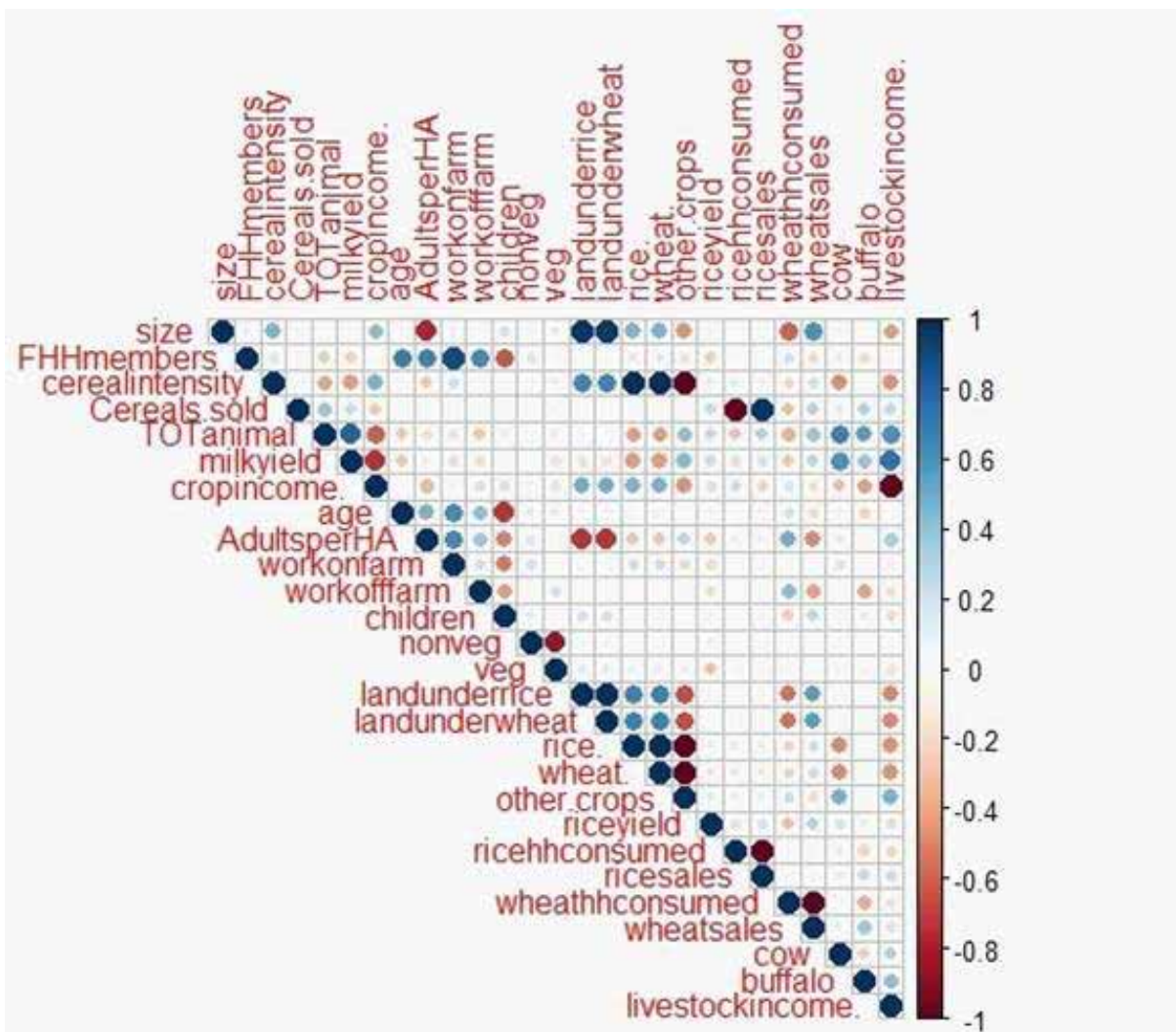
To visualize the plots we need to get back to only one plot per page

```
par(mfrow=c(1,1))
```

We will plot: A first glance

```
corrplot(corMatrix,type="upper")
```

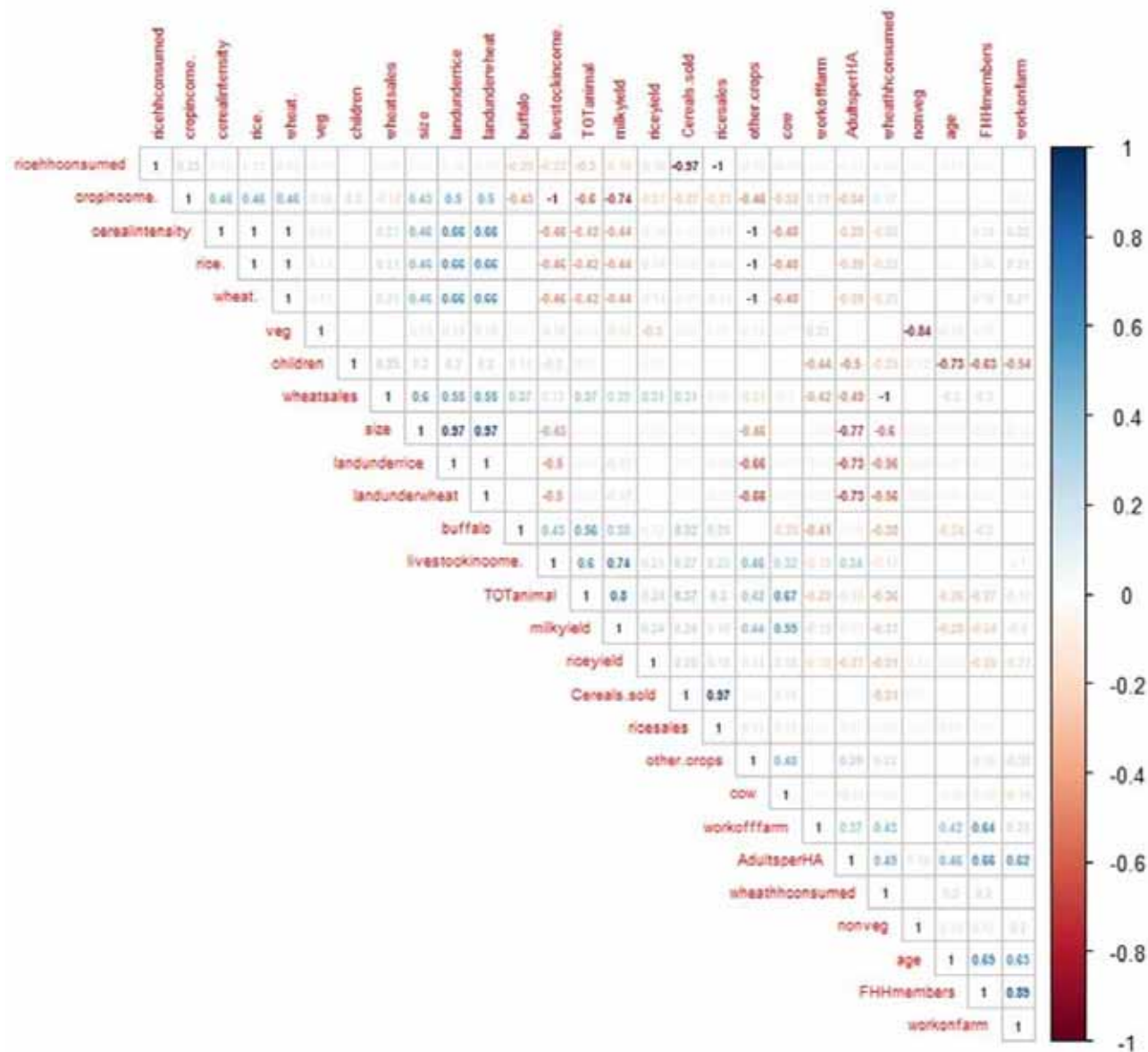
its output



The values of the Pearson's correlation on correlation plot,

```
corrplot(cormatrix,type = "upper", order="hclust",method="number",number.cex = 0.5,tl.cex = 0.5)
```

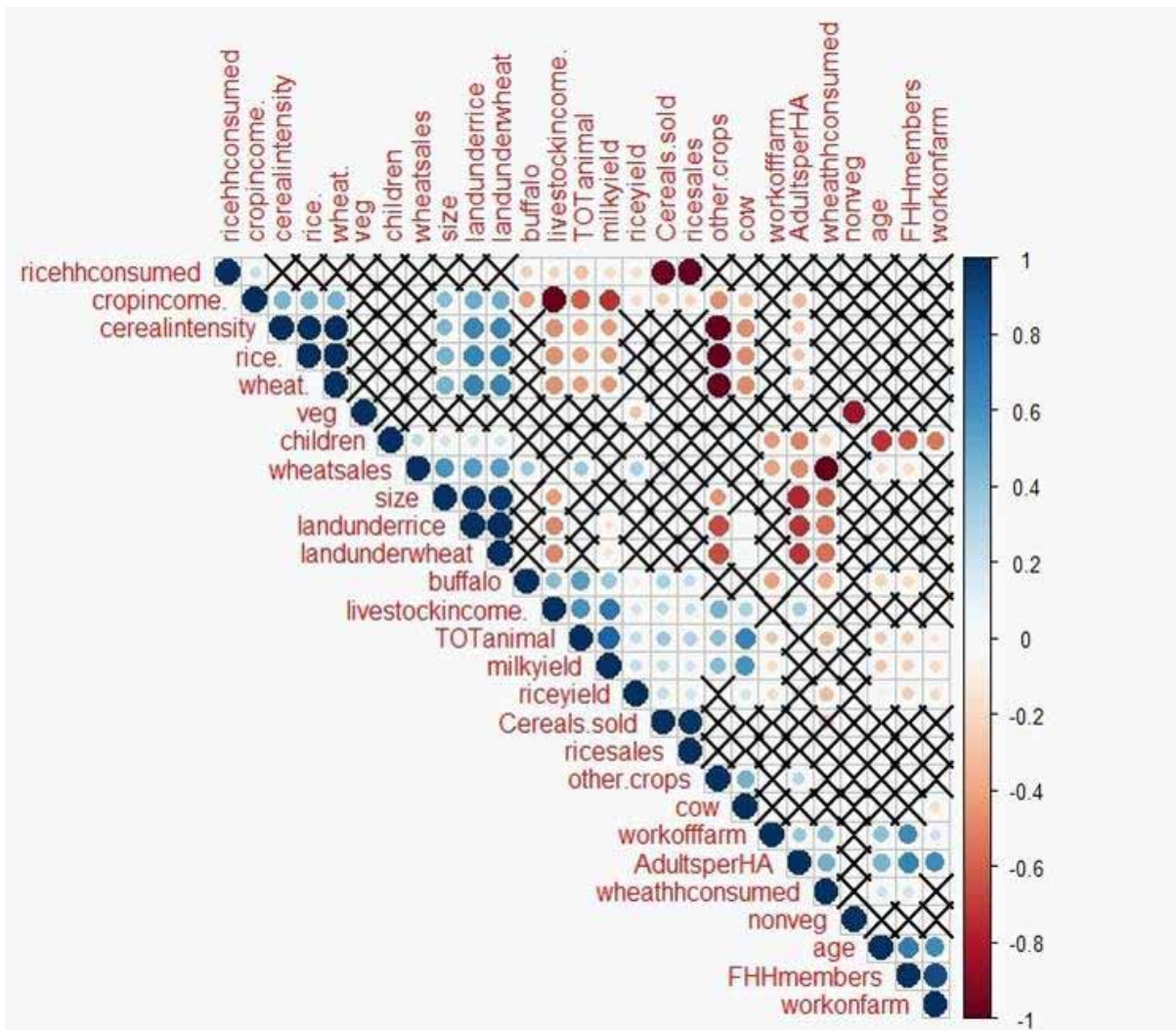
we set the method to "number" a and adjust text size, and cluster variables by its correlations output



We can cluster the variables by their correlations and cross out the non-significant running following codes

```
corrplot(corMatrix,p.mat=p.mat,type="upper",order="hclust")
```

With p.mat we supply the arguments to cross out the non-significant correlations



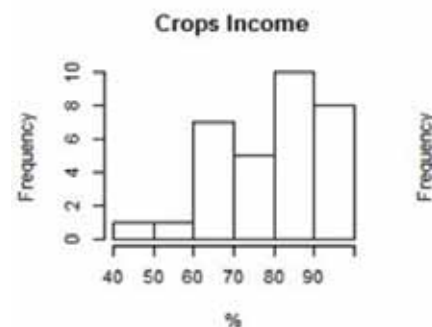
3.1.15 Choosing variables for PCA from the correlation matrix

THUMB RULE:

You must have at least 5 times households than discriminant variables (variables for PCA). In this case we have 27 variables and 32 HH. Thus we must limit ourself to select 6-7 variables for PCA. If for example, we had 150 households, we could choose to use all variables (27) for analysis, but inclusion of those highly correlated variables may add extra weight to some features and mislead its interpretation.

By inspecting the matrix visualization of the significant correlations and the clustered variables along with the histograms and boxplots we will select:

- a. Crops income
- Has a good distribution



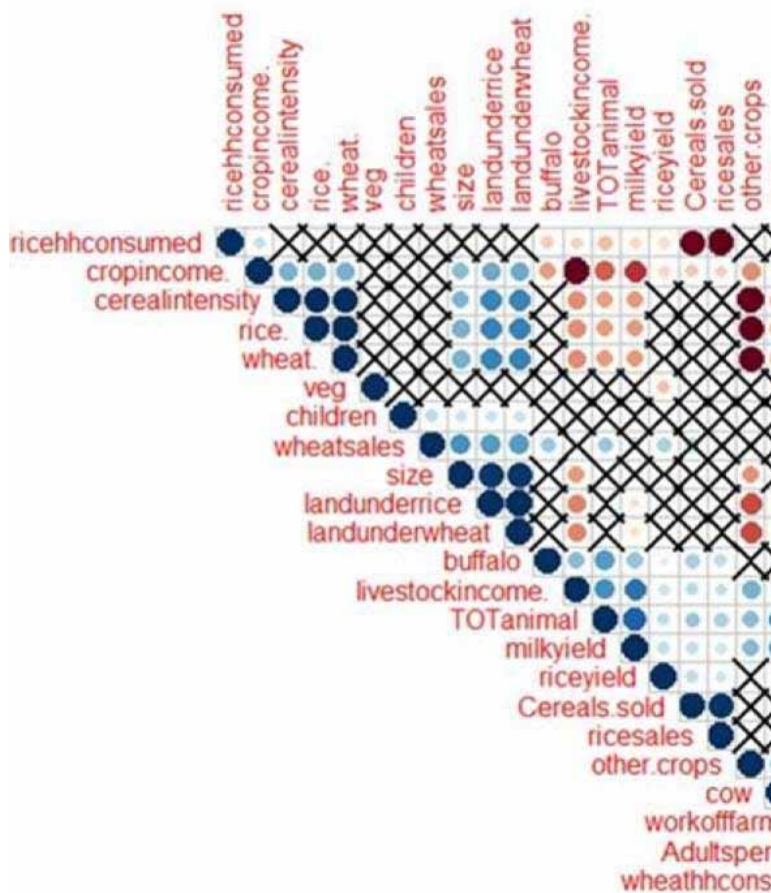
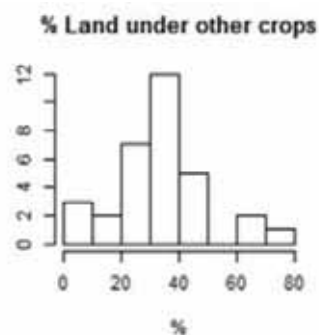
- Has a good correlation with many variables, it has a high correlation with livestock income so we will keep crops income as one variable for PCA.



b. % Land other crops

And similarly:

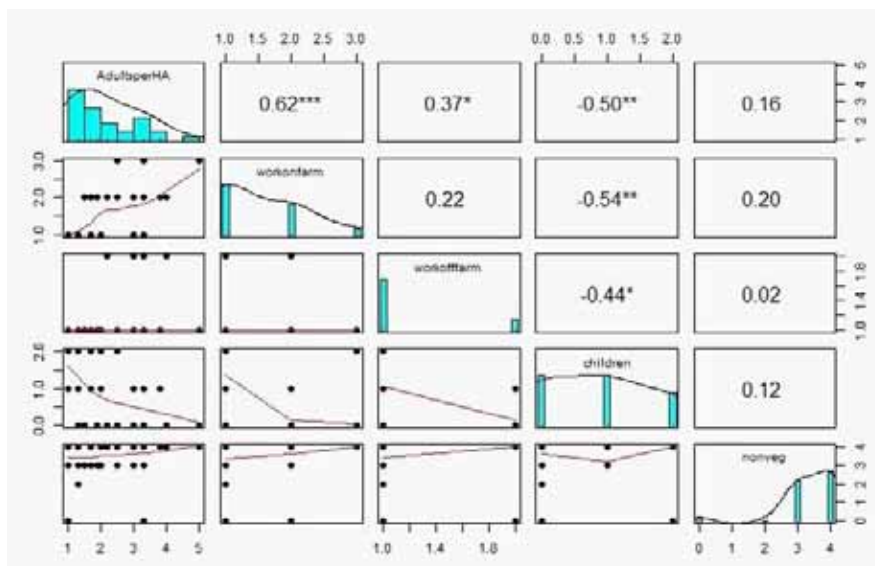
c. Wheat sales, d. Total animals, e. Size, f. Adults per HA.



BONUS: another matrix of correlation

`pairs.panels(d1[,2:10],pch=19)`

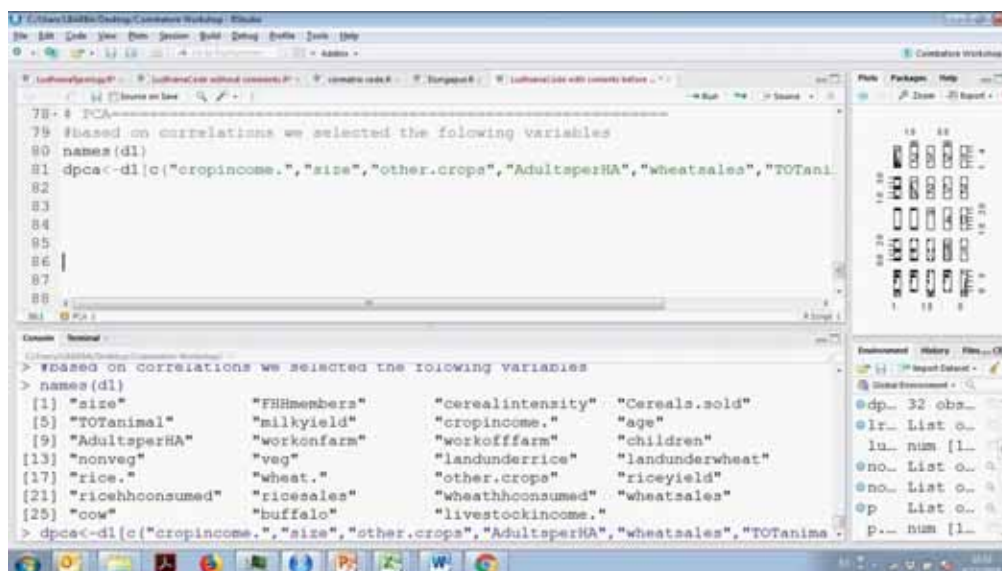
Output also shows stars with significant correlated



3.1.16 Construct a vector with the names of the selected variables for PCA analysis

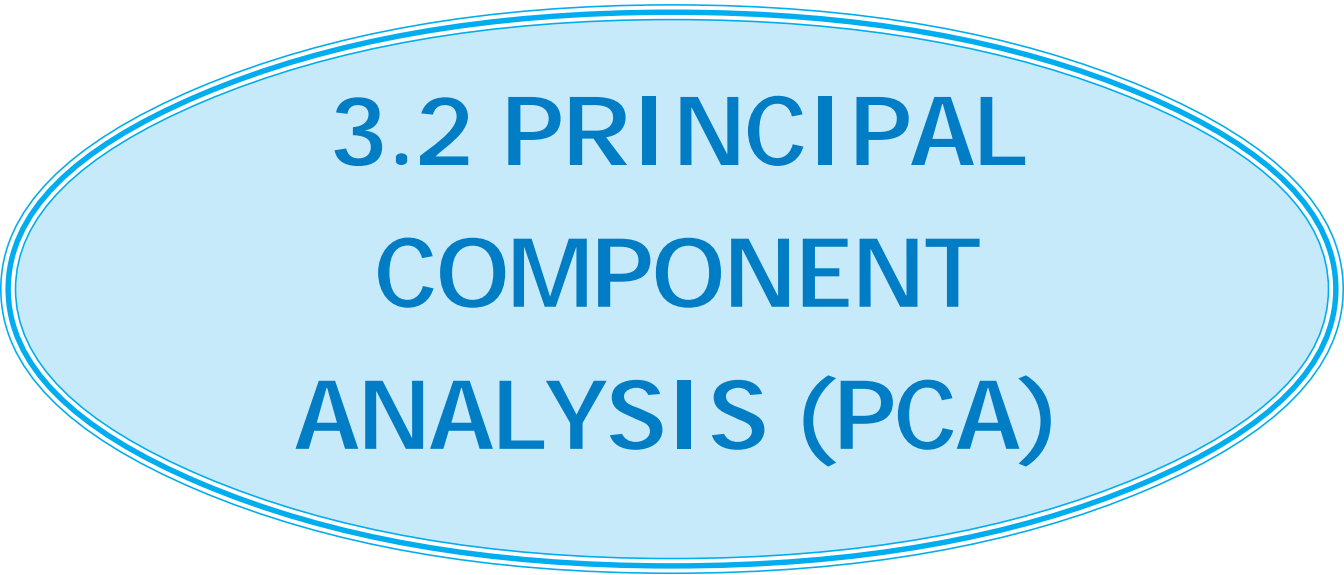
`names(d1)`

This will display the names of the variables so you can copy the exact names to avoid possible errors



A subset of d1, we will call dpca with the variables selected

`dpca<d1[c("cropincome.", "size", "other.crops", "wheatsales", "TOTanimal", "FHHmembers")]`



3.2 PRINCIPAL COMPONENT ANALYSIS (PCA)

3.2.1 Run the FIRST PCA

A first PCA should be run for us to select the number of PCs to retain after inspecting the either eigenvalue criteria (choose the PCs that have eigenvalues >1) or the scree plot test (choose a break point in which variance change, slope, from one PC to another, flattens)

Compute the FIRST PCA

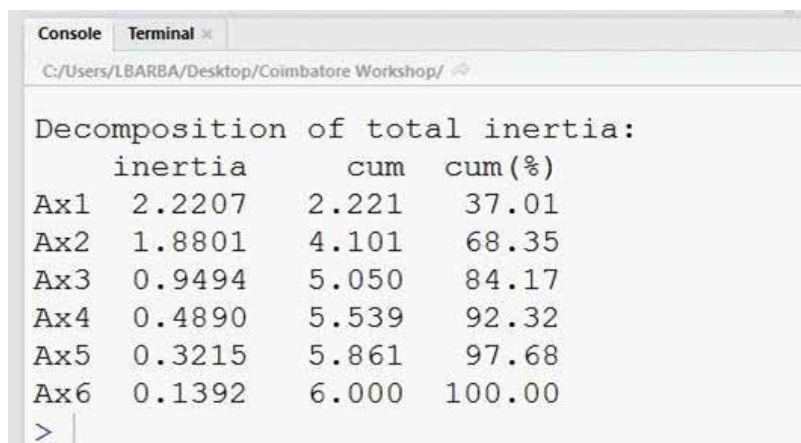
```
pca<-dudi.pca(dpca,center = TRUE, scale = TRUE,scannf = FALSE)
```

Note: Here 'pca' is the result of PCA analysis performed on the data set 'dpca' selected subset of 'd1'

Check the eigenvalues (displayed ass the column "inertia")

```
inertia.dudi(pca)
```

This shows the eigenvalues of PC's



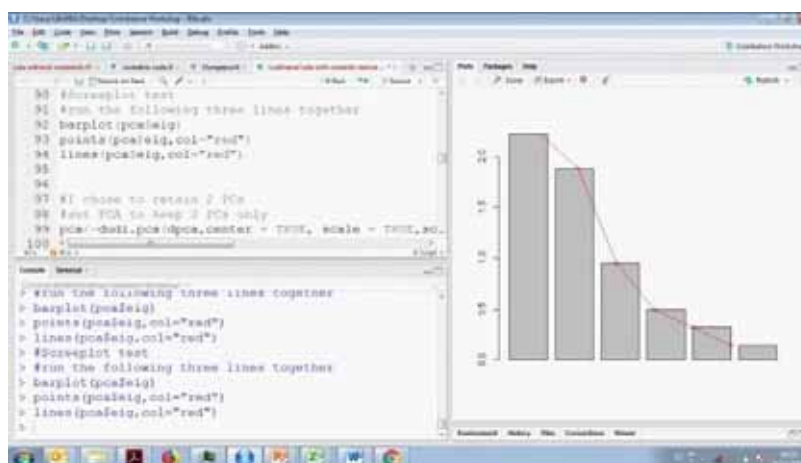
We can say that based on the eigenvalue >1 criterion, it is suggested to take only 2 PCs, this two we will retain 68.35% of the variance

Run the Screeplot test. Run the following three lines together

```
barplot(pca$eig)
```

```
points(pca$eig,col="red")
```

```
lines(pca$eig,col="red")
```



The red lines simplify the “scree” visualization, but the bars also show the eigenvalues

Note: Screeplot test suggest: 3 PCs, eigenvalues criterion: 2

Choose whatever method you want but justify it. Here you have the evidences. Sometimes both methods coincide.

we will choose 2 PCs.

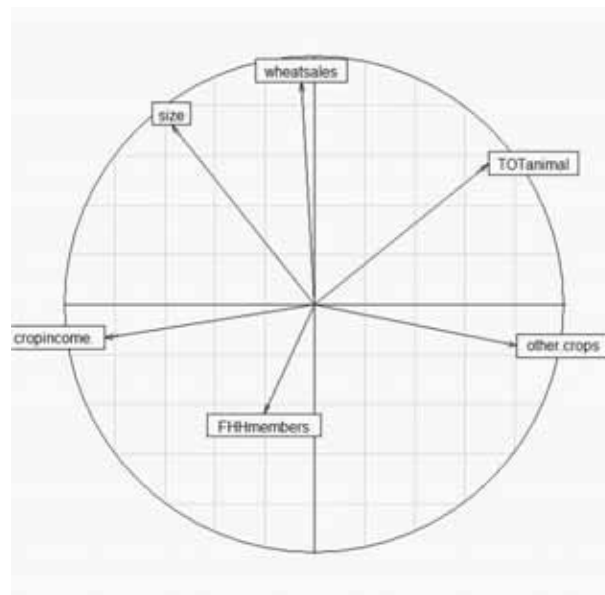
3.2.2 Run the **SECOND PCA**

As we are retaining only two PCs, we must run the PCA again so it will only store that information, as the first PCA kept the 6 PCs, one for each input variable.

```
pca<-dudi.pca(dpca,center = TRUE, scale = TRUE,scannf = FALSE,nf=2 )
```

nf=2 sets the number of PCs to retain (Remember in first PCA 6 PC were retained, now we know that 2 PC's are to be retained based on eigenvalue criteria. So put nf=2)

Now, Plot the correlation circle



Plot the HH against the PCs and the variables

```
scatter (pca,
```

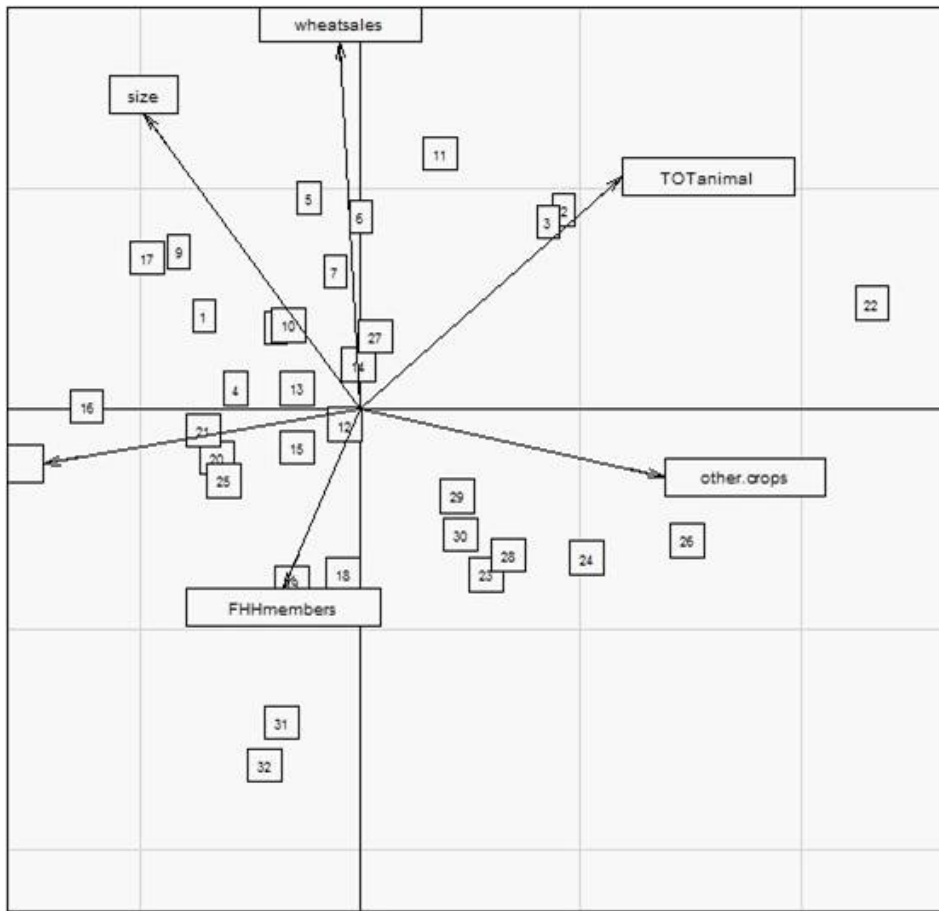
```
  posieig = "none", # Hide the scree plot
```

```
  clab.row = 0.5,   # Hide row labels
```

```
  clab.col = 0.6,
```

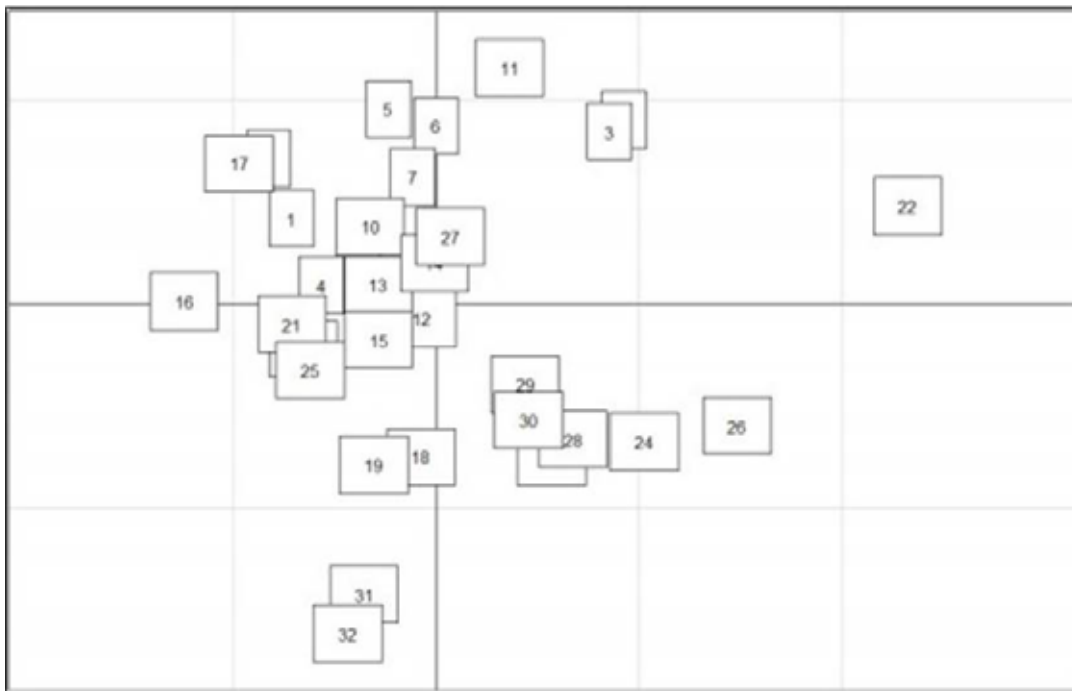
```
  xax = 1, yax = 2, # IF you had more than one PC you may like to graph PC1 PC3
```

```
  sub="PC1,PC2",box = FALSE) #change the PC name accordingly
```



Check for possible outliers

`s.label(pca$li)`



3.2.3 Access the most determinant variables on the PCs

Examine the correlation of the PCs and the variables

`pca$co`

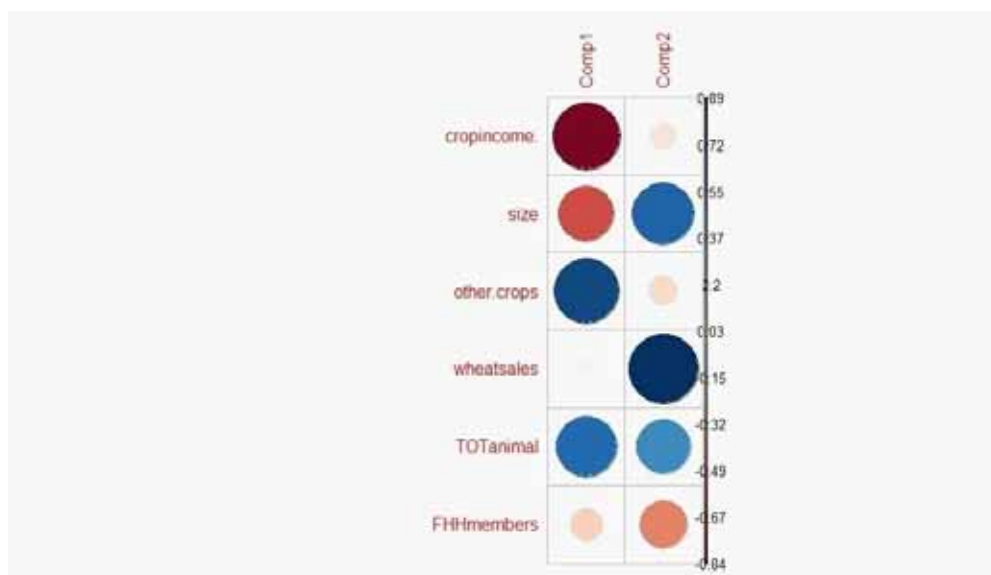
```
Console Terminal x
C:/Users/LBARBA/Desktop/Coimbatore Workshop/
> s.label(pca$li)
> # examine the most important variables
> pca$co
              Comp1      Comp2
cropincome. -0.84231918 -0.1335945
size         -0.57391425  0.7216418
other.crops  0.80889369  -0.1653083
wheatsales  -0.05375975  0.8946027
TOTanimal    0.69516828  0.5670659
FHHmembers  -0.20347969 -0.4385473
> |
```

Note: Here Comp1 = PC1 & Comp2 = PC2

Access the most determinant variables on the PCs

Visualize the most segregating variables

`corrplot (as.matrix(pca$co), is.corr=FALSE)`



Here we can conclude that:

- For PC1 the variables are Crop income and Other crops
- For PC2 Size and Wheat sales

3.3 CLUSTERING

3.3.1 Clustering

First the distance between households based on their PCs coordinates is computed

```
distHH <- dist(pca$li, method = "euclidean")
```

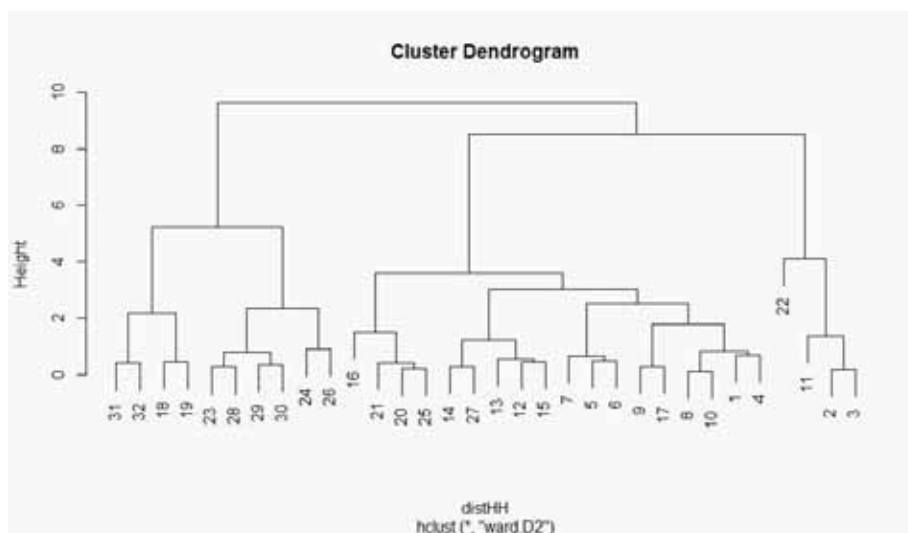
Then hierarchical clustering is performed with those distances using Ward's method:

```
dendo <- hclust(distHH, method = "ward.D2")
```

3.3.2 Visualize the dendrogram

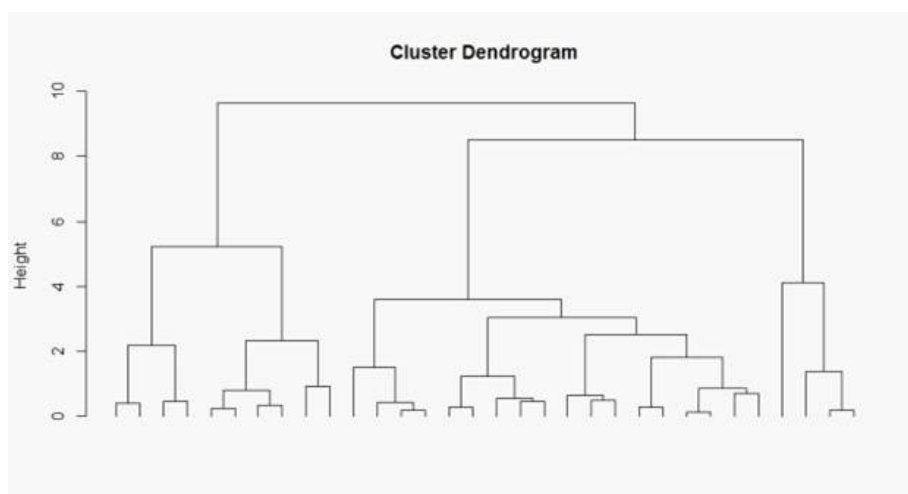
We can visualize the dendrogram with or without labels using "plot(dendo)" on clustered household euclidean distance.

`plot (dendo)`



Without labels

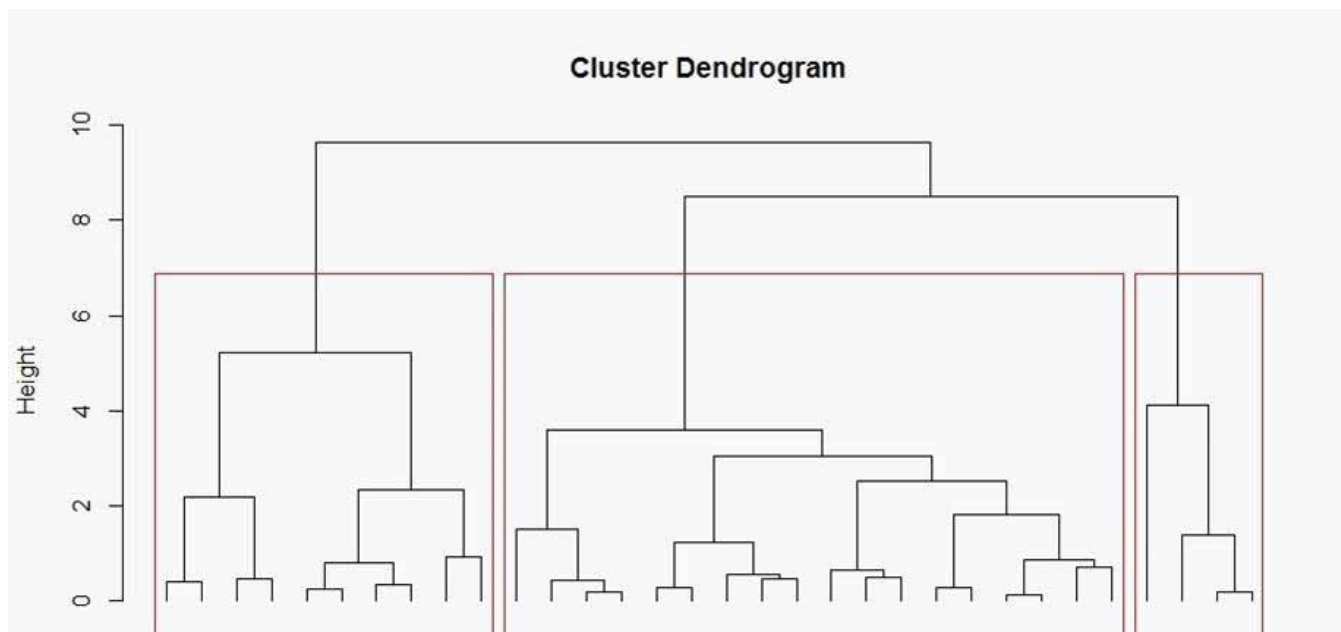
```
plot (dendo, hang=-1, ax = TRUE, ann=TRUE, xlab="", sub="", labels=FALSE)
```



Trace some rectangles surrounding possible clusters for visualization

```
rect.hclust (dendo, k=3, border="red")
```

Note: You can change the k value to see how best the number of clusters best describe the sample households and you can also play with the colour of the cluster border for example border="blue"



3.3.3 Clusters number selection

This can be done subjectively by deciding a good number of clusters by looking at the dendrogram. But we can estimate a good number of cluster based on which cluster number will minimize the within group (clusters) sum of squares (WGSS), this is how many clusters make groups, so that if we measure the distance of its members to the group mean, this is minimal. The mean here is a “multivariate mean” as we are measuring the distance based on PCs space.

First we will write a function to estimate WGSS

```
wss <- function(d) {  
  sum(scale(d, scale = FALSE)^2)  
}  
wss  
wrap <- function(i, hc, x) {  
  cl <- cutree(hc, i)  
  spl <- split(x, cl)  
  wss <- sum (sapply(spl, wss))  
  wss  
}
```

The parameter 'cl' is the result of the hierarchical clustering

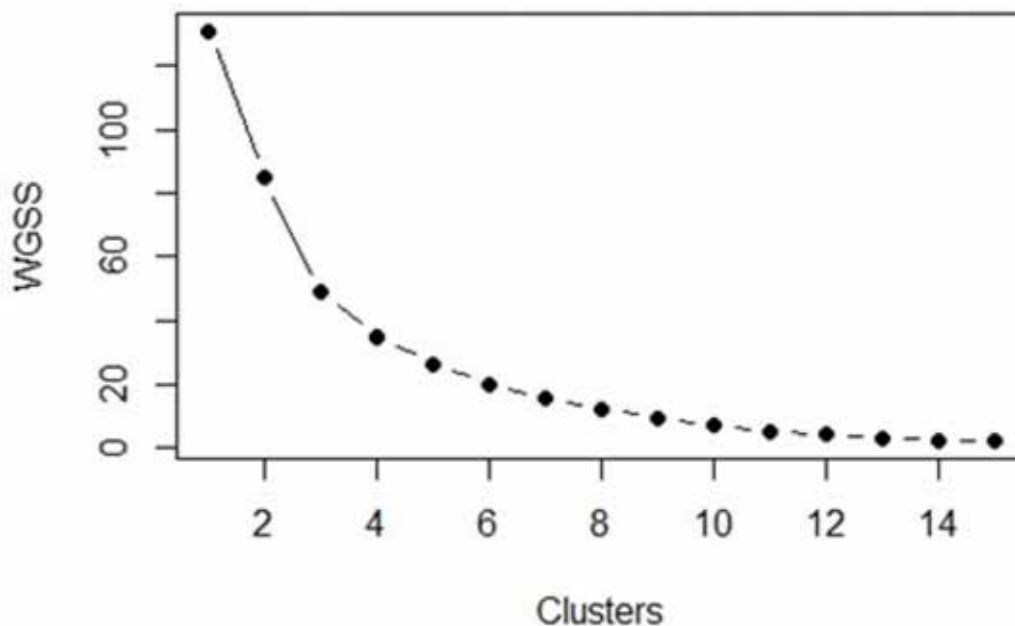
`cl<-dendo` Note: Here, 'dendo' is renamed as 'cl'

We compute the WGSS

`WGSS <- supply(seq.int(1, 15), wrap, h = cl, x = pca$li)`

Plot the number of clusters against the WGSS

`plot (seq_along(WGSS), WGSS, type = "b", pch = 19, xlab = "Clusters")`



We are looking analogously to the Screeplot for PCs selection, for a breaking point or “elbow” in which the number of clusters minimizes WGSS and that, adding a new cluster do not improve, in relative terms the WGSS

In this case 3 clusters are suggested.

Alternatives for choosing cluster number.

Provided you install the package, factoextra:

<https://cran.r-project.org/web/packages/factoextra/index.html>

and load the package: `library(factoextra)`

the following lines, will build some graphs about clusters number and the WGSS:

`fviz_nbclust(pca$li,hcut ,method = “wss”, k.max =8)`

or the silhouette method:

`fviz_nbclust(pca$li,hcut ,method = “silhouette”,k.max = 8)`

3.3.4 Cut the dendrogram with the number of clusters selected

First assign the number of clusters

```
numclust<-3
```

Cut the tree or dendrogram, this will assign each individual one cluster, **clusters** will become a categorical variable

```
clusters <- as.factor(cutree(dendo, k=numclust))
```

You can see the list

```
clusters
```

inspect the households distribution between clusters

To see the no. of households in each of the cluster.

```
table (clusters)
```

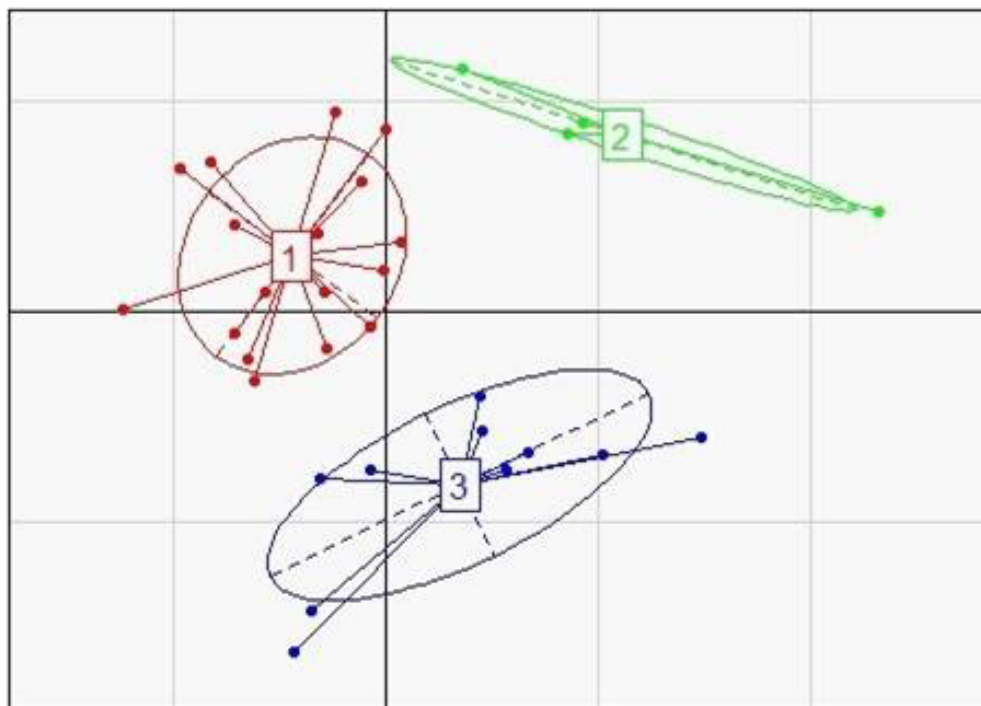
To see the % from total households in each cluster.

```
prop.table (table(clusters))*100
```

3.3.5 Plot the clusters against the PC dimensions

In our case we only have 2 dimensions to plot, so : xax=1,yax=2

```
s.class(pca$li,fac=clusters, col=rainbow(numclust),xax=1,yax=2)
```



3.3.6 Plot the PCs, the Variables and The Clusters

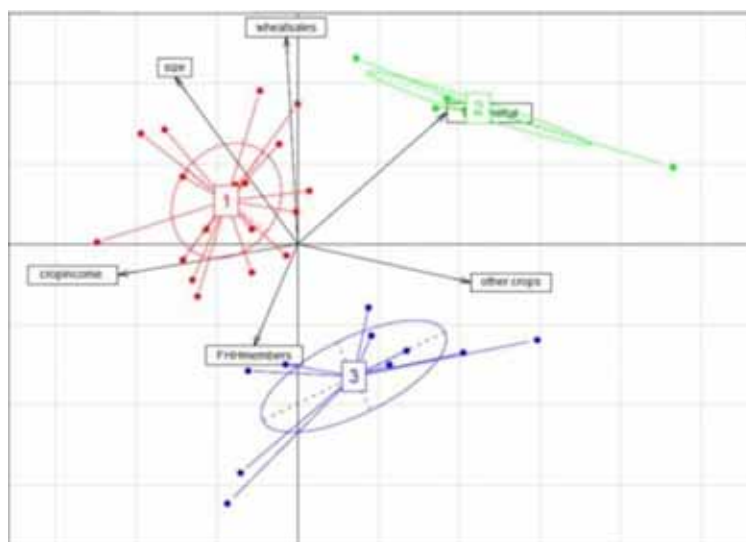
This will print the variables and PCs, changing `clab.col = 0.65` helps in fitting the labels.

```
res <- scatter(pca, clab.row = 0, clab.col = 0.65, posieig = "none")
```

Plot the clusters

```
s.class(pca$li,
        fac = clusters,
        col = rainbow(numclust),
        add.plot = TRUE,
        plot.cstar = 0.95,
        cellipse = 0.95,
        grid = FALSE
    )
```


Add onto the scatter
Remove stars
Remove ellipses



Interpretation:

An easy way to interpret the ordination is to observe the plots as gradients, where the arrows of variables point to the highest value of a certain variable, and an imaginary line pointing towards the opposite direction to the lowest value of same variable. For example, households in the left side of the plot are related with high values of the variable *cropsincome*, while households on the right side of the plot are related to low values of *cropsincome* but also with high values of the variables *othercrops* and *total animals*.

We can say that Type 1 farm HH have larger landholdings and are oriented to the commercialization of their crops, they grow mainly cereals and have possibly no livestock or very few. Type2 households could be considered as livestock farms as they are highly related with a high number of animals, and also the seem to be negatively correlated with crops income, their families are small and may grow other crops. Type 3 farm households are characterized by large household sizes, small cultivated areas, low wheat sales and lower crops income, they grow other crops than cereals.



3.4 DESCRIBING FARM TYPES

3.4.1 Types profiling

With the help of boxplots and descriptive statistics of the variables we can start constructing an argument about the most conspicuous characteristics of our farm types and also a detailed comparison of the behavior of the variables between the types. Some variables might show no evident differences, that is, all types might show similar values. But also some variables will show extreme behaviors among types.

With the help of boxplots we can visualize these behaviors.

3.4.2 Boxplots of variables vs types

We will go back to our original database d1 as we don't want to waste all our surveys precious information.

We will add a new column with the HH cluster identity

```
d1$Types<-clusters
```

And the boxplots

```
par (mfrow=c(3,3)) for (i in 1:27) {  
  boxplot (d1[,i]~d1$Types,main=codebook$Variable[i],ylab =  
  codebook$Unit[i],  
  xlab="Type",col=rainbow(numclust))  
}
```

3.4.3 Descriptive statistics by types and total sample

For the exact numbers

Of the types vs variables

```
descrTypes<-describeBy(d1[,1:27],group = d1$Type,fast=TRUE,mat = FALSE)
```

Of the variables total sample

```
descrTotal<-describe(d1[,1:27],fast=TRUE)
```

Merge both

```
descriptives<-cbind(as.data.frame.list(descrTypes),descrTotal)
```

#save it as a .csv and edit the table with the parameters you want to choose

```
write.csv(descriptives,"descriptives.csv")
```


3.4.4 Kruskal-Wallis and *post hoc* tests

To assess that the differences observed in the boxplots and in the descriptive statistics are statistically significant, we run a non-parametric ANOVA, as most probably our variables must be non-normal distributed and our types might comprehend different sample sizes.

First we run the Kruskal-Wallis over each variable vs types as factor.

```
kw<-lapply(d1[,1:27], function(x) kruskal(x ,d1$Types))
kwlst<-as.data.frame(t(sapply(kw, '[',c("statistics"))))
kwlst
```

kwlst, shows a list of the variables, the test Chi-squared statistic and the p-value of the test. So those with $p.chisq < 0.05$, are considered significant, for example:

	Chisq	Df	p.chisq
size	17.60183	2	0.0001505952
FHHmembers	5.2576	2	0.07216501
cerealintensity	17.44651	2	0.0001627567
Cereals.sold	12.70483	2	0.001742533

size is significant but FHHmembers is not

We then run a post hoc test to detect pair by pair which types are different for each variable

post hoc test Bonferroni

k=NULL

```
for (i in seq_along(1:27)){
  o=kruskal(d1[i], d1$Types, group=T, p.adj="bonferroni")$groups
  names(o)[1]<-names(d1[i])
  o$Types <- as.numeric(rownames(o))
  o<-o[order(o$Types),]
  m<-kruskal(d1[i], d1$Types, group=T, p.adj="bonferroni")$means
  a<-cbind(o[2],m[1])
  names(a)[c(1,2)]<-c(names(d1[i]),"mean")
  a$Types<- as.numeric(rownames(o))
  a<-t(a)
  k[[i]]=a
}
```

#print kruskal-wallis results in .csv, save it with another name as it does not overwrite it

```
out_file <- file("LUDHIANAKW.csv", open="a")
```

#creates a file in append mode

```

for (i in seq_along(k)){
  write.table(k[[i]], file=out_file, sep="," , dec=".", quote=FALSE, col.names=F,
    row.names=T) #writes the data.frames
}
close(out_file)

```

The output file: LUDHIANAKW.csv, show coded in low case: a,b,c, etc. the groups(types) that are significantly different, for example in our case:

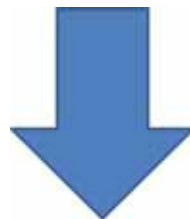
	mean	rank	std	r	Min	Max	Q25	Q50	Q75	Type
size										
a	1.61111	21.4444	0.30271	18	1.2	2	1.3	1.6	1.8	1
a	1.55	19.5	0.41231	4	1	2	1.45	1.6	1.7	2
b	0.92	6.4	0.21499	10	0.6	1.2	0.8	1	1	3
HHInemb										
a	2.66667	16	0.84017	18	2	4	2	2	3	1
a	2	9	0	4	2	2	2	2	2	2
a	3.1	20.4	0.8756	10	2	4	2.25	3	4	3
cerealsinter										
a	175.75	22.4444	13.1859	18	155.6	200	166.7	175	179.45	1
b	145	6	17.3205	4	120	160	142.5	150	152.5	2
b	154	10	12.2726	10	133.3	166.7	150	160	160	3
Cereals.sol										
a	0.90222	19.9167	0.16551	18	0.43	0.99	0.94	0.955	0.97	1
a	0.96	22.625	0.01414	4	0.94	0.97	0.955	0.965	0.97	2
b	0.887	7.9	0.0636	10	0.77	0.94	0.89	0.92	0.92	3
TOTanima										
b	2.22222	15.2778	0.64676	18	1	4	2	2	2	1
a	4.75	30.25	0.95743	4	4	6	4	4.5	5.25	2
b	2	13.2	0.66667	10	1	3	2	2	2	3

Size in ha for types 1 and 2, both are “a” that means they are no different, but type 3 is “b” that means is, significantly different from types 1 and 2, and it even lower.

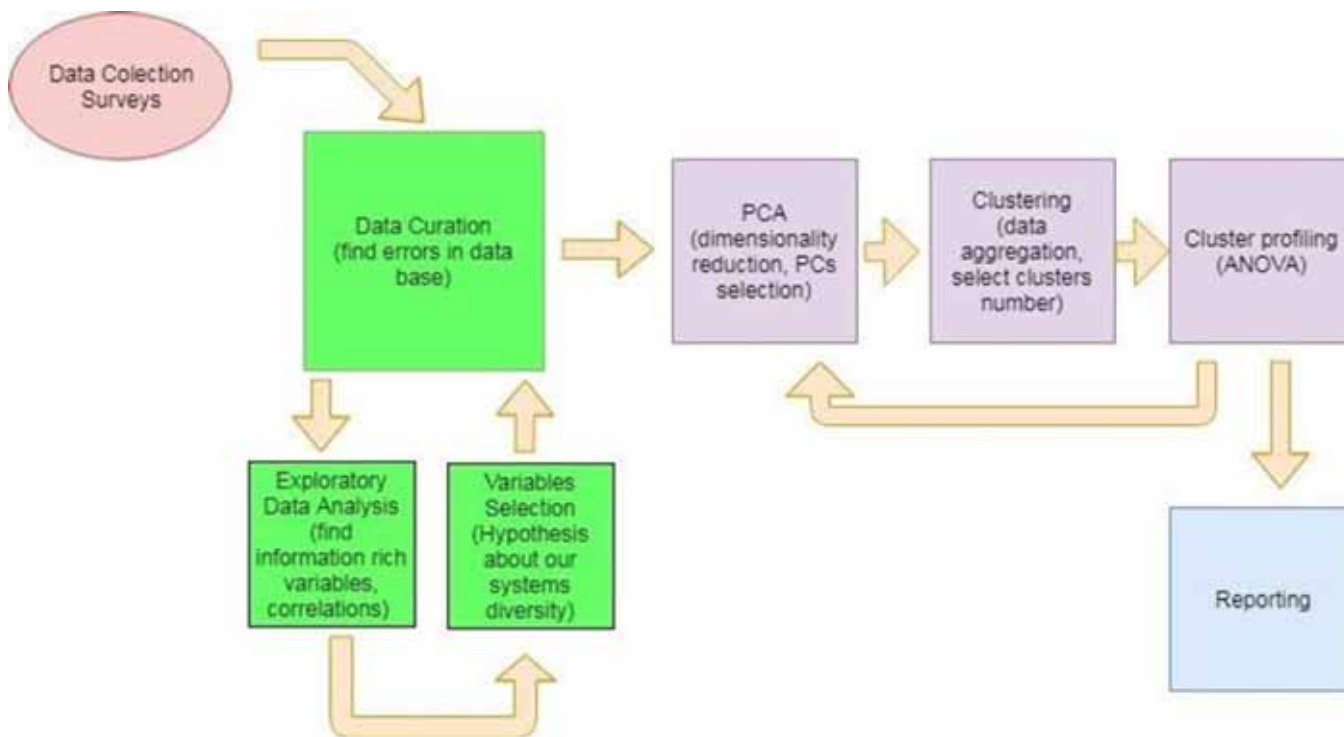
Now you can fill the table of types vs variables (Appendix I)

but....

Bear in mind that:



Farm typologies is an iterative process so, you (certainly) will need to step back to specific stages of the process. Sometimes the types will not make any sense, outliers will show, you may like to try other variables or include more variance by selecting the number of PCs based on the eigenvalues >1 criteria...



The iterative process of farm typology construction

4. References

- Alvarez, S., Paas, W., Descheemaeker, K., Tiftonell, P., Groot, J.C.J., 2014. “*Constructing typologies, a way to deal with farm diversity: general guidelines for the Humidtropics*”. Report for the CGIAR Research Program on Integrated Systems for the Humid Tropics. Plant Sciences Group, Wageningen University, the Netherlands.
- Giller, K. E., Tiftonell, P., Rufino, M. C., Van Wijk, M. T., Zingore, S., Mapfumo, P., ... & Rowe, E. C. (2011). Communicating complexity: integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. *Agricultural systems*, 104(2), 191-203.
- Goswami, R., Chatterjee, S., & Prasad, B. (2014). Farm types and their economic characterization in complex agro-ecosystems for informed extension intervention: study from coastal West Bengal, India. *Agricultural and Food Economics*, 2(1), 5.
- Groot, J. C., Oomen, G. J., & Rossing, W. A. (2012). Multi-objective optimization and design of farming systems. *Agricultural Systems*, 110, 63-77.
- Kostrowicki, J., 1977. Agricultural typology concept and method. *Agric Syst.* 2,33–45.
- Kuivanen K.S., Alvarez S., Michalscheck M., Adjei-Nsiah S., Descheemaeker K., Mellon-Bedi S., Groot J.C.J. 2016. Characterising the diversity of smallholder farming systems and their constraints and opportunities for innovation: A case study from the Northern Region, Ghana. *NJAS - Wageningen Journal of Life Sciences* 78, 153-166.
- Mahapatra, A.K., Mitchell, C.P. 2001. Classifying tree planters and non planters in a subsistence farming system using a discriminant analytical approach. *Agroforest Syst.* 52,41–52.
- Ojiem, J., Ridder, N., Vanlauwe, B., Giller, K.E., 2006. Socio-ecological niche: a conceptual framework for integration of legumes in smallholder farming systems. *Int J Agric Sust.* 4,79–93.
- Soule, M.J., 2001. Soil management and the farm typology: do small family farms manage soil and nutrient resources differently than large family farms? *Agric Resource Econ Rev* 30, 179–188.
- Tiftonell, P., Muriuki, A., Shepherd, K. D., Mugendi, D., Kaizzi, K. C., Okeyo, J., ... & Vanlauwe, B. (2010). The diversity of rural livelihoods and their influence on soil fertility in agricultural systems of East Africa—A typology of smallholder farms. *Agricultural systems*, 103(2), 83-97.

Table. Descriptive statistics of the n farm types identified

HH number	Type 1 Name		Type 2 Name		Type n Name		Total Sample		Selected for PCA
	Mean	S.Dev	Mean	S.Dev	Mean	S.Dev	Mean	S.Dev	
HH % from total									
Family HH characteristics									
HH Size									*
Household Head Age									
Non-Vegetarians									
Vegetarians									
Children									
Labor allocation									
Work on Farm									
Work off Farm									*
Lands Size and land									
Allocated to crops									
Total Land									
Land under Rice									*
Land under Wheat									
% Land under rice									
% Land under wheath									
% Land under other crops									
Cereals Intensity									*
Livestock									
Cows									
Buffalos									
TLU									*
Production									
Crops Income									
Livestock Income									
Milk Yield									
Rice Yield									*
Production allocation									
Rice consumed by HH									
Rice Sales									
Wheat consumed by HH								*	
Wheat sales									
Cereals Sold									

Appendix- II

Four training sessions in different regions of India were organized as per following programme where researchers from neighboring centres and stations joined. The calendar of the course was as follows and the program for each course is presented below followed by list of participants at each workshop.

Zone	Location	Dates	OFR centres from states	Contact Point/ Organizer from IIFSR	Local contact/ Local organizer
Western Zone	AU, Kota, Rajasthan	03-Jul Sept, 2018	Gujarat, Maharsahtra, Rajasthan	Dr. A . K. Prusty	Dr. J. P. Tetarwal
Southern Zone	TNAU, Coimbatore, TN	Oct-14 Sept, 2018	Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Telengana	Dr. N. Ravisankar	Dr. K. R. Latha
North Zone	ICAR-IIFSR, Modipuram, UP	17-21 Sept, 2019	Haryana, Himachal Pradesh, Jammu and Kashmir, Punjab, Uttar Pradesh, Uttarakhand	Dr. A . K. Prusty	Dr. A . K. Prusty
East Zone	ICAR-RC, Patna, Bihar	24-28 Sept, 2018	Assam, Bihar, Jharkhand, Chhatisgarh, Madhya Pradesh, Odisha, West Bengal	Dr. A . K. Prusty	Dr. Sanjeev Kumar Dr. Koteswar Rao

Western Zone, (AU, Kota, Rajasthan)

03-07 Sept, 2018

The training workshop on “Quantitative farming systems typologies applications with R statistical software” for OFR scientists from Western zone was organized at Agricultural University, Kota and was attended by 11 participants from 3 states namely, Gujarat, Maharashtra and Rajasthan. Dr. Luis Barba Escoto from CIMMYT, Mexico and Dr. A. K. Prusty, from ICAR-IIFSR acted as resource persons for the training. During the 5 days training programme participants were undergone both theory and practical classes on R software application for construction of typologies starting from basics like software and package download to visualization of results of their own data. The valedictory function of the training was graced by Hon’ble Vice Chancellor of AU, Kota, Prof. G. L. Keshwa on 07th September, 2018. He had highlighted the importance of this type of training for quantified analysis of benchmark information for delineating meaningful conclusions and for policy planning. On this occasion registrar and other delegates from AU, Kota were also present. The training was coordinated by Dr. J. P. Tatarwal, PI, AICRP on IFS centre, Kota. Following participants attended the training programme.

List of Participants for western zone training workshop at AU, Kota during 03-07 September, 2018 :

S.No	Name & Designation	Institution & address
1	Dr. A. K. Prusty, Scientist	ICAR-IIFSR, AICRP-IFS Coordination Unit, Meerut (U.P.)
2	Dr. Luis Barba Escoto, Scientist	CIMMYT-Mexico
3	Dr. J. P. Tatarwal, Farming Systems Agronomist	AICRP-IFS, Agricultural Research Station (AU, Kota), Ummedganj, Kota (Rajasthan)
4	Dr. L. J. Desai, OFR Agronomist	On - Farm Research Station, SDAU, Adiya, Dist. Patan(Gujarat)
5	Dr. Girish J. Patel, OFR Agronomist	Tribal Research Cum Training Centre, Anand Agriculture University, Devgad Baria (Gujarat)
6	Dr. Y. D. Charjan, OFR Agronomist	OFR Centre, RFRS, Katol, Dr.PDKV, Nagpur (Maharashtra)
7	Dr. Amol Dahiphale, OFR Agronomist	OFR-IFS, RARS-karjat , Dr BSKKV. Raigad (Maharashtra)
8	Dr. Jyotiba Kumbhar, Jr Economist	OFR-IFS, Padegaon, Satara, MPKV Rahuri (Maharashtra)
9	Prof. L. N. Dashora, OFR Agronomist	OFR-IFS, Department of Agronomy, Rajasthan College of Agriculture, MPUA & T, Udaipur (Rajasthan)
10	Dr. Babu Lal Meena, OFR Agronomist	OFR-IFS, College of Agriculture, Lalsot, Dausa SKNAU- Jobner, (Rajasthan)
11	Mr. Narottam Malav, SRF	RKVY-IFS Project, Agricultural Research Station (AU, Kota), Ummedganj, Kota (Rajasthan)
12	Dr Rewati Singh Jatav, Ag Supervisor	AICRP-IFS, Agricultural Research Station (AU, Kota), Ummedganj, Kota (Rajasthan)
13	Mr Dharmendra Suman, FA	RKVY Project, Agricultural Research Station (AU, Kota), Ummedganj, Kota (Rajasthan)



Prof. G. L. Keshwa, Hon'ble Vice Chancellor, AU, Kota, Rajasthan with participants of western zone training



Valedictory function of training workshop at AU, Kota



Training workshop in progress with hands on experience



Dr. Luis Barba and Dr. A. K. Prusty explaining the R statistical application

Southern Zone, (TNAU, Coimbatore, Tamil Nadu)

10-14 Sept, 2018

A training workshop was jointly organized by ICAR – IIFSR – CIMMYT – TNAU at the Department of Agronomy, TNAU, Coimbatore on “Quantitative farming systems typologies applications with the R statistical computing software” during Sept. 10 – 14th, 2018. The objective of the workshop series was to prepare the farming system typology of different mandate districts of on – farm research programme on AICRP on IFS. Participants from southern zone comprising of Andhra Pradesh (Dr.D. Nagarjuna, OFR (Agronomist), AICRP – IFS, ANGRAU), Telungana (Dr. Mohd. Lateef Pasha, OFR (Agronomist), AICRP – IFS, PJTSAU), Karnataka (Dr. M.T.Sanjay, OFR (Agronomist), AICRP – IFS, UAS, Bengaluru, , Kerala (Dr.D. Jacob, OFR (Agronomist), AICRP – IFS, KAU) and Tamil Nadu (Dr.N. Satheesh Kumar and Dr.D. Raja, OFR (Agronomists), AICRP – IFS, TNAU, Dr. V. Vasuki, Jr. Agronomist, Dr.K. Sathiya Bama, Jr. Soil Scientist, Dr.K.R.Latha, Chief Agronomist of AICRP –IFS, Main centre, TNAU, Coimbatore along with three local participants attended the training programme.

The programme was inaugurated on 10th, Sept. 2018 with the welcome address delivered by Dr. C. R. Chinnamuthu, Professor and Head, Dept. of Agronomy, TNAU, Coimbatore and briefed about the usage of various languages and models available for farming system research. Dr. C. Jayanthi, Director (Crop Management) in her presidential address detailed the importance of integrated farming system in today’s agriculture and the On-farm research experiments in IFS to be analysed with R language.

Dr. N. Ravisankar, Principal Scientist, IIFSR, Modipuram presided over the function and addressed the participants. During his address, he elaborated about the training series conducted at Kota, Rajasthan (Western Zone) from 03.09.2018 to 07.09.2018 and similar trainings to be organized at North zone IIFSR, Modipuram and East zone ICAR – RCER, Patna. He also emphasized the importance and applications of R language in OFR experiments of AICRP –IFS and requested the participants to utilize this opportunity and complete the assignments.

Mr. Luis Barba Escoto, System Analyst, CIMMYT, Mexico trained the participants on R statistical computing software. He personally monitored the progress of the participants and cleared their doubts then and there. The participants also co operated and showed utmost interest in learning the programme. The OFR data of individual districts of AICRP –IFS were analysed in R language. A complete package of codes were prepared by Mr. Luis Barba Escoto and handed over to the participants. He also gave his contact address for any clarifications while operating the model later.

The training workshop was completed successfully and certificates were distributed by Dr. C. Jayanthi, Director (Crop Management) during the valedictory function on 14th, Sept, 2018 at Freeman Hall, Department of Agronomy, TNAU, Coimbatore. The vote of thanks was delivered by Dr. K.R. Latha, Chief Agronomist & Professor (Agronomy), AICRP – IFS, Dept. of Agronomy, TNAU, Coimbatore.

List of Participants for Southern Zone training at TNAU, Coimbatore during 10-14 September, 2018:

S.No	Name & Designation	Institution & address
1	Dr. D. Nagarjuna, OFR (Agronomist),	AICRP – IFS, ANGRAU), Andhra Pradesh
2	Dr. Mohd. Lateef Pasha, OFR (Agronomist)	AICRP – IFS, PJTSAU), Telengana
3	Dr. M.T.Sanjay, OFR (Agronomist)	AICRP – IFS, UAS, Bengaluru, Karnataka
4	Dr. V.V. Angadi, OFR (Agronomist)	AICRP – IFS, UAS, Dharwad, Karnataka
5	Dr. D. Jacob, OFR (Agronomist)	AICRP – IFS, KAU, Kerala
6	Dr.N. Satheesh Kumar, OFR (Agronomists)	AICRP – IFS, TNAU, Tamil Nadu
7	Dr.D. Raja, OFR (Agronomists)	AICRP – IFS, TNAU, Tamil Nadu
8	Dr. V. Vasuki, Jr. Agronomist,	AICRP – IFS, Main Centre, TNAU, Tamil Nadu
9	Dr.K. Sathiya Bama, Jr. SoilScientist	AICRP – IFS, Main Centre, TNAU, Tamil Nadu
10	Dr.K.R.Latha, Chief Agronomist	AICRP – IFS, Main Centre, TNAU, Tamil Nadu



Dr. C. R. Chinnamuthu, Professor and Head, Dept. of Agronomy, TNAU, Coimbatore (left) and Dr. Luis Barba, CIMMYT, Mexico (Right) addressing the participants during inaugural session of Training workshop at TNAU, Coimbatore



Dr. C. Jayanthi, Director (Crop Management) (left) and Dr. N. Ravisankar, National PI, AICRP-IFS (Right) addressing the participants during inaugural session



Dr. Luis Barba Escoto explaining R software application for typology during training



Group photo of participants of training programme for Southern zone



Distribution of certificates to participants on successful completion of the training workshop



**Northern Zone, (ICAR-IIFSR, Modipuram, Meerut, Uttar Pradesh)
17-21 Sept, 2018**

A training workshop was jointly organized by ICAR –IIFSR – CIMMYT at ICAR-IIFSR, Modipuram, Meerut on “Quantitative farming systems typologies applications with the R statistical computing software” during Sept. 17 – 21th, 2018. The objective of the workshop series was to prepare the farming system typology of different mandate districts of on – farm research programme on AICRP on IFS. Participants from northern zone comprising of AICRP-IFS and agronomist from Jammu, Haryana, Jammu & Kashmir, Himachal Pradesh, Uttarakhand, Uttar Pradesh besides nominated members from NDUAT, Faizabad and local participants from ICAR-IIFSR attended the training programme.

The programme was inaugurated on 17th, Sept. 2018 graced by Dr. J. S. Sandhu, Hon’ble Vice Chancellor, NDUAT, Faizabad and Dr. Naredra Prakash, Director, ICAR-RC for NEH in the presence of Dr. A. S. Panwar, Director, ICAR-IIFSR and Dr. M. L. Jat, Systems agronomists, CIMMYT, India. Dr. J. S. Sandhu, Chief guest of the function in his presidential address highlighted the importance of integrated farming system in today’s agriculture and the On-farm research

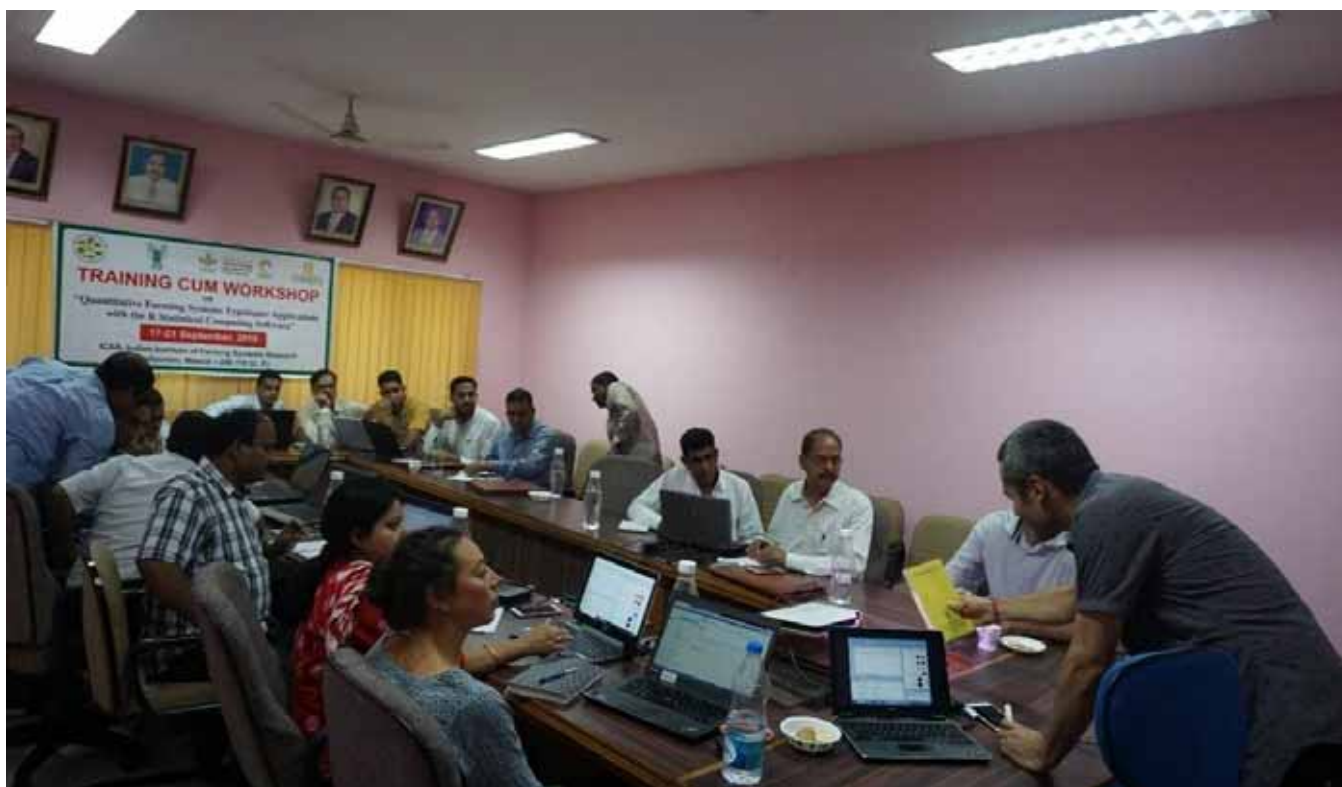
experiments in IFS to be analysed with R language. Dr. Narendra Prakash, Director, ICAR-R for NEH appreciated the initiatives for capacity building of researchers involved in farming systems in farming systems data analysis. Dr. A. S. Panwar, Director, ICAR-IIFSR suggested for quantitative analysis of IFS models developed across country and importance of software based analysis for characterization studies. Dr. M. L. Jat, Systems Agronomist, CIMMYT, India briefed about recent developments in farming systems analysis and different quantitative tools which can be used for characterization of farming systems. Dr. A. K. Prusty, Senior Scientist proposed the vote of thanks, The programme was coordinated by Dr. Poonam Kashyap. During the 5 days training programme participants were give both theoretical knowledge and hands on experience on data analysis in farming systems typology. Dr. Jeroen Groot, Systems Analyst and Dr. Roos de Adelhart, Systems Analyst from Wageningen University of Research, The Netherlands also joined the workshop and shared their experience in farming systems analysis and typology studies. The valedictory function was organized on 21st September, 2018 and participants were distributed with completion certificate.

List of Participants for North Zone training at ICAR-IIFSR, Modipuram during 17-21 September, 2018:

S.No	Name & Designation	Institution & address
1	Dr. A. K. Prusty, Scientist	ICAR-IIFSR, Meerut (U.P.)
2	Dr. Luis Barba Escoto, Scientist	CIMMYT-Mexico
3	Dr Pawan Kumar, ChiefAgronomist	AICRP-IFS, CCSHAU, Hisar
4	Dr Dinesh K. Singh , OFR Agronomist	AICRP-IFS, Pantnagar, Uttarakhand
5	Dr. A. K. Gupta, OFR Agronomist	AICRP-IFS, Chhata, Jammu
6	Dr S.K. Sharma, OFR Agronomist	AICRP-IFS, Palampur, Himachal Pradesh
7	Dr. Alok Kumar, Junior SoilScientist,	NDUAT, Faizabad
8	Sh. Ashutosh Singh, SRF	NDUAT, Faizabad
9	Dr. Neeraj Kumar	AICRP, Dry Land
10	Er. R. C. Tiwari	AICRP, Water Management
11	Dr. R. S. Yadav	AICRP, Forage Crops
12	Sh. Saurabh Dixit	AICRP, Rice
13	Dr. U.S. Tiwari , Junior SoilScientist	CSAUAT, Kanpur
14	Dr.Subash Babu, Scientist(Agronomy),	AICRP-IFS,ICAR-RC, Umiam
15	Dr Peyush Punia	Pr. Scientist, IIFSR, Modipuram
16	Dr. Poonam Kashyap	Scientist, IIFSR, Modipuram
17	Dr. Sunil Kumar	Scientist, IIFSR, Modipuram



Inaugural session of training workshop graced by Dr. J. S. Sandhu, Hon'ble Vice Chancellor, NDUAT, Faizabad and Dr. Narendra Prakash, Director, ICAR-RC-NEH



Participants of the training programme in action during the training programme



Dr. Roos de Adelhart, WUR, The Netherlands interacting with the participants of workshop



Group photo of the participants of the training programme with Chief guest at Modipuram



Distribution of certificates to trainees on successful completion of the training workshop

**Eastern Zone, (ICAR-RCER, Bihar Patna)
17-21 Sept, 2018**

A training workshop was jointly organized by ICAR-RCER–CIMMYT–IIFSR at ICAR-RCER, Patna, Bihar on “Quantitative farming systems typologies applications with the R statistical computing software” during Sept. 24 – 28th, 2018. With an aim to develop capacity building of local scientists in quantitative analysis of benchmark information and farming systems characterization using R statistical software. Participants from northern zone comprising of AICRP-IFS and agronomist from Assam, Bihar, Jharkhand, Chhatishgarh, Madhya Pradesh, Odisha and West Bengal attended the training programme.

The training programme was graced by Dr. B. P. Bhatt, Director, ICAR-RCER, Patna in the presence of Dr. M. L. Jat, Systems Agronomist, CIMMYT, India, Dr. Jeroen Groot, Farming Systems Analyst, WUR, The Netherlands and Dr. Roos de Adelhart, Systems Analyst, WUR, The Netherlands and Dr. Lusi Barba Escoto, Resource person for the training from CIMMYT, Mexico.

During the 5 days training participants were appraised about the importance of quantitative analysis of benchmark information for characterization of farming systems and application of R statistical software as a potential tool for construction of typologies. Both theory and practical hands on training were imparted to the trainees.

List of Participants for Eastern Zone training at ICAR-RCER, Patna during 24-28 September, 2018:

S.No	Name & Designation	Institution & address
1	Dr. Luis Barba Escoto, Scientist	CIMMYT-Mexico
2	Dr. Jay Shankar Borah, OFR Agronomist	AICRP-IFS, Kamrup, assam
3	Dr Sanjeev Kumar, PI	AICRP-IFS, ICAR-RCER, patna
4	Dr D. K. Mahto, OFR Agronomist	AICRP-IFS, Nalanda, Bihar
5	Dr. A. K. Netam, OFR Agronomist	AICRP-IFS, Kanker, Chhatisgarh
6	Dr Sambhu Sharan Kumar, OFR Agronomist	AICRP-IFS, East Singhbhum, Jharkhand
7	Dr. D. N. Shrivesh, OFR Agronomist	AICRP-IFS, Dindori, Madhya Pradesh
8	Dr (Mrs) Namrata Jain, OFR Agronomist	AICRP-IFS, Umaria, MP
9	Dr. Tushar Ranjan Mohanty, OFR Agronomist	AICRP_IFS, Keonjhar, Odisha
10	Dr. Bhawani Shankar Nayak, OFR Agronomist	AICRP_IFS, Kalahandi, Odisha
11	Dr. Soumitra Chatterjee, JuniorEconomist	AICRP-IFS, BCKV, Kalyani, West Bengal
12	Dr. K. Koteswar Rao, Scientist	ICAR-RCER, Patna, Bihar



← **Group photo of the participants of the training programme at Patna**

Inaugural session of typology training workshop at Patna chaired by Dr. B. P. Bhatt, Director, ICAR-RCER, Patna →



Appendix- III (Ludhiana data set)

HH ID	Size	FHH members	Cereal intensity	Cereals sold	TOT animal	Milk yield	Crop income %	Age	Adults per HA	Work on farm	Work off farm	Children	Nonveg	Veg	Land under rice	Rice %	Wheat %	Other crops	Rice yield	Rice hh consumed	Wheat hh consumed	Wheat sales	Cow	Buffalo	Livestock income %	
1	1.6	2	175	0.47	2	1100	100	35	1.3	1	1	2	4	0	1.4	1.4	88	25	3827	100	0	6	94	1	1	0
2	1.6	2	150	0.96	5	5500	70.1	33	1.3	1	1	1	3	0	1.2	1.2	75	50	5208	0	100	8	92	2	3	29.9
3	1.6	2	150	0.97	4	8250	62.8	42	1.3	1	1	0	2	0	1.2	1.2	75	50	6250	0	100	6	94	3	1	37.2
7	1.6	4	175	0.97	2	2475	84.8	45	2.5	3	1	2	4	0	1.4	1.4	88	25	5102	0	100	6	94	0	2	15.2
8	2	2	160	0.98	3	2612.5	89	42	1	1	1	2	4	0	1.6	1.6	80	40	7031	0	100	4	96	1	2	11
9	2	3	160	0.99	3	2200	73.3	50	1.5	2	1	0	3	0	1.6	1.6	80	40	7031	0	100	3	97	1	2	26.7
10	1.6	2	175	0.97	3	1925	81.3	48	1.3	1	1	1	3	0	1.4	1.4	88	25	8163	0	100	7	93	1	2	18.7
11	1.6	2	175	0.96	2	2200	84.2	52	1.3	1	1	1	3	0	1.4	1.4	88	25	7143	0	100	8	92	0	2	15.8
12	2	2	180	0.97	2	1925	92.8	42	1	1	1	2	4	0	1.8	1.8	90	20	3704	0	100	6	94	1	1	7.2
13	1.6	2	175	0.96	2	2750	81.2	36	1.3	1	1	2	4	0	1.4	1.4	88	25	4592	0	100	8	92	0	2	18.8
14	2	2	160	0.97	4	5500	70.3	40	1	1	1	1	3	0	1.6	1.6	80	40	4063	0	100	5	95	0	4	29.7
15	1.2	2	166.	0.94	2	2200	84.4	40	1.7	1	1	1	3	0	1	1	83	33.3	6800	0	100	11	89	1	1	15.6
16	1.2	2	200	0.94	2	3300	69.9	34	1.7	1	1	1	3	0	1.2	1.2	100	0	4722	0	100	11	89	0	2	30.1
17	1.2	4	200	0.95	4	3575	69.1	53	3.3	3	1	0	4	0	1.2	1.2	100	0	3958	0	100	9	91	0	4	30.9
18	1.2	3	166.	0.95	2	1375	90.6	52	2.5	2	1	0	3	0	1	1	83	33.3	6000	0	100	9	91	0	2	9.4
19	1.8	3	200	0.93	2	825	100	50	1.7	2	1	1	4	0	1.8	1.8	100	0	2469	0	100	14	86	1	1	0
20	2	2	180	0.97	2	1375	100	28	1	1	1	2	0	4	1.8	1.8	90	20	2469	0	100	6	94	0	2	0
21	1.2	4	166.	0.92	2	1925	77.5	53	3.3	2	2	0	0	4	1	1	83	33.3	3900	0	100	17	83	1	1	22.5
22	1.2	4	166.	0.92	2	1375	88.9	50	3.3	2	2	0	4	0	1	1	83	33.3	5700	0	100	17	83	1	1	11.1
23	1.8	4	155.	0.94	2	1375	100	70	2.2	2	2	0	4	0	1.4	1.4	78	44.4	3316	0	100	12	88	1	1	0
24	1.8	4	177.	0.94	2	2750	86.2	55	2.2	2	2	0	4	0	1.6	1.6	89	22.2	3594	0	100	13	88	2	0	13.8
26	1	2	120	0.94	6	8250	48.4	35	2	1	1	2	4	0	0.6	0.6	60	80	6500	0	100	13	88	4	2	51.6
27	0.8	4	150	0.94	2	1650	66.6	60	5	3	1	0	4	0	0.6	0.6	75	50	4000	0	100	12	88	1	1	33.4
28	0.6	2	133.	0.92	3	3300	80.5	38	3.3	1	1	1	3	0	0.4	0.4	67	66.7	3000	0	100	17	83	1	2	19.5
29	1.6	3	175	0.43	1	2475	82.1	47	1.9	2	1	0	3	0	1.4	1.4	88	25	4000	100	0	14	86	0	1	17.9
30	0.6	2	133.	0.92	3	2475	57.9	45	3.3	1	1	1	3	0	0.4	0.4	67	66.7	3000	0	100	17	83	1	2	42.1
31	1.2	2	166.	0.98	2	2750	76.4	32	1.7	1	1	2	4	0	1	1	83	33.3	6500	0	100	4	96	0	2	23.6
32	0.8	3	150	0.93	2	3300	65.7	40	3.8	2	1	1	4	0	0.6	0.6	75	50	5000	0	100	14	86	0	2	34.3
33	1	3	160	0.89	2	2750	65	62	3	2	1	0	3	0	0.8	0.8	80	40	4500	11	89	12	88	0	2	35
34	1	4	160	0.89	2	4125	60.1	57	4	2	2	0	4	0	0.8	0.8	80	40	4500	11	89	12	88	0	2	39.9
35	1	2	160	0.77	1	825	100	52	2	1	1	1	3	0	0.8	0.8	80	40	4600	17	83	29	71	0	1	0
36	1	3	160	0.77	1	825	100	38	3	1	2	1	4	0	0.8	0.8	80	40	4600	17	83	29	71	0	1	0

Appendix- III (Ludhiana codebook)

Acronym	Variable	Unit
size	Land Size	ha
FHHmembers	HH Size	Persons
cerealintensity	Cereals Intensity	Ratio
Cereals sold	Cereals Sold	%
TOTanimal	Total Animals	Animals
milkyield	Milk Yield	Liters/Year
cropincome%	Crops Income	%
age	Household Head Age	Years
AdultsperHA	Adults per Ha	persons/ha
workonfarm	Work on Farm	Number
workofffarm	Work off Farm	Number
children	Children	Number
nonveg	Non-Vegetarians	Number
veg	Vegetarians	Number
landunderrice	Land under Rice	ha
landunderwheat	Land under Wheat	ha
rice%	% Land under rice	%
wheat%	% Land under wheath	%
other crops	% Land under other crops	%
riceyield	Rice Yield	t/ha
ricehhconsumed	Rice consumed by HH	%
ricesales	Rice Sales	%
wheathhconsumed	Wheath consumed by HH	%
wheatsales	Wheath sales	%
cow	Cows	Number
buffalo	Buffalo	Number
livestockincome%	Livestock Income	%

Appendix- IV (Ludhiana Typology R script)

You may copy and paste the following code in Rstudio and perform the typology analysis of Ludhiana dataset.

```
# Ludhiana typology
#set the directory in which our data (dataset,codebook, etc.) is and will be stored(graphs,tables)
# IMPORTANT: choose YOUR OWN DIRECTORY
setwd("C:/Users/LBARBA/Documents/CIMMYT/Google Drive/INDIA/India Workshops/Coimbatore Workshop")
#load the packages
library(psych)# for descriptive statistics
library(corrplot)# for correlation plot visualization
library(ade4)# for PCA and CLustering
library(agricolae)# for Kruskal-Wallis test
# load the data
d<-read.csv("Ludhiana.csv")

# the codebook is a file that contains, in columns the variable acronyms,
#the variables names, and the variables units
codebook<-read.csv("codebook.csv")

str(d)# str() stands for structure of the dataframe
#drop the first column as is the HHs id's
d1<-d[,2:28]#depending on the number of columns ,2:28 could change
#check the structure again and now we have only 27 variables str(d1)
# make histograms of each variable

par(mfrow=c(3,4))
for (i in 1:27) {
    hist(d1[,i],main=codebook$Variable[i],xlab = codebook$Unit[i])
# make histograms extracting the main titles
#and x-axis labels from the codebook file

#boxplots
# we only need to change the hist() function by the boxplot()function
#in this case the y-label is the unit
par(mfrow=c(3,4))
for (i in 1:27) {
    boxplot(d1[,i],main=codebook$Variable[i],ylab = codebook$Unit[i])

```

```

# correlation matrix
corMatrix<-cor(d1)# it computes the Pearson's correlation for all pairs of variables
corMatrix #matrix of correlation values its a 27 by 27 matrix, that is
27 variables
#compute the p-value of correlations
#this is the funtion for computing
cor.mtest <- function(mat, ...) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat<- matrix(NA, n, n)
  diag(p.mat) <- 0
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], ...)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
    }
  }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}

# matrix of the p-value of the correlation
p.mat <- cor.mtest(corMatrix)# here we assign to p.mat the matix of p- values
head(p.mat[, 1:5]) # inspect the results

write.csv(p.mat,“CorpvaluesMatrix.csv”)

#correlation matrix plot
par(mfrow=c(1,1))# set the plots panel to only one image
#the correlations as circles
corrplot(corMatrix,type=“upper”)
#the correlations as numbers, export is as 10 X 10 in PDF to see the numbers
corrplot(corMatrix,type =“upper”,order=“hclust”,method
=“number”,number.cex = 0.5,tl.cex = 0.55)
#the correlation plot with significant only correlations and clustered
corrplot (corMatrix, p.mat=p.mat,type=“upper”,order=“hclust”)

# another way to present the correlation matrix
pairs.panels(d1[,9:13],pch=19,ellipses = FALSE,stars=TRUE)

# PCA=====
#based on correlations we selected the folowing variables
names(d1)

```

```

#subset d1, take only the variables in the vector and store them in dpca
dpca<-d1[c("cropincome.", "size", "other.crops", "wheatsales", "TOTanimal", "FHHmembers" )]

#FIRST pca

pca<-dudi.pca(dpca,center = TRUE, scale = TRUE,scannf = FALSE)
inertia.dudi(pca)

#Screeplot test
#run the following three lines together
barplot (pca$eig)
points (pca$eig,col="red")
lines (pca$eig,col="red")

# chose to retain 2 PCs
#set PCA to keep 2 PCs only
pca<-dudi.pca(dpca,center = TRUE, scale = TRUE,scannf = FALSE,nf=2 )
screeplot(pca)

s.corcircle(pca$co)

#Plot HH vs Variables VS PCs

scatter(pca,
        posieig = "none", # Hide the scree plot
        clab.row = 0.5,    # Hide row labels
        clab.col = 0.6,
        yax = 1, yax = 2,# IF you had more than one PC you may like to graph PC1 PC3
        sub="PC1,PC2",box = FALSE)#change the PC name accordingly

#check for outliers
s.label(pca$li)# what about hh 22?

# examine the most important variables
pca$co
# visualization of variables vs PCs
corrplot(as.matrix(pca$co), is.corr=FALSE)

# cluster
distHH <- dist(pca$li, method = "euclidean")
dendo <- hclust(distHH, method = "ward.D2")

```

```

plot(dendo)
plot(dendo, hang=-1, ax = TRUE, ann=TRUE, xlab="", sub="", labels
=FALSE)
rect.hclust(dendo, k=3, border="red")

#elbow method Within Group Sum of Squares

#first the function to compute WGSS
wss <- function(d) {
  sum(scale(d, scale = FALSE)^2)
}
wss
wrap <- function(i, hc, x) {
  cl <- cutree(hc, i)
  spl <- split(x, cl)
  wss <- sum(sapply(spl, wss))
  wss
}
#then compute WGSS for our individual with the PCA coordinates
cl<-dendo
WGSS <- sapply(seq.int(1, 15), wrap, h = cl, x = pca$li)
plot(seq_along(WGSS), WGSS, type = "b", pch = 19,xlab="Clusters")

# wgss and silhouette tests, this will give you more information to decide, #method= "wss", is the
same as WGSS in the previous lines library(factoextra)
fviz_nbclust(pca$li,hcut ,method = "wss", k.max =8)

fviz_nbclust(pca$li,hcut ,method = "silhouette",k.max = 15)
#sometimes both methods may coincide , WGSS tells to cut k=3, silhouette k=4

#choose numclust
numclust<-3
#Cut the tree or dendogram, this will assign each individual one cluster
clusters <- as.factor(cutree(dendo, k=numclust))
clusters
#inspect the HH distribution between clusters
#HH
table(clusters)
#% from total
prop.table(table(clusters))*100
# clusters against PCs
s.class(pca$li,fac=clusters, col=rainbow(numclust),xax=1,yax=2)

```

```

#Biplot clusters vs PCs
res <- scatter(pca, clab.row = 0,clab.col = 0.65, posieig = "none")
s.class(pca$li,
        fac = clusters,
        col = rainbow(numclust),
        add.plot = TRUE,    # Add onto the scatter plot
        cstar = 0.95,      # Remove stars
        cellipse =0.95,    # Remove ellipses
        grid = FALSE
)

# add a new column to thhe original data d1
# with the cluster identity for each HH
d1$Types<-clusters

#boxplots of variables vs Farm Types
par(mfrow=c(3,3))
for (i in 1:27) {
    boxplot(d1[,i]~d1$Types,main=codebook$Variable[i],ylab =
codebook$Unit[i],
           xlab="Type",col=rainbow(numclust))
}

#descriptive statistics
# of the types vs variables
descrTypes<-describeBy(d1[,1:27],group = d1$Type,fast=TRUE,mat = FALSE)
# of the variables total sample
descrTotal<-describe(d1[,1:27],fast=TRUE)

#merge both
descriptives<-cbind(as.data.frame.list(descrTypes),descrTotal)

#save it as a .csv and edit the table with the parameters you want to choose
write.csv(descriptives,"descriptives.csv")
#--- the Kruskal-Wallis test

str(d1)
#KW...Chi-squared and significance
kw<-lapply(d1[,1:27], function(x) kruskal(x ,d1$Types))
kwlst<-as.data.frame(t(sapply(kw,'[',c("statistics"))))
kwlst

```



```

# post hoc test Bonferroni
k=NULL
for (i in seq_along(1:27)){
  o=kruskal(d1[i], d1$Types, group=T, p.adj="bonferroni")$groups
  names(o)[1]<-names(d1[i])
  o$Types <- as.numeric(rownames(o))
  o<-o[order(o$Types),]
  m<-kruskal(d1[i], d1$Types, group=T, p.adj="bonferroni")$means
  a<-cbind(o[2],m[1])
  names(a)[c(1,2)]<-c(names(d1[i]),"mean")
  a$Types<- as.numeric(rownames(o))
  a<-t(a)
  k[[i]]=a
}

#print kruskal-wallis results in .csv, save it with another name as it does not overwrites it
out_file <- file("LUDHIANAKW.csv", open="a") #creates a file in append mode
for (i in seq_along(k)){
write.table(k[[i]], file=out_file, sep="," , dec=".", quote=FALSE, col.names=F, row.names=T)
  #writes the data.frames}
close(out_file)

```



हर कदम, हर उमर
 किसानों का हमसाफर
 भारतीय कृषि अनुसंधान परिषद

AgriSearch with a human touch



RESEARCH PROGRAM ON
 Climate Change,
 Agriculture and
 Food Security



RESEARCH PROGRAM ON
 Wheat

