



# Semester-long Internship Report

on

**FLOSS - R**

submitted by

**Siddhant Raghuvanshi (AITR, Indore)**

under the guidance of

**Prof. Kannan M. Moudgalya**  
Chemical Engineering Department  
IIT Bombay

**Prof. Radhendushka Srivastava**  
Mathematics Department  
IIT Bombay

and supervision of

**Mrs. Smita Wangikar**  
Project Manager,  
R Team, FOSSEE  
IIT Bombay

**Mr. Digvijay Singh**  
Project Research Assistant,  
R Team, FOSSEE  
IIT Bombay

November 07, 2021

# Acknowledgment

I would like to thank my FLOSS mentor, Prof. Radhendushka Srivastava, Department of Mathematics, IIT Bombay, for his immense support, patience, motivation, knowledge & influence throughout this internship. I want to express my sincere gratitude to Prof. Kannan M. Moudgalya, Department of Chemical Engineering, IIT Bombay, for creating the Semester-long Internship program and providing students from all over India an opportunity to participate in it. I would also like to express my heartfelt gratitude to the other mentors of the R FLOSS team, namely Mrs. Smita Wangikar and Mr. Digvijay Singh, for their guidance and valuable inputs throughout the internship and for providing me with an overview on data analysis. I also want to thank my fellow interns, Tanmay Srinath and Aboli Marathe, for their support, intellectual discussions, and enthusiasm.

# Contents

1. Introduction	4
2. Maintenance of R on Cloud	5
3. Analysis of FOSSEE workshop feedback data	8
4. Implementation and visualization of SOM algorithm in R	29
5. R Case Study: Clustering of common goods and commodities based on time-series characteristics of their Wholesale Price Index	33
6. Conclusion	41
7. References	42

# Chapter 1

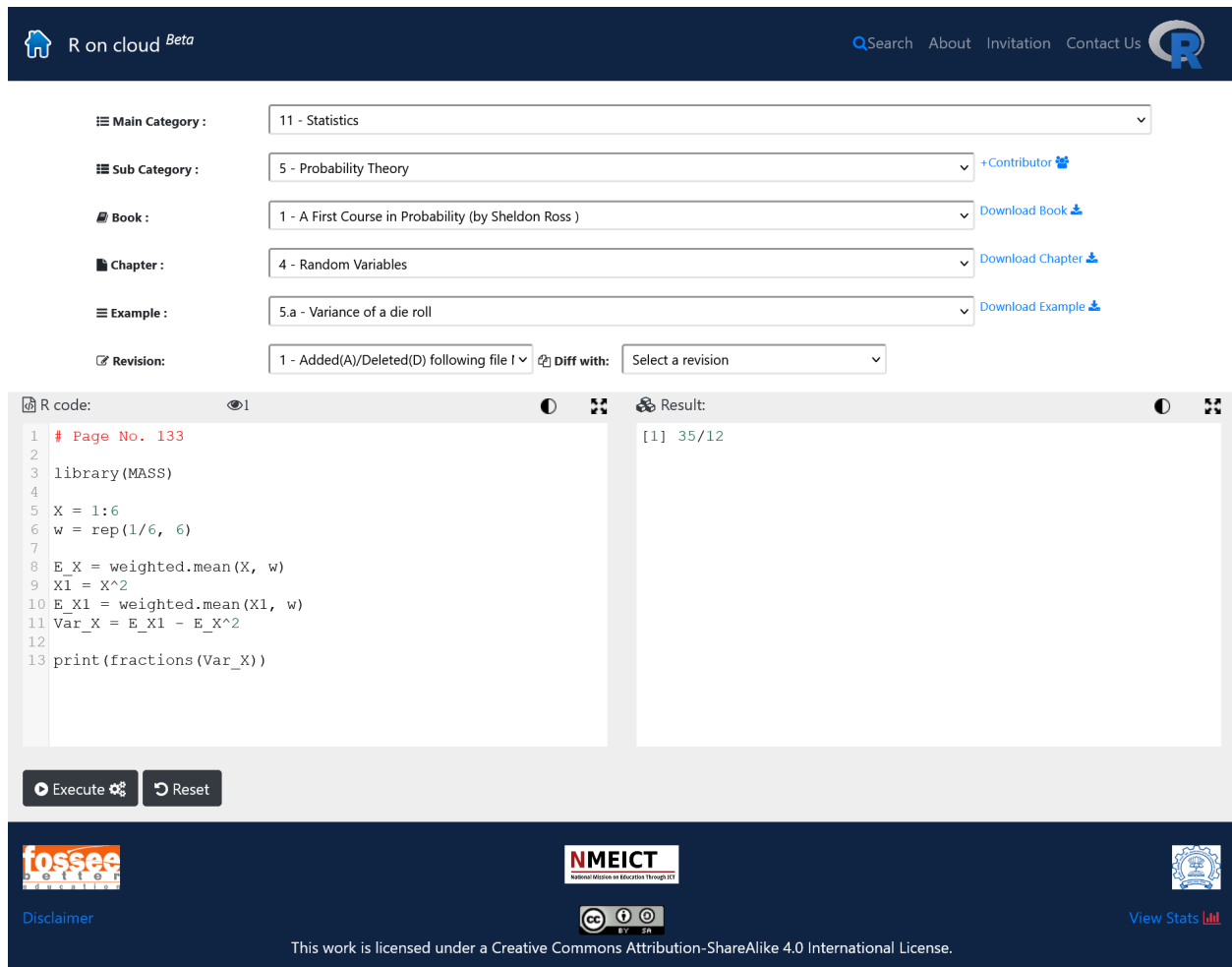
## Introduction

This report shares my contributions to open-source software made during the Semester-long Internship, starting from 7th April 2021 to 7th November 2021. Contributions were made using a FLOSS (Free-Libre/Open Source Software) known as "R" as a part of the [FOSSEE \(Free/Libre and Open Source Software for Education\) project](#) by IIT Bombay and MoE, Government of India. The FOSSEE project is a part of the National Mission on Education through ICT. The thrust area is promoting and creating open-source software equivalent to proprietary software, funded by MoE, based at the Indian Institute of Technology Bombay (IITB). The contributions include the maintenance of R on Cloud, analysis of FOSSEE workshop feedback data, implementation & visualization of SOM algorithm in R, and an R case study on clustering of common commodities based on the time-series characteristics of their Wholesale Price Index.

# Chapter 2

## Maintenance of R on Cloud

The [R on Cloud](#) is an online facility created by FOSSEE which works as a platform for executing R codes. It also allows users to interact with the codes of the [completed textbook companions \(TBCs\)](#), as shown in Figure 2.1.



The screenshot displays the R on Cloud interface. At the top, there is a navigation bar with the logo, 'R on cloud Beta', and links for Search, About, Invitation, and Contact Us. Below this, a sidebar contains filters for Main Category (11 - Statistics), Sub Category (5 - Probability Theory), Book (1 - A First Course in Probability (by Sheldon Ross)), Chapter (4 - Random Variables), Example (5.a - Variance of a die roll), and Revision (1 - Added(A)/Deleted(D) following file I). The main area is split into two panes: 'R code:' and 'Result:'. The 'R code:' pane contains the following code:

```
1 # Page No. 133
2
3 library(MASS)
4
5 X = 1:6
6 w = rep(1/6, 6)
7
8 E_X = weighted.mean(X, w)
9 X1 = X^2
10 E_X1 = weighted.mean(X1, w)
11 Var_X = E_X1 - E_X^2
12
13 print(fractions(Var_X))
```

The 'Result:' pane shows the output: [1] 35/12. Below the code editor are 'Execute' and 'Reset' buttons. The footer includes the FOSSEE logo, NMEICT logo, a Creative Commons Attribution-ShareAlike 4.0 International License notice, and a 'View Stats' link.

Figure 2.1: R on Cloud by FOSSEE.

Because of this feature, it is required to check the completed TBCs over the platform for errors by running their codes. One of the commonly encountered errors is shown in Figure 2.2.

Main Category : 11 - Statistics  
 Sub Category : 3 - Data mining  
 Book : 2 - Data Mining: Concepts and Techniques (by Jiawei Han, Micheline Kamber, and Jian Pei)  
 Chapter : 4 - Data warehousing and online analytical processing  
 Example : 4.10 - A ROLAP data store  
 Revision: 1 - Added(A)/Deleted(D) following file Diff with: Select a revision

```

R code:
1 PK.....iL.....Ex4_10/PK.....YL!
  c1...$.....Ex4_10/Ex4_10.R[0]K...?...\Dj...=...n...
2 [U...oe...7...qR...]=...3...n...s...n...yC...>X...=...7...$X@*
  S...%h...n4...C...Z...y?
  . . . . .j...("w...pd...g8N...>1... . . . .f...=...=...拆
  @...ü...xMa...P...Q...yQP...U...j...|
  c...U...~k...d...H...S...S...jR...':>4...O{...PK...@...LqW...;
  . . . . .Ex4_10/Table4_10_1.csv-ℓ  !... \...)'...1...`...F
  v...{...},1...8...TD...t"b...*...PK...L...z...9...:.....E
  x4_10/Table4_10_2.csv-ℓ
3 . . . . .^b...7...-
  '(...b...^%...ByN...*...@...y3e"...p?|4...>^PK...Q...L...Z7...8...
  . . . . .Ex4_10/Table4_10_3.csv-á
4 . . . . .^b...E)...z...Do{...t'...E...C
5 . . . . .i...^Ka...7p...|...PK...?...i...L...$.....Ex4_10/
6 . . . . .B...5...
7 . . . . .B...5...
  
```

Figure 2.2: Error when loading an R code over the platform from a zip file.

Hence, the assigned task involved checking each code file associated with 13 completed TBCs mentioned in Table 2.1 over the platform, recording the errors obtained, and forwarding the list of errors to the FOSSEE web team for correction.

Table 2.1: List of completed TBCs checked over the R on Cloud platform.

S. No.	Book Title
1	A Textbook of Electrical Engineering Materials by P. L. Kapoor, Khanna Publishers, New Delhi, 2010.
2	An Introduction to Statistical Methods and Data Analysis by RLyman Ott and Michael Longnecker, Cengage Learning, Canada,2010.
3	Applied Statistics and Probability for Engineers by Douglas C.Montgomery and George C. Runger, John Wiley and Sons, USA,2014
4	Biostatistics: Basic Concepts and Methodology for the HealthSciences by Daniel W. Wayne, Chad L. Cross, John Wiley and Sons, Singapore, 2014
5	Business Statistics For Contemporary Decision Making by KenBlack, Wiley, USA, 2010
6	Concepts Of Modern Physics by Arthur Beiser, Mcgraw-hill, New York, 2003

7	Data Mining: Concepts and Techniques by Jiawei Han, Miche-line Kamber, and Jian Pei, Morgan Kaufmann, USA, 2011
8	Elementary Statistics: A Step by Step Approach by Allan G.Bluman, McGraw-Hill, New York, 2009
9	Fundamentals of Mathematical Statistics by S.C. Gupta, V.K.Kapoor, Sultan Chand and Sons, New Delhi, 2008
10	Fundamentals of Matrix Algebra, Third Edition by GregoryHartman, CreateSpace Independent Publishing Platform, 2011
11	Introduction to Linear Algebra by Gilbert Strang, Wellesley -Cambridge Press, Wellesley MA USA, 2009
12	Introduction to Probability by Dimitri P. Bertsekas and John N.Tsitsiklis, Athena Scientific, 2008
13	Introduction to Probability and Statistics by William Menden-hall, Robert J Beaver, and Barbara M Beaver, Brooks Cole, USA, 2008

Following is the list of the type of errors encountered during the process of testing TBC codes over the R on Cloud platform -

1. Missing libraries.
2. Error when loading code from a zip file.
3. Missing R objects.
4. Runtime error.

FOSSEE web team did the following to fix the errors -

1. Installed all missing libraries over the platform.
2. Manually extracted codes from zip files and made them available over the platform.
3. Fixed code chunks causing missing R object/s error or runtime error.

# Chapter 3

## Analysis of FOSSEE workshop feedback data

### 1. Introduction

The FOSSEE project promotes the use of FLOSS tools in academia and research. It conducts regular workshops on different FLOSS to help industry professionals, faculty, researchers, and students from various institutions shift from proprietary to open-source software. These workshops are conducted throughout the year. They generally consist of spoken tutorials, live lectures, assignments, and interactive activities to engage the participants. For the assessment of a workshop's effectiveness, participants are required to fill up a feedback form at the end. The task assigned was to analyze the feedback data to identify the underlying variables called factors that can explain the interrelationships among the variables (questions) of the feedback data using a method known as EFA (Exploratory Factor Analysis) [1]. The obtained factors shall help in determining those aspects of the workshop that contributed more towards its effectiveness. Analysis began after cleaning and processing the obtained data. The complete procedure from data collection to analysis has been described in the following sections.

### 2. Data Collection

The feedback data was acquired from the [3 Day workshop on Chemistry Laboratory Experiments using ChemCollective Virtual Lab](#) conducted from 25th to 27th February 2021. A total of 210 participants attended the workshop, out of which 162 filled the feedback form. The form consisted of 32 questions related to the participants' educational background, job history, and workshop experience. Some questions consisted of sub-sections corresponding to participants' background, software used, workshop activity, and general opinions. The responses to these questions were in the form of Likert scale ratings and subjective comments. Different scales were used for recording responses depending upon the nature of the question. One of the scales used was between 1 and 5, where "1" represented "Strongly Disagree" and "5" represented "Strongly Agree".

### 3. Data Exploration

The feedback data was originally in an XLS format. It was loaded into the R environment using the "read\_excel()" function from "readxl" package [2], which belongs to the "tidyverse" ecosystem of packages [3]. A glimpse of the original dataset can be seen in Figure 3.1.



4. Generally, how comfortable are you in learning a BRAND NEW software? Please select one.	Questions related to the ChemCollective Vlabs background 5A. Do you use any chemistry virtual lab in your School/College/Institute/Organisation?	5B. If the answer to the above question is Yes, please let us know the name of the software used in your School/College/Institute/Organization.	6. For how long have you been using the above-mentioned software?	7A. For what purpose do you use this software? Please select all the relevant options.
NA	NA	NA	NA	NA
Neither comfortable nor uncomfortable (neither fast nor slo...	No	NA	less than 1 year	Research,
Somewhat comfortable (or fast) in learning a new software	No	NA	less than 1 year	Other,
Somewhat comfortable (or fast) in learning a new software	No	NA	less than 1 year	Course work - class or lab,
Extremely comfortable (or fast) in learning a new software	No	NA	less than 1 year	Course work - class or lab,

Figure 3.1: ChemCollective Virtual Lab workshop feedback data used for analysis.

The dataset initially consisted of 163 rows and 85 columns. As the feedback form questions contained sub-sections, therefore resulting in 85 distinct columns instead of 32. It was also observed that some column names were lengthy, whereas some consisted of only numbers that did not provide any information about the question associated with that column, as shown in Figure 3.2.

11. The tutorials were useful in learning the following concepts. (Kindly respond using the scale given below, wherein 1 implies "totally useless" and 5 implies "extremely useful".)	...21	...22	...23	...24	...25
Preparation of Standard Solutions	Dilutions and pH Measurements	Density of Solids and Liquids	Solubility of Salts	Heat of Reaction	Metal Displacement Reactions
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)
3	3	3	3	3	3
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)

Figure 3.2: Column names of the dataset after importing it in R.

It felt necessary to rename the columns for two reasons -

1. To convey information about the associated question in fewer words.
2. To replace the column name containing numbers or characters with information about the associated question.

Therefore the columns were renamed and grouped into the following categories, where each category has sub-divisions based on the associated feedback form questions -

1. Participants' details
2. Participants' technical background
3. Spoken tutorial
4. Tutorial topics
5. Practice problems
6. Live sessions
7. Guest lectures
8. Workshop quality feedback
9. Learning through spoken tutorial v/s conventional workshops
10. Feedback on the use of spoken tutorials
11. Miscellaneous

The following code was used to rename the columns -

```
colNames(Feedback) <- c(# Participants' details
  "Name","Institute","Audience","Age","Background","Background (other)","Name of Institute",
  "Comfortable in learning new software",

  # Participants' technical background
  "Used any chemistry virtual lab software in institute","Name of the chemistry virtual lab software in institute",
  "Duration of use","Purpose of using the software","Purpose of using the software (other)",
  "Reason to learn ChemCollective","Duration of any other chemistry virtual lab workshop (in days)",

  # Spoken tutorial
  "Spoken tutorial well made","Spoken tutorial need improvement","Spoken tutorial unclear",
  "Learned a lot from Spoken tutorial",

  # Tutorial topics
  "(topic) Preparation of standard solutions","(topic) Dilution and pH measurements",
  "(topic) Density of solids and liquids","(topic) Solubility of salts","(topic) Heat of reaction",
  "(topic) Metal displacement reactions","(topic) Buffer Solutions","(topic) Determination of Equilibrium constant",
  "(topic) Determination of solubility product","(topic) Gravimetric Analysis","(topic) Determination of pKa",

  # Practice problems
  "Spoken tutorial helpful with practice problems","ChemCollective improved lab skills",
  "Optional session helpful with practice problem","Difficulty of practice problems",

  # Live sessions
  "Serial dilution problem","Alcohol density problem","Standardization of NaOH",
  "Thermochemistry of coolant problem","DNA Binding problem","Reaction of halogen",
  "Determination of pKa of Acetic Acid (if done with practice)","Other(Live session comment)",

  # Guest lectures
  "FOSSEE and spoken tutorial by Prof. Kannan","Real v/s virtual lab Prof. Lakshmy Ravishankar",
  "Common ion effect Prof. Padmavathy","Determination of pKa of CH3COOH Dr. Gomathi Sridhar",
  "Acid Base mixture Dr. Rama Kanwar","Determination of unknown concentration Dr. Sivaranjana",

  # Workshop quality feedback
  "Quality of instruction material","Learning thorough spoken tutorial",
  "Interaction with ChemCollective team","Live Session learning","Overall quality",
  "Pace of workshop","Did not learn much","Will recommend chem-collective",
  "Will recommend similar workshop to friends","Before workshop","After workshop",
  "Most liked aspect","Most disliked aspect","Cost benefit",

  # Learning through spoken tutorial v/s conventional workshops
  "Did not give self learning opportunities","Managing a large group with strict schedule",
  "Managing fast and slow learners","Workshop for large group with less TA's",
  "Ability to conduct practical workshop with learning material","Enables to contribute content for chem-collective",

  # Feedback on the use of spoken tutorials
  "Managing a large group with strict schedule (spoken)","Managing fast and slow learners (spoken)",
  "Managing participants not fluent in english (spoken)","Managing participants with poor internet (spoken)",
  "Conduct similar workshop for other softwares (spoken)","Conduct similar workshop if spoken tutorial available (spoken)",
  "Workshop for large group with less TA's (spoken)","Contribution to content of chem-collective (spoken)",

  # Miscellaneous
  "Help required from expert/spoken tutorial forums","Only spoken tutorial forums are enough",
  "Only learning material is sufficient","What chem-collective must do next","What to do next (Comment)",
  "Software for next workshop","Want email regarding next workshop",
  "How can participant help promote chem-collective","Overall suggestion"]
```

Figure 3.3: Code to rename the feedback dataset columns.

The updated column names are shown in Figure 3.4.

```

> colnames(feedback)
[1] "Name"
[4] "Age"
[7] "Name of Institute"
[10] "Name of the chemistry virtual lab software in institute"
[13] "Purpose of using the software (other)"
[16] "Spoken tutorial well made"
[19] "Learned a lot from spoken tutorial"
[22] "(topic) Density of solids and liquids"
[25] "(topic) Metal displacement reactions"
[28] "(topic) Determination of solubility product"
[31] "Spoken tutorial helpful with paractice problems"
[34] "Difficulty of practice problems"
[37] "Standardization of NaOH"
[40] "Reaction of halogen"
[43] "FOSSEE and spoken tutorial by Prof. kannan"
[46] "Determination of pKa of CH3COOH. Dr. Gomathi Sridhar"
[49] "Quality of instruction material"
[52] "Live Session Learning"
[55] "Did not learn much"
[58] "Before workshop"
[61] "Most disliked aspect"
[64] "Managing a large group with strict schedule"
[67] "Ability to conduct practical workshop with learning material"
[70] "Managing fast and slow learners (spoken)"
[73] "Conduct similar workshop for other softwares (spoken)"
[76] "Contribution to content of chem-collective (spoken)"
[79] "Only learning material is sufficient"
[82] "Software for next workshop"
[85] "Overall suggestion"

      "Institute"
      "Background"
      "Comfortable in learning new software"
      "Duration of use"
      "Reason to learn ChemCollective"
      "Spoken tutorial need improvement"
      "(topic) Preparation of standard solutions"
      "(topic) Solubility of salts"
      "(topic) Buffer Solutions"
      "(topic) Gravimetric Analysis"
      "ChemCollective improved lab skills"
      "Serial dilution problem"
      "Thermochemistry of coolant problem"
      "Determination of pKa of Acetic Acid (if done with practice)"
      "Real v/s virtual lab Prof. Lakshmy Ravishankar"
      "Acid Base mixture Dr. Rama Kanwar"
      "Learning thorough spoken tutorial"
      "Overall quality"
      "Will recommend chem-collective"
      "After workshop"
      "Cost benefit"
      "Managing fast and slow learners"
      "Enables to contribute content for chem-collective"
      "Managing participants not fluent in english (spoken)"
      "Conduct similar workshop if spoken tutorial available (spoken)"
      "Help required from expert/spoken tutorial forums"
      "What chem-collective must do next"
      "Want email regarding next workshop"

      "Audience"
      "Background (other)"
      "Used any chemistry virtual lab software in institute"
      "Purpose of using the software"
      "Duration of any other chemistry virtual lab workshop (in days)"
      "Spoken tutorial unclear"
      "(topic) Dilution and pH measurements"
      "(topic) Heat of reaction"
      "(topic) Determination of Equilibrium constant"
      "(topic) Determination of pKa"
      "Optional session helpful with practice problem"
      "Alcohol density problem"
      "DNA Binding problem"
      "Other(Live session comment)"
      "Common ion effect Prof. Padmavathy"
      "Determination of unknown concentration Dr. Sivaranjana"
      "Interaction with ChemCollective team"
      "Face of workshop"
      "Will recommend similar workshop to friends"
      "Most liked aspect"
      "Did not give self learning opportunities"
      "Workshop for large group with less TA's"
      "Managing a large group with strict schedule (spoken)"
      "Managing participants with poor internet (spoken)"
      "Workshop for large group with less TA's (spoken)"
      "Only spoken tutorial forums are enough"
      "What to do next (Comment)"
      "How can participant help promote chem-collective"

```

Figure 3.4: Updated column names.

Data Exploration continued after updating the column names using the “skim()” function from the “skimr” package [4]. It was observed that there were several missing values in multiple columns of the dataset as shown in Figure 3.5.

```

> skim(feedback)$n_missing
[1] 1 1 1 1 1 149 13 1 1 92 46 44 152 1 1 0 0 0
[19] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0
[37] 0 0 0 0 1 158 0 0 0 0 0 0 0 0 0 0 0 1 0
[55] 0 0 0 0 54 66 1 0 0 0 0 0 0 0 0 0 0 0 0
[73] 0 22 0 0 13 15 51 120 51 51 112 112 1

> unique(skim(feedback)$n_missing)
[1] 1 149 13 92 46 44 152 0 158 54 66 22 15 51 120 112

```

Figure 3.5: Presence of missing values in the dataset.

Missing values may occur due to various reasons, such as the participants’ unwillingness to respond to optional feedback sections or the occurrence of a data formatting/reformatting error. To move ahead with the exploration, it was necessary to remove the missing values in such a way that the vital information remains unaltered. Therefore, a row and column-wise check for missing values was carried out.

After observing the results of the row-wise check, it was found that the first row contained the maximum number of missing values, i.e., 31. Further examination of the first row entries showed that it was not a valid feedback response as it did not even contain the participant’s name, as shown in Figure 3.6. Hence it was removed.

```

> # Row-wise missing values check
> feedback %>% summarise(na_rows = rowSums(is.na()))
[1] 31 15 9 16 10 14 14 11 12 7 9 10 7 9 9 10 14 8 12 10 13 10 9 14 6 11 6 8 10 11 3 10 10 8 8 8 13 5 8 9 7 9 8 4 12 10 10 9 12 9 10 8 16 4 1 11 5 9 10 9 11 6 11 3 8 10 11 8 9 15 11 0 8
[75] 8 4 6 11 3 10 12 8 4 5 6 3 8 11 8 4 0 4 10 2 6 3 8 9 7 7 6 11 7 7 6 9 9 11 5 10 6 5 5 8 11 9 8 5 6 3 8 5 0 11 9 11 3 6 4 11 7 8 6 11 10 5 3 6 6 12 9 9 6 8 9 11 8 8
[149] 8 8 6 7 10 3 7 3 8 6 4 7 8 9 8

> # Finding the column indexes for each row containing NA
> NA_rows <- apply(as.data.frame(t(apply(feedback, 1, is.na))), 1, function(x){which(x)})
> NA_rows[[1]]
      Name
1
      Institute
1
      Audience
3
      Age
4
      Background
6
      Name of Institute
7
      Comfortable in learning new software
9
      Name of the chemistry virtual lab software in institute
10
      Duration of use
11
      Purpose of using the software (other)
12
      Reason to learn ChemCollective
14
      Duration of any other chemistry virtual lab workshop (in days)
15
      Spoken tutorial helpful with practice problems
16
      ChemCollective improved lab skills
18
      Optional session helpful with practice problem
21
      Difficulty of practice problems
23
      Determination of pKa of Acetic Acid (if done with practice)
25
      Other(Live session comment)
28
      Pace of workshop
34
      Most liked aspect
42
      Cost benefit
54
      What chem-collective must do next
60
      Most disliked aspect
61
      Software for next workshop
82
      Want email regarding next workshop
85
      How can participant help promote chem-collective
84
      Overall suggestion
85

```

Figure 3.6: Row-wise examination of missing values.

Column-wise examination showed that most of the columns containing missing values were non-numeric and contained subjective comments from the participants, as shown in Figure 3.7. Hence the columns were kept unchanged.

	Most liked aspect	Most disliked aspect	How can participant help promote chem-collective	Overall suggestion
1	NA	NA	NA	NA
2	Very interactive	Had difficulty in uploading Java	NA	NA
3	NA	NA	NA	NA
4	The faculty were very patient in their approach to solve eac...	None	NA	NA
5	NA	NA	NA	NA

Figure 3.7: Examining missing values column-wise.

An interesting observation was made during the examination of columns containing missing values, i.e., for some columns, the data entries seem out of place, as shown in figure 3.8.

Help required from expert/spoken tutorial forums	Only spoken tutorial forums are enough	Only learning material is sufficient	What chem-collective must do next
Yes	NA	NA	NA
Yes	By demonstrating	Can include salt analysis	NA
No	NA	NA	NA
Yes	By sharing my experience with other faculty and using the s...	The workshop was very effectively conducted.	NA
Yes	NA	NA	NA
No	NA	NA	NA
Yes	to train our school teachers	NA	NA
Yes	NA	NA	NA
5	4	5	Create more Spoken Tutorials on topics in ChemCollective V...

Figure 3.8: Columns with erroneous entries.

Columns that should contain only string values also had numeric responses. This information was communicated to the mentor. The mentor suggested removing such columns from the dataset. Therefore, the last 23 columns, from “Did not give self learning opportunities” to “Overall suggestion”, were removed.

After dealing with missing values and columns with incorrect data, the dataset was also checked for duplicate entries as they can introduce bias in statistical analyses [5,6]. The Approximate String Matching (Fuzzy Matching) technique was applied over the “Name” data column using the “agrep()” function by Brian Ripley and Kurt Hornik provided in the “base” package of R [7] to check for similar participant names. If matching names are found, then other background details of those participants like their institution name, educational background and profession were examined. In Figure 3.9, it is shown that the result of string matching turns out to be non-zero, indicating the presence of matching names.

```
> name <- feedback[, "Name"]
> unique(lengths(lapply(name, agrep, name, value=TRUE)) - 1)
[1] 0 1 2 3
```

Figure 3.9: String matching results for the “Name” column.

On further examination of the data associated with the matching names, it was found that there were no duplicate entries.

The data exploration process shed light on the structure and format of the original feedback data. It also helped in identifying and removing some errors. It is followed by the data cleaning process as described in the subsequent section.

## 4. Data Cleaning

Data cleaning is the most crucial step performed before analysis, as any result obtained from incorrect data will be unreliable. It involves the steps for identifying and removing erroneous and mislabelled data [8,9]. There is a possibility of incorrect responses in the feedback data due to several reasons, such as inattentiveness of participants while filling the feedback form and lack of understanding of the questions asked. Therefore, it is necessary to carefully examine the data and remove all misleading responses to preserve its reliability.

For cleaning the data, all possible ambiguities in it were systematically checked and recorded. Grouping the responses into categories (performed during data exploration) made the checking process easier. The complete procedure can be broadly divided into three main steps with multiple substeps as listed below -

### 1. Checked participants’ backgrounds and opinions regarding ChemCollective and similar tools -

- Checked for entries where a participant first selected "Yes" when asked whether he/she ever used any chemistry virtual lab software in his/her institute but did not provide the name of the software in the column “Name of the chemistry virtual lab software in institute”.
- Checked for entries where a participant first selected "No" when asked whether he/she ever used any chemistry virtual lab software in his/her institute but added the name of a software in the column “Name of the chemistry virtual lab software in institute”.

Used any chemistry virtual lab software in institute	Name of the chemistry virtual lab software in institute
No	NA
No	NA
No	NA
No	NA
Yes	Olabs

Figure 3.10: Columns with entries related to participants' experience with chemistry virtual lab software.

- Checked for entries where a participant initially responded with “Other” when asked regarding the purpose of using the chemistry virtual lab software but later did not mention any purpose under the column “Purpose of using the software (other)”.

Purpose of using the software	Purpose of using the software (other)
Other,	Virtual lab
Course work - class or lab,	NA
Course work - class or lab,	NA
Course work - class or lab,	NA
Other,	NA

Figure 3.11: Columns indicating participants' purpose of using a chemistry virtual lab software.

## 2. Checked the qualitative and quantitative feedback responses regarding the procedure and quality of the workshop -

- Checked if the level of knowledge regarding ChemCollective for any participant dropped after the workshop, as it is improbable.

Before workshop	After workshop
1	4
1	4
1	2
1	5
1	3

Figure 3.12: Entries associated with the knowledge of ChemCollective.

- Checked for contradicting responses in columns associated with the quality and effectiveness of the workshop; for example, searched and recorded all such entries where a participant had given a high score in both the “Did not learn much” and “Will recommend similar workshop to friends” columns.

Spoken tutorial well made	Spoken tutorial unclear	Did not learn much	Will recommend chem-collective	Will recommend similar workshop to friends
5	5	1	4	4
2	4	1	5	5
3	3	4	3	3
5	5	5	5	5
5	1	1	5	5

Figure 3.13: Column entries associated with the workshop’s quality and effectiveness.

### 3. Removed misleading entries.

- After recording all potentially misleading entries, the mentor was consulted. As per his advice, the row-wise frequency of misleading responses was calculated and then those row entries for which the frequency value was more than one were examined. The “table()” function of R was used to calculate the frequency, as shown in Figure 3.14.

```
> remove <- c('Row entries I', 'Row entries II', 'Row entries III', 'Row entries IV', 'Row entries V', 'Row entries VI')
> Table <- table(remove)
> Table
remove
 1  4  6  7  8  9 14 15 17 18 22 24 25 27 31 33 40 43 59 60 68 73 91 93 109 110 121 125 128 137 147 148 155
 2  3  3  3  3  1  2  3  2  2  3  2  1  1  1  1  4  1  2  1  1  4  2  2  2  1  2  2  1  1  2  1  3
> rowentries <- c(as.numeric(names(Table)[which(Table>1)]))
> rowentries
[1] 1 4 6 7 8 14 15 17 18 22 24 40 59 73 91 93 109 121 125 147 155
```

Figure 3.14: Row entries selected for further examination.

- After careful examination of the remaining twenty-one selected row entries, as shown in Figure 3.14, the mentor suggested removing all of them from the feedback data.

After the removal of misleading responses, the dataset was left with 141 rows and 62 columns.

## 5. Data Pre-processing

As only the Likert scale-based columns were required to apply EFA over the dataset, it was necessary to filter them out [10,11]. While filtering the columns, it was observed that some column entries contained both numeric and string values, as shown in Figure 3.15.

(topic) Preparation of standard solutions	(topic) Dilution and pH measurements	(topic) Density of solids and liquids	(topic) Solubility of salts	(topic) Heat of reaction	(topic) Metal displacement reactions
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)
3	3	3	3	3	3
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)
5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)	5 (extremely useful)

Figure 3.15: Data entries containing both numeric and string values.

Therefore, removing the string portion of those data entries was necessary. Also, all dataset columns were checked for missing values. If any column is found with missing values, it is removed from the dataset. After making all the required changes and filtering the necessary data columns, the remaining dataset contained 141 rows and 48 columns, where each column had the numeric data type.

## 6. Data Analysis

The data analysis was performed using the “EFAtools” package [12]. The “N\_FACTORS()” function from “EFAtools” was used to find the suitable number of factors in the data by finding its correlation matrix. The “N\_FACTORS()” function tested the suitability of the correlation matrix for EFA by applying “Bartlett’s test of sphericity” over it and calculating its “Kaiser-Meyer-Olkin criterion (KMO)” value. Bartlett’s test of sphericity statistically tests the hypothesis that the correlation matrix contains ones on the diagonal and zeros on the off-diagonals. This test should produce a statistically significant chi-square value to justify the application of EFA [13]. The KMO value indicates the proportion of variance in the variables that might be caused by underlying factors [14]. The “N\_FACTORS()” function calculates the appropriate number of factors for the given data only when it obtains a favorable result from Bartlett’s test and a suitable KMO value. Before applying the “N\_FACTORS()” function, the values associated with the column “Overall quality” were stored separately and the column was removed from the dataset. It was done to later perform a regression analysis over the obtained factors and the “Overall quality” variable. Unfortunately, the “N\_FACTORS()” function gave an error, as shown in Figure 3.16.

```
> N_FACTORS(feedback_selected_for_analysis)
✖ ( ) ( ) ( ) ( ) ( ) ( ) Running CDError in eigen(R_samp, symmetric = TRUE, only.values = TRUE) :
infinite or missing values in 'x'
In addition: Warning message:
In stats::cor(samp, method = cor_method) : the standard deviation is zero
```

Figure 3.16: Error given by the “N\_FACTORS()” function.



The error was fixed by removing column number 22, i.e., “Optional session helpful with practice problem”, from the pre-processed data. The “N\_FACTORS()” function was executed again over the remaining data and it gave the following results.

```
> N_FACTORS(feedback_selected_for_analysis)
(*) (*) ✘ ( ) ( ) ( ) ( ) Running HULLi Only CAF can be used as gof if method "PAF" is used. Setting gof to "CAF"

(*) (*) (*) (*) (*) (*) (*) Done!

-- Tests for the suitability of the data for factor analysis -----
Bartlett's test of sphericity
√ The Bartlett's test of sphericity was significant at an alpha level of .05.
  These data are probably suitable for factor analysis.

 $\chi^2(1035) = 5458.57, p < .001$ 

Kaiser-Meyer-Olkin criterion (KMO)
√ The overall KMO value for your data is marvellous with 0.903.
  These data are probably suitable for factor analysis.

-- Number of factors suggested by the different factor retention criteria -----
( ) Comparison data: 1
( ) Empirical Kaiser criterion: 2
( ) Hull method with CAF: 1
( ) Hull method with CFI: NA
( ) Hull method with RMSEA: NA
( ) Kaiser-Guttman criterion with PCA: 10
( ) Kaiser-Guttman criterion with SMC: 6
( ) Kaiser-Guttman criterion with EFA: 4
( ) Parallel analysis with PCA: 3
( ) Parallel analysis with SMC: 6
( ) Parallel analysis with EFA: 4
( ) Sequential  $\chi^2$  model tests: 20
( ) Lower bound of RMSEA 90% confidence interval: 13
( ) Akaike Information Criterion: 17
```

Figure 3.17: Result obtained from the “N\_FACTORS()” function.

Three scree plots associated with PCA-determined, SMC-determined and EFA-determined eigenvalues were also obtained from the analysis, as shown in Figures 3.18, 3.19 and 3.20.

**Scree plot with PCA-determined eigenvalues**

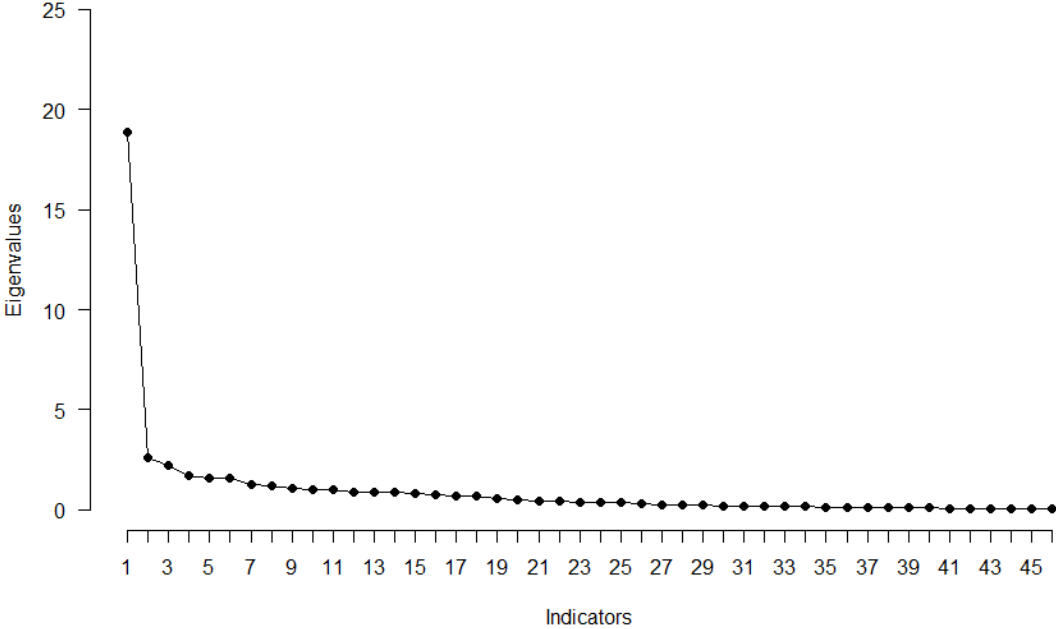


Figure 3.18: Scree plot with PCA-determined eigenvalues.

**Scree plot with SMC-determined eigenvalues**

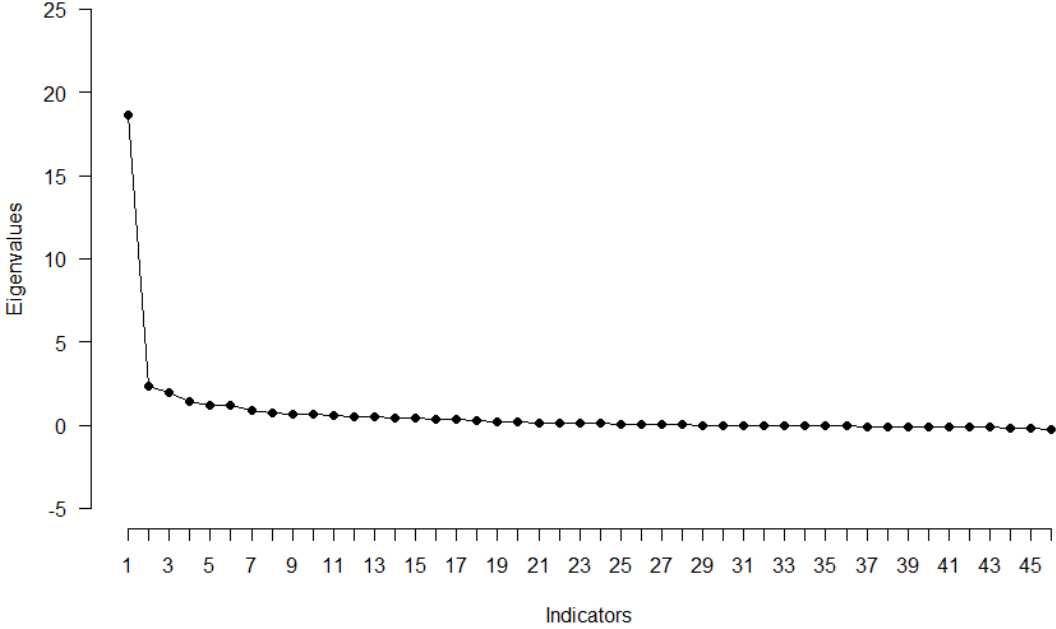


Figure 3.19: Scree plot with SMC-determined eigenvalues.

Scree plot with EFA-determined eigenvalues

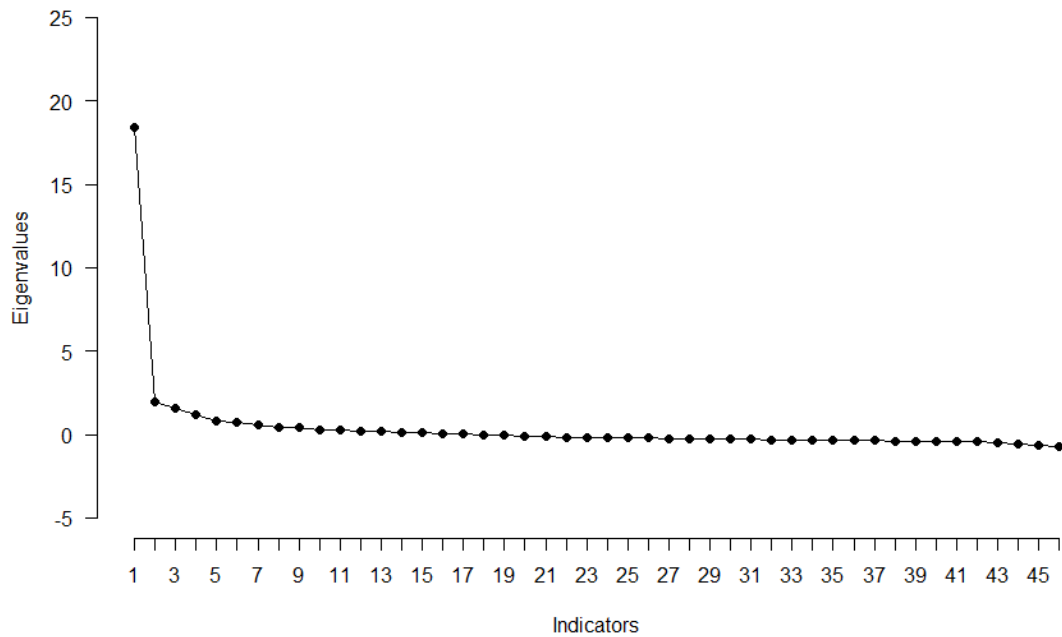


Figure 3.20: Scree plot with EFA-determined eigenvalues.

It was decided to keep the value four as the number of factors based on the “Kaiser-Guttman criterion with EFA” and “Parallel analysis with EFA” from the above results. The following results were obtained when EFA was applied over the pre-processed data (by keeping the number of factors as four) using the “EFA()” function of the “EFAtools” package [12].

```

> # 3) Applying EFA
> EFA(cor(feedback_selected_for_analysis), n_factors = 4)

EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'none'.

-- Unrotated Loadings -----

```

	F1	F2	F3	F4
Background	-.080	-.111	-.012	.021
Comfortabl	-.171	.040	-.226	-.059
Used any c	.152	.166	.128	-.156
Duration o	.008	.203	-.106	-.168
Spoken tut	.491	-.339	-.250	.104
Spoken tut	-.207	.409	.281	.089
Spoken tut	-.269	.443	.292	.016
Learned a	.543	-.133	-.355	.086
(topic) Pr	.770	-.072	.140	.051
(topic) Di	.779	-.257	.199	.022
(topic) De	.706	-.200	.346	.180
(topic) So	.765	-.320	.210	.108
(topic) He	.786	-.251	.285	-.019
(topic) Me	.809	-.238	.166	.089
(topic) Bu	.833	-.228	.134	.071
(topic) De	.815	-.194	.105	.091
(topic) De	.819	-.205	.177	.079
(topic) Gr	.806	-.195	.115	.038
(topic) De	.806	-.082	.190	.096
Spoken tut	.523	-.018	-.210	.078
ChemCollec	.213	-.020	-.152	-.152
Difficulty	.161	.049	.017	.075
Serial dil	.820	.068	-.050	-.378
Alcohol de	.815	.086	.004	-.381
Standardiz	.766	.011	-.092	-.368
Thermochem	.776	.091	.044	-.378
DNA Bindin	.767	.125	.047	-.326
Reaction o	.817	-.053	.023	-.373
Determinat	-.107	-.009	.058	-.105
FOSSEE and	.766	.299	-.081	.109
Real v/s v	.706	.349	.007	.248
Commom ion	.673	.343	-.046	.156
Determinat	.787	.387	.041	.150
Acid Base	.667	.388	.062	.174
Determinat	.710	.460	-.005	.034
Quality of	.797	.161	-.158	.051
Learning t	.748	.009	-.190	.162
Interactio	.750	.107	-.191	.086
Live Sessi	.749	.205	-.099	-.075
Pace of wo	.145	-.132	-.107	.021
Did not le	-.361	.094	.295	.057
Will recom	.668	.037	-.386	.227
Will recom	.600	.010	-.403	.175
Before wor	.044	.205	.405	.018
After work	.499	.021	.220	.060
Cost benef	-.408	.074	.143	.072

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	18.529	2.126	1.725	1.315
Prop Tot Var	0.403	0.046	0.037	0.029
Cum Prop Tot Var	0.403	0.449	0.486	0.515
Prop Comm Var	0.782	0.090	0.073	0.055
Cum Prop Comm Var	0.782	0.872	0.945	1.000

```

-- Model Fit -----

CAF: .45
df: 857

```

Figure 3.21: Results obtained after applying EFA.

From the above-mentioned results, it was not clear how to effectively segregate variables into different factors. Therefore, rotation was applied over the loadings [15]. The type of rotation applied was oblique [16]. Following results were obtained after applying rotation.

```

> # 4) Applying rotation
> EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")

EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.

-- Rotated Loadings -----

```

	F1	F2	F3	F4
Background	-.090	.026	-.071	.072
Comfortabl	.011	-.298	.016	.118
Used any c	-.005	.050	.252	-.166
Duration o	.058	-.234	.227	-.042
Spoken tut	.184	.238	-.019	.477
Spoken tut	.133	-.103	-.082	-.533
Spoken tut	.064	-.153	-.017	-.564
Learned a	.386	.043	.060	.419
(topic) Pr	.266	.513	.183	.057
(topic) Di	.090	.678	.180	.149
(topic) De	.156	.753	-.004	-.029
(topic) So	.105	.733	.069	.171
(topic) He	.023	.744	.227	.089
(topic) Me	.186	.662	.121	.155
(topic) Bu	.204	.637	.151	.177
(topic) De	.250	.586	.130	.169
(topic) De	.201	.653	.141	.126
(topic) Gr	.198	.584	.184	.168
(topic) De	.288	.583	.144	.027
Spoken tut	.383	.081	.084	.232
ChemCollec	.028	-.039	.231	.176
Difficulty	.159	.065	-.018	-.033
Serial dil	.127	.233	.693	.160
Alcohol de	.109	.264	.698	.108
Standardiz	.091	.212	.652	.221
Thermochem	.078	.278	.683	.069
DNA Bindin	.139	.262	.631	.035
Reaction o	.010	.369	.662	.191
Determinat	-.166	-.008	.076	-.037
FOSSEE and	.676	.102	.195	-.051
Real v/s v	.756	.136	.035	-.178
Common ion	.686	.069	.123	-.128
Determinat	.721	.164	.176	-.204
Acid Base	.679	.130	.111	-.241
Determinat	.665	.030	.291	-.217
Quality of	.583	.131	.240	.113
Learning t	.559	.193	.074	.221
Interactio	.567	.121	.177	.163
Live Sessi	.460	.114	.370	.049
Pace of wo	.040	.061	-.001	.189
Did not le	-.188	.032	-.163	-.341
Will recom	.690	-.015	-.017	.323
Will recom	.606	-.050	.012	.351
Before wor	-.015	.230	.035	-.432
After work	.180	.402	.103	-.108
Cost benef	-.139	-.100	-.197	-.225

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	0.468	-0.485	-0.204
F2	0.468	1.000	-0.298	-0.176
F3	-0.485	-0.298	1.000	0.070
F4	-0.204	-0.176	0.070	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	18.529	2.126	1.725	1.315
Prop Tot Var	0.403	0.046	0.037	0.029
Cum Prop Tot Var	0.403	0.449	0.486	0.515
Prop Comm Var	0.782	0.090	0.073	0.055
Cum Prop Comm Var	0.782	0.872	0.945	1.000

Figure 3.22: Results obtained after applying EFA with oblique rotation.

The total variance explained by the factors was 51.5%, as shown in Figure 3.22. It is lower than 60% which is generally expected as satisfactory in social sciences where information is often less precise. To

increase the amount of variance explained, it is required to remove insignificant variables from the data [17]. From Figure 3.22, it can be observed that the variables, “Background”, “Comfortable in learning new software”, “Used any chemistry virtual lab software in institute”, “Duration of any other chemistry virtual lab workshop (in days)”, “ChemCollective improved lab skills”, “Difficulty of practice problems”, “Determination of pKa of Acetic Acid (if done with practice)”, “Pace of workshop” and “Cost benefit”, are not associated with any of the four factors. Hence, they were removed and EFA was again performed over the remaining data, as shown in Figure 3.23.

```
> EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")
EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.
-- Rotated Loadings -----

```

	F1	F2	F3	F4
Spoken tut	.351	.129	-.081	.519
Spoken tut	-.098	.137	-.103	-.495
Spoken tut	-.200	.089	-.020	-.516
Learned a	.140	.342	.019	.439
(topic) Pr	.532	.157	.204	.002
(topic) Di	.686	-.032	.215	.060
(topic) De	.832	.049	-.041	-.101
(topic) So	.845	-.018	.018	.084
(topic) He	.757	-.106	.250	-.020
(topic) Me	.711	.067	.129	.071
(topic) Bu	.707	.074	.145	.106
(topic) De	.695	.147	.071	.091
(topic) De	.733	.096	.096	.060
(topic) Gr	.639	.100	.157	.104
(topic) De	.634	.192	.121	-.035
Spoken tut	.155	.346	.057	.188
Serial dil	.053	.067	.814	.061
Alcohol de	.081	.052	.809	.010
Standardiz	.056	.040	.755	.116
Thermochem	.066	.029	.815	-.022
DNA Bindin	.049	.081	.783	-.054
Reaction o	.196	-.065	.782	.085
FOSSEE and	.110	.630	.197	-.049
Real v/s v	.143	.719	.055	-.181
Common ion	.058	.667	.133	-.130
Determinat	.130	.686	.200	-.209
Acid Base	.111	.661	.116	-.241
Determinat	-.067	.644	.357	-.225
Quality of	.139	.539	.237	.103
Learning t	.297	.489	.036	.204
Interactio	.164	.513	.164	.157
Live Sessi	.063	.425	.394	.010
Did not le	.026	-.169	-.183	-.306
Will recom	.090	.672	-.073	.356
Will recom	.044	.587	-.035	.386
Before wor	.259	-.018	-.059	-.422
After work	.446	.140	.021	-.099

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	0.549	-0.642	-0.191
F2	0.549	1.000	-0.603	-0.141
F3	-0.642	-0.603	1.000	0.197
F4	-0.191	-0.141	0.197	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	18.203	2.044	1.597	1.230
Prop Tot Var	0.492	0.055	0.043	0.033
Cum Prop Tot Var	0.492	0.547	0.590	0.624
Prop Comm Var	0.789	0.089	0.069	0.053
Cum Prop Comm Var	0.789	0.877	0.947	1.000

```

-- Model Fit -----
CAF: .44
df: 524

```

Figure 3.23: Results obtained after removing insignificant variables.

Finally, the results were obtained in a way that all variables got mapped and the total variance explained became 62.4%. In Figure 3.23, multiple cross-loadings can be observed. When a variable is found to have

more than one significant loading, it is known as cross-loading. One can apply different rotation methods to eliminate the cross-loadings, but if that does not work, then the variables causing cross-loading become candidates for deletion [17]. Therefore, EFA was again performed after removing all variables causing cross-loadings and the obtained results can be seen in Figure 3.24.

```
> EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")
EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.

-- Rotated Loadings -----

```

	F1	F2	F3	F4
Spoken tut	-.111	.027	.050	.756
Spoken tut	-.211	-.023	.136	.750
(topic) Pr	.565	.127	.202	.031
(topic) Di	.715	.011	.132	-.125
(topic) De	.791	.140	-.146	-.030
(topic) So	.914	-.109	.037	-.034
(topic) He	.802	-.083	.173	-.021
(topic) Me	.712	.119	.064	-.119
(topic) Bu	.785	-.045	.212	.038
(topic) De	.748	.060	.106	-.021
(topic) De	.773	.049	.100	-.027
(topic) Gr	.678	.077	.147	-.074
(topic) De	.651	.172	.110	.051
Spoken tut	.168	.350	.012	-.246
Serial dil	.120	.130	.742	-.014
Alcohol de	.135	.146	.712	.007
Standardiz	.146	.072	.689	-.070
Thermochem	.152	.086	.712	.039
DNA Bindin	.092	.216	.628	.001
Reaction o	.289	-.047	.741	.003
FOSSEE and	.134	.534	.240	.105
Real v/s v	.042	.849	-.066	.002
Commom ion	-.084	.885	-.007	-.033
Determinat	.040	.815	.093	.080
Acid Base	-.005	.818	-.006	.082
Quality of	.123	.518	.280	-.017
Learning t	.255	.491	.076	-.140
Interactio	.115	.524	.225	-.043
Live Sessi	.030	.503	.355	-.006
Did not le	-.021	-.143	-.193	.227
Before wor	.198	.042	-.111	.302
After work	.415	.179	-.024	.044

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	0.652	0.582	-0.135
F2	0.652	1.000	0.579	-0.067
F3	0.582	0.579	1.000	-0.241
F4	-0.135	-0.067	-0.241	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	16.420	1.755	1.323	1.061
Prop Tot Var	0.513	0.055	0.041	0.033
Cum Prop Tot Var	0.513	0.568	0.609	0.642
Prop Comm Var	0.799	0.085	0.064	0.052
Cum Prop Comm Var	0.799	0.884	0.948	1.000

```

-- Model Fit -----
CAF: .45
df: 374

```

Figure 3.24: Results obtained after removing variables causing cross-loadings.

From Figure 3.24, it can be observed that one variable with the name “Live Session learning” is still causing cross-loading and another variable with the name “Did not learn much” does not belong to any of the four factors. Hence, those two variables were also removed and EFA was again performed over the remaining data. The final results can be seen in Figure 3.25.

```
> feedback_selected_for_analysis <- feedback_selected_for_analysis[,-c(29,30)]
> EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")

EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.

-- Rotated Loadings -----

```

	F1	F2	F3	F4
Spoken tut	-.093	-.020	.042	<b>.840</b>
Spoken tut	-.219	-.045	.112	<b>.719</b>
(topic) Pr	.575	.108	.201	.034
(topic) Di	.722	.003	.133	-.122
(topic) De	<b>.789</b>	.141	-.142	-.029
(topic) So	<b>.905</b>	-.103	.039	-.026
(topic) He	<b>.797</b>	-.087	.175	-.022
(topic) Me	.737	.097	.060	-.114
(topic) Bu	<b>.798</b>	-.065	.207	.050
(topic) De	.752	.050	.111	-.017
(topic) De	<b>.788</b>	.033	.098	-.014
(topic) Gr	<b>.703</b>	.053	.143	-.054
(topic) De	<b>.647</b>	.165	.117	.051
Spoken tut	.191	<b>.336</b>	.025	-.210
Serial dil	.087	.128	<b>.770</b>	-.020
Alcohol de	.094	.150	<b>.743</b>	.000
Standardiz	.120	.074	<b>.712</b>	-.057
Thermochem	.123	.085	<b>.730</b>	.038
DNA Bindin	.069	.213	<b>.647</b>	.001
Reaction o	.255	-.044	<b>.760</b>	-.005
FOSSEE and	.130	<b>.517</b>	.263	.110
Real v/s v	.010	<b>.870</b>	-.032	-.015
Common ion	-.081	<b>.867</b>	.024	-.037
Determinat	.024	<b>.807</b>	.127	.066
Acid Base	-.034	<b>.827</b>	.028	.060
Quality of	.158	<b>.470</b>	.289	-.007
Learning t	<b>.318</b>	<b>.428</b>	.079	-.110
Interactio	.167	<b>.464</b>	.228	-.024
Before wor	.171	.045	-.106	.257
After work	<b>.418</b>	.170	-.023	.039

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	0.649	0.612	-0.100
F2	0.649	1.000	0.566	-0.009
F3	0.612	0.566	1.000	-0.200
F4	-0.100	-0.009	-0.200	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	15.756	1.720	1.279	1.075
Prop Tot Var	0.525	0.057	0.043	0.036
Cum Prop Tot Var	0.525	0.583	0.625	0.661
Prop Comm Var	0.795	0.087	0.064	0.054
Cum Prop Comm Var	0.795	0.881	0.946	1.000

```

-- Model Fit -----
CAF: .44
df: 321

```

Figure 3.25: Results obtained after removing insignificant and cross-loading causing variables.



The obtained results also contained one variable that was causing cross-loading and one which was not associated with any of the four factors, as shown in Figure 3.25. After removing both the variables EFA was once more applied over the remaining data and the results are shown in Figure 3.26.

```
> feedback_selected_for_analysis <- feedback_selected_for_analysis[,-c(27,29)]
> EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")

EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.

-- Rotated Loadings -----

```

	F1	F2	F3	F4
Spoken tut	.057	-.075	.028	.797
Spoken tut	-.065	.018	-.007	.782
(topic) Pr	.629	.163	.077	.065
(topic) Di	.724	.109	-.020	-.122
(topic) De	.839	-.220	.157	-.050
(topic) So	.961	-.044	-.105	-.010
(topic) He	.810	.132	-.099	-.028
(topic) Me	.783	.016	.065	-.082
(topic) Bu	.851	.158	-.101	.077
(topic) De	.803	.066	.016	.007
(topic) De	.850	.037	.007	.015
(topic) Gr	.753	.099	.016	-.017
(topic) De	.727	.044	.152	.078
Spoken tut	.186	.059	.285	-.191
Serial dil	.003	.868	.056	-.033
Alcohol de	.003	.835	.095	-.032
Standardiz	.021	.815	.007	-.081
Thermochem	.039	.807	.040	.001
DNA Bindin	.018	.702	.171	-.010
Reaction o	.180	.826	-.099	-.015
FOSSEE and	.174	.284	.457	.125
Real v/s v	.045	-.022	.850	-.042
Common ion	-.045	.049	.828	-.043
Determinat	.069	.143	.766	.056
Acid Base	.013	.021	.820	.042
Quality of	.176	.345	.379	.019
Interactio	.202	.281	.360	.024
After work	.392	-.007	.158	-.031

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	-0.734	0.629	-0.263
F2	-0.734	1.000	-0.610	0.208
F3	0.629	-0.610	1.000	-0.092
F4	-0.263	0.208	-0.092	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	15.224	1.706	1.216	1.052
Prop Tot Var	0.544	0.061	0.043	0.038
Cum Prop Tot Var	0.544	0.605	0.648	0.686
Prop Comm Var	0.793	0.089	0.063	0.055
Cum Prop Comm Var	0.793	0.882	0.945	1.000

```

-- Model Fit -----
CAF: .44
df: 272

```

Figure 3.26: Results obtained after removing insignificant and cross-loading causing variables.

Again the obtained results contained one variable which was causing cross-loading and one which was not associated with any of the four factors, as shown in Figure 3.26. After removing both the variables EFA was once more applied over the remaining data and the results are shown in Figure 3.27.

```

> feedback_selected_for_analysis <- feedback_selected_for_analysis[,-c(14,26)]
> Results <- EFA(cor(feedback_selected_for_analysis), n_factors = 4, rotation = "oblimin")
> Results

EFA performed with type = 'EFAtools', method = 'PAF', and rotation = 'oblimin'.

-- Rotated Loadings -----

```

	F1	F2	F3	F4
Spoken tut	.038	-.082	.023	<b>.768</b>
Spoken tut	-.070	.019	-.027	<b>.799</b>
(topic) Pr	<b>.636</b>	.147	.080	.050
(topic) Di	.722	.097	-.002	-.139
(topic) De	<b>.831</b>	-.221	.174	-.061
(topic) So	<b>.959</b>	-.048	-.101	-.012
(topic) He	<b>.805</b>	.126	-.090	-.035
(topic) Me	<b>.785</b>	.013	.070	-.086
(topic) Bu	<b>.853</b>	.151	-.101	.074
(topic) De	<b>.814</b>	.070	-.003	.018
(topic) De	<b>.856</b>	.042	-.007	.025
(topic) Gr	<b>.758</b>	.112	-.002	.002
(topic) De	.725	.050	.151	.081
Serial dil	.004	<b>.868</b>	.059	-.033
Alcohol de	.005	<b>.839</b>	.096	-.029
Standardiz	.033	<b>.816</b>	-.006	-.067
Thermochem	.038	<b>.810</b>	.041	.003
DNA Bindin	.013	<b>.713</b>	.174	-.007
Reaction o	.177	<b>.813</b>	-.080	-.027
FOSSEE and	.206	.283	<b>.413</b>	.129
Real v/s v	.056	-.013	<b>.843</b>	-.051
Commom ion	-.038	.065	<b>.822</b>	-.046
Determinat	.083	.156	<b>.747</b>	.054
Acid Base	.015	.026	<b>.827</b>	.027
Interactio	.239	.271	<b>.317</b>	.024
After work	<b>.396</b>	-.010	.156	-.037

```

-- Factor Intercorrelations -----

```

	F1	F2	F3	F4
F1	1.000	-0.733	0.613	-0.236
F2	-0.733	1.000	-0.595	0.186
F3	0.613	-0.595	1.000	-0.049
F4	-0.236	0.186	-0.049	1.000

```

-- Variances Accounted for -----

```

	F1	F2	F3	F4
SS loadings	14.390	1.676	1.211	1.019
Prop Tot Var	0.553	0.064	0.047	0.039
Cum Prop Tot Var	0.553	0.618	0.665	0.704
Prop Comm Var	0.787	0.092	0.066	0.056
Cum Prop Comm Var	0.787	0.878	0.944	1.000

```

-- Model Fit -----
CAF: .44
df: 227

```

Figure 3.27: Results obtained after removing insignificant and cross-loading causing variables.

Finally, the obtained results, as shown in Figure 3.27, became free from cross-loadings and insignificant variables. After obtaining the desired results, segregation of variables was performed based on the four factors, as shown in Figure 3.28.

```
> # 10) Segregating factors
> F1 <- feedback_selected_for_analysis[,c(3:13,26)]
> colnames(F1)
[1] "(topic) Preparation of standard solutions"
[2] "(topic) Dilution and pH measurements"
[3] "(topic) Density of solids and liquids"
[4] "(topic) Solubility of salts"
[5] "(topic) Heat of reaction"
[6] "(topic) Metal displacement reactions"
[7] "(topic) Buffer Solutions"
[8] "(topic) Determination of Equilibrium constant"
[9] "(topic) Determination of solubility product"
[10] "(topic) Gravimetric Analysis"
[11] "(topic) Determination of pKa"
[12] "After workshop"
> F2 <- feedback_selected_for_analysis[,c(14:19)]
> colnames(F2)
[1] "Serial dilution problem"
[2] "Alcohol density problem"
[3] "Standardization of NaOH"
[4] "Thermochemistry of coolant problem"
[5] "DNA Binding problem"
[6] "Reaction of halogen"
> F3 <- feedback_selected_for_analysis[,c(20:25)]
> colnames(F3)
[1] "FOSSEE and spoken tutorial by Prof. Kannan"
[2] "Real v/s virtual lab Prof. Lakshmy Ravishankar"
[3] "Common ion effect Prof. Padmavathy"
[4] "Determination of pKa of CH3COOH Dr. Gomathi Sridhar"
[5] "Acid Base mixture Dr. Rama Kanwar"
[6] "Interaction with ChemCollective team"
> F4 <- feedback_selected_for_analysis[,c(1,2)]
> colnames(F4)
[1] "Spoken tutorial need improvement"
[2] "Spoken tutorial unclear"
```

Figure 3.28: Segregation of variables based on the four factors.

After segregation, the factors were given names depending on the variables they contained [17]. Factor **F1** was given the name “**Quality of spoken tutorials**”, factor **F2** was given the name “**Quality of live sessions**”, factor **F3** was given the name “**Quality of live lectures**” and finally factor **F4** was given the name “**Spoken Tutorial learning experience**”. Later the factor scores associated with each factor were obtained using the “FACTOR\_SCORES()” function of the “EFAtools” package [12], to perform a regression analysis between all four factors and the “Overall quality” variable, which was removed from the pre-processed dataset and separately stored before performing EFA. Regression was performed by taking all factors as independent variables and “Overall quality” as a dependent variable [18]. Results of the regression analysis are shown in Figure 3.29.

```
> lm(data = as.data.frame(Scores), `Overall Quality`~.)  
  
Call:  
lm(formula = `Overall Quality` ~ ., data = as.data.frame(Scores))  
  
Coefficients:  
                (Intercept)  
                3.65248  
  `Quality of spoken tutorials`  
                1.35276  
  `Quality of live sessions`  
                1.83515  
  `Quality of live lectures`  
                0.71274  
  `Spoken Tutorial learning experience`  
               -0.06633
```

Figure 3.29: Results of the regression analysis.

Finally, the factors contributing to the effectiveness of the workshop have been determined. The following equation explains the relationship between them and the overall quality of the workshop.

$$\text{Overall Quality} = 3.65 + 1.35(\text{Quality of spoken tutorials}) + 1.84(\text{Quality of live sessions}) + 0.71(\text{Quality of live lectures}) - 0.07(\text{Spoken Tutorial learning experience})$$

## 7. Conclusion

The results from EFA revealed various underlying factors affecting the overall quality of the workshop. Later, the regression analysis provided an equation to quantify the relationship between the factors and the workshop quality. Various data manipulation steps were performed over the workshop feedback data to implement EFA and regression. The results from this project shall help improve the content of future workshops by focusing more on those factors that will have a significant impact on the overall quality and experience of the workshop.

# Chapter 4

## Implementation and visualization of SOM algorithm in R

### 1. Introduction

SOM is an unsupervised data visualization technique popular among researchers for dimensionality reduction and clustering [19]. This project aims to create an open-source code base for SOM in R to help researchers, students, and professionals understand the working of SOM. The material has been designed to encourage and promote the R programming language among people wanting to learn and apply SOM for their choice of use. The complete code with proper explanation and examples has been made freely available for educational purposes in the form of a document on the [Resources](#) page of the R FOSSEE website.

### 2. Self Organizing Maps

Self Organizing Maps (Kohonen Maps) are a class of artificial neural network created by Dr. Teuvo Kohonen that can map high dimensional input data to a 2D map using unsupervised learning [20-24]. SOMs are utilized for various applications because they provide a low-dimensional representation of a high-dimensional input while maintaining the features of input data in the representation [25,26].

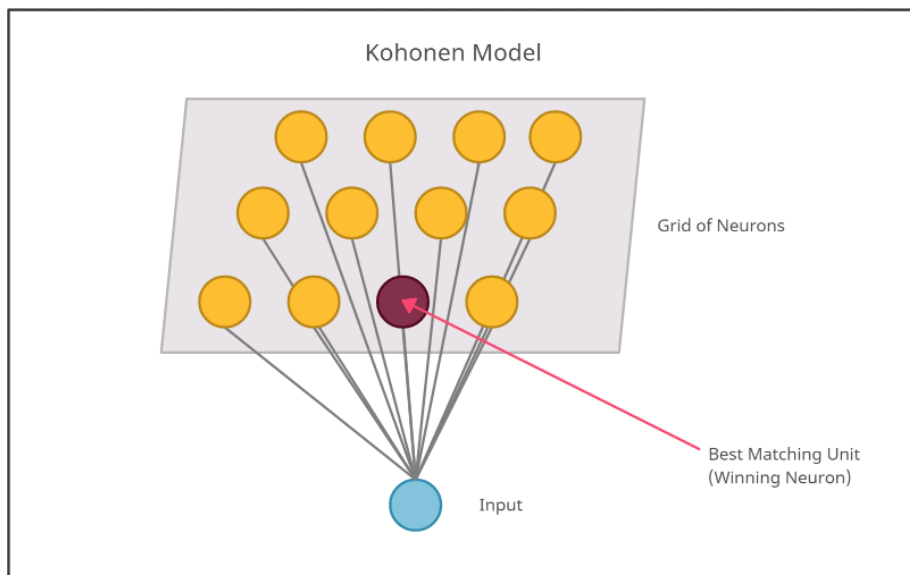


Figure 4.1: Kohonen Model of Self Organizing Map [24].

### 3. Implementation of SOM in R

Due to the project's complexity, the entire process of implementing SOM in R was divided into various tasks and each FOSSEE fellow was assigned a particular task. The first phase of development started with the creation of a prototype of SOM in R to make our understanding better regarding the concept. For testing the prototype we attempted to recreate an example from [Wikipedia](#).

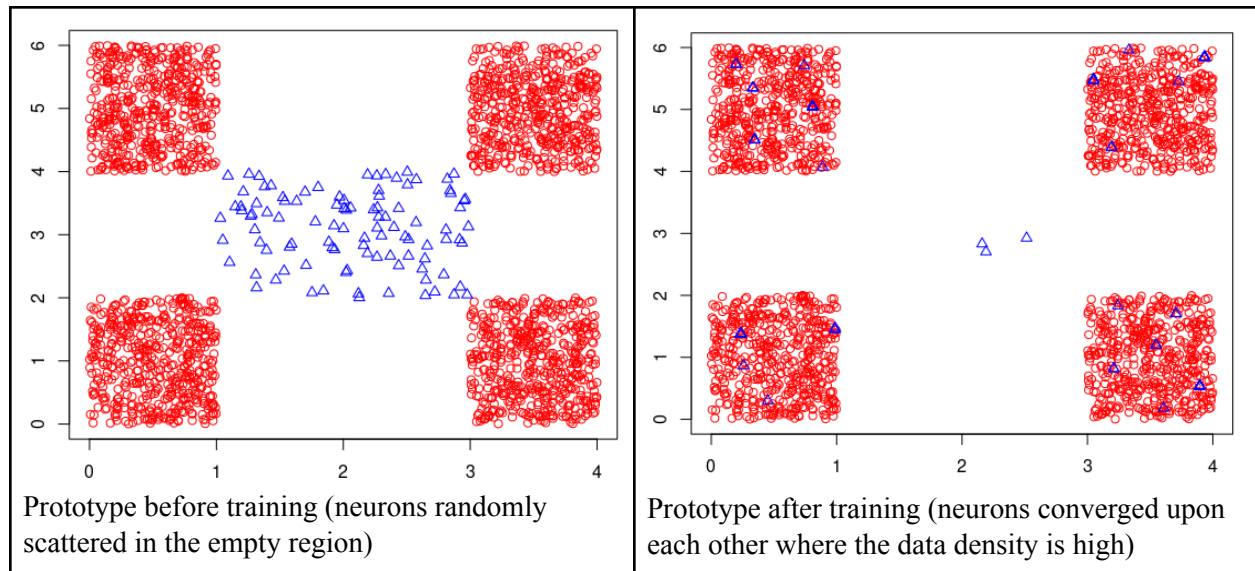


Figure 4.2: SOM prototype created using R.

In the second phase, a basic SOM model was created from the prototype and implemented over the Iris dataset [27] for the purpose of testing.

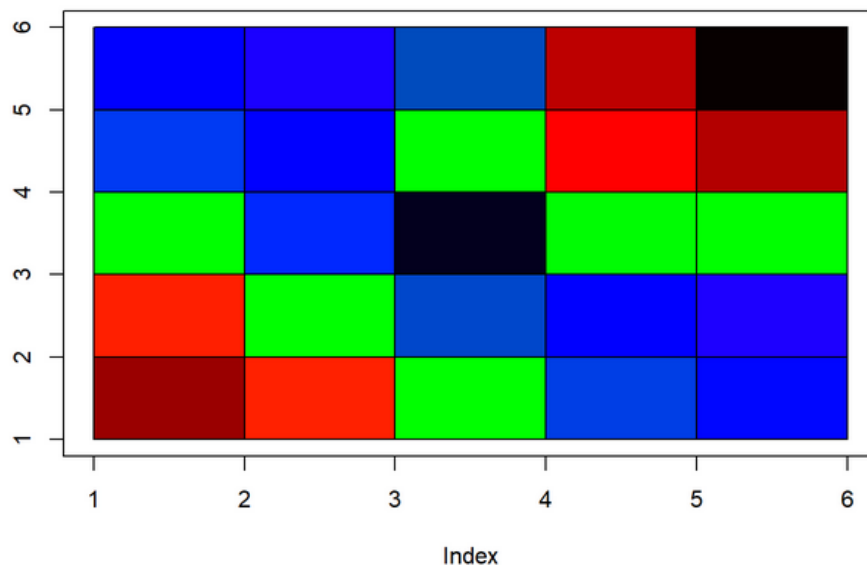


Figure 4.3: Clusters obtained after implementing the SOM model over the Iris dataset.

The results obtained were accurate but the model execution was slow. It was observed that the model did not satisfactorily converge after a single epoch over the complete input data. The model training algorithm was then modified to incorporate multiple epochs. The map converged during the second epoch for most datasets.

The tasks assigned to me were data collection, preprocessing of the collected data, visualization of the SOM neuron grid and creation of the final code script.

## 4. Visualization

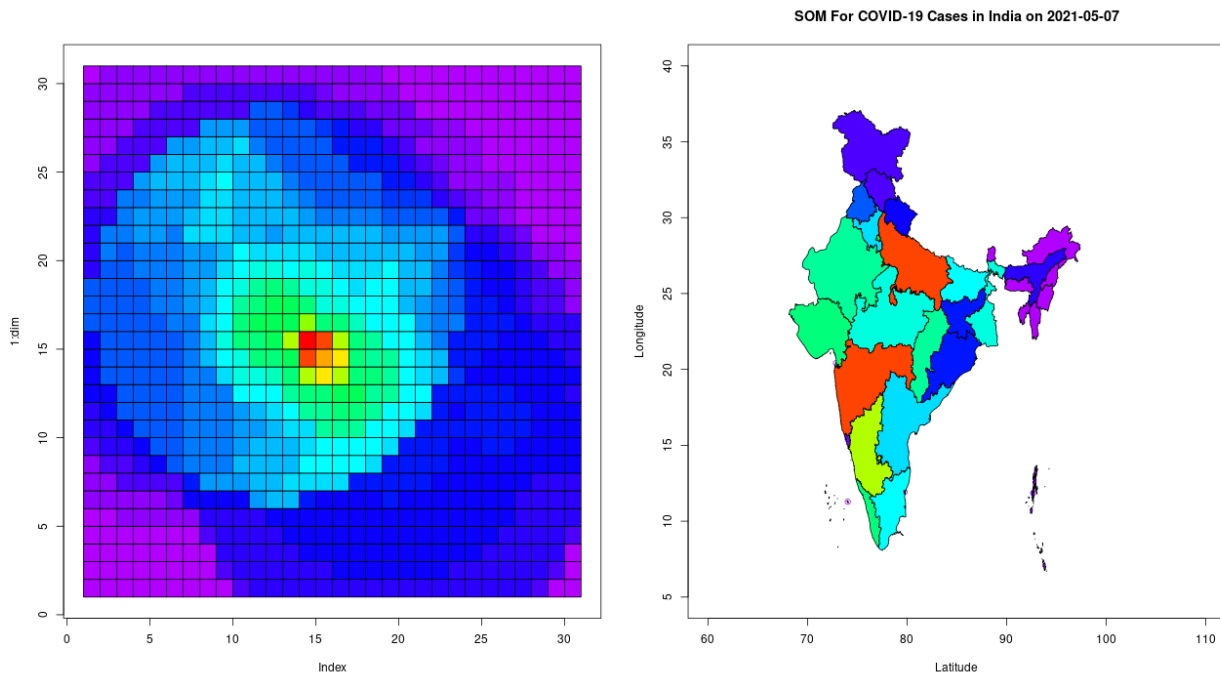


Figure 4.4: SOM feature map along with India map showing the intensity of COVID-19 infection.

### 4.1 SOM feature map

SOM was visualized in the form of a feature map. A feature map is a 2D array of shapes like a circle, hexagon and rectangle, where each shape represents a neuron. In our implementation of SOM, each neuron was assigned a color based on its weights which were obtained after model training. The model training process converted an N-dimensional vector to a single number. This process was repeated for all the input data vectors and the results were mapped to 38 numbers ranging from 10 and 48. For each number in the range of 10 to 48, a color was assigned from the rainbow color palette of the “base” package of R [7].

## 4.2 India map

The KML file associated with the India map was obtained from the GeoServer platform by National Remote Sensing Centre, Department of Space, Government of India [28]. The map was constructed in R using the “rgdal” [29] and “sf” [30] packages. To display the intensity of COVID-19 infection over the map, data of each state and union territory was passed to the BMU function that returned the location of the associated winning neuron from the SOM grid. Using the location of the winning neuron, a color was assigned to it and later that same color was given to the associated state or union territory over the map.



# Chapter 5

## R Case Study: Clustering of common goods and commodities based on time-series characteristics of their Wholesale Price Index

### 1. Introduction

Countries like India and Philippines use Wholesale Price Index (WPI) as a measure of inflation. Due to the usage of WPI as an indicator of inflation associated with the goods and commodities, the Indian government keeps a detailed record of price changes associated with a list of items in the form of WPI. The WPI based inflation estimates also serve as an essential determinant in the formulation of trade, fiscal, and other economic policies by the Government [31]. Due to the importance of this indicator I proposed a case study to perform a clustering analysis over the time-series characteristics of the Wholesale Price Index of various common goods and commodities to obtain insights regarding the underlying similarity between them. The case study project was completed under the guidance of Prof. Radhendushka Srivastava. The entire analysis was performed using the R programming language. The complete case study with code and data has been made available in the [Completed Case studies](#) section of the R FOSSEE website. A brief description of the complete case study is given in the following sections.

### 2. Data

The data for the case study was obtained from the **Wholesale Price Index** catalogue of [data.gov.in](http://data.gov.in) and can be accessed [here](#). The data is freely available for educational purposes. The dataset had a collection of 869 row items divided into two categories: **All Commodities** and **Food Index**. **All Commodities** category is further divided into three subcategories, namely **Primary Articles**, **Fuel & Power** and **Manufactured Products**. The data contains commodity names, commodity code, commodity weight as percentage of total weight of all commodities and their monthly price index for nine years as column entries from April 2011 to December 2020 with year 2011-12 as base year for calculating WPI. This type of data is routinely made available by the Office of Economic Adviser, Ministry of Commerce & Industry, Government of India.

	A	B	C	D	E	F	G
1	COMM_NAME	COMM_CODE	COMM_WINDX0420:	INDX0520:	INDX0620:	INDX0720:	
2	ALL COMMODITIES	1000000000	100	97.2	97.8	98.3	98.7
3	I PRIMARY ARTICLES	1100000000	22.61756	95.6	95.7	97.6	98.6
4	(A). FOOD ARTICLES	1101000000	15.25585	94.5	95.5	98	99.6
5	a. FOOD GRAINS (CEREALS	1101010000	3.46238	97.6	98.7	98.8	99.6
6	a1. CEREALS	1101010100	2.82378	98.5	99.7	99.7	100.5
7	Paddy	1101010101	1.43052	97.5	98.9	99.4	100.1
8	Wheat	1101010102	1.02823	101.1	101	100.4	101.2
9	Jowar	1101010103	0.06764	84.5	91.9	98.3	100.4
10	Bajra	1101010104	0.08637	97.7	102.6	98.3	101.4
11	Maize	1101010105	0.18927	98.2	100.8	99.5	99.1
12	Barley	1101010106	0.01437	91.1	99.1	97.7	101.3
13	Ragi	1101010107	0.00738	94	93.9	95.7	97.9
14	a2. PULSES	1101010200	0.6386	93.7	94.2	94.6	95.9
15	Gram	1101010201	0.26377	79.5	82	84.9	90
16	Arhar	1101010202	0.12914	107.3	103.8	100.9	99.1
17	Moong	1101010203	0.07088	104.6	103.8	102.2	100.7
18	Masur	1101010204	0.05299	101	100.5	98.4	98.7
19	Urad	1101010205	0.09165	105.3	105.5	105.5	101.7
20	Peas/Chawali	1101010206	0.02444	87.2	92.8	97.6	102
21	Rajma	1101010207	0.00573	81.6	83.8	84.8	89.1
22	b. FRUITS & VEGETABLES	1101020000	3.47508	94.7	94	99	101.1
23	b1. VEGETABLES	1101020100	1.87448	81.9	84.2	96.3	101.9
24	Potato	1101020101	0.27737	92	101.6	106.9	115.8

Figure 5.1 : Original dataset with a row-wise arrangement of each category of commodity.

### 3. Data Cleaning

Before proceeding to data analysis, all rows with missing WPI values were removed and the clustering was performed only over the data of individual common goods and commodities.

### 4. Data Analysis

The objective of clustering analysis is to partition the data such that similarity is minimized across the groups and maximized within each group [32]. The clustering was performed over the smooth versions of the original time series which were obtained through kernel smoothing. In this case study, we used the Gaussian kernel and the bandwidth for the kernel was selected using the cross validation method [33]. Kernel smoothing was implemented using the “sm.regression()” function of the “sm” package in R [34]. The objective was to remove noise from raw time series data.

As time series data is dynamic, i.e., it changes with time, we can not directly apply it over the generic clustering algorithms like Agglomerative Hierarchical Clustering (AHC) which work only on static data [35]. Time series data can be converted to static form using a dissimilarity matrix. The aim of using dissimilarity matrix is to obtain a single numeric value expressing the degree of dissimilarity between every unique pairwise combination of the different time series in consideration [36]. To compute a dissimilarity matrix, we used the “diss()” function from the “TSclust” package of R [37]. The output was a triangular matrix. It was now possible to apply AHC with complete linkage over the data by passing the obtained matrix as an input to the “hclust()” function of the “stats” package of R [7].

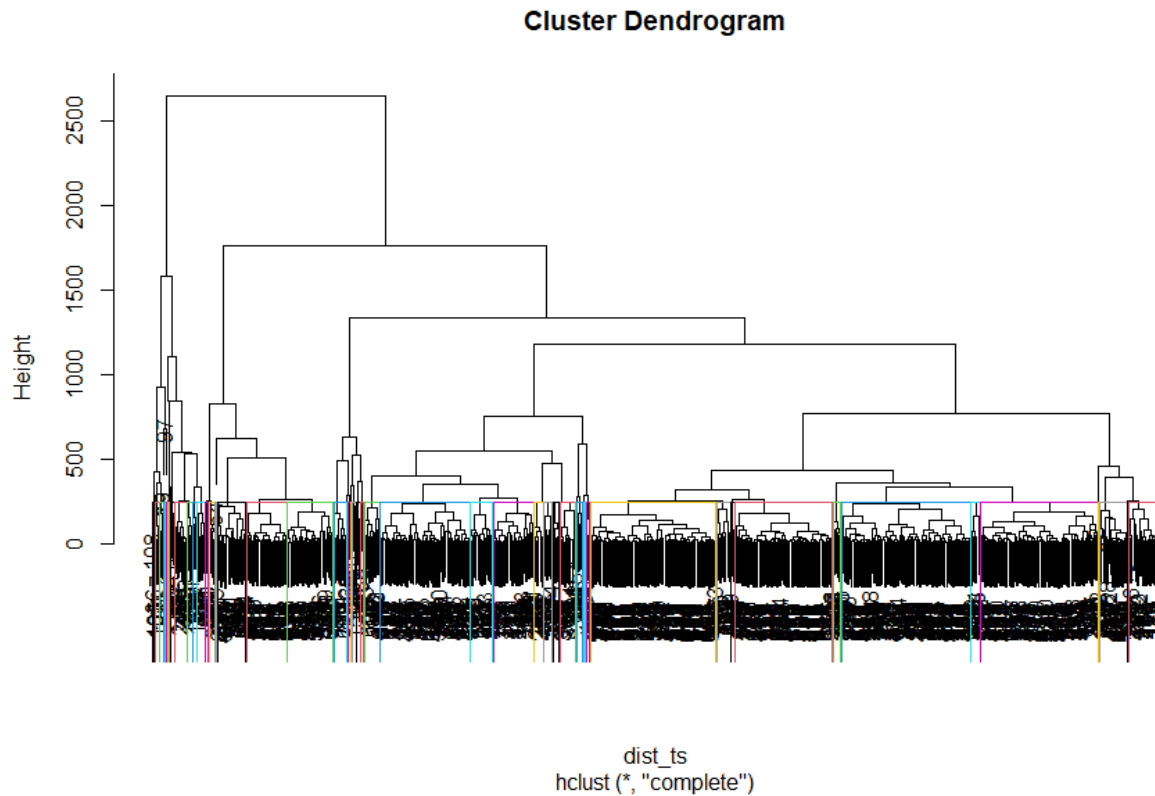


Figure 5.2: Dendrogram obtained from clustering.

The dendrogram obtained from AHC containing around 50 clusters is shown in Figure 5.2. Each cluster is represented by a rectangle with colored border. Finally, only 24 clusters were retained containing five or more elements for further analysis, as suggested by the mentor. Figure 5.3 shows the plot of all the time-series associated with a particular cluster.

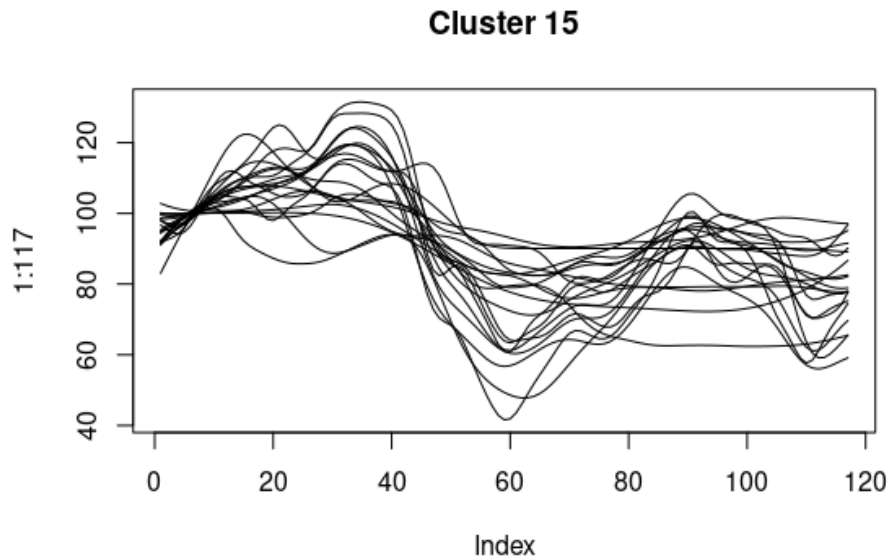


Figure 5.3: Time-series data associated with a cluster.

By examining the retained clusters a central trend was observed in each cluster, but when the items contained in each cluster were examined it was found that in most clusters the items did not belong to a single commodity class or in other words the items were non-homogenous commodity-wise. Figure 5.4 shows that the cluster number 19 of the retained clusters contained edible items like “Basmati rice” along with a variety of non-edible items like industrial goods, clothing materials, etc.

Cluster : 19		
[1] "Raw Cotton"	"Hides (Raw)"	"Electricity"
[4] "Fruit Juice including concentrates"	"Basmati rice"	"Vegetable starch"
[7] "Spirits"	"Cotton Yarn"	"Synthetic yarn"
[10] "Viscose yarn"	"Woollen yarn"	"Texturised and twisted Yarn"
[13] "Synthetic Fabric - Others"	"Fabrics/cloth, rayon"	"Knitted fabrics of cotton"
[16] "Nylon rope"	"Vegetable Tanned Leather"	"Duplex paper"
[19] "Laminated plastic sheet"	"Printed labels/posters/calendars"	"Organic Solvent"
[22] "Aromatic chemicals"	"Ethyl acetate"	"Ethylene Oxide"
[25] "Urea"	"XLPE Compound"	"Printing ink"

Figure 5.4 : List of the items present in cluster 19.

In an attempt to segregate homogeneous items with each cluster, ARIMA modeling was applied to create subclusters [38,39]. The subclusters were obtained by examining the characteristics of noise associated with the original WPI time-series data of each cluster through ARIMA modeling. Noise data was obtained by subtracting the smooth version of the time-series from the original time-series. ARIMA modeling was implemented using the “auto.arima()” function of the “forecast” package in R [40]. Figure 5.5 shows the plot of original values (black) over the fitted values (red).

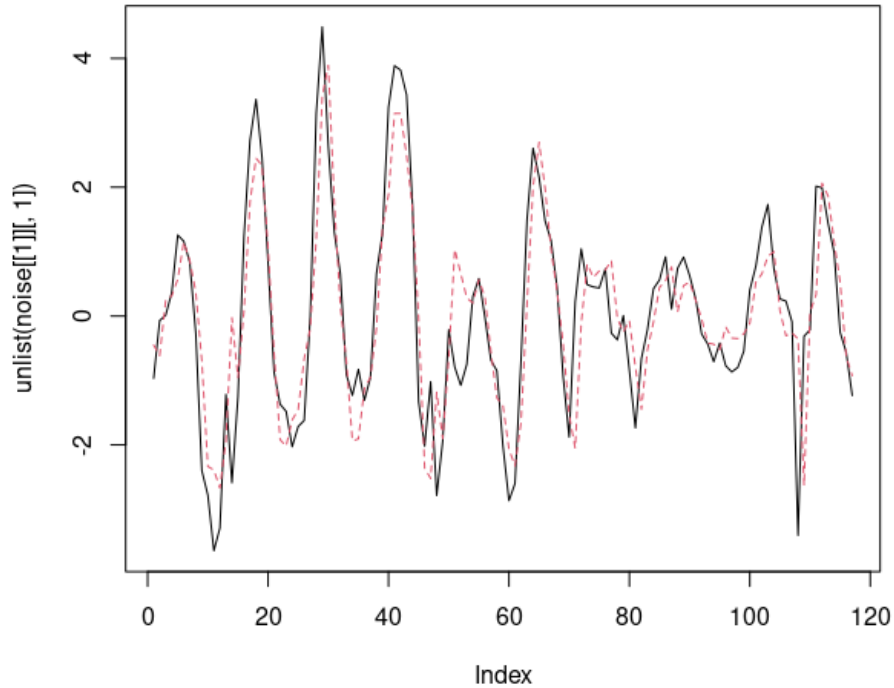


Figure 5.5: Original noise time-series of commodity 1, i.e. ‘paddy’ (black), over the fitted ARIMA model values (red).

Hierarchical clustering was performed over the sum of squared differences between the noise values and the fitted values using euclidean distance as the distance metric.

Subclusters were obtained only for the cluster 19 as it was the largest among all. For others the same procedure can be implemented to obtain subclusters.

## 5. Results

We found that the initial set of clusters obtained after applying AHC over the smooth version of the original time-series were homogeneous concerning the shape of the time-series data. We also found that, for the data used in this case study, the correlation-based metrics efficiently identify time-series with similar shape. Figure 5.6 shows all the retained 24 clusters.

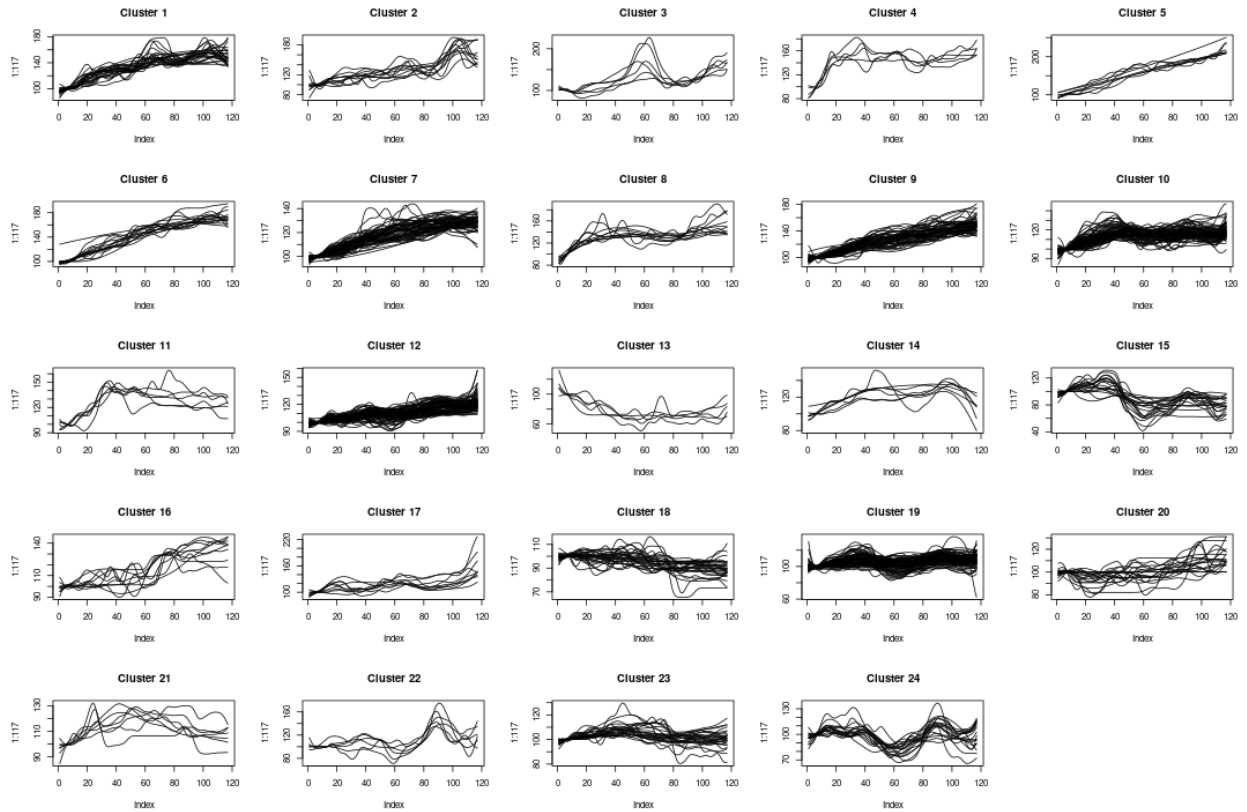


Figure 5.6: Plot containing all the 24 retained clusters.

Despite having a similar shape, the clusters contained elements belonging to different product categories, making them non-homogeneous commodity-wise. To further obtain commodity-wise homogenous clusters, subclusters were generated within the originally obtained clusters by applying ARIMA modeling over the noise associated with the original time-series data belonging to each cluster, which was initially removed using kernel smoothing. The subclusters were relatively more homogeneous in terms of both the time-series shape and the common goods and commodities they contain, but we were not completely homogenous. The time series associated with each subcluster of the cluster 19 are shown in Figure 5.7 and the Figure 5.8 contains the list of goods and commodities associated with each subcluster.

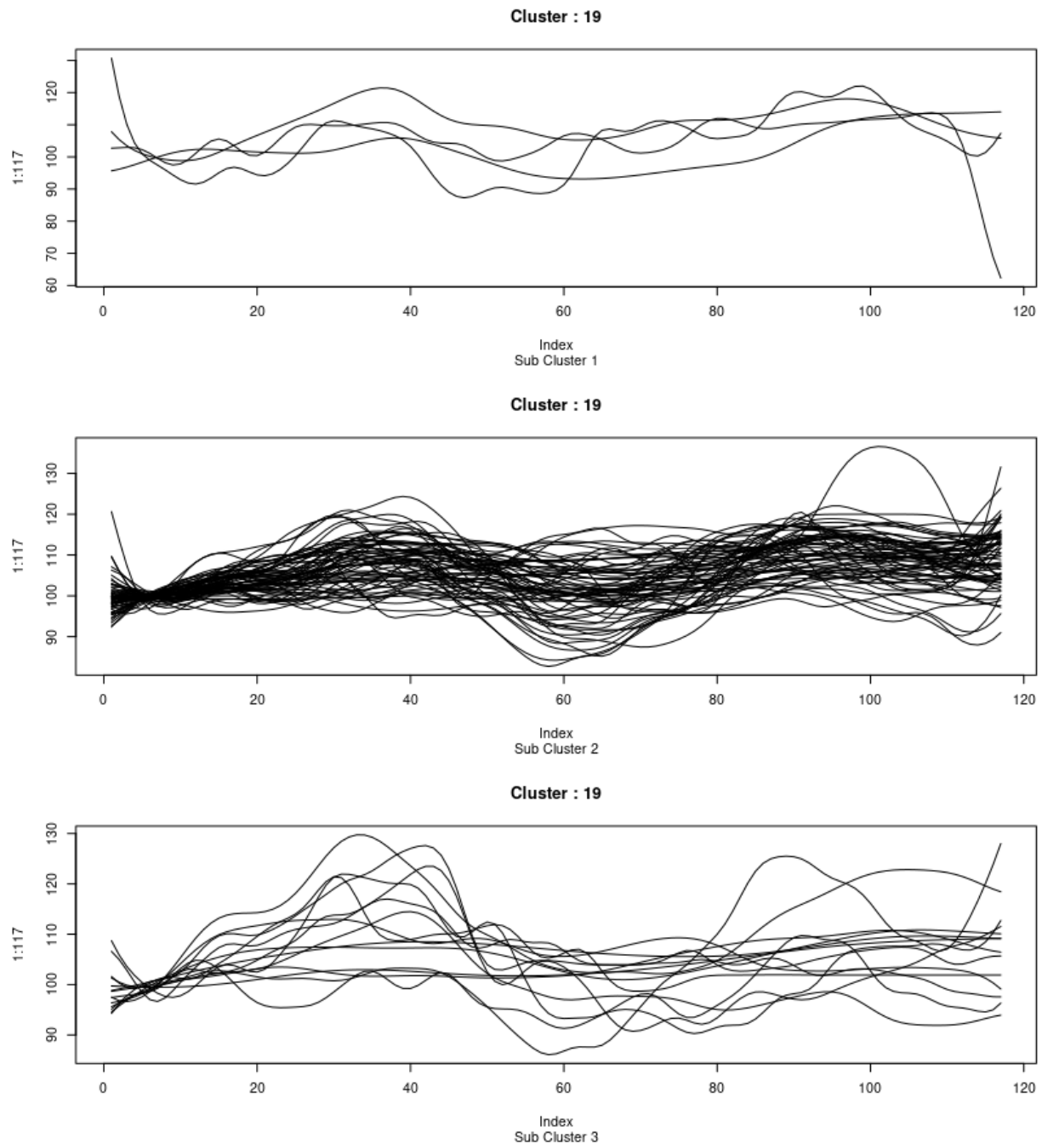


Figure 5.7: Plot of each subcluster associated with the cluster 19.

Sub Cluster : 1		
[1] "Raw Cotton"	"Laminated plastic sheet"	"Sunglasses"
"Cranes"		
Sub Cluster : 2		
[1] "Hides (Raw)"		"Electricity"
[3] "Fruit Juice including concentrates"		"Basmati rice"
[5] "Vegetable starch"		"Spirits"
[7] "Cotton Yarn"		"Synthetic yarn"
[9] "Viscose yarn"		"Woolen yarn"
[11] "Texturised and twisted Yarn"		"Synthetic Fabric - Others"
[13] "Fabrics/cloth, rayon"		"Knitted fabrics of cotton"
[15] "Nylon rope"		"Duplex paper"
[17] "Organic Solvent"		"Aromatic chemicals"
[19] "Ethyl acetate"		"Ethylene Oxide"
[21] "Urea"		"Printing ink"
[23] "Plasticizer"		"Polyester film(metalized)"
[25] "Adhesive excluding gum"		"Epoxy, liquid"
[27] "Rubber Chemicals"		"Organic chemicals"
[29] "Acrylic fibre"		"Polyester fibre fabric"
[31] "Vials/ampoule, glass, empty or filled"		"IV fluids"
[33] "2/3 wheeler Tyre"		"Cycle/Cycle rickshaw tyre"
[35] "Rubber moulded goods"		"Condoms"
[37] "Polypropylene film"		"Plastic bottle"
[39] "Plastic tape"		"Acrylic/plastic sheet"
[41] "Electric insulating material"		"Ferrosilicon"
[43] "Hot Rolled (HR) Coils & Sheets, including Narrow Strip"		"Cold Rolled (CR) Coils & Sheets, including Narrow Strip"
[45] "Galvanized iron pipes"		"Copper metal/Copper Rings"
[47] "Lead ingots, bars, blocks, plates"		"Copper shapes - bars/rods/plates/strips"
[49] "Brass metal/sheet/coils"		"Cast iron, castings"
[51] "MS castings"		"Steel pipes, tubes & poles"
[53] "Boilers"		"Bolts, screws, nuts & nails of Iron & steel"
[55] "Telephone sets including mobile hand sets"		"Microscope"
[57] "Electrical relay/conductor"		"PVC Insulated Cable"
[59] "Copper wire"		"Insulating & flexible wire"
[61] "Flourescent tube"		"Electric filament type lamps"
[63] "Motors & other DC equipment"		"Clutches and shaft couplings"
[65] "Mining, quarrying & metallurgical machinery/parts"		"Wheels/wheels & parts"
[67] "Shock absorbers"		"Gear box and parts"
[69] "Propellers & Blades of Boats/Ships"		
Sub Cluster : 3		
[1] "Vegetable Tanned Leather"	"Printed labels/posters/calendars"	"XLPE Compound"
[4] "Organic surface active agent"	"Antiseptics and disinfectants"	"Polyester film (non-metalized)"
[7] "Porcelain sanitary ware"	"GP/GC sheet"	"Forged Steel Rings"
[10] "Meter (non-electrical)"	"Geyser"	"Material handling, lifting and hoisting equipment"
[13] "Body (for commercial motor vehicles)"	"Cylinder liners"	

Figure 5.8: List of elements belonging to each subcluster of cluster 19.

## 6. Conclusion

This case study attempted to explore one of many approaches to time-series clustering to obtain homogeneous clusters of various common goods and commodities based on the time function of their Wholesale Price Index (WPI) after removing noise using kernel smoothing. The clustering method used in this case study doesn't seem to create homogenous clusters for the given time-series data. One needs a better distance metric that involves the correlation structure of the time-series to create homogenous clusters. The correlation structure of the time-series data can be exploited to obtain homogenous clusters in future work.



# Chapter 6

## Conclusion

The FOSSEE Semester-long Internship has been a fantastic opportunity to learn, apply and collaborate with my fellow interns while working on various projects under the guidance of mentors from the R FOSSEE team. The R on Cloud platform enables individuals from various disciplines to learn and explore R programming through TBC. Reporting bugs and errors on this platform enabled me to contribute towards making R on Cloud robust and improve its quality. Analysis of FOSSEE workshop feedback data allowed me to better grasp the flow of a data science project, it also highlighted the nuances of data cleaning and preprocessing. The SOM project was one of the most challenging projects I worked on during this internship. The task of creating a machine learning model from scratch not only improved my programming skills but also developed my insights on how to prototype, plan and test such projects while working in a team. The SOM project gave me confidence and provided a solid foundation to better approach such challenges in the future. The individually performed case study project pushed me to take up an open source dataset and perform novel analysis over it under the guidance of the R FOSSEE team. The case study project exposed me to the methods of identifying and applying the most suitable statistical operations over a dataset.

FOSSEE Semester-long Internship was a journey to explore the R ecosystem. This internship made me and my fellow interns learn R with hands-on experience and along the way allowed me to develop and inculcate proper research and reporting habits. Through this internship I got to meet a lot of fantastic and smart people and the experience gained here will carry me further in my academic journey.

# References

- [1] A Practical Introduction to Factor Analysis: Confirmatory Factor Analysis. UCLA: Statistical Consulting Group.  
<https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/>
- [2] Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1.  
<https://CRAN.R-project.org/package=readxl>
- [3] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686,  
<https://doi.org/10.21105/joss.01686>
- [4] Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2021). skimr: Compact and Flexible Summaries of Data. R package version 2.1.3.
- [5] Broeck, J., Argeseanu Cunningham, S., Eeckels, R., and Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), p.e267.
- [6] Chu, X., Ilyas, I., Krishnan, S., and Wang, J. (2016). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 international conference on management of data (pp. 2201–2206).
- [7] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [8] Margaret Beaver (2012). Survey Data Cleaning Guidelines: (SPSS and Stata) 1st Edition.  
<https://www.canr.msu.edu/resources/survey-data-cleaning-guidelines-spss-and-stata-1st-edition>
- [9] Krishnan, S., Haas, D., Franklin, M., and Wu, E. 2016. Towards reliable interactive data cleaning: A user survey and recommendations. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics (pp. 1–5).
- [10] Hooper, D. (2012), 'Exploratory Factor Analysis', in Chen, H. (Ed.), Approaches to Quantitative Research – Theory and its Practical Application: A Guide to Dissertation Students, Cork, Ireland: Oak Tree Press.
- [11] Tarka, P. (2015). Likert Scale and Change in Range of Response Categories vs. the Factors Extraction in EFA Model. Acta Universitatis Lodzianis. Folia Oeconomica, 311.
- [12] Steiner, M.D., & Grieder, S.G. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. Journal of Open Source Software, 5(53), 2521.  
<https://doi.org/10.21105/joss.02521>
- [13] Watkins, M. (2018). Exploratory factor analysis: A guide to best practice. Journal of Black Psychology, 44(3), p.219–246.
- [14] KMO and Bartlett's test, SPSS Statistics Subscription - New, SPSS Statistics, IBM Corporation.  
<https://www.ibm.com/docs/en/spss-statistics/version-missing?topic=detection-kmo-bartletts-test>
- [15] Marley W. Watkins (2018). Exploratory Factor Analysis: A Guide to Best Practice. Journal of Black Psychology, 44(3), 219–246.
- [16] Osborne, Jason W. (2015) "What is Rotating in Exploratory Factor Analysis?," Practical Assessment, Research, and Evaluation: Vol. 20, Article 2.
- [17] Joseph F. Hair, Jr., William C. Black, Barry J. Babin, Rolph E. Anderson (2010). 'Multivariate Data Analysis', 7/e. Pearson Prentice Hall.

- [18] Tu, T.T.T. and Dung, N.T.P., 2017. Factors affecting green banking practices: Exploratory factor analysis on Vietnamese banks. *Journal of Economic Development*, (JED, Vol. 24 (2)), pp.4-30.
- [19] Kevin Pang. Self-organizing Maps. <https://www.cs.hmc.edu/~kpang/nn/som.html>
- [20] Kohonen, Teuvo. "The self-organizing map." *Proceedings of the IEEE* 78.9 (1990): 1464-1480.
- [21] Uoolc, A. Bradford. "Self-organizing Map Formation: Foundations of Neural Computation."
- [22] Kohonen, Teuvo. "Essentials of the self-organizing map." *Neural networks* 37 (2013): 52-65.
- [23] Kohonen, Teuvo, and Timo Honkela. "Kohonen network." *Scholarpedia* 2.1 (2007): 1568.
- [24] Sven Krüger. Self-Organizing Maps. [https://www.iikt.ovgu.de/iesk\\_media/Downloads/ks/computational\\_neuroscience/vorlesung/comp\\_neuro8-p-2090.pdf](https://www.iikt.ovgu.de/iesk_media/Downloads/ks/computational_neuroscience/vorlesung/comp_neuro8-p-2090.pdf)
- [25] John A. Bullinaria. (2004). Self Organizing Maps: Fundamentals. <https://www.cs.bham.ac.uk/~jxb/NN/116.pdf>
- [26] Jae-Wook Ahn and Sue Yeon Syn. (2005). Self-Organizing Maps. <https://sites.pitt.edu/~is2470pb/Spring05/FinalProjects/Group1a/tutorial/som.html>
- [27] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179–188.
- [28] GeoServer by National Remote Sensing Centre, Department of Space, Government of India. <https://bhuvan-vec3.nrsc.gov.in/bhuvan/web/wicket/bookmarkable/org.geoserver.web.demo.MapPreviewPage.jsessionid=E704D8D216740FC88D13EC934562D253.worker1?0>
- [29] Roger Bivand, Tim Keitt and Barry Rowlingson (2021). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-27. <https://CRAN.R-project.org/package=rgdal>
- [30] Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10 (1), 439-446, <https://doi.org/10.32614/RJ-2018-009>
- [31] MANUAL ON WHOLESALE PRICE INDEX (Base: 2011-12 = 100). Office of the Economic Adviser, Department of Industrial Policy & Promotion, Ministry of Commerce & Industry, Government of India. [https://eaindustry.nic.in/uploaded\\_files/WPI\\_Manual.pdf](https://eaindustry.nic.in/uploaded_files/WPI_Manual.pdf)
- [32] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*. Third Edition. Elsevier, 2012.0
- [33] Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- [34] Bowman, A. W. and Azzalini, A. (2021). R package 'sm':nonparametric smoothing methods (version 2.2-5.7) URL <http://www.stats.gla.ac.uk/~adrian/sm>
- [35] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11. Elsevier BV, pp. 1857–1874, Nov. 2005. doi: 10.1016/j.patcog.2005.01.025
- [36] C. M. M. Pereira and R. F. de Mello, "Common Dissimilarity Measures are Inappropriate for Time Series Clustering," *Revista de Informática Teórica e Aplicada*, vol. 20, no. 1. Universidade Federal do Rio Grande do Sul, p. 25, Jan. 09, 2013. doi: 10.22456/2175-2745.25070
- [37] Pablo Montero, José A. Vilar (2014). TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software*, 62(1), 1-43. URL <http://www.jstatsoft.org/v62/i01/>.
- [38] Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

- [39] Rupinder Katoch and Arpit Sidhu, “An Application of ARIMA Model to Forecast the Dynamics of COVID-19 Epidemic in India,” *Global Business Review*, 1–14, 2021. DOI: 10.1177/097215092098865
- [40] Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeeen F (2022). *\_forecast: Forecasting functions for time series and linear models\_*. R package version 8.16, URL: <https://pkg.robjhyndman.com/forecast/>