Semester-long Internship Report

on

FLOSS - R

submitted by

Aboli Marathe (SCTR's Pune Institute of Computer Technology)

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 Siddhant Raghuvanshi, for their support, intellectual discussions, and enthusiasm. I am very grateful to be given such a fantastic opportunity.

Contents

Chapter 1
Introduction

This report shares my contributions made to the open-source software community 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) "R" as a part of the FOSSEE (Free/Libre and Open Source Software for Education) project. The FOSSEE project is a part of the National Mission on Education through ICT with the thrust area of creating and promoting FLOSS equivalents to proprietary software. This project is based at the Indian Institute of Technology Bombay (IITB) and funded by the Ministry of Education, Government of India. The contributions include maintenance of R on Cloud, analysis of FOSSEE workshop feedback data, implementation & documentation of SOM algorithm in R, and an R case study on statistical analysis of the Spatio-temporal trends in COVID-19 strains and infections across India during 2020-21.

Chapter 2
Maintenance of R on Cloud

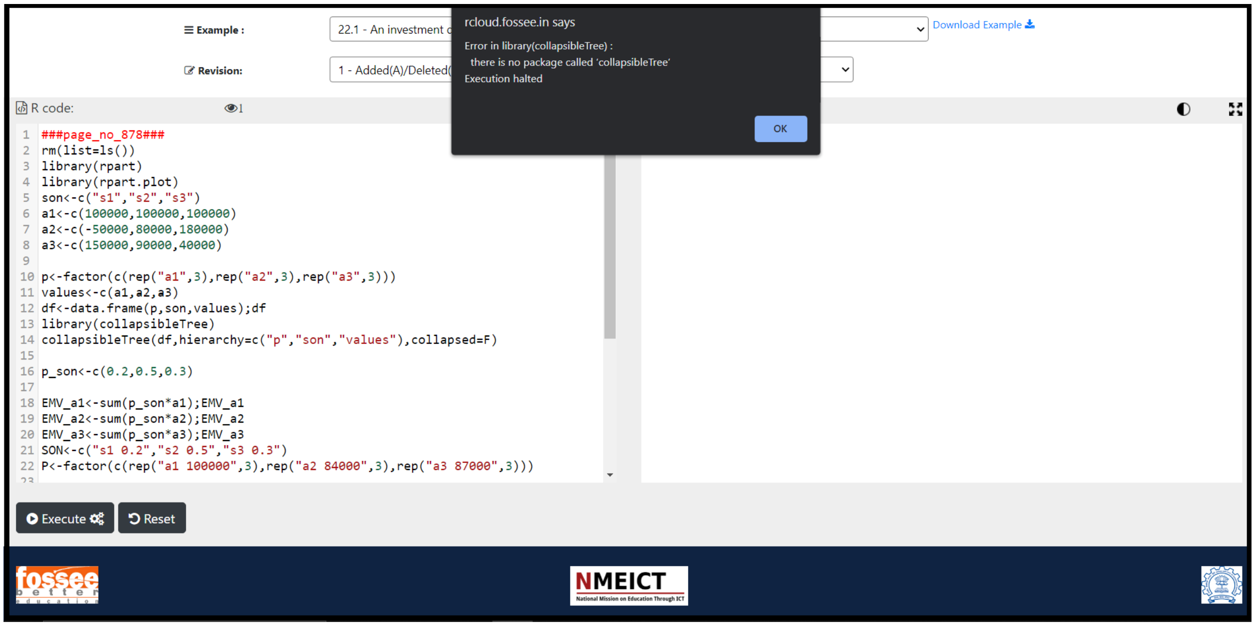
The R on Cloud is an online facility created by FOSSEE which works as a platform for executing R codes. It also allows users to interact with the codes of the completed textbook companions (TBCs). Because of this feature, it is required to check the completed TBCs over the platform for errors by running their codes. Hence, the assigned task involved checking each code file associated with 11 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 Name
1	Introduction to Probability and Statistics for Engineers and Scientists by Sheldon M. Ross, Elsevier Academic Press, USA, 2004
2	Introductory Linear Algebra: An Applied First Course by Bernard Kolman & David R. Hill, Dorling Kindersley, India, 2008
3	Introductory Statistics by Douglas S Shafer and Zhiyi Zhang, Flat World Publishers, 2012
4	Linear Algebra by Jim Hefferon, Orthogonal Publishing L3C, 2017
5	Linear Algebra and Its Applications by David C. Lay, Pearson Addison-Wesley, 2006
6	Managerial Statistics by Gerald Keller, South-western Cenage Learning; Usa, 2009
7	Matrices and Linear Transformations by Charles G. Cullen, Dover Publications, Inc., New York, 1990
8	Miller and Freund's Probability and Statistics for Engineers by Richard A. Johnson, Pearson, USA, 2017
9	Modern Physical Chemistry: A Molecular Approach by George H. Duffey, Springer US, 2000
10	Numerical Methods for Engineers by S. C. Chapra and R. P. Canale, McGraw Hill, New York, 2006
11	Numerical Methods in Finance and Economics: A MATLAB-Based Introduction by Paolo Brandimarte, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006

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.



Example : 22.1 - An investment c

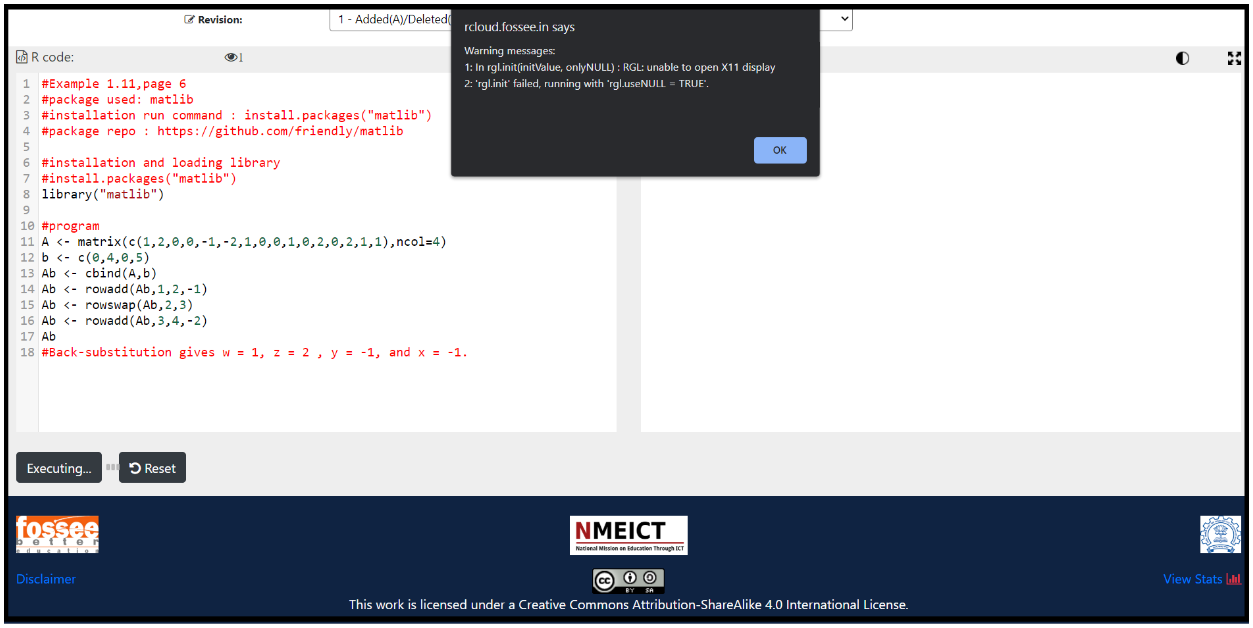
Revision: 1 - Added(A)/Deleted(

```
###page_no_878###
rm(list=ls())
library(rpart)
library(rpart.plot)
son<-c("s1","s2","s3")
a1<-c(100000,100000,100000)
a2<-c(-50000,80000,180000)
a3<-c(150000,90000,40000)

p<-factor(c(rep("a1",3),rep("a2",3),rep("a3",3)))
values<-c(a1,a2,a3)
df<-data.frame(p,son,values);df
library(collapsibleTree)
collapsibleTree(df,hierarchy=c("p","son","values"),collapsed=F)

p_son<-c(0.2,0.5,0.3)

EMV_a1<-sum(p_son*a1);EMV_a1
EMV_a2<-sum(p_son*a2);EMV_a2
EMV_a3<-sum(p_son*a3);EMV_a3
SON<-c("s1 0.2","s2 0.5","s3 0.3")
P<-factor(c(rep("a1 100000",3),rep("a2 84000",3),rep("a3 87000",3)))
```

```
rcloud.fossee.in says
Error in library(collapsibleTree) :
  there is no package called 'collapsibleTree'
Execution halted
```

Figure 2.1: Error due to the missing library “collapsibleTree” [1].

2. Warning messages prevented the execution of code.



Revision: 1 - Added(A)/Deleted(

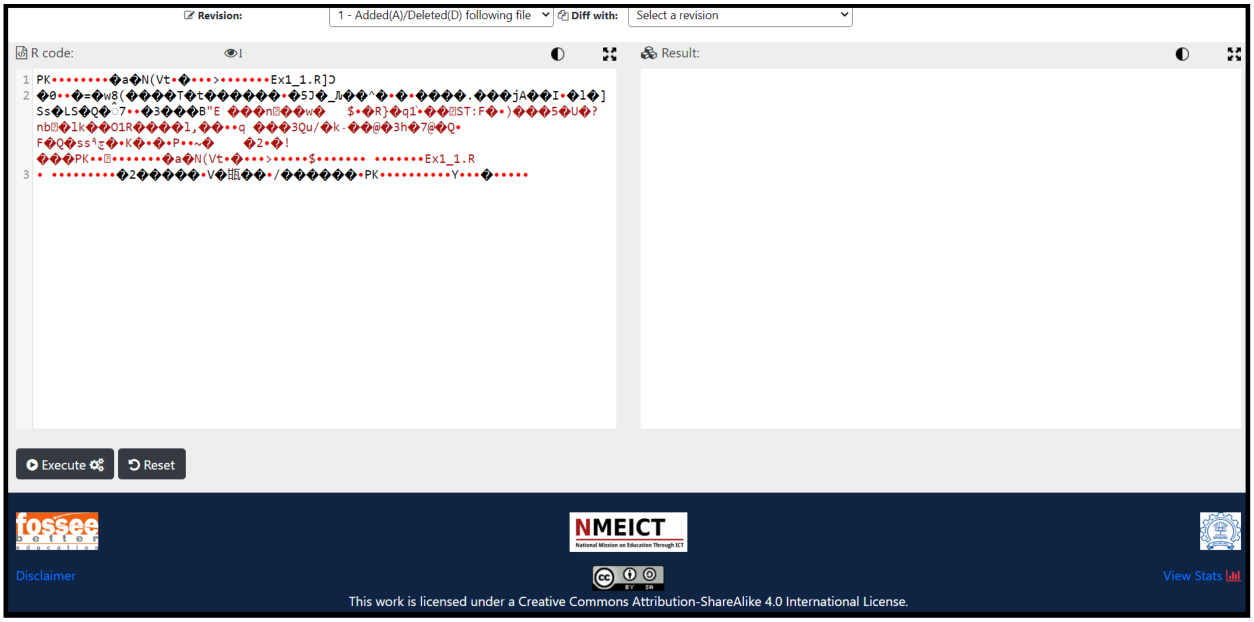
```
#Example 1.11,page 6
#package used: matlib
#installation run command : install.packages("matlib")
#package repo : https://github.com/friendly/matlib

#installation and loading library
#install.packages("matlib")
library("matlib")

#program
A <- matrix(c(1,2,0,0,-1,-2,1,0,0,1,0,2,0,2,1,1),ncol=4)
b <- c(0,4,0,5)
Ab <- cbind(A,b)
Ab <- rowadd(Ab,1,2,-1)
Ab <- rowswap(Ab,2,3)
Ab <- rowadd(Ab,3,4,-2)
Ab
#Back-substitution gives w = 1, z = 2 , y = -1, and x = -1.
```

```
rcloud.fossee.in says
Warning messages:
1: In rgl.init(initValue, onlyNULL) : RGL: unable to open X11 display
2: 'rgl.init' failed, running with 'rgl.useNULL = TRUE'.
```

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure 2.2: Warning message indicating the inability of the cloud platform to open the X11 display.

3. Unable to load R code from a zip file.



1 - Added(A)/Deleted(D) following file

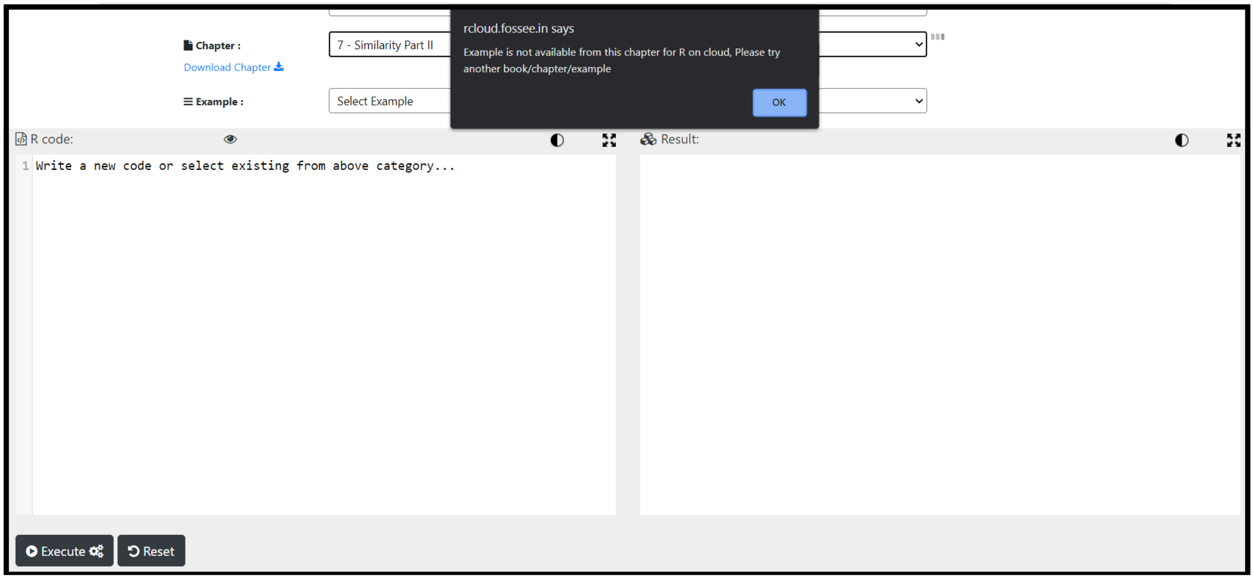
Select a revision

```
PK••••••••�a�N(Vt•�•••>•••••••Ex1_1.R]Ɔ
�0••�=�w8(����T�t�������•�5J�_₯��^�•�•�����.���jA��I•�l�]
Ss�LS�Q�͂7••�3���B"E ���n⍰��w�   $•�R}�q1̀•��⍰ST:F�•)����5�U�?
nb⍰�lk��O1R�����l,��••q ���3Qu/�k-��@�3h�7@�Q•
F�Q�ss꠹�•K�•�•P••~�    �2•�!
���PK••⍰•••••••�a�N(Vt•�•••>•••••$••••••• •••••••Ex1_1.R
• •••••••••�2������•V�瓶��•/�������•PK••••••••••Y•••�•••••
```

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure 2.3: Zip file content displayed as text.

4. Code not available for a TBC example.



7 - Similarity Part II

Select Example

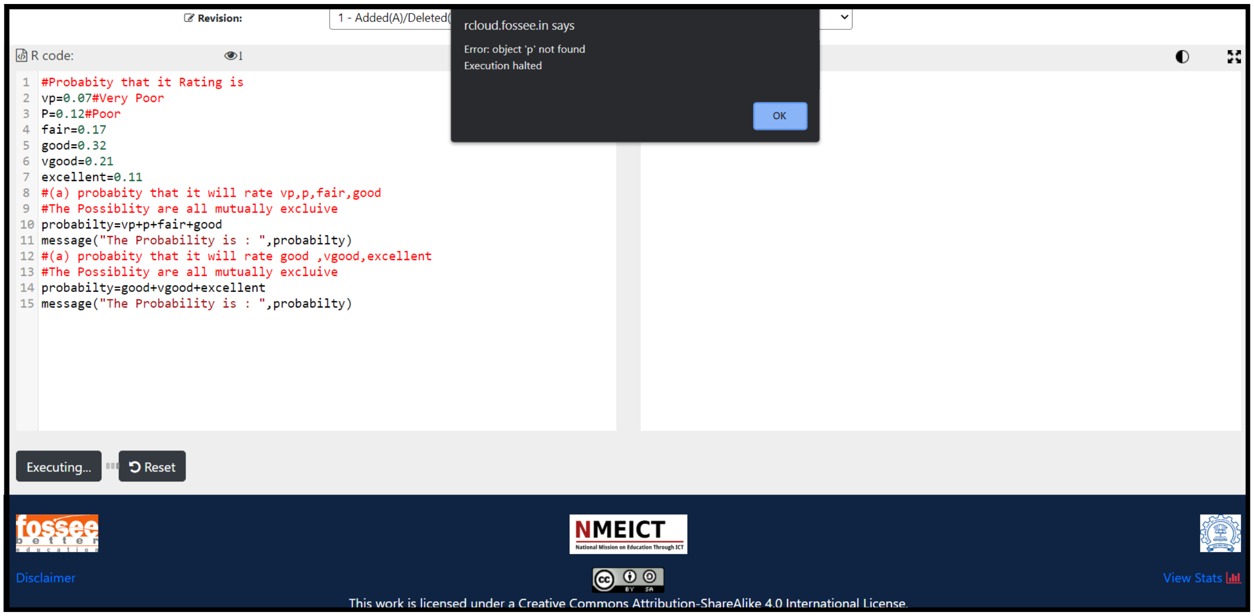
rcloud.fossee.in says

Example is not available from this chapter for R on cloud, Please try another book/chapter/example

```
Write a new code or select existing from above category...
```

Figure 2.4: Message prompt indicating the unavailability of code for the selected example.

5. Missing R objects.



rcloud.fossee.in says

Error: object 'p' not found
Execution halted

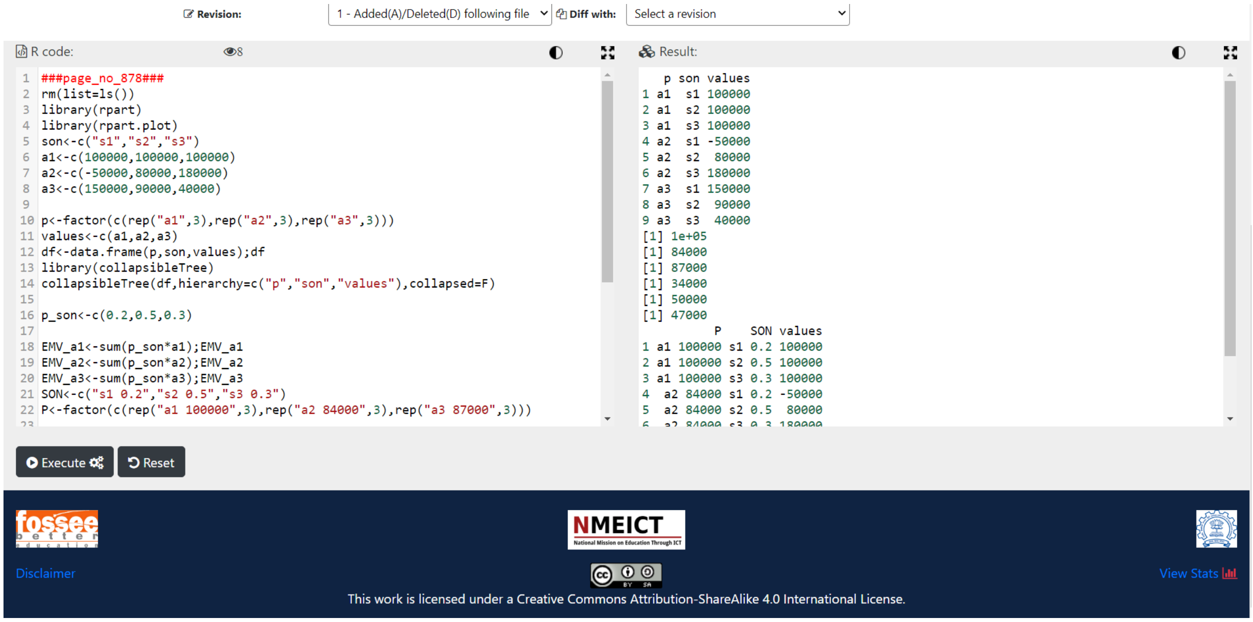
OK

```r
#Probabity that it Rating is
vp=0.07#Very Poor
P=0.12#Poor
fair=0.17
good=0.32
vgood=0.21
excellent=0.11
#(a) probabity that it will rate vp,p,fair,good
#The Possiblity are all mutually excluive
probabilty=vp+p+fair+good
message("The Probability is : ",probabilty)
#(a) probabity that it will rate good ,vgood,excellent
#The Possiblity are all mutually excluive
probabilty=good+vgood+excellent
message("The Probability is : ",probabilty)
```

Figure: 2.5: Error due to a missing R object.

FOSSEE web team did the following to fix the errors -

1. Installed all missing libraries over the platform.

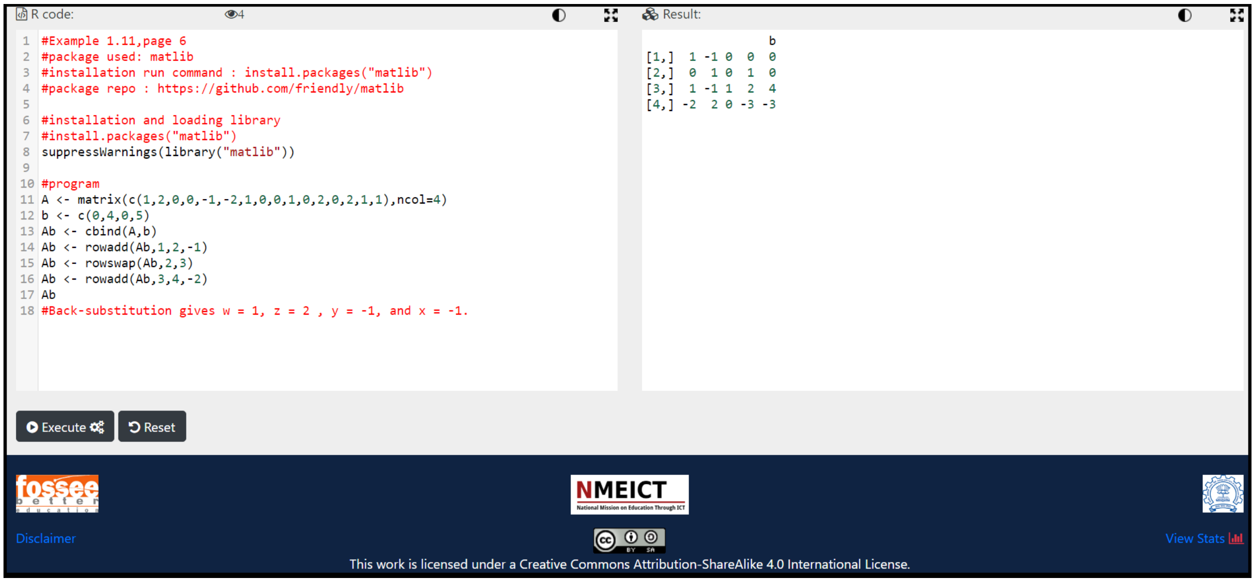


```r
###page_no_878###
rm(list=ls())
library(rpart)
library(rpart.plot)
son<-c("s1","s2","s3")
a1<-c(100000,100000,100000)
a2<-c(-50000,80000,180000)
a3<-c(150000,90000,40000)

p<-factor(c(rep("a1",3),rep("a2",3),rep("a3",3)))
values<-c(a1,a2,a3)
df<-data.frame(p,son,values);df
library(collapsibleTree)
collapsibleTree(df,hierarchy=c("p","son","values"),collapsed=F)

p_son<-c(0.2,0.5,0.3)

EMV_a1<-sum(p_son*a1);EMV_a1
EMV_a2<-sum(p_son*a2);EMV_a2
EMV_a3<-sum(p_son*a3);EMV_a3
SON<-c("s1 0.2","s2 0.5","s3 0.3")
P<-factor(c(rep("a1 100000",3),rep("a2 84000",3),rep("a3 87000",3)))
```

```
   p son values
1 a1  s1 100000
2 a1  s2 100000
3 a1  s3 100000
4 a2  s1 -50000
5 a2  s2  80000
6 a2  s3 180000
7 a3  s1 150000
8 a3  s2  90000
9 a3  s3  40000
[1] 1e+05
[1] 84000
[1] 87000
[1] 34000
[1] 50000
[1] 47000
             P     SON values
1 a1 100000 s1 0.2 100000
2 a1 100000 s2 0.5 100000
3 a1 100000 s3 0.3 100000
4  a2 84000 s1 0.2 -50000
5  a2 84000 s2 0.5  80000
6  a2 84000 s3 0.3 180000
```

Figure: 2.6: Successful TBC code execution after installing “collapsibleTree” library [1].

2. Suppressed unnecessary warning messages by making use of the “suppressWarnings()” function.



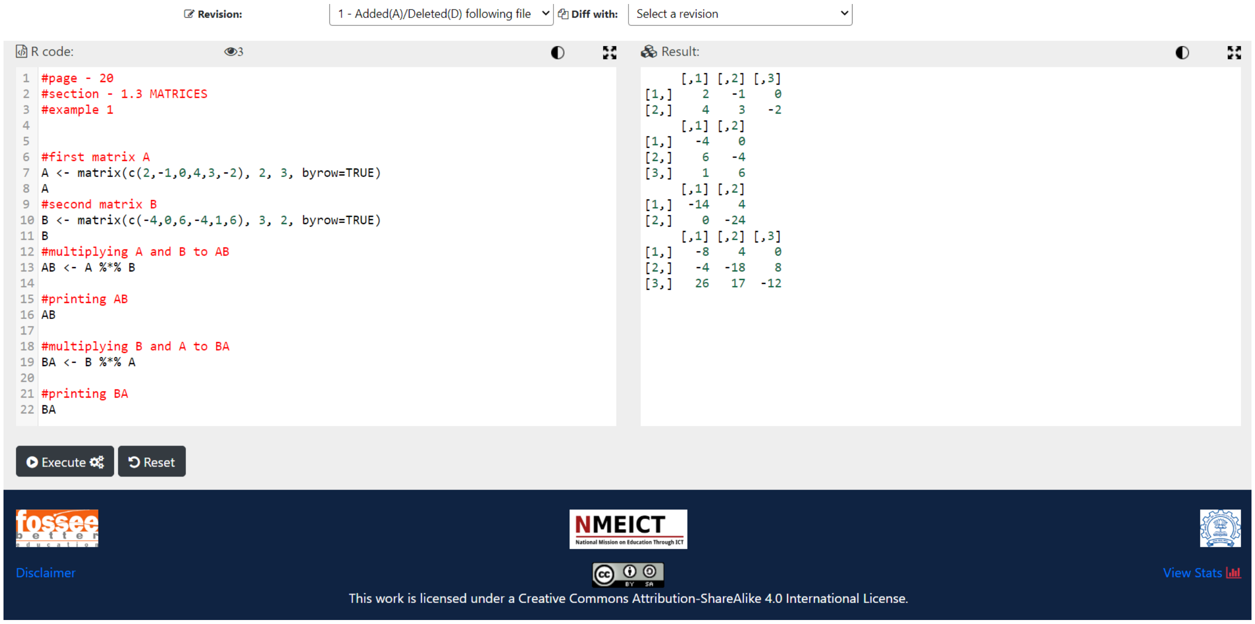
```r
#Example 1.11,page 6
#package used: matlib
#installation run command : install.packages("matlib")
#package repo : https://github.com/friendly/matlib

#installation and loading library
#install.packages("matlib")
suppressWarnings(library("matlib"))

#program
A <- matrix(c(1,2,0,0,-1,-2,1,0,0,1,0,2,0,2,1,1),ncol=4)
b <- c(0,4,0,5)
Ab <- cbind(A,b)
Ab <- rowadd(Ab,1,2,-1)
Ab <- rowswap(Ab,2,3)
Ab <- rowadd(Ab,3,4,-2)
Ab
#Back-substitution gives w = 1, z = 2 , y = -1, and x = -1.
```

```
                b
[1,]  1 -1 0  0  0
[2,]  0  1 0  1  0
[3,]  1 -1 1  2  4
[4,] -2  2 0 -3 -3
```

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure: 2.7: Successful execution of TBC code after suppressing the warning message.

3. R codes were extracted from zip files and made available over the platform.



Revision: 1 - Added(A)/Deleted(D) following file

Diff with: Select a revision

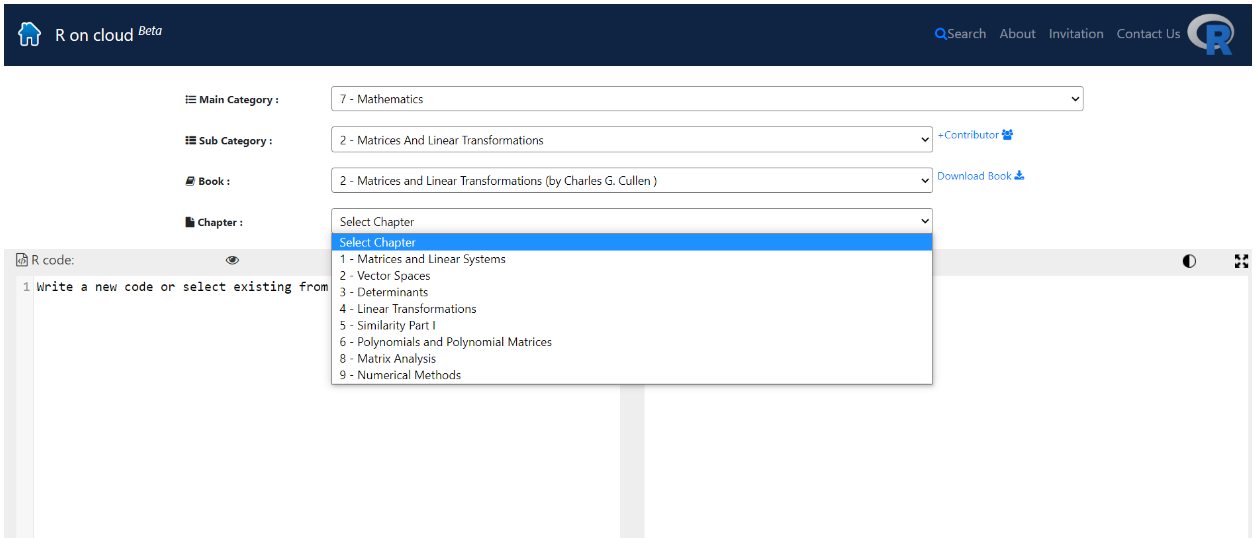
```r
#page - 20
#section - 1.3 MATRICES
#example 1


#first matrix A
A <- matrix(c(2,-1,0,4,3,-2), 2, 3, byrow=TRUE)
A
#second matrix B
B <- matrix(c(-4,0,6,-4,1,6), 3, 2, byrow=TRUE)
B
#multiplying A and B to AB
AB <- A %*% B

#printing AB
AB

#multiplying B and A to BA
BA <- B %*% A

#printing BA
BA
```

```
     [,1] [,2] [,3]
[1,]    2   -1    0
[2,]    4    3   -2
     [,1] [,2]
[1,]   -4    0
[2,]    6   -4
[3,]    1    6
     [,1] [,2]
[1,]  -14    4
[2,]    0  -24
     [,1] [,2] [,3]
[1,]   -8    4    0
[2,]   -4  -18    8
[3,]   26   17  -12
```

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure: 2.8: Execution of R code extracted from zip file.

4. Removed TBC chapters that do not contain any code files.



Main Category : 7 - Mathematics

Sub Category : 2 - Matrices And Linear Transformations

Book : 2 - Matrices and Linear Transformations (by Charles G. Cullen)

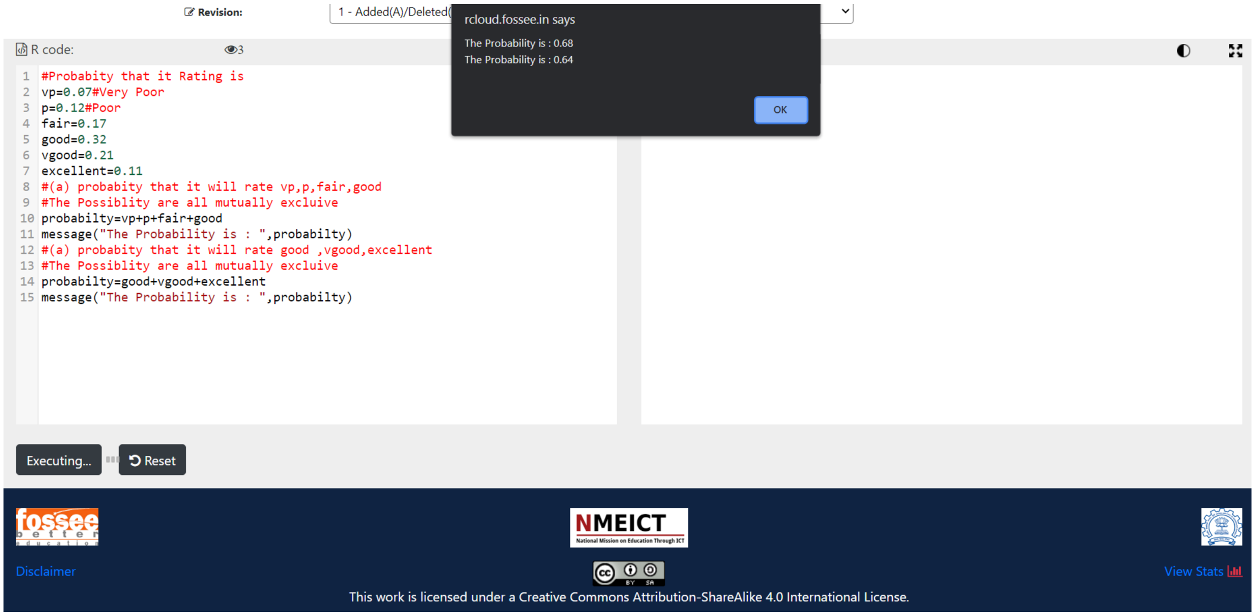
Chapter : Select Chapter

- Select Chapter
- 1 - Matrices and Linear Systems
- 2 - Vector Spaces
- 3 - Determinants
- 4 - Linear Transformations
- 5 - Similarity Part I
- 6 - Polynomials and Polynomial Matrices
- 8 - Matrix Analysis
- 9 - Numerical Methods

```
Write a new code or select existing from
```

Figure: 2.9: Remaining TBC chapters after removal of chapter number 7.

5. Fixed R codes causing the missing object error.



Revision: 1 - Added(A)/Deleted(

```
#Probabity that it Rating is
vp=0.07#Very Poor
p=0.12#Poor
fair=0.17
good=0.32
vgood=0.21
excellent=0.11
#(a) probabity that it will rate vp,p,fair,good
#The Possiblity are all mutually excluive
probabilty=vp+p+fair+good
message("The Probability is : ",probabilty)
#(a) probabity that it will rate good ,vgood,excellent
#The Possiblity are all mutually excluive
probabilty=good+vgood+excellent
message("The Probability is : ",probabilty)
```

rcloud.fossee.in says

```
The Probability is : 0.68
The Probability is : 0.64
```

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure: 2.10: Successful execution of R code after fixing the missing object error.

Chapter 3
Analysis of FOSSEE workshop feedback data

1. Introduction

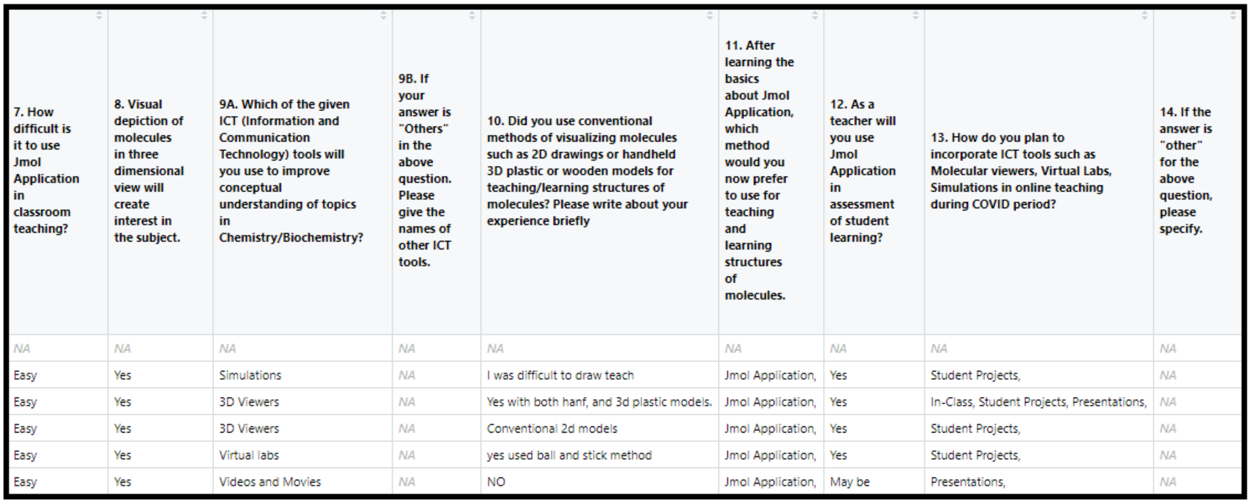
The FOSSEE project promotes the use of FLOSS tools in academia and research. It conducts regular workshops on different FLOSS to help industry professionals, faculty, researchers, and students from various institutions shift from proprietary to open-source software. These workshops are conducted throughout the year and 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) [2]. 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 Jmol Application Beginner Workshop conducted on 19 September 2020. A total of 68 participants attended the workshop, out of which 63 filled the feedback form. The form consisted of 31 questions related to the participants' educational background, job history, and workshop experience. The questions associated with the workshop experience consisted of sub-sections corresponding to workshop activity, practice problems, spoken tutorials, knowledge gained from the workshop, 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 "extremely easy to understand and follow" or "extremely easy" and "5" represented "extremely difficult to understand and follow" or "extremely difficult" depending upon the question.

3. Data Exploration

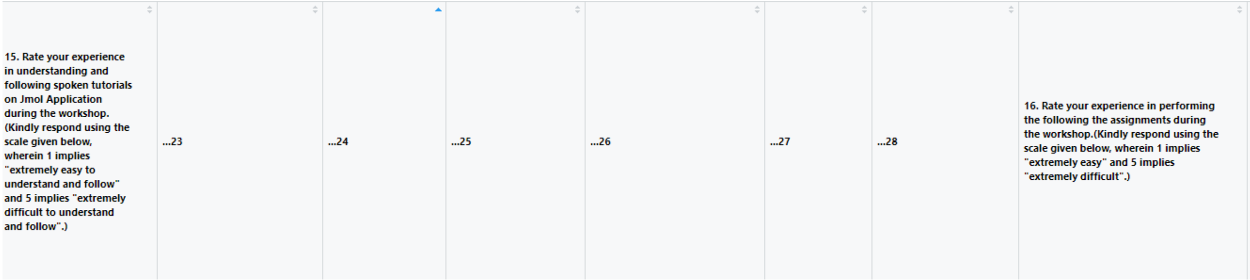
The feedback data was originally in an XLS format. It was loaded into the R environment using the "read_excel()" function from "readxl" package [3], which belongs to the "tidyverse" ecosystem of packages [4]. A glimpse of the original dataset can be seen in Figure 3.1.



7. How difficult is it to use Jmol Application in classroom teaching?	8. Visual depiction of molecules in three dimensional view will create interest in the subject.	9A. Which of the given ICT (Information and Communication Technology) tools will you use to improve conceptual understanding of topics in Chemistry/Biochemistry?	9B. If your answer is "Others" in the above question. Please give the names of other ICT tools.	10. Did you use conventional methods of visualizing molecules such as 2D drawings or handheld 3D plastic or wooden models for teaching/learning structures of molecules? Please write about your experience briefly	11. After learning the basics about Jmol Application, which method would you now prefer to use for teaching and learning structures of molecules.	12. As a teacher will you use Jmol Application in assessment of student learning?	13. How do you plan to incorporate ICT tools such as Molecular viewers, Virtual Labs, Simulations in online teaching during COVID period?	14. If the answer is "other" for the above question, please specify.
NA	NA	NA	NA	NA	NA	NA	NA	NA
Easy	Yes	Simulations	NA	I was difficult to draw teach	Jmol Application,	Yes	Student Projects,	NA
Easy	Yes	3D Viewers	NA	Yes with both hanf, and 3d plastic models.	Jmol Application,	Yes	In-Class, Student Projects, Presentations,	NA
Easy	Yes	3D Viewers	NA	Conventional 2d models	Jmol Application,	Yes	Student Projects,	NA
Easy	Yes	Virtual labs	NA	yes used ball and stick method	Jmol Application,	Yes	Student Projects,	NA
Easy	Yes	Videos and Movies	NA	NO	Jmol Application,	May be	Presentations,	NA

Figure 3.1: Jmol Application Beginner Workshop feedback data used for analysis.

The dataset mentioned above consisted of 64 rows and 67 columns. As stated earlier, the feedback form questions contained sub-sections, resulting in 67 distinct columns instead of 31. It was 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.



15. Rate your experience in understanding and following spoken tutorials on Jmol Application during the workshop. (Kindly respond using the scale given below, wherein 1 implies "extremely easy to understand and follow" and 5 implies "extremely difficult to understand and follow".)	...23	...24	...25	...26	...27	...28	16. Rate your experience in performing the following the assignments during the workshop.(Kindly respond using the scale given below, wherein 1 implies "extremely easy" and 5 implies "extremely difficult".)

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

It felt necessary to rename the columns for two reasons -

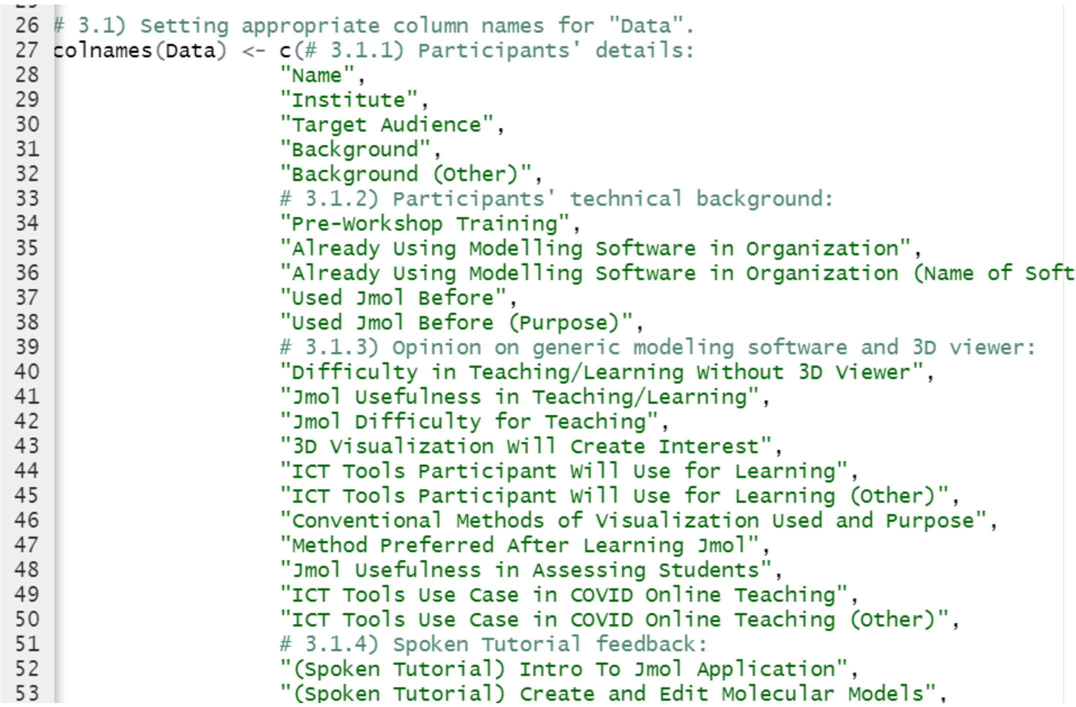
1. To convey information about the associated question in fewer words.
2. To replace the column name containing a number with information about the associated question.

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. Opinion on generic modeling software and 3D viewer
4. Spoken Tutorial feedback
5. Assignment feedback

6. Knowledge of Jmol concepts before and after the workshop
7. Overall workshop feedback

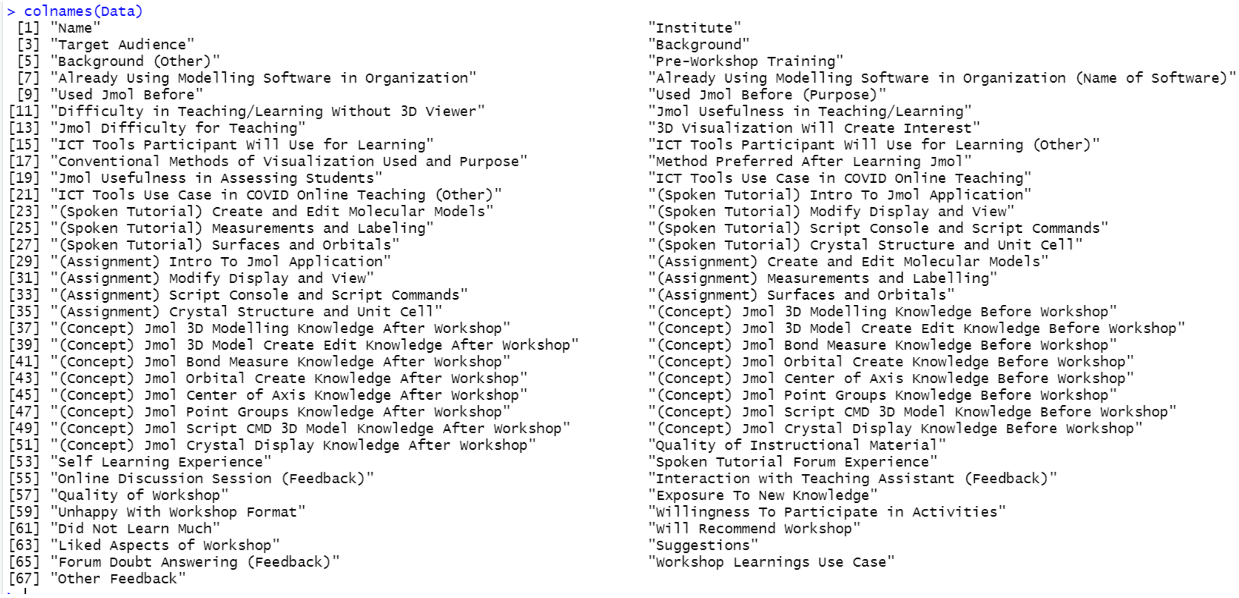
The following code was used to rename the columns -



```r
# 3.1) Setting appropriate column names for "Data".
colnames(Data) <- c(# 3.1.1) Participants' details:
                    "Name",
                    "Institute",
                    "Target Audience",
                    "Background",
                    "Background (Other)",
                    # 3.1.2) Participants' technical background:
                    "Pre-Workshop Training",
                    "Already Using Modelling Software in Organization",
                    "Already Using Modelling Software in Organization (Name of Software)",
                    "Used Jmol Before",
                    "Used Jmol Before (Purpose)",
                    # 3.1.3) Opinion on generic modeling software and 3D viewer:
                    "Difficulty in Teaching/Learning Without 3D Viewer",
                    "Jmol Usefulness in Teaching/Learning",
                    "Jmol Difficulty for Teaching",
                    "3D Visualization Will Create Interest",
                    "ICT Tools Participant Will Use for Learning",
                    "ICT Tools Participant Will Use for Learning (Other)",
                    "Conventional Methods of Visualization Used and Purpose",
                    "Method Preferred After Learning Jmol",
                    "Jmol Usefulness in Assessing Students",
                    "ICT Tools Use Case in COVID Online Teaching",
                    "ICT Tools Use Case in COVID Online Teaching (Other)",
                    # 3.1.4) Spoken Tutorial feedback:
                    "(Spoken Tutorial) Intro To Jmol Application",
                    "(Spoken Tutorial) Create and Edit Molecular Models",
```

Figure 3.3: Code to rename the feedback dataset columns.

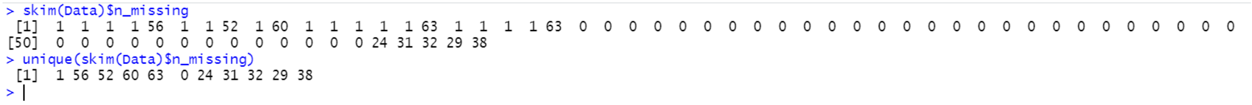
The updated column names are shown in Figure 3.4.



```
> colnames(Data)
 [1] "Name"                                                                 "Institute"
 [3] "Target Audience"                                                      "Background"
 [5] "Background (Other)"                                                   "Pre-Workshop Training"
 [7] "Already Using Modelling Software in Organization"                     "Already Using Modelling Software in Organization (Name of Software)"
 [9] "Used Jmol Before"                                                     "Used Jmol Before (Purpose)"
[11] "Difficulty in Teaching/Learning Without 3D Viewer"                    "Jmol Usefulness in Teaching/Learning"
[13] "Jmol Difficulty for Teaching"                                         "3D Visualization Will Create Interest"
[15] "ICT Tools Participant Will Use for Learning"                          "ICT Tools Participant Will Use for Learning (Other)"
[17] "Conventional Methods of Visualization Used and Purpose"               "Method Preferred After Learning Jmol"
[19] "Jmol Usefulness in Assessing Students"                                "ICT Tools Use Case in COVID Online Teaching"
[21] "ICT Tools Use Case in COVID Online Teaching (Other)"                  "(Spoken Tutorial) Intro To Jmol Application"
[23] "(Spoken Tutorial) Create and Edit Molecular Models"                   "(Spoken Tutorial) Modify Display and View"
[25] "(Spoken Tutorial) Measurements and Labeling"                          "(Spoken Tutorial) Script Console and Script Commands"
[27] "(Spoken Tutorial) Surfaces and Orbitals"                              "(Spoken Tutorial) Crystal Structure and Unit Cell"
[29] "(Assignment) Intro To Jmol Application"                               "(Assignment) Create and Edit Molecular Models"
[31] "(Assignment) Modify Display and View"                                 "(Assignment) Measurements and Labelling"
[33] "(Assignment) Script Console and Script Commands"                      "(Assignment) Surfaces and Orbitals"
[35] "(Assignment) Crystal Structure and Unit Cell"                         "(Concept) Jmol 3D Modelling Knowledge Before Workshop"
[37] "(Concept) Jmol 3D Modelling Knowledge After Workshop"                 "(Concept) Jmol 3D Model Create Edit Knowledge Before Workshop"
[39] "(Concept) Jmol 3D Model Create Edit Knowledge After Workshop"         "(Concept) Jmol Bond Measure Knowledge Before Workshop"
[41] "(Concept) Jmol Bond Measure Knowledge After Workshop"                 "(Concept) Jmol Orbital Create Knowledge Before Workshop"
[43] "(Concept) Jmol Orbital Create Knowledge After Workshop"               "(Concept) Jmol Center of Axis Knowledge Before Workshop"
[45] "(Concept) Jmol Center of Axis Knowledge After Workshop"               "(Concept) Jmol Point Groups Knowledge Before Workshop"
[47] "(Concept) Jmol Point Groups Knowledge After Workshop"                 "(Concept) Jmol Script CMD 3D Model Knowledge Before Workshop"
[49] "(Concept) Jmol Script CMD 3D Model Knowledge After Workshop"          "(Concept) Jmol Crystal Display Knowledge Before Workshop"
[51] "(Concept) Jmol Crystal Display Knowledge After Workshop"              "Quality of Instructional Material"
[53] "Self Learning Experience"                                             "Spoken Tutorial Forum Experience"
[55] "Online Discussion Session (Feedback)"                                 "Interaction with Teaching Assistant (Feedback)"
[57] "Quality of Workshop"                                                  "Exposure To New Knowledge"
[59] "Unhappy With Workshop Format"                                         "Willingness To Participate in Activities"
[61] "Did Not Learn Much"                                                   "Will Recommend Workshop"
[63] "Liked Aspects of Workshop"                                            "Suggestions"
[65] "Forum Doubt Answering (Feedback)"                                     "Workshop Learnings Use Case"
[67] "Other Feedback"
```

Figure 3.4: Updated column names.

Data Exploration continued after updating the column names using the “skim()” function from the “skimr” package [5]. It was observed that there were several missing values in multiple columns of the dataset as shown in Figure 3.5.

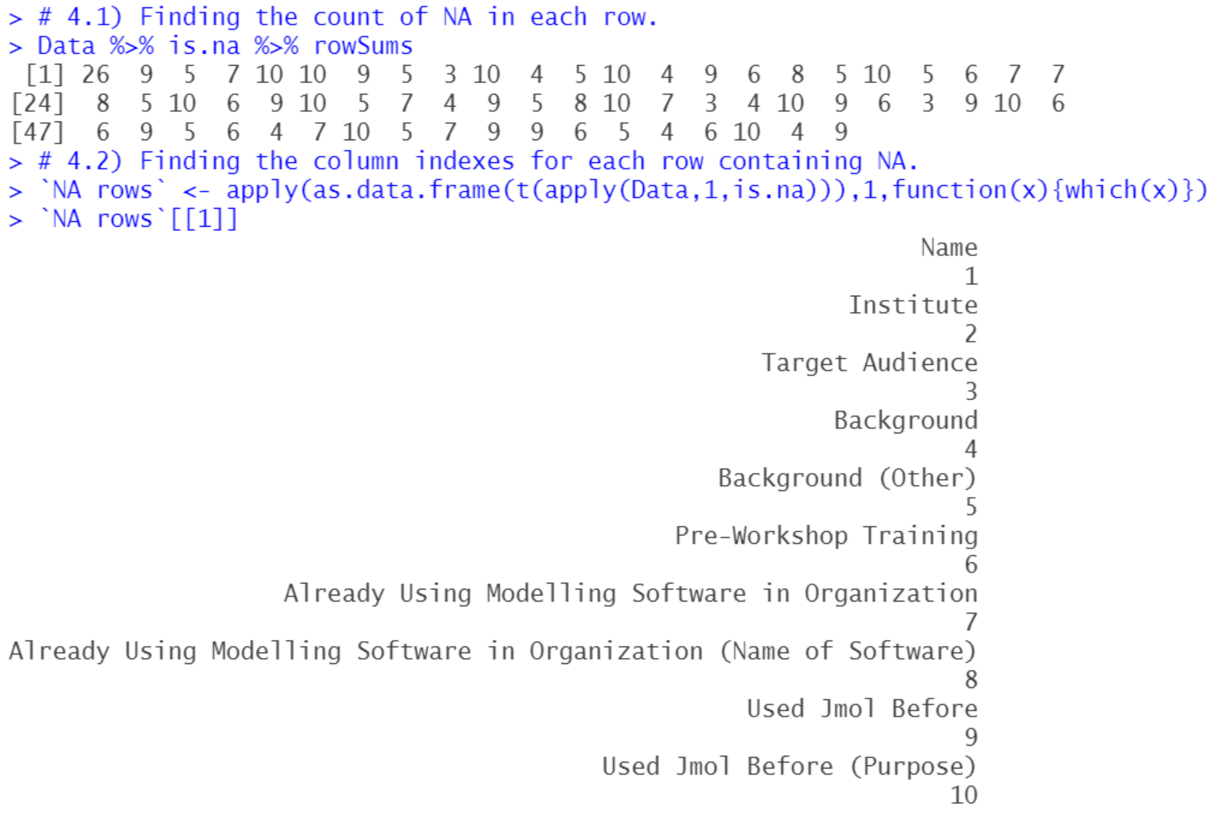


```
> skim(Data)$n_missing
 [1]  1  1  1  1 56  1  1 52  1 60  1  1  1  1  1 63  1  1  1  1 63  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[50]  0  0  0  0  0  0  0  0  0  0  0  0  0 24 31 32 29 38
> unique(skim(Data)$n_missing)
 [1]  1 56 52 60 63  0 24 31 32 29 38
> |
```

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.

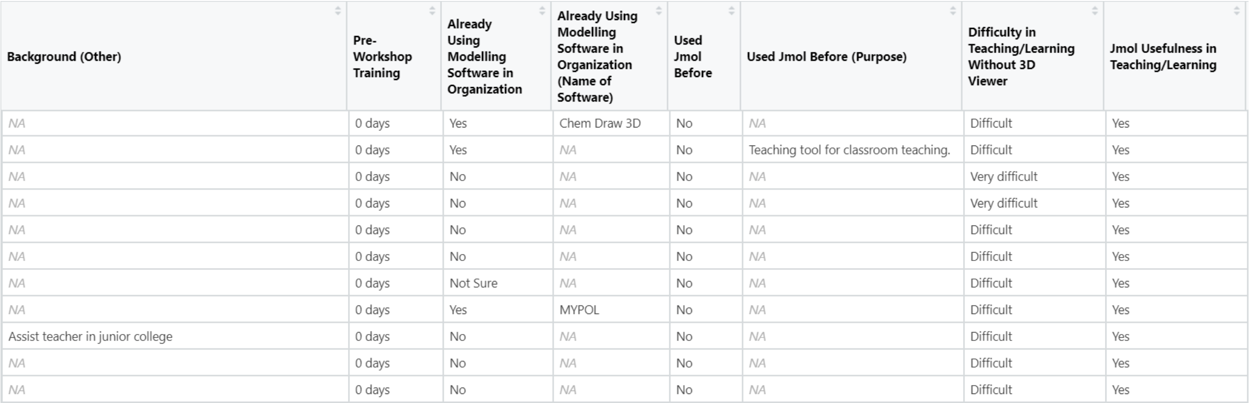
After observing the results of the row-wise check, it was found out that the first row contained the maximum number of missing values, i.e., 26. Further examination of the first row entries showed that it was not a valid feedback response as it did not even contain the participant's name, as shown in Figure 3.6. Hence it was removed.



```
> # 4.1) Finding the count of NA in each row.
> Data %>% is.na %>% rowSums
 [1] 26  9  5  7 10 10  9  5  3 10  4  5 10  4  9  6  8  5 10  5  6  7  7
[24]  8  5 10  6  9 10  5  7  4  9  5  8 10  7  3  4 10  9  6  3  9 10  6
[47]  6  9  5  6  4  7 10  5  7  9  9  6  5  4  6 10  4  9
> # 4.2) Finding the column indexes for each row containing NA.
> `NA rows` <- apply(as.data.frame(t(apply(Data,1,is.na))),1,function(x){which(x)})
> `NA rows`[[1]]
                                                                   Name 
                                                                      1 
                                                              Institute 
                                                                      2 
                                                        Target Audience 
                                                                      3 
                                                             Background 
                                                                      4 
                                                     Background (Other) 
                                                                      5 
                                                  Pre-Workshop Training 
                                                                      6 
                         Already Using Modelling Software in Organization 
                                                                      7 
Already Using Modelling Software in Organization (Name of Software) 
                                                                      8 
                                                        Used Jmol Before 
                                                                      9 
                                                Used Jmol Before (Purpose) 
                                                                     10 
```

Figure 3.6: Row-wise examination of missing values.

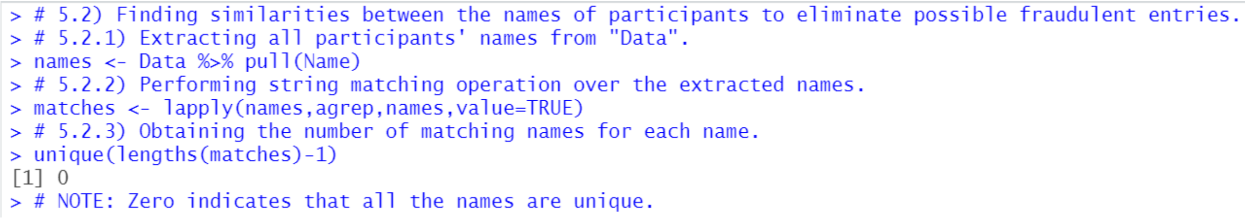
Column-wise examination showed that most of the columns containing missing values were non-numeric and contained subjective comments from the participants, as shown in Figure 3.7. Hence the columns were kept unchanged.



Background (Other)	Pre-Workshop Training	Already Using Modelling Software in Organization	Already Using Modelling Software in Organization (Name of Software)	Used Jmol Before	Used Jmol Before (Purpose)	Difficulty in Teaching/Learning Without 3D Viewer	Jmol Usefulness in Teaching/Learning
NA	0 days	Yes	Chem Draw 3D	No	NA	Difficult	Yes
NA	0 days	Yes	NA	No	Teaching tool for classroom teaching.	Difficult	Yes
NA	0 days	No	NA	No	NA	Very difficult	Yes
NA	0 days	No	NA	No	NA	Very difficult	Yes
NA	0 days	No	NA	No	NA	Difficult	Yes
NA	0 days	No	NA	No	NA	Difficult	Yes
NA	0 days	Not Sure	NA	No	NA	Difficult	Yes
NA	0 days	Yes	MYPOL	No	NA	Difficult	Yes
Assist teacher in junior college	0 days	No	NA	No	NA	Difficult	Yes
NA	0 days	No	NA	No	NA	Difficult	Yes
NA	0 days	No	NA	No	NA	Difficult	Yes

Figure 3.7: Examining missing values column-wise.

After dealing with missing values, the dataset was also checked for duplicate entries as they can introduce bias in statistical analyses [6,7]. 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 [8] 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 3.8, it is shown that the result of string matching turns out to be zero indicating that there are no matching names.



```
> # 5.2) Finding similarities between the names of participants to eliminate possible fraudulent entries.
> # 5.2.1) Extracting all participants' names from "Data".
> names <- Data %>% pull(Name)
> # 5.2.2) Performing string matching operation over the extracted names.
> matches <- lapply(names,agrep,names,value=TRUE)
> # 5.2.3) Obtaining the number of matching names for each name.
> unique(lengths(matches)-1)
[1] 0
> # NOTE: Zero indicates that all the names are unique.
```

Figure 3.8: String matching results for the "Name" column.

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.

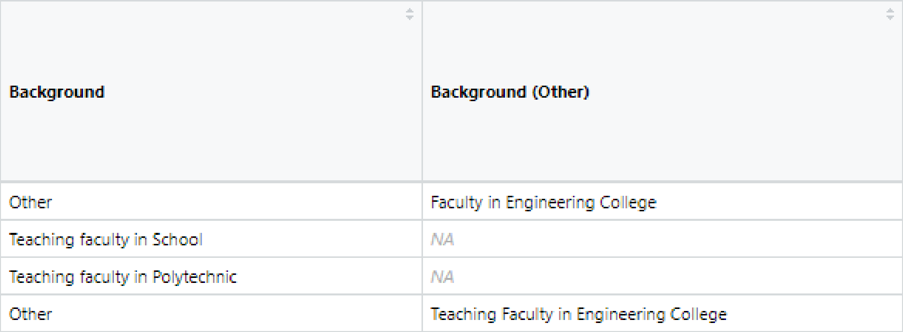
4. Data Cleaning

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 [9,10]. 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 JMOL and similar tools -**

- Checked whether any participant who selected the option "Other" in the "Background" column gave its description in the subsequent column, i.e., "Background (Other)" or not.

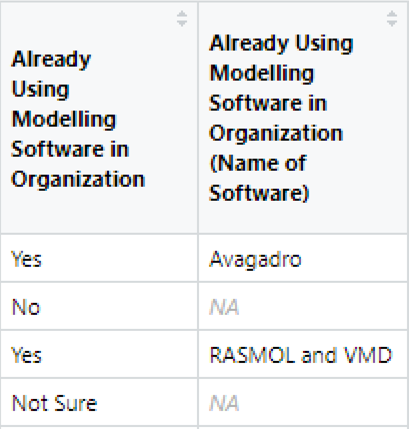


Background	Background (Other)
Other	Faculty in Engineering College
Teaching faculty in School	NA
Teaching faculty in Polytechnic	NA
Other	Teaching Faculty in Engineering College

Figure 3.9: Participants' background-related column entries.

- Checked for entries where a participant first selected "Yes" when asked whether he/she was already using any modeling software but did not provide the name of the software in the column "Already Using Modelling Software in Organization (Name of Software)".

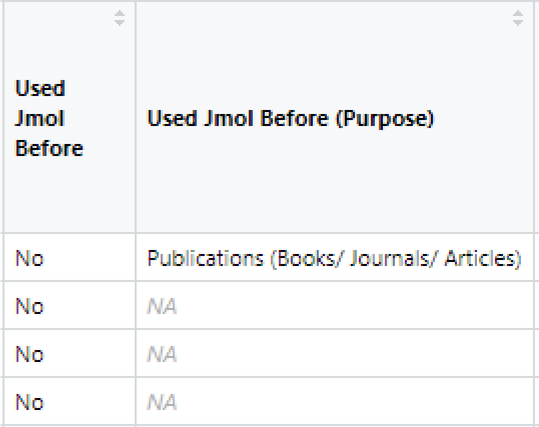
- Checked for entries where a participant first selected "No" when asked whether he/she was already using any modeling software but added the name of a software in the column "Already Using Modelling Software in Organization (Name of Software)".



Already Using Modelling Software in Organization	Already Using Modelling Software in Organization (Name of Software)
Yes	Avagadro
No	NA
Yes	RASMOL and VMD
Not Sure	NA

Figure 3.10: Columns containing entries related to participants' experience with modeling software.

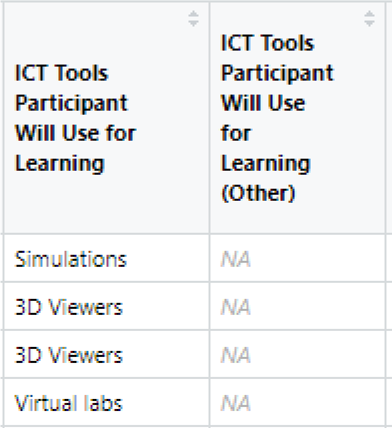
- Checked for entries where a participant initially stated that he/she had never used Jmol before but later mentioned the purpose of using Jmol.



Used Jmol Before	Used Jmol Before (Purpose)
No	Publications (Books/ Journals/ Articles)
No	NA
No	NA
No	NA

Figure 3.11: Columns indicating participants' experience with the Jmol software.

- Checked whether any participant added a response under the "ICT Tools Participant Will Use for Learning (Other)" column without initially adding "Other" in the previous column.

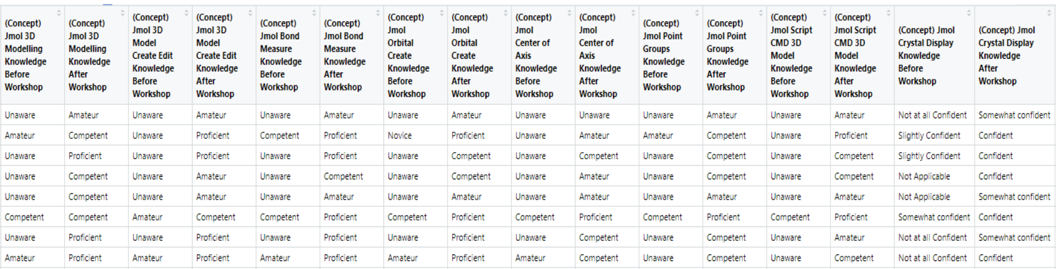


ICT Tools Participant Will Use for Learning	ICT Tools Participant Will Use for Learning (Other)
Simulations	NA
3D Viewers	NA
3D Viewers	NA
Virtual labs	NA

Figure 3.12: Columns associated with entries indicating participants' willingness to use ICT tools.

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

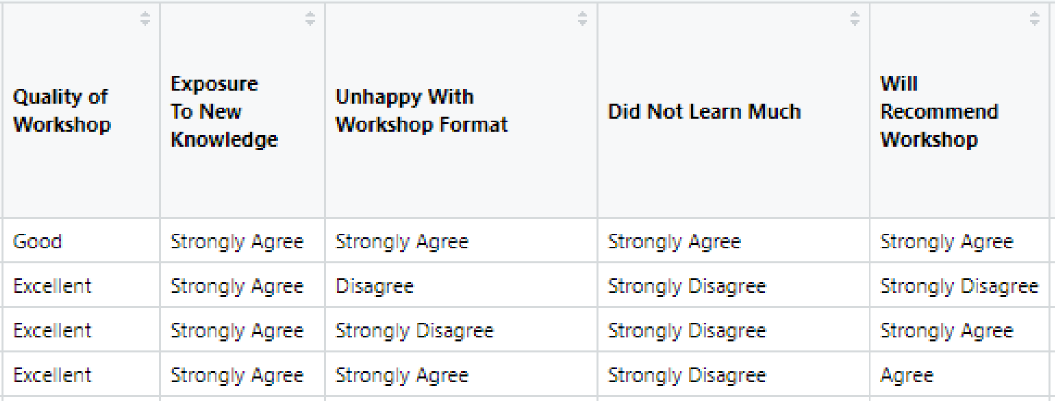
- Checked if the level of knowledge regarding Jmol for any participant dropped after the workshop, as it is improbable.



(Concept) Jmol 3D Modelling Knowledge Before Workshop	(Concept) Jmol 3D Modelling Knowledge After Workshop	(Concept) Jmol 3D Model Create Edit Knowledge Before Workshop	(Concept) Jmol 3D Model Create Edit Knowledge After Workshop	(Concept) Jmol Bond Measure Knowledge Before Workshop	(Concept) Jmol Bond Measure Knowledge After Workshop	(Concept) Jmol Orbital Create Knowledge Before Workshop	(Concept) Jmol Orbital Create Knowledge After Workshop	(Concept) Jmol Center of Axis Knowledge Before Workshop	(Concept) Jmol Center of Axis Knowledge After Workshop	(Concept) Jmol Point Groups Knowledge Before Workshop	(Concept) Jmol Point Groups Knowledge After Workshop	(Concept) Jmol Script CMD 3D Model Knowledge Before Workshop	(Concept) Jmol Script CMD 3D Model Knowledge After Workshop	(Concept) Jmol Crystal Display Knowledge Before Workshop	(Concept) Jmol Crystal Display Knowledge After Workshop
Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Unaware	Unaware	Unaware	Amateur	Unaware	Amateur	Not at all Confident	Somewhat confident
Amateur	Competent	Unaware	Proficient	Competent	Proficient	Novice	Proficient	Unaware	Amateur	Amateur	Competent	Unaware	Proficient	Slightly Confident	Confident
Unaware	Proficient	Unaware	Proficient	Unaware	Proficient	Unaware	Competent	Unaware	Competent	Unaware	Competent	Unaware	Competent	Slightly Confident	Confident
Unaware	Competent	Unaware	Amateur	Unaware	Competent	Unaware	Competent	Unaware	Amateur	Unaware	Competent	Unaware	Competent	Not Applicable	Confident
Unaware	Competent	Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Unaware	Amateur	Not Applicable	Somewhat confident
Competent	Competent	Amateur	Competent	Competent	Proficient	Competent	Proficient	Competent	Proficient	Competent	Proficient	Competent	Proficient	Somewhat confident	Confident
Unaware	Proficient	Unaware	Proficient	Unaware	Proficient	Unaware	Proficient	Unaware	Competent	Unaware	Competent	Unaware	Amateur	Not at all Confident	Somewhat confident
Amateur	Proficient	Amateur	Proficient	Amateur	Proficient	Amateur	Proficient	Amateur	Competent	Unaware	Competent	Unaware	Competent	Not at all Confident	Confident

Figure 3.13: Entries associated with the conceptual knowledge regarding various aspects of Jmol.

- 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 selected the option "Strongly Agree" for both "Exposure To New Knowledge" and "Did Not Learn Much" columns.

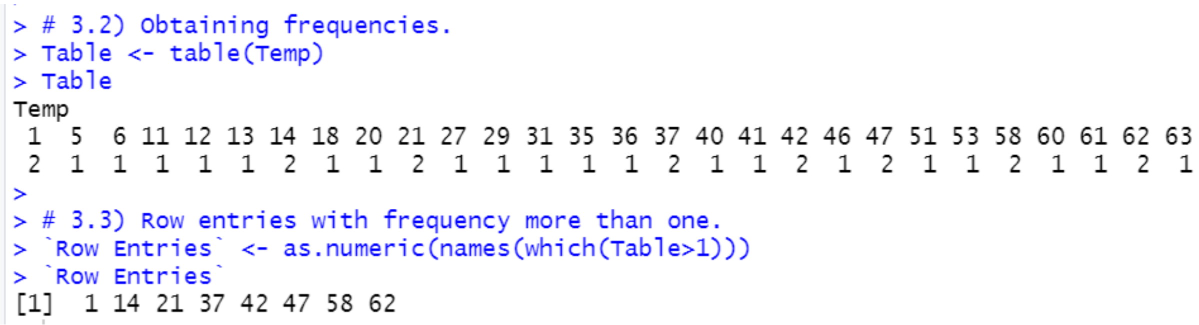


Quality of Workshop	Exposure To New Knowledge	Unhappy With Workshop Format	Did Not Learn Much	Will Recommend Workshop
Good	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree
Excellent	Strongly Agree	Disagree	Strongly Disagree	Strongly Disagree
Excellent	Strongly Agree	Strongly Disagree	Strongly Disagree	Strongly Agree
Excellent	Strongly Agree	Strongly Agree	Strongly Disagree	Agree

Figure 3.14: Column entries associated with the workshop's quality and effectiveness.

3. **Removed misleading entries.**

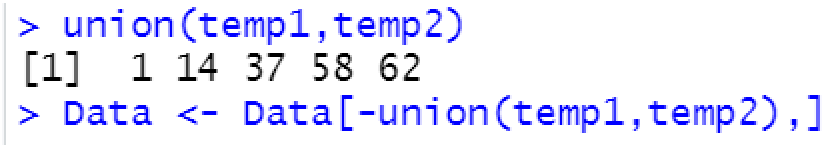
- After recording all potentially misleading entries, the mentor was consulted. As per his advice, the row-wise frequency of misleading responses was calculated and then those row entries for which the frequency value was more than one were examined. The "table()" function of R was used to calculate the frequency, as shown in Figure 3.15.



```
> # 3.2) Obtaining frequencies.
> Table <- table(Temp)
> Table
Temp
 1  5  6 11 12 13 14 18 20 21 27 29 31 35 36 37 40 41 42 46 47 51 53 58 60 61 62 63
 2  1  1  1  1  1  2  1  1  2  1  1  1  1  1  2  1  1  2  1  2  1  1  2  1  1  2  1
>
> # 3.3) Row entries with frequency more than one.
> `Row Entries` <- as.numeric(names(which(Table>1)))
> `Row Entries`
[1]  1 14 21 37 42 47 58 62
```

Figure 3.15: Row entries selected for further examination.

- After carefully examining the remaining eight selected row entries, the mentor suggested to remove five out of them from the feedback data. Those five row entries are shown in Figure 3.16.



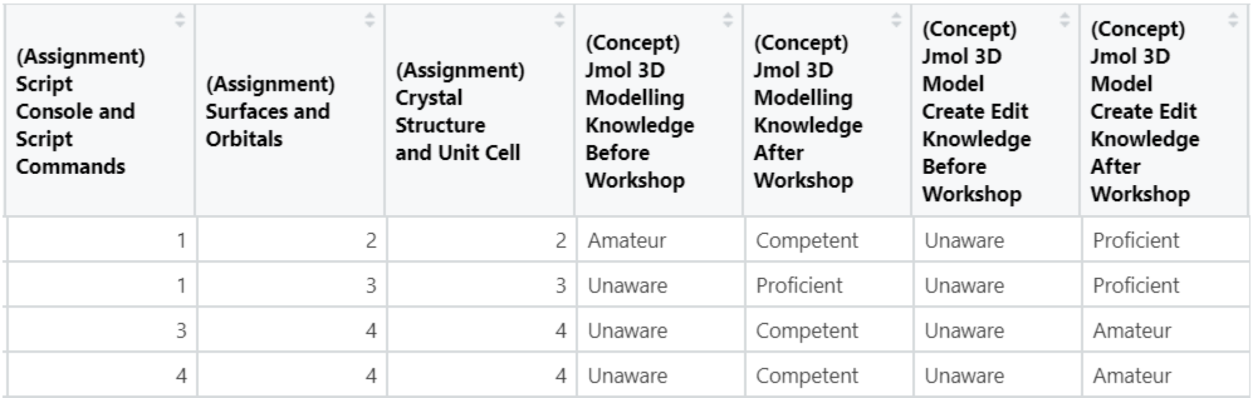
```
> union(temp1,temp2)
[1]  1 14 37 58 62
> Data <- Data[-union(temp1,temp2),]
```

Figure 3.16: Row entries that were removed from the feedback data after careful examination.

After the removal of misleading responses, the dataset was left with 58 rows and 67 columns.

5. Data Pre-processing

Only the Likert scale based columns containing categorical responses were kept from the cleaned Jmol workshop feedback dataset for EFA, as shown in Figure 3.17 [11,12]. Any information related to the participants' backgrounds and their subjective comments was removed from the dataset. The remaining dataset had 58 rows and 49 columns, where each column had the factor data type.

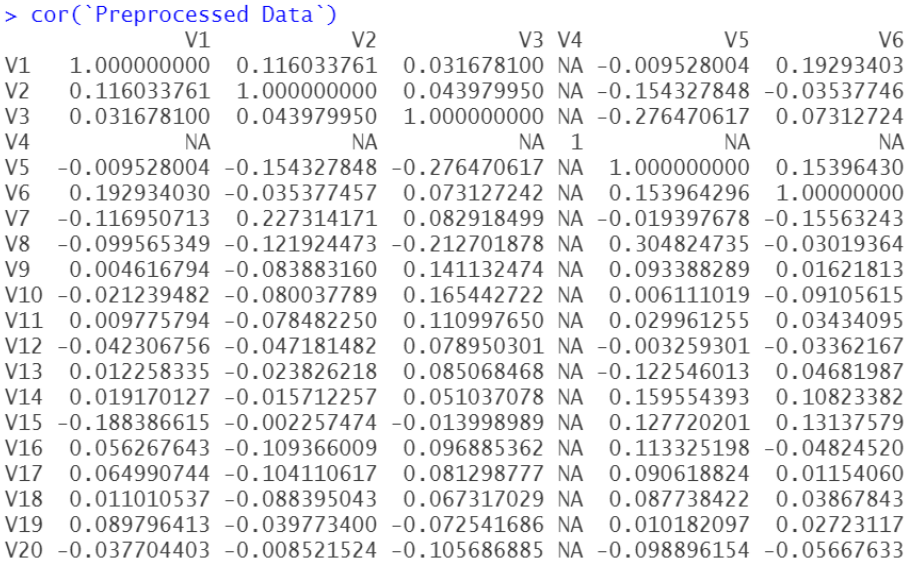


(Assignment) Script Console and Script Commands	(Assignment) Surfaces and Orbitals	(Assignment) Crystal Structure and Unit Cell	(Concept) Jmol 3D Modelling Knowledge Before Workshop	(Concept) Jmol 3D Modelling Knowledge After Workshop	(Concept) Jmol 3D Model Create Edit Knowledge Before Workshop	(Concept) Jmol 3D Model Create Edit Knowledge After Workshop
1	2	2	Amateur	Competent	Unaware	Proficient
1	3	3	Unaware	Proficient	Unaware	Proficient
3	4	4	Unaware	Competent	Unaware	Amateur
4	4	4	Unaware	Competent	Unaware	Amateur

Figure 3.17: Glimpse of the dataset after pre-processing.

6. Data Analysis

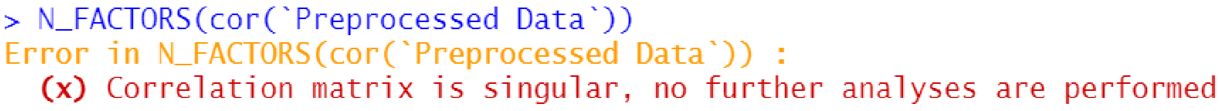
The data analysis was performed using the "EFAtools" package [13]. The "N_FACTORS()" function from "EFAtools" was used to find the suitable number of factors in the data by first converting the data into a numeric format and then finding its correlation matrix. Due to the column "3D Visualization Will Create Interest", the obtained correlation matrix contained NA values, as shown in Figure 3.18.



```
> cor(`Preprocessed Data`)
              V1           V2           V3 V4           V5          V6
V1   1.000000000  0.116033761  0.031678100 NA -0.009528004  0.19293403
V2   0.116033761  1.000000000  0.043979950 NA -0.154327848 -0.03537746
V3   0.031678100  0.043979950  1.000000000 NA -0.276470617  0.07312724
V4            NA           NA           NA  1           NA          NA
V5  -0.009528004 -0.154327848 -0.276470617 NA  1.000000000  0.15396430
V6   0.192934030 -0.035377457  0.073127242 NA  0.153964296  1.00000000
V7  -0.116950713  0.227314171  0.082918499 NA -0.019397678 -0.15563243
V8  -0.099565349 -0.121924473 -0.212701878 NA  0.304824735 -0.03019364
V9   0.004616794 -0.083883160  0.141132474 NA  0.093388289  0.01621813
V10 -0.021239482 -0.080037789  0.165442722 NA  0.006111019 -0.09105615
V11  0.009775794 -0.078482250  0.110997650 NA  0.029961255  0.03434095
V12 -0.042306756 -0.047181482  0.078950301 NA -0.003259301 -0.03362167
V13  0.012258335 -0.023826218  0.085068468 NA -0.122546013  0.04681987
V14  0.019170127 -0.015712257  0.051037078 NA  0.159554393  0.10823382
V15 -0.188386615 -0.002257474 -0.013998989 NA  0.127720201  0.13137579
V16  0.056267643 -0.109366009  0.096885362 NA  0.113325198 -0.04824520
V17  0.064990744 -0.104110617  0.081298777 NA  0.090618824  0.01154060
V18  0.011010537 -0.088395043  0.067317029 NA  0.087738422  0.03867843
V19  0.089796413 -0.039773400 -0.072541686 NA  0.010182097  0.02723117
V20 -0.037704403 -0.008521524 -0.105686885 NA -0.098896154 -0.05667633
```

Figure 3.18: Correlation matrix of the pre-processed dataset.

Therefore, that column was removed and the “N_FACTORS()” function was applied over the correlation matrix obtained from the remaining data. 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 [14]. The KMO value indicates the proportion of variance in the variables that might be caused by underlying factors [15]. 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. Unfortunately, the pre-processed data failed both tests because its correlation matrix was singular, as shown in Figure 3.19.



```
> N_FACTORS(cor(`Preprocessed Data`))
Error in N_FACTORS(cor(`Preprocessed Data`)) :
  (x) Correlation matrix is singular, no further analyses are performed
```

Figure 3.19: Result obtained from the “N_FACTORS()” function.

7. Conclusion

In this project, data exploration, cleaning and preprocessing, were performed over the Jmol Application Beginner Workshop feedback data completely using the R programming language with the objective of applying EFA over it. However, the data turned out to be unsuitable for the proposed analysis as it failed the reliability tests required for EFA. This project could be further extended by using some alternative of EFA, keeping in mind the mixed nature of the feedback data.

Chapter 4
Implementation and documentation of SOM in R

1. Introduction

SOM is an unsupervised data visualization technique popular among researchers for dimensionality reduction and clustering [16]. 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 [17-21]. 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 [22,23].

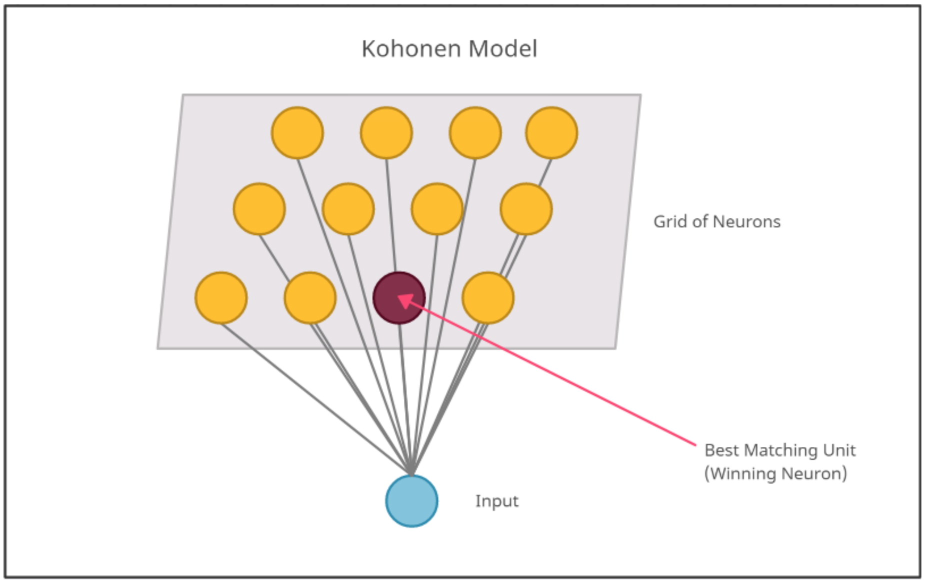


Figure 4.1: Kohonen Model of Self Organizing Map [21].

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. Once the basic SOM model got created, its output was analyzed. 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.

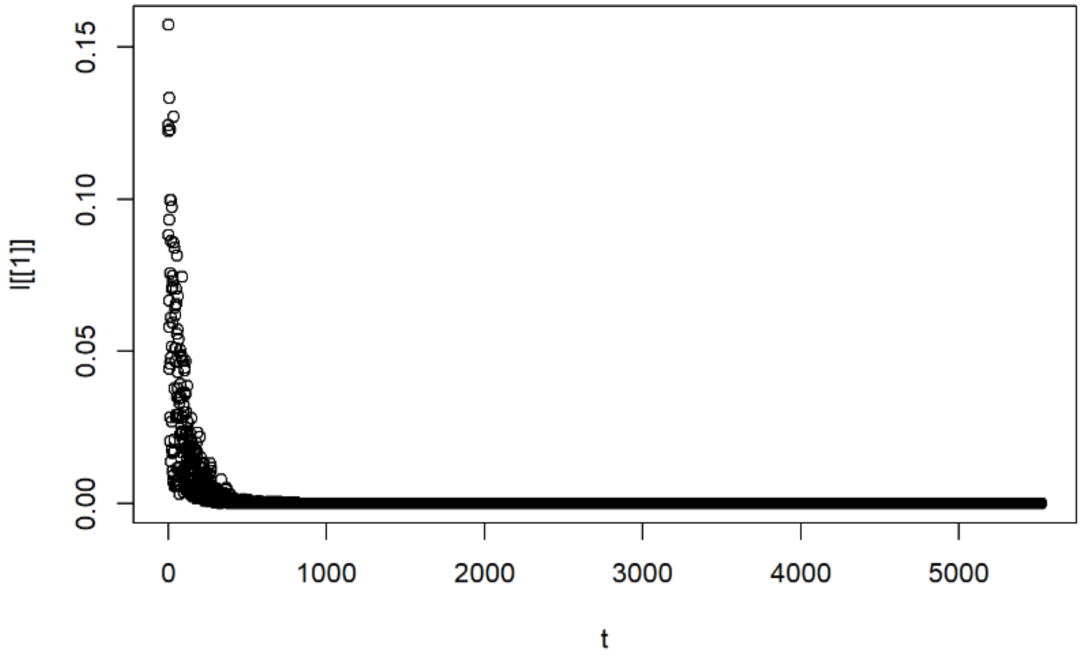


Figure 4.2: Decay in SOM entropy.

The task assigned to me was to identify and implement the convergence criterion of the map. As the SOM ideally is supposed to rapidly change and adjust its weights until it represents the input space accurately, the convergence point is when the map stops adjusting weights and presents the complete lower dimensional representation of the input data. To identify the point of convergence, the difference between weights among two consecutive states of the SOM during training overall epochs was plotted, as shown in Figure 4.2. The downward curve represents a decay in rapidness with which the SOM updates its weights as it adjusts itself more and more to map the input space accurately. As the training progresses, the neighbourhood radii decreases and the map fixates on finer details. However, as a majority of the representation has already been learned, no sudden changes in the plot can be observed. Finally, the convergence of SOM was achieved for the given dataset.

4. Documentation

After the successful implementation of the SOM algorithm in R from scratch, I was assigned to create a document explaining each step in detail with examples. The document was created for the utilization of students, teachers and professionals. The document was in the form of a tutorial series split into four parts for ease of understanding, as shown below -

- Self-Organizing Map in R (Tutorial Introduction)
- Module 1: Introduction to Self-Organizing Map
- Module 2: A solved example of SOM using the Iris dataset
- Module 3: Application of SOM in Real Life: Modeling the COVID-19 Pandemic in India

The first part provides an overview of the tutorial series and the project contributors. The second part or the first module gives an overview of the concept behind SOM and its implementation procedure in R. The third part or the second module deals with the application of SOM over the iris dataset. The final part or the third module concentrates on visualizing India's highly dimensional COVID-19 data on the Indian map as an example of a real-world application of SOM.

Chapter 5

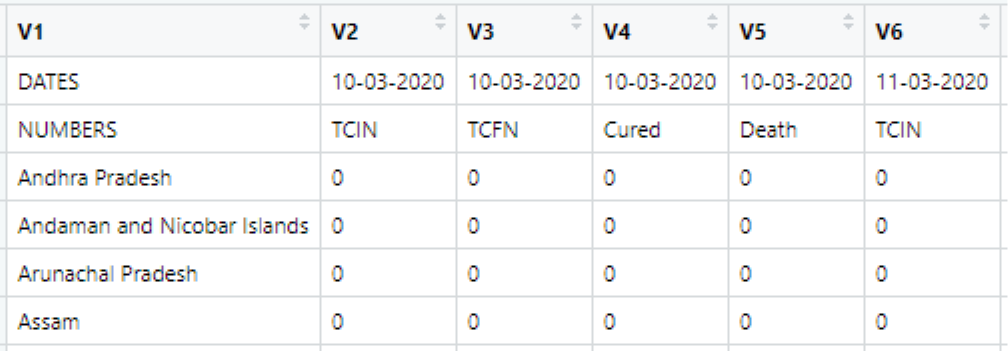
R Case Study: Statistical analysis of the Spatio-temporal trends in COVID-19 strains and infections across India during 2020-21

1. Introduction

The COVID-19 virus spread rapidly across India during 2020-21 and was accompanied by a surge in the number of confirmed cases, deaths and virus mutations. To help researchers and epidemiologists better understand the statistical nature of these trends, I proposed to analyze the data related to the COVID-19 pandemic in India and under the guidance of Prof. Radhendushka Srivastava, created a case study on the same. The entire analysis was performed using the R programming language. A brief description of the complete case study is given in the following sections.

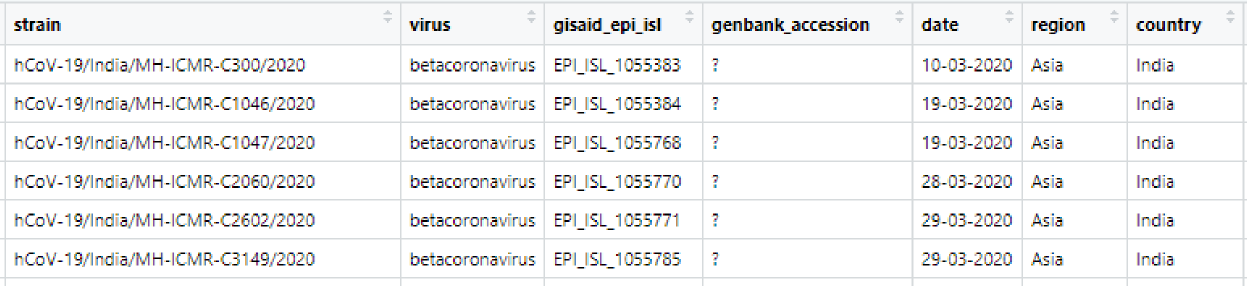
2. Data Collection and Exploration

The COVID-19 data was collected from the Bangalore Centre of the Indian Statistical Institute, Ministry of Family Health and Welfare, Govt. of India and Global Initiative on Sharing Avian Influenza Data (GISAID). The obtained data was in a time-series format and was contained in two separate files. The data was associated with the number of infections and virus strains from 2020 to 2021. Data from both files were later combined based on the date values after performing necessary data cleaning. The data cleaning process involved removing missing values, changing the date format and correcting erroneous values.



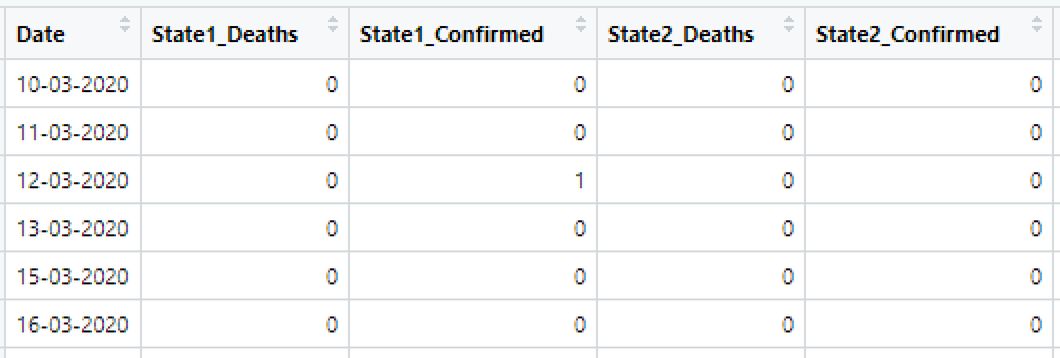
V1	V2	V3	V4	V5	V6
DATES	10-03-2020	10-03-2020	10-03-2020	10-03-2020	11-03-2020
NUMBERS	TCIN	TCFN	Cured	Death	TCIN
Andhra Pradesh	0	0	0	0	0
Andaman and Nicobar Islands	0	0	0	0	0
Arunachal Pradesh	0	0	0	0	0
Assam	0	0	0	0	0

Figure 5.1: COVID-19 infections data.



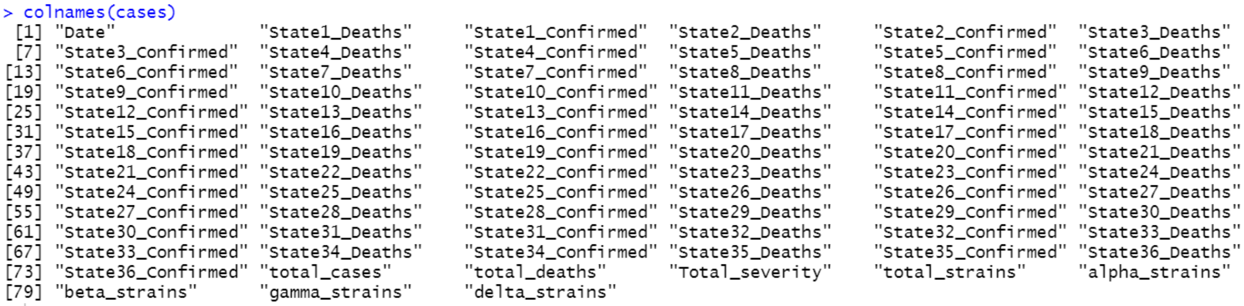
strain	virus	gisaid_epi_isl	genbank_accession	date	region	country
hCoV-19/India/MH-ICMR-C300/2020	betacoronavirus	EPI_ISL_1055383	?	10-03-2020	Asia	India
hCoV-19/India/MH-ICMR-C1046/2020	betacoronavirus	EPI_ISL_1055384	?	19-03-2020	Asia	India
hCoV-19/India/MH-ICMR-C1047/2020	betacoronavirus	EPI_ISL_1055768	?	19-03-2020	Asia	India
hCoV-19/India/MH-ICMR-C2060/2020	betacoronavirus	EPI_ISL_1055770	?	28-03-2020	Asia	India
hCoV-19/India/MH-ICMR-C2602/2020	betacoronavirus	EPI_ISL_1055771	?	29-03-2020	Asia	India
hCoV-19/India/MH-ICMR-C3149/2020	betacoronavirus	EPI_ISL_1055785	?	29-03-2020	Asia	India

Figure 5.2: COVID-19 virus strains data.



Date	State1_Deaths	State1_Confirmed	State2_Deaths	State2_Confirmed
10-03-2020	0	0	0	0
11-03-2020	0	0	0	0
12-03-2020	0	1	0	0
13-03-2020	0	0	0	0
15-03-2020	0	0	0	0
16-03-2020	0	0	0	0

Figure 5.3: Final merged data.



```
> colnames(cases)
 [1] "Date"              "State1_Deaths"     "State1_Confirmed"  "State2_Deaths"     "State2_Confirmed"  "State3_Deaths"
 [7] "State3_Confirmed"  "State4_Deaths"     "State4_Confirmed"  "State5_Deaths"     "State5_Confirmed"  "State6_Deaths"
[13] "State6_Confirmed"  "State7_Deaths"     "State7_Confirmed"  "State8_Deaths"     "State8_Confirmed"  "State9_Deaths"
[19] "State9_Confirmed"  "State10_Deaths"    "State10_Confirmed" "State11_Deaths"    "State11_Confirmed" "State12_Deaths"
[25] "State12_Confirmed" "State13_Deaths"    "State13_Confirmed" "State14_Deaths"    "State14_Confirmed" "State15_Deaths"
[31] "State15_Confirmed" "State16_Deaths"    "State16_Confirmed" "State17_Deaths"    "State17_Confirmed" "State18_Deaths"
[37] "State18_Confirmed" "State19_Deaths"    "State19_Confirmed" "State20_Deaths"    "State20_Confirmed" "State21_Deaths"
[43] "State21_Confirmed" "State22_Deaths"    "State22_Confirmed" "State23_Deaths"    "State23_Confirmed" "State24_Deaths"
[49] "State24_Confirmed" "State25_Deaths"    "State25_Confirmed" "State26_Deaths"    "State26_Confirmed" "State27_Deaths"
[55] "State27_Confirmed" "State28_Deaths"    "State28_Confirmed" "State29_Deaths"    "State29_Confirmed" "State30_Deaths"
[61] "State30_Confirmed" "State31_Deaths"    "State31_Confirmed" "State32_Deaths"    "State32_Confirmed" "State33_Deaths"
[67] "State33_Confirmed" "State34_Deaths"    "State34_Confirmed" "State35_Deaths"    "State35_Confirmed" "State36_Deaths"
[73] "State36_Confirmed" "total_cases"       "total_deaths"      "Total_severity"    "total_strains"     "alpha_strains"
[79] "beta_strains"      "gamma_strains"     "delta_strains"
```

Figure 5.4: Column names of the final dataset.

Before starting the analysis, we visualized the data to explore the features of interest. A plot of data related to the confirmed cases, cured cases and deaths is shown in Figure 5.5.

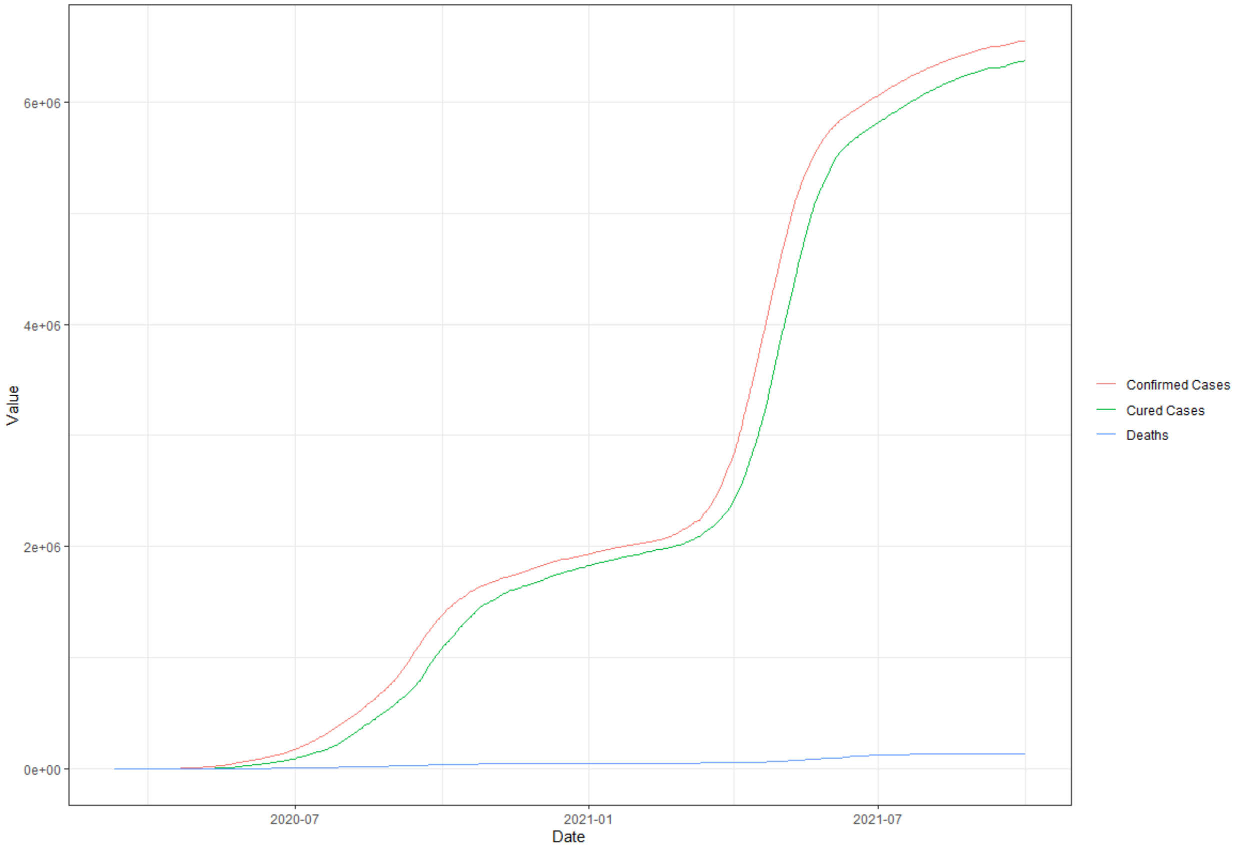


Figure 5.5: Visualization of the primary features of interest.

3. Data Analysis

We first implemented the polynomial curve fitting method [24] over the confirmed cases data to model its underlying pattern, as shown in Figure 5.6. However, it did not yield satisfactory results as the fitted polynomials were too smooth and could not capture the volatility of trends associated with the data.

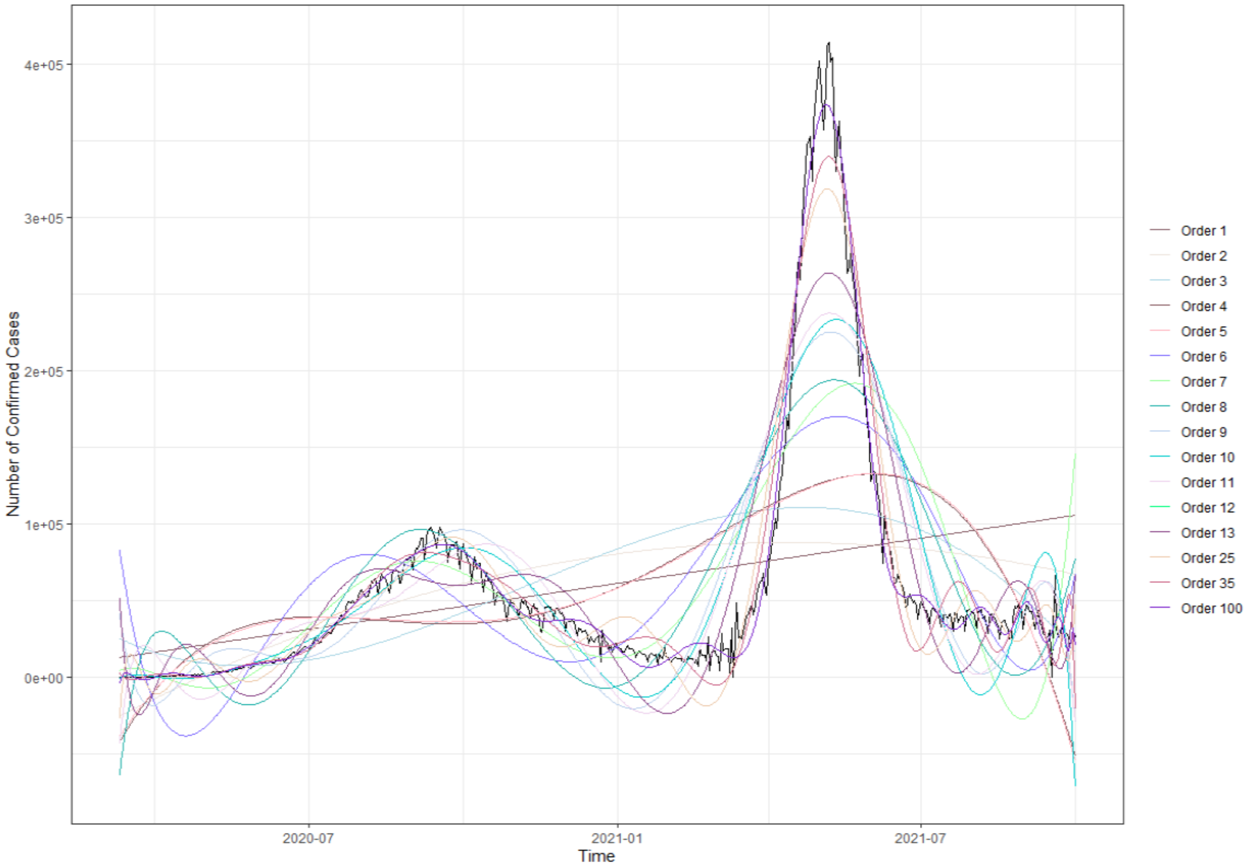


Figure 5.6: Polynomial curve fitting over the confirmed cases data.

Therefore, we experimented with different techniques, namely ARMA-GARCH modeling, piecewise ARMA-GARCH modeling, and Auto-ARIMA forecasting. To apply them, it was necessary first to smoothen the input time series data using a mathematical operator known as Kernel Regression Smoother [25]. The ARMA-GARCH and piecewise ARMA-GARCH models were applied by making use of the "ugarchspec()" and "ugarchfit()" functions from "rugarch" package [26] and the Auto-ARIMA forecasting was done using the "auto.arima()" function from "forecast" package [27].

4. Results

We first applied the ARMA-GARCH model over the entire confirmed cases data, as seen in Figure 5.7. The model performed poorly as the goal was to achieve results with edges preserved. Therefore, we went on to perform piecewise ARMA-GARCH modeling. We observed that the obtained piecewise ARMA-GARCH model gave better results as it successfully captured the volatility in the data. The piecewise ARMA-GARCH model applied on the time series data from 6th August 2020 to 3rd January 2021 is shown in Figure 5.8. However, the variability in the data was not captured fully; thus, we tried the auto-ARIMA forecasting method on the same portion of the data, as shown in Figure 5.9. It was able to forecast the data well. There were differences between the fitted model's forecasted and original values, but the model successfully captured the variability in the trends.

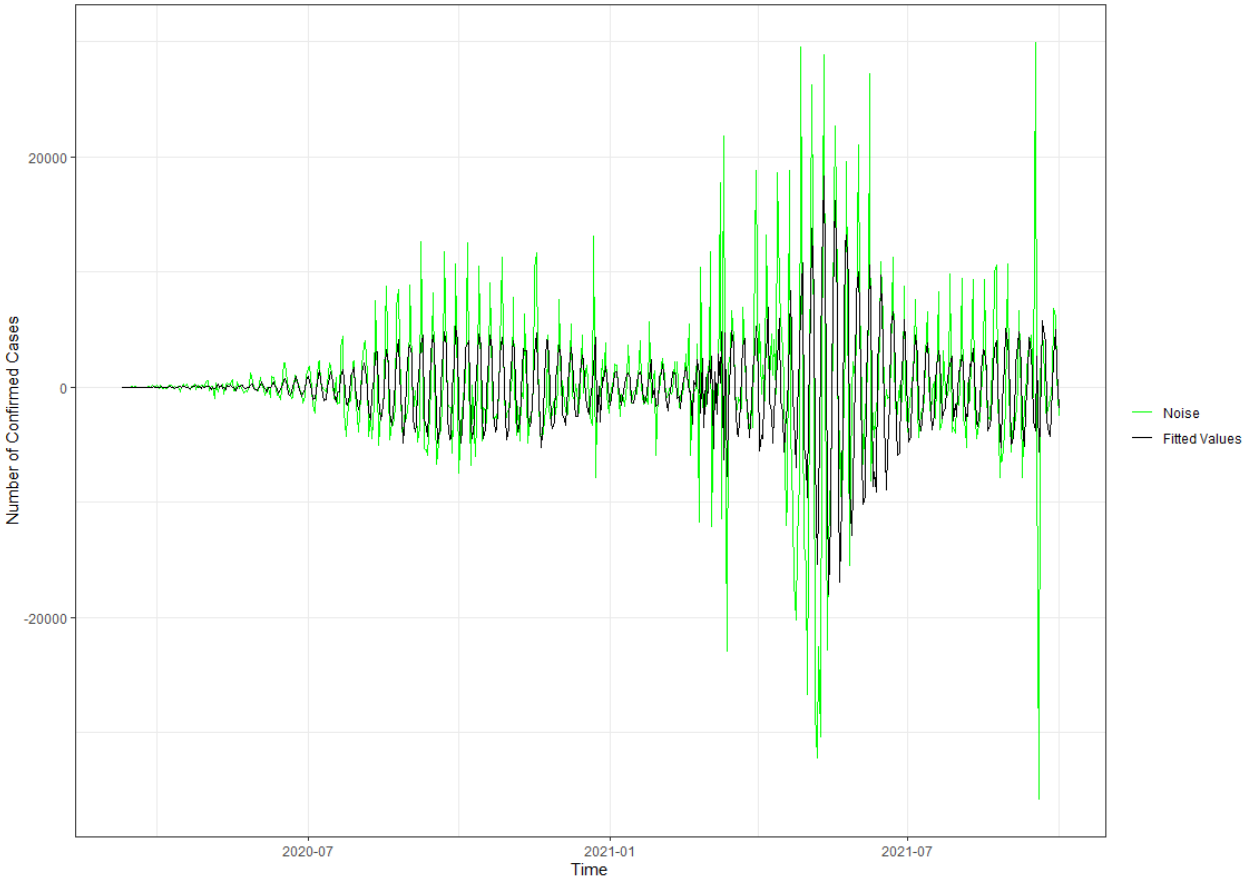


Figure 5.7: ARMA-GARCH model applied over the confirmed cases data.

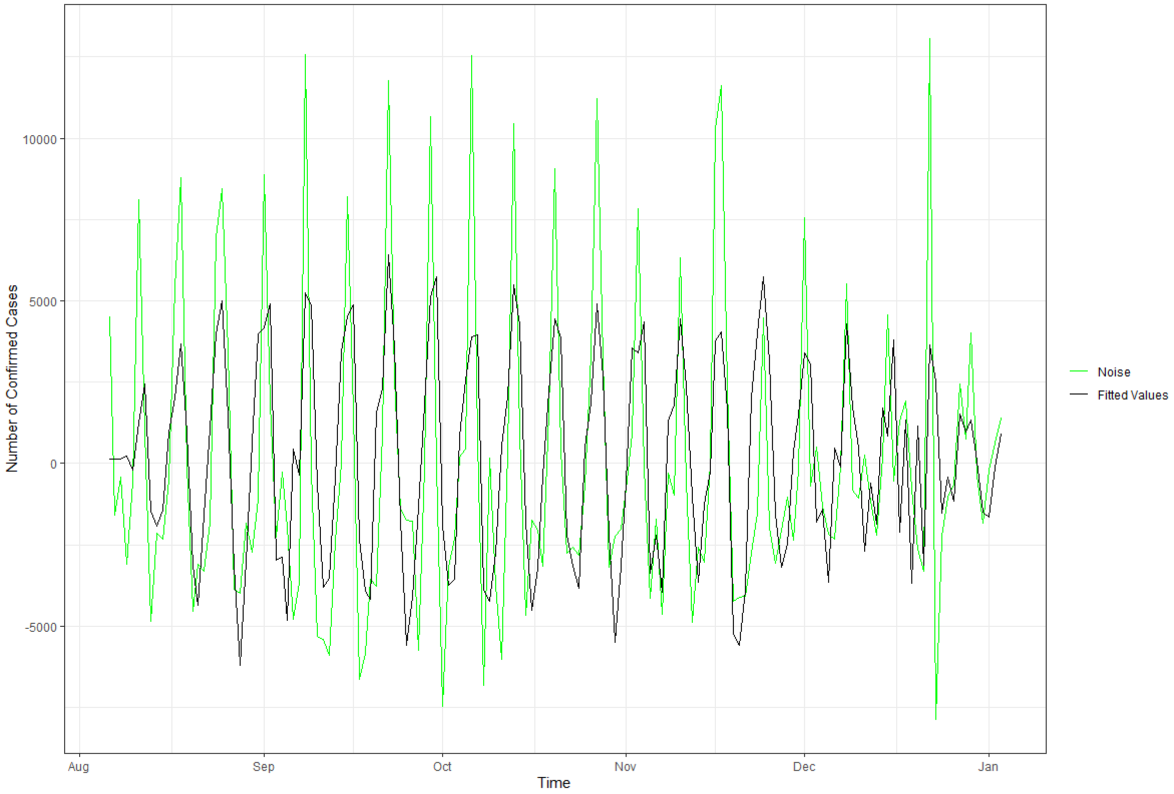


Figure 5.8: Piecewise ARMA-GARCH model applied over a piece of the confirmed cases data.

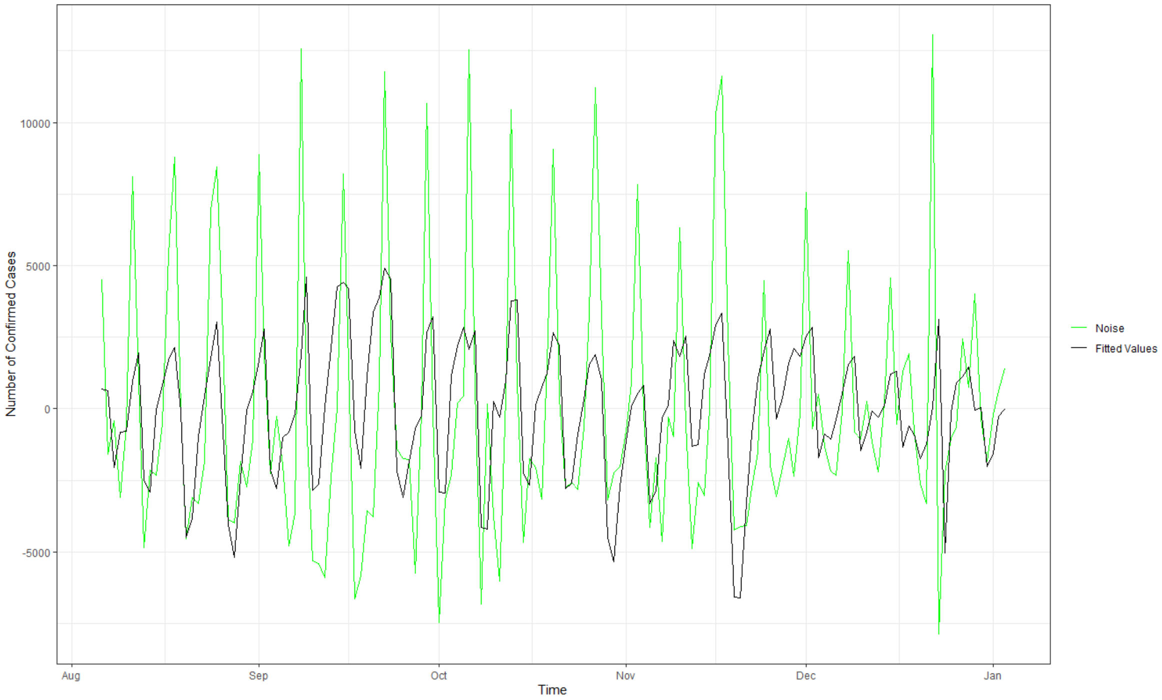


Figure 5.9: Auto-ARIMA forecasting done over a piece of the confirmed cases data.

5. Conclusion

After a thorough procedure consisting of data collection, cleaning, pre-processing and statistical analysis, this case study presents several insights on the COVID-19 pandemic as the applied models were able to capture the underlying pattern in the data. The trend analysis will be helpful to epidemiologists and researchers aiming to study and forecast the pandemic.

Chapter 6
Conclusion

The FOSSEE Semester-long Internship has been a great learning experience for me and my fellow interns, in which we were able to explore several exciting projects in depth. The R on Cloud project encourages the use of R programming language among programmers and assists students engaged in creating TBC of standard textbooks. By maintaining the R on Cloud facility, we hope to enhance user experience and quality of service. In the FOSSEE workshop feedback data analysis project, we realized the importance of data cleaning and achieved it using the various functions provided by R. Creation of the applied tutorial on Self Organizing Map from scratch was a challenging task which pushed us to expand our research and programming skills immensely. The case study project was an independent research project that helped us understand the application and usefulness of statistical modeling.

The FOSSEE internship was a journey of learning and exploration that extended beyond just programming. We were able to utilize the versatility of the R programming language in performing different tasks. Developing interpersonal communication skills, implementing an algorithm from scratch, performing statistical modeling, and learning the documentation process are just a few of the many takeaways from this internship. I hope that my work will benefit society and propagate interest in the R programming language among people from different backgrounds.

References

[1] Adeel Khan (2018). collapsibleTree: Interactive Collapsible Tree Diagrams using 'D3.js'. R package version 0.1.7. https://CRAN.R-project.org/package=collapsibleTree

[2] 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/

[3] Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

[4] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[5] 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. https://CRAN.R-project.org/package=skimr

[6] 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.

[7] 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).

[8] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[9] 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

[10] 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).

[11] 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.

[12] Tarka, P. (2015). Likert Scale and Change in Range of Response Categories vs. the Factors Extraction in EFA Model. Acta Universitatis Lodziensis. Folia Oeconomica, 311.

[13] 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

[14] Watkins, M. (2018). Exploratory factor analysis: A guide to best practice. Journal of Black Psychology, 44(3), p.219–246.

[15] 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

[16] Kevin Pang. Self-organizing Maps. https://www.cs.hmc.edu/~kpang/nn/som.html

[17] Kohonen, Teuvo. "The self-organizing map." Proceedings of the IEEE 78.9 (1990): 1464-1480.

[18] Uoolc, A. Bradford. "Self-organizing Map Formation: Foundations of Neural Computation."

[19] Kohonen, Teuvo. "Essentials of the self-organizing map." Neural networks 37 (2013): 52-65.

[20] Kohonen, Teuvo, and Timo Honkela. "Kohonen network." Scholarpedia 2.1 (2007): 1568.

[21] Sven Krüger. Self-Organizing Maps. https://www.iikt.ovgu.de/iesk_media/Downloads/ks/computational_neuroscience/vorlesung/comp_neuro8-p-2090.pdf

[22] John A. Bullinaria. (2004). Self Organizing Maps: Fundamentals. https://www.cs.bham.ac.uk/~jxb/NN/l16.pdf

[23] Jae-Wook Ahn and Sue Yeon Syn. (2005). Self-Organizing Maps. https://sites.pitt.edu/~is2470pb/Spring05/FinalProjects/Group1a/tutorial/som.html

[24] Nancy Reid. (2010). Statistical Methods for Data Mining and Machine Learning. http://www.utstat.toronto.edu/reid/sta414/Rsession-polys.pdf

[25] Nathaniel E. Helwig. (2021). Nonparametric Regression (Smoothers) in R. http://users.stat.umn.edu/~helwig/notes/smooth-notes.html

[26] Ghalanos, A. "Rugarch: Univariate GARCH models. R package version 1.4-0, accessed 16 January 2019." (2014).

[27] Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i03