Semester-long Internship Report

on

FLOSS - R

submitted by

Tanmay Srinath (BMSCE, Bangalore)

under the guidance of

Prof. Kannan M. Moudgalya	**Prof. Radhendushka Srivastava**
Chemical Engineering Department	Mathematics Department
IIT Bombay	IIT Bombay

and supervision of

Mrs. Smita Wangikar	**Mr. Digvijay Singh**
Project Manager,	Project Research Assistant,
R Team, FOSSEE	R Team, FOSSEE
IIT Bombay	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, and 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 members of the R FLOSS team, namely Mrs. Smita Wangikar and Mr. Digvijay Singh, for helping me shore up my knowledge of data analysis. I also want to thank my fellow interns, Siddhant Raghuvanshi and Aboli Marathe, for their support, intellectual discussions, and enthusiasm.

Contents

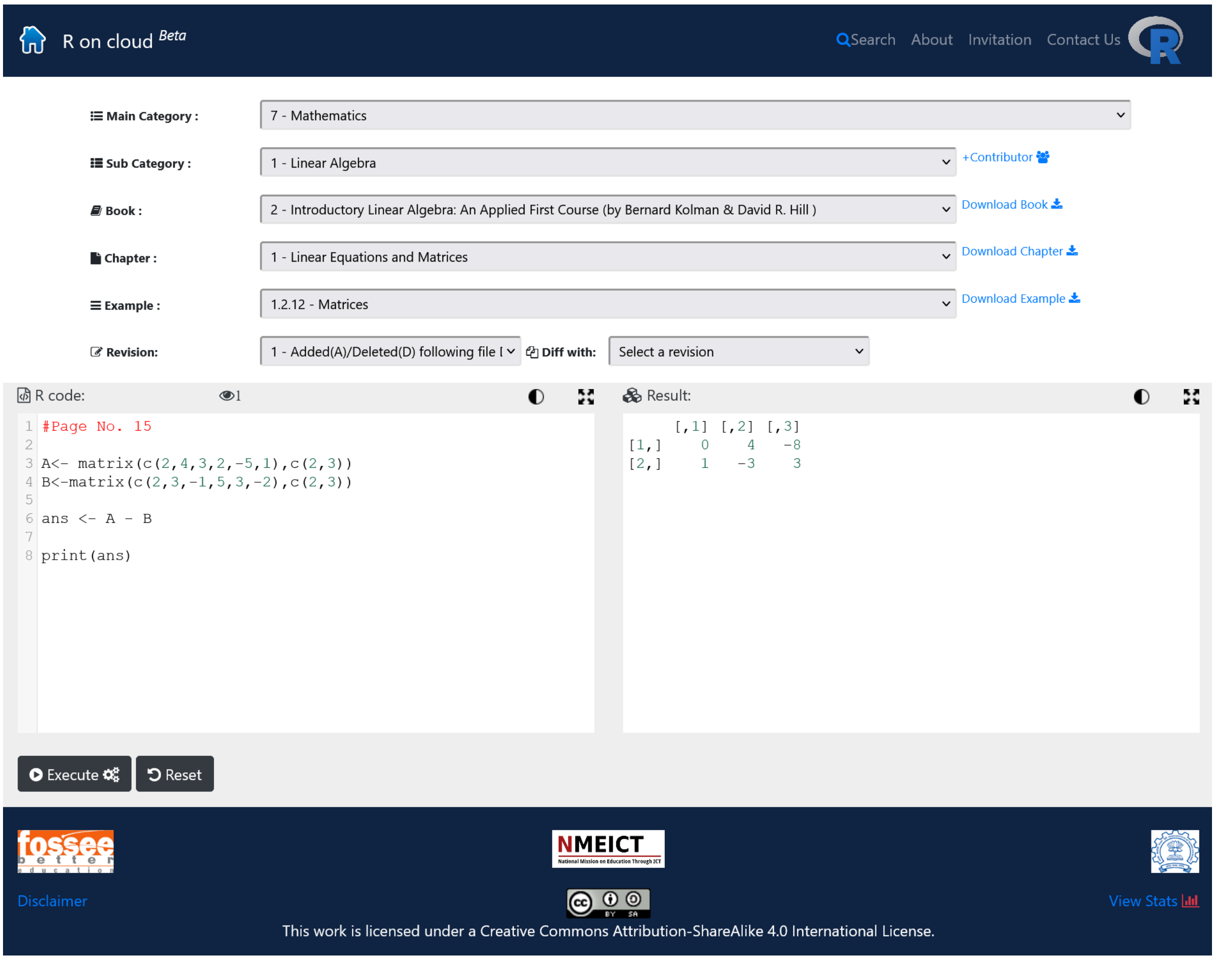
1. Introduction 4
2. Maintenance of R on Cloud 5
3. Analysis of FOSSEE workshop feedback data 7
4. Spoken Tutorial content creation 18
5. Implementation and optimization of SOM algorithm in R 21
6. R Case Study: Analysis and prediction of the impact of COVID-19 on the global economy 23
7. Conclusion 29
8. References 30

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 maintenance of R on Cloud, analysis of FOSSEE workshop feedback data, Spoken Tutorial content creation, implementation and optimization of the SOM algorithm in R, and an R case study on analysis and prediction of the impact of COVID-19 on the global economy.

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.



Main Category : 7 - Mathematics

Sub Category : 1 - Linear Algebra

Book : 2 - Introductory Linear Algebra: An Applied First Course (by Bernard Kolman & David R. Hill)

Chapter : 1 - Linear Equations and Matrices

Example : 1.2.12 - Matrices

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

Diff with: Select a revision

```
#Page No. 15

A<- matrix(c(2,4,3,2,-5,1),c(2,3))
B<-matrix(c(2,3,-1,5,3,-2),c(2,3))

ans <- A - B

print(ans)
```

```
     [,1] [,2] [,3]
[1,]    0    4   -8
[2,]    1   -3    3
```

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

Figure 2.1: R on Cloud by FOSSEE.

For 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 12 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	Operations Research: An Introduction by Hamdy A Taha, Pearson, 2014
2	Probability and Statistics for Engineering and the Sciences by Jay L Devore, Richard Stratton, Boston, USA, 2012

3	Probability and Statistics for Engineers by Richard L. Scheaffer, Madhuri S. Mulekar, James T. McClave, Cengage Learning, USA, 2011
4	Probability and Statistics for Engineers and Scientists by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, Pearson Education, Boston USA, 2016
5	Probability, Random Variables, and Stochastic Processes by Athanasios Papoulis and S. Unnikrishna Pillai, Mcgraw Hill Education (India) Private Limited, 2002
6	Semiconductor Physics And Devices - Basic Principles by D. A. Neamen, Mcgraw-hill, 2003
7	Statistics and Probability Theory by Dr. K.C. Jain and Dr. M.L. Rawat, College Book Centre, Jaipur, 2013
8	Statistics for Business and Economics by Anderson, Sweeney, and Williams, Cengage Learning, USA, 20111
9	Statistics for Management and Economics by Gerald Keller, Cengage Learning USA, 2012
10	Statistics for Psychology by Arthur Aron, Elliot J. Coups, and Elaine N. Aron, Pearson, USA, 2013
11	Statistics in Education and Psychology by P. C. Dash and Bhabagrahi Biswal, Dominant Publishers and Distributors Pvt Ltd, 20091
12	Thermodynamics And Heat Power by I. Granet And M. Bluestein, Addison Wesley (Singapore), New Delhi, 2001

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.

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.

Chapter 3

Analysis of FOSSEE workshop feedback data

1. Introduction

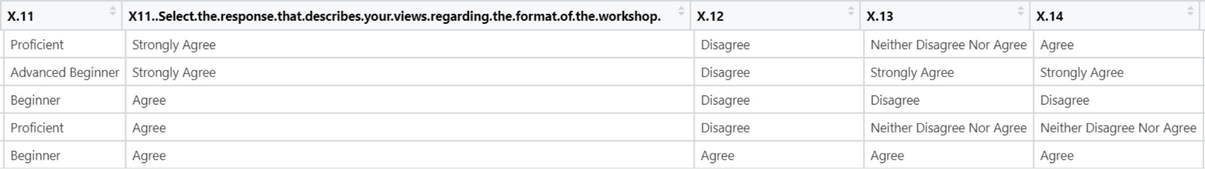
The FOSSEE project promotes the use of FLOSS tools in academia and research. It conducts regular workshops on different FLOSS to help industry professionals, faculty, researchers, and students from various institutions shift from proprietary to open-source software. These workshops are conducted throughout 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) [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

Data was collected through feedback forms for the ChemCollective Virtual Lab Beginner 1 day workshop conducted on 12th December 2020 and the Jmol Application Advanced Workshop conducted on 12th September 2020. Thirty-five people attended the ChemCollective workshop, out of which thirty-two filled the feedback form. Only seventeen people attended the Jmol workshop and all of them filled the feedback form. The feedback form contained questions associated with the workshop experience consisting 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.

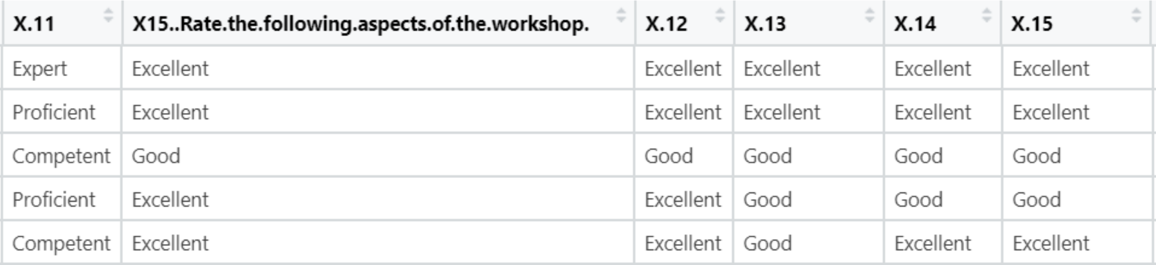
3. Data Exploration

The feedback datasets were loaded into the R environment using the "read.csv()" function. A glimpse of the original datasets can be seen in Figure 3.1 and Figure 3.2.



X.11	X11..Select.the.response.that.describes.your.views.regarding.the.format.of.the.workshop.	X.12	X.13	X.14
Proficient	Strongly Agree	Disagree	Neither Disagree Nor Agree	Agree
Advanced Beginner	Strongly Agree	Disagree	Strongly Agree	Strongly Agree
Beginner	Agree	Disagree	Disagree	Disagree
Proficient	Agree	Disagree	Neither Disagree Nor Agree	Neither Disagree Nor Agree
Beginner	Agree	Agree	Agree	Agree

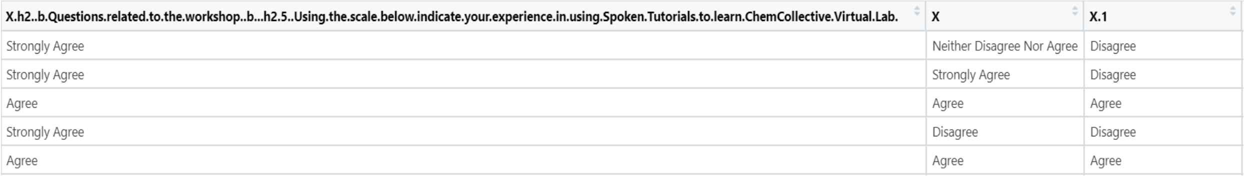
Figure 3.1: Glimpse of the original ChemCollective workshop feedback response dataset.



X.11	X15..Rate.the.following.aspects.of.the.workshop.	X.12	X.13	X.14	X.15
Expert	Excellent	Excellent	Excellent	Excellent	Excellent
Proficient	Excellent	Excellent	Excellent	Excellent	Excellent
Competent	Good	Good	Good	Good	Good
Proficient	Excellent	Excellent	Good	Good	Good
Competent	Excellent	Excellent	Good	Excellent	Excellent

Figure 3.2: Glimpse of the original Jmol workshop feedback response dataset.

The datasets mentioned above contained more columns than their respective feedback form questions because some questions contained sub-sections. The Jmol workshop dataset contained 50 columns, whereas the ChemCollective workshop dataset contained 56 columns. It was also observed that some column names were lengthy and some consisted of only numbers or a few characters that did not provide any information about the question associated with that column, as shown in Figure 3.3.



X.h2..b.Questions.related.to.the.workshop..b...h2.5..Using.the.scale.below.indicate.your.experience.in.using.Spoken.Tutorials.to.learn.ChemCollective.Virtual.Lab.	X	X.1
Strongly Agree	Neither Disagree Nor Agree	Disagree
Strongly Agree	Strongly Agree	Disagree
Agree	Agree	Agree
Strongly Agree	Disagree	Disagree
Agree	Agree	Agree

Figure 3.3: Column names of one of the datasets.

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.

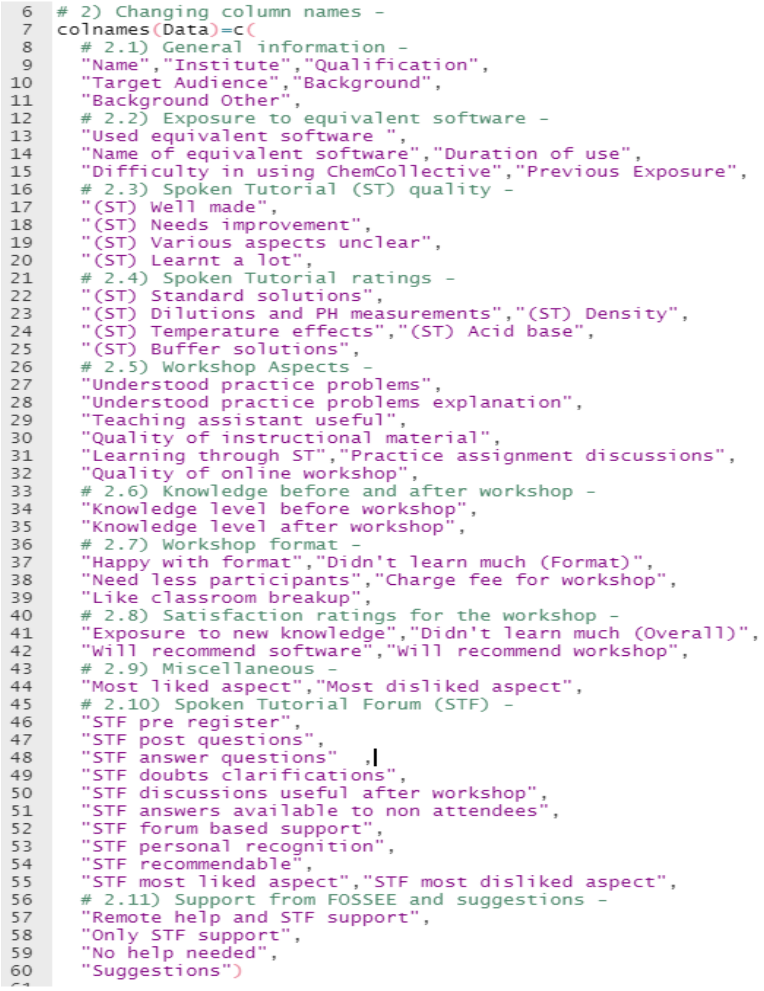
Therefore the columns were renamed and grouped into the following categories, where each category has sub-divisions based on the associated feedback form questions -

1. ChemCollective workshop feedback response dataset -
   - 1.1 General information
   - 1.2 Exposure to equivalent software
   - 1.3 Spoken Tutorial (ST) quality
   - 1.4 Spoken Tutorial ratings
   - 1.5 Workshop Aspects
   - 1.6 Knowledge before and after workshop
   - 1.7 Workshop format
   - 1.8 Satisfaction ratings for the workshop

   1.9 Miscellaneous
   1.10 Spoken Tutorial Forum (STF)
   1.11 Support from FOSSEE and suggestions
2. Jmol workshop feedback response dataset -
   1.1 General information
   1.2 Exposure to Jmol
   1.3 Experience of screening task
   1.4 Spoken Tutorial ratings
   1.5 Live Demonstration ratings
   1.6 Assignment ratings
   1.7 Knowledge before and after workshop
   1.8 Overall quality ratings
   1.9 Workshop satisfaction ratings
   1.10 Miscellaneous

The following code was used to rename the columns -

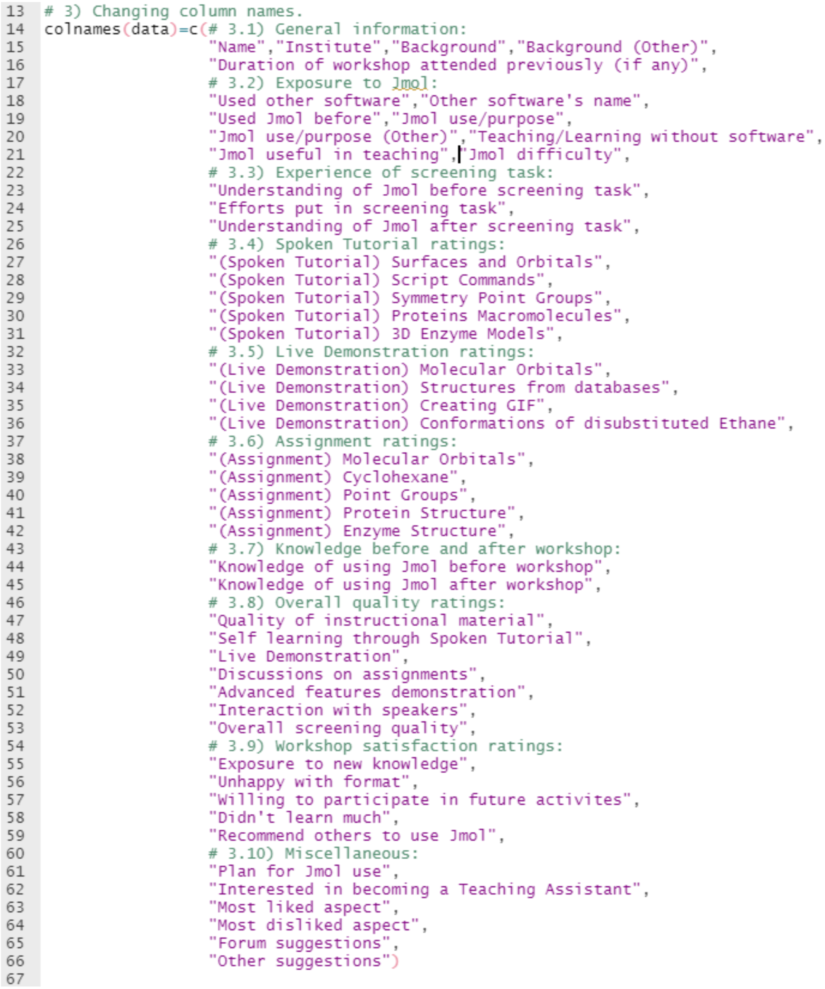
1. ChemCollective workshop feedback response dataset -



```
# 2) Changing column names -
colnames(Data)=c(
  # 2.1) General information -
  "Name","Institute","Qualification",
  "Target Audience","Background",
  "Background Other",
  # 2.2) Exposure to equivalent software -
  "Used equivalent software ",
  "Name of equivalent software","Duration of use",
  "Difficulty in using ChemCollective","Previous Exposure",
  # 2.3) Spoken Tutorial (ST) quality -
  "(ST) Well made",
  "(ST) Needs improvement",
  "(ST) Various aspects unclear",
  "(ST) Learnt a lot",
  # 2.4) Spoken Tutorial ratings -
  "(ST) Standard solutions",
  "(ST) Dilutions and PH measurements","(ST) Density",
  "(ST) Temperature effects","(ST) Acid base",
  "(ST) Buffer solutions",
  # 2.5) Workshop Aspects -
  "Understood practice problems",
  "Understood practice problems explanation",
  "Teaching assistant useful",
  "Quality of instructional material",
  "Learning through ST","Practice assignment discussions",
  "Quality of online workshop",
  # 2.6) Knowledge before and after workshop -
  "Knowledge level before workshop",
  "Knowledge level after workshop",
  # 2.7) Workshop format -
  "Happy with format","Didn't learn much (Format)",
  "Need less participants","Charge fee for workshop",
  "Like classroom breakup",
  # 2.8) Satisfaction ratings for the workshop -
  "Exposure to new knowledge","Didn't learn much (Overall)",
  "Will recommend software","Will recommend workshop",
  # 2.9) Miscellaneous -
  "Most liked aspect","Most disliked aspect",
  # 2.10) Spoken Tutorial Forum (STF) -
  "STF pre register",
  "STF post questions",
  "STF answer questions"   ,
  "STF doubts clarifications",
  "STF discussions useful after workshop",
  "STF answers available to non attendees",
  "STF forum based support",
  "STF personal recognition",
  "STF recommendable",
  "STF most liked aspect","STF most disliked aspect",
  # 2.11) Support from FOSSEE and suggestions -
  "Remote help and STF support",
  "Only STF support",
  "No help needed",
  "Suggestions")
```

Figure 3.4: Code to rename the ChemCollective workshop feedback dataset columns.

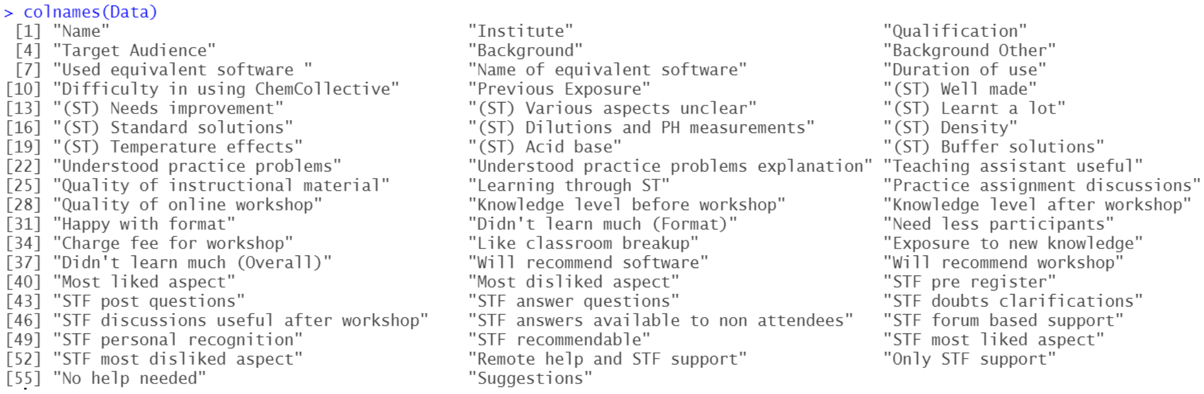
2. Jmol workshop feedback response dataset -



```r
# 3) Changing column names.
colnames(data)=c(# 3.1) General information:
                 "Name","Institute","Background","Background (Other)",
                 "Duration of workshop attended previously (if any)",
                 # 3.2) Exposure to Jmol:
                 "Used other software","Other software's name",
                 "Used Jmol before","Jmol use/purpose",
                 "Jmol use/purpose (Other)","Teaching/Learning without software",
                 "Jmol useful in teaching","Jmol difficulty",
                 # 3.3) Experience of screening task:
                 "Understanding of Jmol before screening task",
                 "Efforts put in screening task",
                 "Understanding of Jmol after screening task",
                 # 3.4) Spoken Tutorial ratings:
                 "(Spoken Tutorial) Surfaces and Orbitals",
                 "(Spoken Tutorial) Script Commands",
                 "(Spoken Tutorial) Symmetry Point Groups",
                 "(Spoken Tutorial) Proteins Macromolecules",
                 "(Spoken Tutorial) 3D Enzyme Models",
                 # 3.5) Live Demonstration ratings:
                 "(Live Demonstration) Molecular Orbitals",
                 "(Live Demonstration) Structures from databases",
                 "(Live Demonstration) Creating GIF",
                 "(Live Demonstration) Conformations of disubstituted Ethane",
                 # 3.6) Assignment ratings:
                 "(Assignment) Molecular Orbitals",
                 "(Assignment) Cyclohexane",
                 "(Assignment) Point Groups",
                 "(Assignment) Protein Structure",
                 "(Assignment) Enzyme Structure",
                 # 3.7) Knowledge before and after workshop:
                 "Knowledge of using Jmol before workshop",
                 "Knowledge of using Jmol after workshop",
                 # 3.8) Overall quality ratings:
                 "Quality of instructional material",
                 "Self learning through Spoken Tutorial",
                 "Live Demonstration",
                 "Discussions on assignments",
                 "Advanced features demonstration",
                 "Interaction with speakers",
                 "Overall screening quality",
                 # 3.9) Workshop satisfaction ratings:
                 "Exposure to new knowledge",
                 "Unhappy with format",
                 "Willing to participate in future activites",
                 "Didn't learn much",
                 "Recommend others to use Jmol",
                 # 3.10) Miscellaneous:
                 "Plan for Jmol use",
                 "Interested in becoming a Teaching Assistant",
                 "Most liked aspect",
                 "Most disliked aspect",
                 "Forum suggestions",
                 "Other suggestions")
```

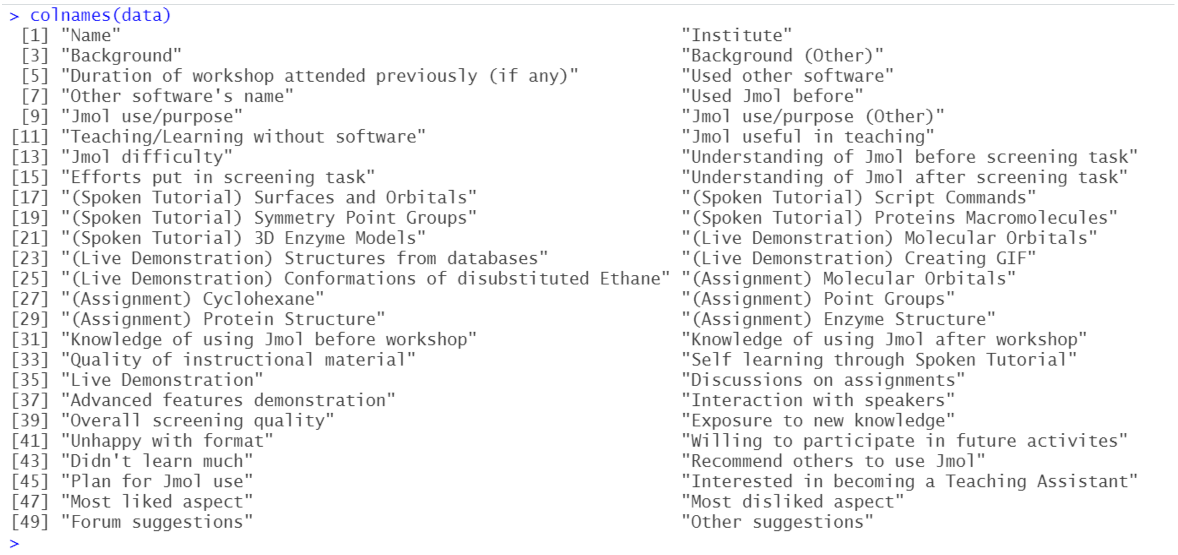
Figure 3.5: Code to rename the Jmol workshop feedback dataset columns.

The updated column names are shown in Figure 3.6 and Figure 3.7.



```
> colnames(Data)
 [1] "Name"                                    "Institute"                                "Qualification"
 [4] "Target Audience"                         "Background"                               "Background Other"
 [7] "Used equivalent software "               "Name of equivalent software"              "Duration of use"
[10] "Difficulty in using ChemCollective"      "Previous Exposure"                        "(ST) Well made"
[13] "(ST) Needs improvement"                  "(ST) Various aspects unclear"             "(ST) Learnt a lot"
[16] "(ST) Standard solutions"                 "(ST) Dilutions and PH measurements"       "(ST) Density"
[19] "(ST) Temperature effects"                "(ST) Acid base"                           "(ST) Buffer solutions"
[22] "Understood practice problems"            "Understood practice problems explanation" "Teaching assistant useful"
[25] "Quality of instructional material"       "Learning through ST"                      "Practice assignment discussions"
[28] "Quality of online workshop"              "Knowledge level before workshop"          "Knowledge level after workshop"
[31] "Happy with format"                       "Didn't learn much (Format)"               "Need less participants"
[34] "Charge fee for workshop"                 "Like classroom breakup"                   "Exposure to new knowledge"
[37] "Didn't learn much (Overall)"             "Will recommend software"                  "Will recommend workshop"
[40] "Most liked aspect"                       "Most disliked aspect"                     "STF pre register"
[43] "STF post questions"                      "STF answer questions"                     "STF doubts clarifications"
[46] "STF discussions useful after workshop"   "STF answers available to non attendees"   "STF forum based support"
[49] "STF personal recognition"                "STF recommendable"                        "STF most liked aspect"
[52] "STF most disliked aspect"                "Remote help and STF support"              "Only STF support"
[55] "No help needed"                          "Suggestions"
```

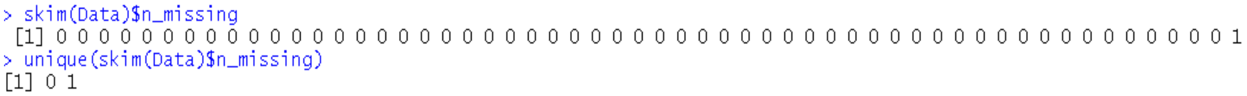
Figure 3.6: Updated column names of the ChemCollective workshop feedback dataset.



```
> colnames(data)
 [1] "Name"                                                "Institute"
 [3] "Background"                                          "Background (Other)"
 [5] "Duration of workshop attended previously (if any)"   "Used other software"
 [7] "Other software's name"                               "Used Jmol before"
 [9] "Jmol use/purpose"                                    "Jmol use/purpose (Other)"
[11] "Teaching/Learning without software"                  "Jmol useful in teaching"
[13] "Jmol difficulty"                                     "Understanding of Jmol before screening task"
[15] "Efforts put in screening task"                       "Understanding of Jmol after screening task"
[17] "(Spoken Tutorial) Surfaces and Orbitals"             "(Spoken Tutorial) Script Commands"
[19] "(Spoken Tutorial) Symmetry Point Groups"             "(Spoken Tutorial) Proteins Macromolecules"
[21] "(Spoken Tutorial) 3D Enzyme Models"                  "(Live Demonstration) Molecular Orbitals"
[23] "(Live Demonstration) Structures from databases"      "(Live Demonstration) Creating GIF"
[25] "(Live Demonstration) Conformations of disubstituted Ethane" "(Assignment) Molecular Orbitals"
[27] "(Assignment) Cyclohexane"                            "(Assignment) Point Groups"
[29] "(Assignment) Protein Structure"                      "(Assignment) Enzyme Structure"
[31] "Knowledge of using Jmol before workshop"             "Knowledge of using Jmol after workshop"
[33] "Quality of instructional material"                   "Self learning through Spoken Tutorial"
[35] "Live Demonstration"                                  "Discussions on assignments"
[37] "Advanced features demonstration"                     "Interaction with speakers"
[39] "Overall screening quality"                           "Exposure to new knowledge"
[41] "Unhappy with format"                                 "Willing to participate in future activites"
[43] "Didn't learn much"                                   "Recommend others to use Jmol"
[45] "Plan for Jmol use"                                   "Interested in becoming a Teaching Assistant"
[47] "Most liked aspect"                                   "Most disliked aspect"
[49] "Forum suggestions"                                   "Other suggestions"
>
```

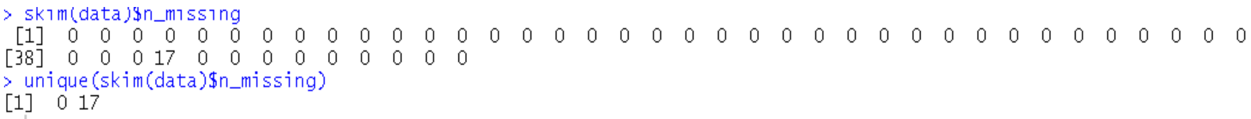
Figure 3.7: Updated column names of the Jmol workshop feedback dataset.

Data Exploration continued after updating the column names using the “skim()” function from the “skimr” package [2]. It was observed that there were missing values in the columns of both the datasets as shown in Figures 3.8 and 3.9.



```
> skim(Data)$n_missing
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
> unique(skim(Data)$n_missing)
[1] 0 1
```

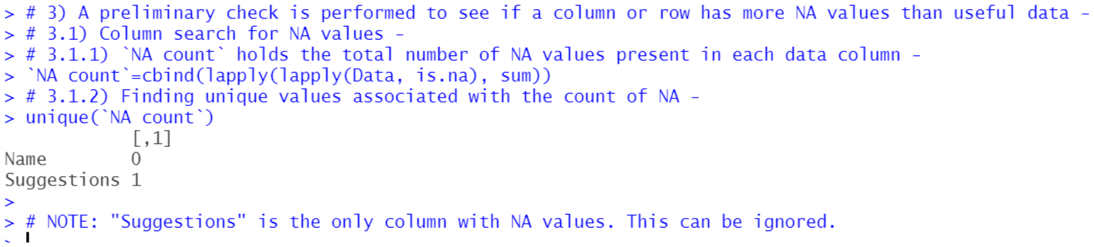
Figure 3.8: Missing values in the ChemCollective dataset



```
> skim(data)$n_missing
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[38]  0  0  0 17  0  0  0  0  0  0  0  0  0
> unique(skim(data)$n_missing)
[1]  0 17
```

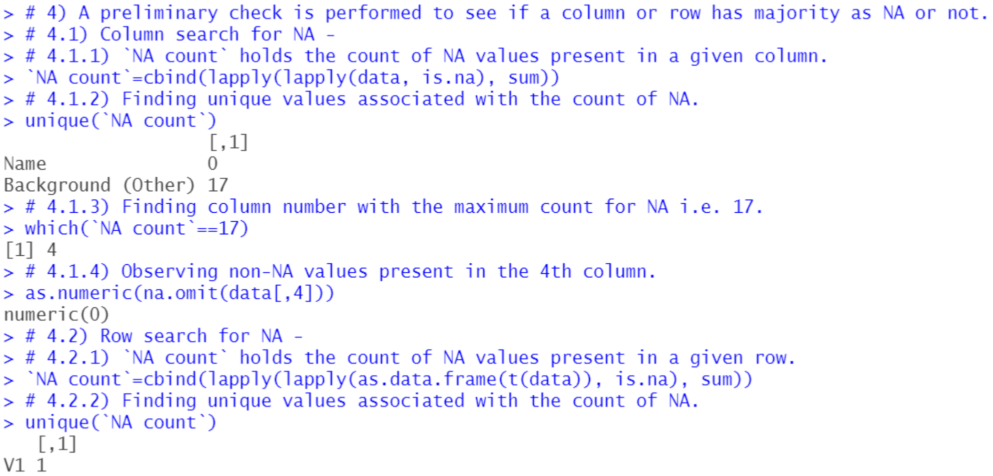
Figure 3.9: Missing values in the Jmol dataset

Missing values may occur due to a variety of reasons. One reason could be the participants’ unwillingness to respond to certain optional feedback sections like providing subjective comments, which is acceptable, but they may also occur due to some data formatting/reformatting issues. To move ahead with the exploration, it was necessary to remove the missing values in a way that we do not lose any vital information. Therefore, a row and column-wise check for missing values was carried out, as shown in Figure 3.10 and 3.11.



```
> # 3) A preliminary check is performed to see if a column or row has more NA values than useful data -
> # 3.1) Column search for NA values -
> # 3.1.1) `NA count` holds the total number of NA values present in each data column -
> `NA count`=cbind(lapply(lapply(Data, is.na), sum))
> # 3.1.2) Finding unique values associated with the count of NA -
> unique(`NA count`)
            [,1]
Name        0
Suggestions 1
>
> # NOTE: "Suggestions" is the only column with NA values. This can be ignored.
```

Figure 3.10: Row and column-wise examination of missing values for ChemCollective data.

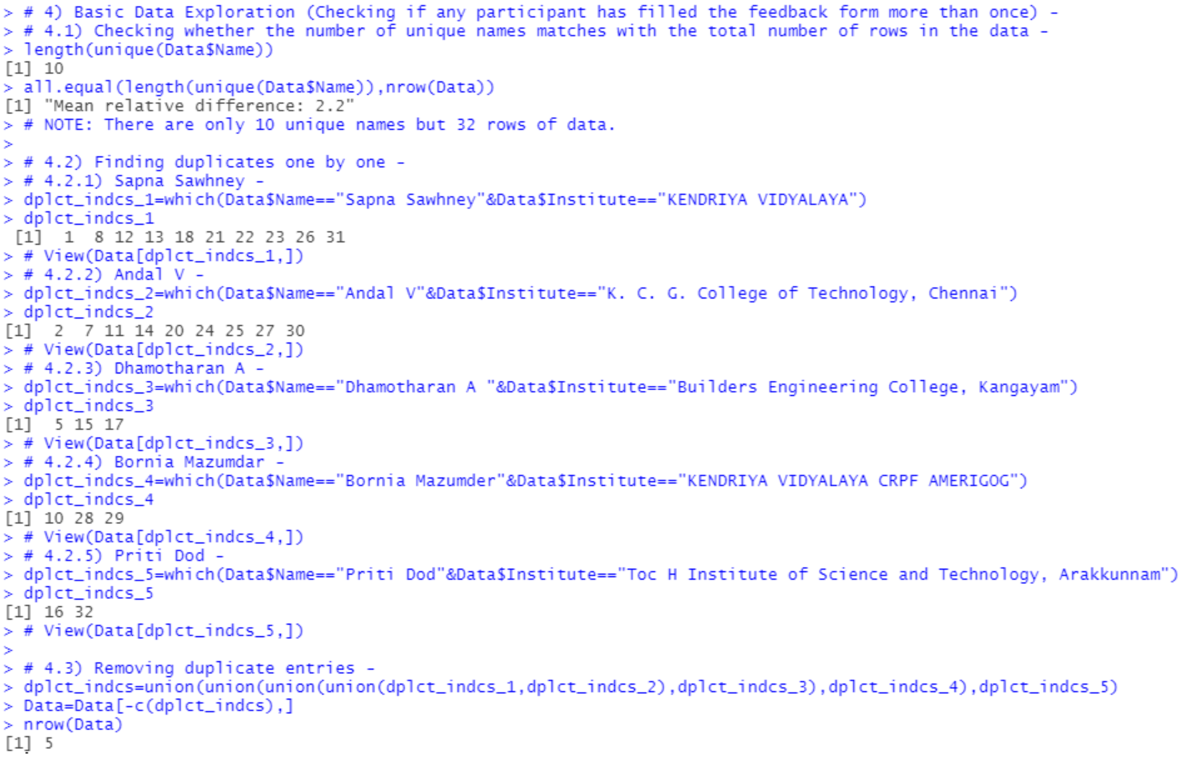


```
> # 4) A preliminary check is performed to see if a column or row has majority as NA or not.
> # 4.1) Column search for NA -
> # 4.1.1) `NA count` holds the count of NA values present in a given column.
> `NA count`=cbind(lapply(lapply(data, is.na), sum))
> # 4.1.2) Finding unique values associated with the count of NA.
> unique(`NA count`)
                   [,1]
Name               0
Background (Other) 17
> # 4.1.3) Finding column number with the maximum count for NA i.e. 17.
> which(`NA count`==17)
[1] 4
> # 4.1.4) Observing non-NA values present in the 4th column.
> as.numeric(na.omit(data[,4]))
numeric(0)
> # 4.2) Row search for NA -
> # 4.2.1) `NA count` holds the count of NA values present in a given row.
> `NA count`=cbind(lapply(lapply(as.data.frame(t(data)), is.na), sum))
> # 4.2.2) Finding unique values associated with the count of NA.
> unique(`NA count`)
   [,1]
V1 1
```

Figure 3.11: Row and column-wise examination of missing values for Jmol data.

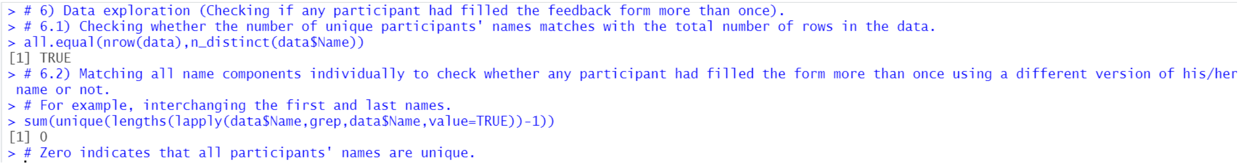
For ChemCollective data, only the “Suggestions” column contained missing values; hence it was ignored because the feedback form question associated with that column was optional. For Jmol data, the row-wise examination showed that at max only one value is missing for a particular row and the column-wise examination showed that only the fourth column contained missing values. Hence the fourth column was removed from the Jmol data.

After dealing with missing values, both datasets were checked for duplicate entries because such entries can introduce bias in statistical analyses [3,4]. The Approximate String Matching (Fuzzy Matching) technique was applied over the “Name” column of both the datasets using the “agrep()” function by Brian Ripley and Kurt Hornik provided in the “base” package of R [5] 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.



```
> # 4) Basic Data Exploration (Checking if any participant has filled the feedback form more than once) -
> # 4.1) Checking whether the number of unique names matches with the total number of rows in the data -
> length(unique(Data$Name))
[1] 10
> all.equal(length(unique(Data$Name)),nrow(Data))
[1] "Mean relative difference: 2.2"
> # NOTE: There are only 10 unique names but 32 rows of data.
>
> # 4.2) Finding duplicates one by one -
> # 4.2.1) Sapna Sawhney -
> dplct_indcs_1=which(Data$Name=="Sapna Sawhney"&Data$Institute=="KENDRIYA VIDYALAYA")
> dplct_indcs_1
 [1]  1  8 12 13 18 21 22 23 26 31
> # View(Data[dplct_indcs_1,])
> # 4.2.2) Andal V -
> dplct_indcs_2=which(Data$Name=="Andal V"&Data$Institute=="K. C. G. College of Technology, Chennai")
> dplct_indcs_2
[1]  2  7 11 14 20 24 25 27 30
> # View(Data[dplct_indcs_2,])
> # 4.2.3) Dhamotharan A -
> dplct_indcs_3=which(Data$Name=="Dhamotharan A "&Data$Institute=="Builders Engineering College, Kangayam")
> dplct_indcs_3
[1]  5 15 17
> # View(Data[dplct_indcs_3,])
> # 4.2.4) Bornia Mazumdar -
> dplct_indcs_4=which(Data$Name=="Bornia Mazumder"&Data$Institute=="KENDRIYA VIDYALAYA CRPF AMERIGOG")
> dplct_indcs_4
[1] 10 28 29
> # View(Data[dplct_indcs_4,])
> # 4.2.5) Priti Dod -
> dplct_indcs_5=which(Data$Name=="Priti Dod"&Data$Institute=="Toc H Institute of Science and Technology, Arakkunnam")
> dplct_indcs_5
[1] 16 32
> # View(Data[dplct_indcs_5,])
>
> # 4.3) Removing duplicate entries -
> dplct_indcs=union(union(union(union(dplct_indcs_1,dplct_indcs_2),dplct_indcs_3),dplct_indcs_4),dplct_indcs_5)
> Data=Data[-c(dplct_indcs),]
> nrow(Data)
[1] 5
```

Figure 3.12: Checking for duplicate entries in the ChemCollective dataset.



```
> # 6) Data exploration (Checking if any participant had filled the feedback form more than once).
> # 6.1) Checking whether the number of unique participants' names matches with the total number of rows in the data.
> all.equal(nrow(data),n_distinct(data$Name))
[1] TRUE
> # 6.2) Matching all name components individually to check whether any participant had filled the form more than once using a different version of his/her
 name or not.
> # For example, interchanging the first and last names.
> sum(unique(lengths(lapply(data$Name,grep,data$Name,value=TRUE))-1))
[1] 0
> # Zero indicates that all participants' names are unique.
```

Figure 3.13: Checking for duplicate entries in the Jmol dataset.

Figure 3.12 and 3.13 show the result of checking duplicate entries for the ChemCollective and Jmol datasets, respectively. In the ChemCollective dataset, only five unique row entries were found, hence it was deemed unfit for further analysis, whereas in the Jmol dataset, no duplicate entries were found.

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

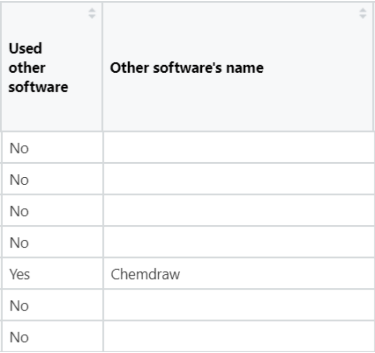
4. Data Cleaning

Data cleaning is the most critical step performed before analyzing the data, as any result obtained from incorrect data will be unreliable. It involves the steps for removing erroneous and mislabelled data [6,7]. 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, lack of understanding of the questions asked, etc. Hence, there is a need to carefully examine the data and remove all misleading responses to preserve its reliability.

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

1. **Checked participants' backgrounds and opinions regarding JMOL and similar tools -**

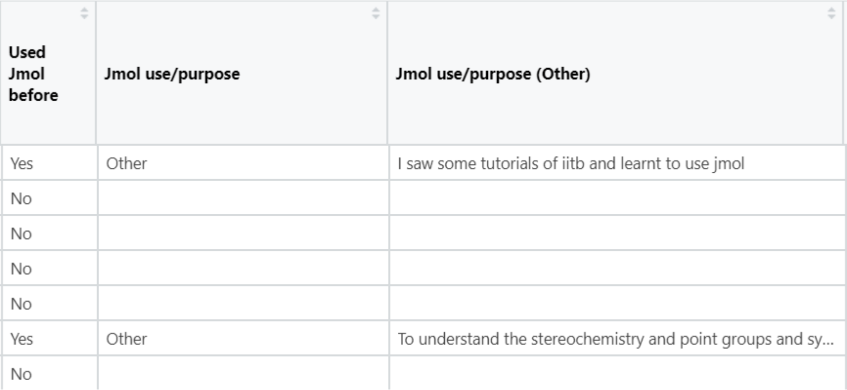
- Checked for entries where participants had given a negative response when asked if they had used any software other than Jmol but entered the name of a software in the subsequent section.
- Checked for entries where participants had given a positive response when asked if they had used any software other than Jmol but did not mention the name of the software.



Used other software	Other software's name
No	
No	
No	
No	
Yes	Chemdraw
No	
No	

Figure 3.14: Columns containing entries related to participants' experience with modeling software.

- Checked for entries where participants had given a positive response when asked if they had used Jmol before but failed to mention the purpose of use.
- Checked for entries where participants had given a negative response when asked if they had used Jmol before but mentioned the purpose of use.

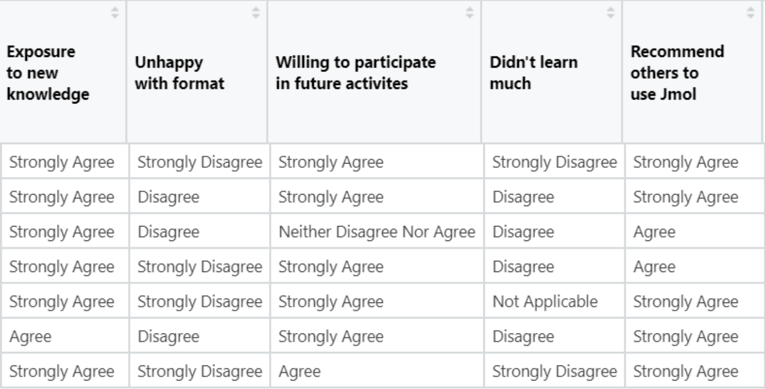


Used Jmol before	Jmol use/purpose	Jmol use/purpose (Other)
Yes	Other	I saw some tutorials of iitb and learnt to use jmol
No		
No		
No		
No		
Yes	Other	To understand the stereochemistry and point groups and sy...
No		

Figure 3.15: Columns indicating participants' experience with the Jmol software.

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

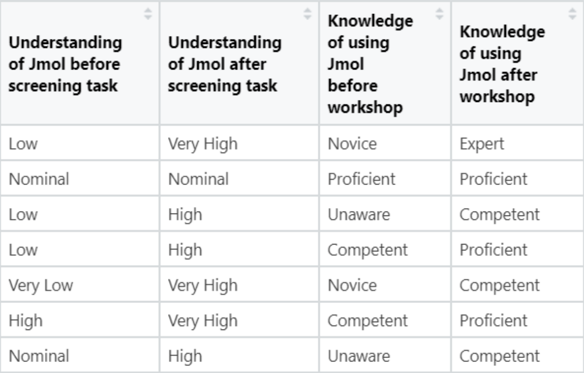
- Checked for contradicting responses in columns associated with the quality and effectiveness of the workshop; for example, searched and recorded all such entries where a participant had selected the option "Strongly Agree" for both "Exposure To New Knowledge" and "Did Not Learn Much" columns.



Exposure to new knowledge	Unhappy with format	Willing to participate in future activites	Didn't learn much	Recommend others to use Jmol
Strongly Agree	Strongly Disagree	Strongly Agree	Strongly Disagree	Strongly Agree
Strongly Agree	Disagree	Strongly Agree	Disagree	Strongly Agree
Strongly Agree	Disagree	Neither Disagree Nor Agree	Disagree	Agree
Strongly Agree	Strongly Disagree	Strongly Agree	Disagree	Agree
Strongly Agree	Strongly Disagree	Strongly Agree	Not Applicable	Strongly Agree
Agree	Disagree	Strongly Agree	Disagree	Strongly Agree
Strongly Agree	Strongly Disagree	Agree	Strongly Disagree	Strongly Agree

Figure 3.16: Column entries associated with the workshop's quality and effectiveness.

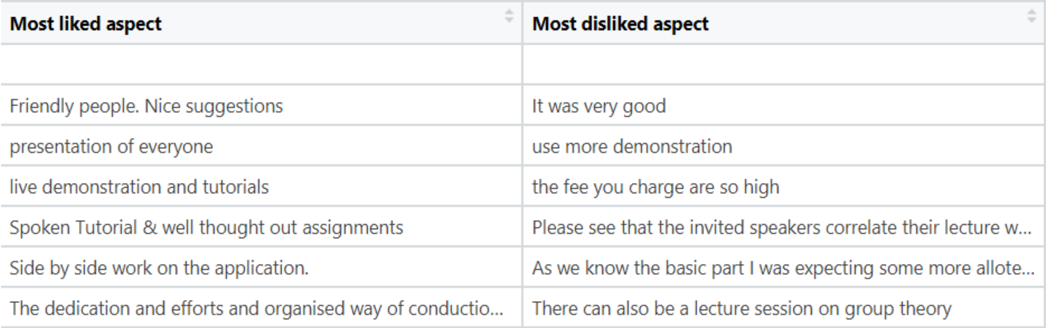
- Checked if the level of knowledge regarding Jmol for any participant dropped after the workshop, as it is improbable.



Understanding of Jmol before screening task	Understanding of Jmol after screening task	Knowledge of using Jmol before workshop	Knowledge of using Jmol after workshop
Low	Very High	Novice	Expert
Nominal	Nominal	Proficient	Proficient
Low	High	Unaware	Competent
Low	High	Competent	Proficient
Very Low	Very High	Novice	Competent
High	Very High	Competent	Proficient
Nominal	High	Unaware	Competent

Figure 3.17: Entries associated with the conceptual knowledge regarding various aspects of Jmol.

- Checked for positive feedback in negative questions and vice versa.



Most liked aspect	Most disliked aspect
Friendly people. Nice suggestions	It was very good
presentation of everyone	use more demonstration
live demonstration and tutorials	the fee you charge are so high
Spoken Tutorial & well thought out assignments	Please see that the invited speakers correlate their lecture w...
Side by side work on the application.	As we know the basic part I was expecting some more allote...
The dedication and efforts and organised way of conductio...	There can also be a lecture session on group theory

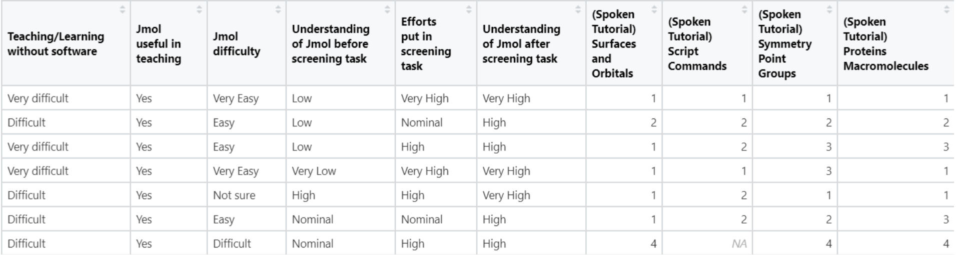
Figure 3.18: Entries associated with the overall workshop's feedback questions.

3. **Removed misleading entries -**

- After performing all possible checks, only a single misleading entry was found and removed. The final dataset contained 16 rows and 49 columns.

5. Data Preprocessing

Only the Likert scale based columns containing categorical responses were kept from the cleaned Jmol workshop feedback dataset for EFA, as shown in Figure 3.19 [8,9]. Any information related to the participants' backgrounds and their subjective comments was removed from the dataset. All entries containing the string "Not Attempted" were replaced with NA. The remaining dataset had 16 rows and 34 columns, where each column had the factor data type.

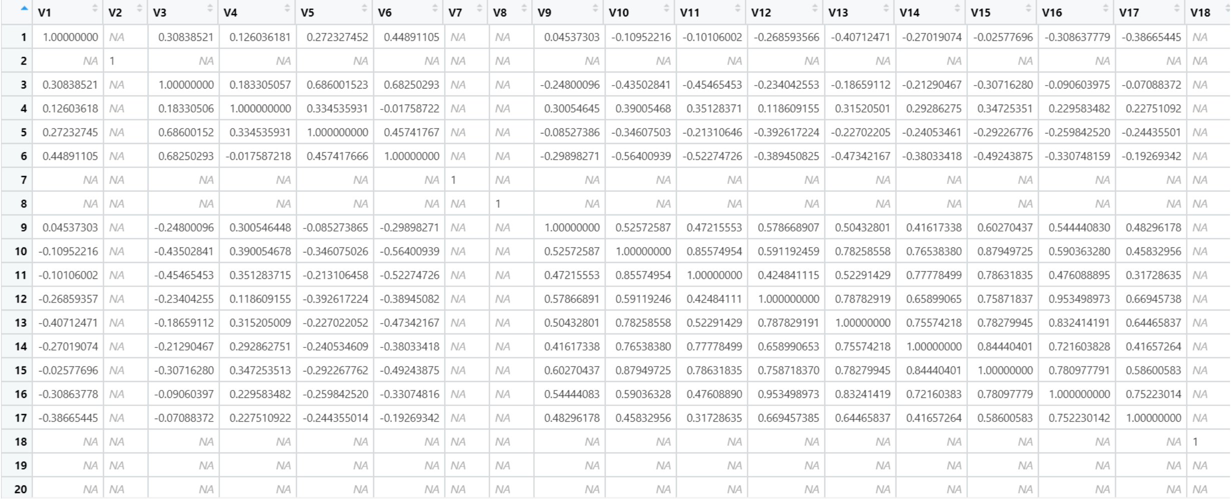


Teaching/Learning without software	Jmol useful in teaching	Jmol difficulty	Understanding of Jmol before screening task	Efforts put in screening task	Understanding of Jmol after screening task	(Spoken Tutorial) Surfaces and Orbitals	(Spoken Tutorial) Script Commands	(Spoken Tutorial) Symmetry Point Groups	(Spoken Tutorial) Proteins Macromolecules
Very difficult	Yes	Very Easy	Low	Very High	Very High	1	1	1	1
Difficult	Yes	Easy	Low	Nominal	High	2	2	2	2
Very difficult	Yes	Easy	Low	High	High	1	2	3	3
Very difficult	Yes	Very Easy	Very Low	Very High	Very High	1	1	3	1
Difficult	Yes	Not sure	High	High	Very High	1	2	1	1
Difficult	Yes	Easy	Nominal	Nominal	High	1	2	2	3
Difficult	Yes	Difficult	Nominal	High	High	4	NA	4	4

Figure 3.19: Glimpse of the dataset after pre-processing.

6. Data Analysis

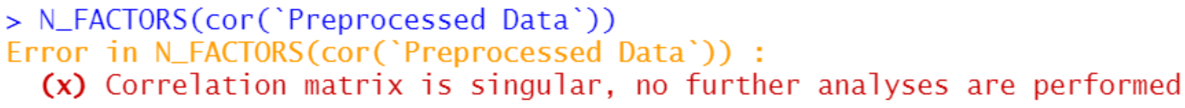
The data analysis was performed using the “EFAtools” package [10]. The “N_FACTORS()” function from “EFAtools” was used to find the suitable number of factors in the data by first converting the data into a numeric format and then finding its correlation matrix. Due to the columns “Jmol useful in teaching”, “(Spoken Tutorial) Surfaces and Orbitals”, “(Spoken Tutorial) Script Commands”, “(Assignment) Point Groups”, “(Assignment) Protein Structure” and “(Assignment) Enzyme Structure”, the obtained correlation matrix contained NA values, as shown in Figure 3.20.



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
1	1.00000000	NA	0.30838521	0.126036181	0.272327452	0.44891105	NA	NA	0.04537303	-0.10952216	-0.10106002	-0.268593566	-0.40712471	-0.27019074	-0.02577696	-0.308637779	-0.38665445	NA
2	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	0.30838521	NA	1.00000000	0.183305057	0.686001523	0.68250293	NA	NA	-0.24800096	-0.43502841	-0.45465453	-0.234042553	-0.18659112	-0.21290467	-0.30716280	-0.090603975	-0.07088372	NA
4	0.12603618	NA	0.18330506	1.000000000	0.334535931	-0.01758722	NA	NA	0.30054645	0.39005468	0.35128371	0.118609155	0.31520501	0.29286275	0.34725351	0.229583482	0.22751092	NA
5	0.27232745	NA	0.68600152	0.334535931	1.000000000	0.45741767	NA	NA	-0.08527386	-0.34607503	-0.21310646	-0.392617224	-0.22702205	-0.24053461	-0.29226776	-0.259842520	-0.24435501	NA
6	0.44891105	NA	0.68250293	-0.017587218	0.457417666	1.00000000	NA	NA	-0.29898271	-0.56400939	-0.52274726	-0.389450825	-0.47342167	-0.38033418	-0.49243875	-0.330748159	-0.19269342	NA
7	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	0.04537303	NA	-0.24800096	0.300546448	-0.085273865	-0.29898271	NA	NA	1.00000000	0.52572587	0.47215553	0.578668907	0.50432801	0.41617338	0.60270437	0.544440830	0.48296178	NA
10	-0.10952216	NA	-0.43502841	0.390054678	-0.346075026	-0.56400939	NA	NA	0.52572587	1.00000000	0.85574954	0.591192459	0.78258558	0.76538380	0.87949725	0.590363280	0.45832956	NA
11	-0.10106002	NA	-0.45465453	0.351283715	-0.213106458	-0.52274726	NA	NA	0.47215553	0.85574954	1.00000000	0.424841115	0.52291429	0.77778499	0.78631835	0.476088895	0.31728635	NA
12	-0.26859357	NA	-0.23404255	0.118609155	-0.392617224	-0.38945082	NA	NA	0.57866891	0.59119246	0.42484111	1.000000000	0.78782919	0.65899065	0.75871837	0.953498973	0.66945738	NA
13	-0.40712471	NA	-0.18659112	0.315205009	-0.227022052	-0.47342167	NA	NA	0.50432801	0.78258558	0.52291429	0.787829191	1.00000000	0.75574218	0.78279945	0.832414191	0.64465837	NA
14	-0.27019074	NA	-0.21290467	0.292862751	-0.240534609	-0.38033418	NA	NA	0.41617338	0.76538380	0.77778499	0.658990653	0.75574218	1.00000000	0.84440401	0.721603828	0.41657264	NA
15	-0.02577696	NA	-0.30716280	0.347253513	-0.292267762	-0.49243875	NA	NA	0.60270437	0.87949725	0.78631835	0.758718370	0.78279945	0.84440401	1.00000000	0.780977791	0.58600583	NA
16	-0.30863778	NA	-0.09060397	0.229583482	-0.259842520	-0.33074816	NA	NA	0.54444083	0.59036328	0.47608890	0.953498973	0.83241419	0.72160383	0.78097779	1.000000000	0.75223014	NA
17	-0.38665445	NA	-0.07088372	0.227510922	-0.244355014	-0.19269342	NA	NA	0.48296178	0.45832956	0.31728635	0.669457385	0.64465837	0.41657264	0.58600583	0.752230142	1.00000000	NA
18	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figure 3.20: Correlation matrix of the pre-processed dataset.

Therefore, those columns were 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 [11]. The KMO value indicates the proportion of variance in the variables that might be caused by underlying factors [12]. 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 3.21.



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

Figure 3.21: Result obtained from the “N_FACTORS()” function.

7. Conclusion

In this project, data exploration, cleaning and preprocessing, were performed over the ChemCollective Virtual Lab Beginner 1 day workshop and the Jmol Application Advanced Workshop feedback datasets completely using the R programming language with the objective of applying EFA over them. However, the datasets turned out to be unsuitable for the proposed analysis as the ChemCollective dataset was too small and the Jmol dataset 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 Jmol feedback data.

Chapter 4

Spoken Tutorial content creation

1. Introduction

The Spoken Tutorial project aims to make video tutorials on Free and Open Source Software (FOSS) available in several Indian languages. The goal is to enable the use of spoken tutorials to teach in any Indian language to learners of various levels of expertise - Beginner, Intermediate or Advanced. Every tutorial has to go through a series of checks to ensure that it is perfect for its audience, which is crucial for achieving the goal of this project. I was given the opportunity to contribute to the creation of eleven scripts and corresponding slides associated with the Advanced R spoken tutorial series. The tutorial topics were related to machine learning and are listed below.

2. List of tutorial topics -

2.1 Supervised Learning

Supervised Learning is a branch of Machine Learning where the goal is to learn a mapping from inputs “x” to outputs “y”, when given a labeled set of input-output pairs, $D = \{(xi,\ yi)\}_{i=1}^{N}$. Here, “D” is called the training set, and “N” is the number of training examples [13]. The tutorial explains how supervised learning works by applying a Naive Bayes Classifier on the Iris dataset. It also introduces the concept of a confusion matrix to calculate the accuracy of the resulting model. Two packages were used in the tutorial, namely “e1071” [14] and “caret” [15].

2.2 Unsupervised Learning

Unsupervised Learning is a branch of Machine Learning where we are only given inputs “x” in a set, $D = \{xi\}_{i=1}^{N}$, and the goal is to find “interesting patterns” in the data. Here, “D” is called the training set, and “N” is the number of training examples. It is sometimes also known as knowledge discovery [13]. The tutorial explains how unsupervised learning works by performing K-means Clustering on the Iris dataset. Then it introduces the Adjusted RAND index; a metric used to measure the accuracy of the obtained clustering. Two packages were used in the tutorial, namely “ggplot2” [16] and “mclust” [17].

2.3 Data Cleaning

Data Cleaning is the process of detecting, diagnosing, and editing erroneous data [3]. It is one of the essential steps performed before analyzing the data [4]. The tutorial aims to demonstrate the steps performed while cleaning a dataset. The dataset used was the text version of the AirQuality dataset. In the tutorial, the following operations were performed -

- Conversion of a text file to CSV file.
- Removal of NA values.

- Encoding of categorical variables into factors.

2.4 Linear Discriminant Analysis

Linear Discriminant Analysis or LDA is a linear combination of features that separates two or more classes of objects or events. It assumes multivariate normality and multicollinearity in the data [18]. In the tutorial, LDA was applied over the Iris dataset and its performance was measured using a confusion matrix. Four packages were used in the tutorial, namely “MASS” [19], “e1071” [14], “caret” [15] and “ggplot2” [16].

2.5 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis or QDA is a quadratic combination of features that separates two or more classes of objects or events. Compared to LDA, QDA assumes that the covariance structures of the classes of objects are different [20]. In the tutorial, QDA was implemented over the Iris dataset. A single package was used in the tutorial, i.e., “MASS” [19].

2.6 Support Vector Machine

Support Vector Machine or SVM is a supervised machine technique. It constructs a hyperplane to separate n-dimensional data into different classes. It is used for classification, regression and outlier detection [20]. In the tutorial, SVM was implemented over the Iris dataset and its accuracy was computed using a confusion matrix. Two packages were used in the tutorial, namely “e1071” [14] and “caret” [15].

2.7 Logistic Regression

Logistic regression is an important machine learning algorithm. The goal is to model the probability of a random variable “Y” being 0 or 1 given experimental data [21]. In the tutorial, logistic regression was implemented over the Iris dataset and its accuracy was computed using a confusion matrix. Three packages were used in the tutorial, namely “stats4” [5], “splines” [5] and “VGAM” [22].

2.8 Decision Tree

In data mining, a decision tree is a predictive model that can represent both classifiers and regression models. When a decision tree is used for classification tasks, it is more appropriately referred to as a classification tree. When it is used for regression tasks, it is called a regression tree [23]. In the tutorial, a decision tree was constructed using the Iris dataset. Two packages were used in the tutorial, namely “rpart” [24] and “rpart.plot” [25].

2.9 Random Forest

The general principle of a random forest is to aggregate a collection of random decision trees. The goal is, instead of seeking to optimize a predictor “at once” as for a CART tree, to pool a set of predictors (not necessarily optimal) [26]. In the tutorial, a random forest was created using the Iris dataset and its performance was measured using a confusion matrix. The package used in the tutorial was “randomforest” [27].

2.10 K-means Clustering

The k-means method is a widely used clustering technique that minimizes the average squared distance between points in the same cluster [28]. In the tutorial, an optimized version of k-means called kmeans++ was implemented on the Iris dataset and the accuracy of the obtained results was measured using a confusion matrix. The package used in the tutorial was “LICORS” [29].

2.11 Hierarchical Clustering

Hierarchical clustering is a clustering approach that does not require the user to choose the number of clusters beforehand. It has an added advantage over K-means clustering in that it results in an attractive tree-based representation of the observations, called a dendrogram. The most common type of hierarchical clustering is bottom-up or agglomerative clustering [20]. In the tutorial, agglomerative clustering was implemented over the Iris dataset. The package used in the tutorial was “ggplot” [16].

Chapter 5

Implementation and optimization of SOM algorithm in R

1. Introduction

SOM is an unsupervised data visualization technique popular among researchers for dimensionality reduction and clustering [30]. 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 [31-35]. 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 [36,37].

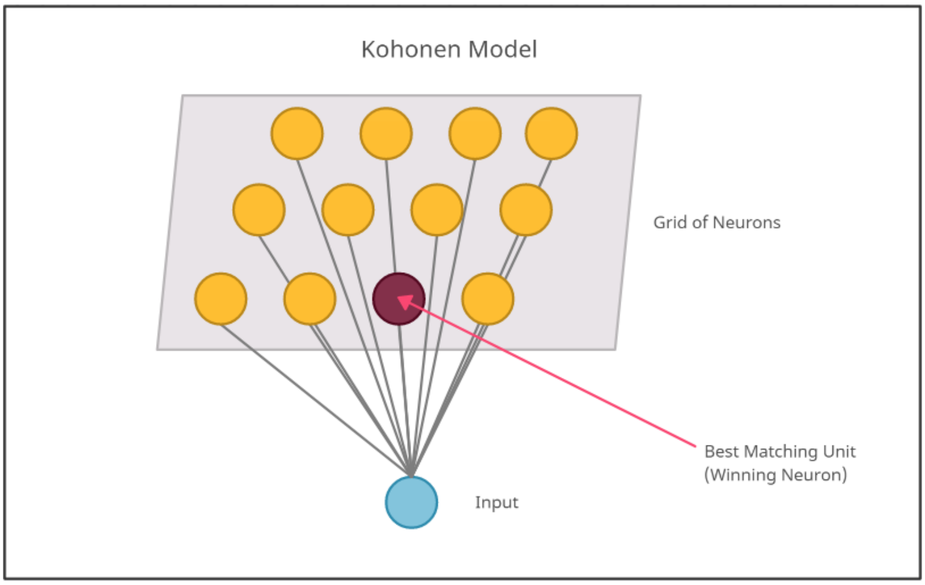


Figure 4.1: Kohonen Model of Self Organizing Map [35].

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 intern 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.

4. Optimization of the SOM algorithm

The base algorithm took over 40 seconds to run on the UCLA Graduate School Admissions dataset [38] that contained only 400 rows. Thus, efforts were made to optimize the algorithm while retaining its fundamental logic. Mainly the following two operations were performed to optimize the code -

1. Eliminating as many for loops as possible.
2. Converting every potential operation to a matrix operation.

With the help of the R profiler, which was implemented using the "profvis()" function of "profvis" package [39], it was determined that the BMU and SOM functions had the highest time complexity. Finally, the following changes were made -

BMU: The entire function was reduced to just three lines of code. In the 1st line, the "sweep()" function was used to find the euclidean distance between every neuron in the grid and the given data point. The 2nd line contained the "which.min()" function to find the winning neuron in the grid. The final line of code returned the index of the winning neuron.

SOM: A similar approach was applied to the SOM function. The lateral distances from the winning neuron were computed using the "sweep()" function. The weights were updated by first finding the required indices using the "which()" function and then applying matrix operation instead of a for loop.

Finally, the entire algorithm sped up by approximately ten times.

Chapter 6

R Case Study: Analysis and prediction of the impact of COVID-19 on the global economy

1. Introduction

COVID-19 has had a profound impact on the lives of each individual. However, for countries, the economy has taken the hardest hit. To analyze the effects of the COVID-19 pandemic on the Gross Domestic Product (GDP) and employment in countries worldwide, I proposed a case study project 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 Collection

The COVID-19 Economic Impact Assessment data was collected for this case study from an online repository known as the ADB Data Library. The ADB Data Library is a platform that hosts publicly available data from the Asian Development Bank. The data obtained contains a measure of the potential economy and sector specific impact of the COVID-19 outbreak [40].

3. Data Exploration

The original dataset had the dimension 1566 x 10 and contained the following columns -

- **Economy -** Contains the country name.
- **ADB Country Code -** Contains the country code as assigned by ADB.
- **Sector -** Contains the economic sector from where the data was collected.
- **Country 2018 GDP -** Contains a country's GDP for the year 2018.
- **Scenario -** Contains the scenario based on which the GDP drop is predicted.
- **as \% of total GDP -** Contains the GDP loss as a percentage of the total GDP.
- **in \$ Mn -** Contains the total income in denominations of \$1 million.
- **Employment (in 000) -** Contains total number of people employed in counts of 1000s.
- **as \% of sector GDP -** Contains percentage of sector GDP loss.
- **as \% of sector employment -** Contains percentage of sector employment loss.

4. Data Cleaning and Preprocessing

Data cleaning and preprocessing involved various steps that were performed to make data adequate for analysis. Some of the steps involved were data reformatting and searching for missing values that were later removed using the function “na.omit()”. Data reformatting involved changing the data type of columns depending on the analysis to be performed over the data. After data cleaning and reformatting, the remaining data was split in a ratio of 3:1 for training and testing, respectively. Later the training and testing datasets were used for statistical modeling.

5. Data Analysis

The data analysis step involved the application of linear regression and artificial neural network over the data -

5.1 Linear Regression

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables, respectively). The case of a single explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression [41]. For the case study, multiple linear regression was utilized. The regression model was trained to predict GDP loss, given the sector, scenario and sector-wise GDP loss. Depending upon the obtained results, the model was further enhanced by removing insignificant data columns.

5.2 Artificial Neural Network

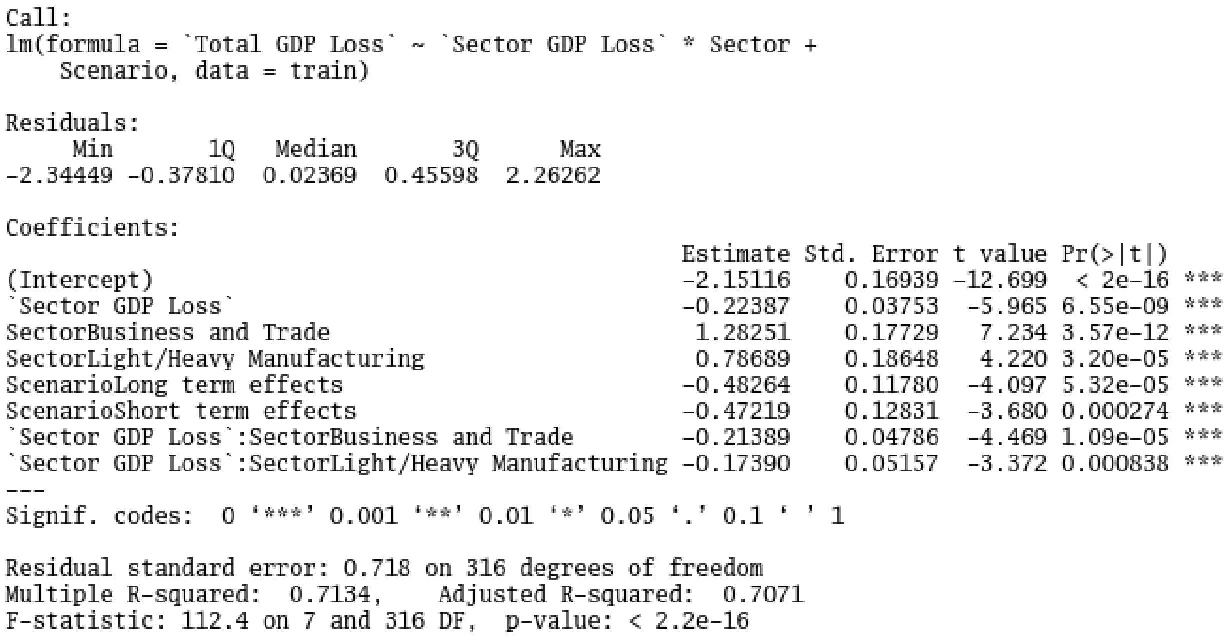
An artificial neural network or ANN takes an input vector of “p” variables “X = (X1,X2,...,Xp)” and builds a non-linear function “f(X)” to predict the response “Y” [42]. The “nnet” package of R [43] was used to create the ANN model. The ANN model was trained over the data left after removing all insignificant variables (columns). The number of hidden units was set to six after numerous experiments. To ensure reproducibility of the results, the command to train an ANN was made to run 1000 times with the iteration number set as the random seed. Finally, the best ANN model with RMSE lower than the linear regression model was selected and saved in an RDS file for later use.

6. Results

6.1 Linear Regression

6.1.1 Model Summary

Following is the summary of the final linear regression model obtained by making use of the “summary()” function.



```
Call:
lm(formula = `Total GDP Loss` ~ `Sector GDP Loss` * Sector +
    Scenario, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.34449 -0.37810  0.02369  0.45598  2.26262

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -2.15116    0.16939 -12.699  < 2e-16 ***
`Sector GDP Loss`                                 -0.22387    0.03753  -5.965 6.55e-09 ***
SectorBusiness and Trade                           1.28251    0.17729   7.234 3.57e-12 ***
SectorLight/Heavy Manufacturing                    0.78689    0.18648   4.220 3.20e-05 ***
ScenarioLong term effects                         -0.48264    0.11780  -4.097 5.32e-05 ***
ScenarioShort term effects                        -0.47219    0.12831  -3.680 0.000274 ***
`Sector GDP Loss`:SectorBusiness and Trade        -0.21389    0.04786  -4.469 1.09e-05 ***
`Sector GDP Loss`:SectorLight/Heavy Manufacturing -0.17390    0.05157  -3.372 0.000838 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.718 on 316 degrees of freedom
Multiple R-squared:  0.7134,    Adjusted R-squared:  0.7071
F-statistic: 112.4 on 7 and 316 DF,  p-value: < 2.2e-16
```

Figure 6.1: Summary of the final linear regression model.

6.1.2 Q-Q Plot

Below is the q-q plot of the final linear regression model's residuals.

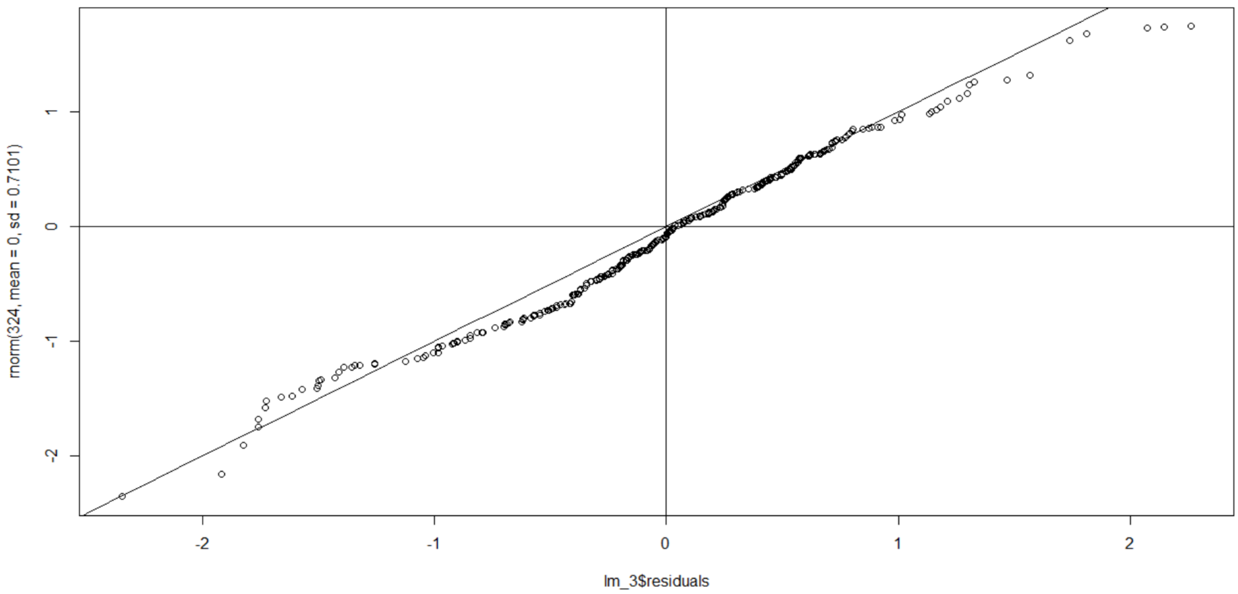


Figure 6.2: q-q plot of the final linear regression model.

6.1.3 Squared Residuals Plot

Following is the squared residuals plot associated with the regression model.

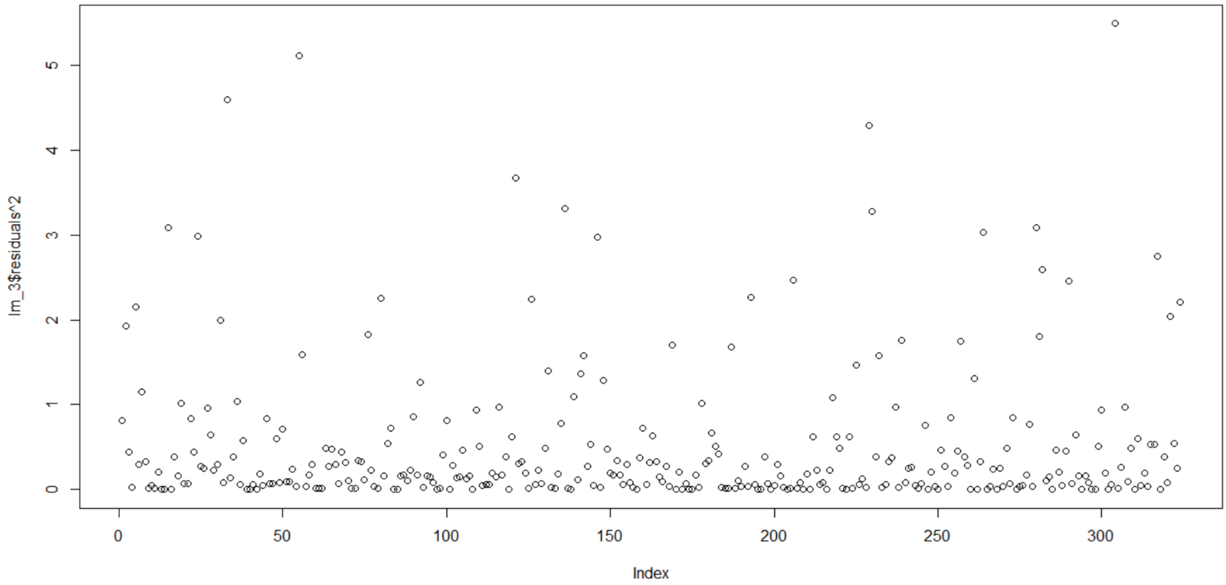
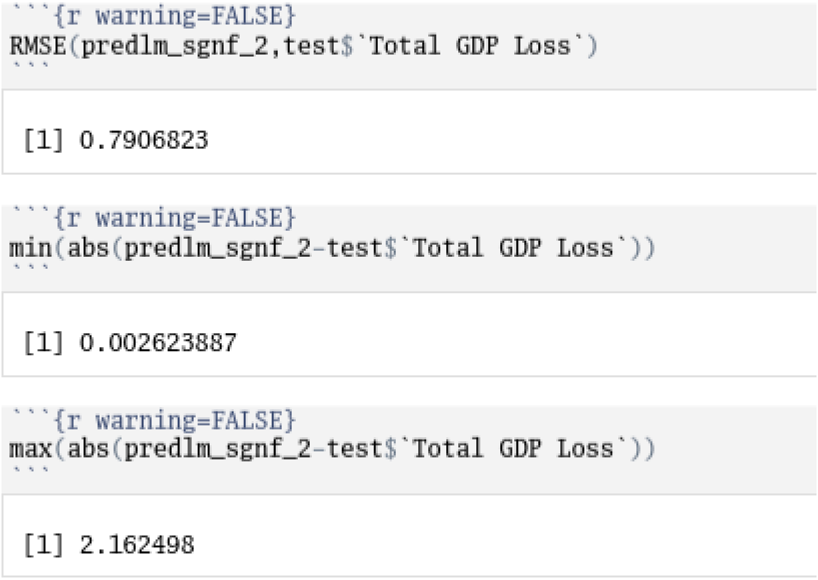


Figure 6.3: Squared residual plot of the final linear regression model.

6.1.4 Accuracy Measurement Results

Following are the accuracy measurement results associated with the regression model.



~~~
```{r warning=FALSE}
RMSE(predlm_sgnf_2,test$`Total GDP Loss`)
```
~~~

```
[1] 0.7906823
```

~~~
```{r warning=FALSE}
min(abs(predlm_sgnf_2-test$`Total GDP Loss`))
```
~~~

```
[1] 0.002623887
```

~~~
```{r warning=FALSE}
max(abs(predlm_sgnf_2-test$`Total GDP Loss`))
```
~~~

```
[1] 2.162498
```

Figure 6.4: Accuracy measurement results of the final linear regression model.

6.1.5 Predicted v/s Original

Below is a plot comparing the predicted with original values.

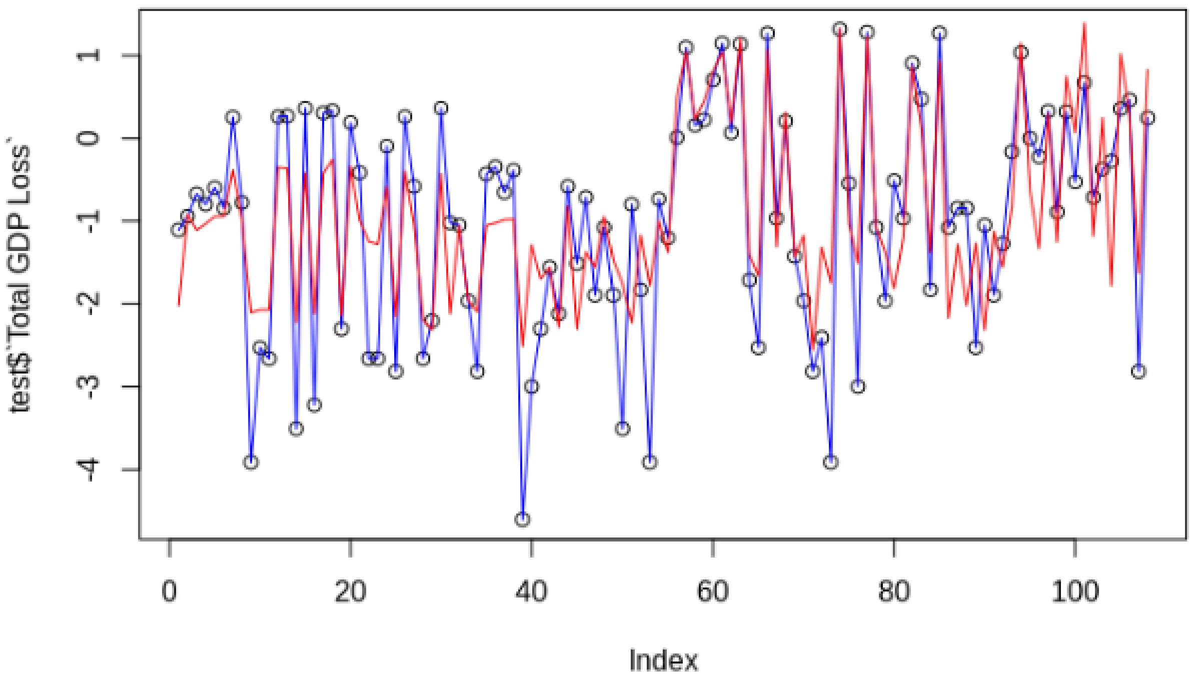


Figure 6.5: Final regression model's predicted values versus original values.

6.2 Artificial Neural Network

6.2.1 Squared Residuals Plot

Following is the squared residuals plot associated with the ANN model.

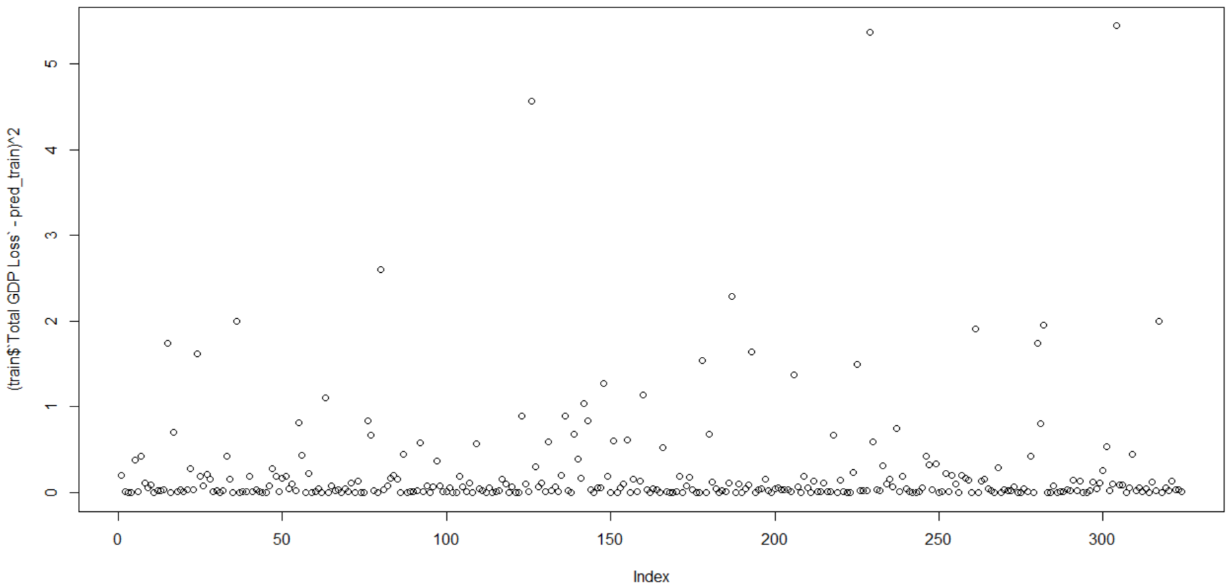
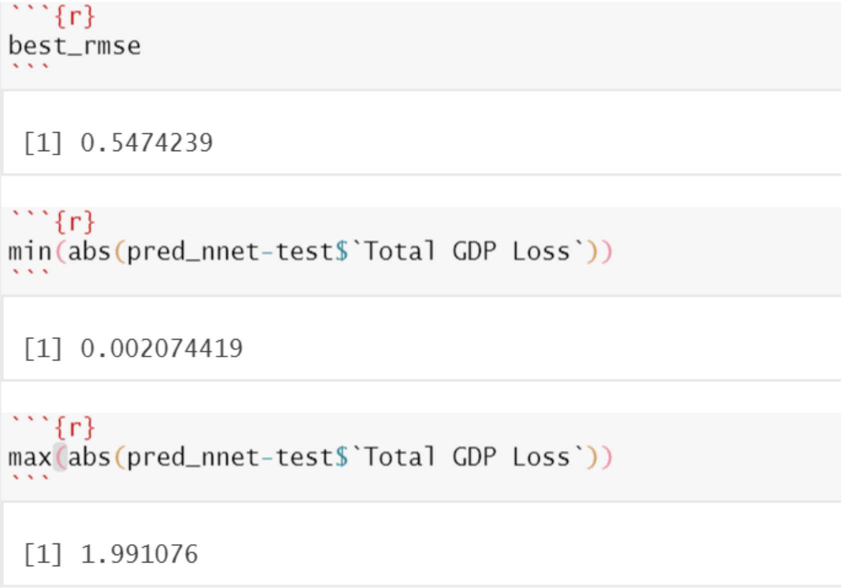


Figure 6.6: Squared residuals plot associated with the neural network model.

6.2.2 Accuracy Measurement Results

Following are the accuracy measurement results associated with the ANN model.



~~~
```{r}
best_rmse
~~~

 [1] 0.5474239

```{r}
min(abs(pred_nnet-test$`Total GDP Loss`))
```

 [1] 0.002074419

```{r}
max(abs(pred_nnet-test$`Total GDP Loss`))
```

 [1] 1.991076
```

Figure 6.7: ANN model's accuracy measurement results.

6.2.3 Predicted v/s Original

Below is a plot comparing the predicted with original values.

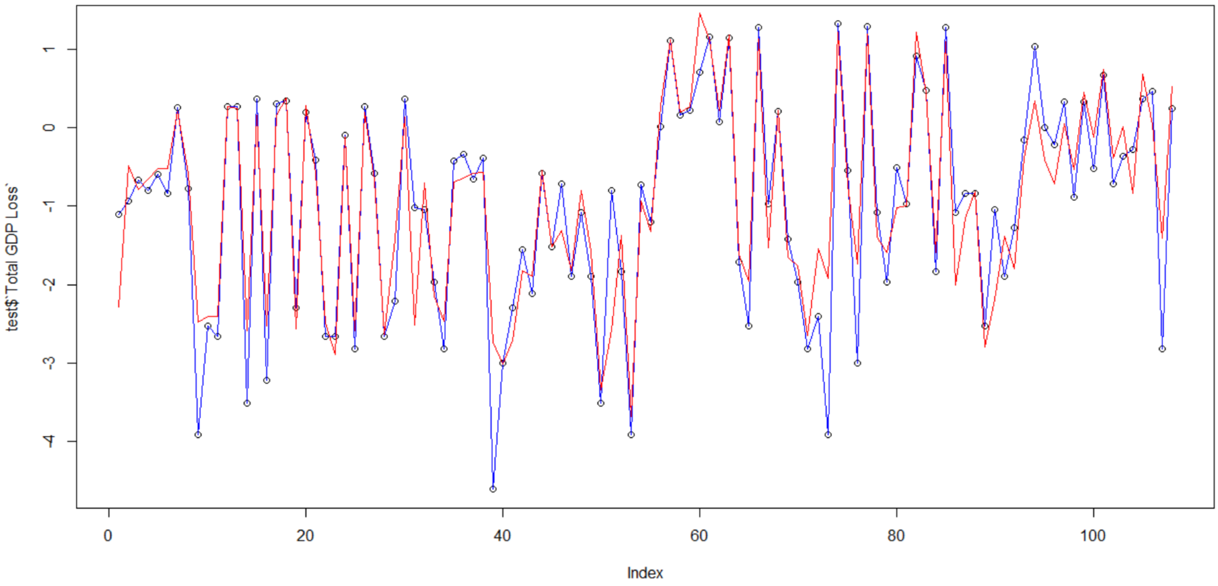


Figure 6.8: ANN model's predicted values versus original values.

7. Conclusion

The case study attempted to explore the impact of COVID-19 on the GDP of various countries. The obtained statistical models have accurately predicted the GDP loss. The final linear regression model predicted the actual values with a minor error. It seems like a good choice for predicting the GDP loss as it has the advantage of being a white-box approach. On the other hand, a neural network may provide better results, but it is a black-box approach.
```

# Chapter 7
# Conclusion

The FOSSEE Semester-long Internship provided me an opportunity to learn a lot, both from my fellow interns and my mentors, while working on a variety of projects. The R on Cloud project allows its users to access all TBC codes over a cloud platform. The task of maintaining it by finding and fixing bugs and errors helped me contribute towards promoting education via a digital medium. Writing scripts related to various machine learning topics for the R Spoken Tutorial series allowed me to strengthen my fundamentals of machine learning and contribute to the noble cause of providing free and high-quality education for all. Analysis of the FOSSEE workshop feedback data helped me to deepen my understanding of the workflow in a typical data science project. It helped me understand why cleaning and preprocessing of data is extremely important to obtain reliable results from an analysis. The SOM project was the most challenging endeavor that I have undertaken to date. It not only helped improve my programming skills in R, but it also taught me how I should break down a complex problem into simpler achievable steps. Coding an intricate machine learning model from scratch proved to be enlightening and helped me expand my frontiers. The Case Study project that I undertook helped me grasp the nuances of research work and taught me the importance of machine learning models for both researchers and their users.

Overall, the FOSSEE Semester-long Internship was much more than what the title seems to suggest. It taught me how to approach a problem, how I should collaborate with my teammates, how I should apply my critical thinking to discern and solve problems, and how I should constantly elevate my current skill level, among various other things. This was my first professional work experience, and it opened up a whole new world for me. Most importantly, my internship at FOSSEE allowed me to transcend to an enhanced level of skill while constructively contributing to society at large, paving the way for my future successes. I hope that my work helps promote the R programming language.

# 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] 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

[3] 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.

[4] 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).

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

[6] 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

[7] 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).

[8] 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.

[9] 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.

[10] 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

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

[12] 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

[13] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

[14] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071

[15] Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret

[16] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[17] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 289-317

[18] Kanti Mardia, J. Kent, J. Bibby. Multivariate Analysis, 1st Edition. December 14, 1979.

[19] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[20] Gareth, J., Daniela, W., Trevor, H. and Robert, T., 2013. An introduction to statistical learning: with applications in R. Spinger.

[21] Ng, Andrew (2000). "CS229 Lecture Notes" (PDF). CS229 Lecture Notes: 16–19.

[22] Thomas W. Yee (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

[23] Rokach, L., & Maimon, O. (2014). Data Mining With Decision Trees: Theory and Applications. World Scientific Publishing Co., Inc.

[24] Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

[25] Stephen Milborrow (2021). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.1.0. https://CRAN.R-project.org/package=rpart.plot

[26] Genuer, R. and Poggi, J.-M. (2020) Random forests with R. Cham: Springer International Publishing.

[27] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22.

[28] Arthur, D. and S. Vassilvitskii (2007). “k-means++: The advantages of careful seeding.” In H. Gabow (Ed.), Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms [SODA07], Philadelphia, pp. 1027-1035. Society for Industrial and Applied Mathematics.

[29] Georg M. Goerg (2013). LICORS: Light Cone Reconstruction of States - Predictive State Estimation From Spatio-Temporal Data. R package version 0.2.0. https://CRAN.R-project.org/package=LICORS

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

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

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

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

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

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

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

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

[38] UCLA Graduate School Admissions Data. https://stats.idre.ucla.edu/stat/data/binary.csv

[39] Winston Chang, Javier Luraschi and Timothy Mastny (2020). profvis: Interactive Visualizations for Profiling R Code. R package version 0.3.7. https://CRAN.R-project.org/package=profvis

[40] Covid-19 economic impact assessment template. https://data.adb.org/dataset/covid-19-economic-impact-assessment-template

[41] D. A. Freedman, Statistical Models: Theory and Practice. Cambridge University Press, 2009.

[42] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R (Second Edition), 2021.

[43] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0