Summer Fellowship Report

On

R Programming

Submitted by

Varshit Dubey and Shaik Sameer

Under the guidance of

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

July 3, 2018

Acknowledgment

First we would like to thank Prof. Kannan M. Moudgalya (Prof. Department of Chemical Engineering, IIT Bombay), the PI of FOSSEE, for giving us the opportunity to do an internship within the organization. We would like to thank Mrs. Usha Viswanathan (Senior Project Manager, FOSSEE) for managing our work place and creating an enjoyable working environment. We thank Mrs. Nancy Varkey (Senior Project Manager, IIT Bombay) for guiding us through different stages in Spoken Tutorial Project. We would also like to thank our mentor Mr. Sudhakar Kumar (Research Assistant, IIT Bombay) who was there for us all the time through out the internship guiding us at all stages of the project. Furthermore we want to thank all our fellow colleagues, with whom we completed the fellowship. We experienced great things together.

Contents

Figure 1: R Programming logo[**1**]

1 Introduction to R, Why R?

1.1 Introduction

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand,[**2**] and is currently developed by the R Development Core Team. R's pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka)[**2**].

Some of the interesting features of R are listed below:

1. R supports procedural programming with functions and object-oriented programming with generic functions.
2. Packages are part of R programming. Hence, they are useful in collecting sets of R functions into a single unit.
3. R has many packages for various functions.
4. R is an interpreted language. So we can access it through command line interpreter.
5. R has a very active community.
6. R has effective data handling and storage facilities.
7. R's programming features include:
   (a) database input
   (b) exporting data
   (c) variable labels
   (d) missing data

1.2 Data Visualization in R

With ever increasing volume of data in today's world, it is impossible to tell stories without these visualizations. While there are dedicated tools like Tableau, nothing can replace a modeling / statistics tools with good visualization capability. It helps tremendously in doing any exploratory data analysis as well as feature engineering. This is where R offers incredible help.

The following are some examples of data visualization in R:

1. **Pie Chart**
   A pie-chart is a representation of values as slices of a circle with different colors. The slices are labeled and the numbers corresponding to each slice is also represented in the chart.In R the pie chart is created using the pie() function which takes positive numbers as a vector input.

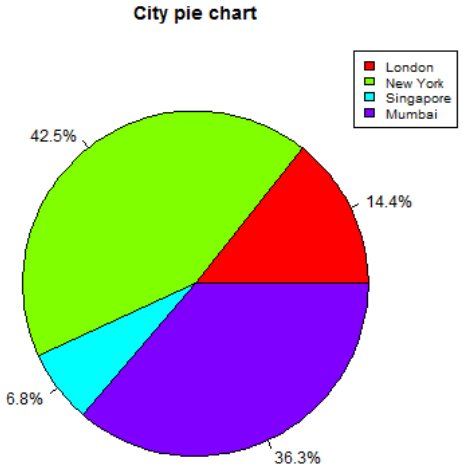


Figure 2: Pie Chart

2. **Histogram**
   A histogram represents the frequencies of values of a variable bucketed into ranges. Each bar in histogram represents the height of the number of values present in that range. R creates histogram using hist() function. This function takes a vector as an input and uses some more parameters to plot histograms.

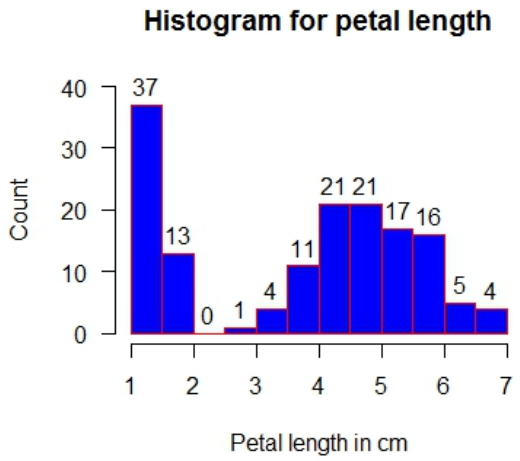


Figure 3: Histogram

3. **Bar Plot**
   A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function barplot() to create bar charts.

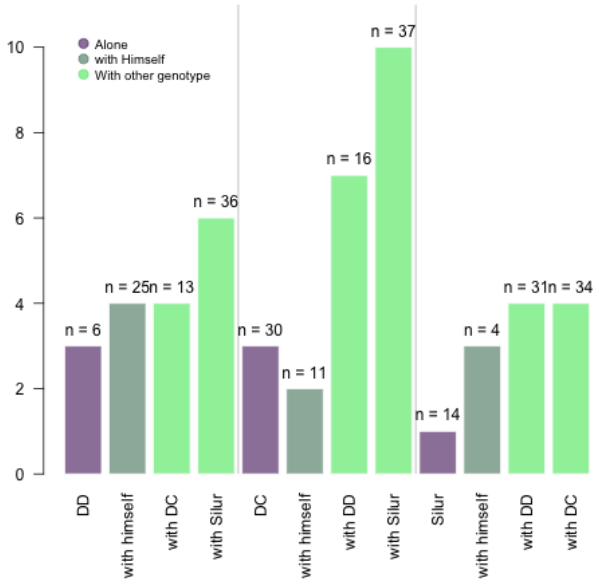


Figure 4: Bar Plot

4. **Box Plot**

Boxplots are a measure of how well distributed is the data in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set. Boxplots are created in R by using the boxplot() function.

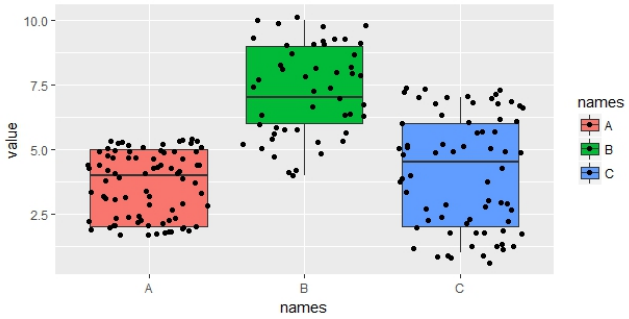


Figure 5: Box Plot

The above graphs can also be drawn using ggplot() which is also a plotting system for R but one which is much more fine and has granular control of everything.

1.3 Why R?

It is time to understand the importance of R Programming. We personally recommend that learning R is always a good option if one is into the Data Science field. Given below are some of the points to justify our recommendation:

- R programming language is best for statistics, data analysis and machine learning[**3**]. By using this language we can create objects, functions, and packages. It is platform-independent, so it can be applied to all operating systems. It's free, so anyone can install it in any organization without purchasing a license.

- R is open source. By using R, we can create any form of statistics and data manipulation. Furthermore, it can be used in almost every field like finance, marketing, sports etc.

- R, SAS, and SPSS are three statistical languages. Of these three statistical languages, R is the only Open Source language. R Programming is extensible and hence, R groups are noted for its energetic contributions. Lots of R's typical features can be written in R itself and hence, R has gotten faster over time and serves as a language to which everyone is glued to. The following table compares general and technical information of R with a selection of commonly used programming languages.

Table 1: Comparison Table

Language	**Intended Use**	**Imperative**	**OOP**	**Functional**	**Procedural**	**Generic**
C	Application, system, general purpose, low-level operations	Yes	No	No	Yes	No
Java	Application, business, client-side, general, mobile development, server-side, web	Yes	Yes	Yes	Yes	Yes
Python	Application, web, scripting, artificial intelligence, scientific computing	Yes	Yes	Yes	Yes	No
R	Application, statistics	Yes	Yes	Yes	Yes	No

2 Textbook Companion Project

2.1 Aim

- The Textbook Companion Project (TBC) aims to port solved examples from standard textbooks using R, so that it makes easy for users of such textbooks to start using R[**4**].
- To improve the documentation available for R.

2.2 Project Description

In this project we have to select one standard book which is approved by AICTE and propose to book to TBC team for their approval. Once the book is approved we start coding only the solved examples of the selected textbook. We chose statistical books as R is great for statistical computing. We learned a great deal of syntax, explored new packages for performing different types of statistical tests as we progressed on to finish our coding.

After the completion we have to submit all codes which are then reviewed by TBC team. The codes are reviewed as per the TBC guidelines and all the faulty codes are disapproved. We receive all the disapproved codes along with the reason for their disapproval. Then we correct our mistakes and resubmit the disapproved codes. Once our codes are approved by TBC team our codes are then made available on the R Textbook Companion website for public use. Later we have to review our fellow students work and help them in successful completion of their Textbook Companion Project. The whole process is depicted in the below shown flow chart.

We saw others codes and reviewed them based on the:

- syntax
- accuracy of final answer
- complexity of codes
- packages and built-in functions used

2.3 Conclusion

While we were making codes for solved examples, we got stuck at many places but R has a great built-in documentation which helped us throughout our project. We could learn how to optimize codes with built-in functions instead of hard coding, which is the case with other programming languages like C, C++ etc. We also learned better ways of coding through this fellowship.

While reviewing the codes we could learn great things like knowing some new packages. And also we shared our knowledge of how one can optimize his/her code. Reviewing helped us a lot as we became more confident in R programming.

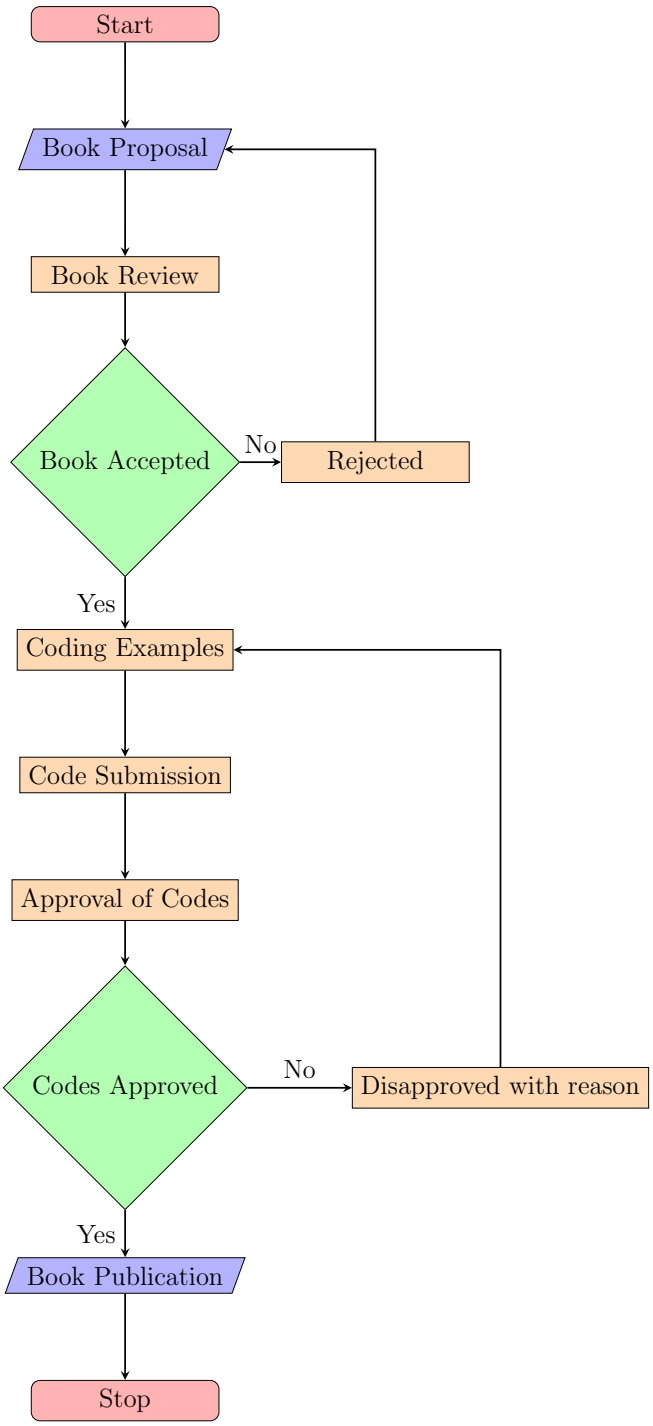


Figure 6: R Textbook Companion Project Flowchart

Figure 7: Spoken Tutorial logo[**5**]

3 Spoken Tutorial

3.1 Aim

The Spoken Tutorial project aims to make spoken tutorials on Free and Open Source Software (FOSS) available in several Indian languages, for the learner to be able to learn in the language he/she is comfortable in. The goal is to enable the use of Spoken Tutorials to teach in any Indian language, and to be taught to learners of all levels of expertise- Beginner, Intermediate or Advanced[**6**]. To achieve all this every tutorial has to go through a series of stages to ensure that it is perfect for its audience.

3.2 Project Description

The project is for the community and by the community. Through the portal, they aim to reach out to like-minded individuals to collaborate with them and with each other to create Spoken Tutorials. The first step for any intended contributor is to take the check list test. It is a small test of five minutes consisting of ten questions. It is intended to make the person know and remember the rules one has to abide to while contributing for the tutorials. Then we go through the process of creating the outline, charting out the script for each tutorial, recording and editing. The next step is to get each Spoken Tutorial dubbed into as many Indian languages as possible. Each of the Tutorials, whether original or dubbed, go through a strict review procedure, after which they are uploaded on the public domain to ensure high quality.

3.3 Conclusion

Working with the Spoken Tutorial team and creating the tutorials was an amazing experience. We realized that in order to make an effective tutorial, a lot of effort and time has to be put into it. Helping in creating these tutorials also helped us to make our basics strong and polish our fundamentals. There was also a feeling of satisfaction as our work would be part of something that would help thousands of people of our country.

4 Applications of R language

R programming allows to integrate with other languages (C/C++, Java, Python) and enables to communicate with many data sources like Excel and other statistical packages like SAS, SPSS, Minitab, Stata[**7**]. This makes it easy and interesting to use it in many prospects.

It's utilized in almost every field that you could think of. Nonetheless the widespread ones comprise - Finance, Bio Science, Supply chain, Sports, Retail, Marketing and Manufacturing. R has turned into the most prevalent language for data science and a fundamental tool for Finance and analytic-driven organizations, for example, Google, Facebook, and LinkedIn[**7**]. But it would be a mistake if one thinks that the usage of R is limited only to the IT segment of the society.

R is not offered by IT companies but all types of companies are hiring highly paid R candidates including:

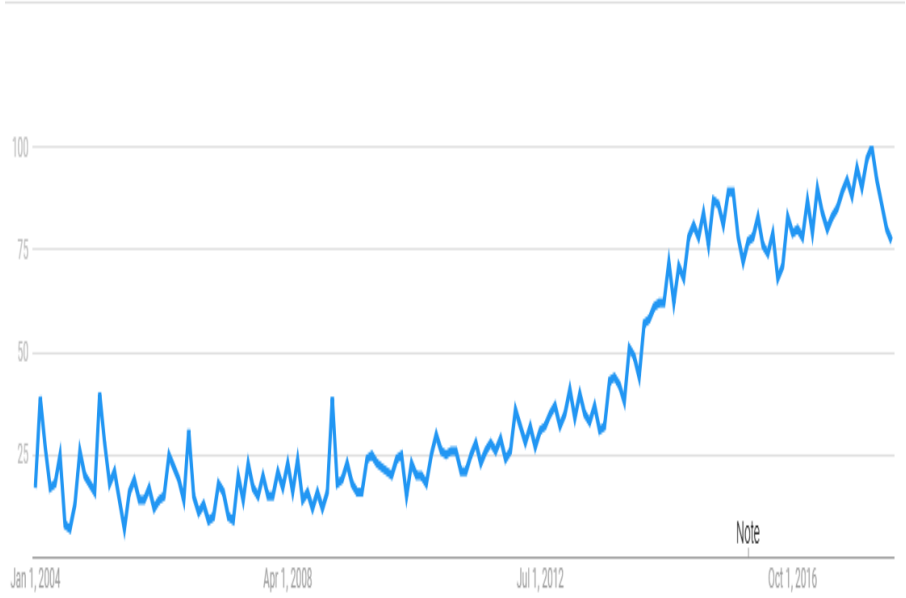
- Financial firms
- Retail organizations
- Banks
- Healthcare organizations etc.

Here are just a few examples of how R is used by some multinational companies[**8**]:

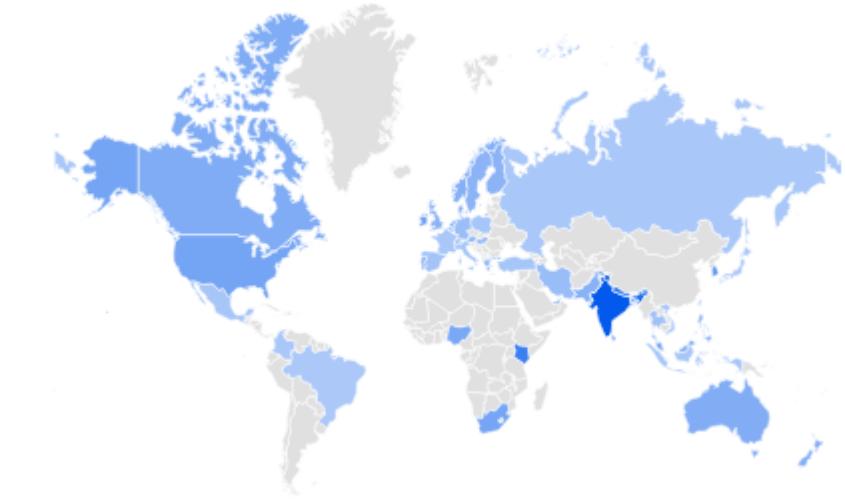
- **Google**: Basically, Google uses R to calculate Returns On Investment (ROI) on advertising campaigns and to predict economic activity. R is also used to improve the efficiency of online advertising.
- **Twitter**: R is part of Twitter's Data Science toolbox for sophisticated statistical modeling.
- **Facebook**: Basically, Facebook uses R to update Facebook status updates and its social network graph. It is also used for predicting colleague interactions.
- **Microsoft**: Microsoft uses R for the Xbox matchmaking service. It's also used as a statistical engine within the Azure Machine Learning (ML) framework.
- **Foodborne Chicago**: It is system which is entirely automated, and uses real-time text analysis implemented with R language to identify those tweets posted in the Chicagoland area that are about a specific case of food poisoning. R package called *textcat* and an algorithm based on n-grams is used to classify the tweets.[**9**]

What makes R so useful and one that gains quick acceptance is that statisticians, engineers and scientists can improve the software code or write variations for specific tasks. Packages written for R add advanced algorithms, colored and textured graphs and mining techniques to dig deeper into databases.

5 Future of R



(a) By Time



(b) By Region

Figure 8: Interest on R over time and by region[**10**]: (a) By Time, and (b) By Region.

IBM predicts that demand for Data Scientists will soar 28% by 2020[**11**]. Annual demand for the fast growing new roles of data scientist, data developers and data engineers will reach nearly 700,000 openings by 2020[**11**]. With such a high demand, companies are increasingly looking at people with expertise in R. Figure 8 (a) depicts the worldwide interest on R from the year 2004 to present. One can see that the popularity of R has significantly increased over time among the world. Figure 8 (b) shows the locations where R programming was most popular during the time period of 2004 to present. The figure shows that among the whole world India is one of the few countries where interest on R is very significant. Companies have several R job openings with various position like[**12**]:

- R Data Scientist
- Analyst Manager
- Business Analyst
- Senior Data Analyst.

A lot of companies have got a boost to innovation because of R's open source nature . In today's data-centric world, even a small bit of analysis that is used to predict customer needs or financial returns can mean the difference between success and failure. All these show a sign of tremendous increase in the usage of R in coming days.

The future scope of R language is very bright. R is very simple and easy to learn for peoples who are new to R. The recent average salary of R is best so you can think how high it will reach in future[**13**].

6 Conclusion

On the whole, this internship was a useful experience. We have gained new knowledge, skills and met many new people. We achieved several of our learning goals, and have moved a step further in achieving other.

We got insight into professional practice. We learned the different facets of working within an organization which is keen on promoting free and quality education. Social education is not one sided, but it is a way of sharing knowledge, ideas and opinions.

At last this internship has given us new insights and motivation to do something better than before for uplifting the society.

References

[1] [Online]
`https://www.r-project.org/logo`

[2] [Online]
https://en.wikipedia.org/wiki/R_(programming_language)

[3] [Online]
https://data-flair.training/blogs/introduction-to-r-programming/

[4] [Online]
https://r.fossee.in/textbook-companion-project

[5] [Online]
http://spoken-tutorial.org/static/spoken/images/logo.png

[6] [Online]
http://spoken-tutorial.org/mission/

[7] [Online]
https://elearningindustry.com/applications-r-programming-r-eal-world

[8] [Online]
https://data-flair.training/blogs/r-applications/

[9] [Online]
http://blog.revolutionanalytics.com/2013/08/foodborne-chicago.html

[10] [Online]
https://trends.google.com/trends/explore?date=all&q=r%20programming

[11] [Online]
https://www.forbes.com/sites/louiscolumbus/2017/05/13/
ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#275a883b7e3b

[12] [Online]
https://data-flair.training/blogs/career-growth-in-r-programming/

[13] [Online]
https://www.datacamp.com/community/blog/comparing-r-programmer-wages-by-industry