

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI-590 018



An Internship Report On

“Data Science with R”

Submitted in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

PRITAM RAJ

USN: 1KS15CS072

Internship Carried Out
at

iPRIMED Solutions Pvt. Ltd.

No.62B, 2nd Floor, 2nd Cross Road, Opposite Electronic City Post, Electronics City Phase 1,
Electronic City, Bengaluru, Karnataka 560100

Internal Guide

Mrs. Vijayalaxmi.Mekali

Assistant Professor

K.S. Institute of Technology

Bengaluru-560109

External Guide

Mr. Prasad

Managing Director

iPrimed Solutions Pvt. Ltd.

Bengaluru-560 100



Department of Computer Science & Engineering

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Road, Bengaluru-560 109

2018-2019

K.S. INSTITUTE OF TECHNOLOGY

#14, Raghuvanahalli, Kanakapura Road, Bengaluru-560 109

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING



CERTIFICATE

This is to certify that the Internship Training entitled “**Data Science with R**” presented by **PRITAM RAJ, USN: 1KS15CS072** of 8th semester in partial fulfillment of the award of Bachelor of Engineering in CSE in Visvesvaraya Technological University, Belagavi during the academic year **2018-19**. The Internship Training has been approved as it satisfies the academic requirements in respect of Internship Training work prescribed for the Bachelor of Engineering degree.

Internal Guide

HOD

Principal

Mrs. Vijayalaxmi.Mekali

Dr. Rekha B. Venkatapur

Dr.T.V.Govindaraju

**Assistant professor,
Dept. of CSE, KSIT**

**Head of the Department,
Dept. of CSE, KSIT**

**Principal/Director
K.S.I.T., Bengaluru**

External Viva

Name of the Examiners

Signature with date

- 1.
- 2.



iPRIMED Solutions Pvt. Ltd.

No.62B, 2nd Floor, 2nd Cross Road, Opposite Electronic City Post, Electronics City Phase 1,
Electronic City, Bengaluru, Karnataka 560100

CERTIFICATE

Certified that the internship program was successfully completed by **Mr. PRITAM RAJ**
USN: 1KS15CS072 a bonafide student of **K.S. INSTITUTE OF TECHNOLOGY** for
4 weeks from 9th July, 2018 to 7th August, 2018.

It is certified that, he has completed the internship satisfactorily.

Name & Signature

iPRIMED Solutions Pvt. Ltd.

**No.62B, 2nd Floor, 2nd Cross Road,
Opposite Electronic City Post,
Electronics City Phase 1, Electronic
City**

Bengaluru, Karnataka

DECLARATION

I, PRITAM RAJ(USN: 1KS15CS072) student of eight semester B.E, Computer Science and Engineering, K. S. Institute of Technology, hereby declare that the Internship Report entitled “Data Modelling” submitted to the Visvesvaraya Technological University, Belagavi during the academic year 2018-2019, is a record of original work based on the Internship carried out at iPRIMED Solutions. Under the External Guidance of Mr. Niranjan Balaji, Managing Director, iPRIMED Solutions and Internal Guidance of Mrs. Vijayalaxmi Mekali, Assistant Professor, Computer Science & Engineering, K. S. Institute of Technology. The Internship Report has been submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree.

Date:

Place: Bengaluru

PRITAM RAJ

1KS15CS072

EXECUTIVE SUMMARY

The internship training at iPRIMED Solutions was a storehouse of knowledge on communication related activities. The university provided us an opportunity to work in any reputed organization as a part of the curriculum to get hands on experience of practical issues in real time situations. I went through iPRIMED Solutions organization as I had an interest on communication. As a part of training, I was introduced to various departments of the company and meeting great resource persons of the organization. I was allotted a developer as my guide and I was following all the orders made by him. The guide made me more comfortable and the environment was student friendly. All the technical and non-technical staff was very helpful and very co-operative.

I was exposed to various tasks and was made to learn all the concepts and made me a good learner. The internship program consisted of learning all the technical activities of the organization. I learnt all the basics of the software's programming and server configuration. All the present and olden techniques involved in the organization were made to learn in the internship program.

Finally, the company people let me to get a real time experience on working and study of R programming. All the people in the organization helped me to successfully complete my internship training.

PRITAM RAJ

1KS15CS072

ACKNOWLEDGEMENT

The successful internship training would be incomplete without the mention of the people who made it possible and whose constant guidance crowned my effort with success.

I take this opportunity to express my sincere gratitude to our **Management K S Institute of Technology**, Bengaluru for providing the environment to present the Internship.

I would express my gratitude to **Dr. K.V.A. Balaji** C.E.O. K.S. Institute of Technology, Bengaluru, for facilitating me to present the Internship.

I would like to extend my gratitude to the Principal/Director, **Dr. T. V. Govindaraju**, K.S. Institute of Technology, Bengaluru, for facilitating me to present the Internship.

I thank **Dr. Rekha B. Venkatapur**, Professor and Head Department of Computer Science and Engineering K.S. Institute of Technology, Bengaluru for her encouragement.

I would also like to thank **Mr. K. VenkataRao**, Associate Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for his constant guidance and inputs.

I thank Internship Coordinator, **Mrs.Beena K** and **Mr. Prashanth H S**, Assistant Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for their constant support and guidance.

I would like to thank my internship guide, **Mrs. Vijayalaxmi.Mekali**, Assistant Professor, Department of Computer Science and Engineering, K.S. Institute of Technology, Bengaluru, for her constant support.

I would also like to thank my external guide, **Mr. Niranjan Balaji**, Managing Director, iPRIMED Solutions for his constant guidance and input.

I would like to thank all the teaching and non-teaching staff of the college for their co-operation.

Finally, I extend my heart-felt gratitude to my **family** for their encouragement and support without which I wouldn't have come so far. Moreover, I thank all my **friends** for their invaluable support and co-ordination.

Name of the Student

PRITAM RAJ

University Seat Number

1KS15CS072

ABSTRACT

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. Data Science is one of the core subjects that were thought at iPRIMED. The company provided us with an opportunity to work with R programming language. R is a programming language and software environment for statistical analysis, graphics representation and reporting. Logistic Regression was one of the most important topics introduced. It is a classification algorithm that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome. Large datasets were used and Logistic Regression was implemented on those datasets to get various outcomes. An example of a dataset containing BreastCancer data was implemented using Logistic Regression to find the best accuracy of whether the given specimen is benign or malignant based on nine other cell features.

Keywords: Data Science, R programming, Logistic Regression, Data Modelling

LIST OF CONTENTS

CHAPTER	CONTENT	PAGE NO
	ACKNOWLEDGEMENT	I
	ABSTRACT	II
1	INTRODUCTION	1
1.1	WHAT IS AN INTERNSHIP?	1
1.2	BENEFITS OF DOING AN INTERNSHIP	2
1.3	WHAT IS DATA SCIENCE?	2
1.4	WHAT IS R PROGRAMMING?	3
1.5	COMPARISON OF R	4
2	COMPANY PROFILE	5
2.1	IPRIMED	5
2.2	NASSCOM	6
3	ABOUT THE COMPANY	7
3.1	LEADERSHIP	7
3.2	MISSION OF NASSCOM	9
3.3	DEVELOPMENT OF INDIAN IT BY NASSCOM	9
3.4	PROGRAMS AND INTERVENTIONS	10
4	TASK PERFORMED	11
4.1	IMPOERTANCE OF DATA SCIENCE	15
4.2	APPLICATION OF DATA SCIENCE	15
4.3	WORKFLOW	17
5	REFLECTION	19
6	RESULTS/CODE SNIPPETS	20
	CONCLUSION	23
	REFERENCES	24

LIST OF FIGURES

FIGURE NO	FIGURE	PAGE NO
1.3	Data Science	3
4.1	Data Science and Applications	11
4.2	Loading Dataset	14
4.3	Setting Columns	14
6.1	Adding Vectors	20
6.2	Loading Iris Dataset	20
6.3	Geom Smooth Plot	21
6.4	Geom Smooth Plot=False	21
6.5	Statistics	22

INTERNSHIP CERTIFICATE



The Academic Council of iPRIMED certifies
Britam Raj from KSTIT, Bangalore
for successfully completing
SAP Yuva Yuga,

Training and Internship Programme conducted for
a period of 30 days between 9th July 2018 and 11th August 2018
under the aegis of NASSCOM Foundation.

We take the pleasure in recognizing the achievement with
the award of

Internship Certificate

In
Data Science with R
given on the 12th day of January 2019.

N. R. Rakesh

COO &
Chief Operating Officer

iPRIMEDTM
Building Industry Ready Professionals

A. J. J. J.

Head of
Academic Transformation

This certificate remains the property of iPRIMED Education Solutions Pvt. Ltd. to whom it must be returned on request.

Chapter 1

INTRODUCTION

1.1 What is an Internship?

Experience is becoming a crucial factor for employers when deciding who gets their foot in the door. It's strongly advised that students and graduates take the opportunity to complete a period of work experience to ensure they have a competitive advantage over their peers; and that's where an internship can make all the difference.

An internship is a period of work experience offered by an employer to give students and graduates exposure to the working environment, often within a specific industry, which relates to their field of study.

Internships can be as short as a week or as long as 12 months. They can be paid or voluntary; however, before you start an internship it's important to know your rights with regards to getting paid.

Internships can be done in a range of sectors, including sales, marketing, engineering, graphic design, management, I.T. and many, many more. Throughout an internship you will develop a variety of soft skills, including communication skills, personal effectiveness, presentation skills, creative problem solving and influencing skills.

'On-the-job' experience can be as valuable as anything learned in your studies. After all, you cannot really understand what a job is all about until you have worked in that environment. Internships are great opportunities to speak directly to people who have experience in the role you aspire to; and their knowledge of the job and working environment will give you a greater understanding of what it's all about and what you need to do to progress.

Your career aspirations may change when you're faced with the true realities of a role. Internships can therefore be used as a 'try before you buy' option, before you embark on a career and confirm if this is what you want to do in the long term.

An internship can give you a real insight into the world of work, allowing you to build on the theory you learned at university and helping you to gain practical skills that will help strengthen your CV and make you more employable. Internships offer you the chance to test your skills in real-life situations, explore your career options and gain an insight into an organization or career path.

1.2 Benefits of Doing an Internship?

Following a successful internship it is not unusual for employers to make a full-time job offer to their intern. Many employers use internships as a trial period and will already have plans to recruit on a permanent basis. Therefore, it's vital that you make a good impression; turn up on time, be enthusiastic and show your flexibility, adaptability and commitment.

Results from a recent survey conducted by Graduate Advantage prove that internships do create jobs for graduates. It showed that 81% of interns are now employed and 74% of those are either in permanent employment or are on a long-term contract. Of these, 68% believe their internship helped them to gain their current position and an impressive 33% are still working with their internship organization.

Michael Ellender of Birmingham Forward said of his internship: "I am a very proactive person and was keen to only take a role where I could use my graduate skills. In my experience, if you are willing to show initiative, enthusiasm and work hard, you will be given further opportunities to develop. I was pleased to stay on after the placement and have now been promoted to a higher level role that I enjoy."

1.3 What is data science?

Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems. Data science is a detailed study of the flow of information from colossal amounts of data present in an organization's repository. It involves obtaining meaningful insights from raw and unstructured data which is processed through analytical, programming, and business skills.

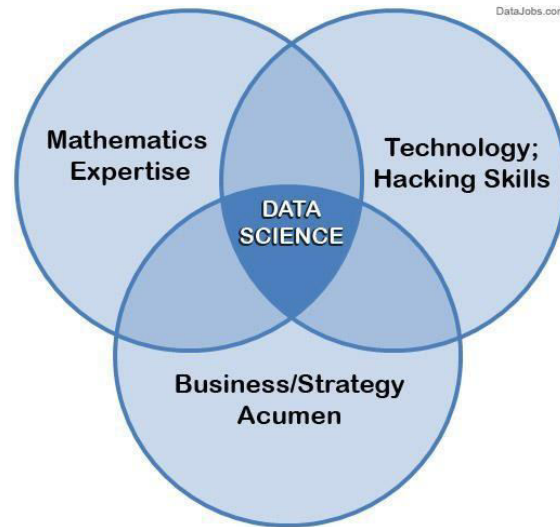


Fig. 1.3 Data Science

For any company that wishes to enhance their business by being more data-driven, data science is the secret sauce. Data science projects can have multiplicative returns on investment, both from guidance through data insight, and development of data product. Though, hiring people who carry this potent mix of different skills is easier said than done. There is simply not enough supply of data scientists in the market to meet the demand (data scientist salary is sky high). Thus, when you manage to hire data scientists, nurture them. Keep them engaged. Give them autonomy to be their own architects in how to solve problems. This sets them up in the company to be highly motivated problem solvers, there to tackle the toughest analytical challenges.

1.4 What is R programming?

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, and is currently developed by the R Development Core Team. The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is an implementation of the S programming language combined with lexical scoping semantics, inspired by Scheme. S was created by John Chambers in 1976, while at Bell Labs. R is an interpreted language user typically access it through a command-line

interpreter. There are some important differences, but much of the code written for S runs unaltered. R is named partly after the first names of the first two R authors and partly as a play on the name of S. The project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000. Like other similar languages such as APL and MATLAB, R supports matrix arithmetic.

1.5 Comparison of R with Other Technologies

- **Data handling Capabilities** – Good data handling capabilities and options for parallel computation.
- **Availability / Cost** – R is an open source and we can use it anywhere.
- **Advancement in Tool** – If you are working on latest technologies, R gets latest features.
- **Ease of Learning** – R has a learning curve. R is a low-level programming language. As a result, simple procedures can take long codes.
- **Job Scenario** – It is a better option for start-ups and companies looking for cost efficiency.
- **Graphical capabilities** – R is having the most advanced graphical capabilities. Hence, it provides you with advanced graphical capabilities.
- **Customer Service support and community** – R is the biggest online growing community.

Chapter 2

COMPANY PROFILE

2.1 iPRIMED

iPRIMED, a private limited company based out of Bangalore, is an NSDC partner and a licensed training provider of NASSCOM. With 9+ years of experience in the industry, 6 delivery hubs across the country, 50+ corporate clients, and proprietary learning methodology (patent pending), we are dedicated to enhancing the employability of students and graduates by driving intrinsic transformation.

We tackle the hardest area of skill development – bringing a significant behavioral change in learners. Our unique learning model, conceptualized in 2009, has demonstrated significant “behavioral shift,” resulting in the learners getting transformed – mature, holistic and always learning. The version 2.0 of the model is software-led but human-like.

Our digital platform, iPRIMED Athena (www.iprimedathena.com), can personalize the curriculum and learning pace at an individual level. Our solutions, powered by an optimal blend of human and technology interface, make use of gamified content, movies, and games in the context of business to make learning exciting and engaging and productive.



Fig. 2.1 Logo of the Company

iPRIMED is driven by a multi-dimensional team of 140+ members from premier institutions like IIMB and industry experts from Tier 1 Companies like Infosys, Wipro with rich and varied global experience.

- Website-<http://www.iprimed.com>
- Headquarters-Bangalore, Karnataka
- Year Founded-2009
- Company Type-Privately Held

- Size-51-200 employees

2.1.1 Specialties

Licensed NASSCOM Training Provider, Transforming Talent Supply Chain, NSDC Partner, Unique and innovative learning methodology, Contextual Learning, Blend of Human interface and Technology

2.2 NASSCOM

NASSCOM is a not-for-profit industry association committed towards supporting the IT BPM industry in becoming an indispensable powerhouse. NASSCOM, a not-for-profit industry association, is the apex body for the 154-billion-dollar IT BPM industry in India, an industry that had made a phenomenal contribution to India's GDP, exports, employment, infrastructure and global visibility.

Established in 1988 and ever since, NASSCOM's relentless pursuit has been to constantly support the IT BPM industry, in the latter's continued journey towards seeking trust and respect from varied stakeholders, even as it reorients itself time and again to remain innovative, without ever losing its humane and friendly touch.

NASSCOM is focused on building the architecture integral to the development of the IT BPM sector through policy advocacy, and help in setting up the strategic direction for the sector to unleash its potential and dominate newer frontiers.

NASSCOM's members, 2200+, constitute 90% of the industry's revenue and have enabled the association to spearhead initiatives at local, national and global levels. In turn, the IT BPM industry has gained recognition as a global powerhouse. In India, this industry provides the highest employment in the private sector.

Chapter 3

ABOUT THE COMPANY

The National Association of Software and Services Companies (NASSCOM) is a trade association of Indian information technology (IT) and business process outsourcing (BPO) industry. Established in 1988, NASSCOM is a non-profit organization is the apex body for the 154-billion-dollar IT BPM industry in India, an industry that has made a phenomenal contribution to India's GDP, exports, employment, infrastructure and global visibility.

NASSCOM's relentless pursuit has been to constantly support the IT BPM industry in India, in the latter's continued journey towards seeking trust and respect from varied stakeholders, even as it reorients itself time and again to remain innovative, without ever losing its humane and friendly touch. NASSCOM is focused on building the architecture integral to the development of the IT BPM sector through policy advocacy, and help in setting up the strategic direction for the sector to unleash its potential and dominate newer frontiers.

NASSCOM's members, 2200+, constitute 90% of the industry's revenue and have enabled the association to spearhead initiatives at local, national and global levels. In turn, the IT BPM industry has gained recognition as a global powerhouse. In India, this industry provides the highest employment in the private sector. The vision statement "To help the IT and IT enabled products and services industry in India to be a trustworthy, respected, innovative and society friendly industry in the world".

3.1 Leadership Team

Navin Kumar (Founder and CEO) – He is a BE from BITS, Ranchi (1993) and holds an MBA from IIM Bangalore (PGP 1999). He worked at Tata Steel for 4 years before joining MBA program at IIM Bangalore.

Nitish Raikar (COO) - Nitish holds Bachelor of Engineering from Goa University (1997) and MBA from XLRI Jamshedpur (2008). In his 20+ years of industry experience, he has been associated with companies like Infosys, IBM, Mindtree, Microland in various capacities

handling large projects and managing teams.

Nitish works as the Chief Operating Officer and ensures that iPRIMED is designed for scale and ready for growth. He does this by streamlining the internal operations and manages three business verticals – T&P, Content Development, and English.

MukundJhunhunwala(head-campus) - Mukund completed his B.Sc. from Kelly School of Business and is an entrepreneur by heart. Prior to joining iPRIMED, he has over 10+ years of experience as a self-grown entrepreneur. He has been the owner and promoter of companies in the manufacturing, power, infrastructure and renewable energy sector.

Mukund handles the digital segment of iPRIMED and focuses on expanding iPRIMED digital portfolio across campuses.

BashaShaik(head-technology) A tech evangelist by nature and having a rich 20+ years of experience in the technology industry, Basha comes with the aim of transforming iPRIMED into a technology driven organization. In his 20+ years of experience, he has lead large scale software solutions and products across industries - Banking, Supply Chain, Healthcare and IT. He has proven hands on experience in and around US, UK, Europe, Australia, Middle East and India as CIO, CTO and Delivery Head in corporates like Infosys and HCL.

Ratheesh Sebastian(head-finance) - Ratheesh is a Chartered accountant and handle the crucial role of ensuring that while we run a tight ship, finance doesn't become the bottleneck for growth. Prior to iPRIMED, he has been part of multiple startups, helping them successfully close series A round including FDI. A Aakillunisa(head-center of excellence) - Aakila comes with a rich experience of 16 years in the Learning Domain as an educator with a proven track record in Learning Methodologies. She has designed, developed and implemented learning models that have driven effective and efficient training delivery. She has brought innovation in practices of Business, Technology, People and Process – aided by her research-driven approach and skills that have helped the organization achieve more within a short span of time. She has worked spractices. Aakila heads the Centre of Excellence at iPRIMED. She holds the focal point of knowledge and content management and brings in new practices and business intelligence to the ongoing projects.

3.2 Mission of Nasscom

Recounting the adages NASSCOM to be associated with

- NASSCOM's ubiquitous raison d'être - 'Transform Business, Transform India'.
- Be a conduit of change through thought leadership, research, market intelligence and membership engagement.
- Establish India as a hub for innovation, products and start-ups.
- Work with the government to shape policy in key areas such as, skill development, trade, digital economy and business services.
- Increase the industry's outreach in its core markets and beyond, through strategic alliances.
- Be an industry platform for sharing and building best practices and collaborative engagement.
- Facilitate growth, and maintain India's leadership position as a trusted business destination.
- Expand the country's pool of relevant and skilled talent to drive inclusive and balanced growth.

3.3 Development of Indian It by Nasscom

NASSCOM is dedicated to expanding India's role in the global IT order by creating a conducive business environment, simplifying policies, procedures, promoting intellectual capital and strengthening the talent pool. Some of the Developments in Indian IT include

3.3.1 Global Trade Development

The Global Trade Development (GTD) Initiative at NASSCOM has two broad slivers:

- **Policy Advocacy**

In today's continuously evolving global regulatory environment, we work to ensure that Indian IT-BPM players remain abreast of various policy developments to try and reduce bottlenecks that have the propensity to impact business, and participate across geographies while conforming to their new laws and modified policies. NASSCOM actively works to

make representation on key policy challenges faced by industry mainly in developed markets including and not limited to US, UK, EU, Australia, Canada, South Africa and Singapore.

● **Market Development**

Indian IT-BPM companies have been expanding their geographic footprint for several years now. In addition to nurturing existing markets, NASSCOM is also focusing on building inroads into newer areas – geographies, verticals and customer segments. Several high growth and under-penetrated regions look promising for the IT-BPM business e.g., Nordics, Latin America, Africa, Middle East, ASEAN, China, Japan among few. While supporting member companies in creating a favourable eco-system to promote business growth, we also create suitable platform for forging Technology Partnerships & Alliances that are likely to bring long-term strategic benefit.

3.4 Programs & Interventions

Market Development programs are conceived with an objective to provide a composite exposure to participating companies including awareness on business landscape in general and ICT in specific, exploration of market & investment opportunities, identifying suitable partnerships and one to one networking with probable clients & decision makers such as CIOs/CTOs/ Digital Leaders etc. These programs constitute of large, medium & small sized companies with a high percentage of SME participation.

Particularly for small & medium enterprises, such programs have helped shrink their learning curve and get a 360 degree view on a particular geography and plan their go-to-market strategy.

NASSCOM has been organizing market development programs in various formats such as

- Overseas business Delegations
- Participation in International Expos
- Creating region specific reports to highlight in-depth market opportunities
- Interactive learning sessions with market experts & industry leaders

Chapter 4

TASK PERFORMED

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data which uses the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problem. Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science and computer science.

Data Science is a field that covers data cleansing, preparation, and analysis. It includes several scientific methods, such as mathematics, statistics, and many other tools data scientists apply to extract knowledge from data sets as shown in Figure 4.1.

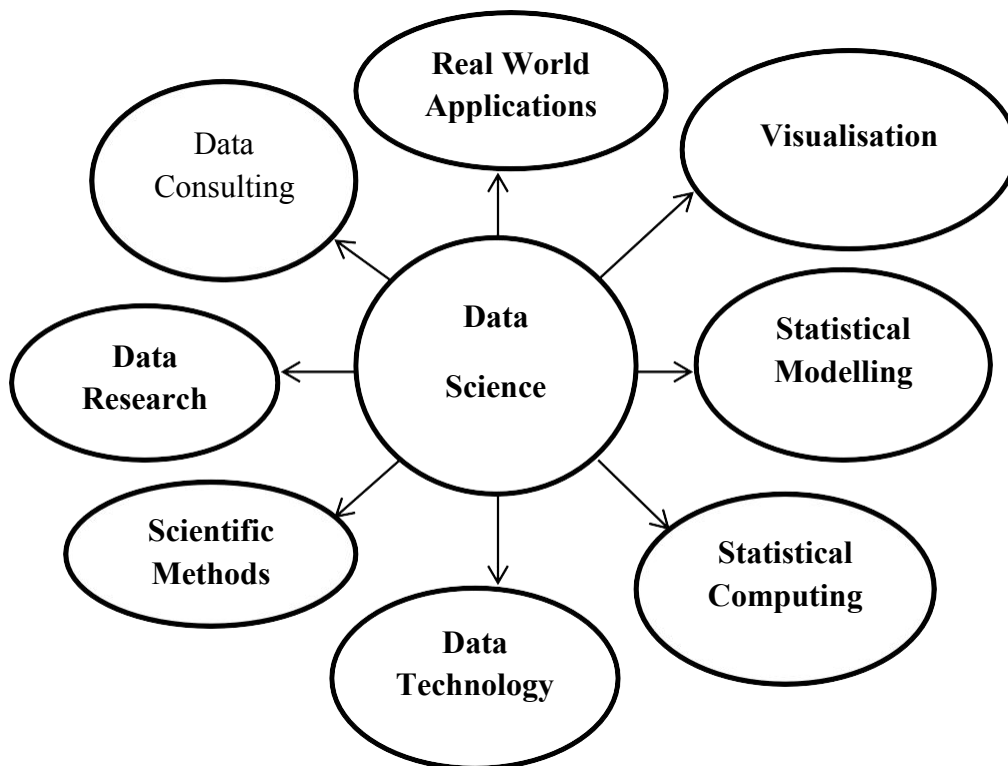


Fig. 4.1 Data Science Applications

First, the Data scientist gathers datasets from multi-disciplines and compiles it. Next, he applies machine learning, predictive and sentimental analysis to analyze the data and eventually deciphers a meaningful pattern out of the huge datasets.

A data scientist makes use of tools and languages like R, MATLAB, and DB Management for data analysis and machine learning. Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education.

- Computer Science- Pattern recognition, visualization, data warehousing, Artificial Intelligence, High performance computing, Databases.
- Mathematics- Mathematical Modeling.
- Statistics- Statistical and Stochastic modeling, Probability.

Machine Learning in R: Step-By-Step

- Installing the R platform.
- Loading the dataset.
- Summarizing the dataset.
- Visualizing the dataset.
- Evaluating some algorithms.
- Making some predictions.

Downloading Installing and Starting R

Here is what we are going to cover in this step:

- Download R.
- Install R.
- Start R.
- Install R Packages.

- **Download R**

You can download R from The R Project webpage.

When you click the download link, you will have to choose a mirror. You can then choose R for your operating system, such as Windows, OS X or Linux.

- **Install R**

R is easy to install and I'm sure you can handle it. There are no special requirements. If you have questions or need help installing see R Installation and Administration.

- **Start R**

You can start R from whatever menu system you use on your operating system. For me, I prefer the command line.

Open your command line, change (or create) to your project directory and start R.

- **Install Packages**

Install the packages we are going to use today. Packages are third party add-ons or libraries that we can use in R.

- **UPDATE:**

We may need other packages, but caret should ask us if we want to load them. If you are having problems with packages, you can install the caret packages and all packages that you might need by typing.

Now, let's load the package that we are going to use in this tutorial, the caret package.

The caret package provides a consistent interface into hundreds of machine learning algorithms and provides useful convenience methods for data visualization, data resampling, model tuning and model comparison, among other features. It's a must have tool for machine learning projects in R.

For more information about the caret R package see the caret package homepage.

- **Load The Data**

We are going to use the iris flowers dataset. This dataset is famous because it is used as the "hello world" dataset in machine learning and statistics by pretty much everyone. The dataset contains 150 observations of iris flowers. There are four columns of measurements of the flowers in centimeters. The fifth column is the species of the flower observed. All observed flowers belong to one of three species.

You can learn more about this dataset on Wikipedia. Here is what we are going to do in this step:

- Load the iris data the easy way.
- Load the iris data from CSV (optional, for purists).
- Separate the data into a training dataset and a validation data.

- **Load Data the Easy Way**

Fortunately, the R platform provides the iris dataset for us. Load the dataset as follows:
You now have the iris data loaded in R and accessible via the dataset variable.

```
1# attach the iris dataset to the environment
2data(iris)
3# rename the dataset
4dataset <- iris
```

Fig. 4.2 Loading dataset

I like to name the loaded data “dataset”. This is helpful if you want to copy-paste code between projects and the dataset always has the same name.

- **Load From CSV**

Maybe you are a purist and you want to load the data just like you would on your own machine learning project, from a CSV file.

- Download the iris dataset from the UCI Machine Learning Repository (here is the direct link).
- Save the file as iris.csv your project directory. Load the dataset from the CSV file as follows:

You now have the iris data loaded in R and accessible via the dataset variable.

```
1# define the filename
2filename <- "iris.csv"
3# load the CSV file from the local directory
4dataset <- read.csv(filename, header=FALSE)
5# set the column names in the dataset
6colnames(dataset) <-
  c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Species")
```

Fig. 4.3 Setting Columns

- **Create a Validation Dataset**

We need to know that the model we created is any good.

Later, we will use statistical methods to estimate the accuracy of the models that we create on unseen data. We also want a more concrete estimate of the accuracy of the best model on unseen data by evaluating it on actual unseen data.

That is, we are going to hold back some data that the algorithms will not get to see and we will use this data to get a second and independent idea of how accurate the best model might actually be.

We will split the loaded dataset into two, 80% of which we will use to train our models and

20% that we will hold back as a validation dataset.

4.1 IMPORTANCE OF DATA SCIENCE

Data science helps brands to understand their customers in a much enhanced and empowered manner. Customers are the soul and base of any brand and have a great role to play in their success and failure. With the use of data science, brands can connect with their customers in a personalized manner, thereby ensuring better brand power and engagement.

- One of the reasons why data science is gaining so much of attention is because it allows brands to communicate their story in such an engaging and powerful manner. When brands and companies utilize this data in a comprehensive manner, they can share their story with their target audience, thereby creating better brand connect. After all, nothing connects with consumers like an effective and powerful story, that can inculcate all human emotions.
- Big-Data is a new field that is constantly growing and evolving. With so many tools being developed, almost on a regular basis, big data is helping brands and organizations to solve complex problems in IT, human resource, and resource management in an effective and strategic manner. This means effective use of resources, both material and non-material.
- One of the most important aspect of data science is that its findings and results can be applied to almost any sector like travel, healthcare and education among others.
- Data science is accessible to almost all sectors. There is a large amount of data available in the world today and utilizing them in a proper manner can spell success and failure for brands and organizations. Utilizing data in a proper manner will hold the key for achieving goals for brands, especially in the coming times.

4.2 APPLICATIONS OF DATA SCIENCE

- **Fraud and Risk Detection**-Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default.

- **Healthcare**-health care applications include Medical Image Analysis, Genetics and Genomics, Drug Development, Virtual Assistant for Patients and Customer Support.
- **Internet Search**-There are many other search engines like Google, Yahoo, Bing, Ask, AOL, and so on. All these search engines make use of data science algorithms to deliver the best result for our searched query in a fraction of seconds.
- **Targeted Advertising**-The entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms.
- **Website Recommendations**-Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, iamb and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.
- **Advanced Image Recognition**-When You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. In their latest update, Facebook has outlined the additional progress they've made in this area, making specific note of their advances in image recognition accuracy and capacity.
- **Speech Recognition**-Some of the best examples of speech recognition products are Google Voice, Siri, Cortana etc. Using speech-recognition feature, even if you aren't in a position to type a message, your life wouldn't stop.
- **Airline Route Planning**- Now using data science, the airline companies can Predict flightdelay, decides which class of airplanes to buy, whether to directly land at the destination or take a halt in between and Effective drive customer loyalty programs.
- **Gaming**-Games are now designed using machine learning algorithms which improve/upgrade themselves as the player moves up to a higher level. EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science.

4.3 WORKFLOW

Allow me to briefly list down the workflow that I have gone through as these are what that has built my foundation in Data Science. And I hope you will find it useful in some ways.

1. Understanding the Business Problem

The project chosen was about Short Term Freeway Travel Time Prediction. However, like I said, asking the right questions is very important for a Data Scientist. A lot of questions were raised to really understand the real business problem before the project was finalized, be it data sources available, the end goals of the project (even after I left) etc. Essentially, our objective was to predict travel time for a freeway in Singapore N minutes ahead more accurate than the current baseline estimation.

2. Collecting Data Source

Excited with the new project, I started collecting data sources from database and colleagues (basically walking around the office to ask questions on data sources). Collecting the right data source is similar to the case where you are scraping data from different websites for data preprocessing later. It is so important that it could affect the accuracy of the models that you are building in the later stage.

3. Data Preprocessing

Real world data is dirty. We can't expect a nicely formatted and clean data as provided by Kaggle. Therefore, data preprocessing (other people might call it data munging or data cleaning) is so crucial that I can't stress enough how important it is. It is the most important stage as it could occupy 40%-70% of the whole workflow, just to clean the data to be fed to your models.

Garbage in, Garbage out

One of the things that I like about data science is that you have to be honest to yourself. When you don't know what you don't know, and you think the data preprocessed is already clean enough and ready to feed to your models, therein lies a risk of building the correct models with the wrong data. In other words, always try to question yourself if the data is technically correct with the domain knowledge that you have, scrutinize the data with stringent threshold to check for any other outliers, missing or inconsistent data in the whole datasets.

I was particularly careful about this after I made a mistake of feeding the models with the wrong data, just because of a simple flaw in one of the preprocessing steps.

4. Building Models

After some research, I proposed four models to be used in our project, which were Support Vector Regression (SVR), Multilayer Perceptron (MLP), Long Short Term Memory (LSTM), and State Space Neural Networks (SSNN). For the sake of brevity, you can find detailed explanation of each model on various websites.

Building different models from scratch was a steep learning curve for me as a person who was still learning from MOOCs and textbooks. Fortunately, Scikit-learn and Keras (with Tensorflow backend) came to my rescue as they are easy to learn for fast models prototyping and implementation in Python. In addition, I also learned how to optimize the models and fine-tuned the hyperparameters for each model using several techniques.

5. Models Evaluation

To evaluate the performance of each model, I used mainly a few metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Coefficient of Determination (R²)

At this stage, Steps 3–5 were repeated (interchangeably) until the best model was determined that could outperform the baseline estimation.

Chapter 5

REFLECTION

Well, the internship has definitely reaffirmed my passion in Data Science and I am grateful that my works did leave some traction for future works. The research and development phase, the communication skills required to talk to different stakeholders, the curiosity and passion to solve business problems using data (just to name a few) have all contributed to my interest in this field.

Data Science industry is still very young and its job description could somehow seem vague and ambiguous to job seekers like us. It's perfectly normal to not possess all the skills needed as most job description is idealistically created to align with their best expectation .

When in doubt, just learn the fundamentals from MOOCs, books, and articles (which I am still doing) and apply what you have learned through your own personal projects or internships. Be patient. The learning journey does take time. Learn from your journey with relish.

Internship impact:

- Knowledge gain
- Experience
- Leadership quality
- Presentation skills
- Opportunity to learn from failures.
- Get exposed to workplace
- Enhance your resume

Chapter 6

RESULTS

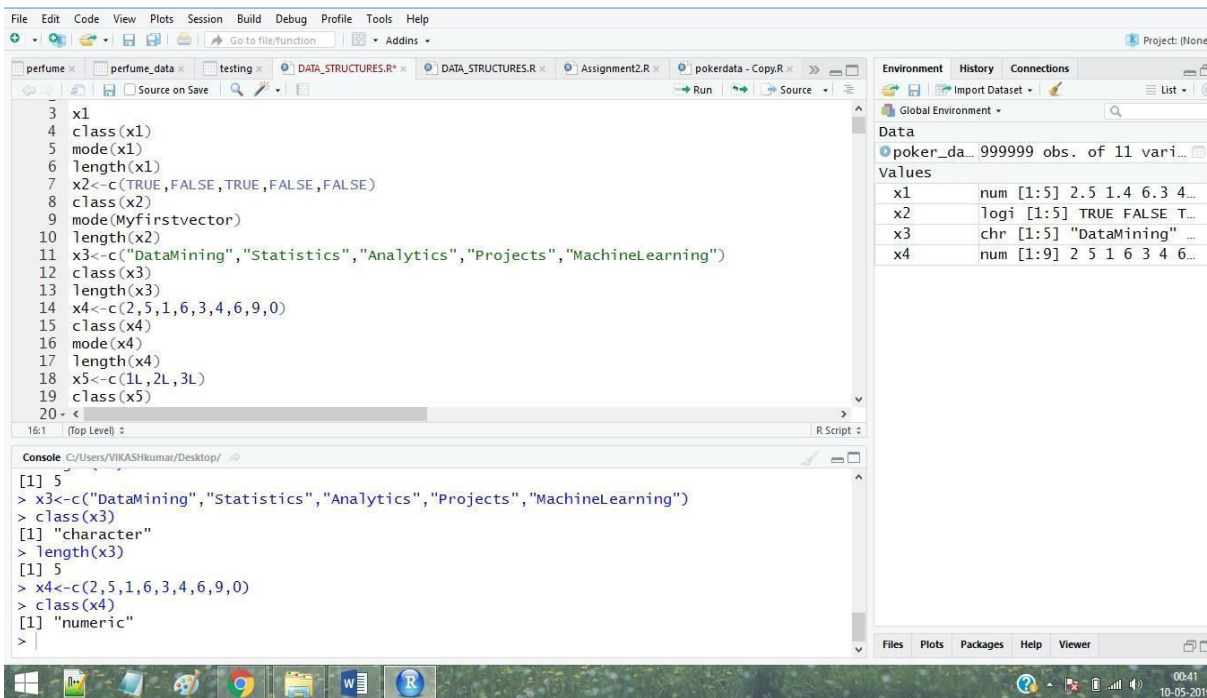


Fig. 6.1 Adding Vectors

The above figure shows functions to add vectors.

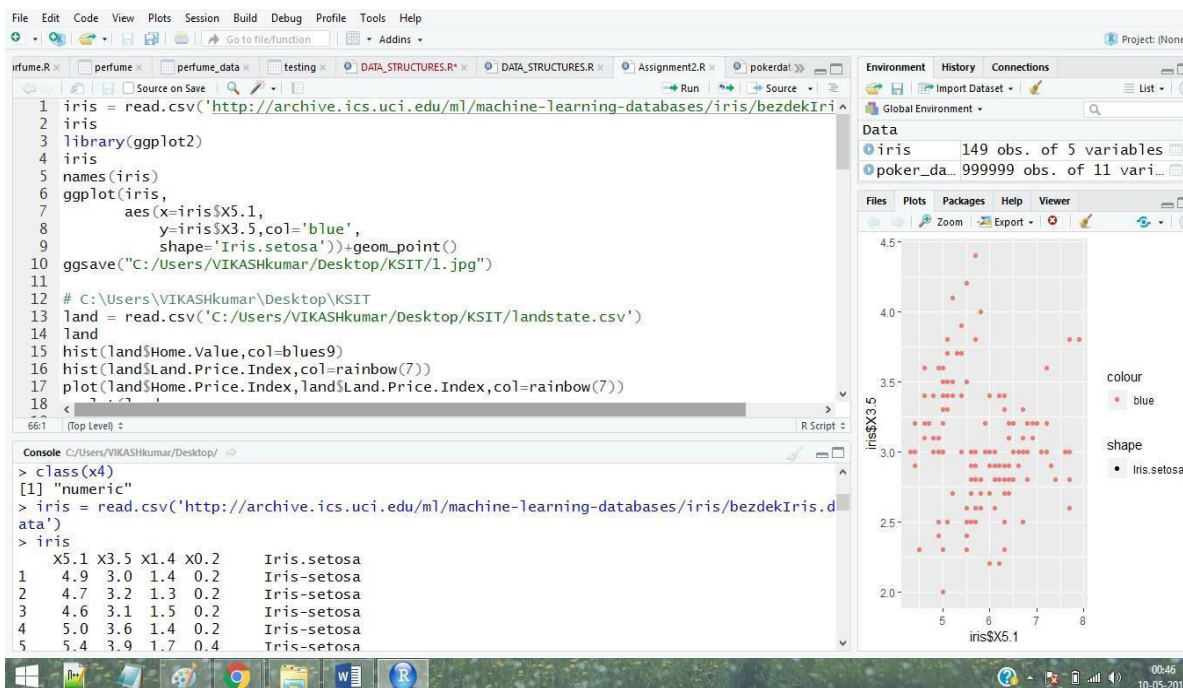


Fig. 6.2 Loading Iris dataset

The above figure explains how to load dataset.

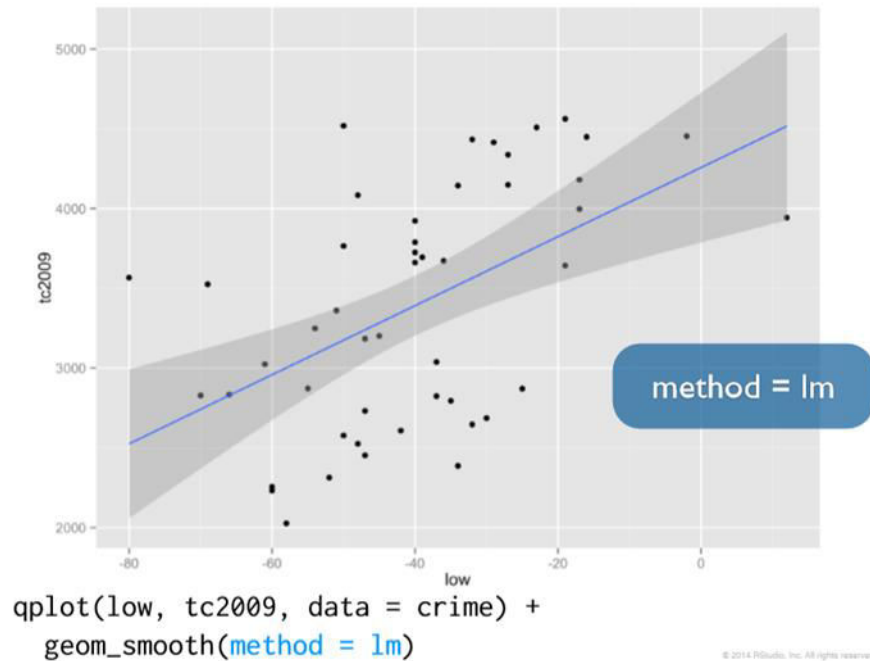


Fig. 6.3 Geom smooth plot

The above figure shows geometrical smooth plotting of variables.

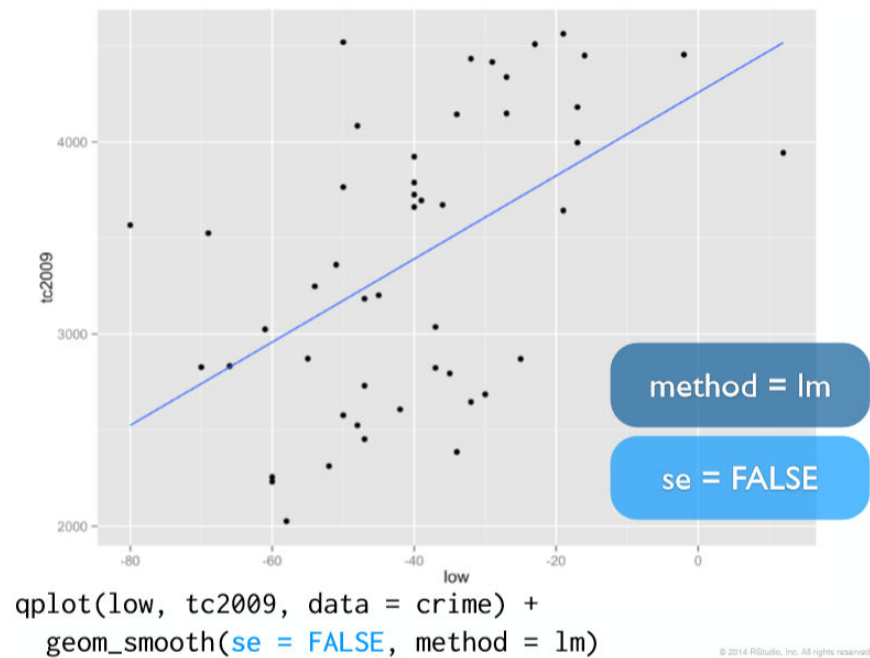
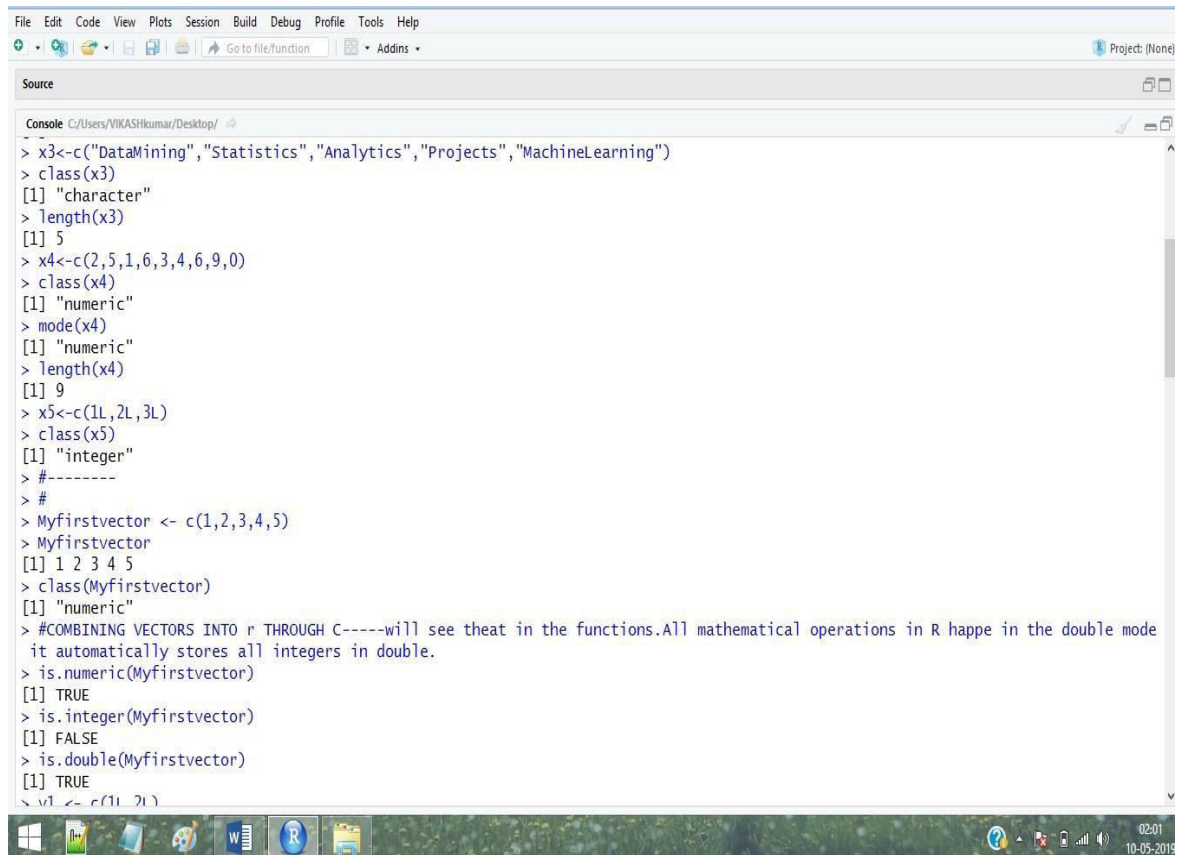


Fig. 6.4 Geom smooth plot=false

The above figure explains geometrical smooth plot when condition is false.



```
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)

Source

Console C:/Users/VIKASHKumar/Desktop/
> x3<-c("DataMining","Statistics","Analytics","Projects","MachineLearning")
> class(x3)
[1] "character"
> length(x3)
[1] 5
> x4<-c(2,5,1,6,3,4,6,9,0)
> class(x4)
[1] "numeric"
> mode(x4)
[1] "numeric"
> length(x4)
[1] 9
> x5<-c(1L,2L,3L)
> class(x5)
[1] "integer"
> #-----
> #
> Myfirstvector <- c(1,2,3,4,5)
> Myfirstvector
[1] 1 2 3 4 5
> class(Myfirstvector)
[1] "numeric"
> #COMBINING VECTORS INTO R THROUGH C-----will see that in the functions. All mathematical operations in R happen in the double mode
it automatically stores all integers in double.
> is.numeric(Myfirstvector)
[1] TRUE
> is.integer(Myfirstvector)
[1] FALSE
> is.double(Myfirstvector)
[1] TRUE
> v1 <- c(1L, 2L)
```

Fig. 6.5 Statistics

The above figure explains statistics for dataset.

CONCLUSION

Data science is important it empowers professionals with data management technologies like Hadoop, R, Flume, Sqoop, Machine learning, Mahout etc. The knowledge and expertise of the skills is an added advantage for a better and competitive career. By migrating current database apps to MYSQL, enterprises are enjoying significant cost saving on new projects. Finally, I conclude by saying that R is great tool to explore and investigate the data, it elaborates the analysis like clustering, correlation and redundant data reduction. Helped me to define what skills and knowledge I have to improve in parallel with the development, Manage time. R is free and open-source, making it possible for anyone to have access to world-class statistical analysis tools. R performs a wide variety of functions, such as data manipulation, statistical modelling, and graphics. It is used widely in academia and the private sector and is the most popular statistical analysis programming language today.

REFERENCES

- [1] H. Yusuff¹, N. Mohamad², U.K. Ngah³ & A.S. Yahaya⁴, breast cancer analysis using logistic regression, IJRRAS 10 (1) January 2012
- [2] Austin, J. T., Yaffee, R. A., & Hinkle, D. E. Logistic regression for research in higher education. Higher Education: Handbook of Theory and Research, 8, 379–410, (1992).
- [3] Archer, K. J., S. Lemeshow, and Hosmer, D. W., Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. Computational Statistics & Data Analysis 51(9) (2007): 4450-4464.
- [4] DesJardins, S.L. A comment on interpreting odds-ratios when logistic regression coefficients are negative. The Association for Institutional Research, 81, 1-10. (2001).
- [5] Harrell, F.E. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer Science+Business Media, Inc. (2001).