Summer Internship 2020

Daily Log

“Data Analysis using R”

Submitted by:

Bhartendu Dubey (9917103102)

Under the supervision of:

Ms. Ambalika Sarkar

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

June 2020

Phase1 (*01/05/2020-21/05/2020*)

DATE: 01/05/2020

- What is Data Science & Data Analytics?
- Exploring the need of Data Analytics.
- Understanding scope of Big Data Analytics.

DATE: 02/05/2020

- Understanding the R Environment
- Installation of R
- Installation of R-Studio.

DATE: 03/05/2020

- Basic Data Types of R.
- Variables
- Practicing Some simple examples on iris dataset.

DATE: 04/05/2020

- Vectors
- Matrices & Arrays
- 3D-Arrays
- List

DATE: 05/05/2020

- Data Frames in R.
- Data Sets in R.
- Importing the Data in R-Studio.

DATE: 06/05/2020

- Exploring R Packages
- Understanding concept of Session Management & Realization

DATE: 07/05/2020

- Dealing with Expressions & Objects.
- Understanding Factors & their role in handling categorical data.

DATE: 08/05/2020

- Working of Functions in R
- Logical Comparisons

DATE: 09/05/2020

- Conditional Selection (Sub-setting)
- Modifying Objects

DATE: 10/05/2020

- Indexing by a vector of positive Integer.
- Indexing by a vector of negative Integer.
- Indexing by a logical vector.
- Indexing by a vector of names.

DATE: 11/05/2020

- Loops (if, ifelse, switch, for, while)
- Loop Control (Break & Next)

DATE: 12/05/2020

- Functions & Data Frames
- Combining Data Frames (cbind)

DATE: 13/05/2020

- Understanding Data Manipulation
- Implementation of Data Manipulation with dplyr

DATE: 14/05/2020

- Understanding Difference of Qualitative and Quantitative Data
- Understanding the need of Qualitative and Quantitative Data

DATE: 15/05/2020

- Data Visualization - Continuous One Variable
- Data Visualization - Discrete One Variable

DATE: 16/05/2020

- Data Visualisation - Continuous X and Y

DATE: 16/05/2020

- Plotting of Histograms on Iris dataset.
- Plotting of Density plots on Iris dataset.
- Plotting of Pie chart on Iris dataset.
- Plotting of Bar chart on Iris dataset.

DATE: 17/05/2020

- Plotting of Box Plot on Iris dataset.
- Plotting of Scatter Plot on Iris dataset.
- Plotting of Scatter Plot with jitter.
- Plotting a matrix of Scatter Plots.

DATE: 18/05/2020

- Plotting of Heat Maps on Iris dataset.
- Plotting of Level Plot on Iris dataset.
- Plotting of Contours.

DATE: 19/05/2020

- Plotting 3D-surface diagrams.
- Plotting parallel coordinates.
- Simple Descriptive Statistics (mean, median, Inter quartile range)
- Understanding Covariance & Correlation.

DATE: 20/05/2020

- Understanding Data Mining, it’s need & data warehousing
- R Packages and Functions for Data Mining

DATE: 21/05/2020

- Implementing Clustering (K-Means Clustering) in R.

Phase2

(*22/05/2020-15/06/2020*)

DATE: 22/05/2020-26/05/2020

- Finalising the Title of Project work.

DATE: 27/05/2020

- Explored the required libraries (tidyverse, ggthemes, RColorBrewer, kableExtra, knitr, ggrepel, scales, gridExtra, tidytext, wordcloud, lubridate, igraph, ggraph) for working on project.

DATE: 28/05/2020

- Installed required libraries (tidyverse, ggthemes, RColorBrewer, kableExtra, knitr, ggrepel, scales, gridExtra, tidytext, wordcloud, lubridate, igraph, ggraph).

DATE: 29/05/2020

- Built functions for calculating ratios(fractions) & percent.
- Built functions for calculating age.

DATE: 30/05/2020

- Started EDA by finding ratio of Male: Female Award winners.
- Visualised the plot for gender ratio.

DATE: 31/05/2020

- Visualised category wise Male: Female Award Winners count.

DATE: 01/06/2020

- Plotted a visualisation for Prizes won in each category.

DATE: 02/06/2020

- Visualised the trend for “Female Laureates Proportion per decade”.
- Visualised the trend for “Male Laureates Proportion per decade”.

DATE: 03/06/2020

- Plotted visuals for “Laureates count Per Prize”.

DATE: 04/06/2020

- Identified Laureates with multiple Nobel Prizes.
- Tabulated the records.

DATE: 05/06/2020

- Plotted scatter plot for “Age vs Year of Nobel Prize Win”.

DATE: 06/06/2020

- Box Plot for “Age Distribution by Category”.
- Scatter Plot for “Age Trend for receiving Nobel price per Category”.

DATE: 07/06/2020

- Identified youngest & oldest Nobel laureates.
- Analysed data for finding first female laureate in each category.

DATE: 08/06/2020

- Visualised “Life Span of Nobel Laureates” in form of a histogram.
- Plotted histogram for “Life Span By category”.

DATE: 09/06/2020

- Box plot for “LifeSpan Distribution by Category”.
- Box plot for “LifeSpan Vs Awarded age” for each category.

DATE: 10/06/2020

- Bar chart for visualising “Prizes won by each country”.
- Line chart for “USA-Year wise proportion for awards”.
- Line chart for “INDIA-Year wise proportion for awards”.

DATE: 11/06/2020

- Plotted world maps for Nobel Laureates distribution in world (for each category).

DATE: 12/06/2020

- Visualised the “Most frequently used words by Nobel laureates” from their motivation quotes in form of a bar chart.
- Created a Word Cloud of motivation key-terms from motivation quotes of nobel laureates.

DATE: 13/06/2020-15/06/2020

- Documentation of R-Notebook code with proper self-explanatory comments.
- Project Report documentation.