EX NO: 4 DATE:	**Extract text data by using RScrapee based on the term 'A Free Online Data Science Courses' and perform EDA**

AIM: To extract text data by using Rscrapee based on the term 'A Free Online Data Science Courses'

ABOUT DATASET:

This dataset is scrapped using WebScrappe - Rvest funcion from www.greatlearning.com and it contains columns such as Courses, Duration, Level, Enrollments, Ratings, etc. Duration of the course will be in hours, Level will be in beginner or intermediate, Enrollments contains number of people enrolled, Ratings contain the rating of the course in floating. The dataset contains 57 rows and 6 columns.

ALGORITHM:

Step 1: Start the program by including all the required packages and call them using library().

Step 2: Create an R object that hold the url of the page from which the data will be scrapped.

Step 3: Using map_df() create a loop that should iterate one by one till to the last page of the url.

Step 4: Create a Data frame with required fields.

Step 5: Using a Selector gadget extension, copy the required element's Xpath/css and pass it inside the "html_nodes () as an argument.

Step 6: Run the code and View the Data frame created.

Step 7: Stop the Program

CODE:

```
---
title: "WebScrappe"
author: "Akshaya"
output:
  html_document:
    df_print: paged
---
```{r load libraries, include=FALSE}
library(rvest)
library(purrr)
library(xml2)
library(stringr)
library(knitr)
library(skimr)
```
```{r}
url1 = "https://www.mygreatlearning.com/data-science/free-courses?p=%d"

map_df(1:6, function(i){
  page <- read_html(sprintf(url1, i))

  data.frame(Courses = html_text(html_nodes(page, ".course-name")),
             Duration = html_text(html_node(page, ".course-info div:nth-child(1)")),
             Level = html_text(html_node(page, ".dot:nth-child(2)")),
             Enrollments = html_text(html_node(page, ".dot+ .dot")),
             Ratings = html_text(html_node(page, ".course-ratings-label"))
  )
}) -> course
View(course)
colnames(course)
str(course)
```
```

```

### Statistical inference

```{r, include=FALSE}
skim(course)
```

```{r}
course$Courses <- trimws(course$Courses, which = c("both"))
course$Duration <- trimws(course$Duration, which = c("both"))
course$Level <- trimws(course$Level, which = c("both"))
course$Enrollments <- trimws(course$Enrollments, which = c("both"))
head(course)
```

Type conversion

```{r}
course$Ratings <- as.numeric(course$Ratings)
```

Ratings column data type has been converted from character to numeric for analysis

Filtering conditions

```{r}
table(course$Level)
```

Histogram

```{r}
hist(course$Ratings, col = rainbow(5), main = "Histogram of Course Ratings")
```

Many courses is rated from 4.4 to 4.6
```

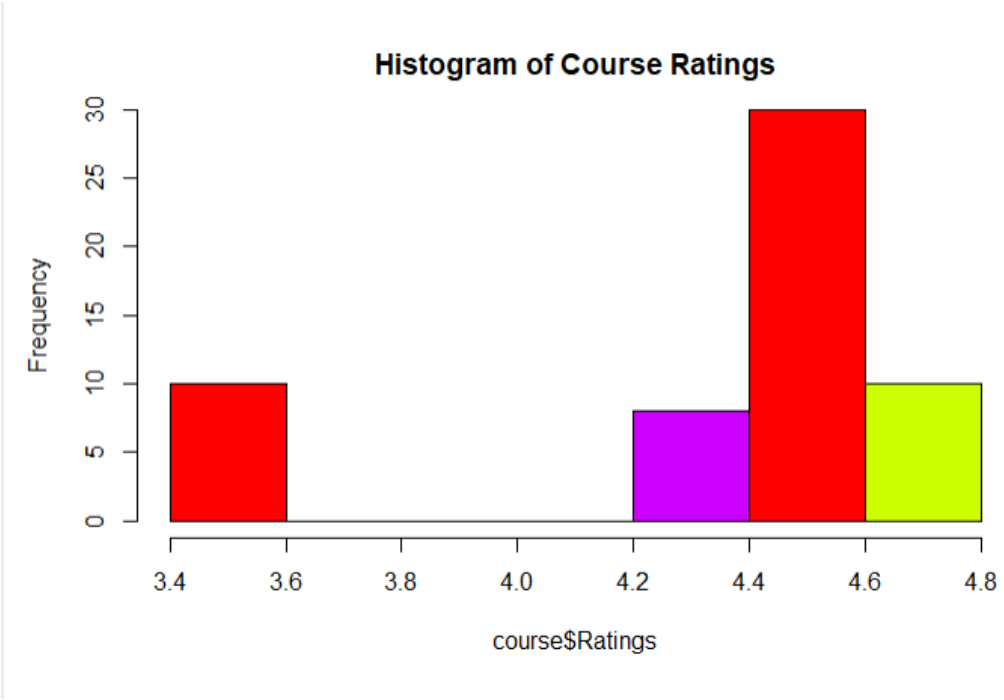
OUTPUT:

	Courses	Duration	Level	Enrollments	Ratings
1	Popular Applications of Data Science	1 hrs	Beginner	2720 people enrolled	4.53
2	Data Science Foundations	1 hrs	Beginner	2720 people enrolled	4.53
3	Career in Data Science	1 hrs	Beginner	2720 people enrolled	4.53
4	Introduction to Data Science	1 hrs	Beginner	2720 people enrolled	4.53
5	R for Data Science	1 hrs	Beginner	2720 people enrolled	4.53
6	Data Science Mathematics	1 hrs	Beginner	2720 people enrolled	4.53
7	Probability for Data Science	1 hrs	Beginner	2720 people enrolled	4.53
8	Statistical Methods for Decision Making	1 hrs	Beginner	2720 people enrolled	4.53
9	Measures of Dispersion	1 hrs	Beginner	2720 people enrolled	4.53
10	Measures of Central Tendency	1 hrs	Beginner	2720 people enrolled	4.53
11	Central Limit Theorem	2 hrs	Beginner	435 people enrolled	4.75
12	Analysis of Variance	2 hrs	Beginner	435 people enrolled	4.75
13	Chi-Square Test	2 hrs	Beginner	435 people enrolled	4.75
14	Data Science with Python	2 hrs	Beginner	435 people enrolled	4.75
15	Data Visualization using Tableau	2 hrs	Beginner	435 people enrolled	4.75
16	Statistics for Data Science	2 hrs	Beginner	435 people enrolled	4.75
17	Data Visualization With Power BI	2 hrs	Beginner	435 people enrolled	4.75
18	Basics of EDA with Python	2 hrs	Beginner	435 people enrolled	4.75
19	Feature Engineering	2 hrs	Beginner	435 people enrolled	4.75
20	Autocorrelation	2 hrs	Beginner	435 people enrolled	4.75
21	Predictive Modeling and Analytics - Regression	3 hrs	Beginner	18076 people enrolled	4.5
22	Intro to Graphic Design with Photoshop	3 hrs	Beginner	18076 people enrolled	4.5

Description: df [6 x 5]

	Courses <chr>	Duration <chr>	Level <chr>	Enrollments <chr>	Ratings <chr>
1	Popular Applications of Data Science	1 hrs	Beginner	2720 people enrolled	4.53
2	Data Science Foundations	1 hrs	Beginner	2720 people enrolled	4.53
3	Career in Data Science	1 hrs	Beginner	2720 people enrolled	4.53
4	Introduction to Data Science	1 hrs	Beginner	2720 people enrolled	4.53
5	R for Data Science	1 hrs	Beginner	2720 people enrolled	4.53
6	Data Science Mathematics	1 hrs	Beginner	2720 people enrolled	4.53

6 rows



INSIGHTS:

From the data set we can find beginner level courses are more when compared to intermediate level. More number of courses contain the word ‘data’ in it as the dataset has been scrapped using the sentence ‘data science’. Many courses were offered 4.4-4.6 rating by the users. Majority of the courses are 2 hours.

From this we can infer there are 50 Beginner and 7 intermediate courses.

From this we can infer there are 50 Beginner and 7 intermediate courses

From this we can infer that many courses is rated from 4.4 to 4.