Department of Information Technology, FAMT Ratnagiri

ASSIGNMENT-I

BE(IT), Sem-VIII(CBCGS) Sub.- R Programming Lab (ITL804)

Extracting data from larger data set and performing exploratory analysis

		Module	Level	CO
1)	List different spreadsheet file formats used for storing the larger data-set. Explain any two in short.	M3	R	LO3
2)	Give package and libraries required for- extracting the data from CSV and excel file format. Also, give a sample codes.	M1	U	LO1
3)	What is exploratory analysis of the data? Explain with a sample code.	M6	U	LO6

.....................X........................

Assignment-I

1. List different spreadsheet file formats used for storing the larger data-set. Explain any two in short.

ANS. Following is a list of common spreadsheet file formats along with their file extensions.

File Extension	File Format
CSV	Comma Separated Values File
DIF	Microsoft Data Interchange Format
ODS	OpenDocument Spreadsheet
OTS	OpenDocument Spreadsheet Template
TSV	Tab Separated Values File
XLM	Microsoft Excel Macro File
XLS	Microsoft Excel Binary File Format
XLSB	Microsoft Excel Binary Spreadsheet File
XLSM	Microsoft Excel Open XML Macro-Enabled Spreadsheet
XLSX	Microsoft Excel Open XML Spreadsheet
XLT	Microsoft Excel Template File
XLTM	Microsoft Excel Open XML Macro-Enabled Spreadsheet Template
XLTX	Microsoft Excel Open XML Spreadsheet Template

1. CSV: Files with CSV (Comma Separated Values) extension represent plain text files that contain records of data with comma separated values. Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently. Almost all spreadsheet applications such as Microsoft Excel or OpenOffice Calc can import CSV without much effort. Data imported from such files is arranged in cells of a spreadsheet for representation to user.

2. XLSX: XLSX is well-known format for Microsoft Excel documents that was introduced by Microsoft with the release of Microsoft Office 2007. Based on structure organized according to the Open Packaging Conventions as of the OOXML standard ECMA-376, the new format is a zip package that contains a number of XML files. The underlying structure and files can be examined by simply unzipping the .xlsx file.

2. Give package and libraries required for- extracting the data from CSV and excel file format. Also, give a sample code.

ANS. Steps to loading and extracting data from csv and excel file format is as follows.

1. Save the CSV/Excel file in the same location as of the script.
2. For CSV file use

```
fr=read.csv("file.csv")
```

where fr is data frame object
          file.csv is the dataset

3. For Excel file use

```
fr=read.xls("file.xlsx")
```

where fr is data frame object
          file.xlsx is the data set

e.g.

```
df=read.csv("https://s3.amazonaws.com/assets.datacamp.com/blog_assets/scores_timed.csv")
print(df)
```

OUTPUT:

```
  X1.6.12.01.03.0        X50.WORST
1 2;16;07:42:51;0          32;BEST
2 3;19;12:01:29;0               50
3 4;13;03:22:50;0 14; INTERMEDIATE
4  5;8;09:30:03;0         40;WORST
```

3. What is exploratory analysis of the data? Explain with a sample code.

ANS. EDA: In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

SAMPLE CODE:

```
basic_eda <- function(data)
{
  glimpse(data)
  df_status(data)
  freq(data)
  profiling_num(data)
  plot_num(data)
  describe(data)
}
```

Glimpse gives the no of observations (rows) and variables and a head of the first cases.
freq function runs for all factor or character variables automatically
profiling_num runs for all numerical/integer variables automatically
Describe is useful to have a quick picture for all the variables.