Finolex Academy of Management and Technology, Ratnagiri

Department of Information Technology

Subject:	**R Programming Lab. (ITL804)**		
Class:	**BE IT / Semester – VIII (Rev-2016) / Academic year: 2019-20**		
Name of Student:	**Kazi Jawwad A Rahim**		
Roll No:	**28**	**Date of performance (DOP) :**	
Assignment/Experiment No:	**06**	**Date of checking (DOC) :**	
Title: Working with larger data-sets and introduction to ggplot2 graphics.			
Marks:		**Teacher's Signature:**	

1. Aim: To understand the exploratory data analysis and the methods required to do it in R.

2. Prerequisites:

1. Data-frames, tables, basic graphical functions.

3. Hardware Requirements:

1. PC with minimum 2GB RAM

4. Software Requirements:

1. Windows / Linux OS.
2. R version 3.6 or higher

5. Learning Objectives:

1. To understand the sources of larger data sets..
2. To understand how the larger data-sets are maintained and managed.
3. To understand the basic usages of ggplot2 graphics package.

6. Learning Objectives Applicable: LO 3, LO 5

7. Program Outcomes Applicable: PO 4, PO 5

8. Program Education Objectives Applicable: PEO 4, PEO 6

Theory:

Working with larger datasets:

R reads entire dataset into RAM all at once. Other programs can read file sections on demand. R object live in memory, entirely. We can categorize large data sets in 2 broad categories -

1) Medium sized:

Medium sized files can be load in R within memory limit (typically in the 1-2 GB range)

2) Larged sized:

Large files can't be loaded in R due to R's limitations.

If you are loading csv or text file, you can't select specific columns. You might want to preprocess the data in command line using commands of alter data.

10. Results:

```
setwd("f:/exp6")

fr = read.csv("data.csv")

print(fr)
```

```
>>>    Sr.  Name Age Gender Marks
1   1 Jawwad 21     M    80
2   2  Sahil 22     M    82
3   3 Aniket 22     M    84
4   4  Sagar 22     M    86
```

mode(fr)
[1] "list"

class(fr)
[1] "data.frame"

fr$Name
[1] Jawwad Sahil Aniket Sagar

fr$Age
[1] 21 22 22 22

fr$Marks
[1] 80 82 84 86

mode(fr$Name)
[1] "numeric"

class(fr$Name)
[1] "factor"

After adding **"header = FALSE"** as a parameter in **read.csv(....)**, we got

```
  V1    V2 V3    V4    V5
1 Sr.  Name Age Gender Marks
2  1 Jawwad 21     M    80
3  2  Sahil 22     M    82
4  3 Aniket 22     M    84
5  4  Sagar 22     M    86
```

Now, I'm using a large data set **"lendingdata.csv"** of about 15 columns and 27518 rows.

fr = read.csv("lendingdata.csv")

ncol(fr)
[1] 15

nrow(fr)
[1] 27518

Now, I'm listing one of the columns data as follows

```
fr$country
  [1] Cambodia        Philippines
  [3] Peru            Tajikistan
  [5] Uganda          Jordan
  [7] Tajikistan      Cambodia
  [9] Nicaragua       Nigeria
 [11] Colombia        Nicaragua
 [13] Colombia        Philippines
 [15] Ecuador         Colombia
```

And so on

mode(fr$country)
[1] "numeric"

class(fr$country)
[1] "factor"

Now, for demonstrating the GGPlot, first we need to install the **ggplot2** package as

install.packages(ggplot2)

After successfully installing the ggplot2 package and its dependencies, I'm ready to demonstrate.

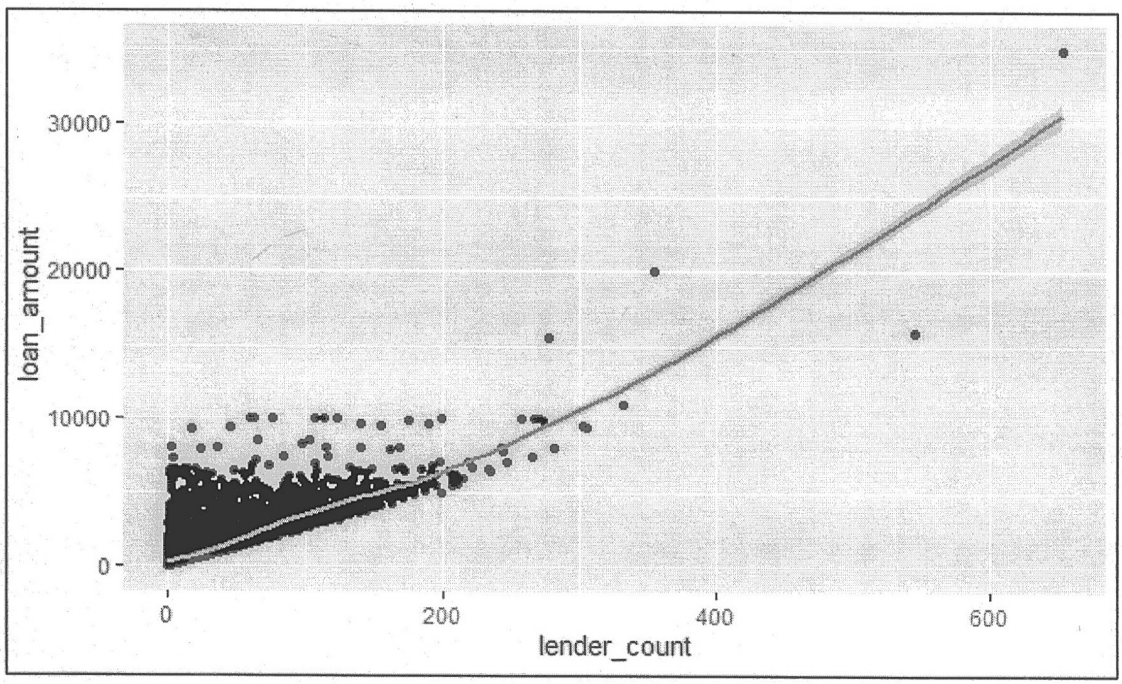
Source code:

```
library(ggplot2)

setwd("f:/exp6")

fr = read.csv("lendingdata.csv")

ggplot(fr,aes(x=lender_count,y=loan_amount))+geom_point()+geom_smooth()
```



11. Learning Outcomes Achieved:

1. We understood the sources of larger data sets.
2. We understood how the larger data-sets are maintained and managed.
3. We understood the basic usages of ggplot2 graphics package.

12. Conclusion:

We understood the exploratory data analysis and the methods required to do it in R. Also, we have done operations on larger data sets and performed GGplot of the data set to analyze the relativity of the data.

13. Experiment/Assignment Evaluation

Experiment/Assignment Evaluation:				
Sr. No.	**Parameters**		**Marks obtained**	**Out of**
1	Technical Understanding (Assessment may be done based on Q & A or any other relevant method.) Teacher should mention the other method used -			6
2	Neatness/presentation			2
3	Punctuality			2
Date of performance (DOP)		**Total marks obtained**		10
Date of checking (DOC)		**Signature of teacher**		

References:

1. URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf (Online Resources)
2. R Cookbook Paperback – 2011 by Teetor Paul O Reilly Publications
3. Beginning R: The Statistical Programming Language by Dr. Mark Gardener, Wiley Publications
4. R Programming For Dummies by Joris Meys Andrie de Vries, Wiley Publications

Viva Questions

1. What are different ways to store larger data-set?
2. What are names of packages required to extract data from data-set stored in standard spreadsheet.
3. What are various plotting functions in ggplot2?