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:	**07**	**Date of checking (DOC) :**	
Title: Program to demonstrate regression and correlation in tabular data including categorical data.			
Marks:		**Teacher's Signature:**	

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

2. Prerequisites:

1. Working with larger data-sets.

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 basic elements of larger data-sets.
2. To understand numerical and categorical variables in larger data-sets.
3. To understand how to apply regression to design decision model on the larger data-sets.

6. Learning Objectives Applicable: LO 5, LO 6

7. Program Outcomes Applicable: PO 4, PO 5

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

Theory:

Regression:

Linear regression specifies one variable as the independent variable and another as dependent variable. The resultant modes relates the variable with a linear regression, are parametric and assume normality, homoscedacity and independence of residuals.

Correlation -

Correlation determines of one variable varies systematically as another variable and another variable is independent variable. It is useful to look at which variables are correlated to others in a dataset.

10. Results:

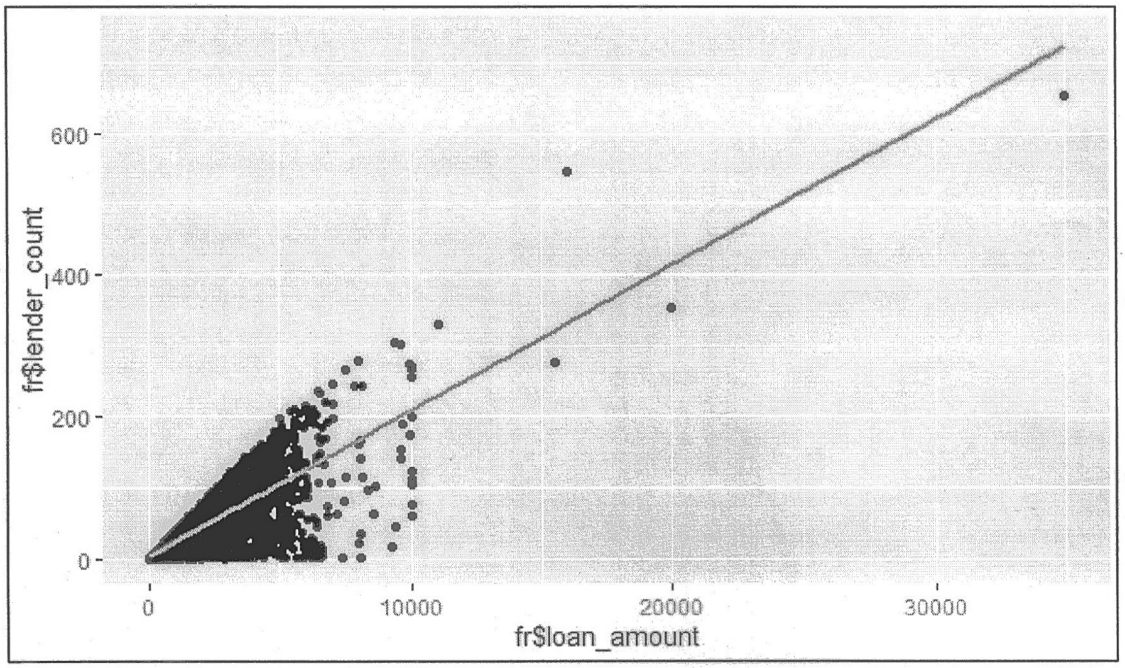
Here we have considered a large data set *"lendingdata.csv"* of 15 columns and 27518 rows.

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

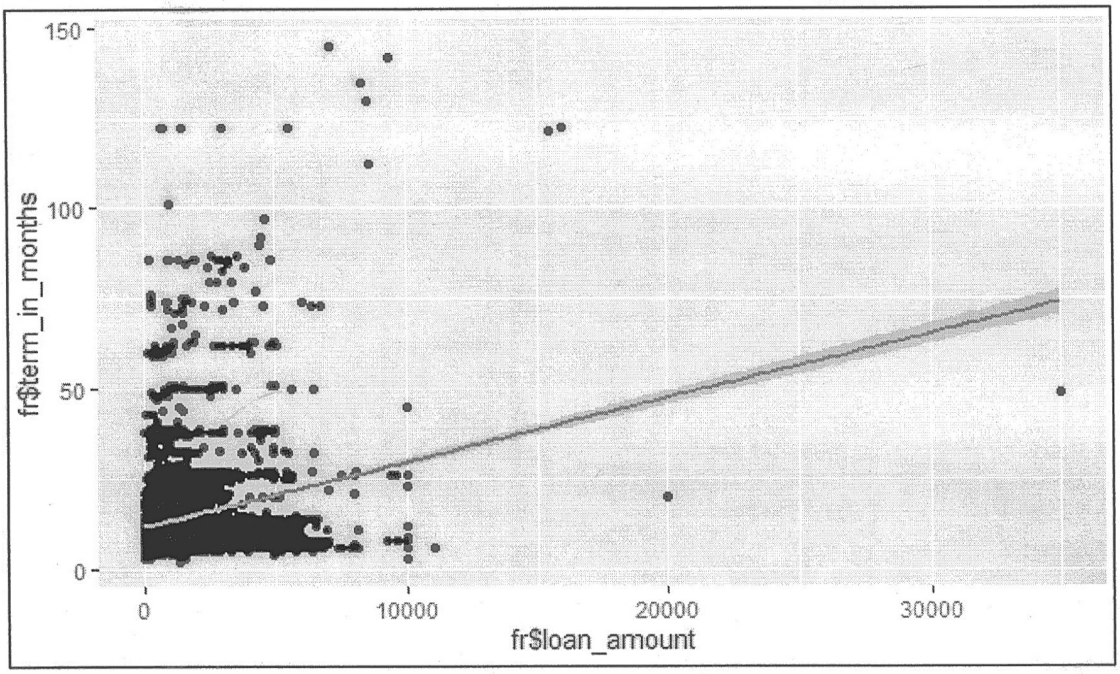
We are now considering three columns namely *loan_amount*, *lender_count* and *term_in_months*.

We will now plot regression line for above mentioned columns in pair of any two columns.

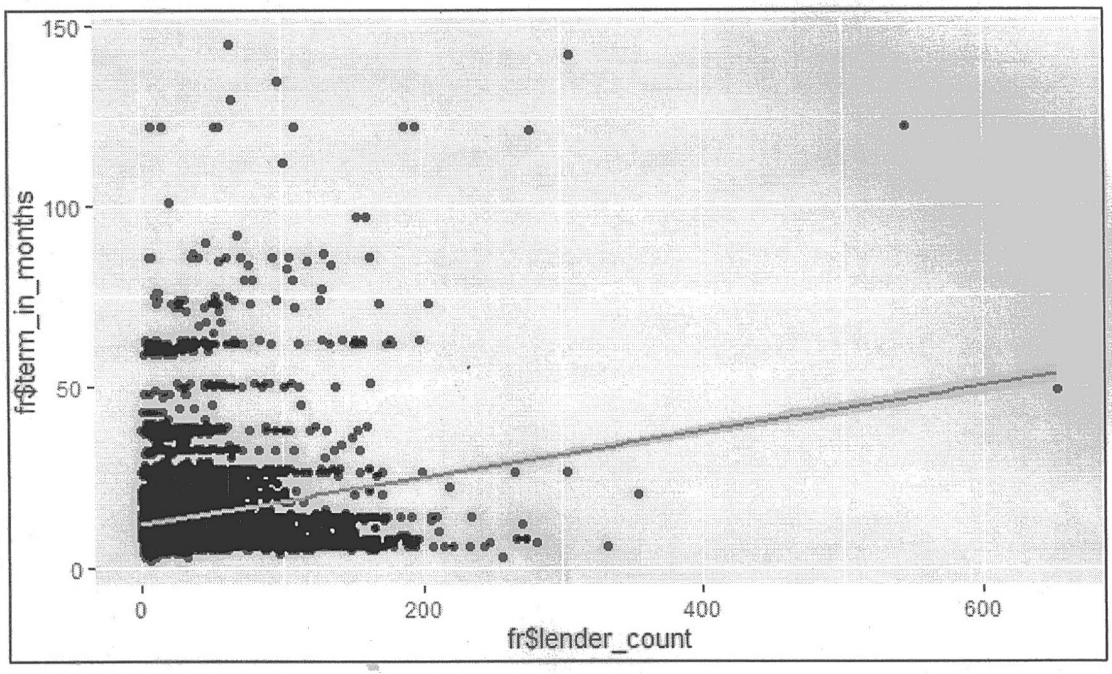
ggplot(fr,aes(x=fr$loan_amount,y=fr$lender_count))+geom_point()+geom_smooth(method=l
m,formula=y~x)



```
ggplot(fr,aes(x=fr$loan_amount,y=fr$term_in_months))+geom_point()+geom_smooth(method
=lm,formula=y~x)
```




```
ggplot(fr,aes(x=fr$lender_count,y=fr$term_in_months))+geom_point()+geom_smooth(method
=lm,formula=y~x)
```

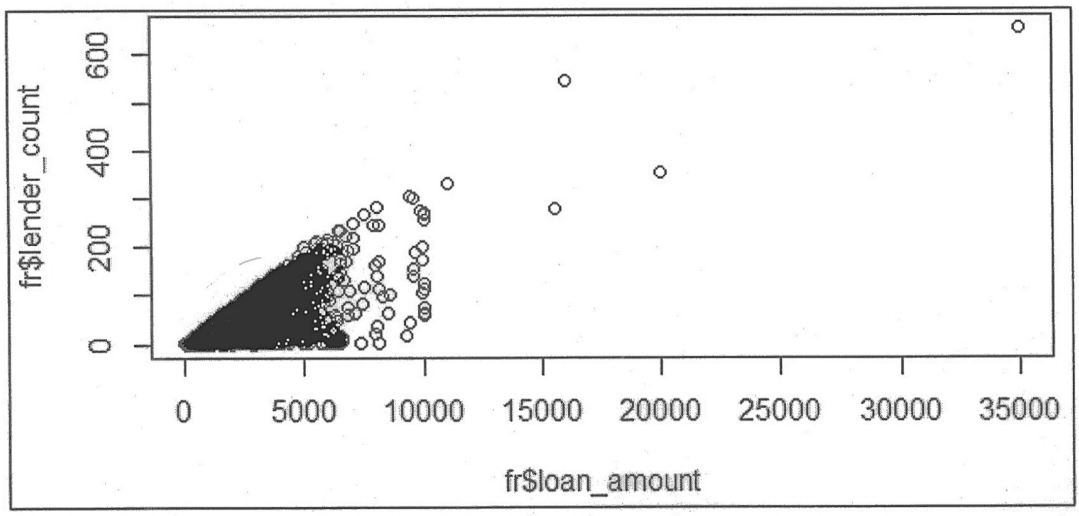


Following are the correlations and their visualization.

cor(fr$loan_amount,fr$lender_count)

>>>0.8151209

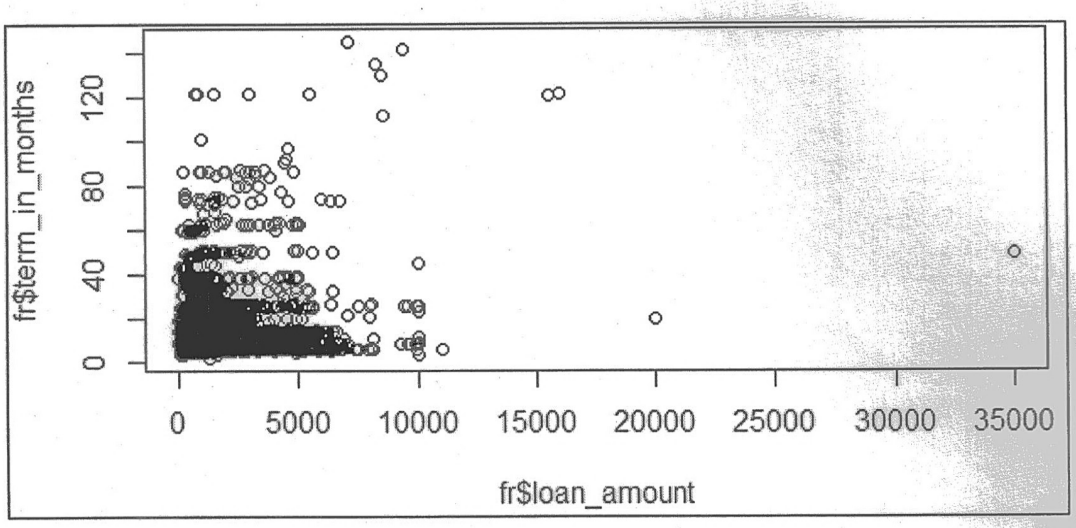
plot(fr$loan_amount,fr$lender_count)



cor(fr$loan_amount,fr$term_in_months)

>>>0.2063649

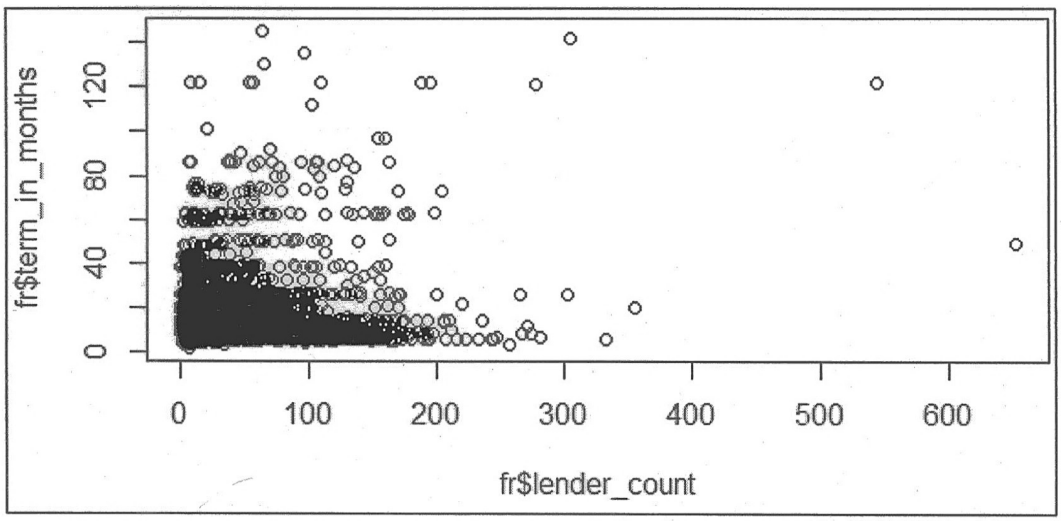
plot(fr$loan_amount,fr$term_in_months)



cor(fr$lender_count,fr$term_in_months)

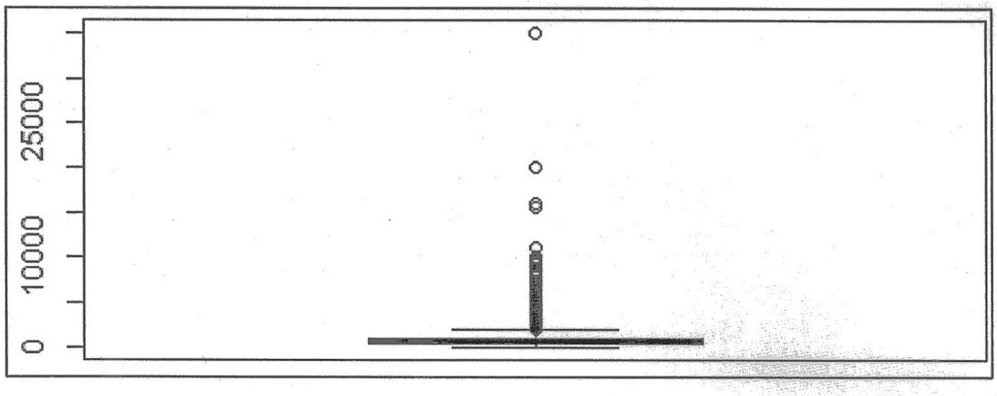
>>>0.1846157

plot(fr$lender_count,fr$term_in_months)

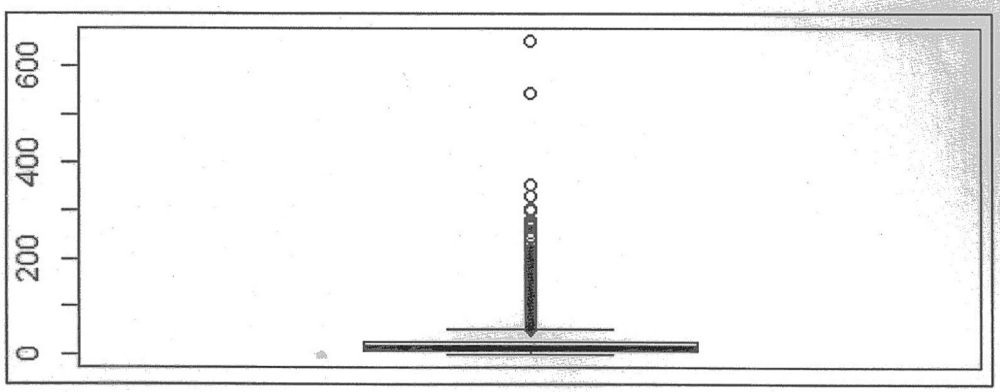


Now we will visualize correlation of categorical variable with a numeric variable using Boxplot for above mentioned three columns.

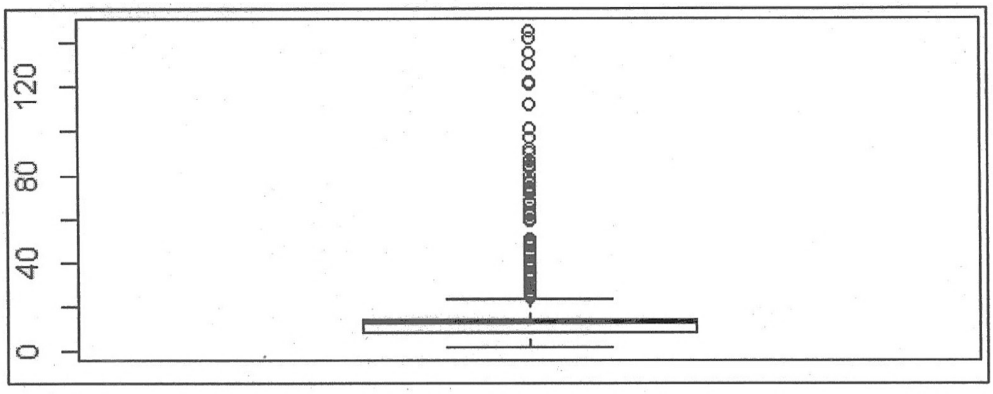
boxplot(fr$loan_amount)



boxplot(fr$lender_count)

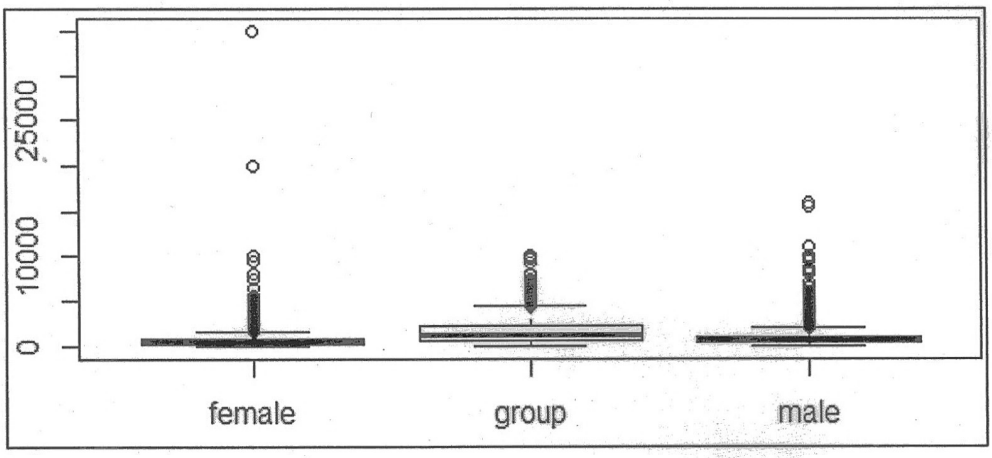


boxplot(fr$term_in_months)

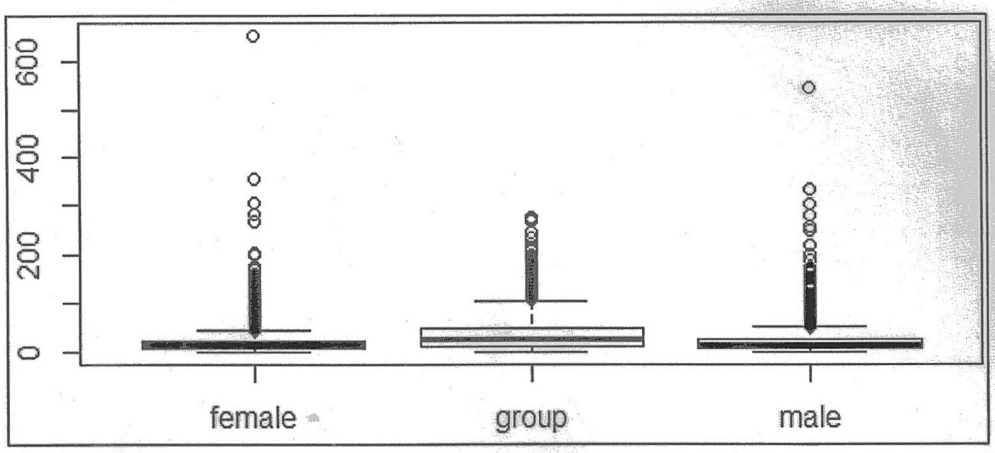


Now let's apply a function by splitting the *loan_amount*, *lender_count* and *term_in_months* as per the genders each, it will display multiple boxplots for different possible genders.

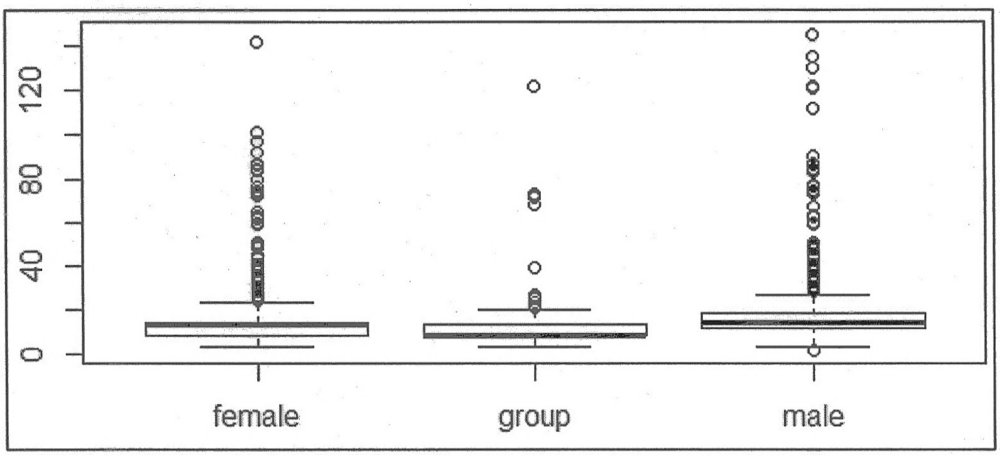
boxplot(split(fr$loan_amount,fr$borrower_genders))



boxplot(split(fr$lender_count,fr$borrower_genders))

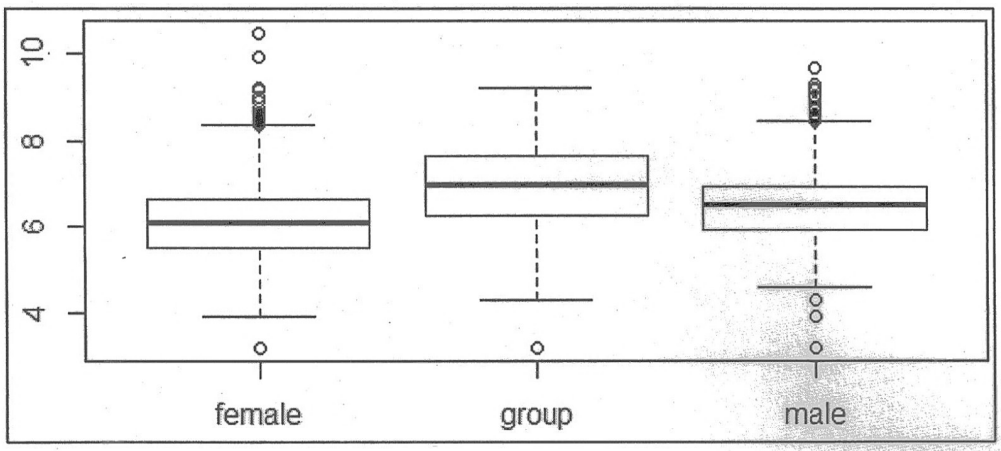


boxplot(split(fr$term_in_months,fr$borrower_genders))

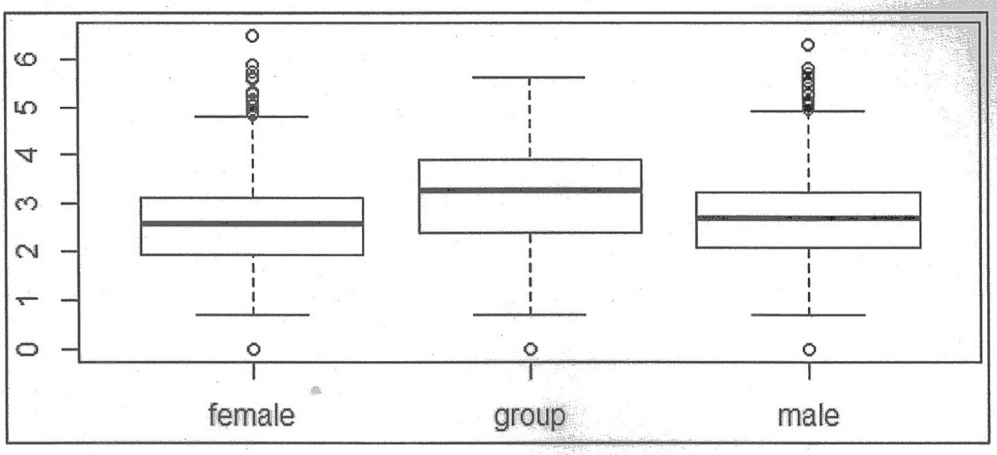


We can take log of *loan_amount, lender_count* and *term_in_months* each to have a broader view.

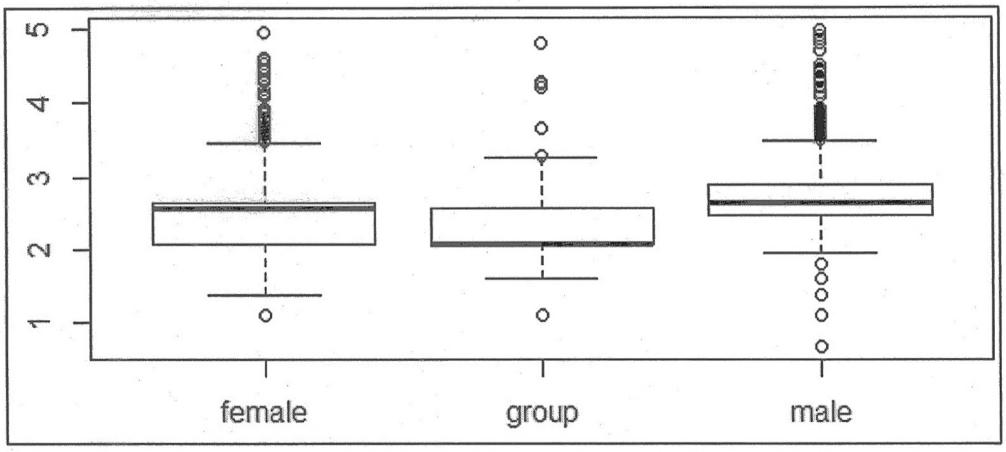
boxplot(split(log(fr$loan_amount),fr$borrower_genders))



boxplot(split(log(fr$lender_count),fr$borrower_genders))



boxplot(split(log(fr$term_in_months),fr$borrower_genders))



11. Learning Outcomes Achieved:

1. We understood the basic elements of larger data-sets.
2. We understood numerical and categorical variables in larger data-sets.
3. We understood how to apply regression to design decision model on the larger data-sets.

12. Conclusion:

We have successfully demonstrated the exploratory data analysis and the methods required to do it in R. Also, we have plotted the regression line, correlations between columns and boxplots.

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 does it mean by categorical variables in data-sets?
2. What does it mean by regression?
3. What is correlation and how is it useful in data-science?