Experiment- 7

Program to demonstrate regression and correlation in tabular data including categorical data

- Regression is used to plot a best fit line given a plot of one variable v/s other variable.
- The regression line can be plotted on the given plot using linear model as follows-
- **plot(x,y)** *# plot x vs y*
  abline(lm(y~x)) *# plot as per the linear model obtained between relation x & y*
  OR
  ggplot(fr,aes(x=*col1*,y=*col2*))+geom_point()+geom_smooth(method=lm,formula=y~x) *# col1 & col2 are required column names of the frame fr*

- Given a data frame **fr** containing a data in the **csv** file **"lendingdata.csv"**
  Here the regression line can be plotted between numeric variables of the frame, i.e. **fr$loan_amount~fr$lender_count** or **fr$loan_amount~fr$term_in_months**
- Students are advised to revise the concepts of correlation they student in mathematics, or revise from the URL:
  https://en.wikipedia.org/wiki/Correlation_and_dependence
- Understand what does it mean by ***highly correlated***, and ***uncorrelated*** pair of variables.
- The function **cor(x,y)** is used to find correlation between variables **x** and **y**, where **x** and **y** are lists of numeric data.
- Find the correlation values between possible two variables among **fr$loan_amount, fr$lender_count, fr$term_in_months,** and at the same time display corresponding **plot**s between these two variables. Try to understand how the correlation is related to the relation between variables.
- To visualize correlation of categorical variable with a numeric variable, box-plot is the best way. For instance, if we use **boxplot(fr$loan_amount)**, it will display

boxplot showing ranges of **fr$loan_amount**.. However we may apply the said function by splitting the loan amount as per the genders, it will display multiple boxplots for different possible genders. Note here, the borrower's gender is a categorical variable of the data-frame. Here is the way-

- **boxplot(split(log(fr$loan_amount),fr$borrower_genders))**
- If it unable to display the box-plot properly due to wide spread in the *loan-amount*, you may take its log before using it in the **boxplot(…)** function.
