Assignment-II

1. Write a code to read the larger data-set contains in the file at http://famt.ac.in/eResource/it/lendingdata.csv.

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

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

OUTPUT:

```
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

2. What is data cleaning? Explain in detail.

ANS. Data Cleaning is the process of transforming raw data into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability. Data cleaning may profoundly influence the statistical statements based on the data. R has a set of comprehensive tools that are specifically designed to clean data in an effective and comprehensive manner. It mainly has three steps as follows-

STEP 1: Initial Exploratory Analysis: The first step to the overall data cleaning process involves an initial exploration of the data frame that you have just imported into R. It is very important to understand how you can import data into R and save it as a data frame.

STEP 2: Visual Exploratory Analysis: There are 2 types of plots that you should use during your cleaning process –The Histogram and the BoxPlot

Histogram: The histogram is very useful in visualizing the overall distribution of a numeric column. We can determine if the distribution of data is normal or bi-modal or unimodal or any other kind of distribution of interest. We can also use Histograms to figure out if there are outliers in the particular numerical column under study.

BoxPlot: Boxplots are super useful because it shows you the median, along with the first, second and third quartiles. BoxPlots are the best way of spotting outliers in your data frame.

STEP 3: Correcting the errors: This step focuses on the methods that you can use to correct all the errors that you have seen.

3. List various regression models used in statistics for estimating the result.

ANS. List of various regression models-

1. Histogram
2. Boxplot
3. Correlation
4. Plot
5. GGPlot