Name : Prajakta Kaur
Roll No : 33231

ASSIGNMENT 5

Aim : Perform different operations using R/Python.

Problem Statement : Perform the following operations using R/Python on the Amazon Book Review and Facebook metric data sets

a) Create data subsets b) Merge Data c) Sort Data
d) Transposing Data e) Melting data to long format
f) Casting data to wide format

Objectives :

- To learn R/Python programming
- To learn different data preprocessing techniques.

Theory : R Programming Language : Developed by Ross Ihaka and Robert Gentleman in 1993. R possess an extensive catalog of statistical and graphical methods. It include ML algorithms, linear regression, time series, statistical reference, etc.

Dataframes : It is a table or a 2D array-like structure in which each column contains values of one variable and each row contains one set of variables from each column.

Characteristics Of DF : 1) Column names should not be empty
2) Row names should be unique
3) The data stored can be of numeric, factor or character type. 4) Each col. should contain same no. of data items.

Reading CSV files :

```
read.csv2 (file, header = True, sep = ";", dec = ",", ...)
```

- file : path to the file containing data to be imported
- sep : the field seperator character
- dec : the character used in the file for decimal points.

Eg : `myData = read.csv2 ("basic.csv")`

Subetting Data : 1) Selecting variables

```
sub1 = data[c("v1", "v2", "v3")]
```

2) Using subset function

```
sub2 = subset(data, age >= 20 | age < 10)
```

Merge Data : 1) Adding columns

```
total = merge(dataA, dataB, by = "ID").
```

2) Adding Rows

```
total = rbind(dataA, dataB)
```

Sort Data : #sort by mpg (ascending) and cyl (descending)

```
newdata = cars[order(mpg, -cyl)]
```

Transpose Data :

```
trasddta = t(data)
```

Melting Data : We have to use the reshape package

eg:

```
library(reshape)
mdata = melt(mydata, id = c("id", "time"))
```

Melting is done to organize the data. Using melt(), dataframe is converted into long format and stretches the data frame.

Casting Data : cast() is used to convert long format data back to some aggregated form.

eg:

```
subjmeans = cast(mdata, id~variable, mean)
```

Conclusion : I learnt various preprocessing techniques in R programming language and performed operations like sorting, merging, melting and casting on the datasets.