NAME:Aditya Somani CLASS:TE-A ROLL NO:T1851061

ASSIGNMENT-1

Part B: Assignments based on R and Python

Aim:

Perform the following operations using R/Python on the Amazon book review and facebook metrics data sets
1) Create data subsets
2) Merge Data
3) Sort Data
4) Transposing Data
5) Melting Data to long format
6) Casting data to wide format

Introduction

What is R?

- R is a programming language and software environment for statistical analysis, graphics representation and reporting.

- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

- R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

- This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions.

- R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

- R is free software distributed under a GNU-style copy left, and an official part of the GNU project called GNU S.

Evolution of R

- R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. R made its first appearance in 1993. – A large group of individuals has contributed to R by sending code and bug reports. – Since mid1997 there has been a core group (the "R Core Team") who can modify the R source code archive.

Features of R

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities. • R has an effective data handling and storage facility, • R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

R Studio

- RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.
- RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio.
- RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server.
- Prepackaged distributions of RStudio Desktop are available for Windows, OS X, and Linux.

Download R Studio

- Windows: –https://download1.rstudio.org/RStudio-0.99.893.exe
- Ubuntu: –https://download1.rstudio.org/rstudio-0.99.893-i386.deb
- Fedora: –https://download1.rstudio.org/rstudio-0.99.893-i686.rpm
- Linux flavors differentiates 32bit and 64bit as well as .deb and .rpm packages.

Python

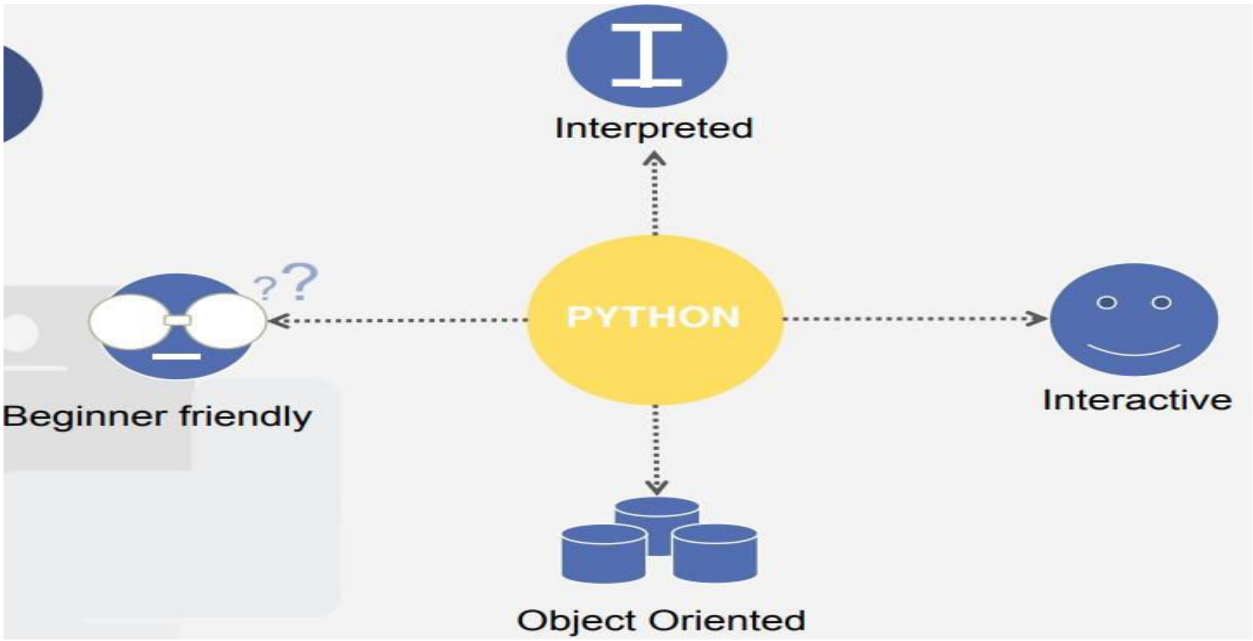
Python is aninterpretedhigh-level programming languageforgeneral-purpose programming. Created byGuido van Rossumand first released in 1991, Python has a design philosophy that emphasizescode readability, and asyntaxthat allows programmers to express concepts in fewerlines of code,[25][26]notably usingsignificant whitespace. It provides constructs that enable clear programming on both small and large scales.

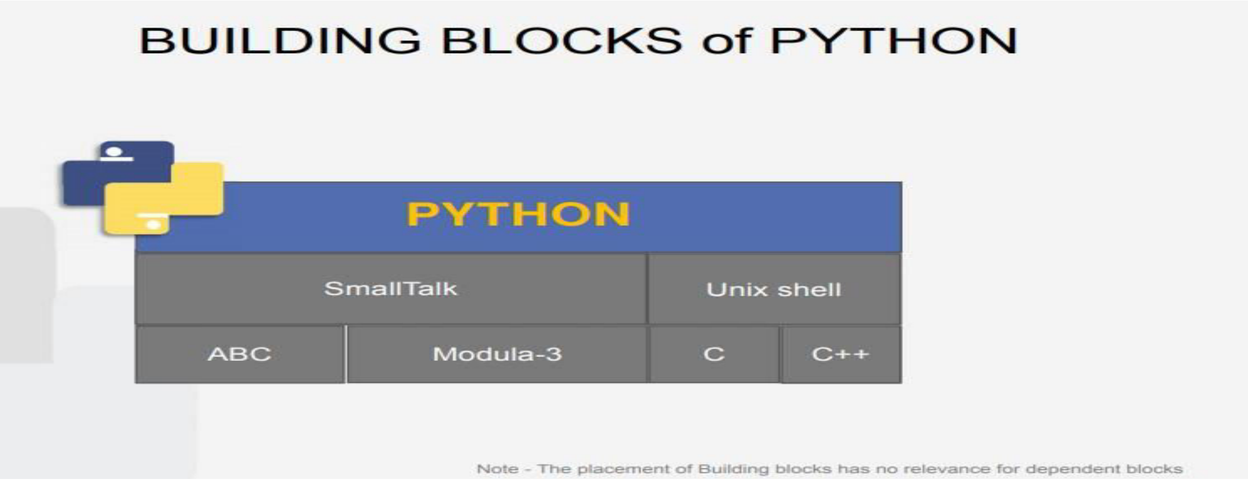
Python Features

Python's features include −

- **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

- **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
- **A broad standard library** − Python's bulk of the library is very portable and crossplatform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases**− Python provides interfaces to all major commercial databases.
- **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** − Python provides a better structure and support for large programs than shell scripting.





BUILDING BLOCKS of PYTHON

PYTHON

SmallTalk

Unix shell

ABC

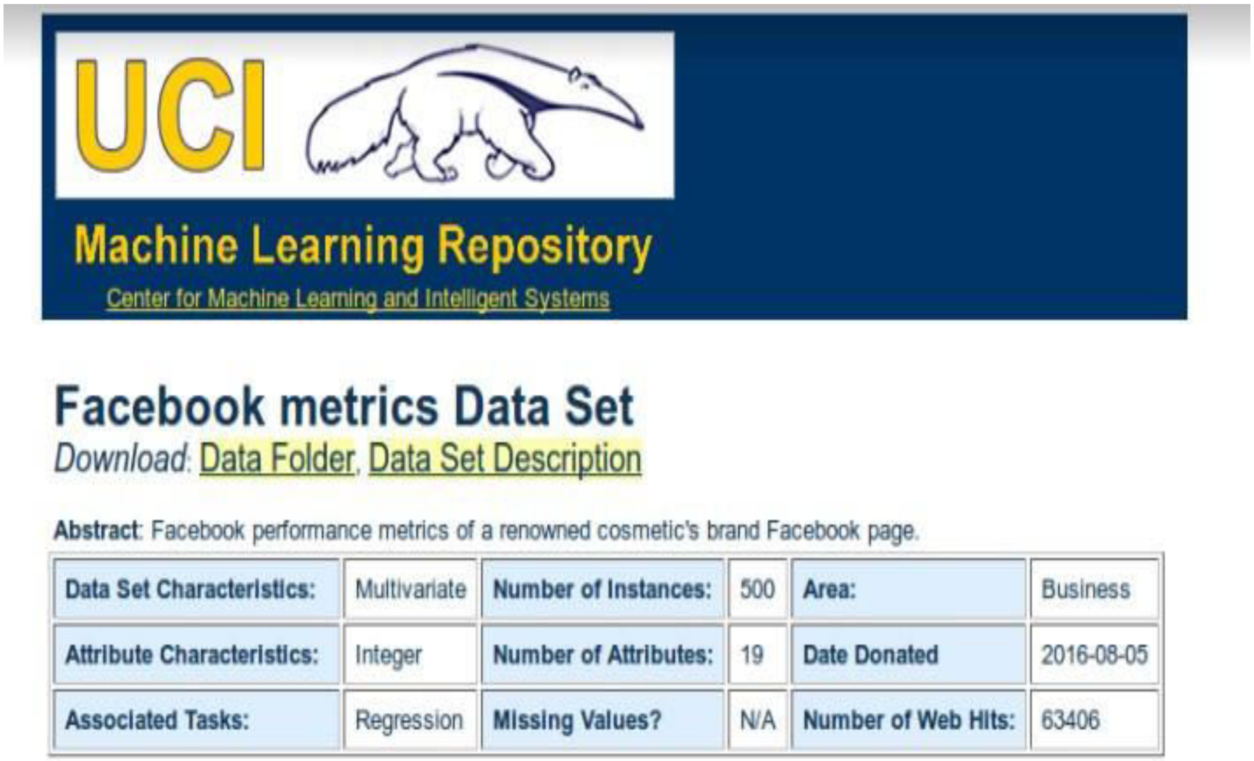
Modula-3

C

C++

Note - The placement of Building blocks has no relevance for dependent blocks

Assignment Details 1.Download Datasets



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

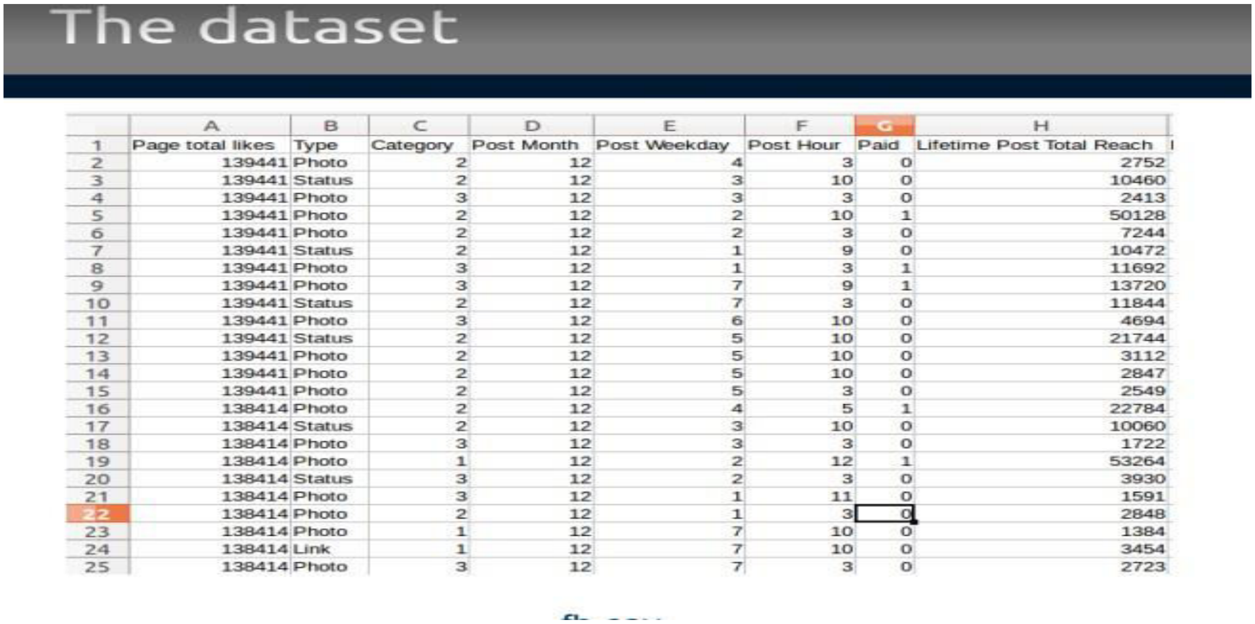
Facebook metrics Data Set

Download: Data Folder, Data Set Description

Abstract: Facebook performance metrics of a renowned cosmetic's brand Facebook page.

Data Set Characteristics:	Multivariate	**Number of Instances:**	500	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	19	**Date Donated**	2016-08-05
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	63406

2.The Dataset:

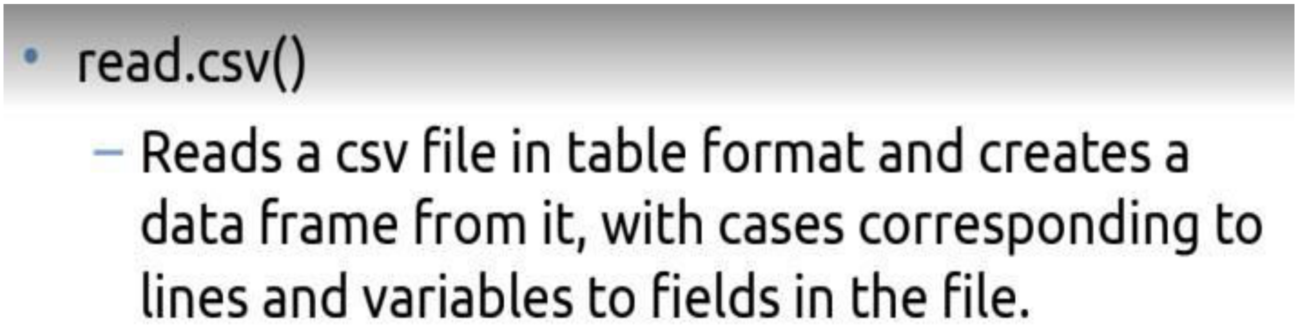


The dataset

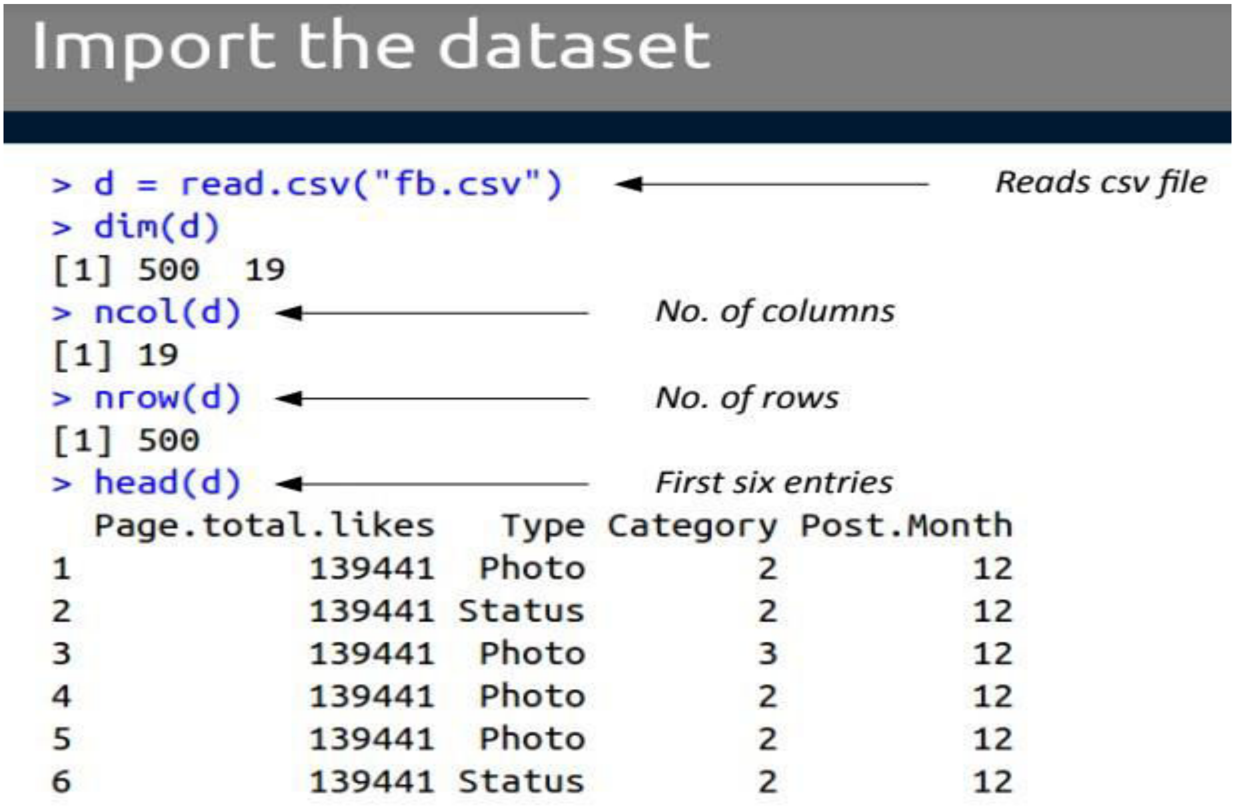
	A	B	C	D	E	F	G	H
1	Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post Total Reach
2	139441	Photo	2	12	4	3	0	2752
3	139441	Status	2	12	3	10	0	10460
4	139441	Photo	3	12	3	3	0	2413
5	139441	Photo	2	12	2	10	1	50128
6	139441	Photo	2	12	2	3	0	7244
7	139441	Status	2	12	1	9	0	10472
8	139441	Photo	3	12	1	3	1	11692
9	139441	Photo	3	12	7	9	1	13720
10	139441	Status	2	12	7	3	0	11844
11	139441	Photo	3	12	6	10	0	4694
12	139441	Status	2	12	5	10	0	21744
13	139441	Photo	2	12	5	10	0	3112
14	139441	Photo	2	12	5	10	0	2847
15	139441	Photo	2	12	5	3	0	2549
16	138414	Photo	2	12	4	5	1	22784
17	138414	Status	2	12	3	10	0	10060
18	138414	Photo	3	12	3	3	0	1722
19	138414	Photo	1	12	2	12	1	53264
20	138414	Status	3	12	2	3	0	3930
21	138414	Photo	3	12	1	11	0	1591
22	138414	Photo	2	12	1	3	0	2848
23	138414	Photo	1	12	7	10	0	1384
24	138414	Link	1	12	7	10	0	3454
25	138414	Photo	3	12	7	3	0	2723

fb.csv

3. Read the Downloaded CSV File



- read.csv()
  - Reads a csv file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.



Import the dataset

```
> d = read.csv("fb.csv")
> dim(d)
[1] 500  19
> ncol(d)
[1] 19
> nrow(d)
[1] 500
> head(d)
  Page.total.likes   Type Category Post.Month
1           139441  Photo        2         12
2           139441 Status        2         12
3           139441  Photo        3         12
4           139441  Photo        2         12
5           139441  Photo        2         12
6           139441 Status        2         12
```

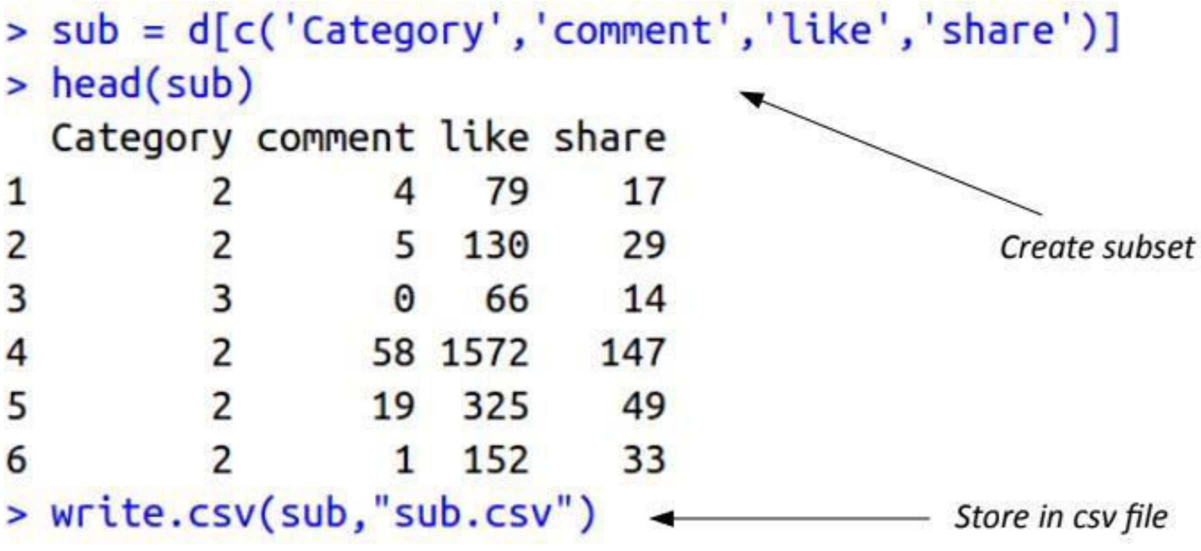
Reads csv file

No. of columns

No. of rows

First six entries

3.Create Subset

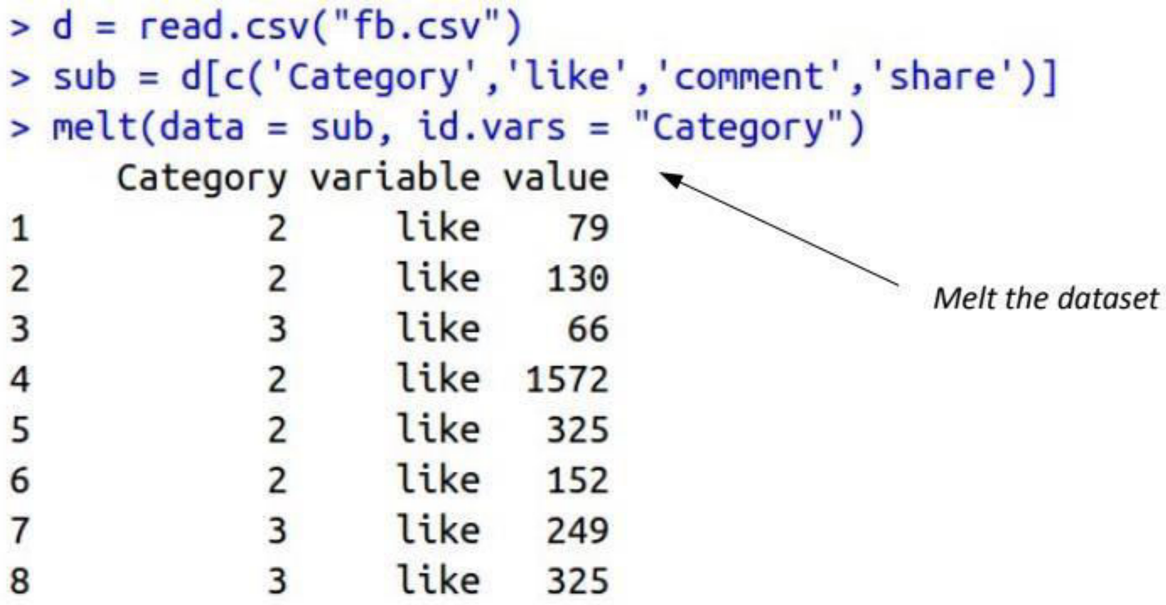


```
> sub = d[c('Category','comment','like','share')]
> head(sub)
  Category comment like share
1        2       4   79    17
2        2       5  130    29
3        3       0   66    14
4        2      58 1572   147
5        2      19  325    49
6        2       1  152    33
> write.csv(sub,"sub.csv")
```

Create subset

Store in csv file

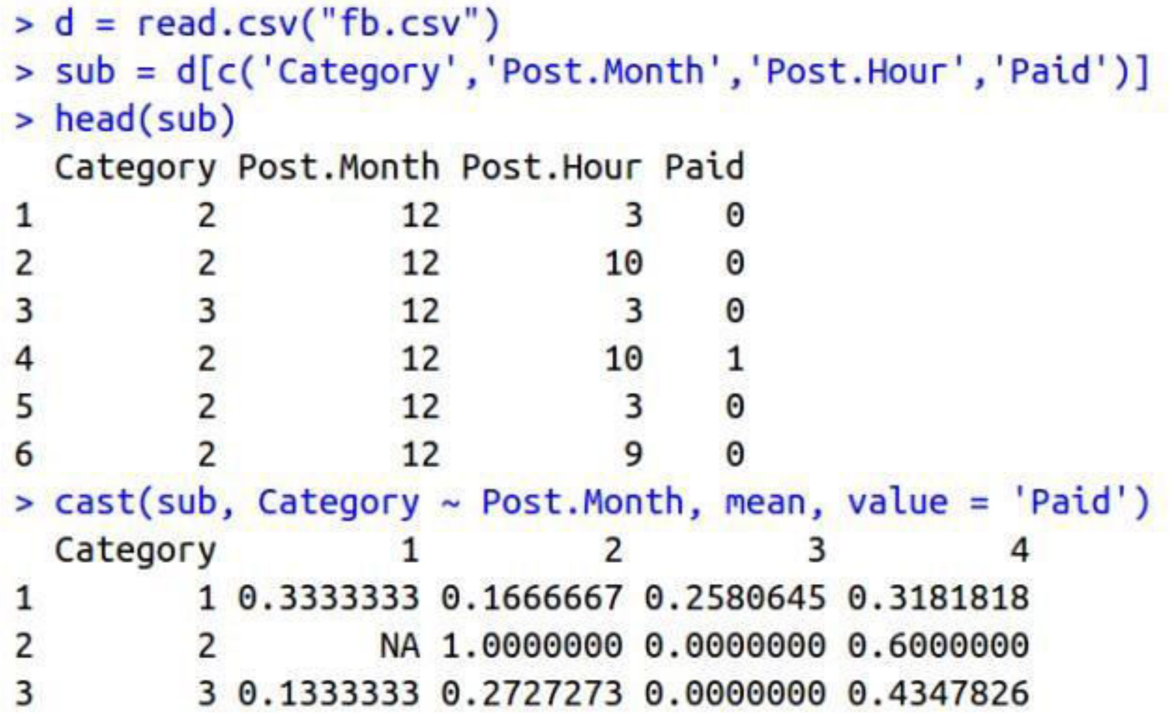
4.Melt Dataset



```
> d = read.csv("fb.csv")
> sub = d[c('Category','like','comment','share')]
> melt(data = sub, id.vars = "Category")
     Category variable value
1           2     like    79
2           2     like   130
3           3     like    66
4           2     like  1572
5           2     like   325
6           2     like   152
7           3     like   249
8           3     like   325
```

Melt the dataset

5.Casting Dataset



```
> d = read.csv("fb.csv")
> sub = d[c('Category','Post.Month','Post.Hour','Paid')]
> head(sub)
  Category Post.Month Post.Hour Paid
1        2         12         3    0
2        2         12        10    0
3        3         12         3    0
4        2         12        10    1
5        2         12         3    0
6        2         12         9    0
> cast(sub, Category ~ Post.Month, mean, value = 'Paid')
  Category         1         2         3         4
1        1 0.3333333 0.1666667 0.2580645 0.3181818
2        2        NA 1.0000000 0.0000000 0.6000000
3        3 0.1333333 0.2727273 0.0000000 0.4347826
```

Conclusion: Thus we have learnt various operations of (Creating data subsets, Merge Data, Sort Data, Transposing Data, Melting Data to long format, Casting data to wide format)with **R Language in RStudio**.