WhatsApp Group Data Analysis with R

Sanchita Patil
MCA Department
Vivekanand Education
Society's Institute of Technology
Chembur, Mumbai – 400074.

ABSTRACT
The means of communication has changed over time according to the situation and advancements in technology. The process of transferring data from one individual to another such as audio, video and images have grown beyond texting and evolved to enable the transmission of media not only between two individuals but also in a group where huge number of people can interact and have a talent to connect worldwide. WhatsApp is such an application which is used widely for transferring media, text, files as well as audio calling. This research paper predicts the level of addiction of an individual to the WhatsApp group as per the age group and gender with the help of R statistics software programme.

Keywords
Data Analysis, R Programming, Visualization, WhatsApp.

1. INTRODUCTION
In the current time, due to various factors such as ease of use, essential features, the usage of WhatsApp has accelerated. In 2009, WhatsApp was founded by Brian Action and Jon Koum [1]. WhatsApp's user base had increased to about 210 million active users by February 2013, due to the user's ability to interact with others through audio calling, texting, and transferring media as well as group chat [1]. The objective of the paper is to classify the number of users as those addicted and not addicted to WhatsApp group chat and thus predicting the level of addiction. In particular, this paper mainly emphasizes on the usage of R statistical software programme, and how it can be used to extract and work with a particular dataset. R is an open-source data analysis environment and programming language [2].It has a wide user base in academia specifically and is also supported by email and web groups.

1.1 Data analysis
Process of cleaning, transforming, inspecting and modelling data with the goal of uncovering useful information, indicating conclusions, and thus supporting decision-making is Data Analysis [4]. It has multiple facts and approaches encompassing multiple techniques under a variety of names in disparate business, science, and social science domains. Analysis refers to breaking a whole component into its separate components for individual examination. Data analysis is a process for acquiring raw data and transforming it into information useful for decision-making by users. Data is collected and analysed for testing hypothesis or answering the questions. Statistician John Tukey defined data analysis in 1961 as: "Procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analysing data [4]. Phases of data analysis are iterative and thus feedback from later phases may result in additional work in earlier phases.

This Research provides the basic idea of Statistical analysis done on a particular WhatsApp group data. Following are the sections in which research has been carried:

- To find what type of communication medium people prefer the most in WhatsApp group chat.
- To find most active day of week.
- To find which age group participants are more active on WhatsApp group and number of messages send by each age group participants per month, day, hour.
- To find whether Males are more addicted to the WhatsApp group or Females.
- Total number of messages send as per Timestamp.

2. LITERATURE SURVEY
The dataset of WhatsApp Group chat used for analysis is of 1 year (May, 2015 –May, 2016) which consists of 5,563 records in total and comprises of certain characteristics that define how much a particular person is using WhatsApp Chat Group, such as the years of usage, duration of usage in a day, the response levels, type of messages posted by each individual in the group (Smiley, Text, Multimedia), which age group people are more active and so on. The main attributes set for this analysis are type of messages been send, duration of use per year/month/week/day /hour, timestamp (AM/PM), age group of sender, gender (Male/Female). RStudio the most favoured IDE for R is been used to perform exploratory data analysis and visualization for the collected data largely because of its open source nature.

3. SYSTEM ARCHITECTURE

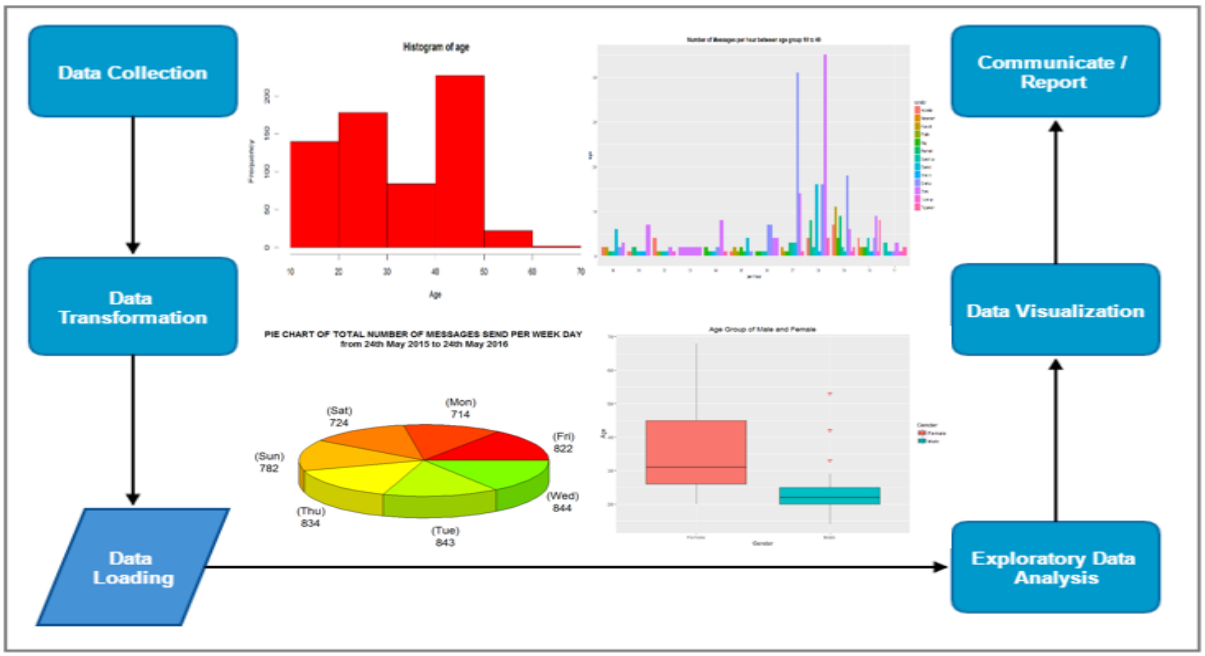

Fig 1: Overview of the stages in Data Analysis

Data Analysis process includes Data Collection, Data Transformation, Data Loading, Exploratory Data Analysis, Data Visualization and Communicate/Report. Fig 1 displays the process of Data Analysis.

4. METHODOLOGY

Now let's have a glance on different stages and procedures in retrieval of insightful results, gathering relevant data, importing it into RStudio and finally analysing it.

4.1 Data Collection

Data Collection is the first stage of the model which includes idea, defining project objective, setting up machine and lastly knowing your data. Fig 2 illustrates these processes.

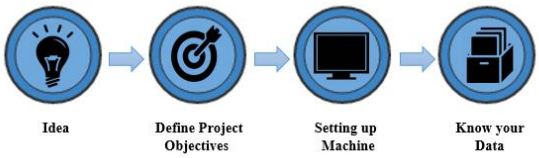

Fig 2: Processes involved in Data Collection

Accurate data collection is essential in order to ensure the integrity of research. The purpose of collecting data is to answer questions in which the answers are not immediately obvious and thus helps in decision making [5].

4.1.1 Idea

The person must have absolute idea withregards to the actual source of data, methods of extraction and usefulness of data. In this case the data been extracted is historic data that holds the key to understand data over time. A copy of the history of a group chat is been extracted, using the Email chat feature:

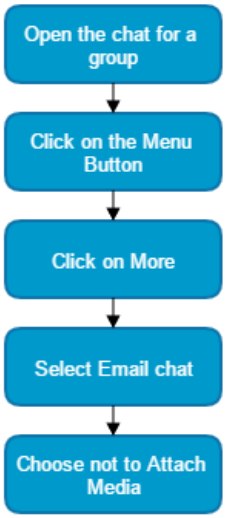

Fig 3: Steps to extract data from WhatsApp using email chat feature.

A .txt document of your chat history will thus be attached and an email will be composed to the specified sender.

4.1.2 Define Project Objective

The aim of this research paper is to predict whether a particular individual is said to be addicted to WhatsApp group or not and this is done using the R statistics software programme. The conclusion is expected to expose the level of addiction of an individual to WhatsApp Group and Association of Gender and Age with Frequency of use.

4.1.3 Setting up Machine

Once R is installed you can choose to work with an integrated development environment (IDE) RStudio. It is the most popular IDE for R and supports debugging, workspace management, plotting and much more [6].The Window of

RStudio is divided into four panes namely source pane, console pane, workspace pane and plots pane.

4.1.3.1 Source Pane

On the left side on top is the source pane where you can write and edit your R programs and documents.

4.1.3.2 Console Pane

It is located on the left side at the bottom, where results are displayed.

4.1.3.3 Workspace Pane

It is located on the right side on top and allows quick access to additional tools. It is used to perform the following functions

- **Environment** that exhibits data objects defined in the current R session.
- **History** is previous commands list that needs to be executed.

4.1.3.4 Plots Pane

Located on the right side on bottom, allowing quick access to additional tools and performing the following functions

- **Files** to browse folders.
- **Plots** illustrates plots created by the user.
- **Packages** option from where packages can be installed and loaded.
- **Help** to get help on R commands.

4.2 Data Transformation

Once the data on which analysis needs to be performed is known, it's time to transform it from raw data (.txt) to useable data (.csv) as shown in Fig 4.

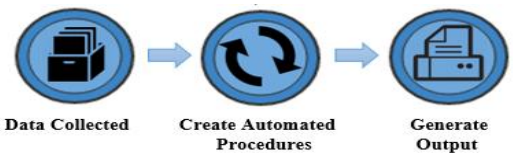

Fig 4: Process involved in Data Transformation

4.2.1 Data cleansing

Data cleansing is also known as Data Scrubbing in which inaccurate records from a particular dataset are corrected and eliminated. The purpose of data cleansing is to detect incorrect, irrelevant or insufficient parts of the data to either alter or delete it to ensure that a given set of data is accurate and consistent with other sets in the system.

Validations to be performed on the text file to avoid any issues while reading your CSV file (Spreadsheet) into R:

- In case of spreadsheet the first row is usually reserved for the header
- Avoid blank space in names, values or fields, else each word will be treated as a separate variable, resulting in errors that are related to the number of elements per line in your data set.
- To concatenate words, do this by making using of a dot (.) .For example Sender.Age
- Short names are preferred over long names.
- To avoid special symbols such as !, @, #, $, ^, ***, ,(,) ,- , ?,,,< ,> , / , | , \ , [,] ,{ , and }.
- Values missing in the data set are tend to be indicated with NA.

Performing all such validations on text file and hence converting it into csv file via the help of excel proved to be a tedious job and hence automating the process of conversion of text file into csv file by writing few lines of code proved to be more efficient, less time consuming as well as reduced manual work. The process of conversion with all the above mentioned validations thus shifted from hours to minutes.

4.2.2 Create Automated Procedures and Generate output:

This step involves automated conversion of text files

to csv files with the help of Microsoft Visual Studio 2010.

- With Visual Studio 2010, a form is been designed consisting of a text field labelled as Upload File, an upload button and a large text area in order to display the result.
- To upload a text file, the user needs to click on the text field which will allow the user to browse the files on the system and hence select the required text file.
- After the required text file is been selected, click on the upload button. On click on the upload button the specified text file is thus converted into csv file and the results are displayed in the large text area below as shown in Fig. 5

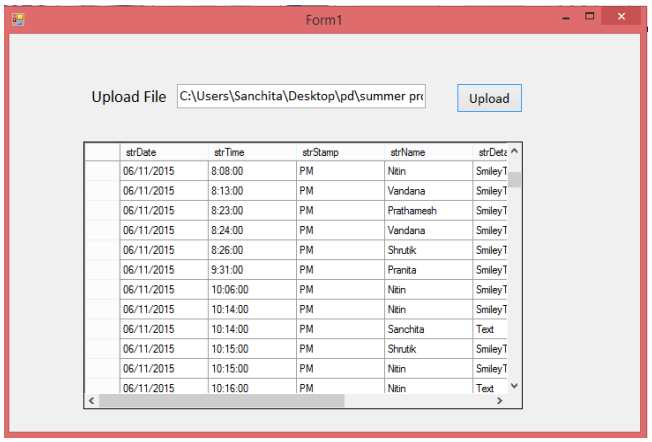

Fig 5: Form design

The resultant csv file is stored in the format filename_date-month-year-hour-minutes-second as shown in Fig. 6, in order to distinguish between multiples csv files.

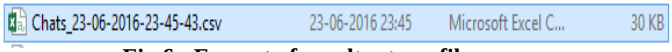

Fig 6: Format of resultant csv file

4.3 Data Loading

Data loading stage includes importing resultant csv file into RStudio as shown in Fig. 7

Fig 7: Process involved in Data Loading

4.3.1 Import Data File

In RStudio, click on the Workspace tab, and then click on "Import Dataset". Select "From local file". A file browser will open up, locate the .csv file and click Open. A dialog box will appear that would mention a few options on the import. Make sure if you have column names in your file then the Header is set to "Yes". Mention the separator as used for a particular csv file. In this case semicolon is been used as a separator. Finally Click "Import". Note that the name of your data appears in the Workspace pane. A preview of the data opens in the file-viewing pane.

4.4 Exploratory Data Analysis (EDA) and Visualization

EDA is an approach to data analysis for summarising and visualising the important characteristics of a data set [9].

Fig 8: Exploratory Data Analysis and Visualization

The aim of this research is to classify the number of users as those addicted and not addicted to WhatsApp group chat and thus predicting the level of addiction as well as to find a way to answer the below questions :

4.4.1 To find what type of communication medium people prefer the most in WhatsApp Group chat.

As our dataset consists of records from 24 May, 2015 to 28 May, 2016 the communication medium people preferred more in both the years is Smiley and Text as shown in Fig. 9

	Multimedia	SmileyText	Text
Female	661	1583	567
Male	800	1402	550

Fig 9: Most frequently used communication medium.

The Communication medium in the given dataset is divided into three parts mainly Multimedia (Audio, Video),Text , Smiley and Text.

From the above analysis, the type of communication medium preferred by both males and females is Smiley and Text.

Graphical presentation of the given analysis is shown in Fig. 10

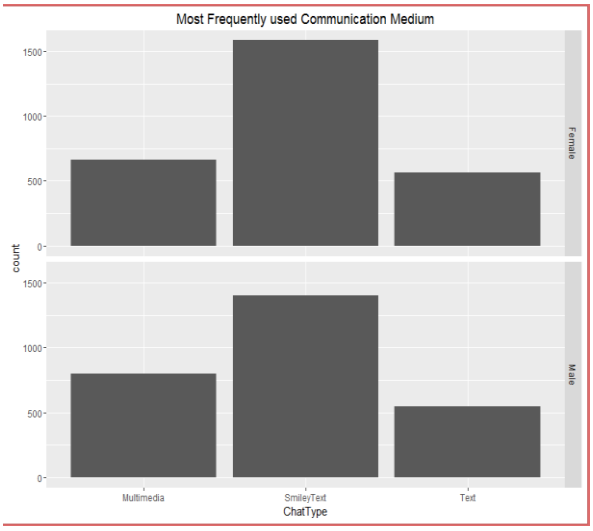

Fig 10: Graphical Representation of Frequently used Communication medium by Males and Females.

4.4.2 To find most active day of week.

As per the performed analysis and visualisation illustrated in Fig. 11 the most active day of the week is 'Monday' with total number of messages send as well as received are 520.So maximum participation of the senders on the WhatsApp family group chat takes place on Monday.

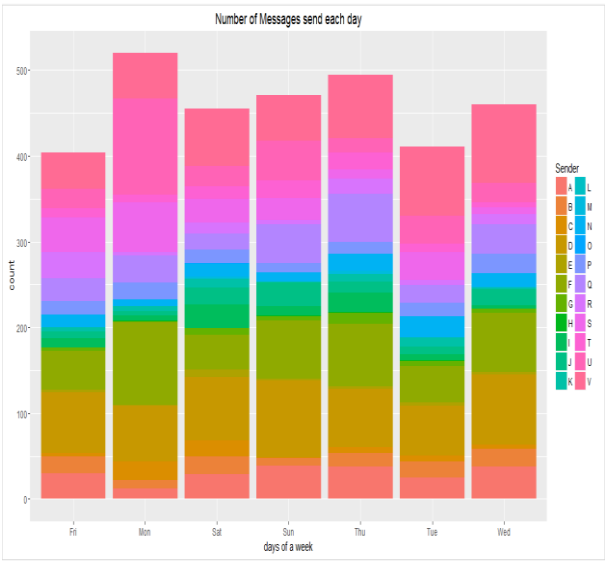

Fig 11: Most active day of week

4.4.3 To find which age group participants are more active on WhatsApp group and number of messages send by each age group participants per month, day, hour.

The age group of the participants in the Dataset ranges from 14 to 68. So the youngest person of the group is of age 14 and the most elder person has an age of 68.

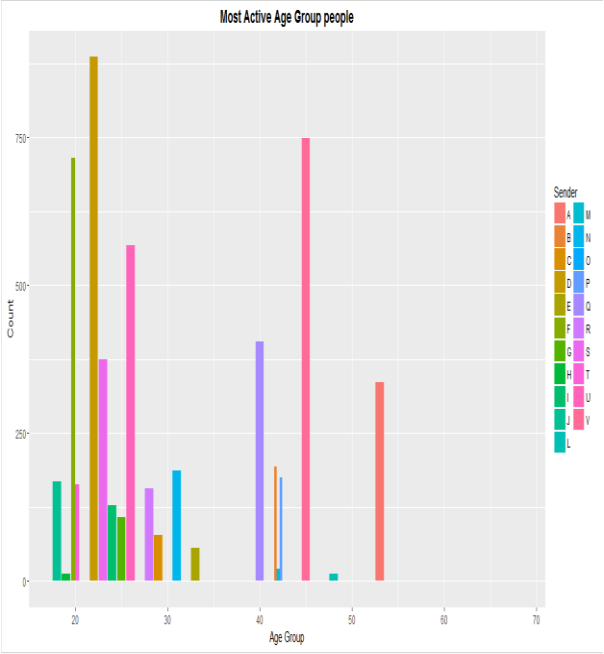

Fig 12: Graphical Representation of most active Age group.

According to the graphical representation as shown in Fig. 12, it is clear enough that the most active participants on the WhatsApp Group are of Age group 20 to 30 and least active participants are of Age group 50 to 60.

As per the analysis, total number of messages send by every age group person per hour is shown in Fig. 13 ,where 00,01 … 11 are number of hours and the column on the left depicts names of participants between age group 14 to 68. The filled rows and columns are the number of messages send by every participants per hour. Similar analysis can be performed as per months and per days instead of hours.

	00	01	02	03	04	05	06	07	08	09	10	11
A	17	8	16	14	20	20	21	47	65	57	37	14
B	20	18	2	4	3	11	16	19	27	31	31	12
C	3	1	2	1	0	2	6	7	7	42	4	3
D	21	27	30	19	18	44	72	130	179	166	147	33
E	11	7	1	0	0	3	1	1	5	15	10	2
F	24	21	22	17	32	16	34	114	155	147	78	55
G	2	1	5	0	1	5	2	6	4	11	34	37
H	0	0	0	0	2	2	1	1	0	4	2	0
I	16	3	5	0	1	7	3	6	18	30	20	19
J	5	10	6	1	12	9	21	13	19	26	22	24
K	13	0	3	0	0	0	3	8	28	6	9	1
L	0	0	8	1	0	0	0	0	1	1	1	0
M	1	4	0	1	0	1	1	5	1	5	2	0
N	19	14	14	7	4	5	11	16	21	25	42	8
O	0	0	0	0	0	0	0	0	0	0	1	0
P	9	8	2	8	8	8	12	21	53	23	13	10
Q	25	33	17	8	7	15	30	68	58	61	49	34
R	7	26	6	1	8	6	6	24	19	20	27	6
S	24	0	4	0	9	11	45	30	26	103	103	19
T	21	6	3	6	4	8	12	16	24	37	22	5
U	24	4	16	0	2	7	54	74	145	169	61	12
V	49	33	57	35	17	23	53	104	115	107	109	47

Fig 13: Analysis of total number of messages send by each age group participants per hour.

	x	freq
1	00	311
2	01	224
3	02	219
4	03	123
5	04	148
6	05	203
7	06	404
8	07	710
9	08	970
10	09	1086
11	10	824
12	11	341

Fig 14: Total Number of Messages send by each age group Participants per hour

In Fig .14 x stands for per hour and frequency is total number of messages send.

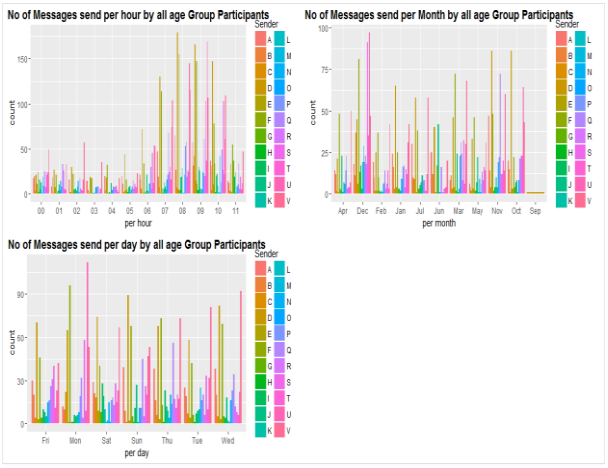

Fig 15: Total Number of Messages send by each age group Participants per hour/per day/per month.

4.4.4 To find whether Males are more addicted to the WhatsApp group or Females.

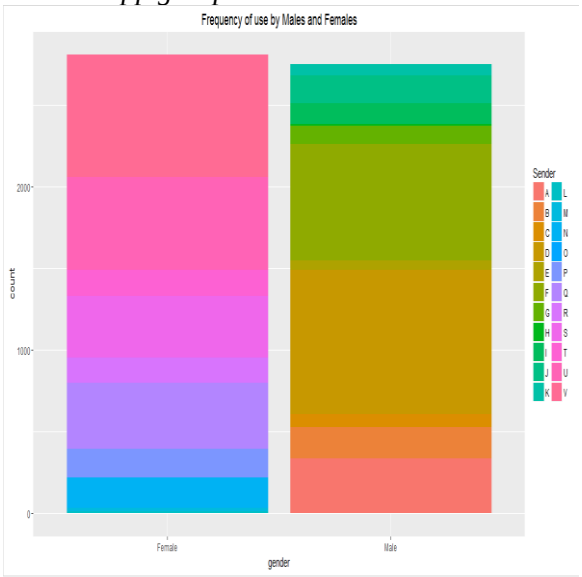

Fig 16: Frequency of using WhatsApp Group by males and females.

According to the analysis, Females are likely to be more addicted to the WhatsApp group as compared to Males. Total number of messages send by the Females are 2811 and by Males are 2752 as shown in Fig. 16 .This clearly concludes that Females are more involved in the Group.

4.4.5 Total number of messages send as per timestamp.

A total of 3674 messages had been delivered after noon while a whole of 1889 messages were delivered before noon. As a result maximum amount of interactions took place after noon as shown in Fig 17 with the help of Pie chart.

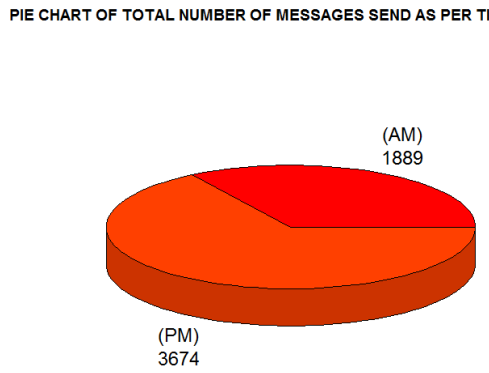

Fig 17: Total Number of Messages send as per timestamp

4.5 Results and Discussion

The following are the finding from the above analysis:

- Dataset consisted of equal number of male and female respondents.
- Majority of the respondents were in the age group of 20 to 30 years representing a young sample.
- Least respondents were in the age group of 50 to 60 years representing a senior sample.
- The most active day of the week is Monday.
- Females been more addicted to WhatsApp group than Males.
- Users mostly were in the favour to share Smiley and text as compared to other multimedia items.
- According to the Timestamp majority of messages were send after noon (PM).

5. CONCLUSION

From the performed analysis and visualization it is found that total number of active users in WhatsApp group chat are 22 consisting of equal number of males and females. Majority of the female users tend to be more addicted to WhatsApp group chat as compared to male users, due to various features provided by WhatsApp such as multimedia, Smiley and Text. The most addicted respondents were in the age group of 20 to 30 years representing a young sample. So as to conclude WhatsApp is one of the best communication platform whose pros and cons are decided by the user itself .If used positively then it's a boon for the users and if addicted then a ban and thus this research paper classified the level of addiction of users to the WhatsApp group chat so as to limit the time spend on it and to explore the group whenever necessary.

6. REFERENCES

[1] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp. 23-26, April 2016.

[2] Sanchita Patil," Big data analytics using R**",** International Research Journal of Engineering and Technology (IRJET) Volume 3, Issue 7 July 2016,pp. 78-81.

[3] Sagar Deshmukh, "Analysis of WhatsApp Users and Its Usage worldwide", International Journal of Scientific and Research Publications, Volume 5, Issue 8, pp. 1-3,August 2015 1 ISSN 2250-3153.

[4] https://en.wikipedia.org/wiki/Data_analysis#cite_note-O.27Neil_and_Schutt_2014-3

[5] https://www.reference.com/education/purpose-collecting-data-8d8be32cc477eb45#

[6] Tal Galili, "R-bloggers", December 10, 2015. [Online] Available: http://www.r-bloggers.com/how-to-learn-r-2/ [Accessed: 23- July- 2016]

[7] SSCC (social science computing cooperative), "R for Researchers: Projects". [Online]Available:http://www.ssc.wisc.edu/sscc/pubs/RFR/RFR_Projects.html [Accessed: 23- July- 2016]

[8] "This R Data Import Tutorial Is Everything You Need",July 21st, 2015 in R Programming. [Online]Available:https://www.datacamp.com/community/tutorials/r-data-import-tutorial#gs.EYeqhvc .[Accessed: 23- July- 2016]

[9] Jovial, "Exploratory data analysis with R", [Online] Available: https://rpubs.com/Jovial/R [Accessed: 23- July- 2016]

7. AUTHOR PROFILE

Miss Sanchita Patil.is currently a final year M.C.A student in Vivekanand Education Society's Institute of Technology (V.E.S.I.T), Mumbai.

She had completed her BSc (I.T) graduation in the year 2014.She has an abiding interest in Database programming and Data analysis.