INDIA INTERNET USAGE

Hardika Muni
School of Information Technology
Illinois State University
hmuni@ilstu.edu

October 21, 2014

1 Beginning of Internet in India

Internet first came to India in the year 1995. It was then growing slowly as many people were not aware of the Internet. Slowly the population was getting aware of the Internet and they started using it gradually. In 1998, it was just 0.01% of the population of India was aware of internet whereas in United States at that time 31% of the popluation of USA was using Internet. In 2006, internet in India took a revolution change and thyere were more than 33 million of the population who were using Internet. By the end of 2012, internet had become one of the important commodity of the people in day-to-day life. But still only 12% of the population is using internet whereas in USA 82% of the population is using internet.

2 Reading in the data

Reading the data from the Quandl source for India Internet Usage. Here we download the data from Quandl library and upload here in the R studio. We retreive Excel file using read.csv function and for that we need plyr library to be downloaded.

2.1 Uncleaned Data

Data before getting cleaned. This is the data which we retreived from Quandl library which is not yet cleaned.

```
         YEAR    Users Population X..Pen. Usage.Source
1  2012-12-31 1.37e+08 1205073612    11.4           NA
2  2010-12-31 1.00e+08 1173108018     8.5           NA
3  2009-12-31 8.10e+07 1156897766     7.0           NA
4  2007-12-31 4.20e+07 1129667528     3.7           NA
5  2006-12-31 4.00e+07 1112225812     3.6           NA
```

```
6  2005-12-31 5.06e+07 1112225812     4.5           NA
7  2004-12-31 3.92e+07 1094870677     3.6           NA
8  2003-12-31 2.25e+07 1094870677     2.1           NA
9  2002-12-31 1.65e+07 1094870677     1.6           NA
10 2001-12-31 7.00e+06 1094870677     0.7           NA
11 2000-12-31 5.50e+06 1094870677     0.5           NA
12 1999-12-31 2.80e+06 1094870677     0.3           NA
13 1998-12-31 1.40e+06 1094870677     0.1           NA
```

2.2 Cleaned Data

Cleaned data after the uncleaned data is processed to get a proper data. Here , We remove the last column of the data which is not required as there is no data coming from that. Also, the population and users value is too large and so it displays in logarithmic form. To convert that, we used as.integer to convert the numeric into an integer form and hence it will display the complete data. This is also called scrubbing the data. Removing the unwanted columns and retaining the columns and rows which are most important.

```
> data.new<-as.data.frame(data[,1:4])
> data.new <- rename(data.new, c("X..Pen."="Percentage"))
> data.new$Users <- as.integer(as.numeric(data.new$Users))
> data.new$Population <- as.integer(as.numeric(data.new$Population))
> head(data.new,15)
```

```
         YEAR     Users Population Percentage
1  2012-12-31 137000000 1205073612       11.4
2  2010-12-31 100000000 1173108018        8.5
3  2009-12-31  81000000 1156897766        7.0
4  2007-12-31  42000000 1129667528        3.7
5  2006-12-31  40000000 1112225812        3.6
6  2005-12-31  50600000 1112225812        4.5
7  2004-12-31  39200000 1094870677        3.6
8  2003-12-31  22500000 1094870677        2.1
9  2002-12-31  16500000 1094870677        1.6
10 2001-12-31   7000000 1094870677        0.7
11 2000-12-31   5500000 1094870677        0.5
12 1999-12-31   2800000 1094870677        0.3
13 1998-12-31   1400000 1094870677        0.1
```

3 Data Section

Here, the class function displays the data frame. Str function gives us the total number of observations, number of variables that is columns , each columns data type and their value. The summary function gives the Minimum and the

Maximum value of the column , also it gives the median of the columns. This is very useful when we want to know about the values and the data types.

```
> class(data.new)

[1] "data.frame"

> str(data.new)

'data.frame':        13 obs. of  4 variables:
 $ YEAR      : Factor w/ 13 levels "1998-12-31","1999-12-31",..: 13 12 11 10 9 8 7 6 5 4 ...
 $ Users     : int  137000000 100000000 81000000 42000000 40000000 50600000 39200000 2250000
 $ Population: int  1205073612 1173108018 1156897766 1129667528 1112225812 1112225812 109487
 $ Percentage: num  11.4 8.5 7 3.7 3.6 4.5 3.6 2.1 1.6 0.7 ...

> summary(data.new)

         YEAR       Users              Population          Percentage
 1998-12-31:1   Min.   :  1400000   Min.   :1.095e+09   Min.   : 0.100
 1999-12-31:1   1st Qu.:  7000000   1st Qu.:1.095e+09   1st Qu.: 0.700
 2000-12-31:1   Median : 39200000   Median :1.095e+09   Median : 3.600
 2001-12-31:1   Mean   : 41961538   Mean   :1.119e+09   Mean   : 3.662
 2002-12-31:1   3rd Qu.: 50600000   3rd Qu.:1.130e+09   3rd Qu.: 4.500
 2003-12-31:1   Max.   :137000000   Max.   :1.205e+09   Max.   :11.400
 (Other)   :7
```

We need a graph below which shows two lines one for population and other for users. So, we use the melt function which will group the values according to YEAR and Percentage. To use the melt funtion we need to download reshape2 package and load that library before using the melt function.

```
> library(reshape2)
> data.long <- melt(data.new,id.vars=c("YEAR","Percentage"))
> head(data.long,27)

         YEAR Percentage   variable      value
1  2012-12-31       11.4      Users  137000000
2  2010-12-31        8.5      Users  100000000
3  2009-12-31        7.0      Users   81000000
4  2007-12-31        3.7      Users   42000000
5  2006-12-31        3.6      Users   40000000
6  2005-12-31        4.5      Users   50600000
7  2004-12-31        3.6      Users   39200000
8  2003-12-31        2.1      Users   22500000
9  2002-12-31        1.6      Users   16500000
10 2001-12-31        0.7      Users    7000000
11 2000-12-31        0.5      Users    5500000
12 1999-12-31        0.3      Users    2800000
```

```
13 1998-12-31        0.1      Users    1400000
14 2012-12-31       11.4 Population 1205073612
15 2010-12-31        8.5 Population 1173108018
16 2009-12-31        7.0 Population 1156897766
17 2007-12-31        3.7 Population 1129667528
18 2006-12-31        3.6 Population 1112225812
19 2005-12-31        4.5 Population 1112225812
20 2004-12-31        3.6 Population 1094870677
21 2003-12-31        2.1 Population 1094870677
22 2002-12-31        1.6 Population 1094870677
23 2001-12-31        0.7 Population 1094870677
24 2000-12-31        0.5 Population 1094870677
25 1999-12-31        0.3 Population 1094870677
26 1998-12-31        0.1 Population 1094870677
```

4 Result

As per the graphs and the table below, it shows that there are not much users in India using Internet with respect to the population. It might be that many people are not aware of the internet facility and they might not know the advantages of it. Also, it is possible that some people cannot afford internet services because of lower class society. Also it is seen that the usage of internet is increasing every year and the graph predicts that in coming few years, there will be more than 50% of the population in India who will be connected to internet and technology.

```
> head(data.new,13)

         YEAR     Users Population Percentage
1  2012-12-31 137000000 1205073612       11.4
2  2010-12-31 100000000 1173108018        8.5
3  2009-12-31  81000000 1156897766        7.0
4  2007-12-31  42000000 1129667528        3.7
5  2006-12-31  40000000 1112225812        3.6
6  2005-12-31  50600000 1112225812        4.5
7  2004-12-31  39200000 1094870677        3.6
8  2003-12-31  22500000 1094870677        2.1
9  2002-12-31  16500000 1094870677        1.6
10 2001-12-31   7000000 1094870677        0.7
11 2000-12-31   5500000 1094870677        0.5
12 1999-12-31   2800000 1094870677        0.3
13 1998-12-31   1400000 1094870677        0.1
```

5 Graphs

Here we plot the Percentage of Internet users by Year. We do this using ggplot. For using ggplot, we eed to install ggplot2 package and then setup ggplot library. When using ggplot function, we first give the data frame name as given here in Figure 1 . Then in aes function we give the X axis and the Y axis and group it to 1 since we want a single line graph. Then we use a plus sign to concatinate and use geomline function which gives a straight line . By using geompoint function, at every value it also gives a point. We can use color function to specify the color of the line.

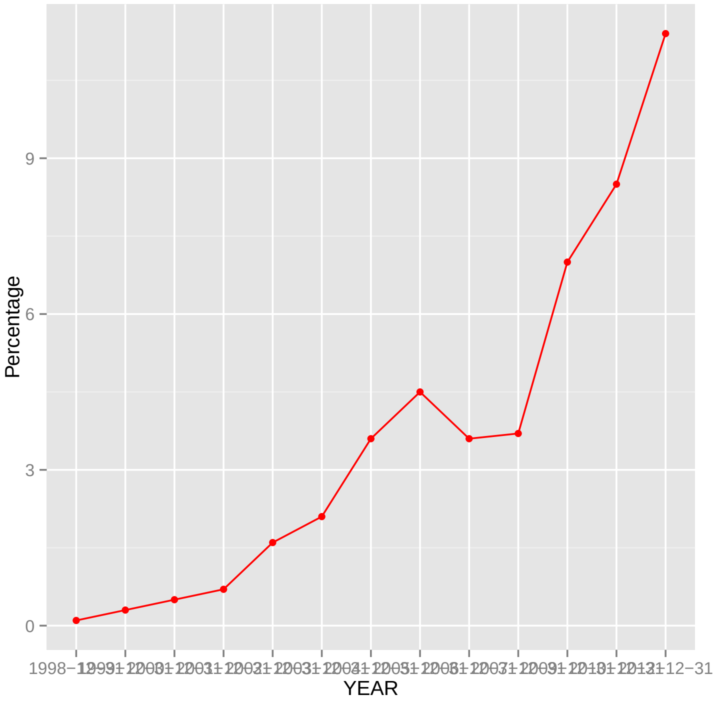

Figure 1: Percentage of users every Year

This is a 2 line graph in a single graph. For this we used the melt command before this. That will group the values according to Year and Percentage. The values we got there , can be used here in ggplot function. The X axis remain the same which is Year and y changes to the value. We group it by the variable and also give different color to both the lines.

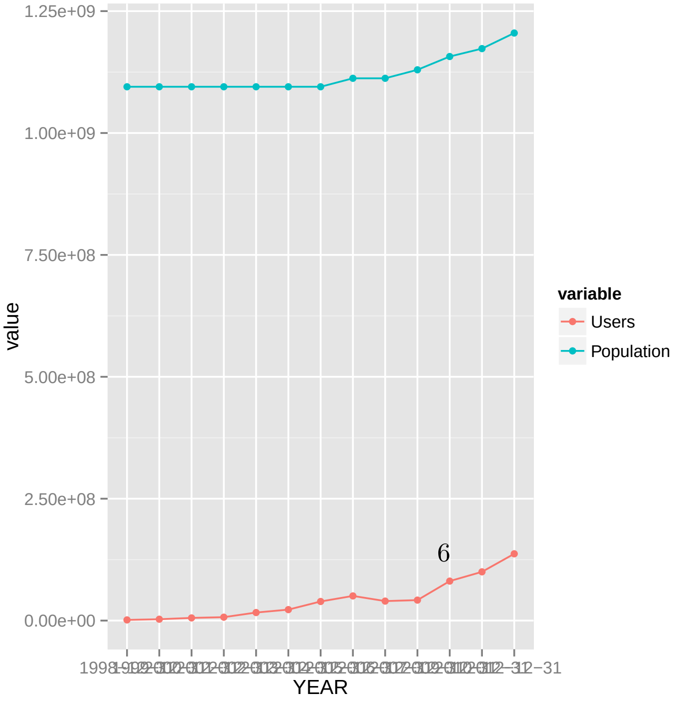

Figure 2: Population And Users of India