DATA INTENSIVE COMPUTING

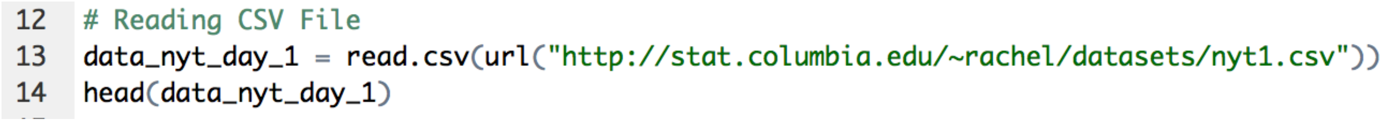
Simple EDA on New York Times Data

Ramanpreet Singh Khinda | DIC 587 | March 5, 2016
rkhinda@buffalo.edu
50169622

Problem 2A: Simple EDA on single day data

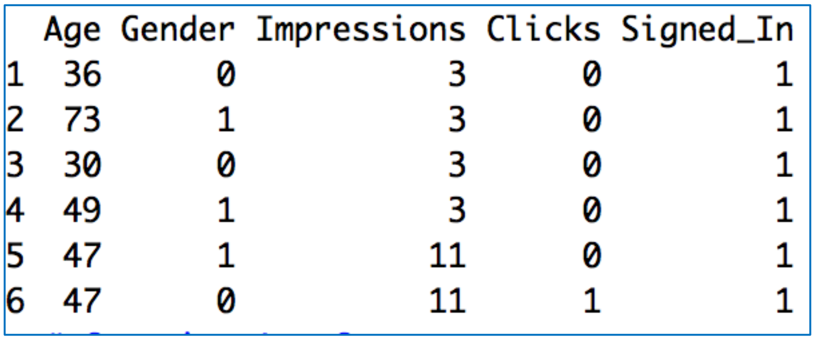
Step 1: Reading the data

The first step in the process of EDA is to read the data which in our case is stored in the csv file.



```
# Reading CSV File
data_nyt_day_1 = read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))
head(data_nyt_day_1)
```

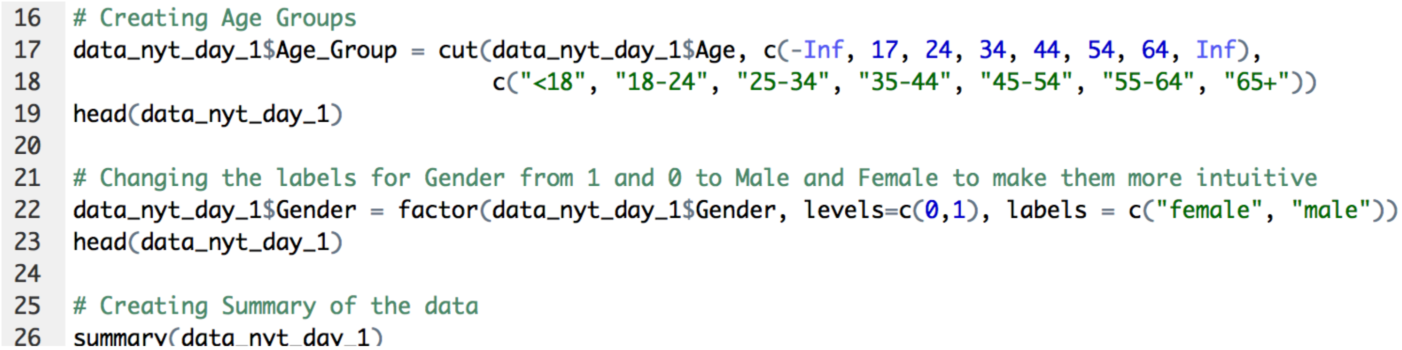
The data looks something like this



```
  Age Gender Impressions Clicks Signed_In
1  36      0           3      0         1
2  73      1           3      0         1
3  30      0           3      0         1
4  49      1           3      0         1
5  47      1          11      0         1
6  47      0          11      1         1
```

Step 2: Cleaning the data

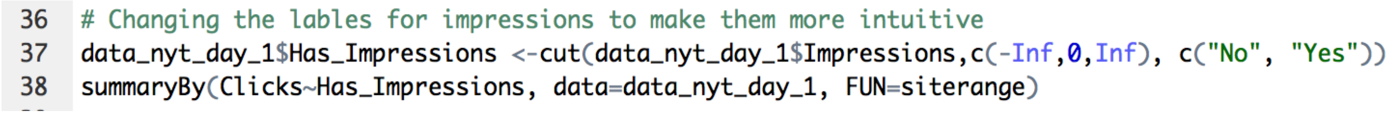
In this step we will perform some cleaning operation to make the data suitable for EDA



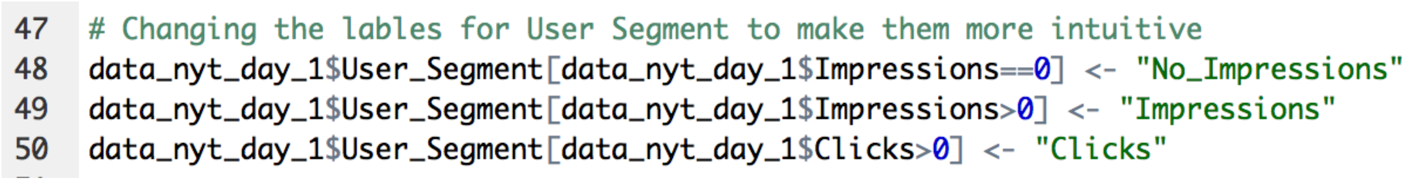
```
# Creating Age Groups
data_nyt_day_1$Age_Group = cut(data_nyt_day_1$Age, c(-Inf, 17, 24, 34, 44, 54, 64, Inf),
                               c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))
head(data_nyt_day_1)

# Changing the labels for Gender from 1 and 0 to Male and Female to make them more intuitive
data_nyt_day_1$Gender = factor(data_nyt_day_1$Gender, levels=c(0,1), labels = c("female", "male"))
head(data_nyt_day_1)

# Creating Summary of the data
summary(data_nyt_day_1)
```



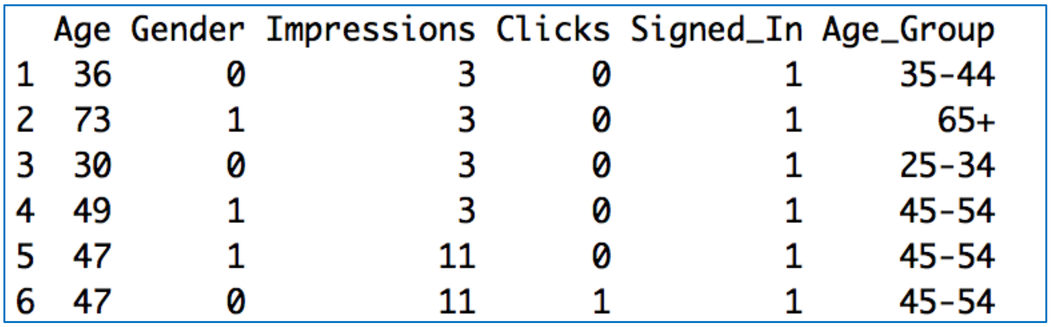
```
# Changing the lables for impressions to make them more intuitive
data_nyt_day_1$Has_Impressions <-cut(data_nyt_day_1$Impressions,c(-Inf,0,Inf), c("No", "Yes"))
summaryBy(Clicks~Has_Impressions, data=data_nyt_day_1, FUN=siterange)
```



```
# Changing the lables for User Segment to make them more intuitive
data_nyt_day_1$User_Segment[data_nyt_day_1$Impressions==0] <- "No_Impressions"
data_nyt_day_1$User_Segment[data_nyt_day_1$Impressions>0] <- "Impressions"
data_nyt_day_1$User_Segment[data_nyt_day_1$Clicks>0] <- "Clicks"
```

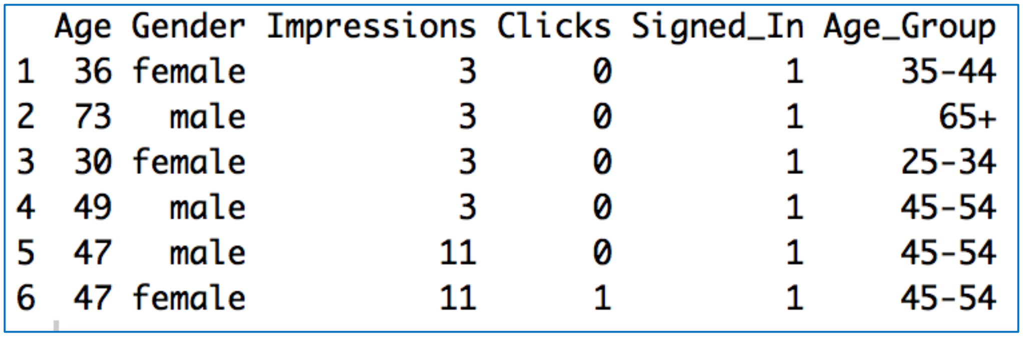
After following very simple cleaning process as described above we are able to get a beautiful data

Grouping based on Age Groups



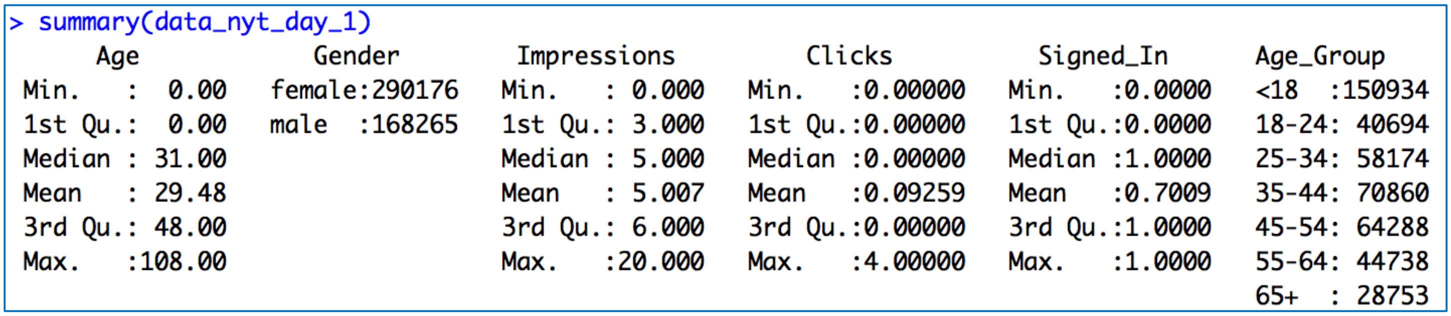
```
  Age Gender Impressions Clicks Signed_In Age_Group
1  36      0           3      0         1     35-44
2  73      1           3      0         1       65+
3  30      0           3      0         1     25-34
4  49      1           3      0         1     45-54
5  47      1          11      0         1     45-54
6  47      0          11      1         1     45-54
```

Changing labels for Gender

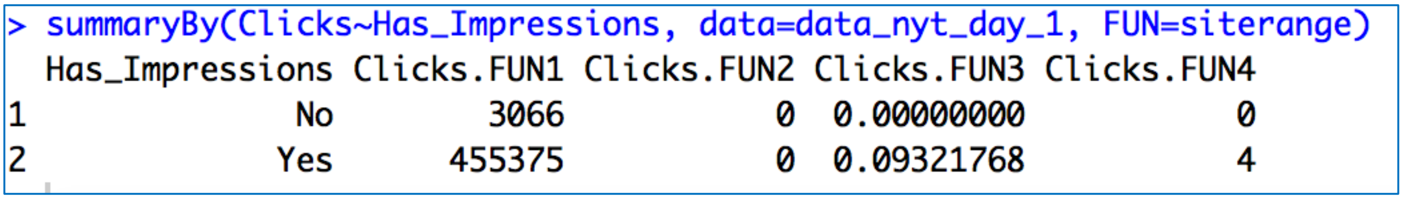


```
  Age Gender Impressions Clicks Signed_In Age_Group
1  36 female           3      0         1     35-44
2  73   male           3      0         1       65+
3  30 female           3      0         1     25-34
4  49   male           3      0         1     45-54
5  47   male          11      0         1     45-54
6  47 female          11      1         1     45-54
```

Summary Report



```
> summary(data_nyt_day_1)
      Age              Gender         Impressions         Clicks          Signed_In         Age_Group
 Min.   :  0.00   female:290176   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000   <18   :150934
 1st Qu.:  0.00   male  :168265   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000   18-24: 40694
 Median : 31.00                   Median : 5.000   Median :0.00000   Median :1.0000   25-34: 58174
 Mean   : 29.48                   Mean   : 5.007   Mean   :0.09259   Mean   :0.7009   35-44: 70860
 3rd Qu.: 48.00                   3rd Qu.: 6.000   3rd Qu.:0.00000   3rd Qu.:1.0000   45-54: 64288
 Max.   :108.00                   Max.   :20.000   Max.   :4.00000   Max.   :1.0000   55-64: 44738
                                                                                      65+  : 28753
```

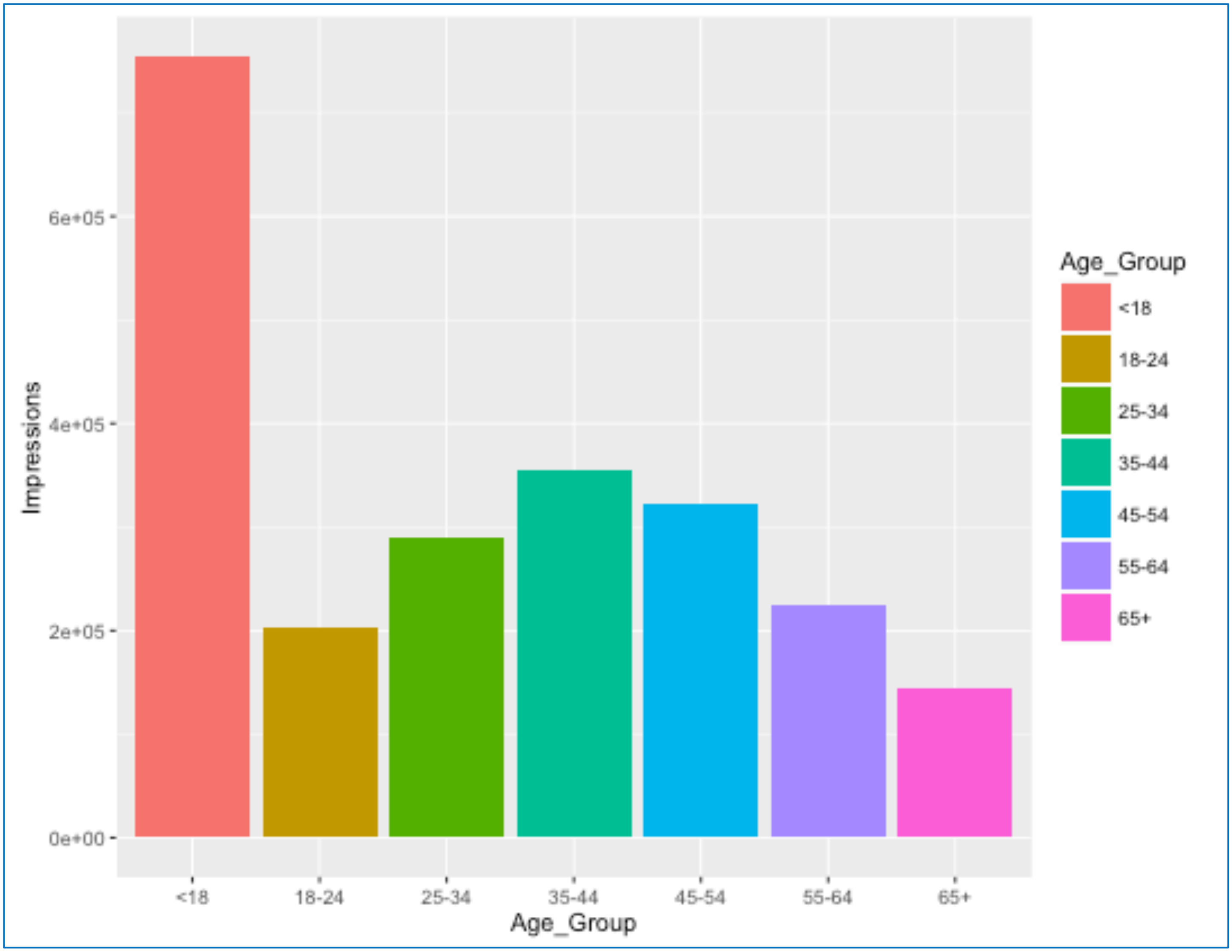


```
> summaryBy(Clicks~Has_Impressions, data=data_nyt_day_1, FUN=siterange)
  Has_Impressions Clicks.FUN1 Clicks.FUN2 Clicks.FUN3 Clicks.FUN4
1              No        3066           0  0.00000000           0
2             Yes      455375           0  0.09321768           4
```

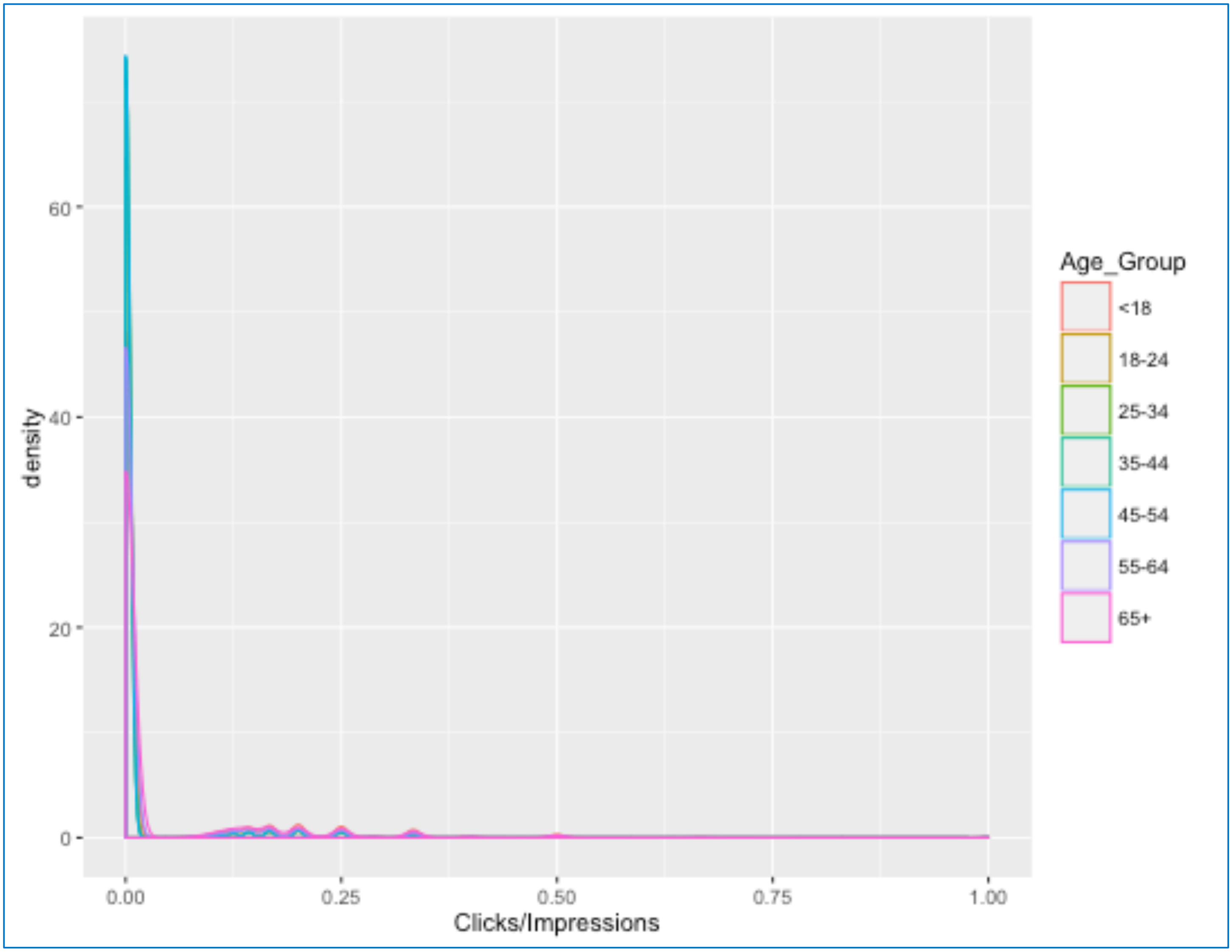
Step 3: Generating Plots and capture meaningful information

In the next step we have to capture meaningful information from the data that we cleaned. Graphs are very intuitive way of depicting such information in a very elegant way.

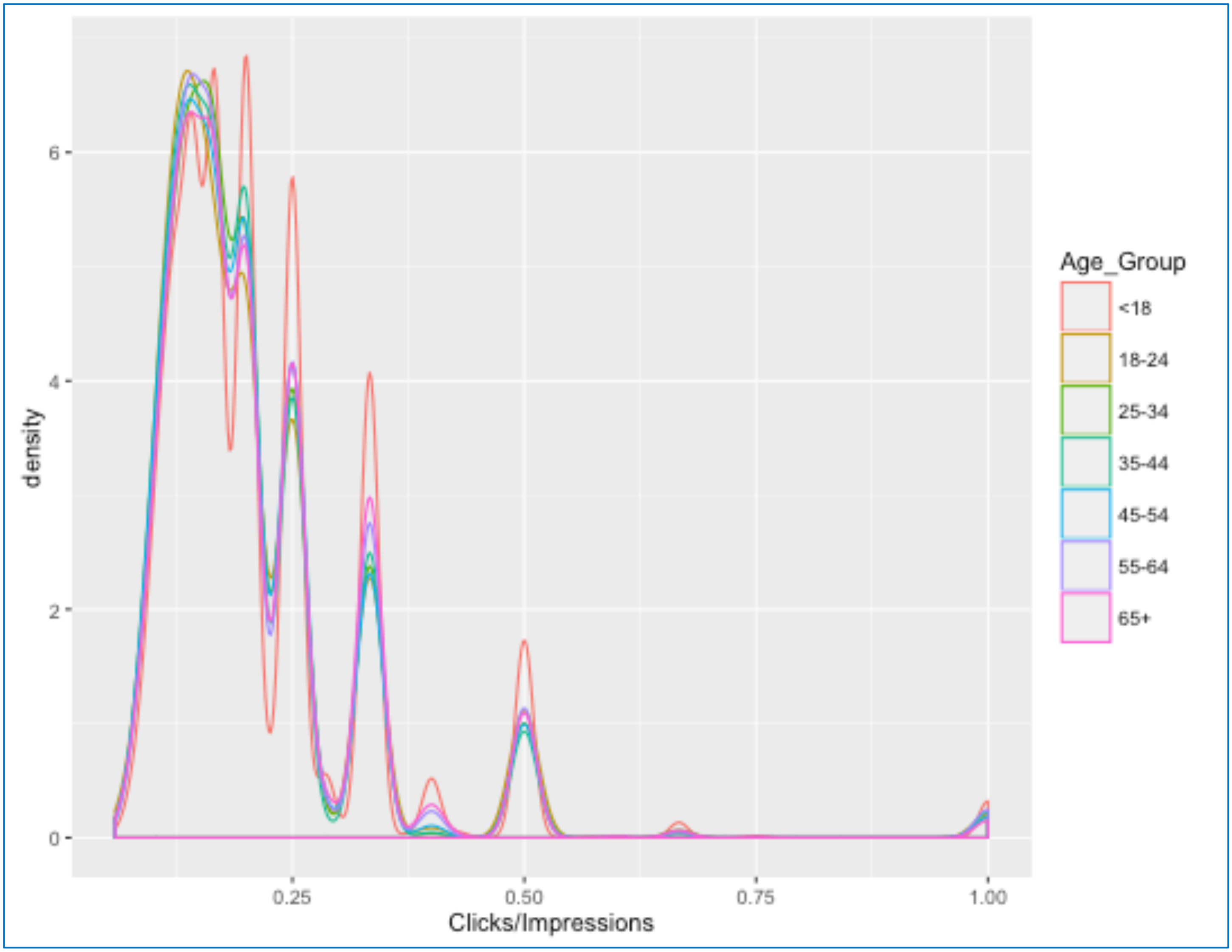
Graph on Age Groups vs No. of Impressions



Graph on Click thru Rate



Graph on Click thru Rate

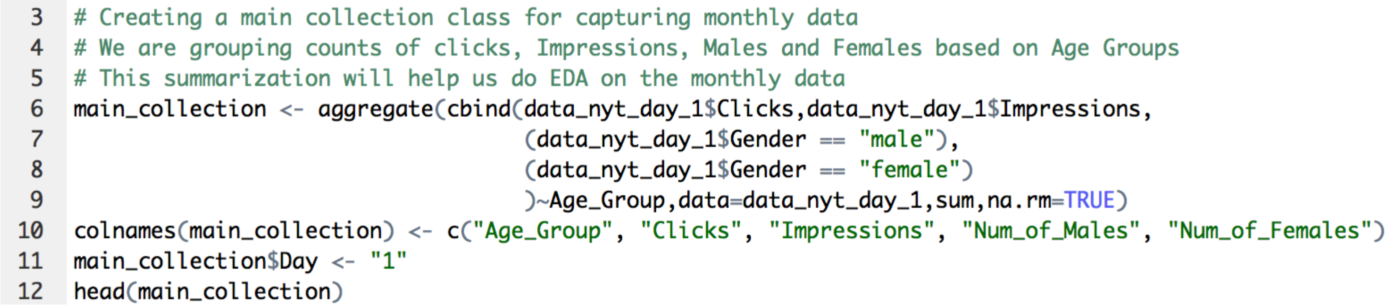


Problem 2B: Simple EDA on monthly data

Step 1:

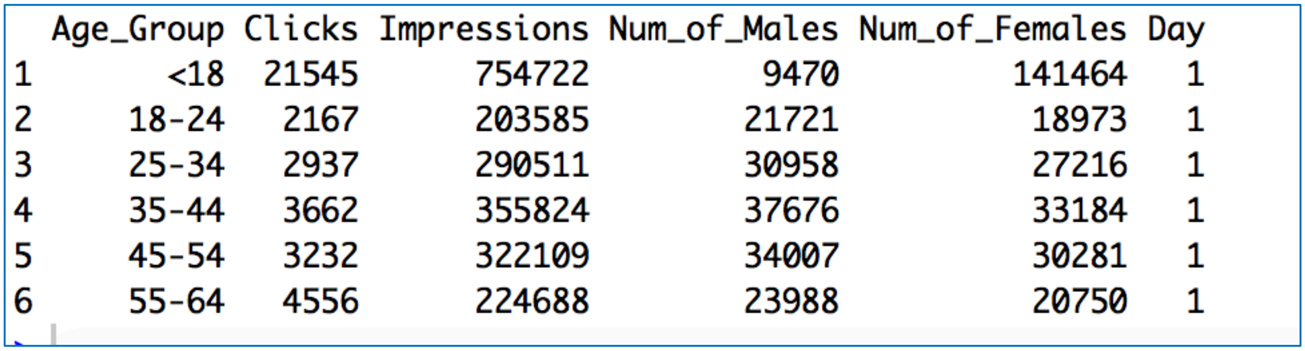
Now to do analysis on monthly data we will do grouping of counts of clicks, Impressions, Males and Females based on Age Groups.

This summarization will help us do EDA on the monthly data



```
# Creating a main collection class for capturing monthly data
# We are grouping counts of clicks, Impressions, Males and Females based on Age Groups
# This summarization will help us do EDA on the monthly data
main_collection <- aggregate(cbind(data_nyt_day_1$Clicks,data_nyt_day_1$Impressions,
                                   (data_nyt_day_1$Gender == "male"),
                                   (data_nyt_day_1$Gender == "female")
                                   )~Age_Group,data=data_nyt_day_1,sum,na.rm=TRUE)
colnames(main_collection) <- c("Age_Group", "Clicks", "Impressions", "Num_of_Males", "Num_of_Females")
main_collection$Day <- "1"
head(main_collection)
```

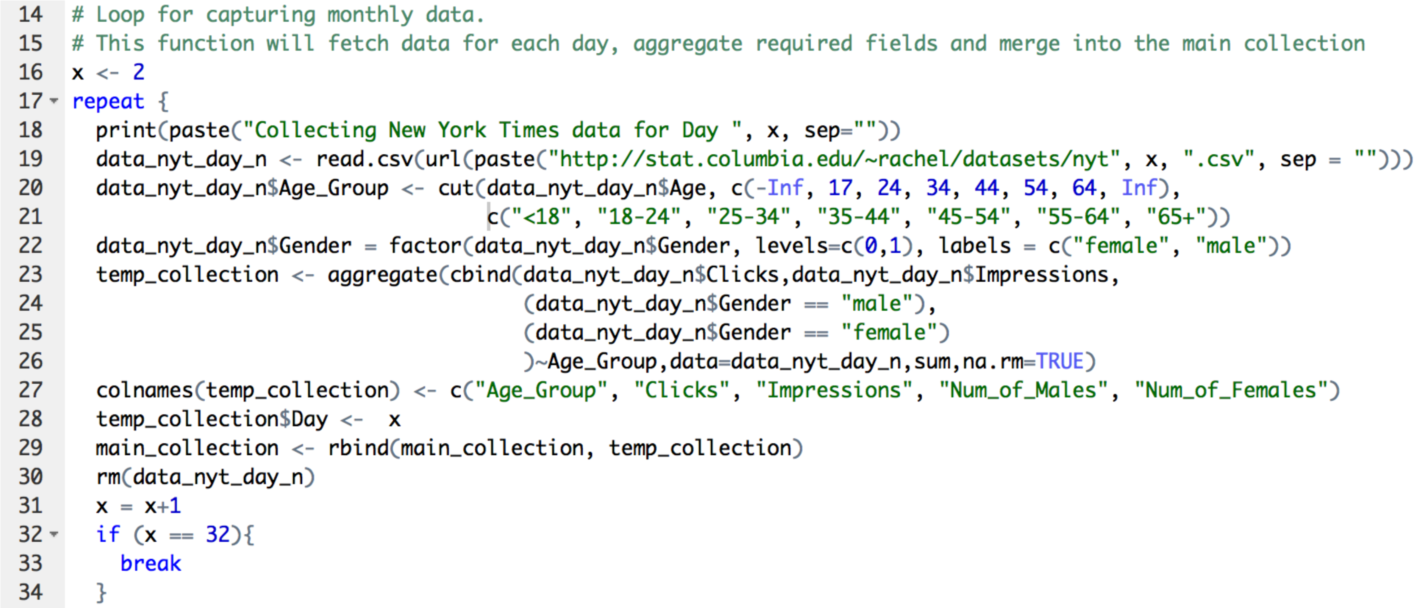
After aggregating required data, we will get something like this



```
  Age_Group Clicks Impressions Num_of_Males Num_of_Females Day
1       <18  21545      754722         9470         141464   1
2     18-24   2167      203585        21721          18973   1
3     25-34   2937      290511        30958          27216   1
4     35-44   3662      355824        37676          33184   1
5     45-54   3232      322109        34007          30281   1
6     55-64   4556      224688        23988          20750   1
```

Step 2:

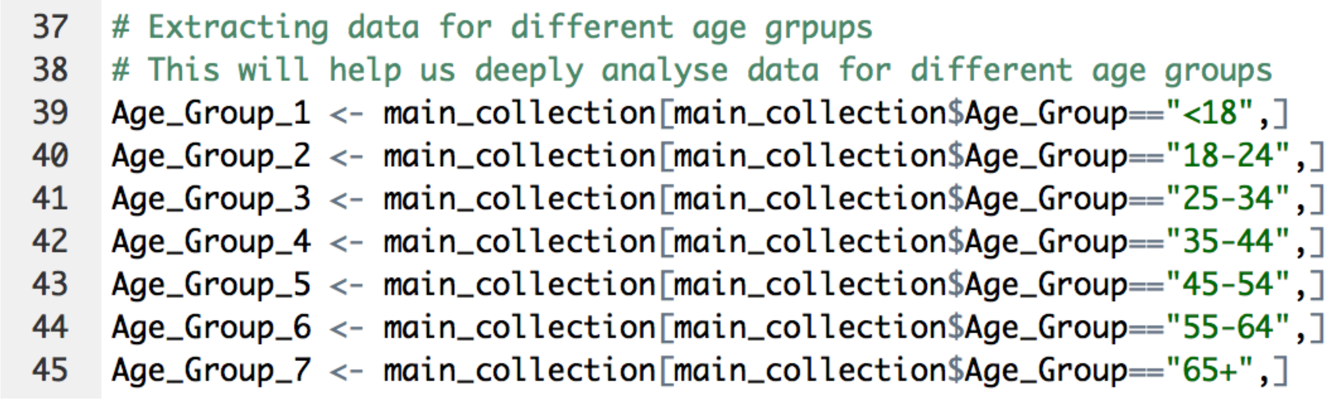
In the next step we will iterate over monthly data and perform the same steps as we performed previously and collect the data into the main collection



```r
# Loop for capturing monthly data.
# This function will fetch data for each day, aggregate required fields and merge into the main collection
x <- 2
repeat {
  print(paste("Collecting New York Times data for Day ", x, sep=""))
  data_nyt_day_n <- read.csv(url(paste("http://stat.columbia.edu/~rachel/datasets/nyt", x, ".csv", sep = "")))
  data_nyt_day_n$Age_Group <- cut(data_nyt_day_n$Age, c(-Inf, 17, 24, 34, 44, 54, 64, Inf),
                                  c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"))
  data_nyt_day_n$Gender = factor(data_nyt_day_n$Gender, levels=c(0,1), labels = c("female", "male"))
  temp_collection <- aggregate(cbind(data_nyt_day_n$Clicks,data_nyt_day_n$Impressions,
                                     (data_nyt_day_n$Gender == "male"),
                                     (data_nyt_day_n$Gender == "female")
                                     )~Age_Group,data=data_nyt_day_n,sum,na.rm=TRUE)
  colnames(temp_collection) <- c("Age_Group", "Clicks", "Impressions", "Num_of_Males", "Num_of_Females")
  temp_collection$Day <-  x
  main_collection <- rbind(main_collection, temp_collection)
  rm(data_nyt_day_n)
  x = x+1
  if (x == 32){
    break
  }
```

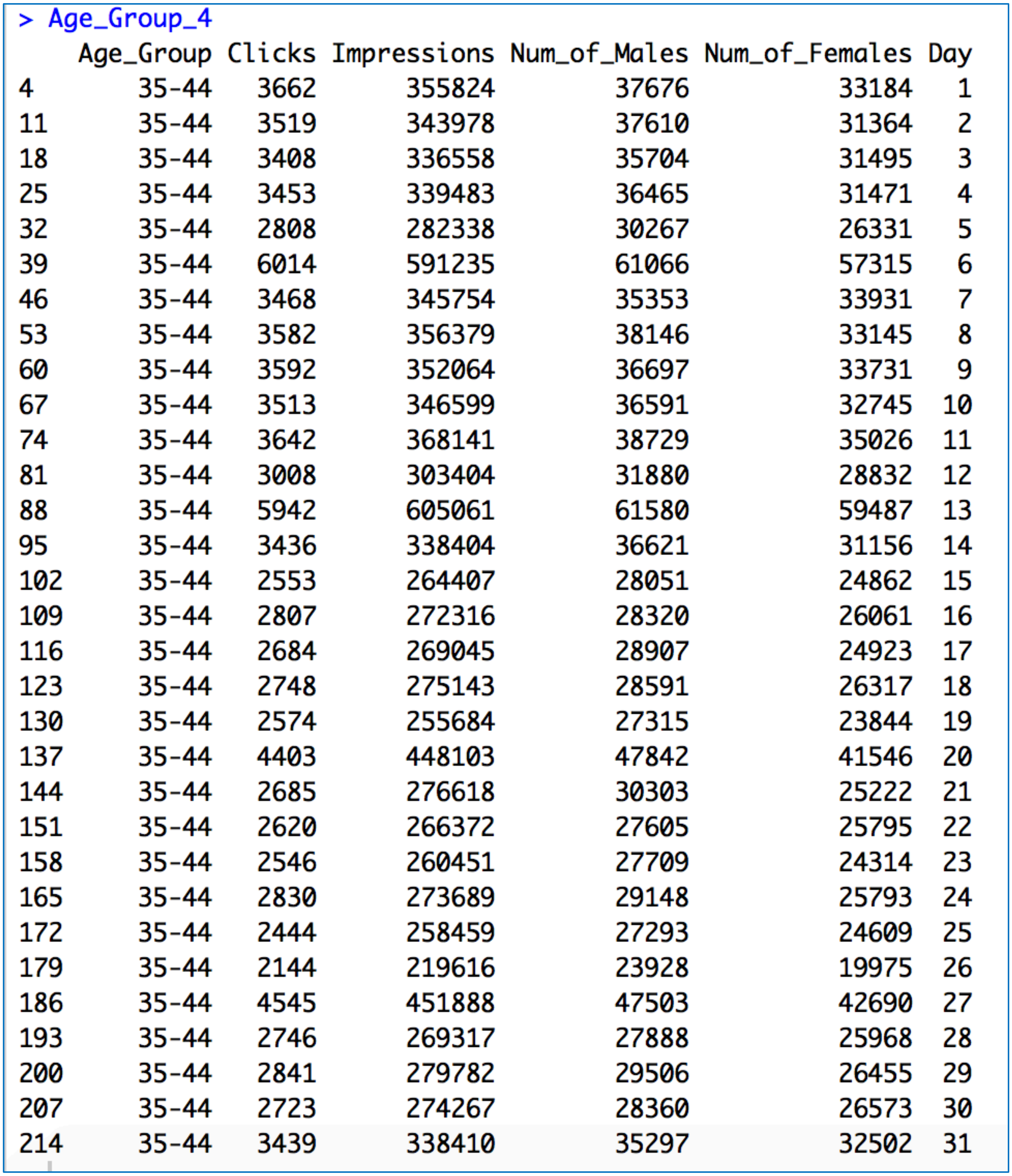
Step 3:

Next we will group data based on different age groups to analyze them more deeply



```r
# Extracting data for different age grpups
# This will help us deeply analyse data for different age groups
Age_Group_1 <- main_collection[main_collection$Age_Group=="<18",]
Age_Group_2 <- main_collection[main_collection$Age_Group=="18-24",]
Age_Group_3 <- main_collection[main_collection$Age_Group=="25-34",]
Age_Group_4 <- main_collection[main_collection$Age_Group=="35-44",]
Age_Group_5 <- main_collection[main_collection$Age_Group=="45-54",]
Age_Group_6 <- main_collection[main_collection$Age_Group=="55-64",]
Age_Group_7 <- main_collection[main_collection$Age_Group=="65+",]
```

The aggregated monthly data for users falling under Age Group 35-44 will look like this

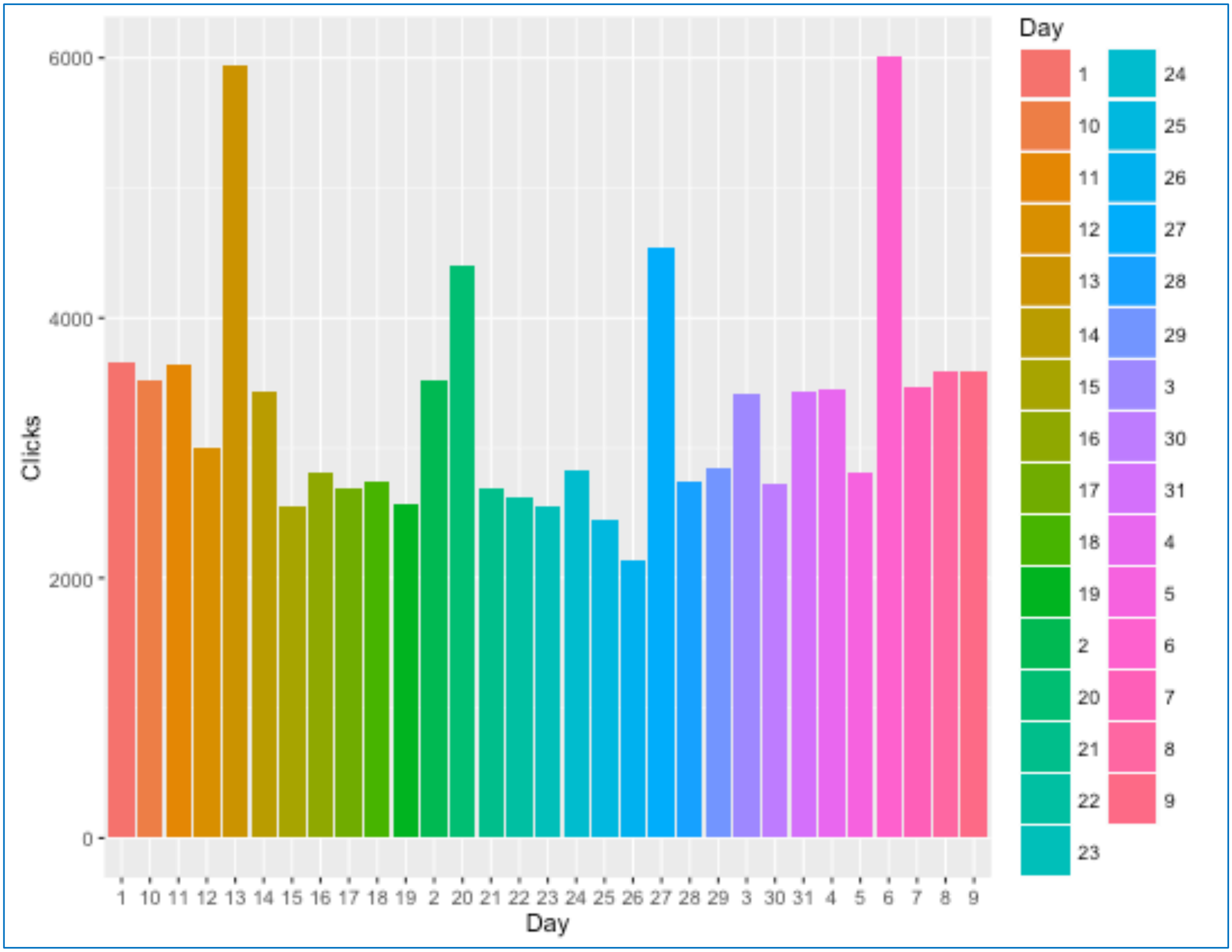


```
> Age_Group_4
    Age_Group Clicks Impressions Num_of_Males Num_of_Females Day
4       35-44   3662      355824        37676          33184   1
11      35-44   3519      343978        37610          31364   2
18      35-44   3408      336558        35704          31495   3
25      35-44   3453      339483        36465          31471   4
32      35-44   2808      282338        30267          26331   5
39      35-44   6014      591235        61066          57315   6
46      35-44   3468      345754        35353          33931   7
53      35-44   3582      356379        38146          33145   8
60      35-44   3592      352064        36697          33731   9
67      35-44   3513      346599        36591          32745  10
74      35-44   3642      368141        38729          35026  11
81      35-44   3008      303404        31880          28832  12
88      35-44   5942      605061        61580          59487  13
95      35-44   3436      338404        36621          31156  14
102     35-44   2553      264407        28051          24862  15
109     35-44   2807      272316        28320          26061  16
116     35-44   2684      269045        28907          24923  17
123     35-44   2748      275143        28591          26317  18
130     35-44   2574      255684        27315          23844  19
137     35-44   4403      448103        47842          41546  20
144     35-44   2685      276618        30303          25222  21
151     35-44   2620      266372        27605          25795  22
158     35-44   2546      260451        27709          24314  23
165     35-44   2830      273689        29148          25793  24
172     35-44   2444      258459        27293          24609  25
179     35-44   2144      219616        23928          19975  26
186     35-44   4545      451888        47503          42690  27
193     35-44   2746      269317        27888          25968  28
200     35-44   2841      279782        29506          26455  29
207     35-44   2723      274267        28360          26573  30
214     35-44   3439      338410        35297          32502  31
```

Step 4: Analyzing monthly data by plotting graphs

Next we will plot some beautiful graphs which will give us intuition about behavior change in different Age Groups.

For example, for the users falling under Age Group 35-44 we see that the maximum number of clicks were on Day 6th and 13th


Clicks
Day
Day