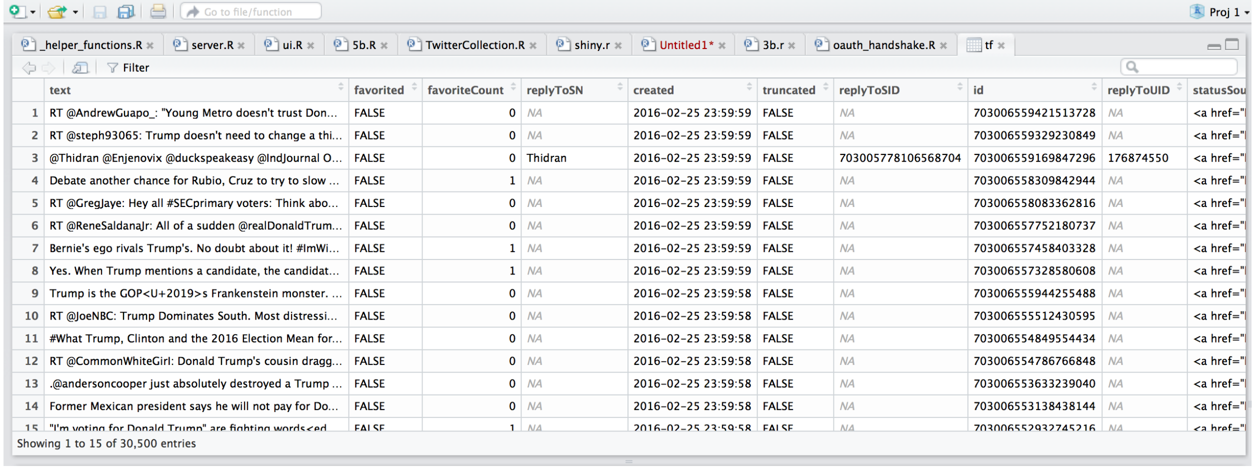
Introduction

In this project, exploratory data analysis (EDA) was conducted on large, structured and semi-structured datasets. With the help of EDA, important analysis about data and its relation with various parameters was studied. The different types of datasets analyses were twitter data set, New York Times ad click dataset, Real Direct dataset and live twitter data.

Problem 1: Data Acquisition

For this problem, the data source used was Twitter data using Twitter Search API, collected over a period of ten days. Twitter Streaming API is used in Problem 5. TwitteR and ROAuth were the packages that were used to carry out twitter data collection. TwitteR provided functions like setup_twitter_oauth which creates a connection to Twitter's Search API. The resulting list of retrieved tweets is stored in a data frame by using the function twListToDF(). Over a period of ten days, I created different data frames that stored that day's tweets. At the end, I combined all these data frames into one unit and stored this collective data as both cvs file and Rdata by using write.cv and save commands respectively.
tf is the local data frame containing collected tweets. These are the column variables it has.

```
> colnames(tf)
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"     "created"       "truncated"     "replyToSID"
 [8] "id"            "replyToUID"    "statusSource"  "screenName"    "retweetCount"  "isRetweet"     "retweeted"
[15] "longitude"     "latitude"
>
```



	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSou
1	RT @AndrewGuapo_: "Young Metro doesn't trust Don…	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559421513728	NA	<a href="
2	RT @steph93065: Trump doesn't need to change a thi…	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559329230849	NA	<a href="
3	@Thidran @Enjenovix @duckspeakeasy @IndJournal O…	FALSE	0	Thidran	2016-02-25 23:59:59	FALSE	703005778106568704	703006559169847296	176874550	<a href="
4	Debate another chance for Rubio, Cruz to try to slow …	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006558309842944	NA	<a href="
5	RT @GregJaye: Hey all #SECprimary voters: Think abo…	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006558083362816	NA	<a href="
6	RT @ReneSaldanaJr: All of a sudden @realDonaldTrum…	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006557752180737	NA	<a href="
7	Bernie's ego rivals Trump's. No doubt about it! #ImWi…	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557458403328	NA	<a href="
8	Yes. When Trump mentions a candidate, the candidat…	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557328580608	NA	<a href="
9	Trump is the GOP<U+2019>s Frankenstein monster. …	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555944255488	NA	<a href="
10	RT @JoeNBC: Trump Dominates South. Most distressi…	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555512430595	NA	<a href="
11	#What Trump, Clinton and the 2016 Election Mean for…	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554849554434	NA	<a href="
12	RT @CommonWhiteGirl: Donald Trump's cousin dragg…	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554786766848	NA	<a href="
13	.@andersoncooper just absolutely destroyed a Trump …	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553633239040	NA	<a href="
14	Former Mexican president says he will not pay for Do…	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553138438144	NA	<a href="
15	"I'm voting for Donald Trump" are fighting words<ed	FALSE	1	NA	2016-02-25 23:59:58	FALSE	NA	703006552932745216	NA	<a href="

Showing 1 to 15 of 30,500 entries

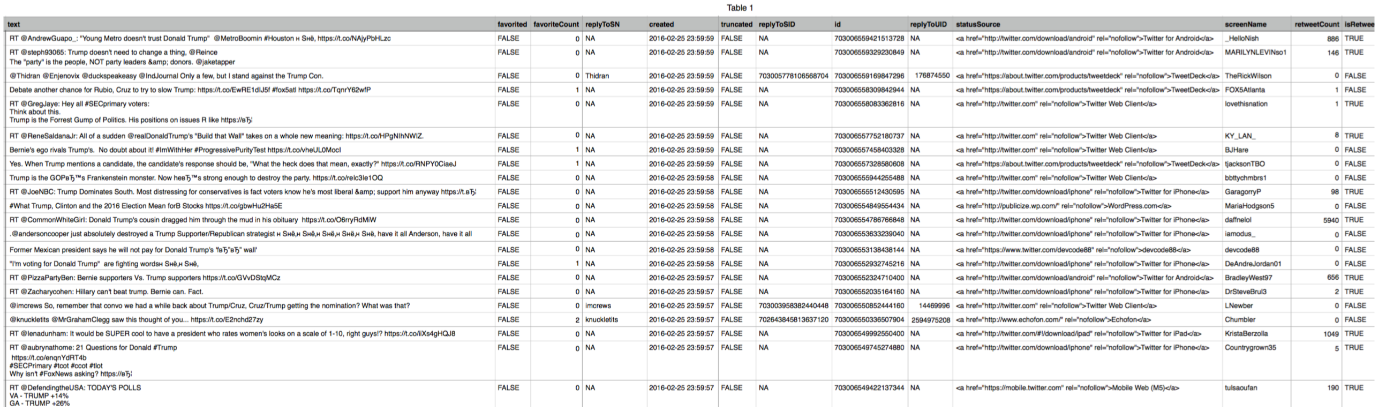


Table 1

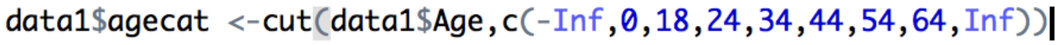
text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource	screenName	retweetCount	isRetwee
RT @AndrewGuapo_: "Young Metro doesn't trust Donald Trump" @MetroBoomin #Houston н Sнё, https://t.co/NAjyPbHLzc	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559421513728	NA	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	_HelloNish	886	TRUE
RT @steph93065: Trump doesn't need to change a thing, @Reince The "party" is the people, NOT party leaders &amp; donors. @jaketapper	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559329230849	NA	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	MARILYNLEVINso1	146	TRUE
@Thidran @Enjenovix @duckspeakeasy @IndJournal Only a few, but I stand against the Trump Con.	FALSE	0	Thidran	2016-02-25 23:59:59	FALSE	703005778106568704	703006559169847296	176874550	<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>	TheRickWilson	0	FALSE
Debate another chance for Rubio, Cruz to try to slow Trump: https://t.co/EwRE1dIJ5f #fox5atl https://t.co/TqnrY62wfP	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006558309842944	NA	<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>	FOX5Atlanta	1	FALSE
RT @GregJaye: Hey all #SECprimary voters: Think about this. Trump is the Forrest Gump of Politics. His positions on issues R like https://вЂ¦	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006558083362816	NA	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	lovethisnation	1	TRUE
RT @ReneSaldanaJr: All of a sudden @realDonaldTrump's "Build that Wall" takes on a whole new meaning: https://t.co/HPgNIhNWIZ.	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006557752180737	NA	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	KY_LAN_	8	TRUE
Bernie's ego rivals Trump's. No doubt about it! #ImWithHer #ProgressivePurityTest https://t.co/vheUL0MocI	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557458403328	NA	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	BJHare	0	FALSE
Yes. When Trump mentions a candidate, the candidate's response should be, "What the heck does that mean, exactly?" https://t.co/RNPY0CiaeJ	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557328580608	NA	<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>	tjacksonTBO	0	FALSE
Trump is the GOPвЂ™s Frankenstein monster. Now heвЂ™s strong enough to destroy the party. https://t.co/relc3le1OQ	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555944255488	NA	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	bbttychmbrs1	0	FALSE
RT @JoeNBC: Trump Dominates South. Most distressing for conservatives is fact voters know he's most liberal &amp; support him anyway https://t.вЂ¦	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555512430595	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	GaragorryP	98	TRUE
#What Trump, Clinton and the 2016 Election Mean forB Stocks https://t.co/gbwHu2Ha5E	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554849554434	NA	<a href="http://publicize.wp.com/" rel="nofollow">WordPress.com</a>	MariaHodgson5	0	FALSE
RT @CommonWhiteGirl: Donald Trump's cousin dragged him through the mud in his obituary https://t.co/O6rryRdMiW	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554786766848	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	daffnelol	5940	TRUE
.@andersoncooper just absolutely destroyed a Trump Supporter/Republican strategist н Sнё,н Sнё,н Sнё,н Sнё,н Sнё, have it all Anderson, have it all	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553633239040	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	iamodus_	0	FALSE
Former Mexican president says he will not pay for Donald Trump's 'fвЂ"вЂ" wall'	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553138438144	NA	<a href="https://www.twitter.com/devcode88" rel="nofollow">devcode88</a>	devcode88	0	FALSE
"I'm voting for Donald Trump" are fighting wordsн Sнё,н Sнё,	FALSE	1	NA	2016-02-25 23:59:58	FALSE	NA	703006552932745216	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	DeAndreJordan01	0	FALSE
RT @PizzaPartyBen: Bernie supporters Vs. Trump supporters https://t.co/GVvDStqMCz	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006552324710400	NA	<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	BradleyWest97	656	TRUE
RT @Zacharycohen: Hillary can't beat trump. Bernie can. Fact.	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006552035164160	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	DrSteveBrul3	2	TRUE
@imcrews So, remember that convo we had a while back about Trump/Cruz, Cruz/Trump getting the nomination? What was that?	FALSE	0	imcrews	2016-02-25 23:59:57	FALSE	703003958382440448	703006550852444160	14469996	<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>	LNewber	0	FALSE
@knuckletits @MrGrahamClegg saw this thought of you... https://t.co/E2nchd27zy	FALSE	2	knuckletits	2016-02-25 23:59:57	FALSE	702643845813637120	703006550336507904	2594975208	<a href="http://www.echofon.com/" rel="nofollow">Echofon</a>	Chumbler	0	FALSE
RT @lenadunham: It would be SUPER cool to have a president who rates women's looks on a scale of 1-10, right guys!? https://t.co/iiXs4gHQJ8	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006549992550400	NA	<a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad</a>	KristaBerzolla	1049	TRUE
RT @aubrynathome: 21 Questions for Donald #Trump https://t.co/enqnYdRT4b #SECPrimary #tcot #ccot #tlot Why isn't #FoxNews asking? https://вЂ¦	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006549745274880	NA	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	Countrygrown35	5	TRUE
RT @DefendingtheUSA: TODAY'S POLLS VA - TRUMP +14% GA - TRUMP +26%	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006549422137344	NA	<a href="https://mobile.twitter.com" rel="nofollow">Mobile Web (M5)</a>	tulsaoufan	190	TRUE

This is the cvs file with tweets, stored on disk.

Problem 2: Simple EDA

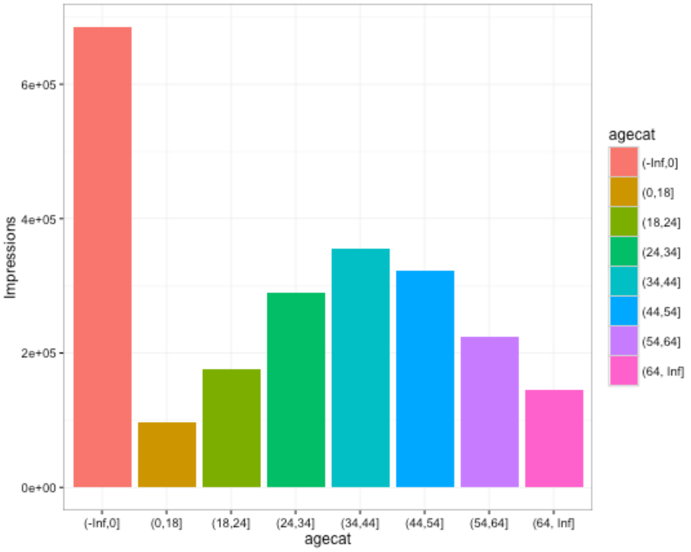
In this problem, we were provided with the New York Times dataset which had the information about the ad clicking behavior of readers with attributes such as age, gender, impressions, number of clicks, signed in or not.

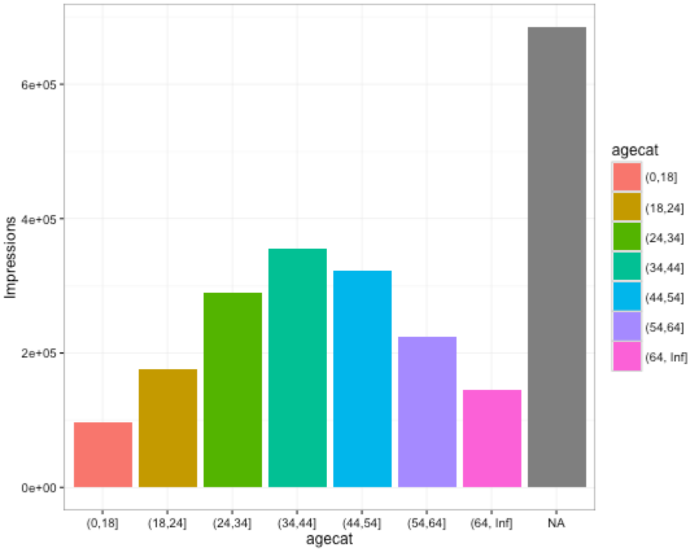
First part of the problem is to categorize the age variable into intervals of age.



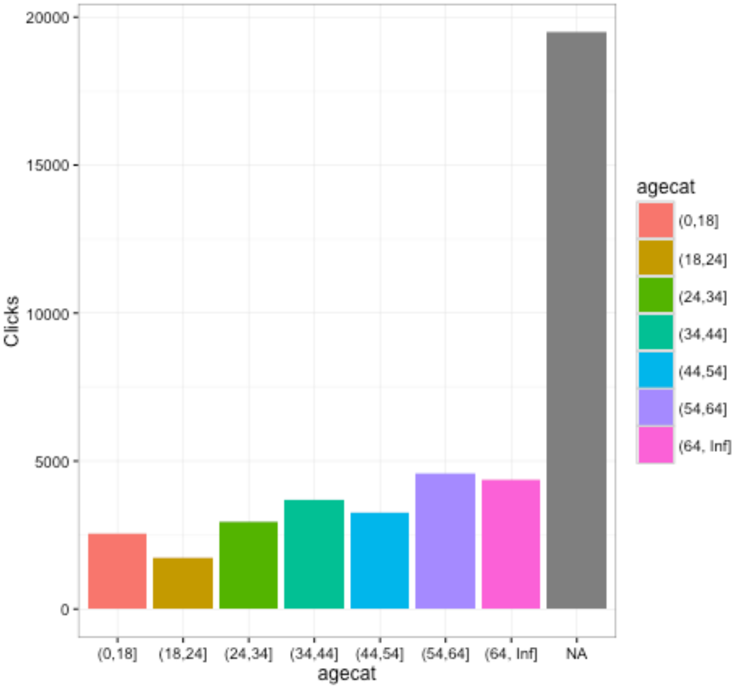
```
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
```

Next, plotting impressions and click-through-rates for this interval of age. Here, upon noticing a high value for(Inf, 0], I realized that these users were not logged in. So setting the signed in value for these as NA and generated the bar plot again.



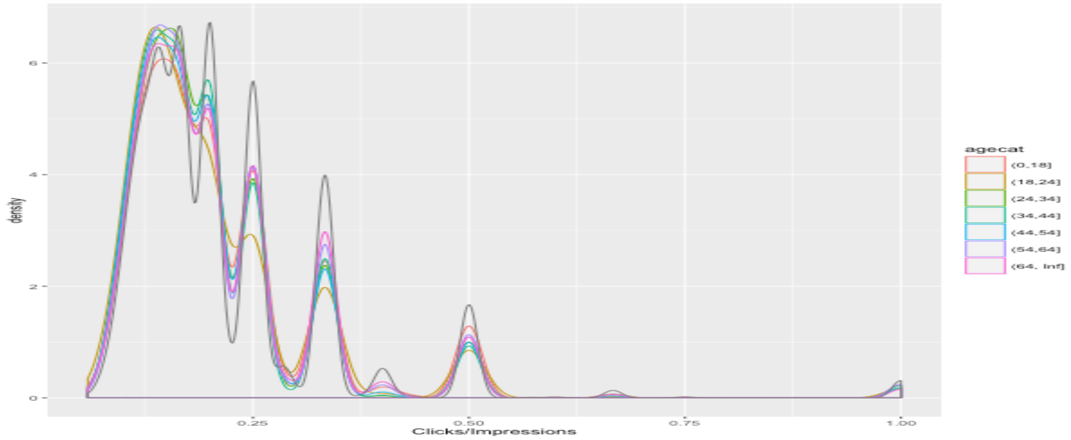


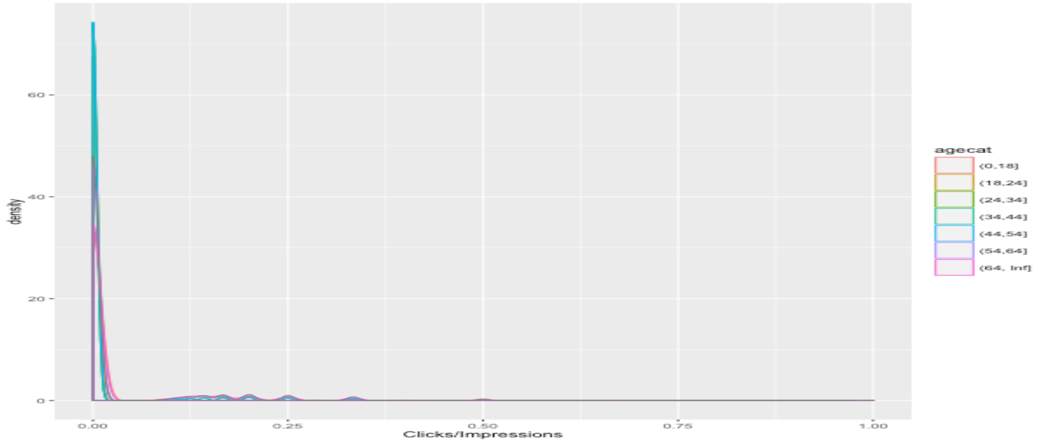
As is seen from the figure, the bar plot shows the number of impressions for the age intervals which is highest for age group (34,44].



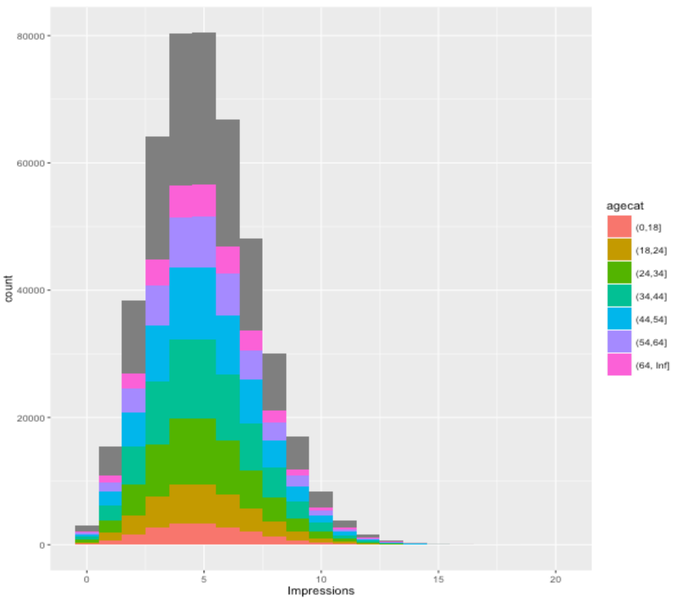
This figure shows the number of clicks for each age group. It can be inferred that people from age group (54,64] appeared to click on the ads the most, followed by (64,Inf] (not counting the unsigned users). This shows that elderly people are more interested in (or easily distracted by) online advertisements. Similarly age group (18,24] seemed to click on ads the least which is probably because most of the people in this age group are students or young working professionals who might not spend much time clicking on ads.

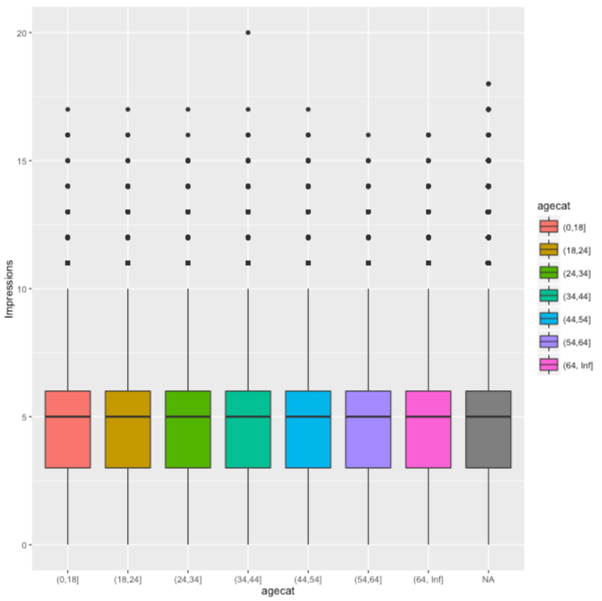
Following shows the density distribution of CTR with respect to age. First figure is for Clicks>0 and second figure is for Impressions>0.

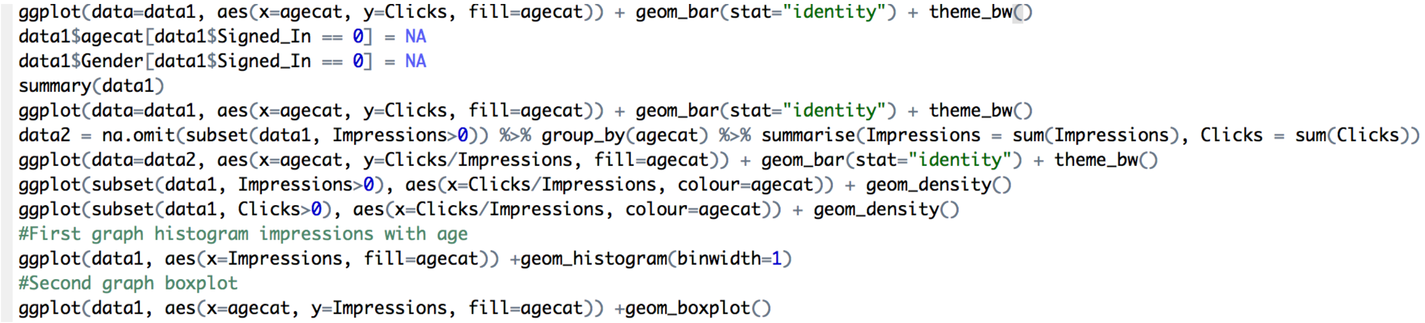




Following are a histogram and box plot that show the relation between CTR and age group.

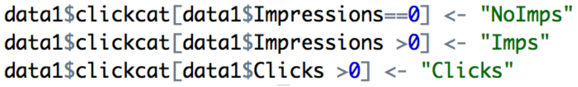






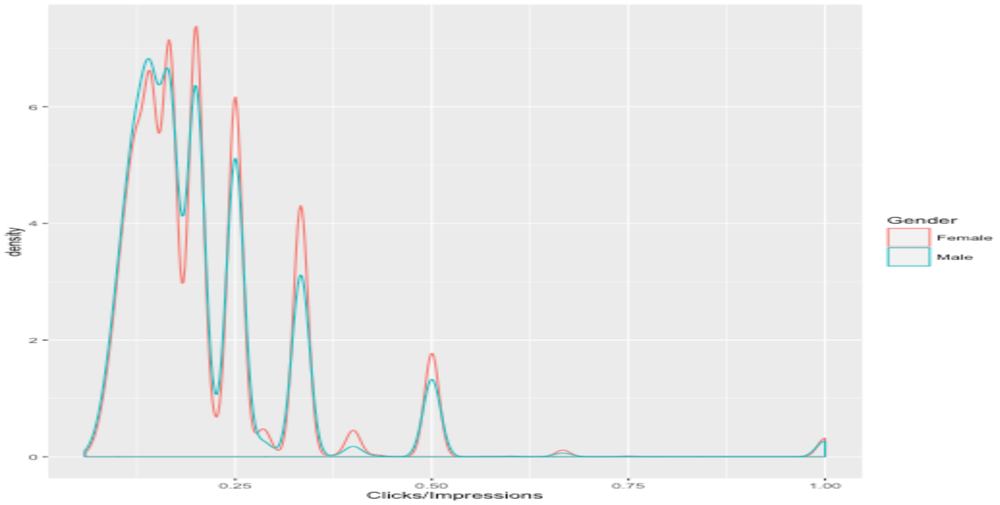
```
ggplot(data=data1, aes(x=agecat, y=Clicks, fill=agecat)) + geom_bar(stat="identity") + theme_bw()
data1$agecat[data1$Signed_In == 0] = NA
data1$Gender[data1$Signed_In == 0] = NA
summary(data1)
ggplot(data=data1, aes(x=agecat, y=Clicks, fill=agecat)) + geom_bar(stat="identity") + theme_bw()
data2 = na.omit(subset(data1, Impressions>0)) %>% group_by(agecat) %>% summarise(Impressions = sum(Impressions), Clicks = sum(Clicks))
ggplot(data=data2, aes(x=agecat, y=Clicks/Impressions, fill=agecat)) + geom_bar(stat="identity") + theme_bw()
ggplot(subset(data1, Impressions>0), aes(x=Clicks/Impressions, colour=agecat)) + geom_density()
ggplot(subset(data1, Clicks>0), aes(x=Clicks/Impressions, colour=agecat)) + geom_density()
#First graph histogram impressions with age
ggplot(data1, aes(x=Impressions, fill=agecat)) +geom_histogram(binwidth=1)
#Second graph boxplot
ggplot(data1, aes(x=agecat, y=Impressions, fill=agecat)) +geom_boxplot()
```

Next, a new variable "clickcat" was defined to differentiate between users based on their click behavior which would hold three values: No Impressions, Only Impressions(No Click) and Clicks.

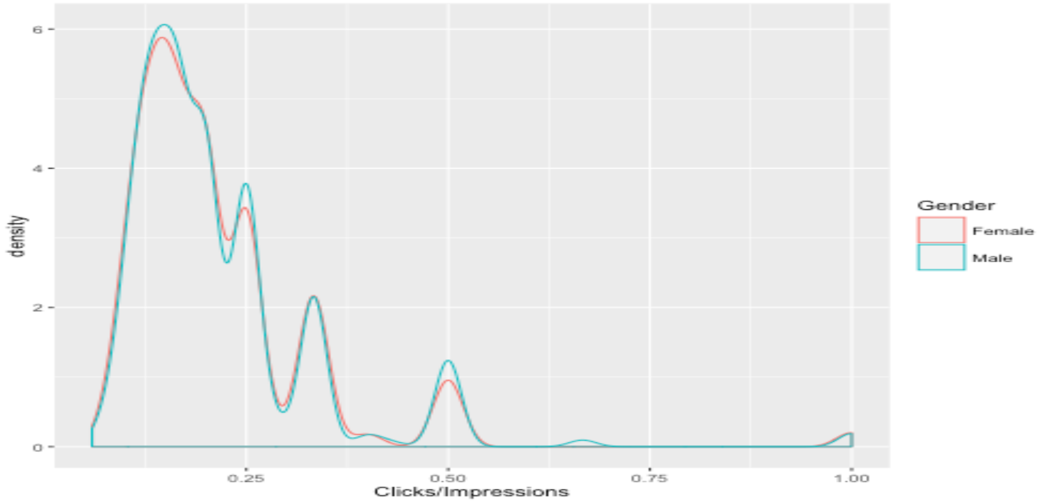


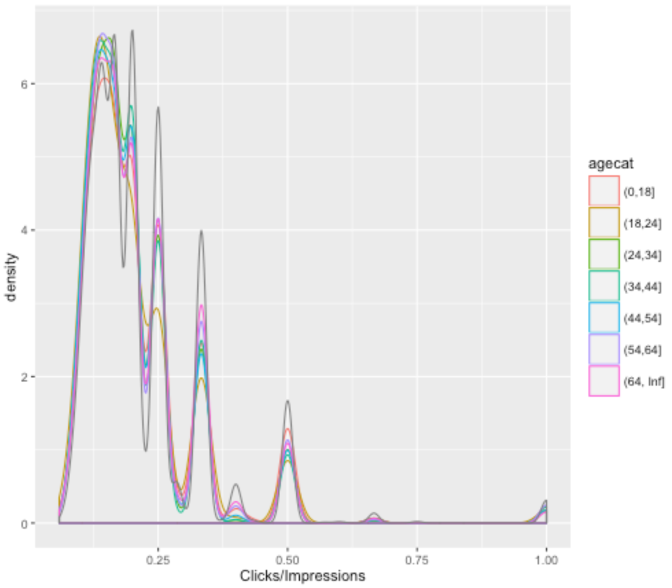
```
data1$clickcat[data1$Impressions==0] <- "NoImps"
data1$clickcat[data1$Impressions >0] <- "Imps"
data1$clickcat[data1$Clicks >0] <- "Clicks"
```

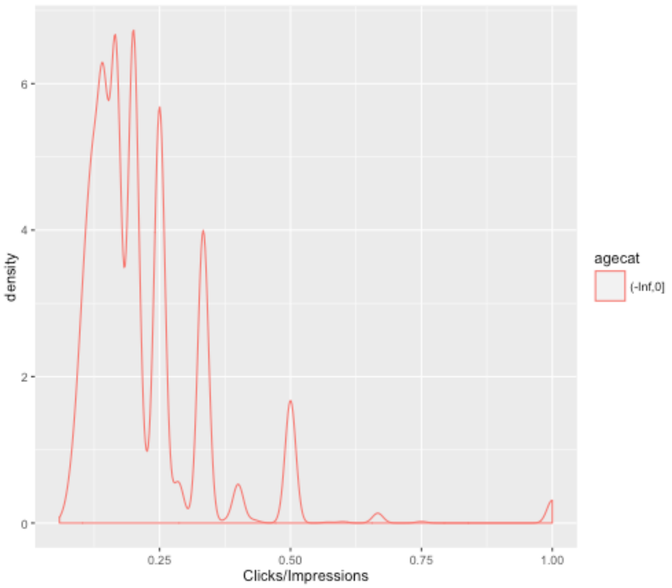
Next part of the problem was to explore the data and make visual and quantitative comparisons across user segments. Here are some observations. This following graph shows a normal density distribution graph for click through rate patterns of males and females of all age groups



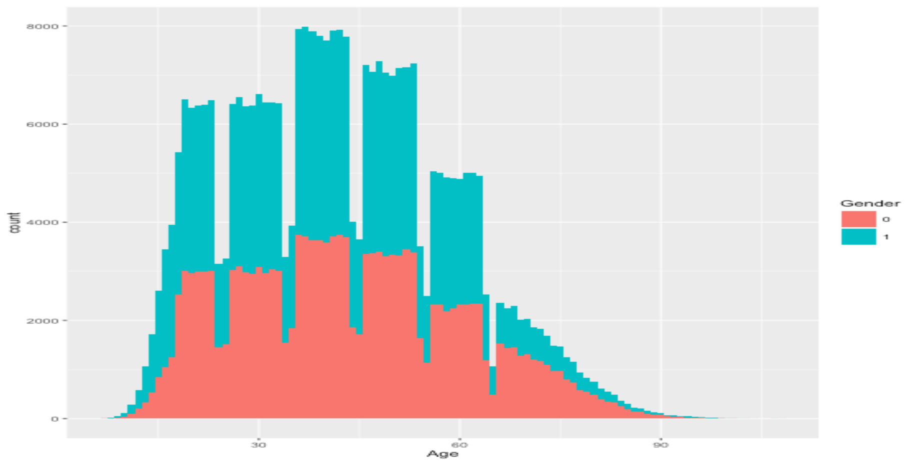
The following graph compares the Click Through Rate patterns of less than 18 year old male & females.





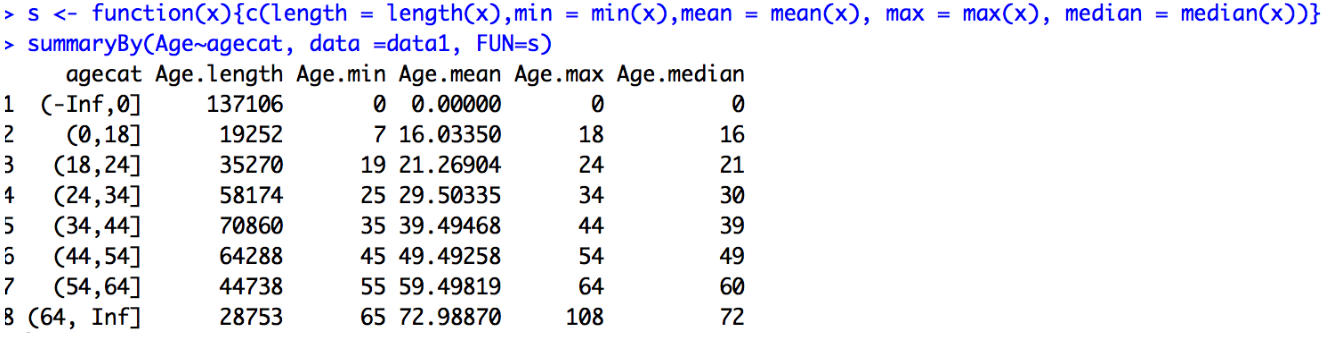


In the above graphs, figure 1 shows the density distribution for logged in users and figure 2 shows density distribution for not logged in users.

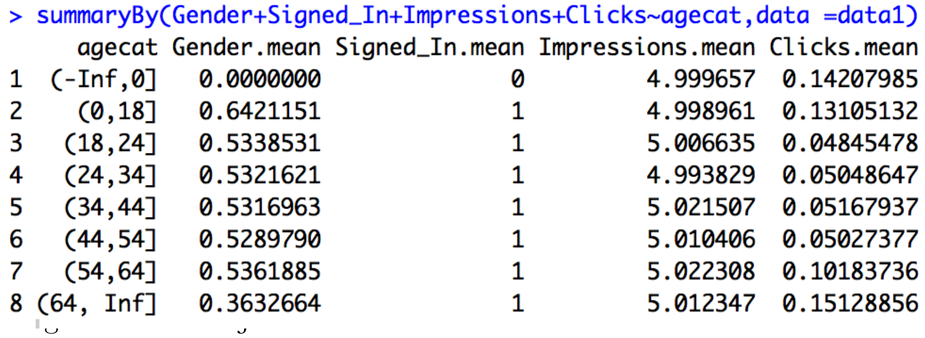


Here, there is a normal distribution of age with males usually twice the number of females except for >65 where difference starts to reduce.

Now looking at some statistics.

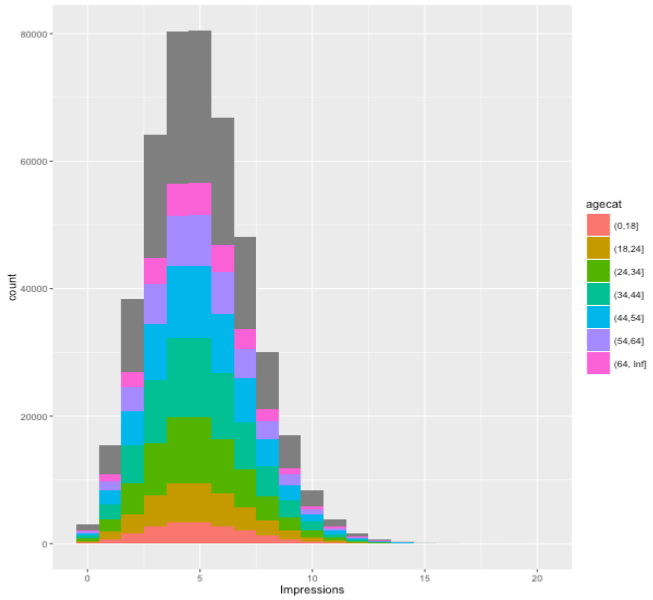


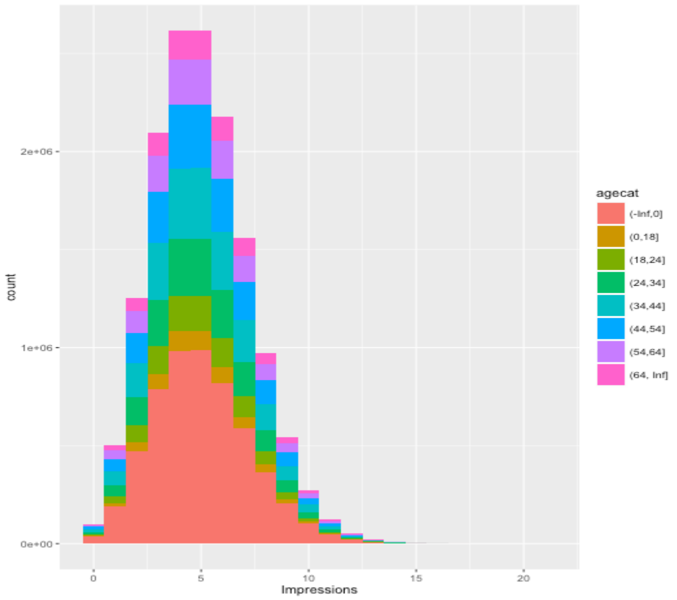
```
> s <- function(x){c(length = length(x),min = min(x),mean = mean(x), max = max(x), median = median(x))}
> summaryBy(Age~agecat, data =data1, FUN=s)
     agecat Age.length Age.min Age.mean Age.max Age.median
1  (-Inf,0]     137106       0  0.00000       0          0
2    (0,18]      19252       7 16.03350      18         16
3   (18,24]      35270      19 21.26904      24         21
4   (24,34]      58174      25 29.50335      34         30
5   (34,44]      70860      35 39.49468      44         39
6   (44,54]      64288      45 49.49258      54         49
7   (54,64]      44738      55 59.49819      64         60
8 (64, Inf]      28753      65 72.98870     108         72
```



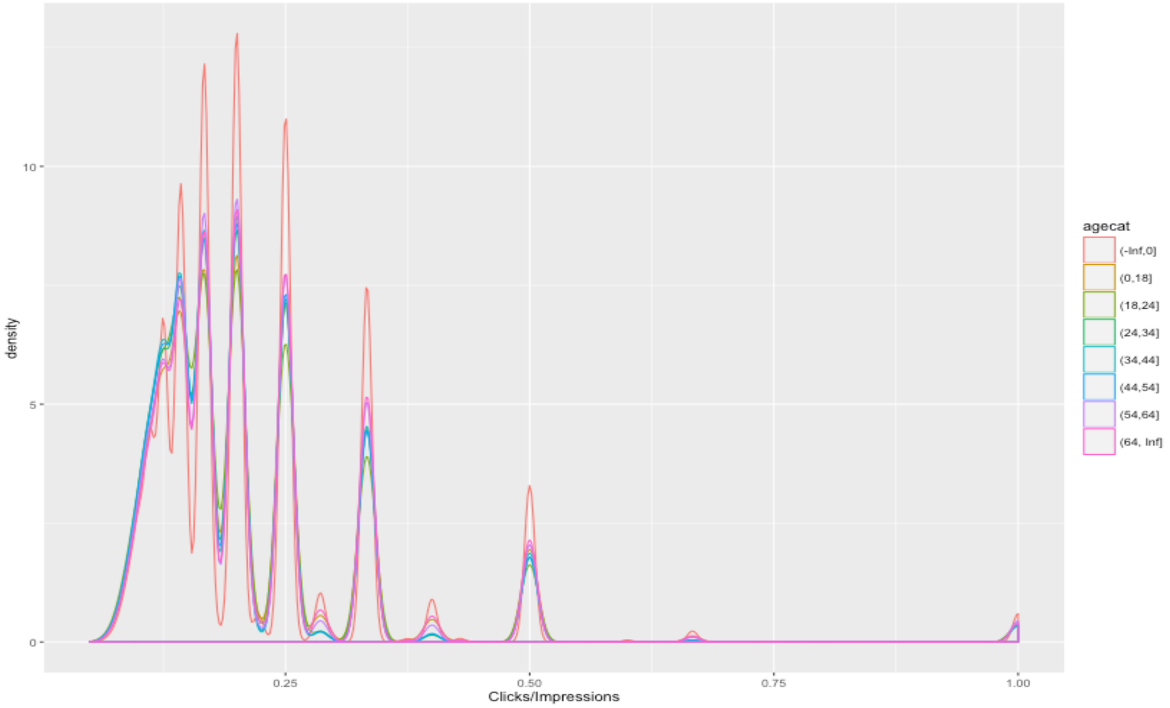
```
> summaryBy(Gender+Signed_In+Impressions+Clicks~agecat,data =data1)
     agecat Gender.mean Signed_In.mean Impressions.mean Clicks.mean
1  (-Inf,0]   0.0000000              0         4.999657  0.14207985
2    (0,18]   0.6421151              1         4.998961  0.13105132
3   (18,24]   0.5338531              1         5.006635  0.04845478
4   (24,34]   0.5321621              1         4.993829  0.05048647
5   (34,44]   0.5316963              1         5.021507  0.05167937
6   (44,54]   0.5289790              1         5.010406  0.05027377
7   (54,64]   0.5361885              1         5.022308  0.10183736
8 (64, Inf]   0.3632664              1         5.012347  0.15128856
```

Last part of the problem, was to extend the findings and calculation analysis from one day to a month. As we can see from the graphs below, the data is normally distributed over the month too.





The following graph shows the density distribution over 30 days. It gives a click through rate pattern of users with respect to age over a period of 30 days.



Falguni Bharadwaj