Problem 2
Part A

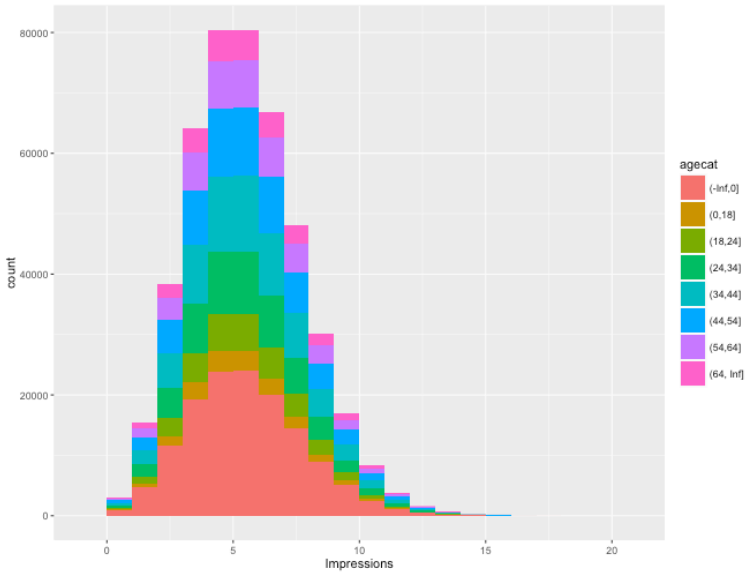
Problem Description:

In this part of the problem we analyze the "nyt1.csv" dataset and try to find patterns. The data is downloaded from the URL "http://stat.columbia.edu/~rachel/datasets/nyt1.csv" using the read.csv() function in R. After reading the data into a data frame in R we first categorize each row based on the age using the cut() function. Each row is also divided based on the number of Impressions and the Click behavior.

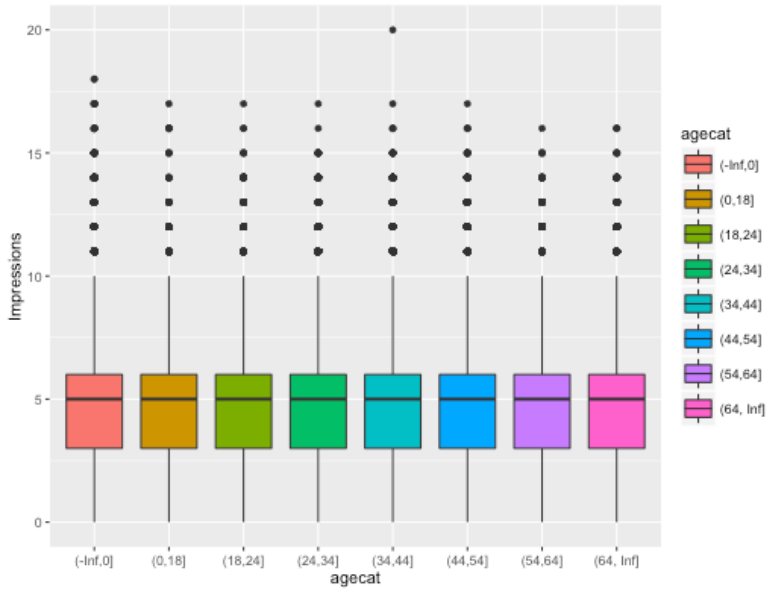
Analysis of Plots and Summaries:

	agecat	Age.FUN1	Age.FUN2	Age.FUN3	Age.FUN4
1	(-Inf,0]	137106	0	0	0.00000
2	(0,18]	19252	7	18	16.03350
3	(18,24]	35270	19	24	21.26904
4	(24,34]	58174	25	34	29.50335
5	(34,44]	70860	35	44	39.49468
6	(44,54]	64288	45	54	49.49258
7	(54,64]	44738	55	64	59.49819
8	(64, Inf]	28753	65	108	72.98870

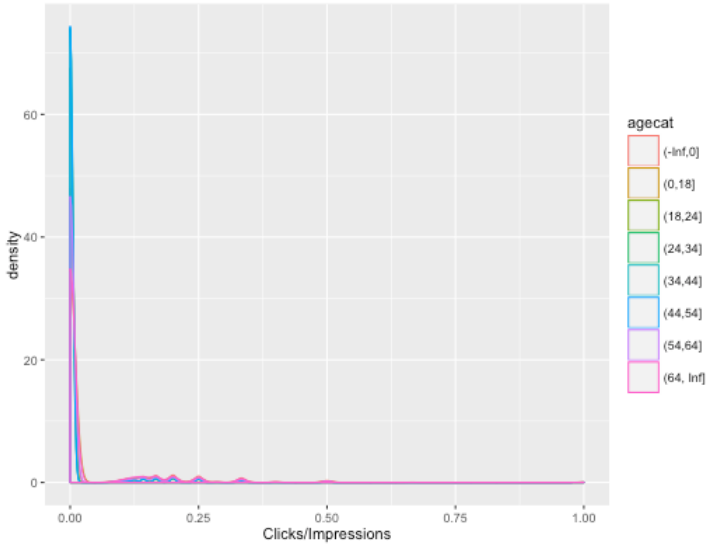
The table above gives us information about the number of entries in each age category. The highest number of entries has a value of 0 set as age that indicates that these users did not enter their age. The next highest number of entries is in the 34 - 44 age category. We can see that the oldest person is 108 years old and the youngest is 7 years old.



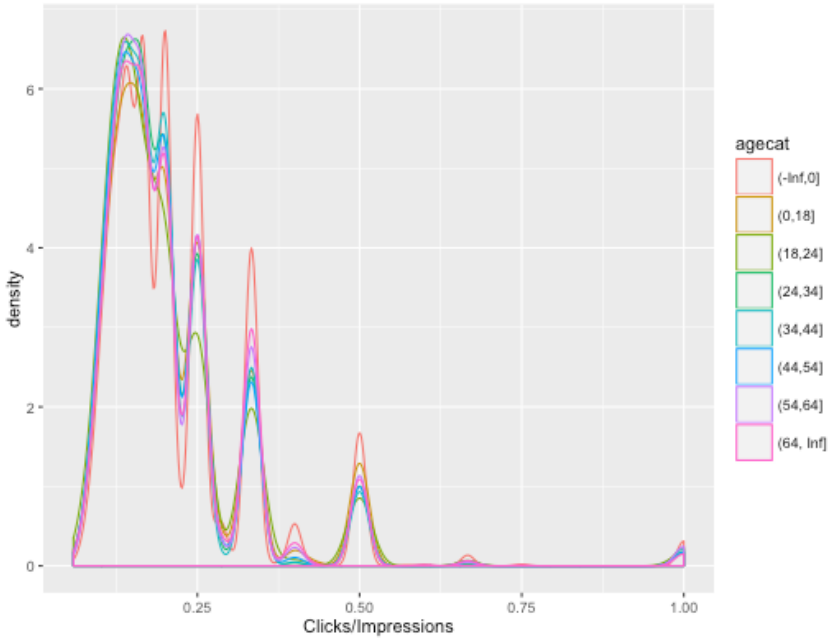
The histogram plot that has been superimposed by the age category above tells us that about 80000 people have seen an average of 5 impressions. It also tells us that the people who have not entered their age have received the maximum number of impressions.



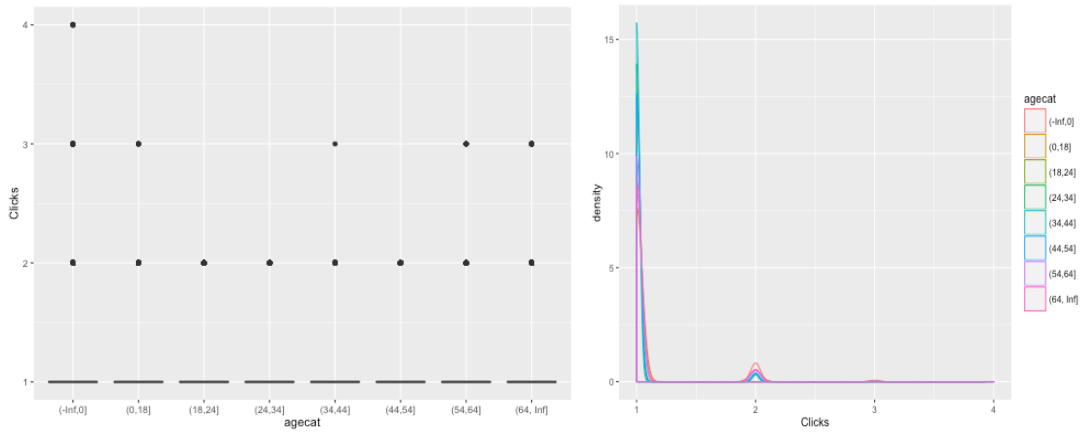
The boxplot above also tells us that the median for all age groups is 5 impressions. The highest number of impressions for a user was 20.



The above density plot is a visualization of all entries where Impressions are greater than 0. We can see that most points lie close to 0. This is because even though the number of impressions is positive, the number of clicks is 0.



This plot gives a better visualization of the click through rate with respect to the age categories compared to the previous plot. We can see that the density of click through rate is highest between 0 and 0.25 for all age categories.



The boxplot above doesn't give much information other than that the median clicks for all age categories is close to 1. The density plot beside gives us a better visualization of the clicks for each age category.

The summary tables below give us a lot of details about the different age categories, gender and number of Impressions. We can see that number of clicks is highest for the 64 and above age group in females with 2598 clicks. The highest number of impressions is for 34-44 year old males.

	scode	Gender	agecat	Impressions.clen
1	Clicks	0	(-Inf,0]	17776
2	Clicks	0	(0,18]	846
3	Clicks	0	(18,24]	779
4	Clicks	0	(24,34]	1361
5	Clicks	0	(34,44]	1675
6	Clicks	0	(44,54]	1494
7	Clicks	0	(54,64]	2006
8	Clicks	0	(64, Inf]	2598
9	Clicks	1	(0,18]	1525
10	Clicks	1	(18,24]	890
11	Clicks	1	(24,34]	1509
12	Clicks	1	(34,44]	1917
13	Clicks	1	(44,54]	1645
14	Clicks	1	(54,64]	2331
15	Clicks	1	(64, Inf]	1486

	scode	Gender	agecat	Impressions.clen
16	Imps	0	(-Inf,0]	118401
17	Imps	0	(0,18]	6001
18	Imps	0	(18,24]	15538
19	Imps	0	(24,34]	25690
20	Imps	0	(34,44]	31290
21	Imps	0	(44,54]	28563
22	Imps	0	(54,64]	18626
23	Imps	0	(64, Inf]	15585
24	Imps	1	(0,18]	10754
25	Imps	1	(18,24]	17807
26	Imps	1	(24,34]	29241
27	Imps	1	(34,44]	35512
28	Imps	1	(44,54]	32143
29	Imps	1	(54,64]	21499
30	Imps	1	(64, Inf]	8887

Part B

Problem Description:

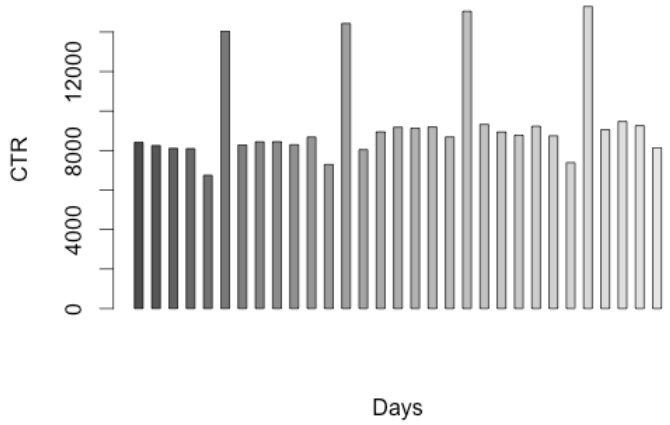
In the second part of the problem, we analyze data over a 31-day period and try to find patterns. The data was downloaded from "http://stat.columbia.edu/~rachel /datasets". The data is then loaded into a data frame to be analyzed in R. Like the previous part, we categorize the rows based on Age, Clicks and Impressions.

Analysis of Plots:

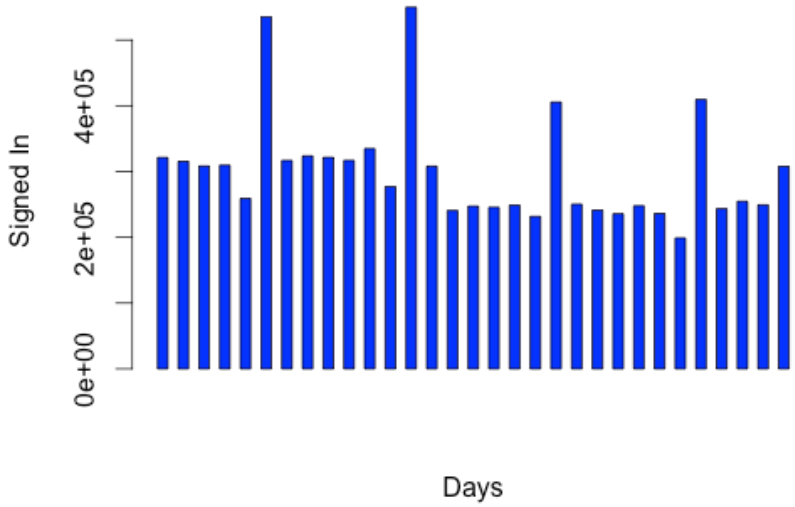
	scode	Gender	agecat	Impressions.clen
1	Clicks	0	(-Inf,0]	729705
2	Clicks	0	(0,18]	24085
3	Clicks	0	(18,24]	23626
4	Clicks	0	(24,34]	38768
5	Clicks	0	(34,44]	47452
6	Clicks	0	(44,54]	43354
7	Clicks	0	(54,64]	58780
8	Clicks	0	(64, Inf]	74417
9	Clicks	1	(0,18]	45007
10	Clicks	1	(18,24]	26146
11	Clicks	1	(24,34]	42974
12	Clicks	1	(34,44]	52481
13	Clicks	1	(44,54]	47640
14	Clicks	1	(54,64]	64902
15	Clicks	1	(64, Inf]	42234

	scode	Gender	agecat	Impressions.clen
16	Imps	0	(-Inf,0]	4846193
17	Imps	0	(0,18]	175876
18	Imps	0	(18,24]	457347
19	Imps	0	(24,34]	748257
20	Imps	0	(34,44]	912661
21	Imps	0	(44,54]	832274
22	Imps	0	(54,64]	552416
23	Imps	0	(64, Inf]	457395
24	Imps	1	(0,18]	308282
25	Imps	1	(18,24]	508218
26	Imps	1	(24,34]	832370
27	Imps	1	(34,44]	1018230
28	Imps	1	(44,54]	923724
29	Imps	1	(54,64]	614492
30	Imps	1	(64, Inf]	256559

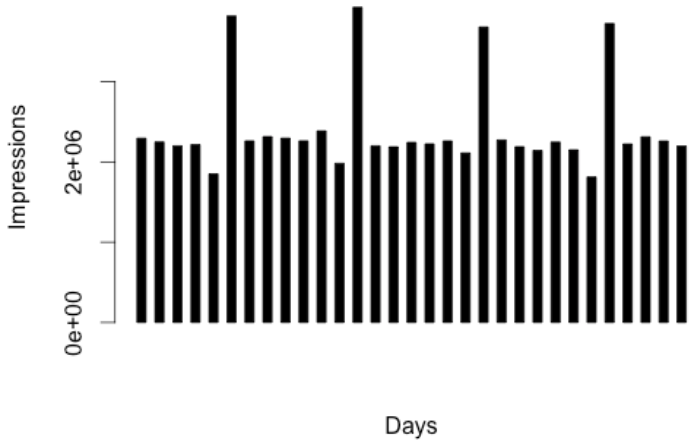
From the summary above we can see that, females aged 64 and above have the most number of clicks. The number of impressions is highest for males aged 34-44.



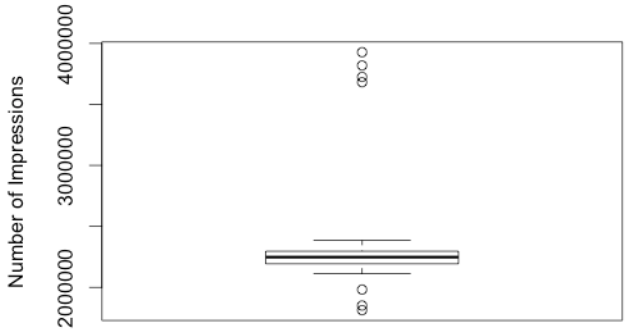
From the above bar plot we can see that the Click through rate peaks every 7 days. The calendar for May 2012 shows that the peaks occur on Sundays. Hence we can conclude that users click more on the ads on Sundays.



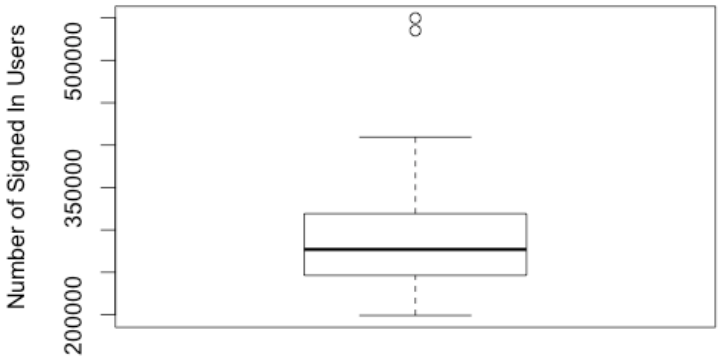
The plot above agrees with the previous plot and shows that most users login on Sundays.



This plot also shows that the number of Impressions is highest in Sundays. We can also see that the number of Clicks increased with increase in Impressions, hence the increase in CTR.

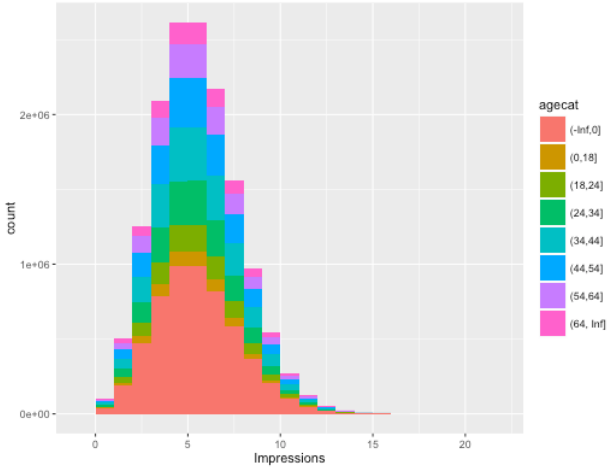


The boxplot above gives us more details about the impressions shown each day. The median number of impressions is 2247927.

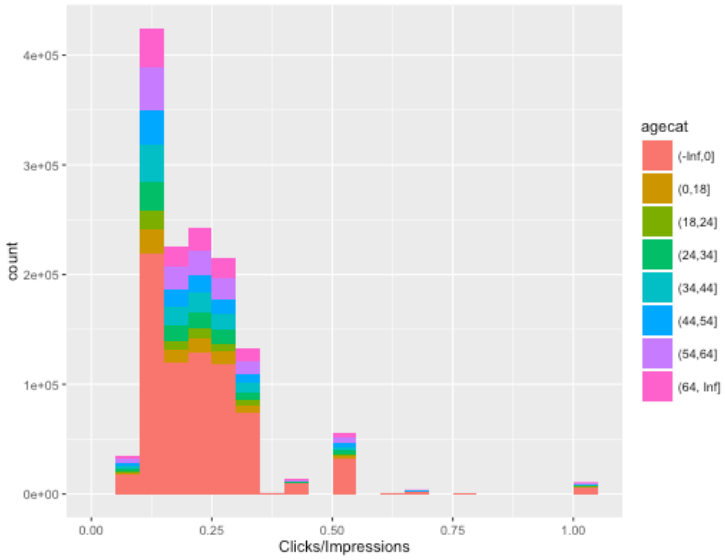


The boxplot above gives us information about the number of signed in users. The average number of signed in users on any day is 299751.

The histogram below gives a better visualization of the Impressions in different age groups.



The histogram below shows the CTR for different age groups. We can ignore the values beyond CTR>1 since for these values since the data is inconsistent.



The density plot below gives information about the click through rate for different age groups. The Click Through Rate lies mostly between 0 and 0.75.