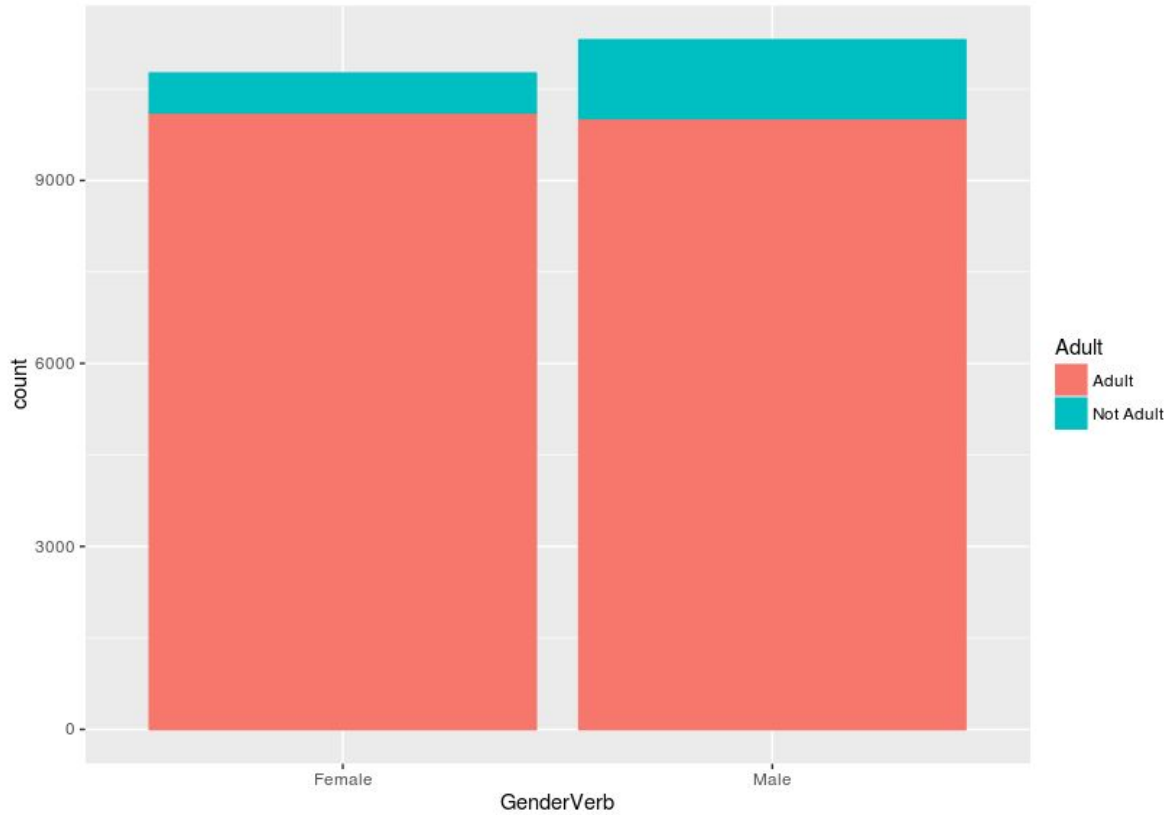
CSE 587 - Data Intensive Computing

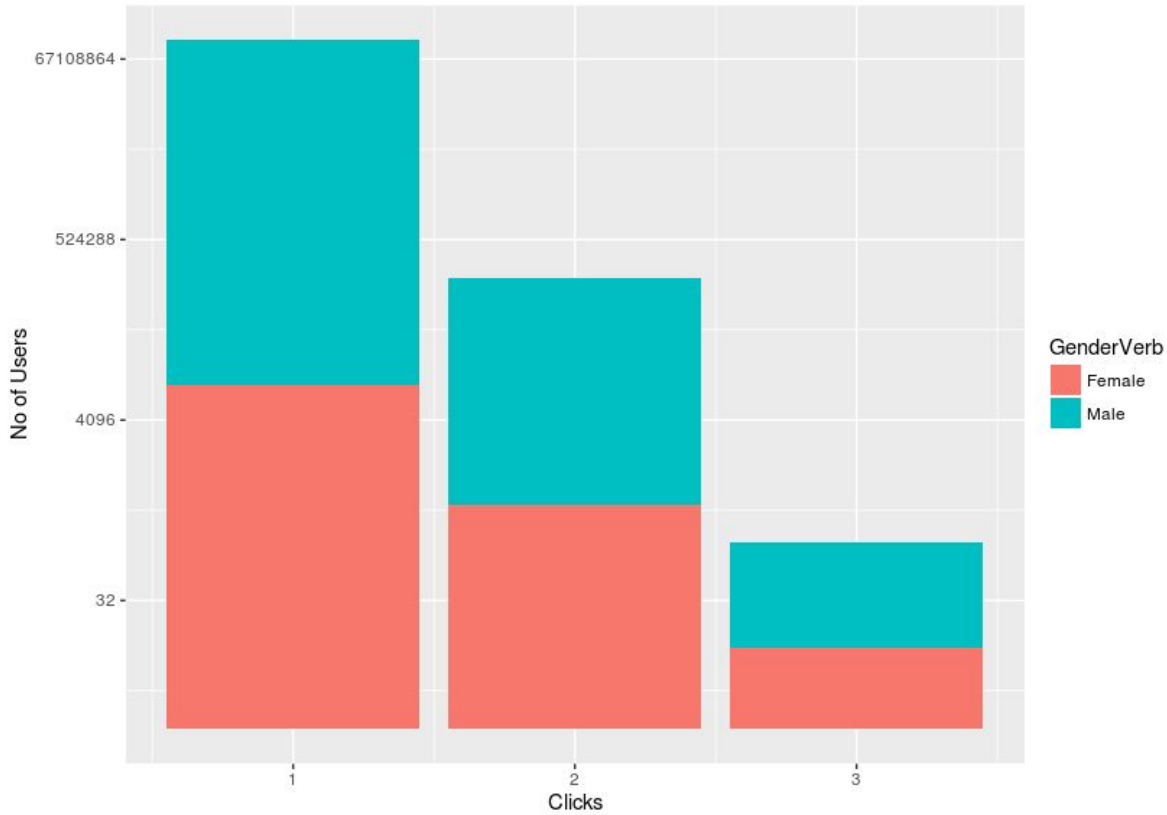
Problem 2 : Simple EDA

Goal: The goal of this problem is to analyze sample data from New York Times for a single day and multiple days using R .
Data Source : NYT data for 1 month
Data Format: CSV files
Period of Time: 2012 May

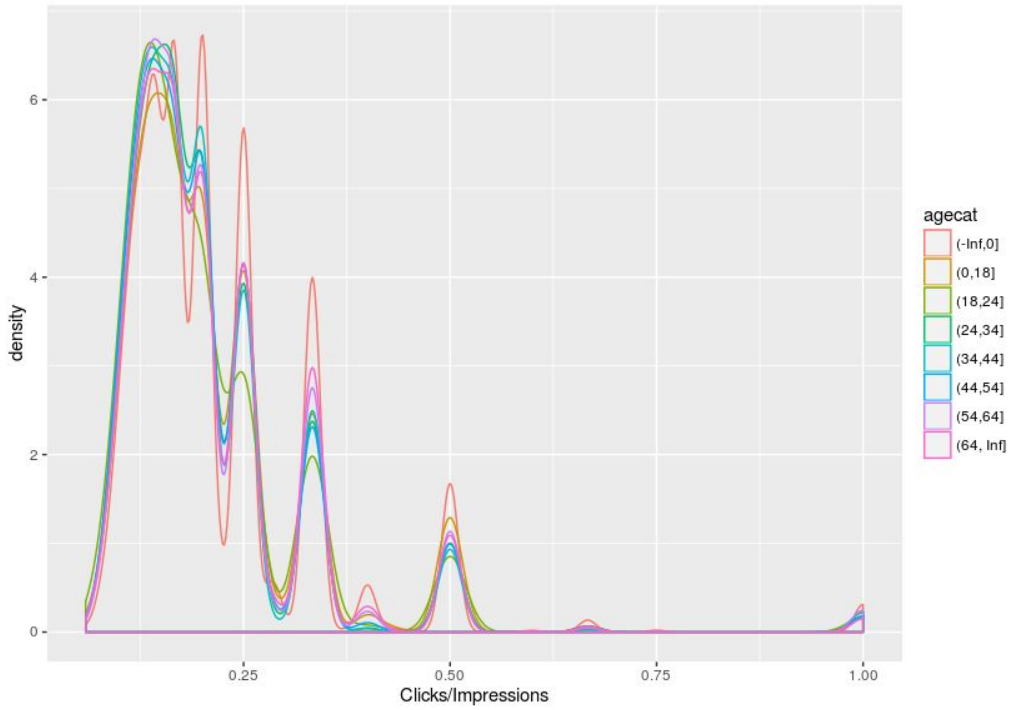
Part A : For Single Day - Following are the plots



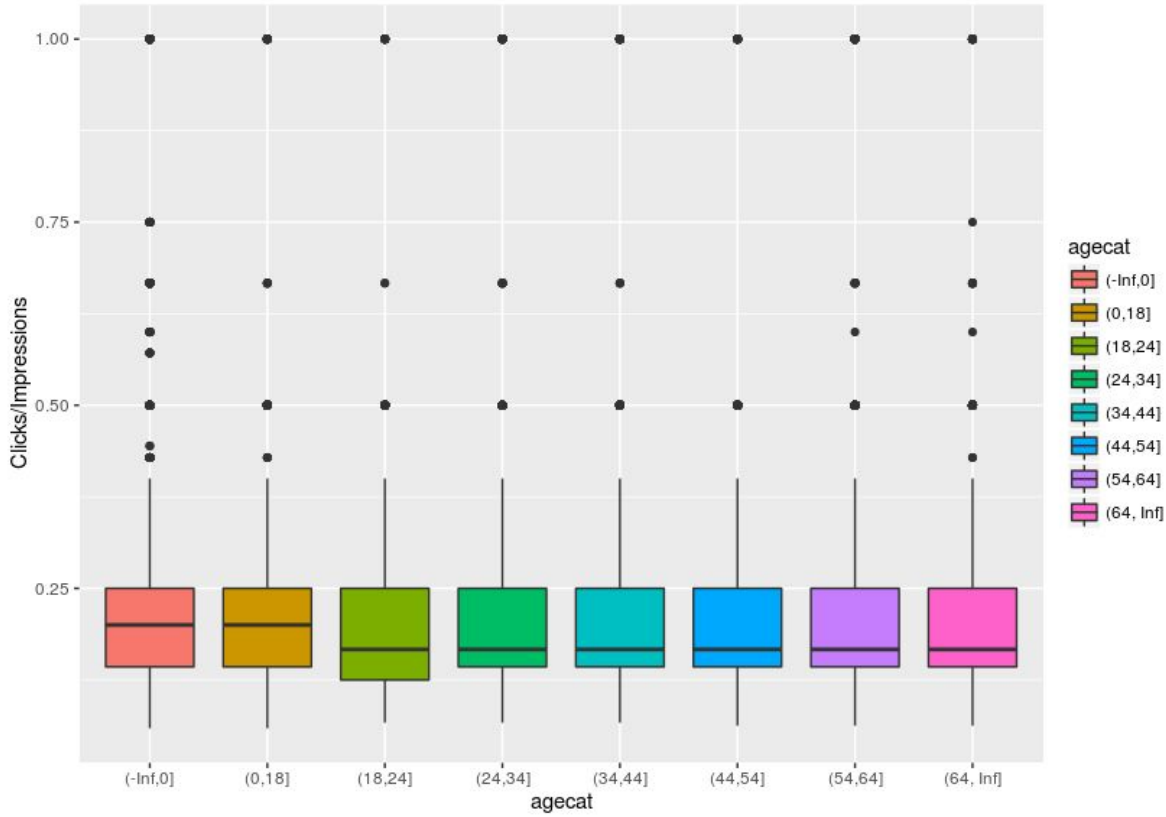
Above Plot gives the the count of males and females who are Adults(Age >= 18) and Not Adults (age < 18). The Plots implies that there more Adults in the data and the count of females are almost equal to males but slightly less.



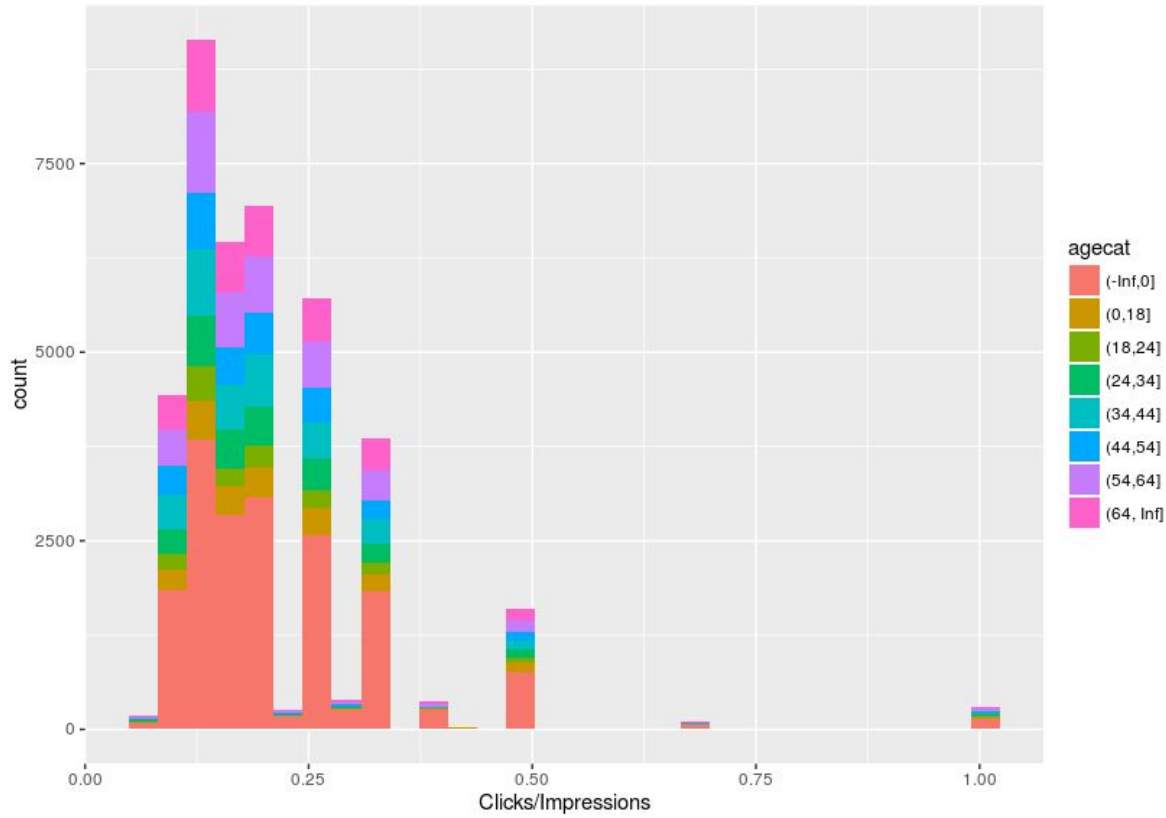
Above Plot gives the No of Users and their Clicks and also their Gender . It shows that there are more No of Users with Clicks = 1.



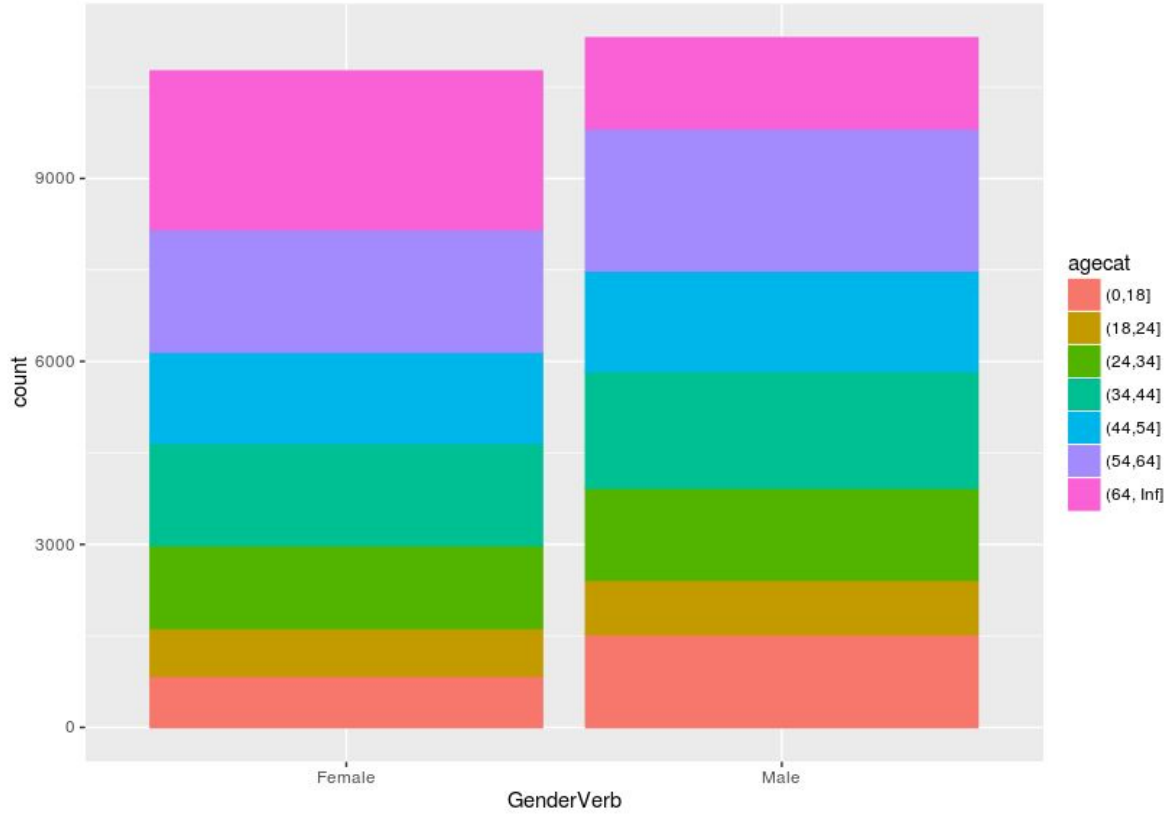
Above Plot gives the density Vs Clicks/Impression (CTR) plot with agecat filled.



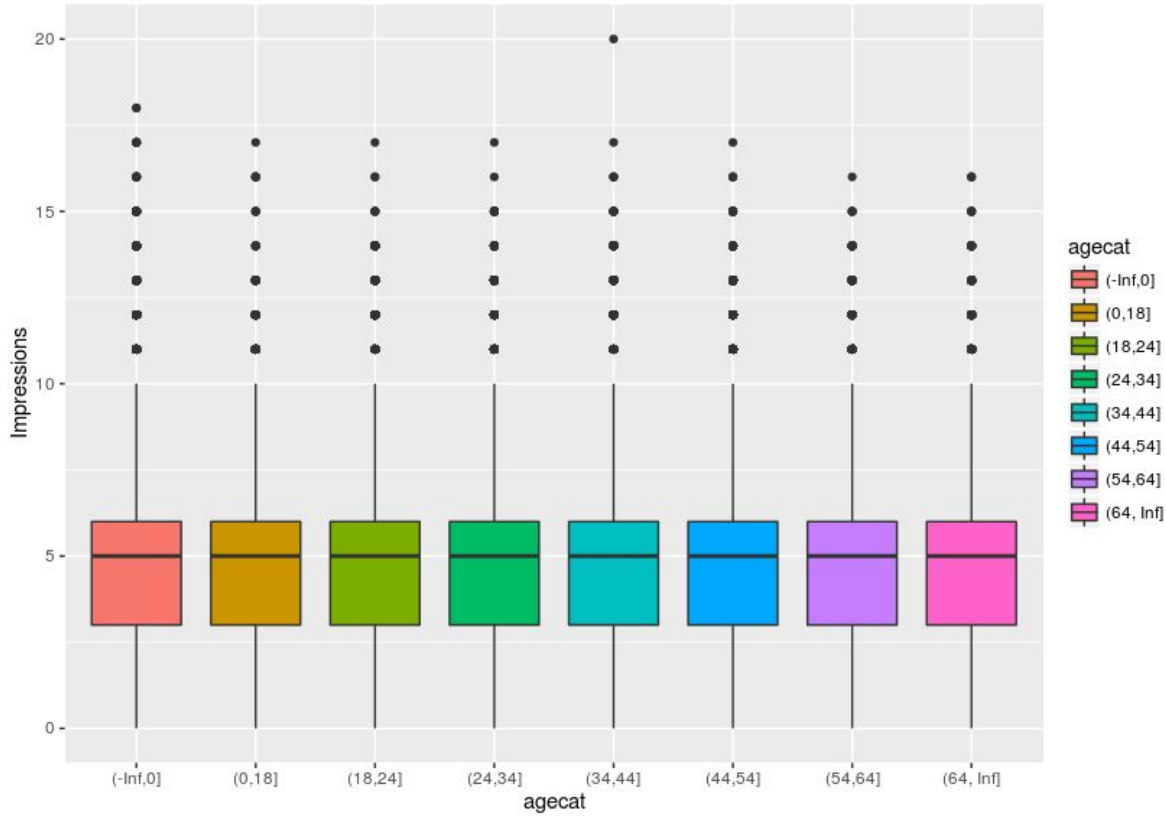
Above Box Plot gives the CTR and agecat analysis



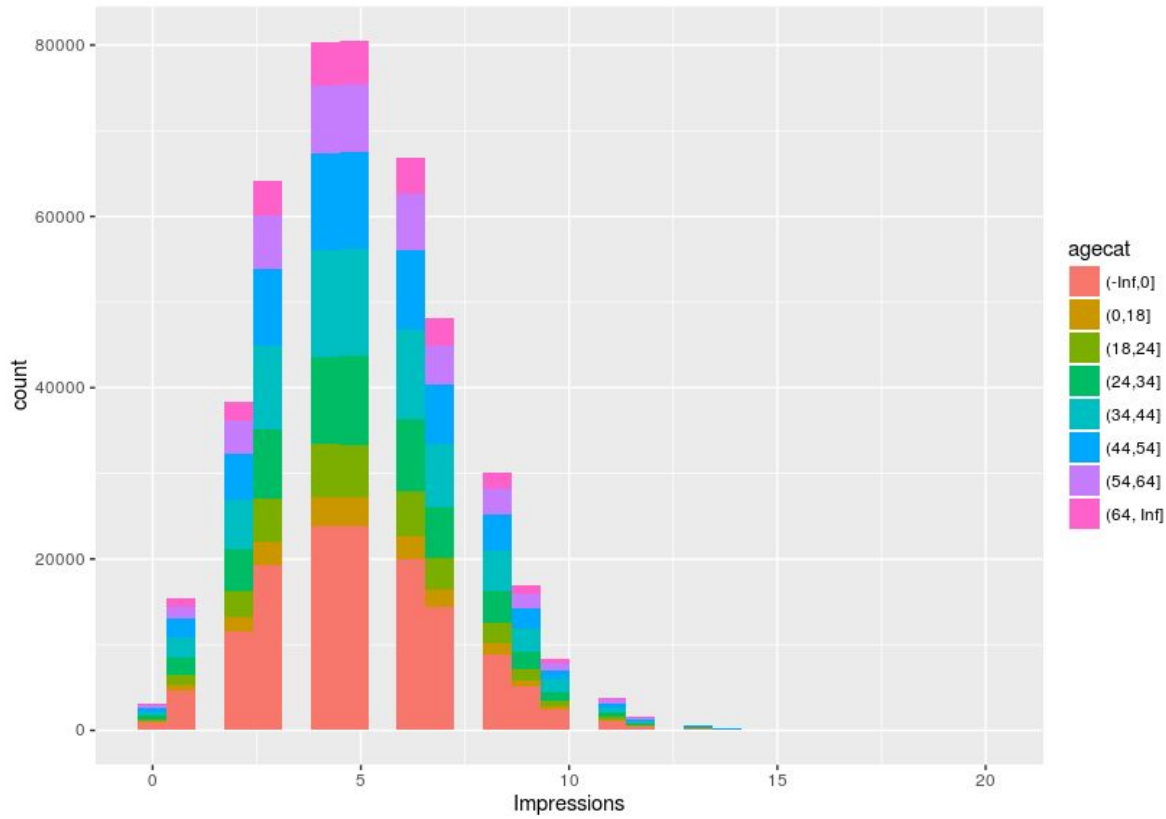
Above Plot gives the Analysis between the count of No of Users Vs CTR



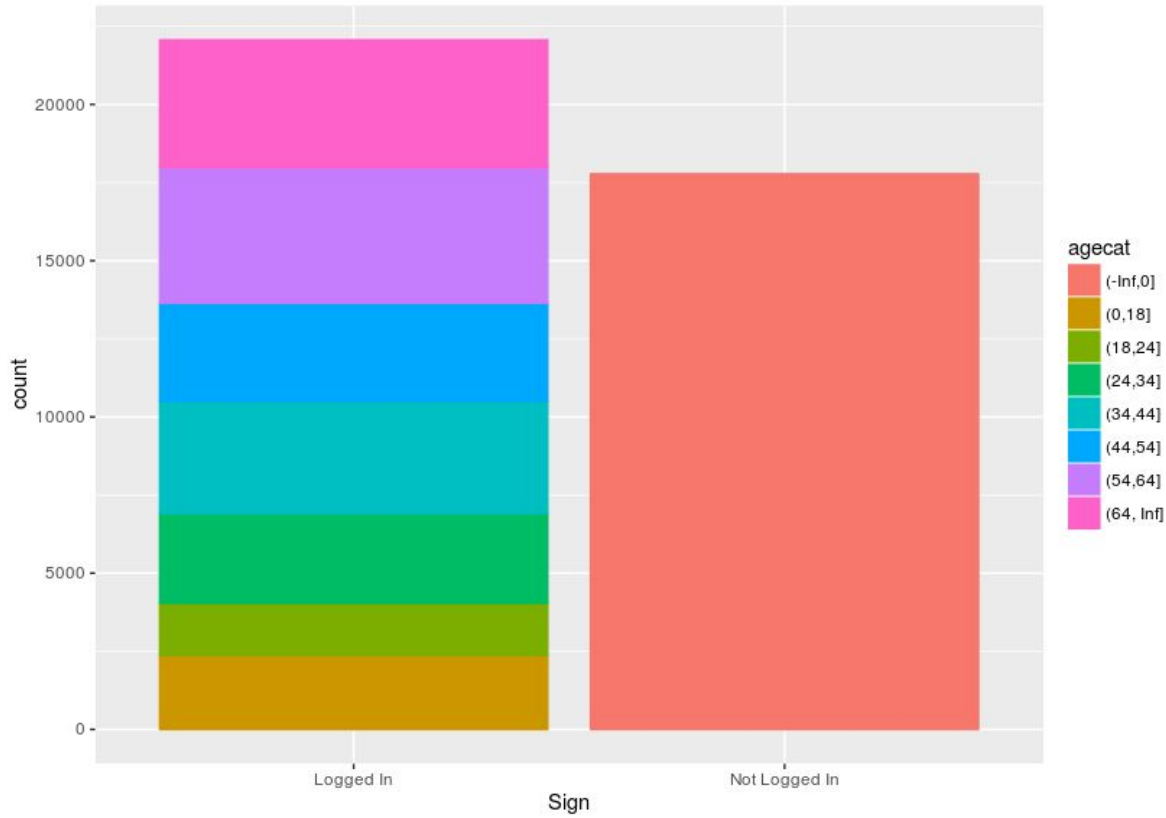
Above Plot gives the Analysis between the count of data Vs Gender and agecat



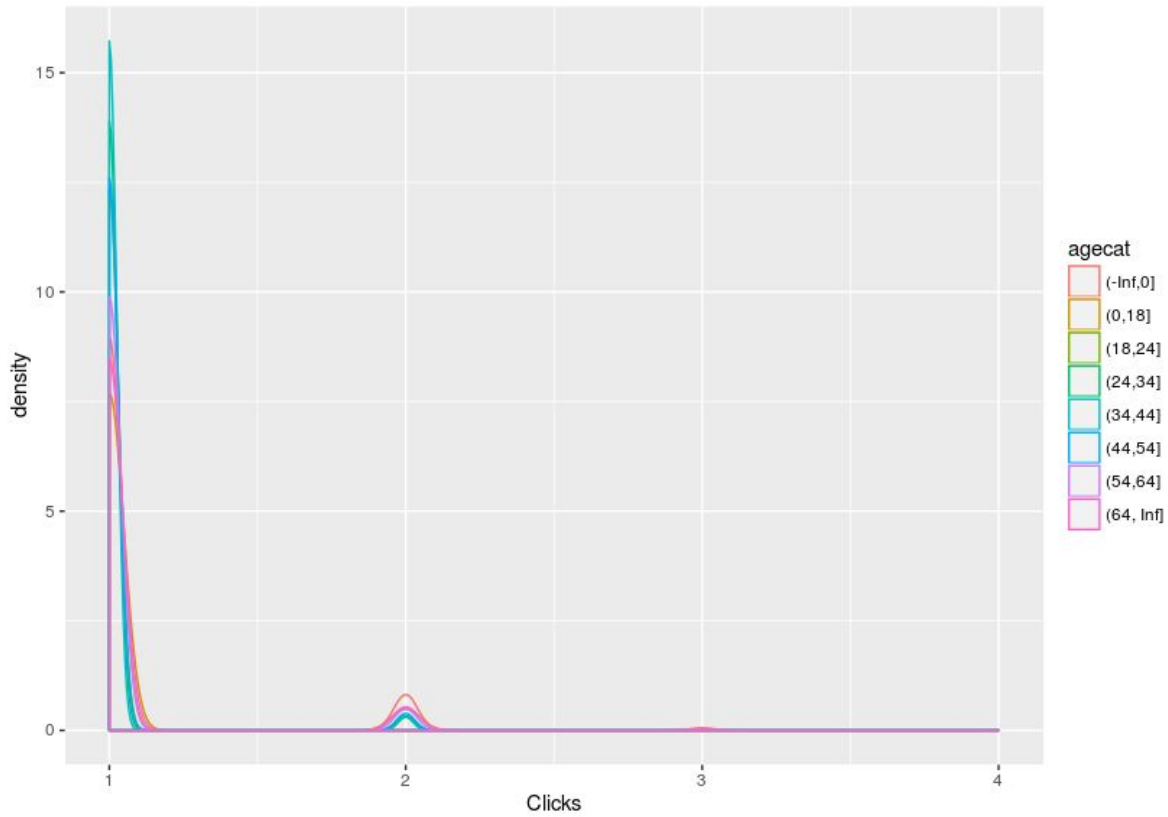
Above Plot gives box plot of Impressions Vs agecat



Above Plot gives graph of count of data Vs Impressions and agecat



Above Plot gives graph of count of data Vs Logged In and Not Logged In and agecat

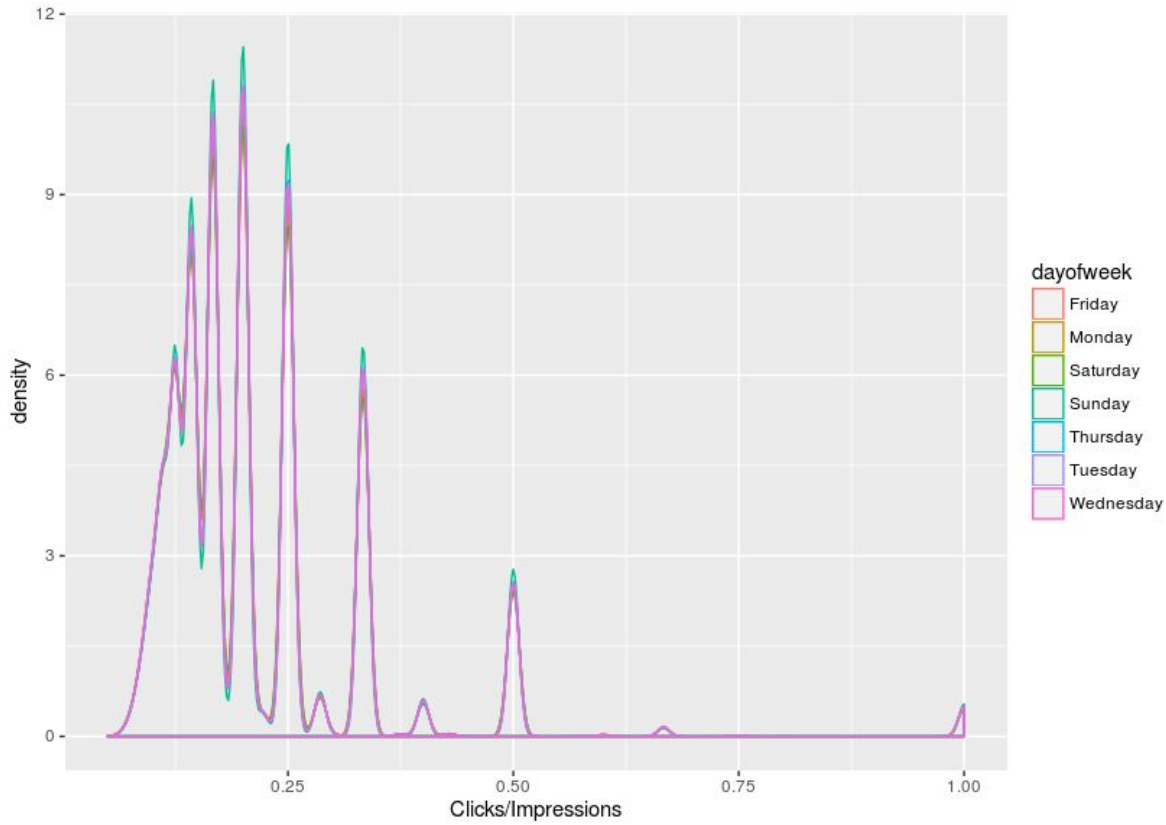


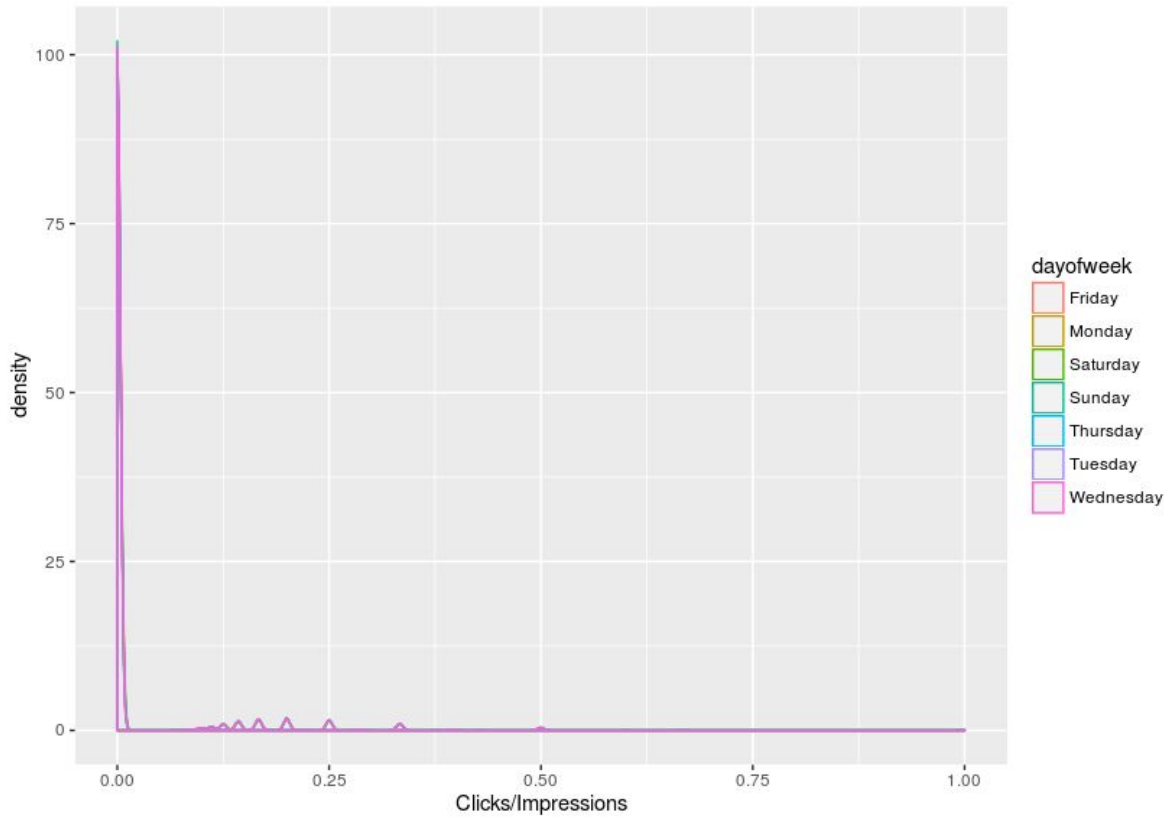
Above Plot - Density Vs No of Click and agecat

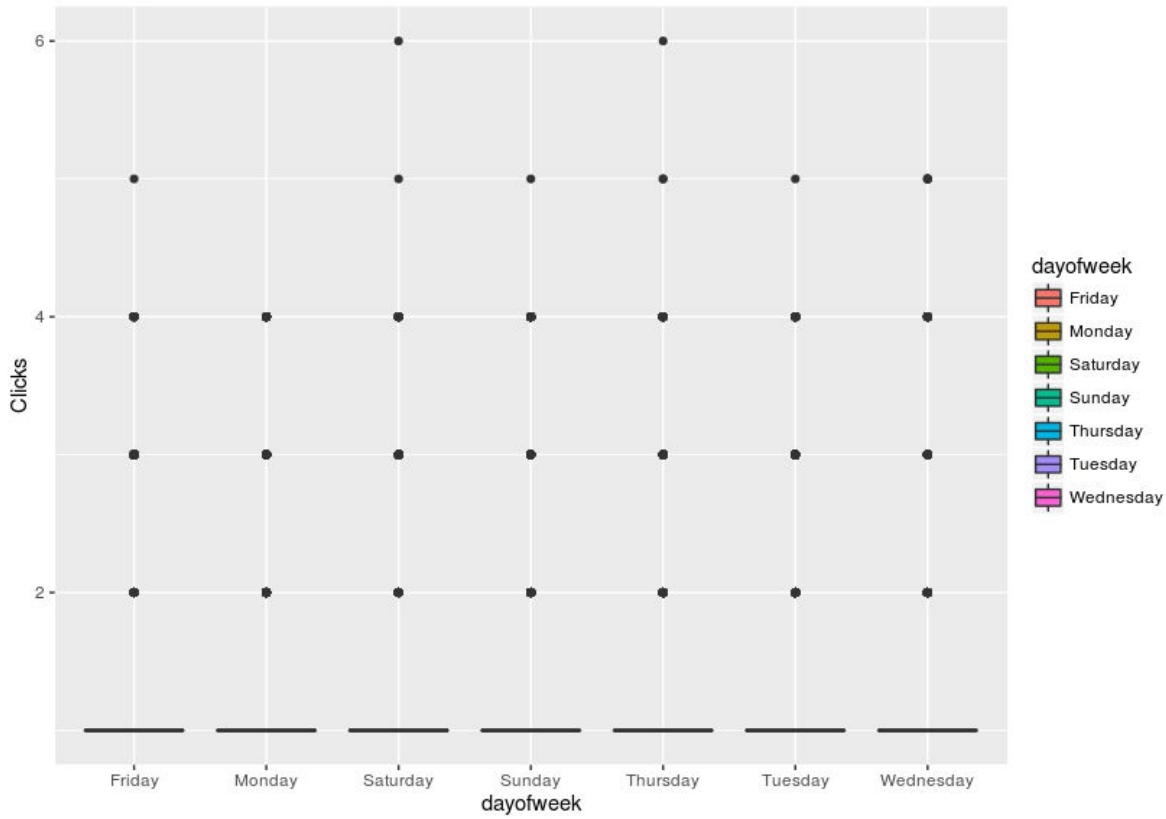
Observations for Single Day Data :

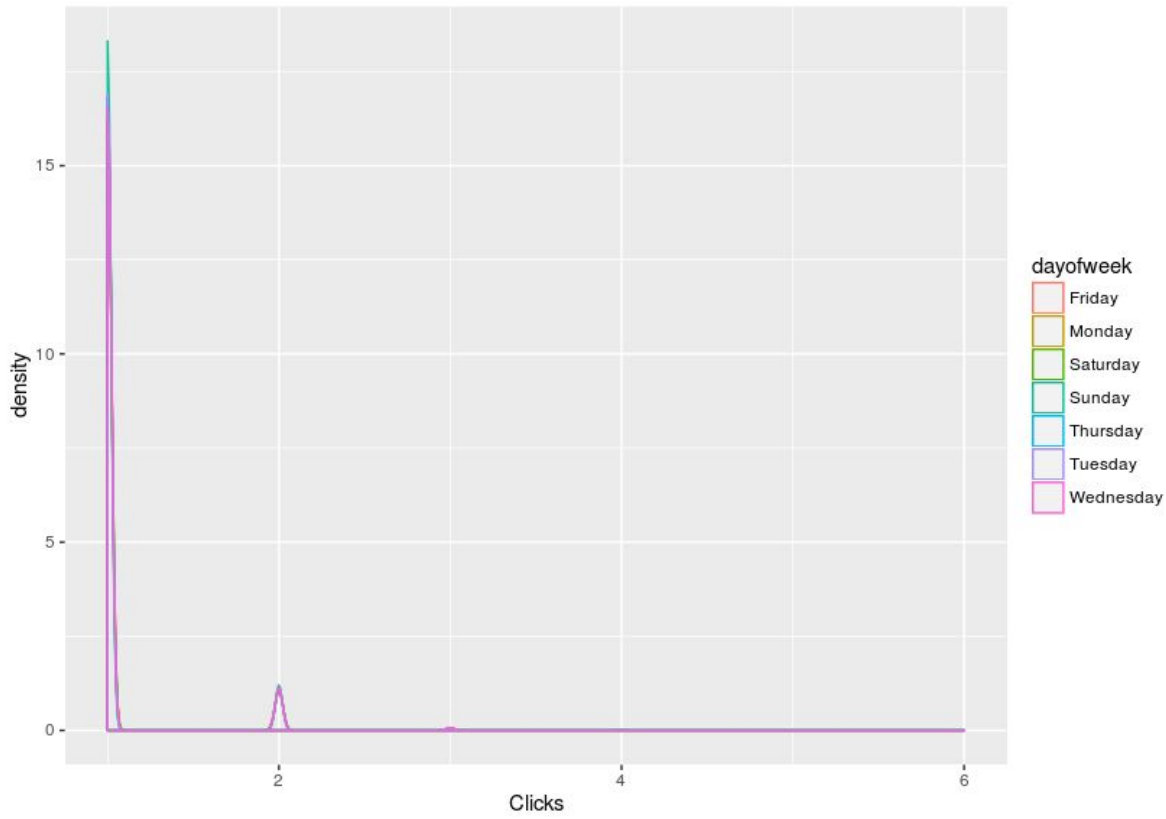
1. Majority of the Users generating the impressions are in the age group (18,24],(24,34],(34,44]) , and very few Users in the age group (64,Inf] generate impressions .
2. Most of the impressions are not converted into clicks which is evident from the density Vs CTR graph
3. All age groups follow the same distribution for Density Vs CTR graph
4. There are more number of Users who are logged In than the number of Users who are not Logged In.
5. There more Adults in the data and the count of females are almost equal to males but slightly less.
6. The Gender Ratio for different age groups are almost equal.

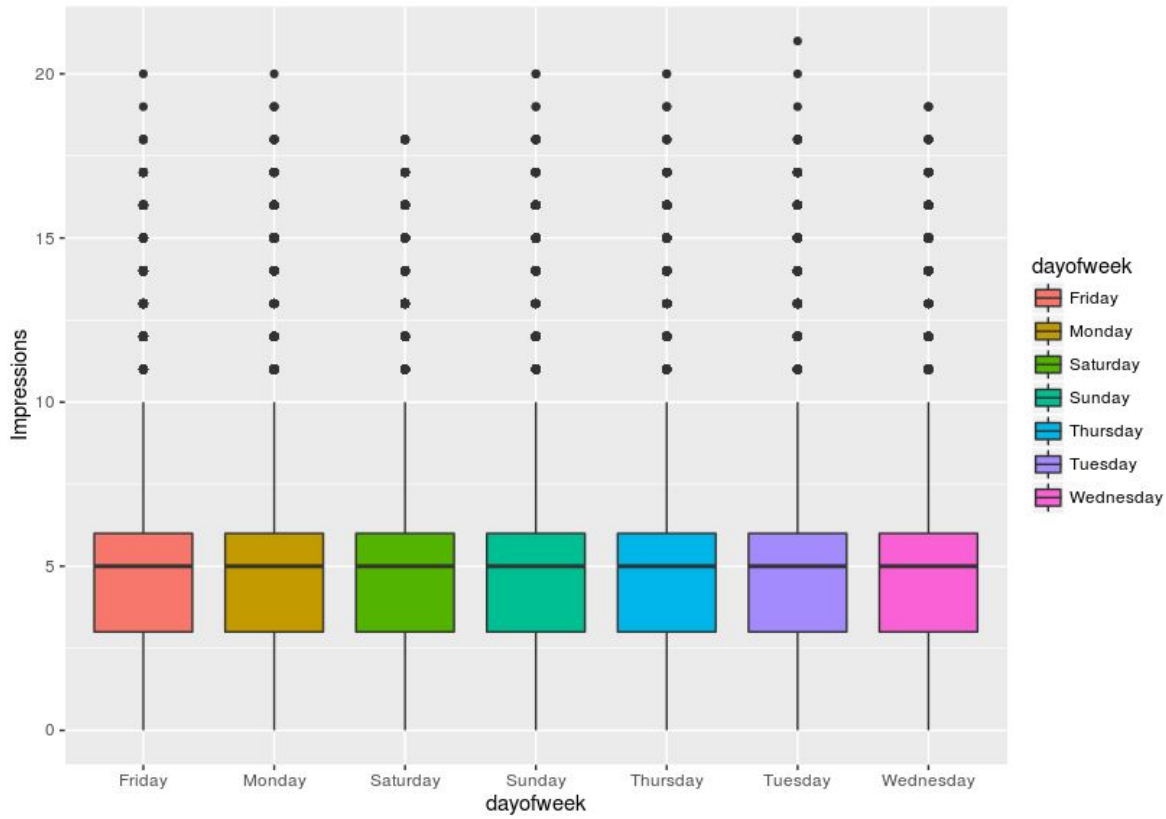
Part B : For Multiple Days - Following are the plots

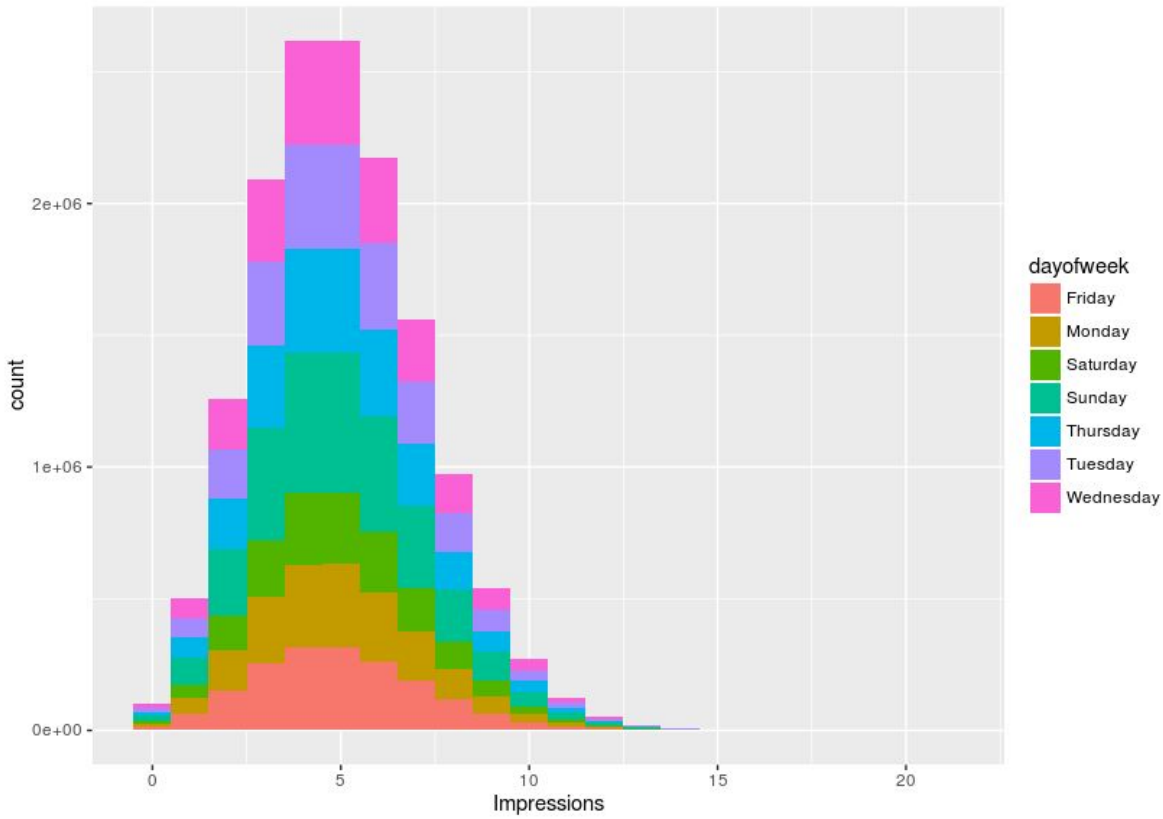


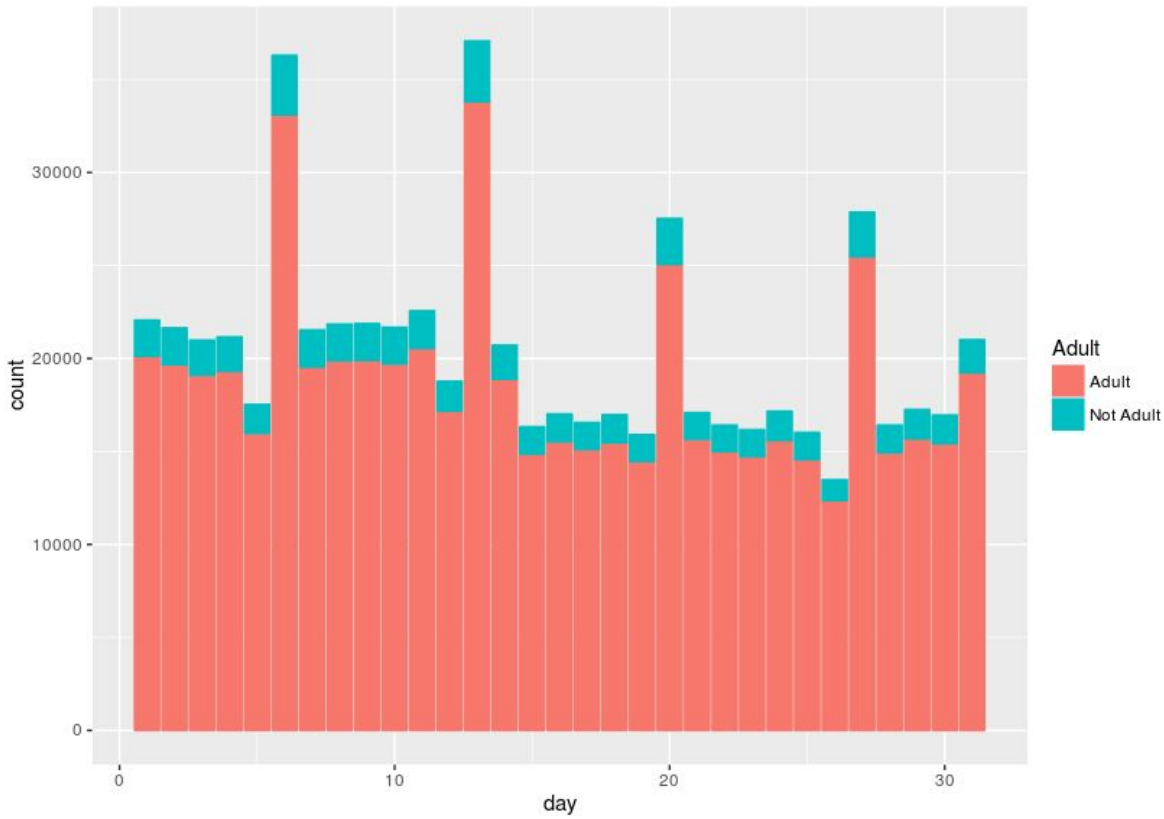

100
75
50
25
0
density
0.00
0.25
0.50
0.75
1.00
Clicks/Impressions
dayofweek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday


6
4
2
Clicks
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
dayofweek
dayofweek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

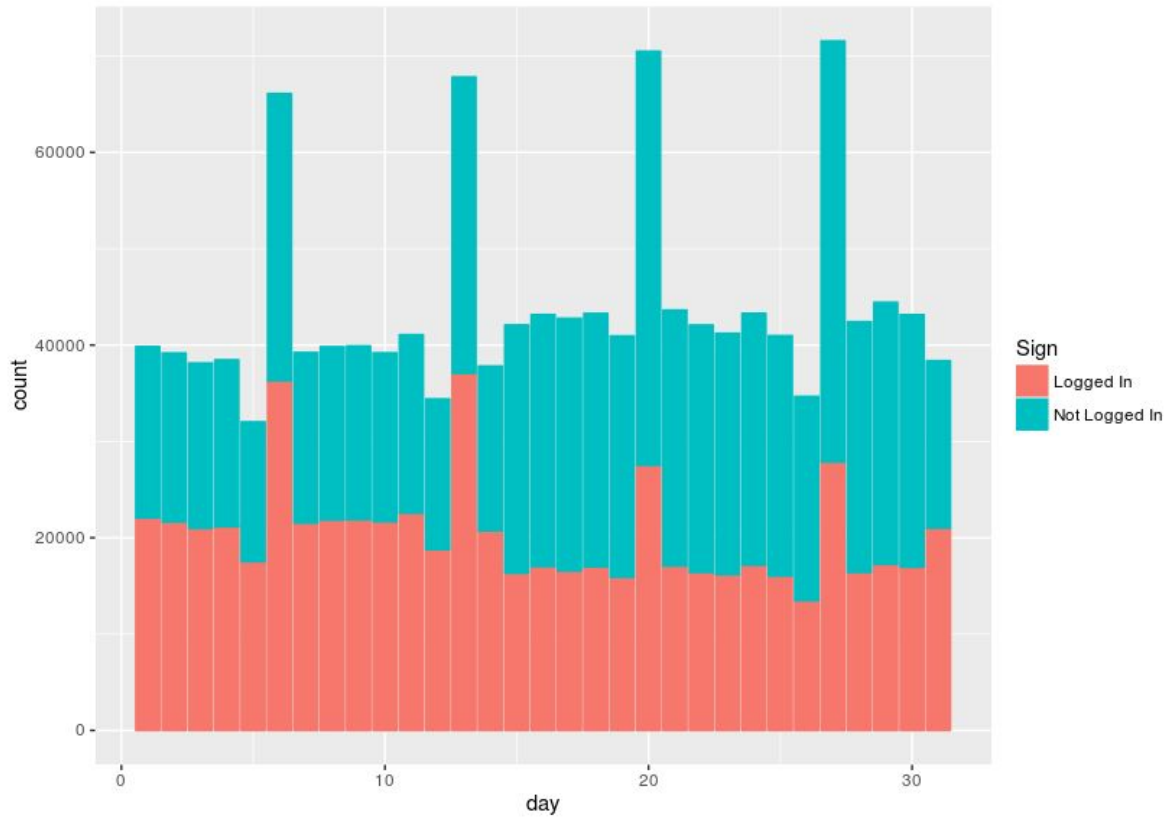

density
15
10
5
0
2
4
6
Clicks
dayofweek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

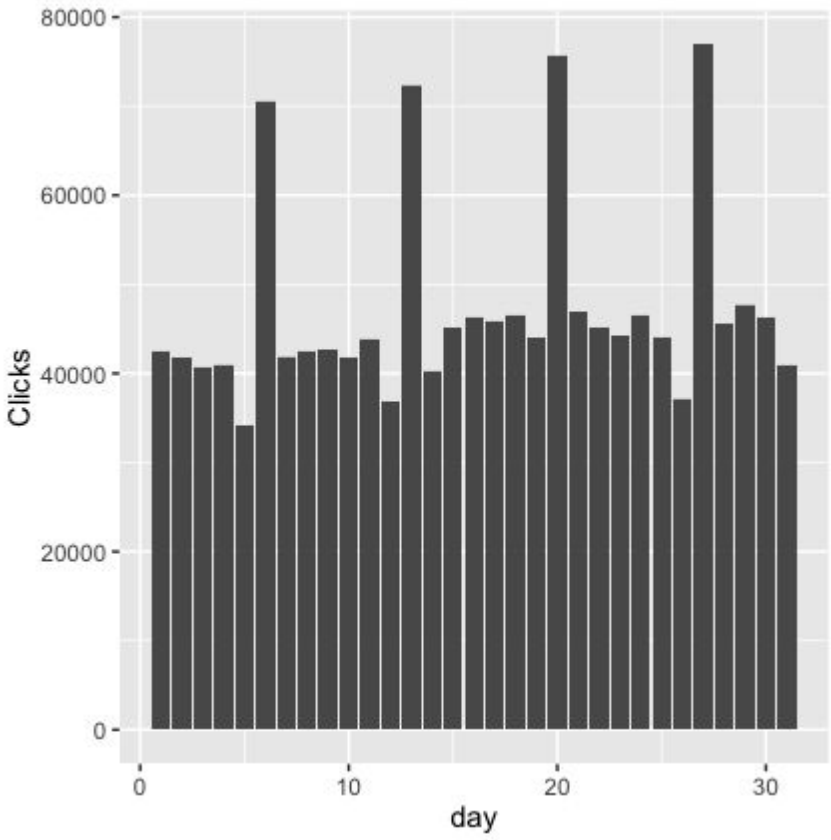

Impressions
20
15
10
5
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
dayofweek
dayofweek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

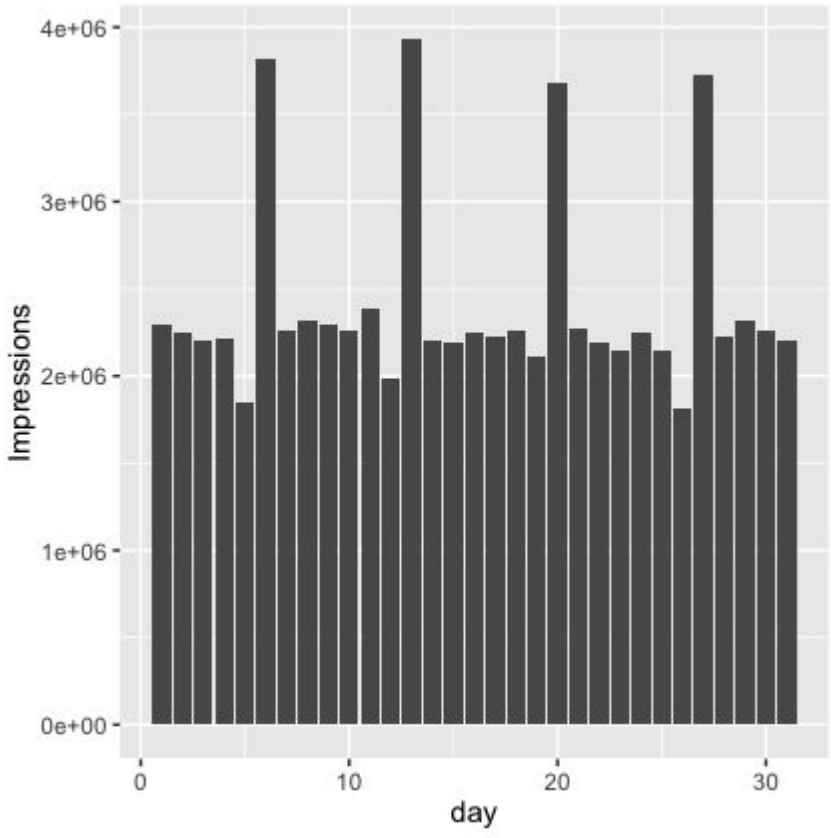

count
2e+06
1e+06
0e+00
0
5
10
15
20
Impressions
dayofweek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday


count
30000
20000
10000
0
0
10
20
30
day
Adult
Adult
Not Adult


30000
20000
10000
0
count
0
10
20
30
day
GenderVerb
Female
Male


60000
40000
20000
0
count
0
10
20
30
day
Sign
Logged In
Not Logged In


80000
60000
40000
20000
0
Clicks
0
10
20
30
day


4e+06
3e+06
2e+06
1e+06
0e+00
Impressions
0
10
20
30
day

Observations for Multiple Days Data :

1. The Count of Clicks and Impressions are high on Sundays which are evident from the graphs.
2. Around 60% of the Users are signed In
3. The number of impressions is almost the same on all the days of the week.
4. There is some fluctuation in the density of Click Through Rate on Wednesday and Sunday which is evident from the graph.
5. There are more males than females .
6. Higher age range implied that CTR tends more towards certain values.