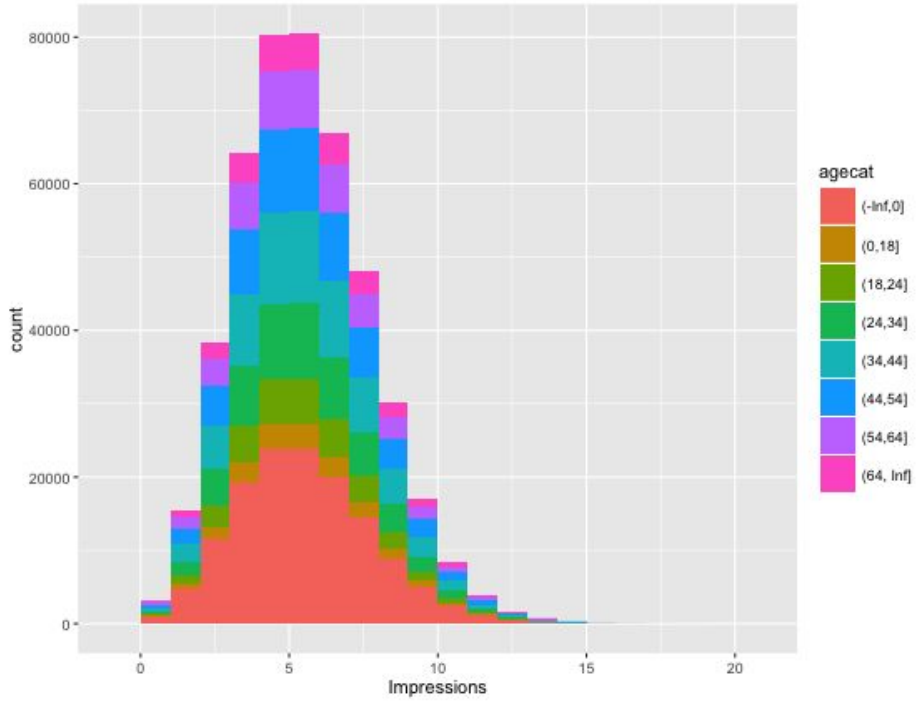
Problem 2: Simple EDA

The exploratory data analysis is performed on data of clicks and impressions of users on New York Times website. First the analysis is done on a single day data and then extended to a monthly data of 31 days. The columns data contains are age, gender, number of impressions, number of clicks and logged in or not. More columns are added for example age category, has impressions and scode by:

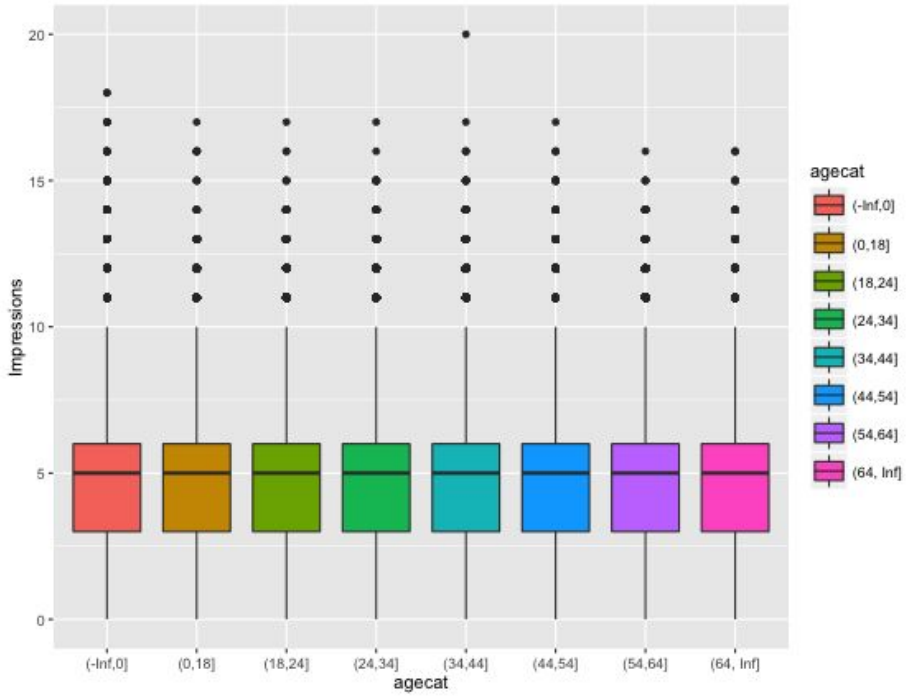
```
data1$agecat<- cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
data1$hasimps<-cut(data1$Impressions, c(-Inf,0,Inf))

data1$scode[data1$Impressions==0]<-"No Imps"
data1$scode[data1$Impressions>0]<-"Imps"
data1$scode[data1$Clicks>0]<-"Clicks"
```

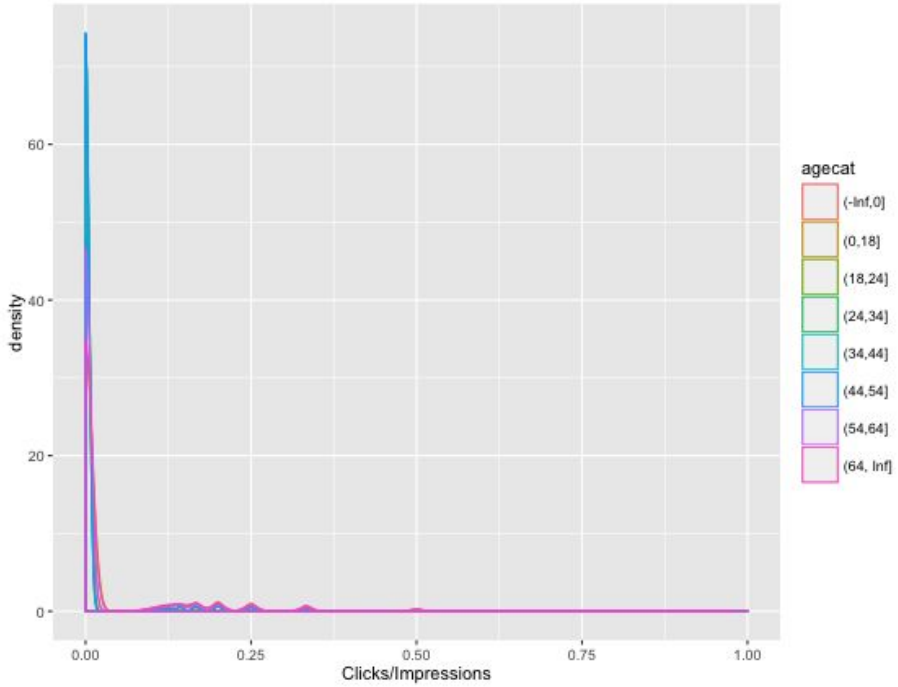
Charts for single day are following:



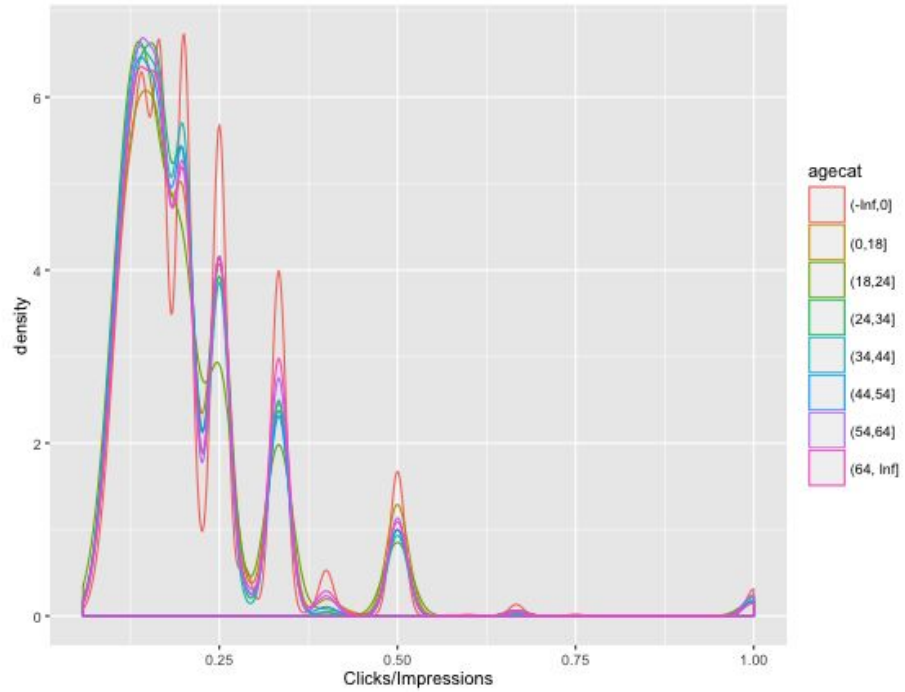

```
ggplot(data1, aes(x=Impressions, fill=agecat))+geom_histogram(binwidth=1)
```



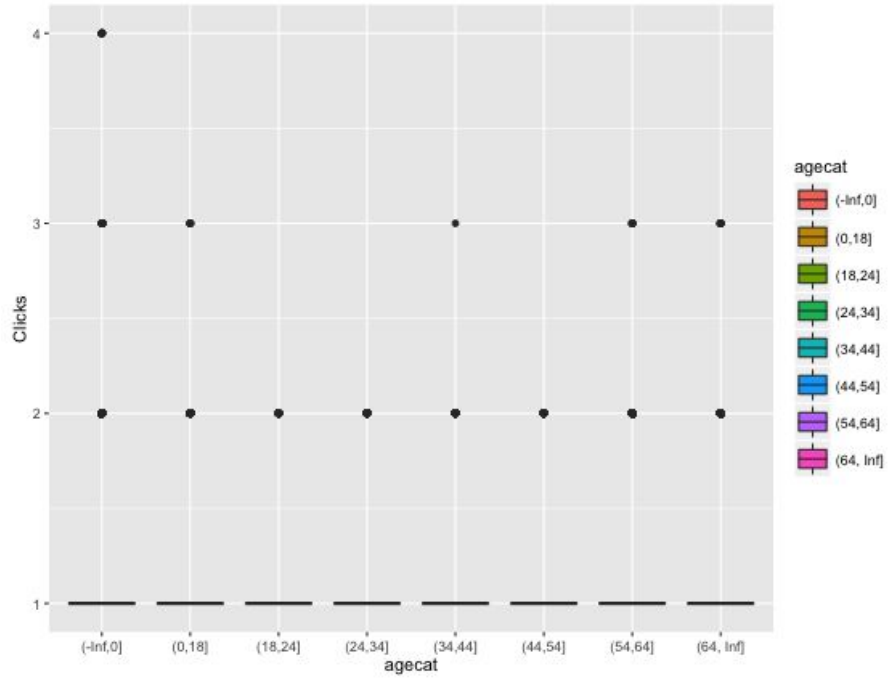
ggplot(data1, aes(x=agecat, y= Impressions, fill=agecat))+geom_boxplot()



ggplot(subset(data1,Impressions>0), aes(x=Clicks/Impressions, colour= agecat))+geom_density()

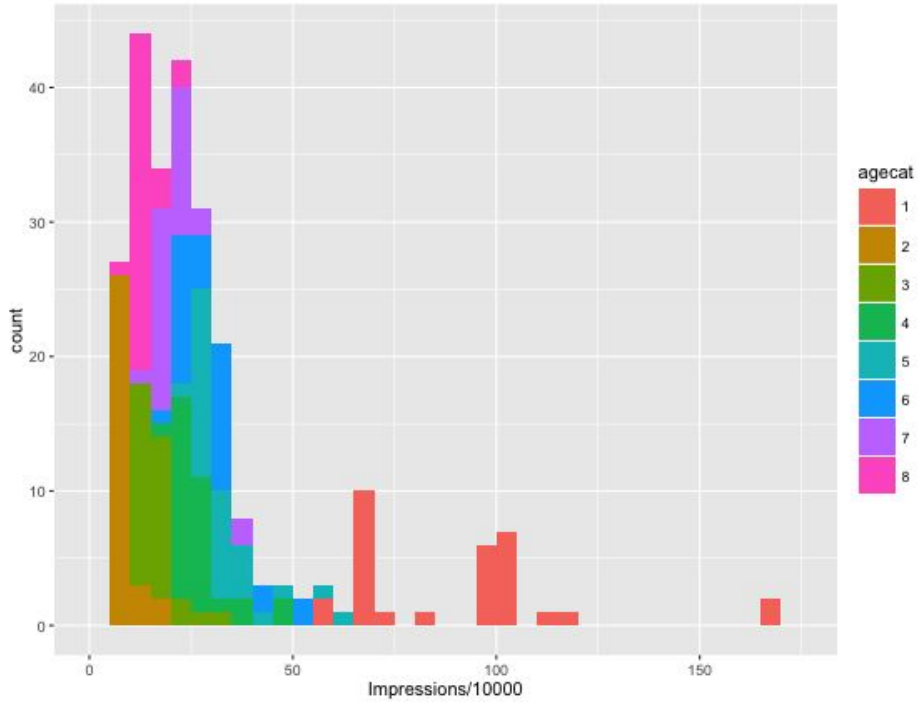


ggplot(subset(data1,Clicks>0), aes(x=Clicks/Impressions, colour= agecat))+geom_density()

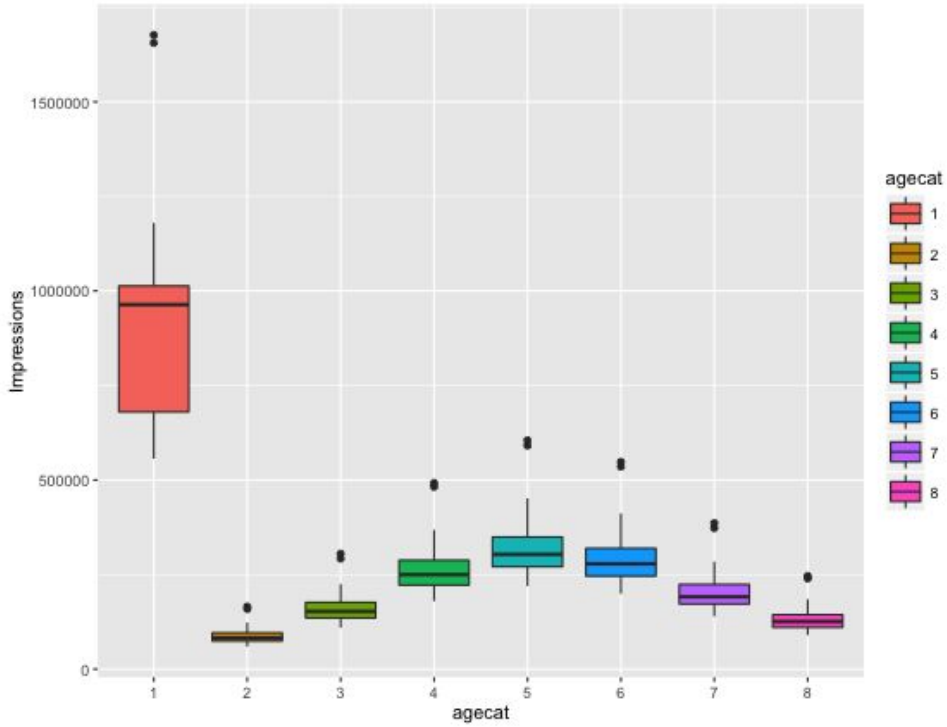


ggplot(subset(data1,Clicks>0), aes(x=agecat, y= Clicks, fill= agecat))+geom_boxplot()

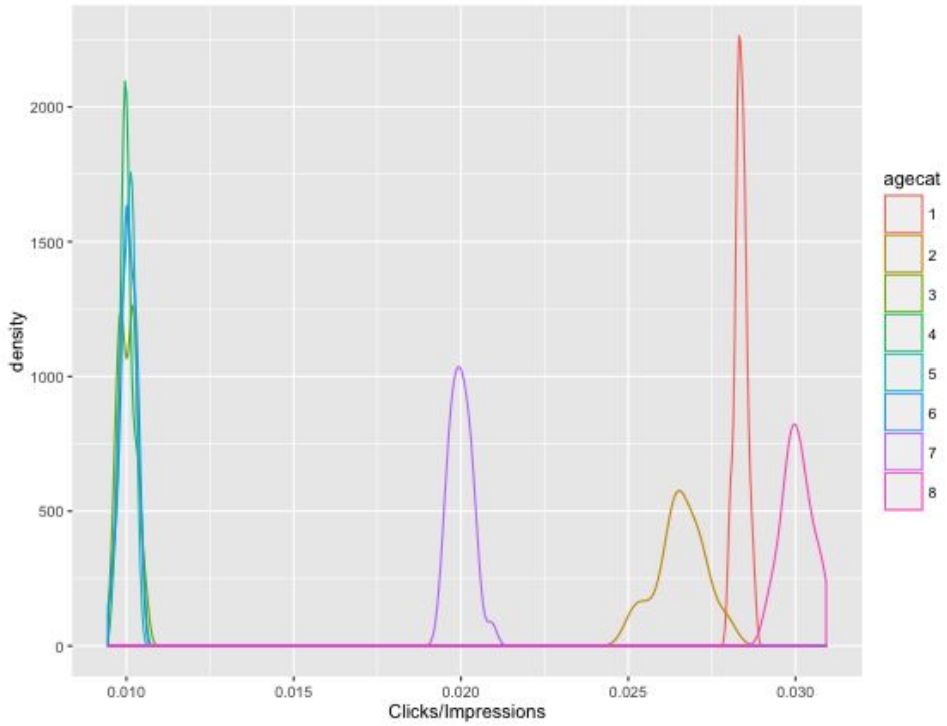
Extending the analysis to monthly data:



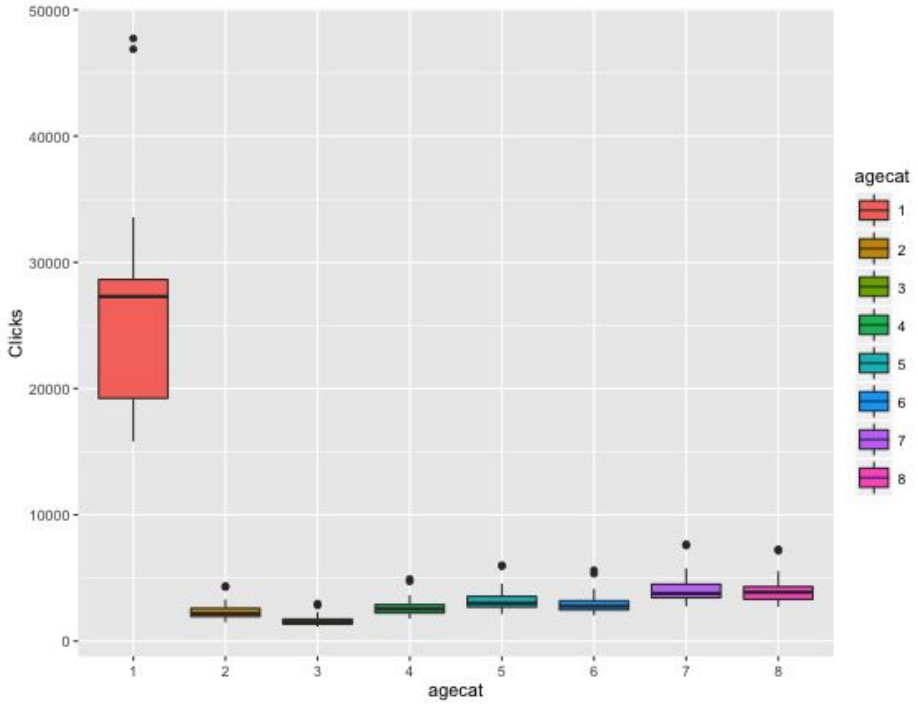
ggplot(collect, aes(x=Impressions/10000, fill=agecat))+geom_histogram(binwidth=5)



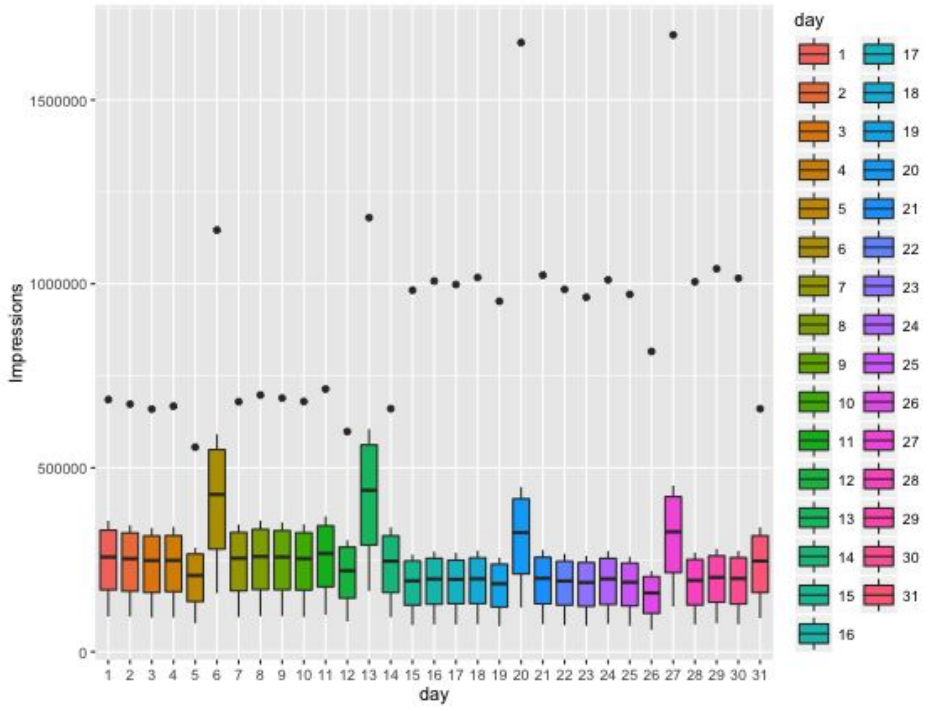
ggplot(collect, aes(x=agecat, y= Impressions, fill=agecat))+geom_boxplot()



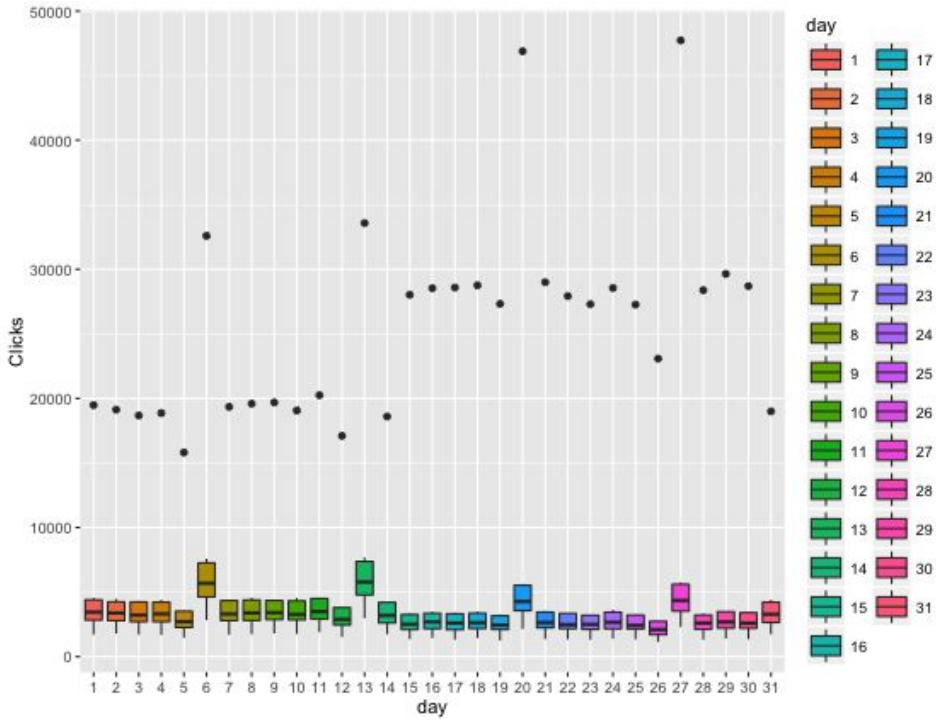
ggplot(collect, aes(x=Clicks/Impressions, colour= agecat))+geom_density()



ggplot(collect, aes(x=agecat, y= Clicks, fill= agecat))+geom_boxplot()



ggplot(collect, aes(x=day, y= Impressions, fill=day))+geom_boxplot()



ggplot(collect, aes(x=day, y= Clicks, fill= day))+geom_boxplot()

By looking at the charts it can be deduced that people 34-44 years of age are most active or interested in the website if we consider impressions and 54-64 years of age if we consider clicks therefore, making them the target audience. Also, looking at the charts showing day wise distribution it can be seen that it follows a certain pattern. Every seventh day the clicks and impressions are considerably more than the rest of the days. It can be assumed that those are the weekends and by looking at the number of clicks make them the target days for any new marketing strategy.