Pratima Kshetry, pkshetry@umd.edu

Part 1

a. Create a new variable, age_group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64" and "65+".

Source Code:

```
setwd("E:/Rdataset/HW1-ExploratoryData-Nyt")
Data <-read.csv("E:/Rdataset/HW1-ExploratoryData-Nyt/nyt5.csv",
                stringsAsFactors=FALSE,    strip.white=TRUE, na.strings=c("NA",""))
summary(data)  ###Showed age (max)=106
break <- c(0,18,25,35,45,55,65,106)   ##Create breaking point vector
label <- c('<18','18-24','25-34','35-44', '45-54','55-64','65+')   ##Create labels
data$Age_Group<-cut(data$Age,
                    breaks=break,
                    labels=label,
                    right=F,
                    ordered_result = T)
```

---

b. For a single day,

i) Plot the distributions of number impressions and click-through-rate (CTR=# clicks/# impressions), for these 6 age categories. (Submitted)

---

ii) Define a new variable to segment or categorize users based on their click behavior.

Ans: data$Users_clicked <- ifelse(data$Clicks>=1, 1, 0)

---

iii) Explore the data and make visual and quantitative comparisons across user segments/ demographics (<18 year old male vs < 18 year old females or logged-in vs not, for example).

Ans:

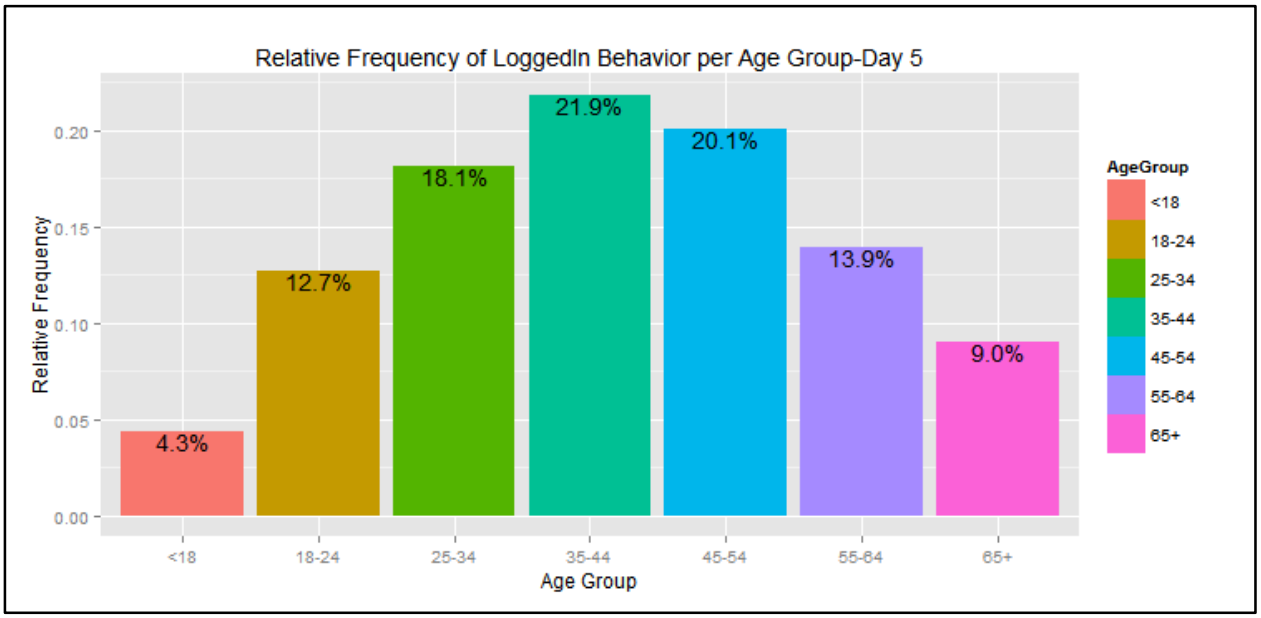


Fig: Comparision of Logged In Behavior across Age Groups

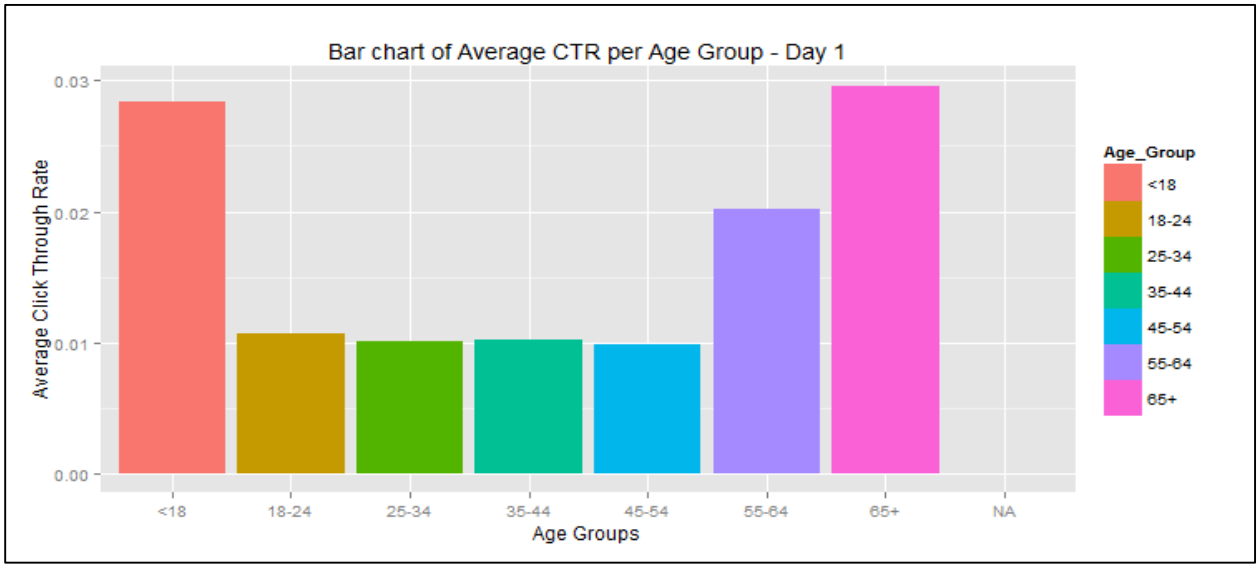


Fig. Comparision of Avge CTR across Age Group

c. Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time; what will compress the data, but still capture user behavior.

Ans: meanAge_Group<-mean(data$Age_Group) ####Gave mean of Age Groups

Also , calcuated average page impressions for day 5 and plotted Average Impressions of the page across age groups and gender.

Fig. Interaction Plot of Gender and Age Group across Impressions.

c. Extend your analysis across days.

Pratima Kshetry, pkshetry@umd.edu

###Plotted Page Impressions across 7days

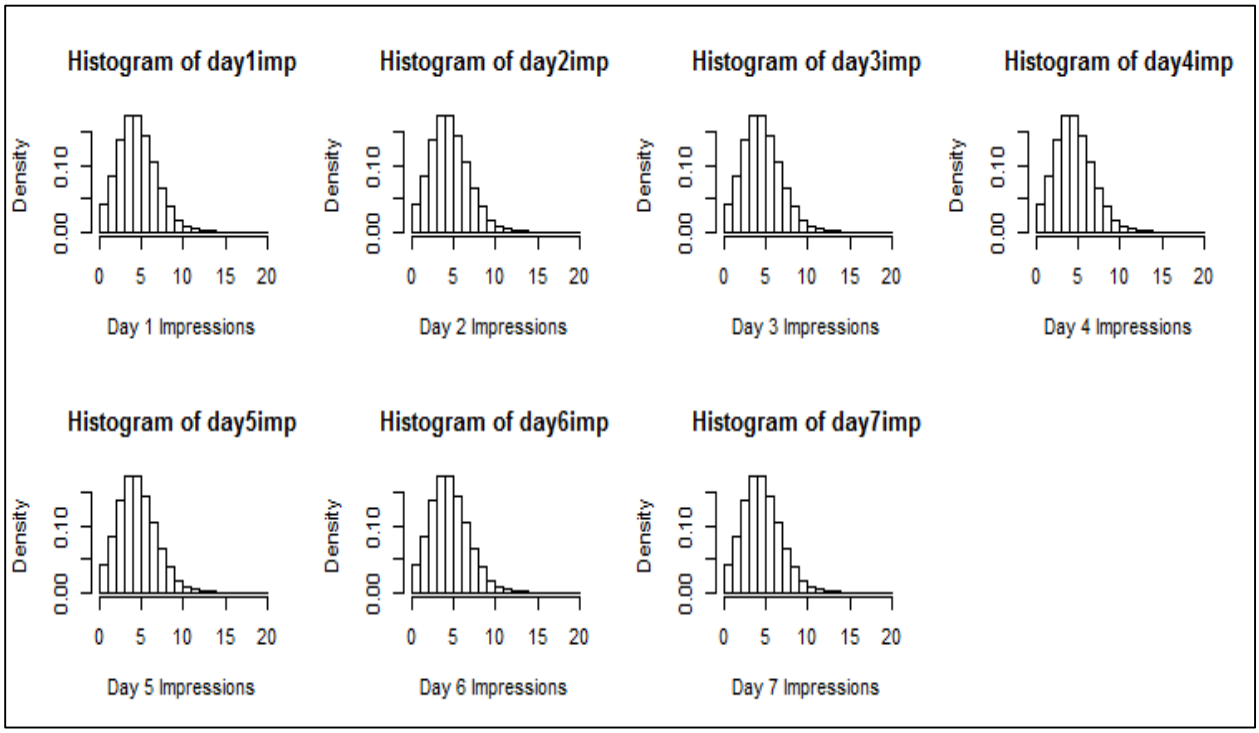
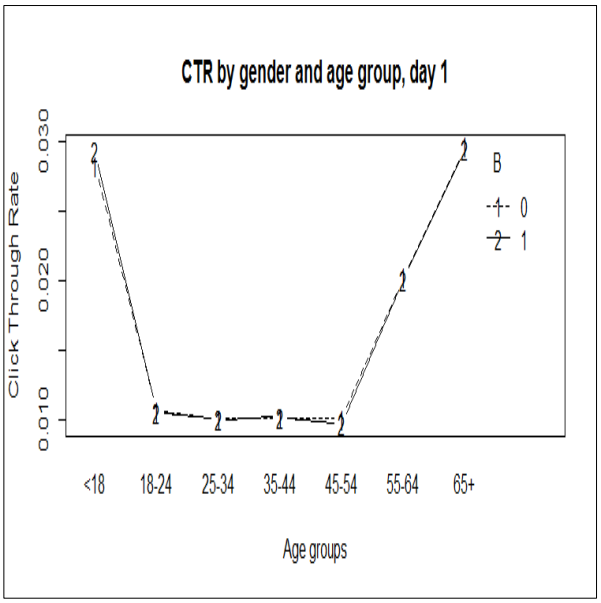


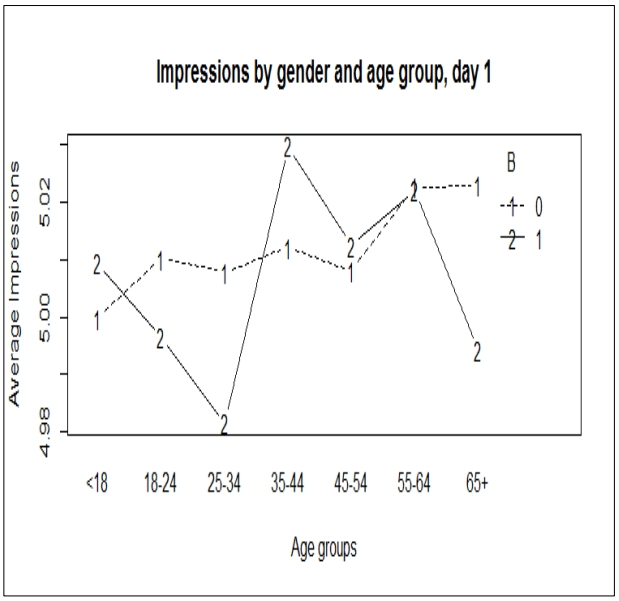
Fig. Page Impressions across 7 days

Also Plotted Interaction between gender and Age group across Impression , and CTR

(Please scroll down)

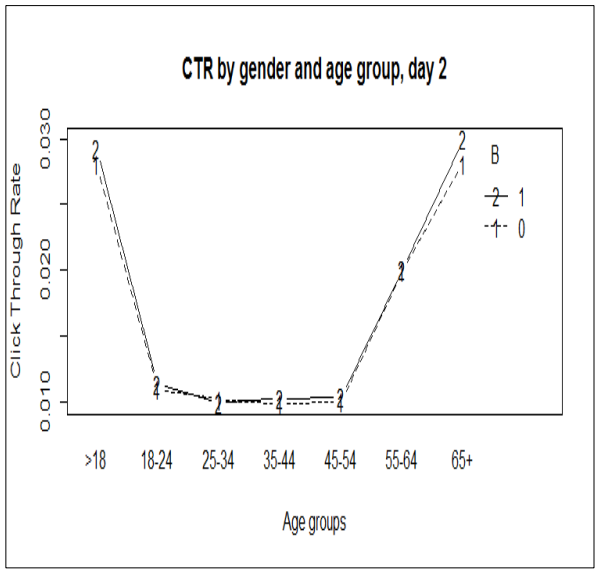


Day1: CTR (Interaction Plot)

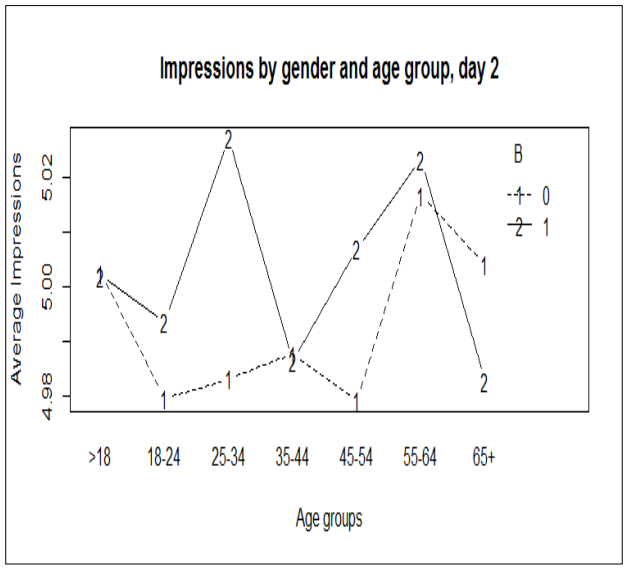


Day1: Impressions (Interaction Plot)

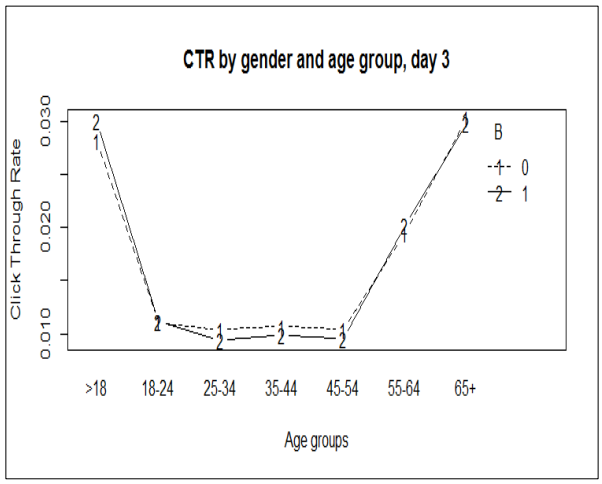
Pratima Kshetry, pkshetry@umd.edu



Day2: CTR (Interaction Plot)

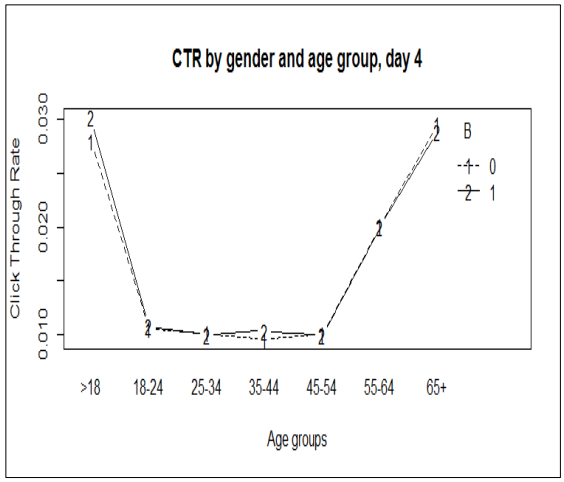


Day2: Impressions (Interaction Plot)

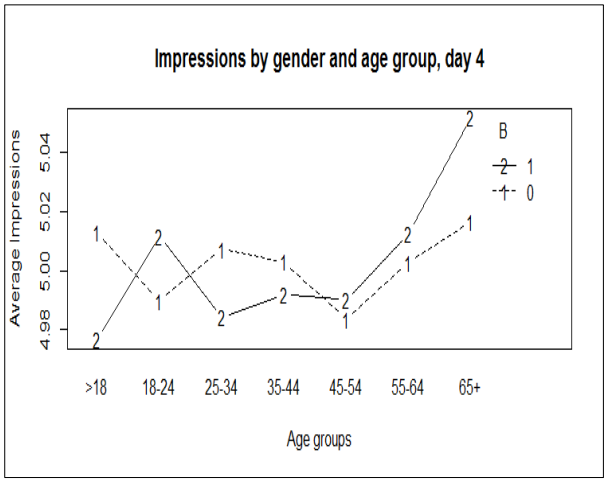


Day3: CTR (Interaction Plot)

Day3: Impressions (Interaction Plot)

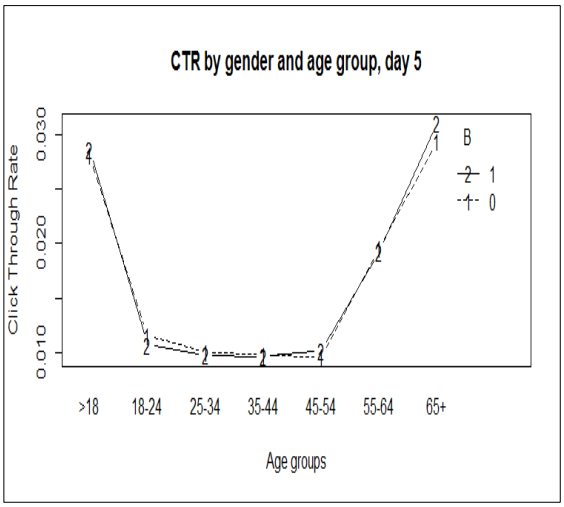


Day4: CTR (Interaction Plot)

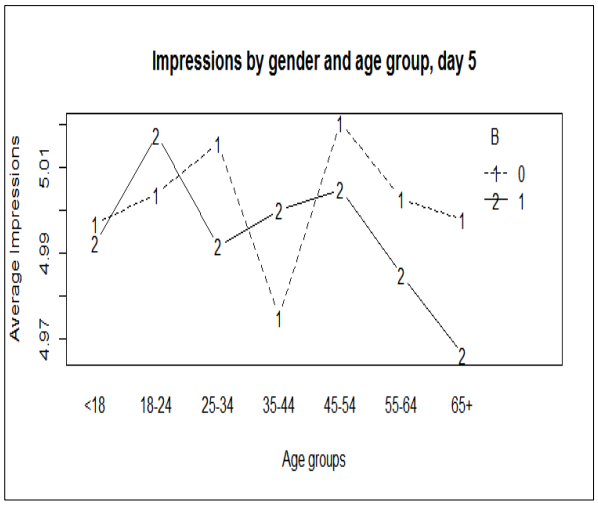


Day4: Impressions (Interaction Plot)

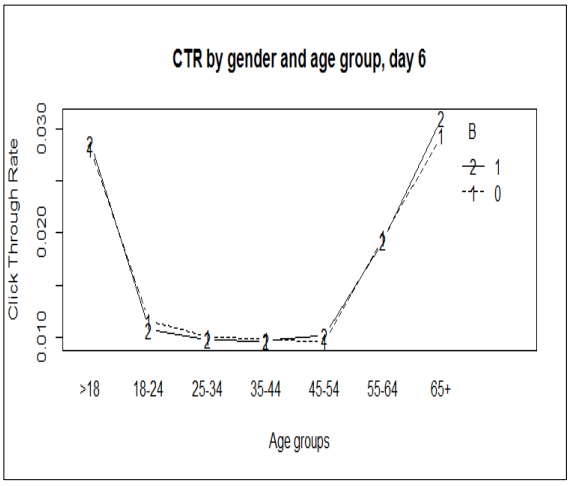
Pratima Kshetry, pkshetry@umd.edu



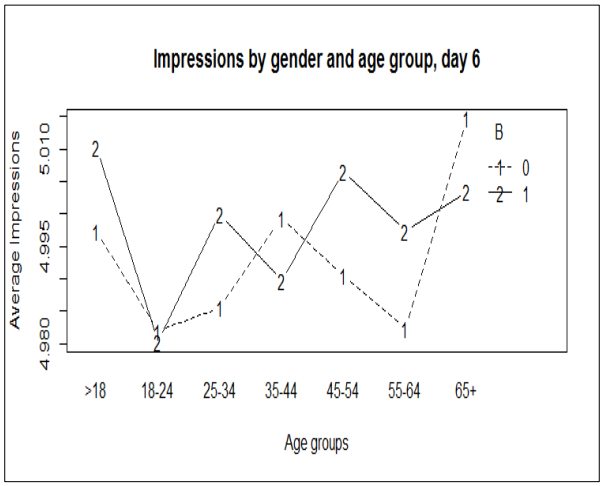
Day5: CTR (Interaction Plot)



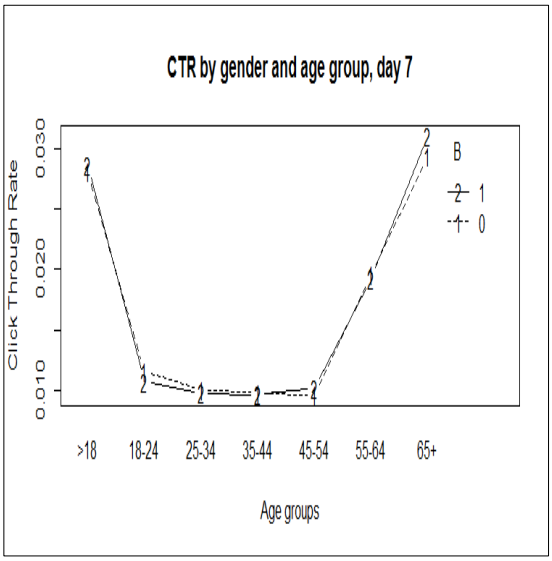
Day5: Impressions (Interaction Plot)



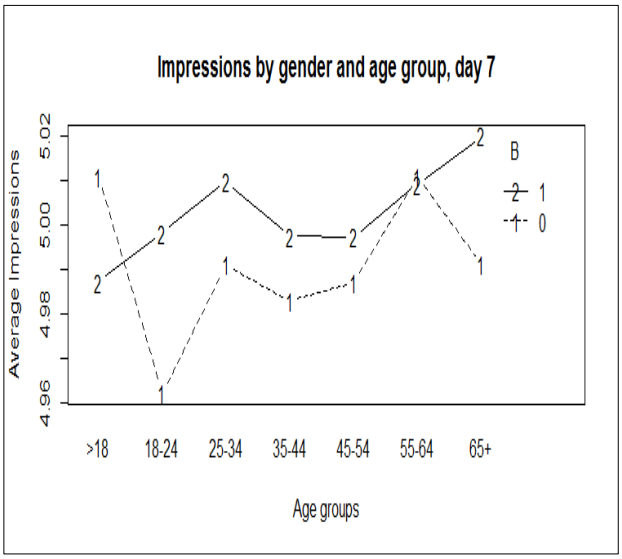
Day6: CTR (Interaction Plot)



Day6: Impressions (Interaction Plot)



Day7: CTR (Interaction Plot)



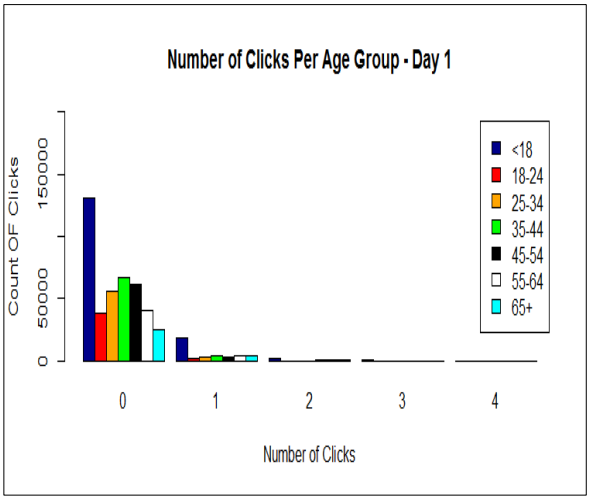
Day7: Impressions (Interaction Plot)

Pratima Kshetry, pkshetry@umd.edu

d. Describe and interpret any patterns you find.

Following are the few observations made on the given dataset:

-In the above data, we see that people from the age group less than 18 years are most likely (and most frequently) to click an advertisement than any other age group.



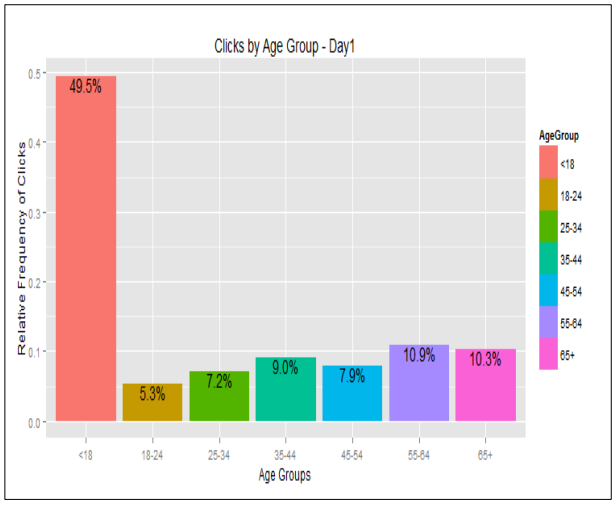
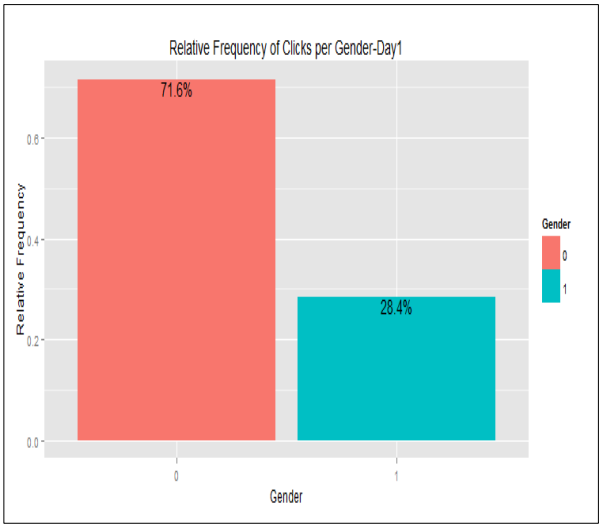


Fig. Number of Clicks Per Age Group

Fig. Relative Frequency of Clicks Per Age group

-It appears that females are most likely to click an advertisement than a male, however males are more likely view the page with a LoggedIn status.



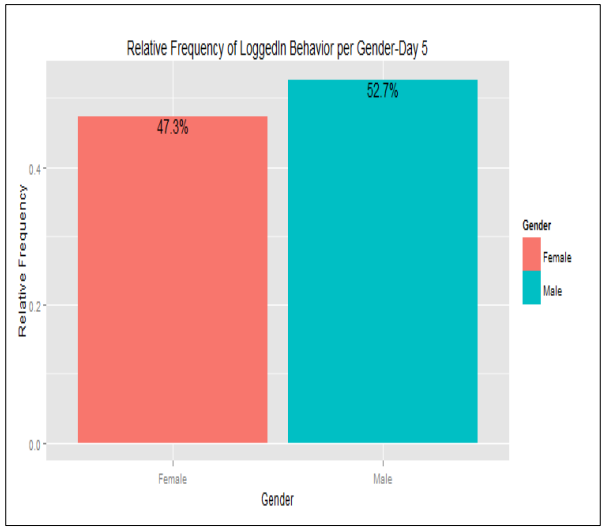
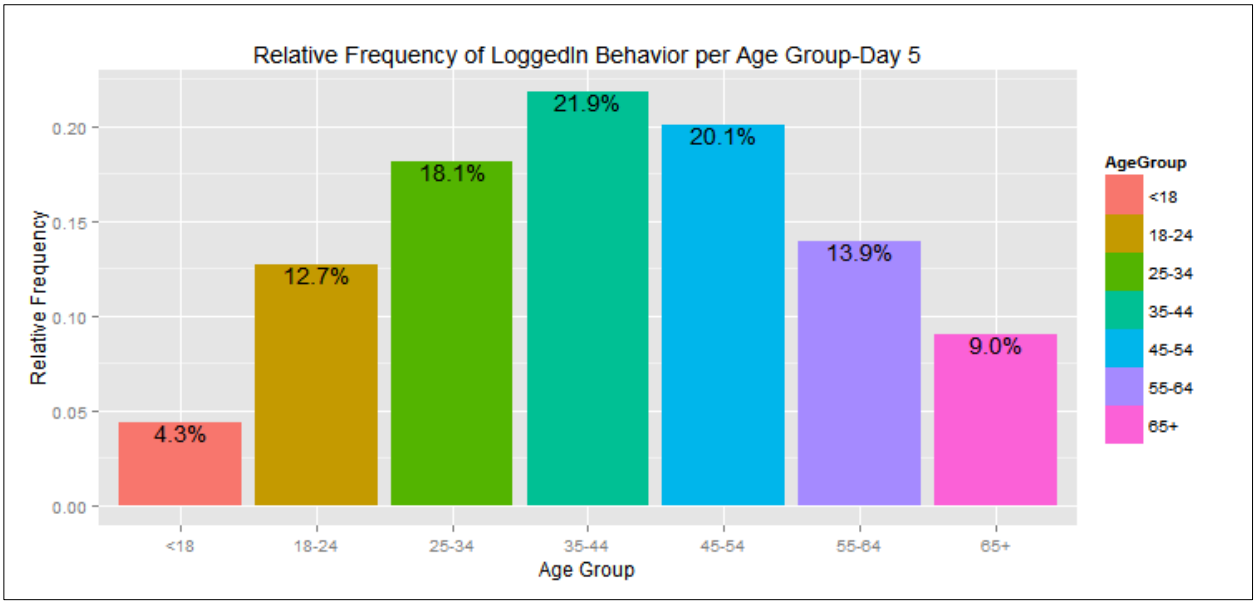


Fig. Freq of Clicks by Gender

Fig. Freq of Clicks by Logged In (status) by gender

-People from the age group of 35-44 show more Signed In behavior. We can also note that the data for the LoggedIn Behavior per age group is **normally distributed**. We see no anomalies in the given data.



-From the interaction plots as showed in the earlier question, the difference between CTR is higher between males and females for the age group >18 and 65 + while the CTR behavior is almost the same for male and female across all age groups. Also, from the interaction plots, we see that the advert impression across gender varies as per age groups. However, people from the age group of less than 18 years have more Page views followed by people from the age group of 35-44. As seen from the bar chart below, the median impression (page views) of the New York Times web page is 4-5. (Note: The outliers page impressions have been excluded from the chart).

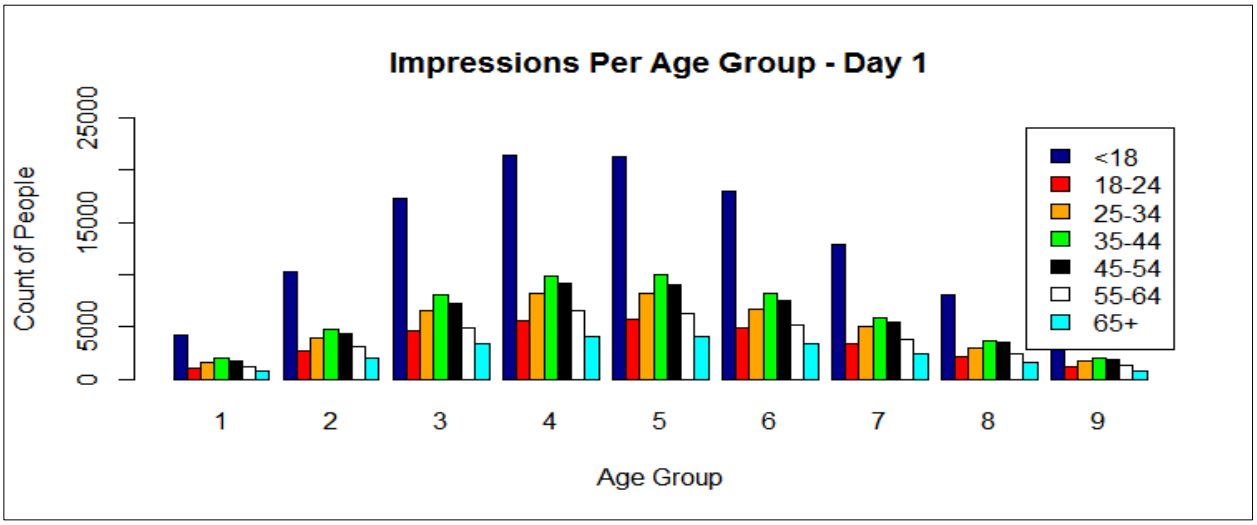


Fig. Impressions Per Age Group

Pratima Kshetry, pkshetry@umd.edu

Part 2 (Find your own Dataset)
Dataset: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

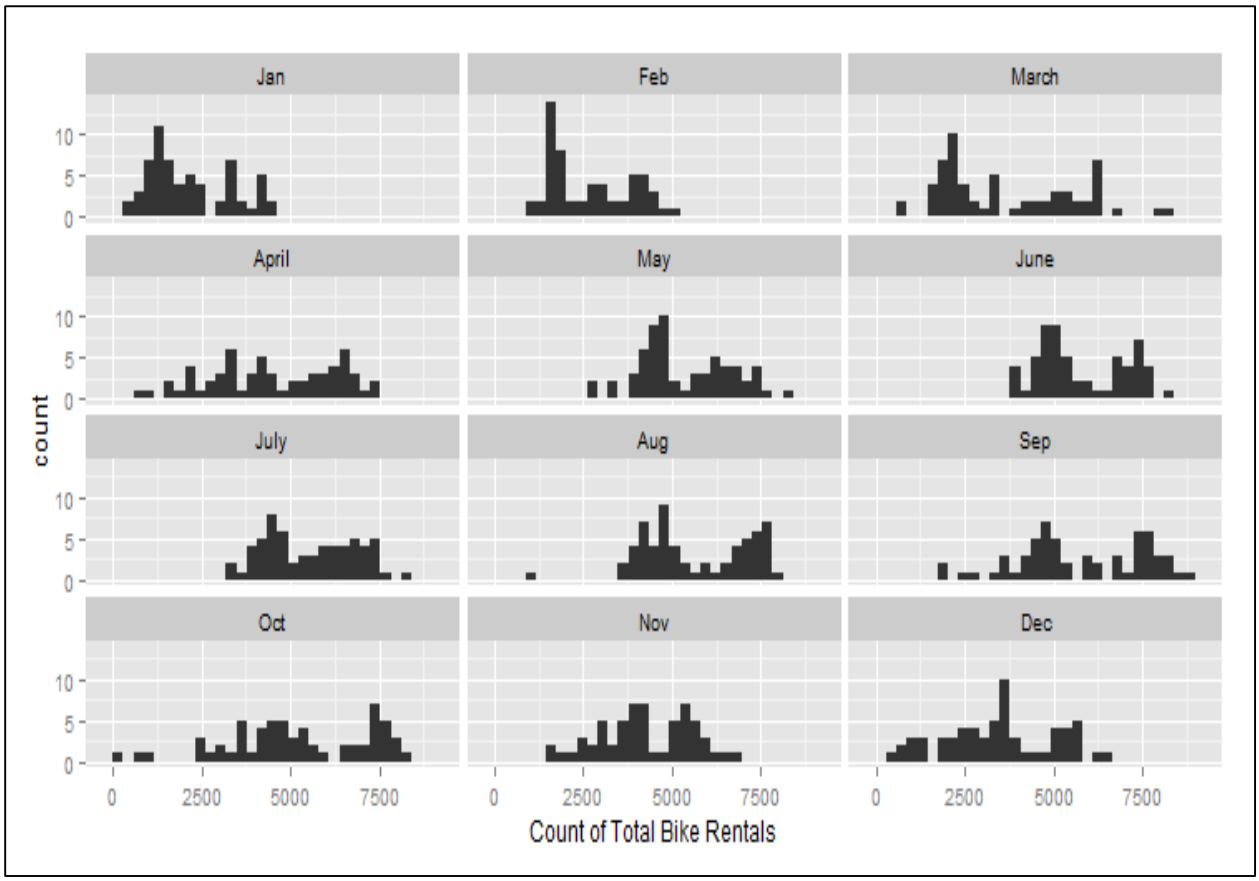


Fig: Bike Rentals by Month

Pratima Kshetry, pkshetry@umd.edu

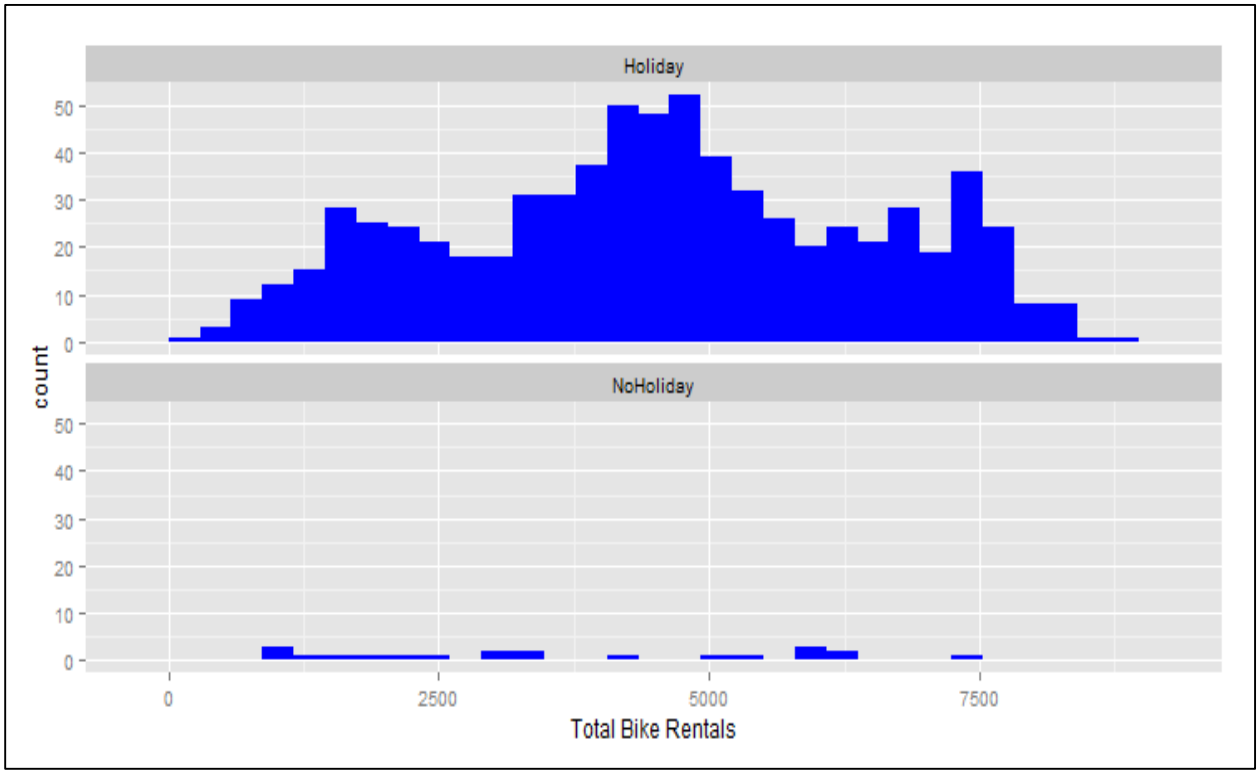


Fig. Bike Rentals on Holidays VS Non-Holidays

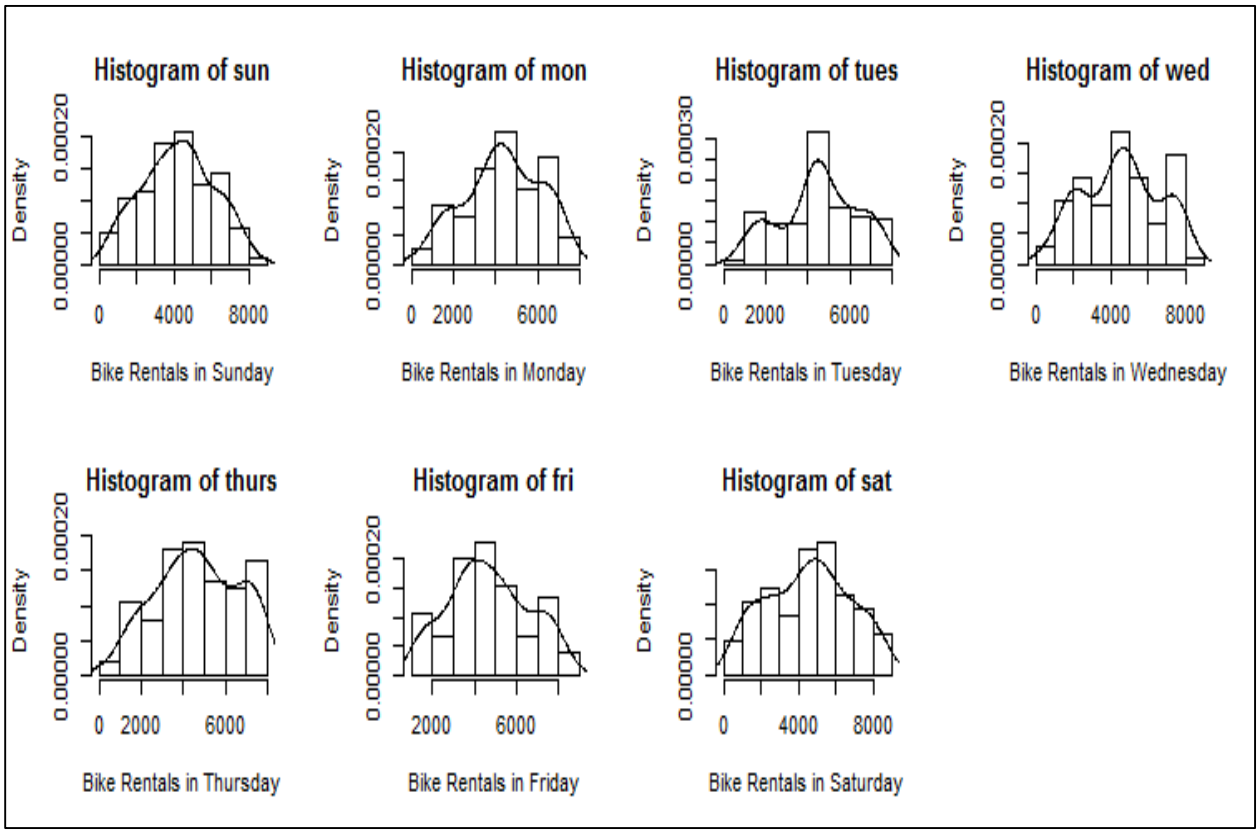


Fig. Bike Rentals by WeekDay

Pratima Kshetry, pkshetry@umd.edu

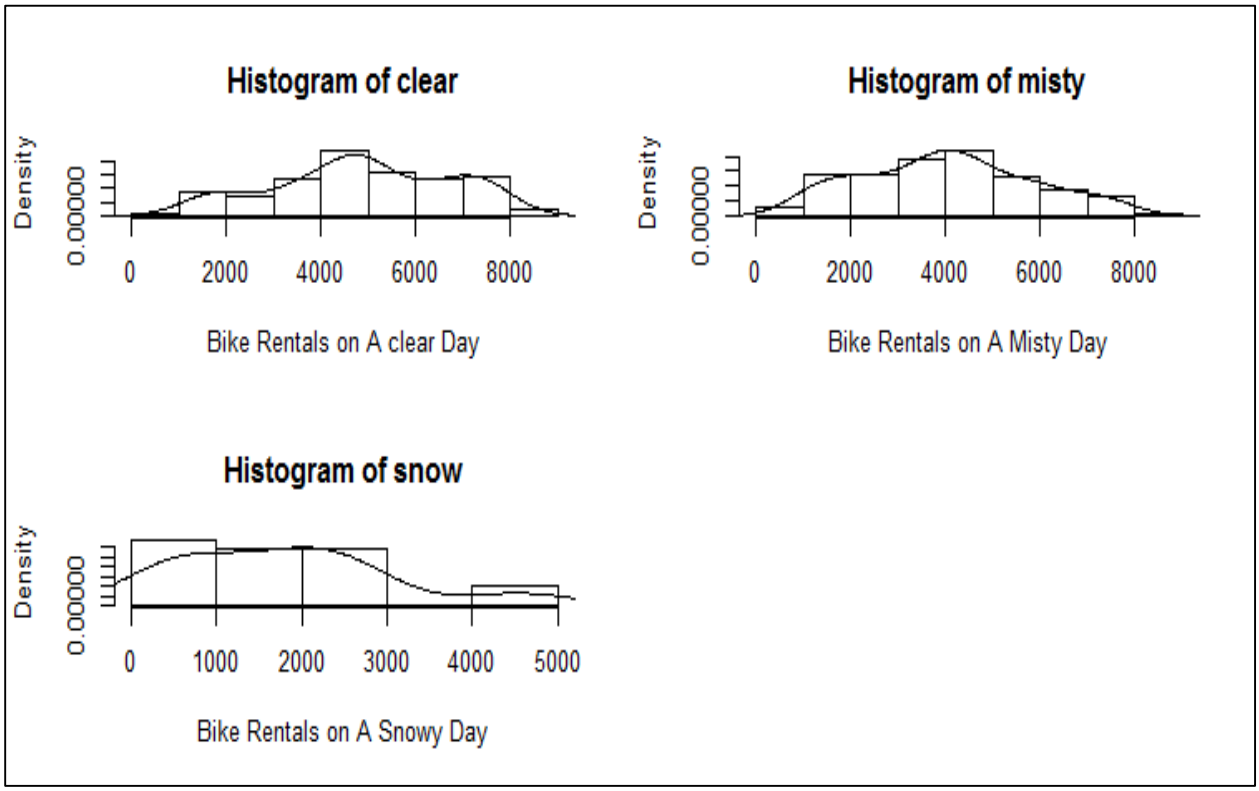


Fig. Bike Rentals by Weather type

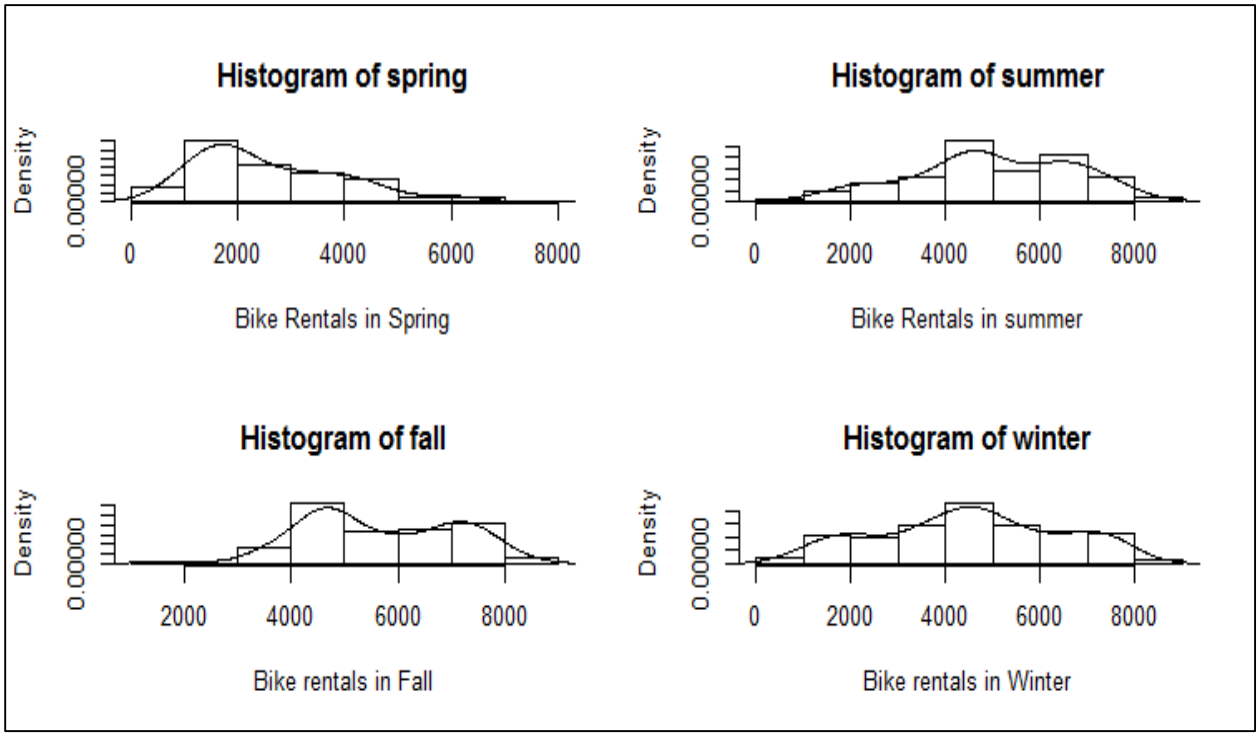


Fig. Bike Rentals by Season

b. Explain what (if anything) you had to do to make the data usable. This will likely be influenced by what visualizations analyses you want to do with your data.

Ans:

I found this dataset at the link : https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

The dataset was already clean, so I need not use or write any program code to clean the dataset. The only additional step I used while loading to dataset was to strip white spaces (if any in the dataset) using the command line as:

```
strip.white= TRUE
```

I also replaced the empty columns with NA values using the command as :

```
na.strings=c("NA",""))
```

The only challenge I faced during this assignment was to produce the multiple histograms side by side. I spent quite a few days to achieve this. Ultimately, I was able to find the solution with the codeline:

```
Par(mfrow=c(2,3))  ##Produces plotting area with 2 rows and 3 columns.
```