XI'AN JIAOTONG-LIVERPOOL UNIVERSITY

西 交 利 物 浦 大 学

YEAR 4

COURSE WORK SUBMISSION

Name	Lu	Kai-Yu
ID Number	*1614649*	
Programme	*Computer Science of Technology*	
Module Title	*Big Data Analytics*	
Module Code	*CSE313*	
Assignment Title	*Assignment 1: BDA process with R*	
Submission Deadline	*2019.10.11*	
Lecturer Responsible	*Gangmin Li*	

I certify that:

- I have read and understood the University's definitions of COLLUSION and PLAGIARISM (available in the Student Handbook of Xi'an Jiaotong-Liverpool University).

With reference to these definitions, I certify that:

- I have not colluded with any other student in the preparation and production of this work;
- this document has been written solely by me and in my own words except where I have clearly indicated and acknowledged that I have quoted or used figures from published or unpublished sources (including the web);
- where appropriate, I have provided an honest statement of the contributions made to my work by other people including technical and other support staff.

I understand that unauthorised collusion and the incorporation of material from other works without acknowledgement (plagiarism) are serious disciplinary offences.

Signature ……Kai-Yu Lu…… Date ……*2019.10.11*……

For Academic Office use:	Date Received	Days Late	Penalty

1. What I have found: (20 marks)

A. Day does not influence CTR much. For example, CTR of working days Day1- Day5 is similar to the one of the weekends Day6-Day7. This condition does not contain the factor of Gender or Age, it is the CTR with comprehensive factors like gender and age group.

B. Female have more interest in the advertisement then Male in week one generally. However, CTR of female and of male are very close.

C. Both gender of people who are in the Agecut (age) (3,18), (55,65) and (65,110) has larger CTR than people are in other ones, which means there is a higher possibility that juveniles and old people have more interests in the advertisement.

D. The distribution of CTR for each day are very similar.

2. Here my running source code with comments shows what I did in each step of BDA process (50 marks)

```
#This is the assignment one: BDA process with R.
#Student: Kai-Yu Lu. ID: 1614649.
#My process has 5 steps:
#Step 1: Data Acquisition. Read the data of a week, which is from Day1 to Day7.

#Step 2: Data Understanding. Impressions means the number of advertisements
#             requested by a web page to the server.
# #          Clicks means the number of the advertisements clicked by
#             users after seeing. CTR = Clicks/Impressions, which reflects
#             the level of interest in the advertisements on the web page.
##           This assignment will focus on CTR and explore its influence
#             factors (Day Gender and Age group).
#        ***KEY: This assignment will focus on CTR because it is not quite meaningful to solely
#        ***     explore Clicks or Impressions. CTR is a reasonable metrics for exploring the
#        ***     level of interest by people during this assignment.

#Step 3: Data Processing.
#        A. Combine the data of the 7 days.
#        B. Edit the attribute and values of data properly.
#            Attribute Day and CTR is added and the values of Gender
#             are changed from number into character.
#           CTR = Clicks/Impressions. Male represents 1 while Female represents 0.
#        C. Find the error data and select the reasonable data. People whose ages
#           are >=4 and <=110 and whose Impressions >0 are the desired/valid data.
#        D.  Gender and age group classification. Gender has male and female.
#             7 age categories: ("(3,18], (18,25], (25,35], (35,45], (45,55],
#             (55,65], (65,110]").

#Step 4. Data Analysis. Use ggplot and summaryBy for exploratory
#        and descriptive analyses.
#        It plots the bar chart of mean CTR for each day by Gender to
#         observe whether factors of Gender, Age group or Day influence CTR.
```

```r
#Step 5: Data Interpretation. Summary the data and write the report. The conclusion
#         are:
#          A. Day does not influence CTR much. For example, CTR of working days
#              Day1- Day5 is similar to the one of the weekends Day6-Day7.
#              This condition does not contain the factor of Gender or Age, it is
#              the CTR with comprehensive factors like gender and age group.
#          B. Female have more interest in the advertisement then Male in week one
#              generally. However, CTR of female and of male are very close.
#          C. Both gender of people who are in the Agecut(age) (3,18), (55,65) and
#              (65,110) has larger CTR than people are in other ones, which means
#              there is a higher possibility that juveniles and old people have more
#              interests in the advertisement.
#          D. The distribution of CTR for each day are very similar.

#These are the library will be used during this assignment.
install.packages("gridExtra")
install.packages("doBy")
install.packages("ggplot2")

library("doBy")
library("gridExtra")
library("ggplot2")

## Read the data of a week. (from Day1 to Day7)
nyt1 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt1.csv")
nyt2 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt2.csv")
nyt3 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt3.csv")
nyt4 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt4.csv")
nyt5 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt5.csv")
nyt6 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt6.csv")
nyt7 <- read.csv("C:/Users/dell/Desktop/dds_ch2_nyt/nyt7.csv")

## Attribute Day is added to each data.
nyt1$Day <- "Day1"
nyt2$Day <- "Day2"
nyt3$Day <- "Day3"
nyt4$Day <- "Day4"
nyt5$Day <- "Day5"
nyt6$Day <- "Day6"
nyt7$Day <- "Day7"

##Combine the data of the 7 days.
nyt <- rbind(nyt1,nyt2,nyt3,nyt4,nyt5,nyt6,nyt7)

##Change the values of the attribute Gender. Male represents 1 while Female represents 0.
nyt$Gender <-ifelse(nyt$Gender == 1, "Male","Female")

##People who are under 4 years old or over 110 years old are assumed as invalid users.
```

```
nyt <- subset(nyt,Age>=4&Age<=110)

##People whose Impressions are 0 are assumed as invalid users. Later CTR will not have NA.
nyt <- subset(nyt,Impressions>=1)

## Group 7 age categories according to the requirement. ("(3,18], (18,25], (25,35], (35,45], (45,55], (55,65] ,(65,110]")
##Here it avoids errors. For example, if the maxmimum of the Age was very large, user can edit the maximum age until a reasonable one is satisfied.
nyt$Agecut <- cut(nyt$Age, c(3,18,25,35,45,55,65,110))

## Add attribute CTR = Clicks/Impressions.
nyt$CTR <- (nyt$Clicks)/(nyt$Impressions)

##CTR set as NA can be assumed as 0.
nyt$CTR[is.na(nyt$CTR)]<-0

##Plot the bar chart of mean CTR for each day by Gender and Age group.
d1 <- ggplot(subset(nyt,nyt$Day=="Day1"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day1: Distributions of number CTR by Gender and Age groups.")
d2 <- ggplot(subset(nyt,nyt$Day=="Day2"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day2: Distributions of number CTR by Gender and Age groups.")
d3 <- ggplot(subset(nyt,nyt$Day=="Day3"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day3: Distributions of number CTR by Gender and Age groups.")
d4 <- ggplot(subset(nyt,nyt$Day=="Day4"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day4: Distributions of number CTR by Gender and Age groups.")
d5 <- ggplot(subset(nyt,nyt$Day=="Day5"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day5: Distributions of number CTR by Gender and Age groups.")
d6 <- ggplot(subset(nyt,nyt$Day=="Day6"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day6: Distributions of number CTR by Gender and Age groups.")
d7 <- ggplot(subset(nyt,nyt$Day=="Day7"), aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Day7: Distributions of number CTR by Gender and Age groups.")

##Combine the 7 bar charts in one figure. More convenient for observing CTR for
# each day by Gender and Age group. Shown in Figure 1.
grid.arrange(d1,d2,d3,d4,d5,d6,d7,ncol=2,nrow=4)

##Plot the performance of CTR during one week by Day. Shown in Figure 2.
ggplot(nyt, aes(x=Day, y=CTR)) +geom_bar(stat="summary",fun.y="mean")+labs(title = "Distributions of number CTR by Day")
##Plot the performance of CTR during one week by Gender. Shown in Figure 3.
```

```r
ggplot(nyt, aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+labs(title = "Distributions of number CTR for week one by Gender")

##Plot the total performance of CTR during one week by Gender and Age group.
# Shown in Figure 4.
ggplot(nyt, aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Agecut)+labs(title = "Distributions of number CTR for 7 days by Gender and Age groups")

##Plot the total performance of CTR during one week by Day and Gender.
Shown in Figure 5.
ggplot(nyt, aes(x=Gender,y=CTR,fill=Gender)) +geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Day)+labs(title = "Distributions of number CTR for 7 days by Day and Gender")
##Plot the total performance of CTR during one week by Day and Age group.
Shown in Figure 6.
ggplot(nyt, aes(x=Agecut,y=CTR))
+geom_bar(stat="summary",fun.y="mean")+facet_grid(.~Day)+labs(title = "Distributions of number CTR for 7 days by Day and Age groups")

##Print the mean CTR for each day by Day.
summaryBy(CTR~Day, data = nyt, FUN=mean)

##Print the mean CTR for each day by Gender.
summaryBy(CTR~Gender, data = nyt, FUN=mean)

##Print the mean CTR for each day by Gender and Age group.
summaryBy(CTR~Gender+Agecut, data = nyt, FUN=mean)

##Print the mean CTR for each day by Day and Gender.
summaryBy(CTR~Day+Gender, data = nyt, FUN=mean)

##Print the mean CTR for each day by Day and Age group.
summaryBy(CTR~Day+Agecut, data = nyt, FUN=mean)
```

3. Plots produced from the code and data evidence with explanation

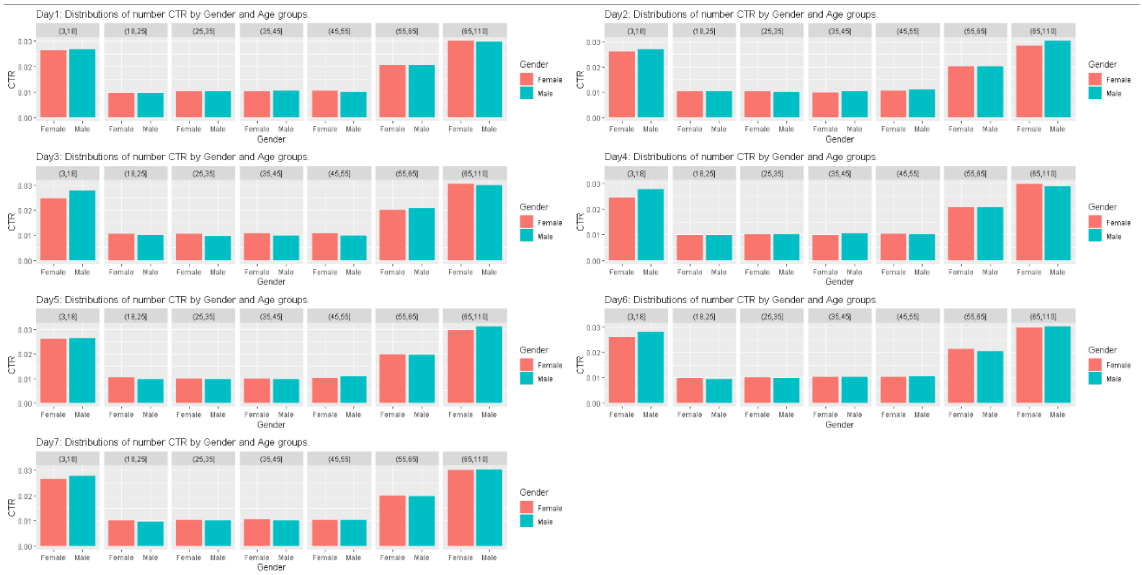

Figure 1: Bar chart of mean CTR for each day by Gender and Age group.

Figure 1 generally shows that distribution of CTR for each day are very similar. Additionally, CTR of female and of male are very close. CTR of male in (3,18) is a little higher than the one of female in (3,18). Both gender of people who are in the Agecut (age) (3,18), (55,65) and (65,110) has larger CTR than people are in other ones. In order to obtain more accurate summaries, below will explore the relations of CTR between Age group, Day and Gender.

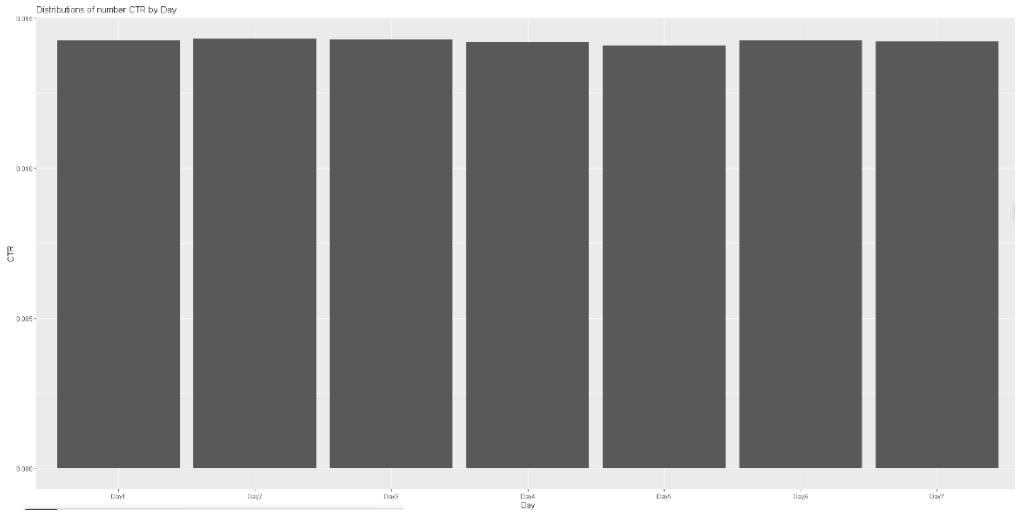

Figure 2: Bar chart of mean CTR from Day1 to Day7 by Day.

Day	CTR.mean
Day1	0.01425364
Day2	0.01430817
Day3	0.01430817
Day4	0.01420272
Day5	0.01407678
Day6	0.01426051
Day7	0.01422619

Table 1: Mean CTR from Day1 to Day7 by Day.

Table 1 and Figure 2 show that Day does not influence CTR much. For example, CTR of

working days Day1- Day5 is similar to the one of the weekends Day6-Day7. This condition does not contain the factor of Gender or Age, it is the CTR with comprehensive factors like gender and age group.

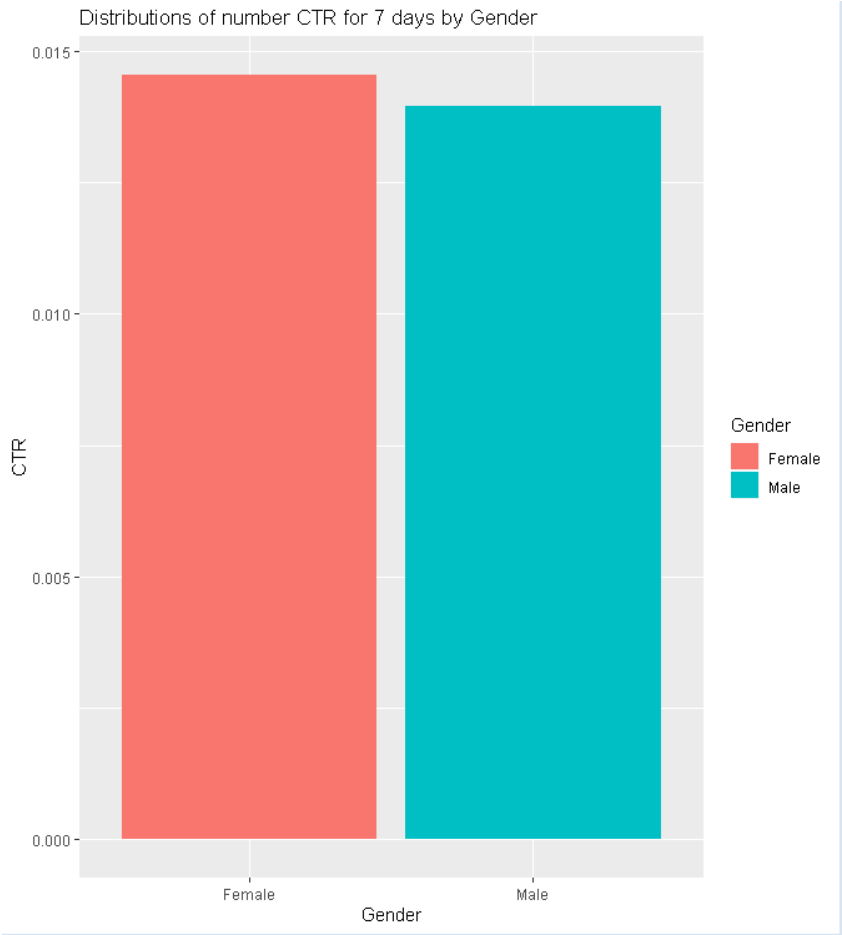

Figure 3: Bar chart of mean CTR for week one by Gender.

Gender	CTR.mean
Female	0.01454922
Male	0.01395263

Table 2: Mean CTR for week one by Gender.

Table 2 and Figure 3 show that Female have more interest in the advertisement then Male in week one in general. However, CTR of female and of male are very close in general.

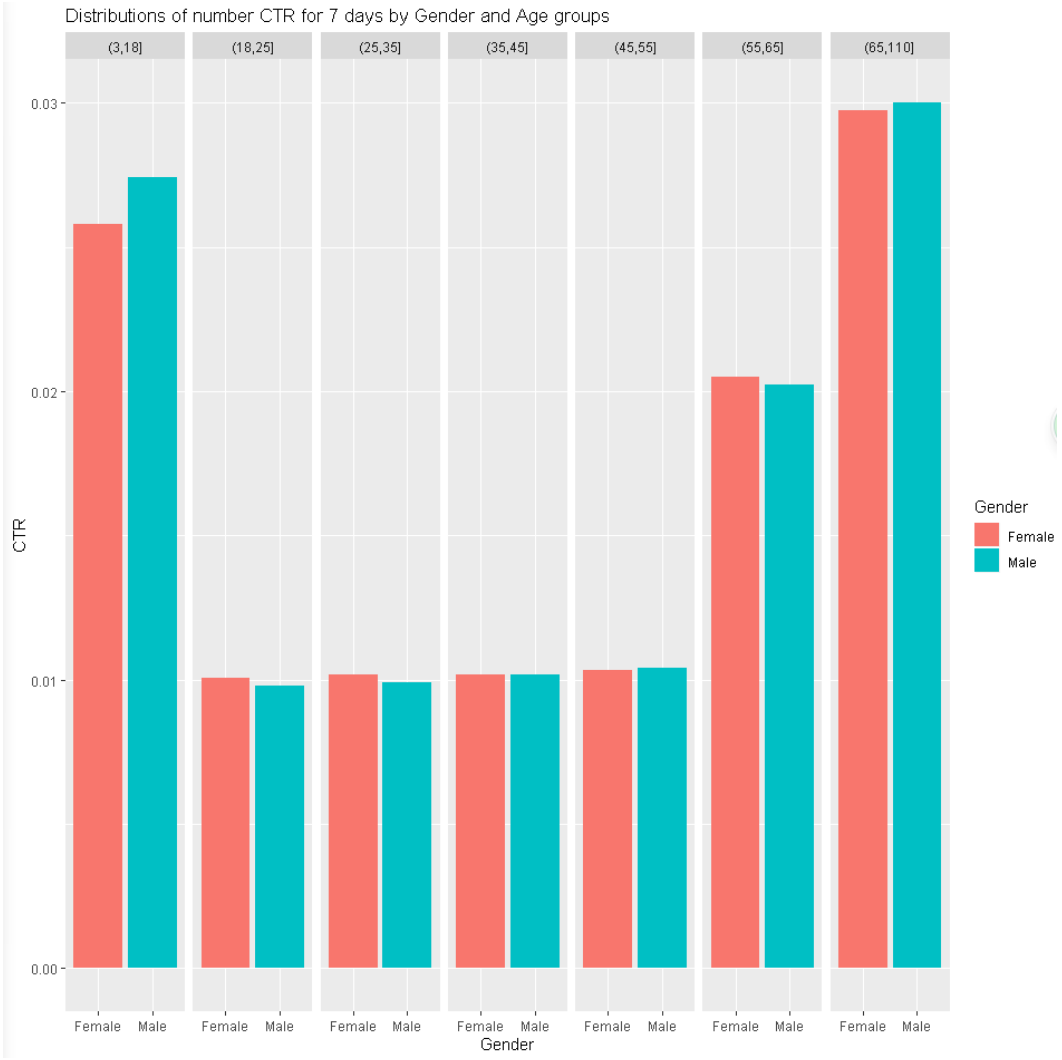

Table 4: Bar chart of mean CTR for week one by Gender and Age group.

Gender	Agecut	CTR.mean
Female	(3,18)	0.025802946
Female	(18,25)	0.010050765
Female	(25,35)	0.010188386
Female	(35,45)	0.010180128
Female	(45,55)	0.010339914
Female	(55,65)	0.020495570
Female	(65,110)	0.029752458
Male	(3,18)	0.027405119
Male	(18,25)	0.009779047
Male	(25,35)	0.009902555
Male	(35,45)	0.010164314
Male	(45,55)	0.010417945
Male	(55,65)	0.020247506
Male	(65,110)	0.030025755

Table 3: Mean CTR for week one by Gender and Age group

Table 3 and Figure 4 shows both gender of people who are in the Agecut (age) (3,18), (55,65) and (65,110) has larger CTR than people are in other ones, which means there is a higher

possibility that juveniles and old people have more interests in the advertisement.

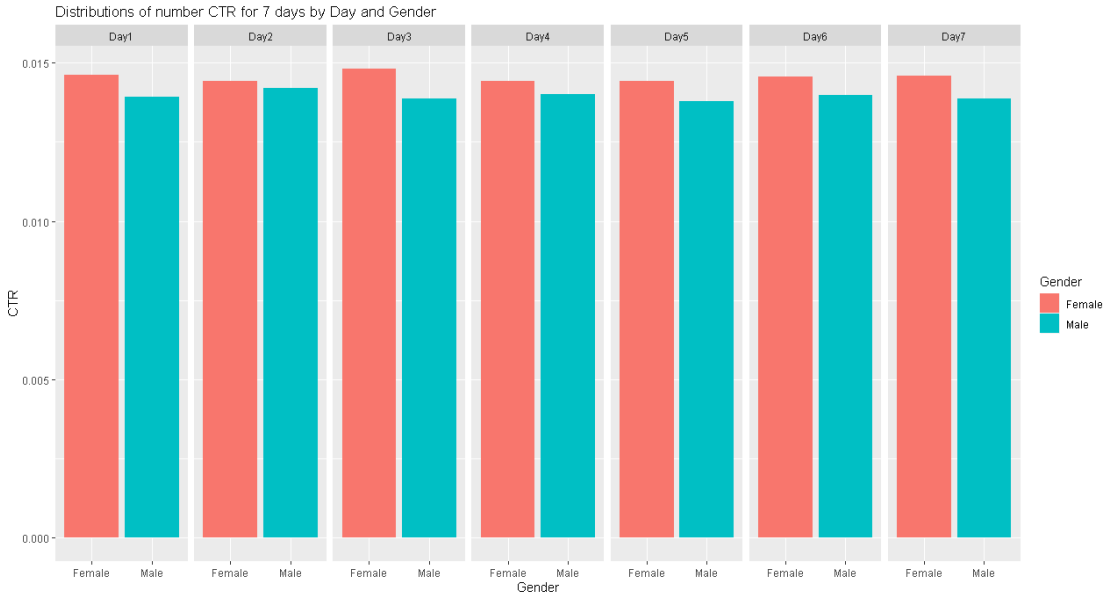

Figure 5: Bar chart of mean CTR for week one by Day and Gender.

Day	Gender	CTR.mean
Day1	Female	0.01462201
Day1	Male	0.01391852
Day2	Female	0.01443014
Day2	Male	0.01420175
Day3	Female	0.01479656
Day3	Male	0.01385505
Day4	Female	0.01441671
Day4	Male	0.01401181
Day5	Female	0.01441556
Day5	Male	0.01377241
Day6	Female	0.01454940
Day6	Male	0.01398000
Day7	Female	0.01458570
Day7	Male	0.01387119

Table 4: Mean CTR for week one by Day and Gender

Table 4 and Figure 5 further shows that Female have more interest in the advertisement then Male in week one, which improves the persuasion of the observation from Table 2.

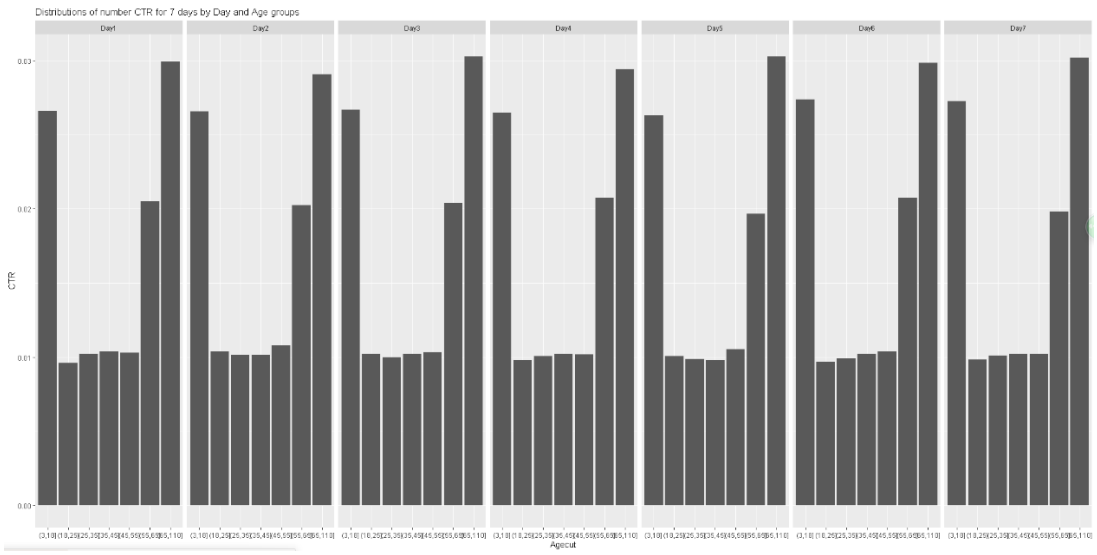

Figure 6: Bar chart of mean CTR for week one by Day and Age group.

Day	Agecut	CTR.mean
Day1	(3,18)	0.026620504
Day1	(18,25)	0.009593848
Day1	(25,35)	0.010216306
Day1	(35,45)	0.010352815
Day1	(45,55)	0.010276707
Day1	(55,65)	0.020521777
Day1	(65,110)	0.029924905
Day2	(3,18)	0.026587556
Day2	(18,25)	0.010378736
Day2	(25,35)	0.010150432
Day2	(35,45)	0.010148746
Day2	(45,55)	0.010792029
Day2	(55,65)	0.020235591
Day2	(65,110)	0.029072809
Day3	(3,18)	0.026680313
Day3	(18,25)	0.010212934
Day3	(25,35)	0.009991738
Day3	(35,45)	0.010204675
Day3	(45,55)	0.010321056
Day3	(55,65)	0.020388236
Day3	(65,110)	0.030272173
Day4	(3,18)	0.026476149
Day4	(18,25)	0.009798576
Day4	(25,35)	0.010070849
Day4	(35,45)	0.010215623
Day4	(45,55)	0.010179142
Day4	(55,65)	0.020735766
Day4	(65,110)	0.029421246
Day5	(3,18)	0.026313913
Day5	(18,25)	0.010070474

Day5	(25,35)	0.009866112
Day5	(35,45)	0.009772040
Day5	(45,55)	0.010512846
Day5	(55,65)	0.019677917
Day5	(65,110)	0.030263825
Day6	(3,18)	0.027386361
Day6	(18,25)	0.009668980
Day6	(25,35)	0.009917196
Day6	(35,45)	0.010195918
Day6	(45,55)	0.010383663
Day6	(55,65)	0.020762342
Day6	(65,110)	0.029855455
Day7	(3,18)	0.027271391
Day7	(18,25)	0.009834782
Day7	(25,35)	0.010098257
Day7	(35,45)	0.010219978
Day7	(45,55)	0.010219994
Day7	(55,65)	0.019834934
Day7	(65,110)	0.030197550

Table 5: Mean CTR for week one by Day and Age group

Table 5 and Figure 6 does not have the consideration of Gender and it shows people who are in the Agecut (age) (3,18), (55,65) and (65,110) has larger CTR than people are in other ones, which enhances the stringency of conclusion that there is a higher possibility that juveniles and old people have more interests in the advertisement.