HW1

Answer 1

a) `setwd("C:/Users/Yashika Bajaj/Downloads/all_nyt")`

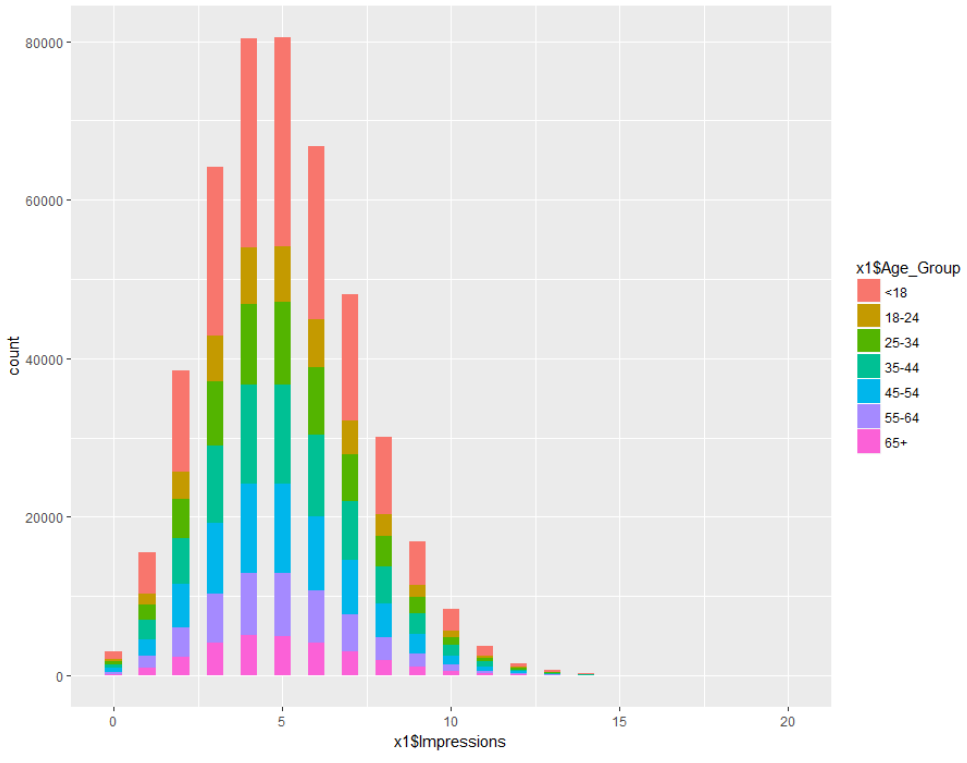
```
> Y=lapply(dir(),read.csv)
> data=do.call("rbind",Y)
>attach(data)
> data$Age_Group<- ifelse(Age<18, "<18", ifelse((Age>=18) & (Age<25), "18-24", ifelse((Age>24) & (Age<35), "25-34", ifelse((Age>34) & (Age<45), "35-44", ifelse((Age>44) & (Age<55), "45-54", ifelse((Age>54) & (Age<65), "55-64", "65+"))))))
```

data × 2.R × S ×

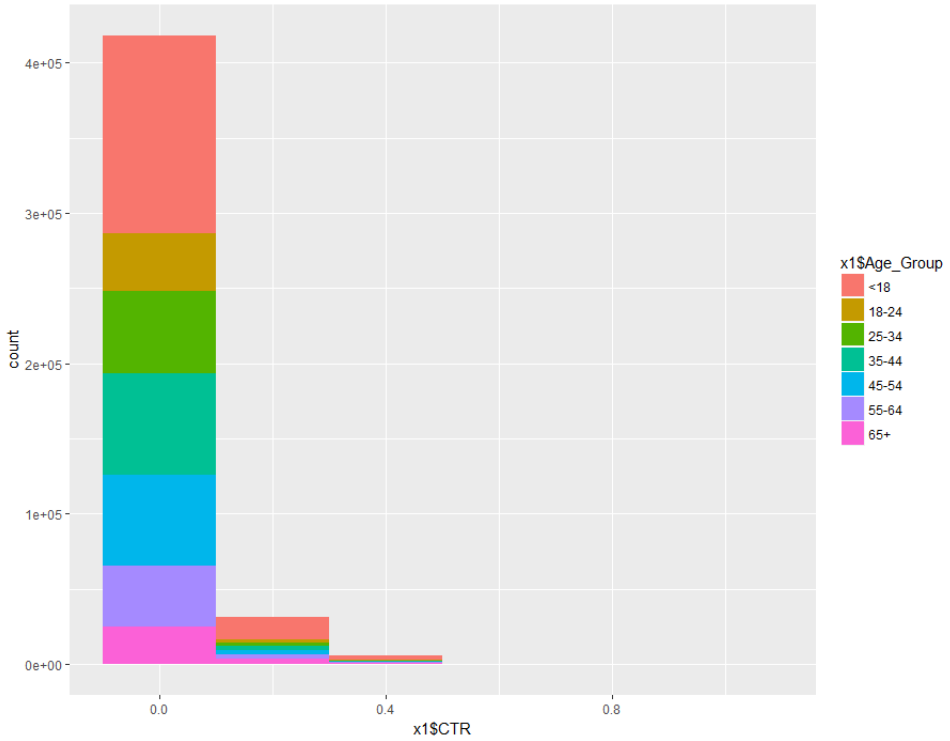
Filter

	Age	Gender	Impressions	Clicks	Signed_In	Day	Age_Group
21	59	1	4	0	1	1	55-64
22	61	0	6	0	1	1	55-64
23	48	0	7	0	1	1	45-54
24	29	1	2	0	1	1	25-34
25	0	0	4	0	0	1	<18
26	19	1	4	0	1	1	18-24
27	19	0	3	0	1	1	18-24

b) i) `ggplot(x1, aes(x=x1$Impressions, fill=x1$Age_Group)) + geom_histogram(binwidth=.5)`



```
ggplot(x1, aes(x=x1$CTR, fill=x1$Age_Group)) + geom_histogram(binwidth=.2)
```



ii) `data$Click_behaviour<- ifelse(Clicks==0, 0, 1)`

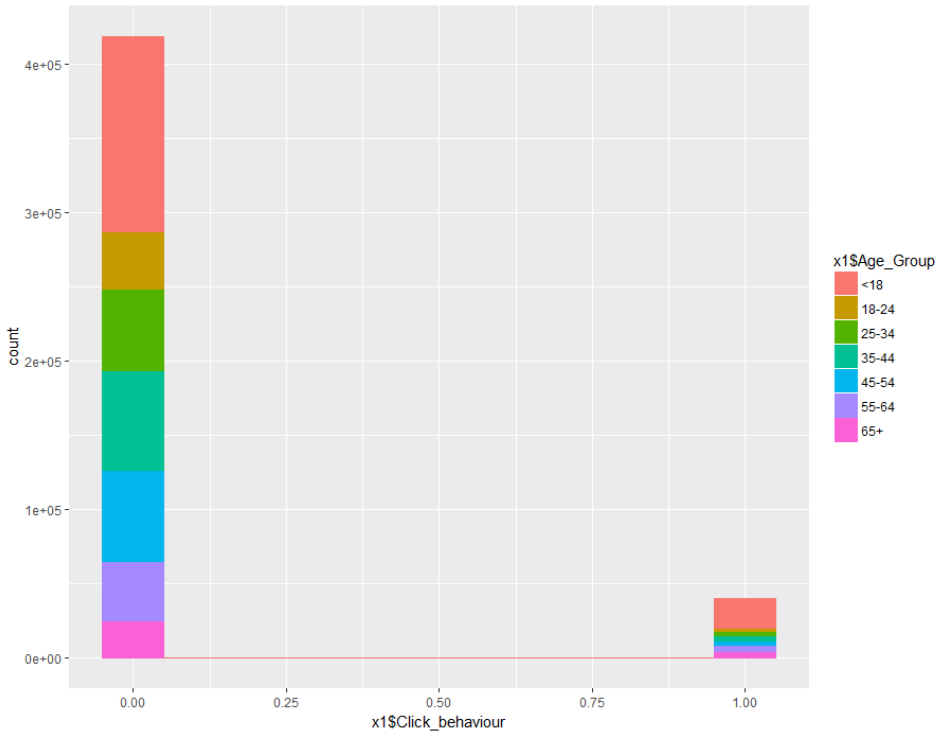
data × 2.R × S ×

Filter

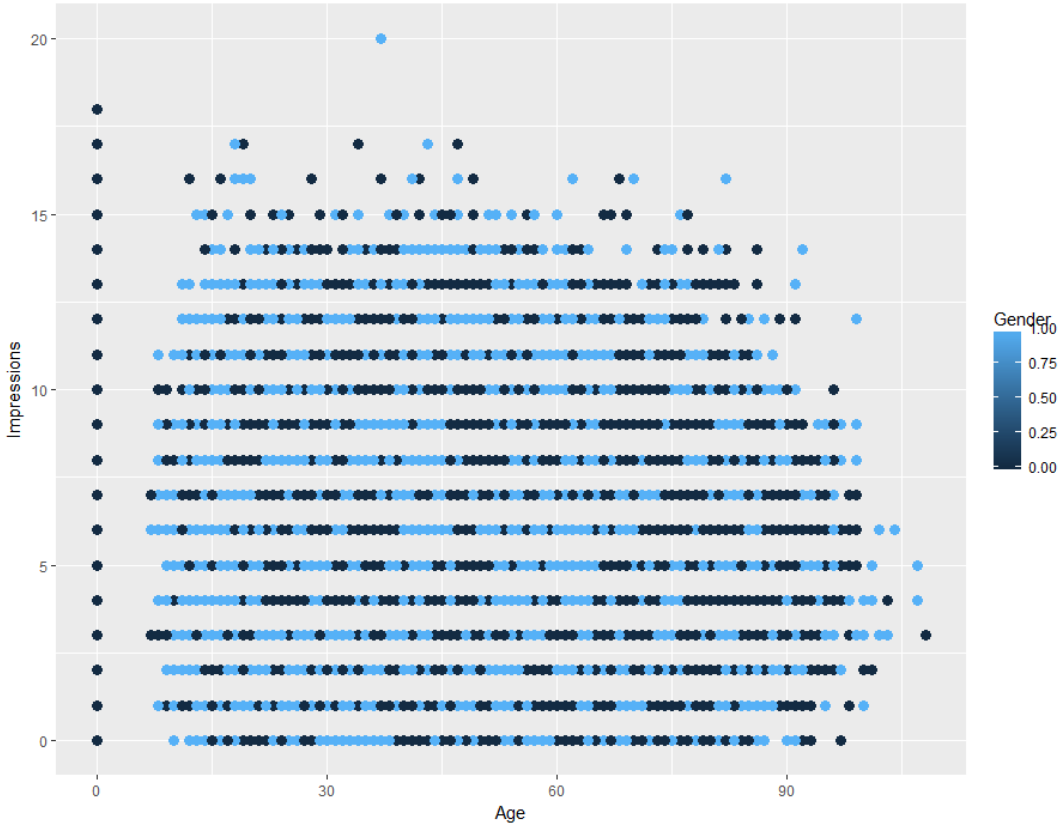
	Age	Gender	Impressions	Clicks	Signed_In	Day	Age_Group	Click_behaviour
20	0	0	5	0	0	1	<18	0
21	59	1	4	0	1	1	55-64	0
22	61	0	6	0	1	1	55-64	0
23	48	0	7	0	1	1	45-54	0
24	29	1	2	0	1	1	25-34	0
25	0	0	4	0	0	1	<18	0
26	19	1	4	0	1	1	18-24	0
27	19	0	3	0	1	1	18-24	0
28	48	1	9	0	1	1	45-54	0
29	48	1	4	0	1	1	45-54	0
30	21	1	5	0	1	1	18-24	0
31	23	0	4	0	1	1	18-24	0

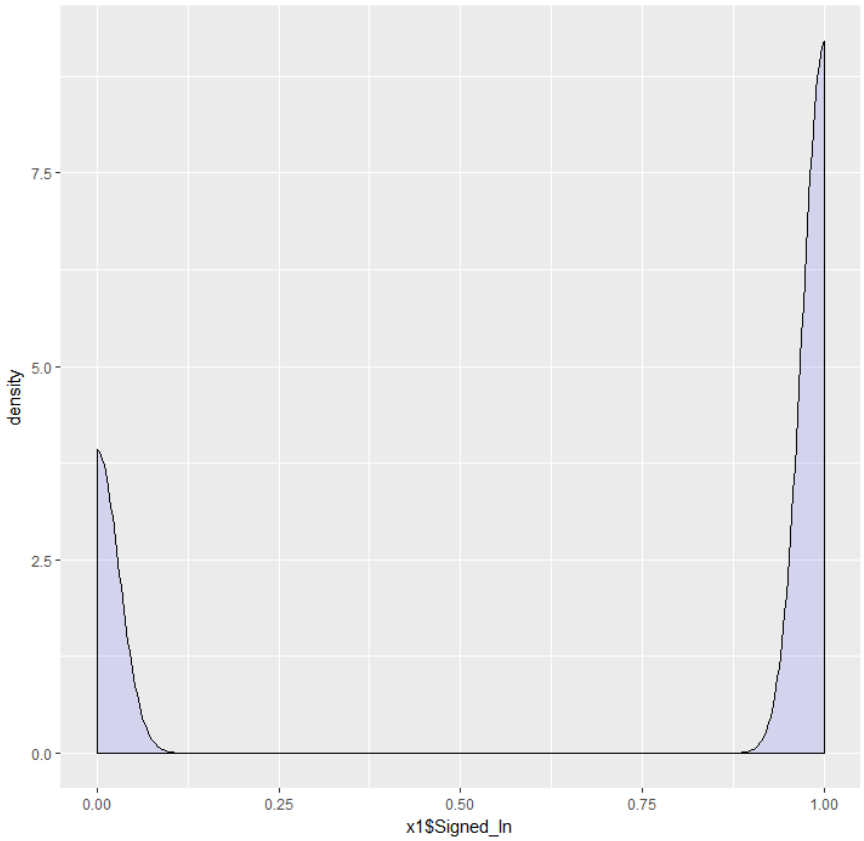
iii)

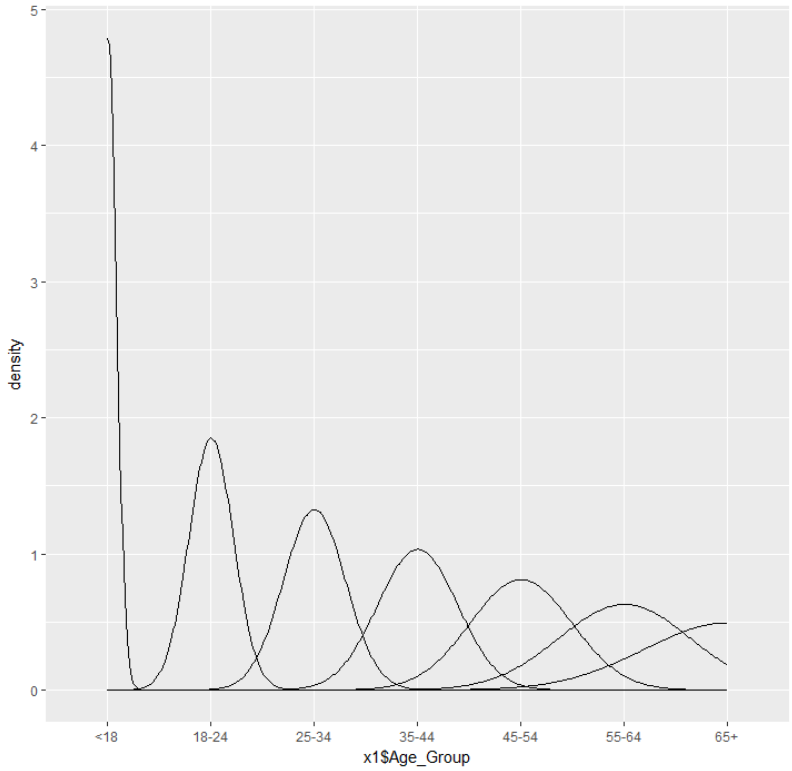
```
ggplot(x1, aes(x=x1$Clicked, color=x1$Age_Group,fill=d$Age_Group)) +geom_hist
ogram(binwidth=0.1)
```




```
> ggplot(x1,aes(Age,Impressions,color=Gender))+ geom_point(size=3)
```




7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
x1$Signed_ln

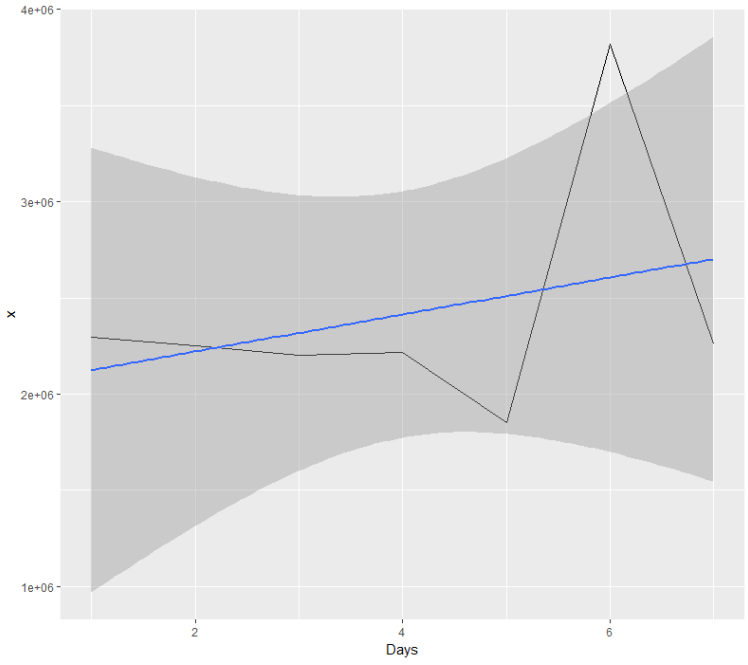

5
4
3
2
1
0
density
<18
18-24
25-34
35-44
45-54
55-64
65+
x1$Age_Group

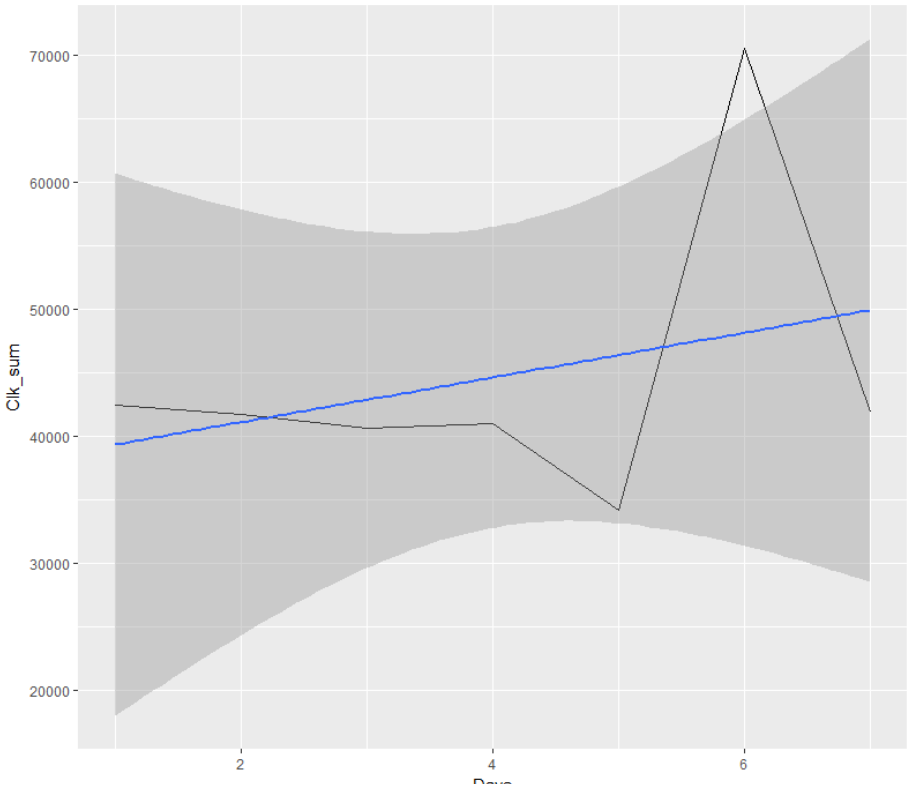
c) Metrics/measurements/statistics that summarize the data

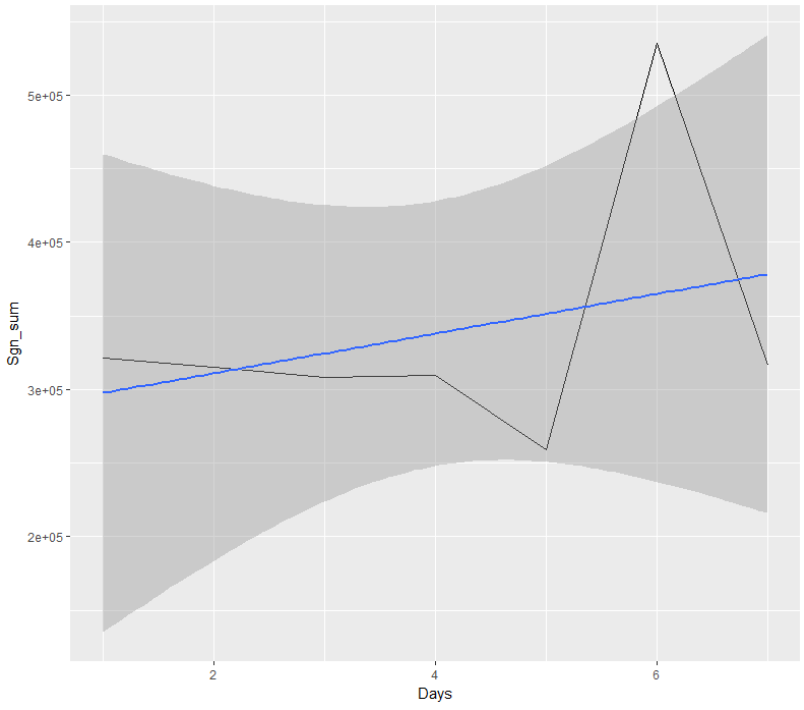
```
> summary(data)
      Age            Gender        Impressions      Clicks          Signed_In           Day          Age_Group
 Min.   :  0.00   Min.   :0.0000   Min.   : 0   Min.   :0.00000   Min.   :0.0000   Min.   : 1.00   Length:14905865
 1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.: 3   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 8.00   Class :character
 Median : 26.00   Median :0.0000   Median : 5   Median :0.00000   Median :1.0000   Median :16.00   Mode  :character
 Mean   : 26.24   Mean   :0.3231   Mean   : 5   Mean   :0.09773   Mean   :0.6234   Mean   :15.98
 3rd Qu.: 46.00   3rd Qu.:1.0000   3rd Qu.: 6   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:24.00
 Max.   :115.00   Max.   :1.0000   Max.   :21   Max.   :6.00000   Max.   :1.0000   Max.   :31.00

 Click_behaviour        CTR          weekday
 Min.   :0.00000   Min.   :0.00    Length:14905865
 1st Qu.:0.00000   1st Qu.:0.00    Class :character
 Median :0.00000   Median :0.00    Mode  :character
 Mean   :0.09134   Mean   :0.02
 3rd Qu.:0.00000   3rd Qu.:0.00
 Max.   :1.00000   Max.   :1.00
                   NA's   :1e+05
```

Metrics and distributions for one week.

```
setwd("C:/Users/Yashika Bajaj/Downloads/all_nyt")
> Y=lapply(dir(),read.csv)
> data=do.call("rbind",Y)
> attach(data)
> data$Age_Group<- ifelse(Age<18, "<18", ifelse((Age>=18) & (Age<25), "18-24"
, ifelse((Age>24) & (Age<35), "25-34", ifelse((Age>34) & (Age<45), "35-44", i
felse((Age>44) & (Age<55), "45-54", ifelse((Age>54) & (Age<65), "55-64", "65+
"))))))
> data$Click_behaviour<- ifelse(Clicks==0, 0, 1)
> data$CTR<- (Clicks/Impressions)
> View(data)
> Imp_sum <- aggregate(data$Impressions, by=list(data$Day),sum)
Clk_sum <- aggregate(data$Clicks, by=list(data$Day),sum)
> Sgn_sum <- aggregate(data$Signed_In, by=list(data$Day),sum)
> library(ggplot2)
ggplot(Imp_sum, aes(Group.1, x)) + geom_line() + xlab("Days") + geom_smooth(m
ethod=lm)
ggplot(Clk_sum, aes(Group.1, x)) + geom_line() + xlab("Days") + ylab("Clk_sum
")+ geom_smooth(method=lm)
> ggplot(Sgn_sum, aes(Group.1, x)) + geom_line() + xlab("Days") + ylab("Sgn_s
um")+ geom_smooth(method=lm)
```







d) All the 3 graphs show a pattern in impressions, signed in and clicks with all of the three highest on the same day and lowest on the same day. Since it's the 6th day it will be Saturday

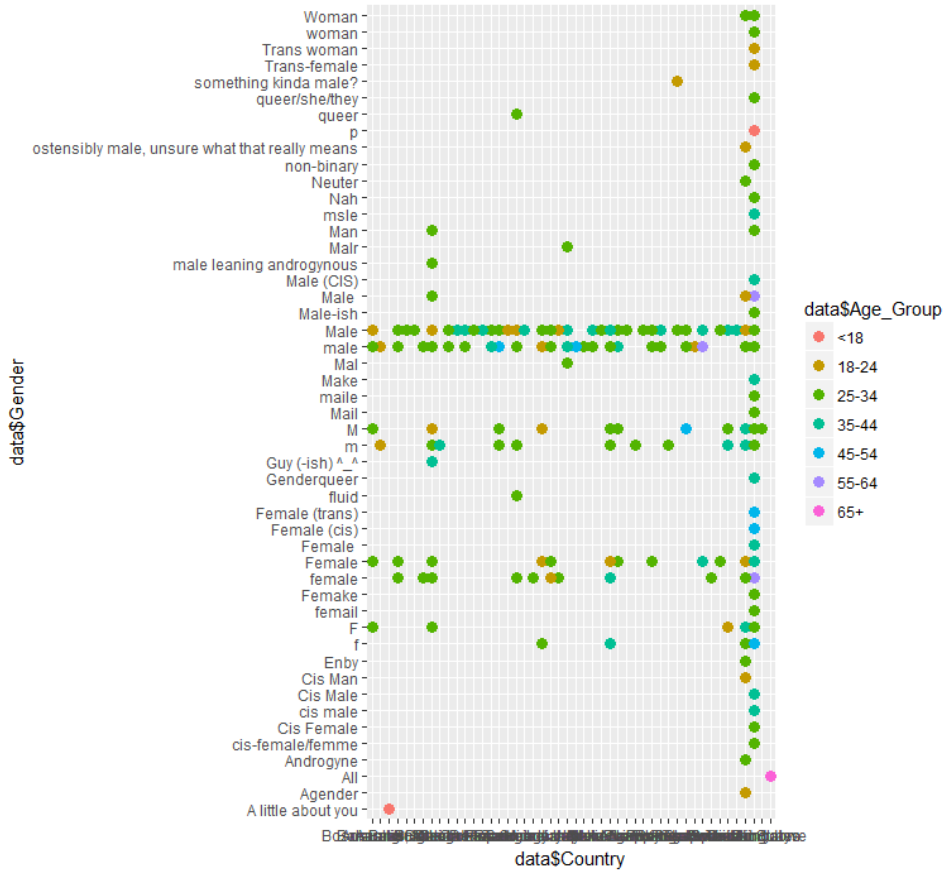
which is quite sensible because it's a holiday that day. People are free, and they want toget updated therefore they might read news.

Answer 2

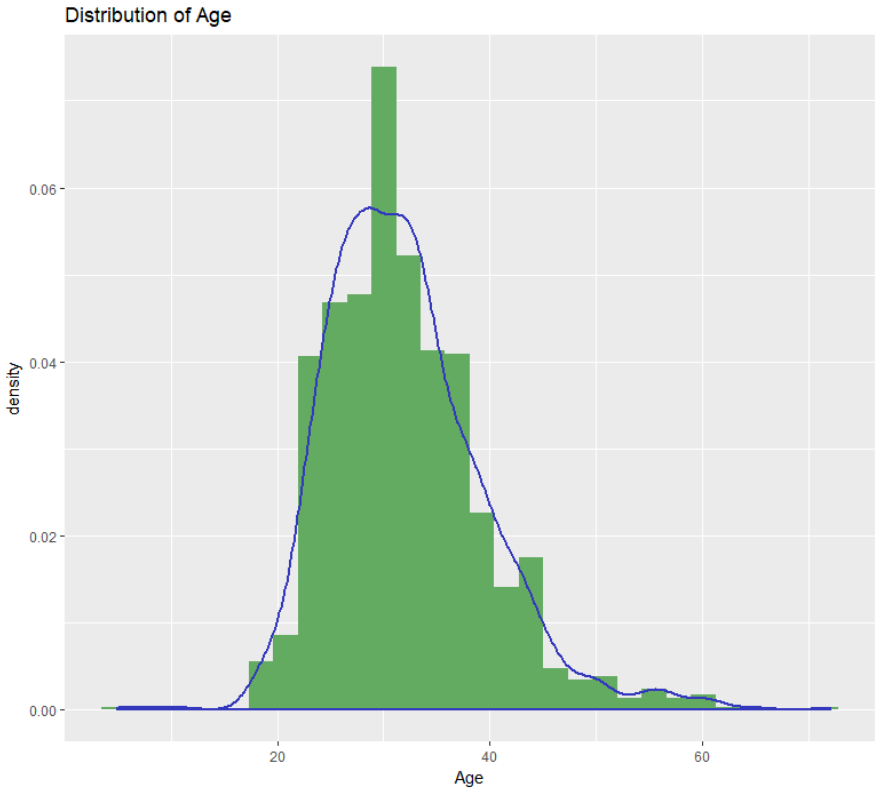
a) I have selected my data about the Mental Health in Tech Survey – 2014 from Kaggle. My data had 27 variables. The primary objective of my study was to see if there is a relationship between the family history of people having mental problems and people undergoing treatment right now (when the survey was done). Also, I wanted to see that people involved in which profession have undergone the highest treatment. And, if anybody is undergoing a treatment does it effect his/her work?

b) Following are some of the visualizations describing my data: -

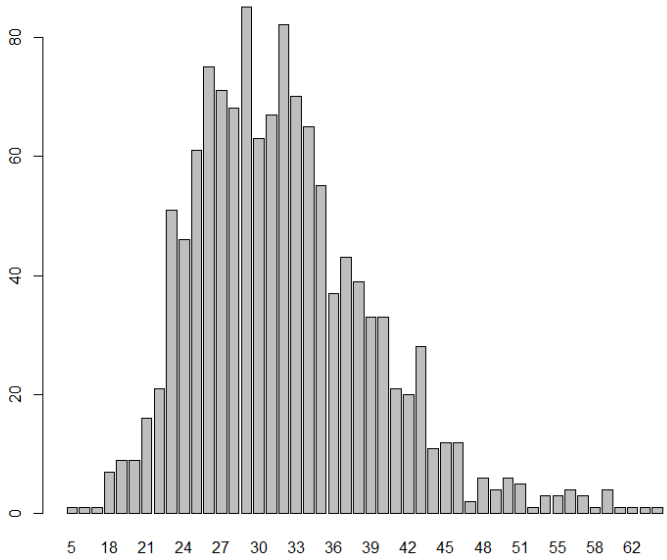
```
  ggplot(data, aes(data$Country, data$Gender, color= data$Age_Group))+geom_po
int(size=3)
```




```
ggplot(A,aes(x=Age))+geom_histogram(aes(y=..density..),fill="#62AB61")+geom_d
ensity(col="#3438BD",size=1)+labs(x="Age",title="Distribution of Age")
```



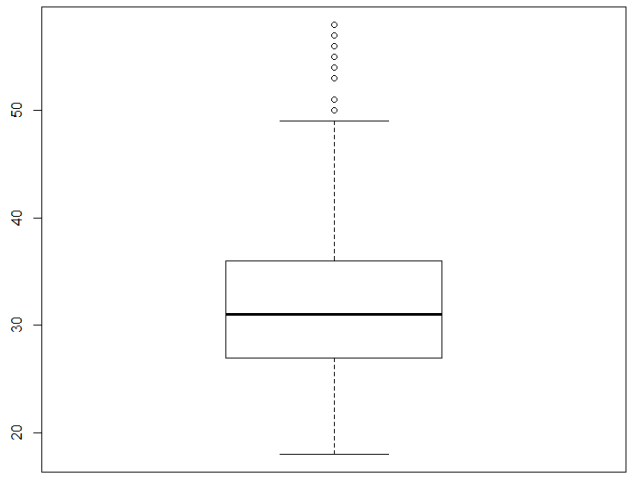

```
Barplot(table(A$Age))
```



There are few outliners here.

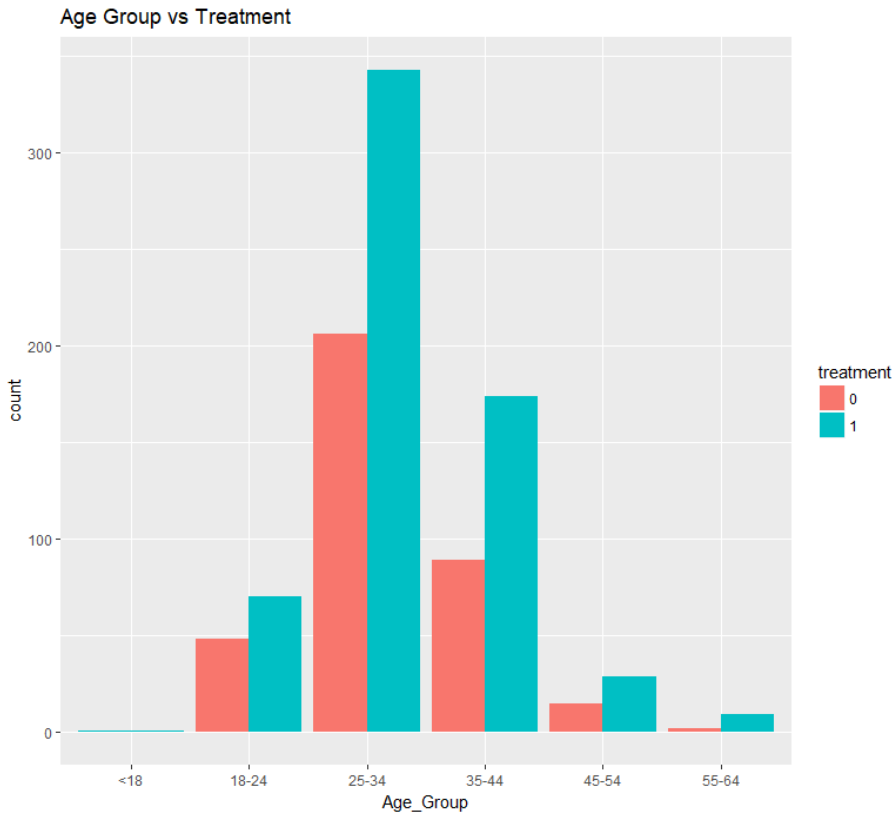
I had to clean the data for it to make some sense.

```
outlier_age <- subset(A[2:4], Age < 16 | Age > 75 )
> nrow(outlier_age)
[1] 3
> age_clean <- subset(A, Age > 16 & Age <60)
> dim(age_clean)
[1] 1248   29
> boxplot(age_clean$Age)
```

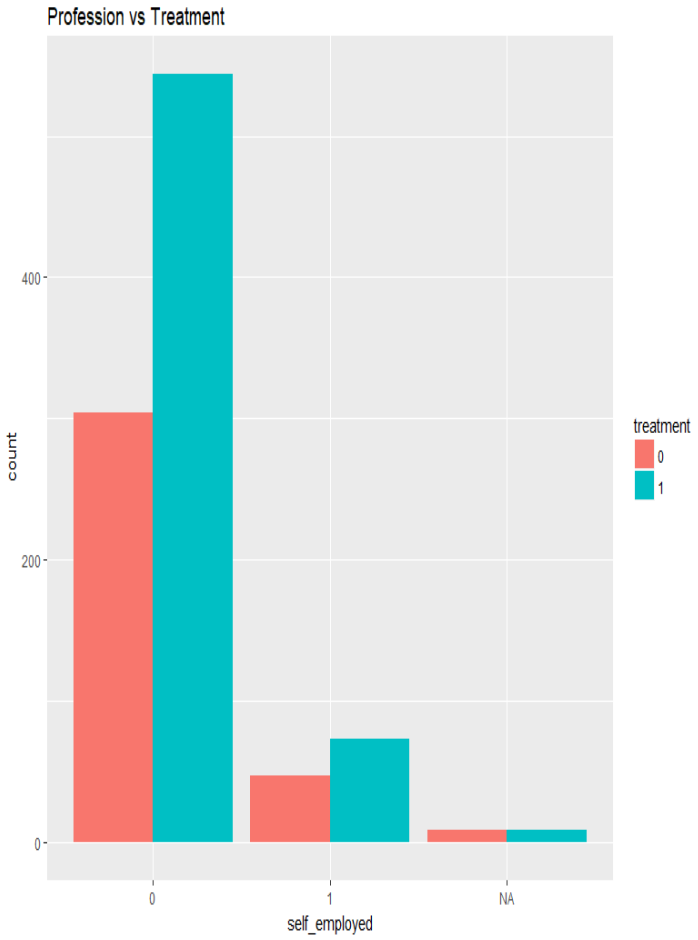


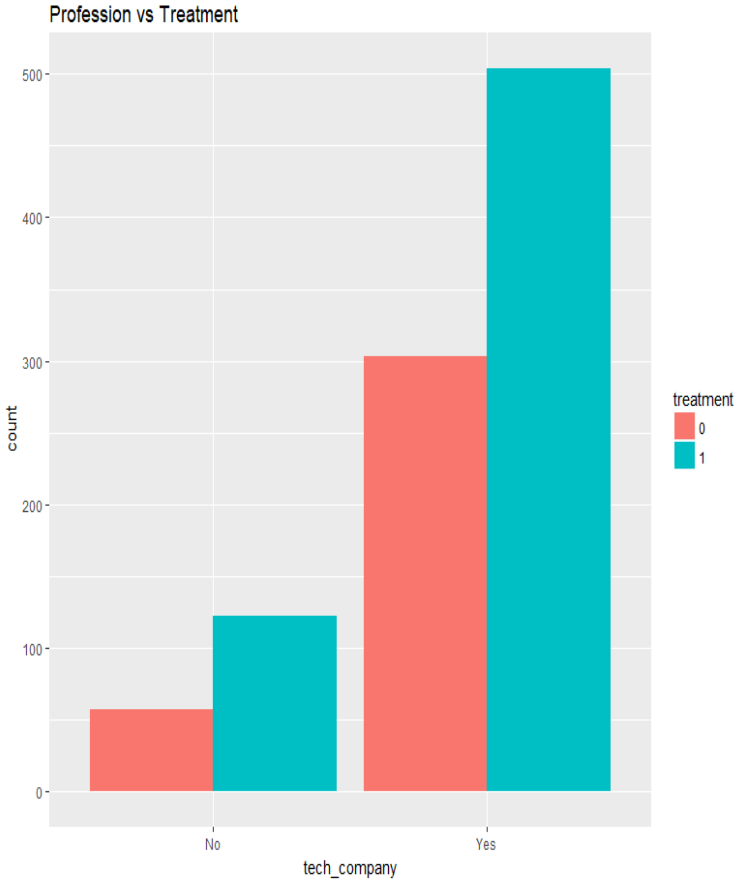
The median age is clearly visible near 30.

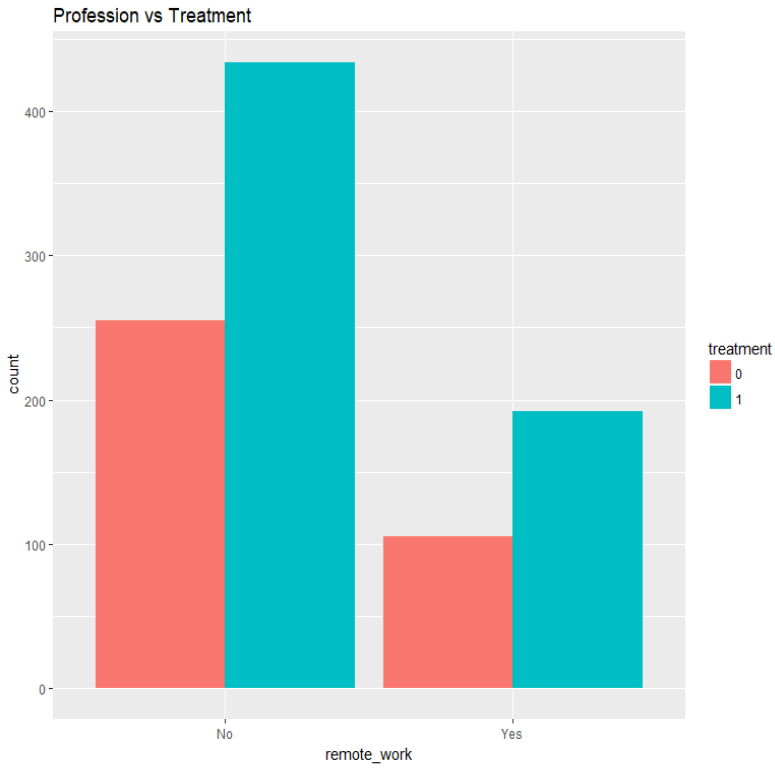
```
prop.table(table(as.factor(A$have_history))
+ )

        0         1
0.6092137 0.3907863

summary(A$Age)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.726e+03  2.700e+01  3.100e+01  7.943e+07  3.600e+01  1.000e+11
> A$Age[which(A$Age<0)]<-20
>
> A$Age[which(A$Age>100)]<-60
> summary(A$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   27.00   31.00   32.03   36.00   72.00

data <- age_clean
```



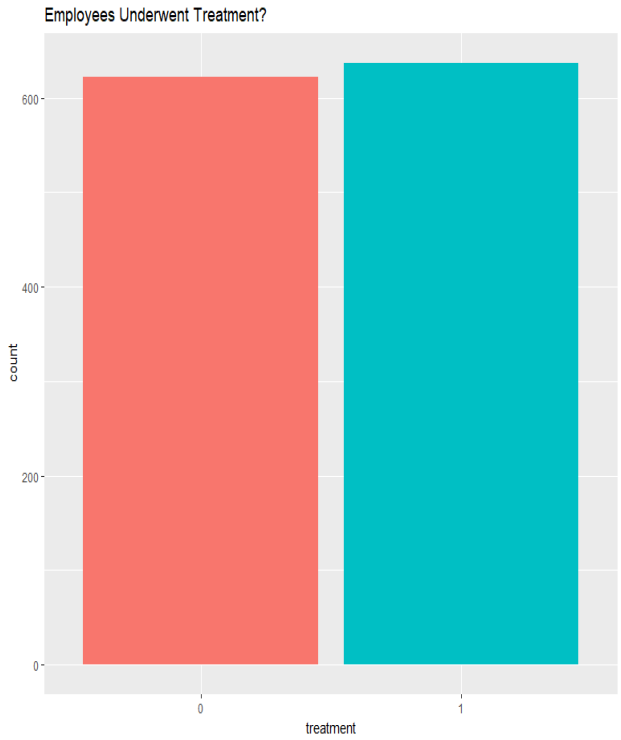
The graph shows that the Age Group 25-34 have the maximum patients undergoing treatment of mental illness.

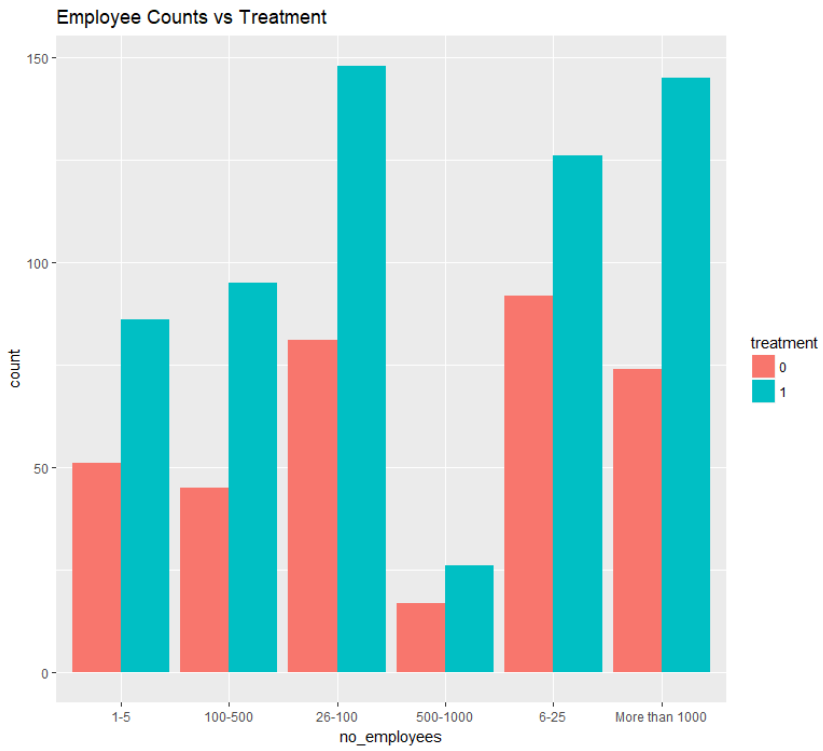




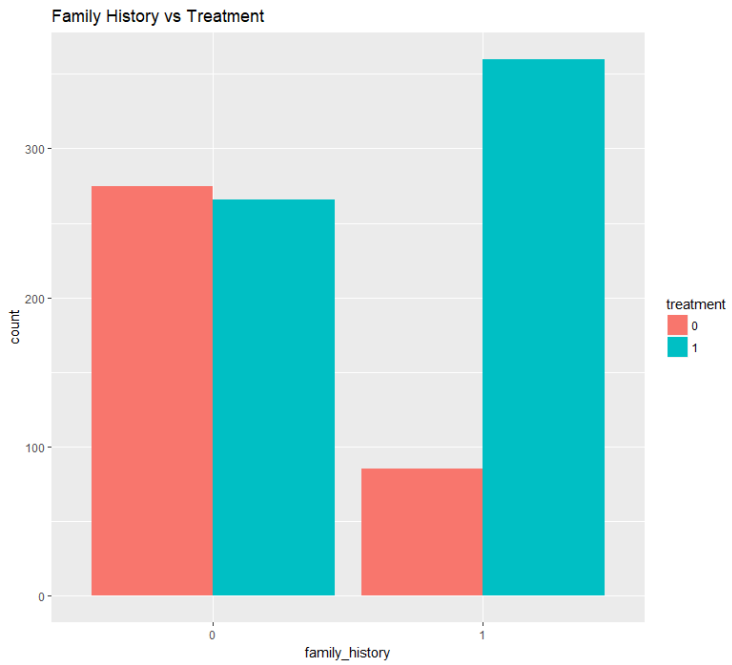


The above three graphs show that the people working in tech companies are the ones with highest number of treatments going on.

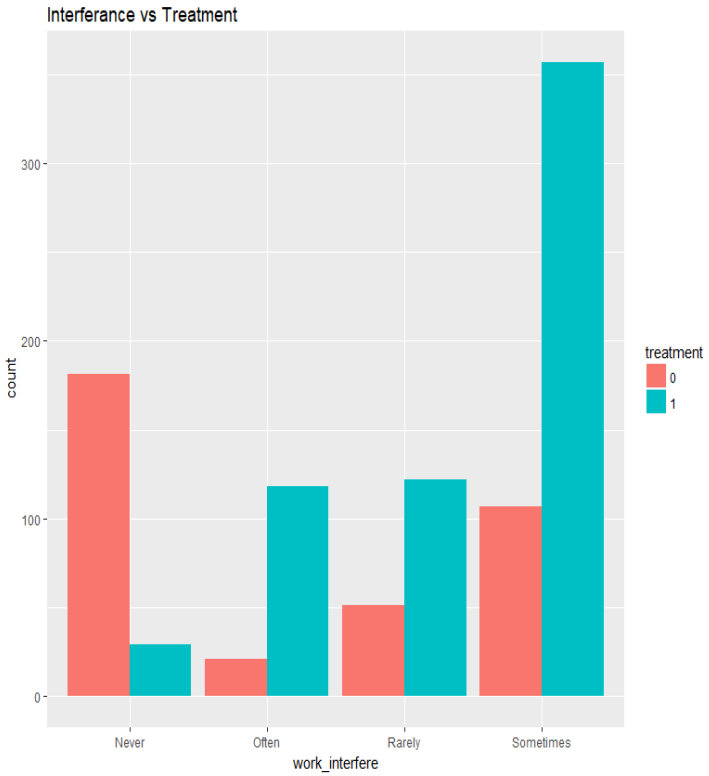




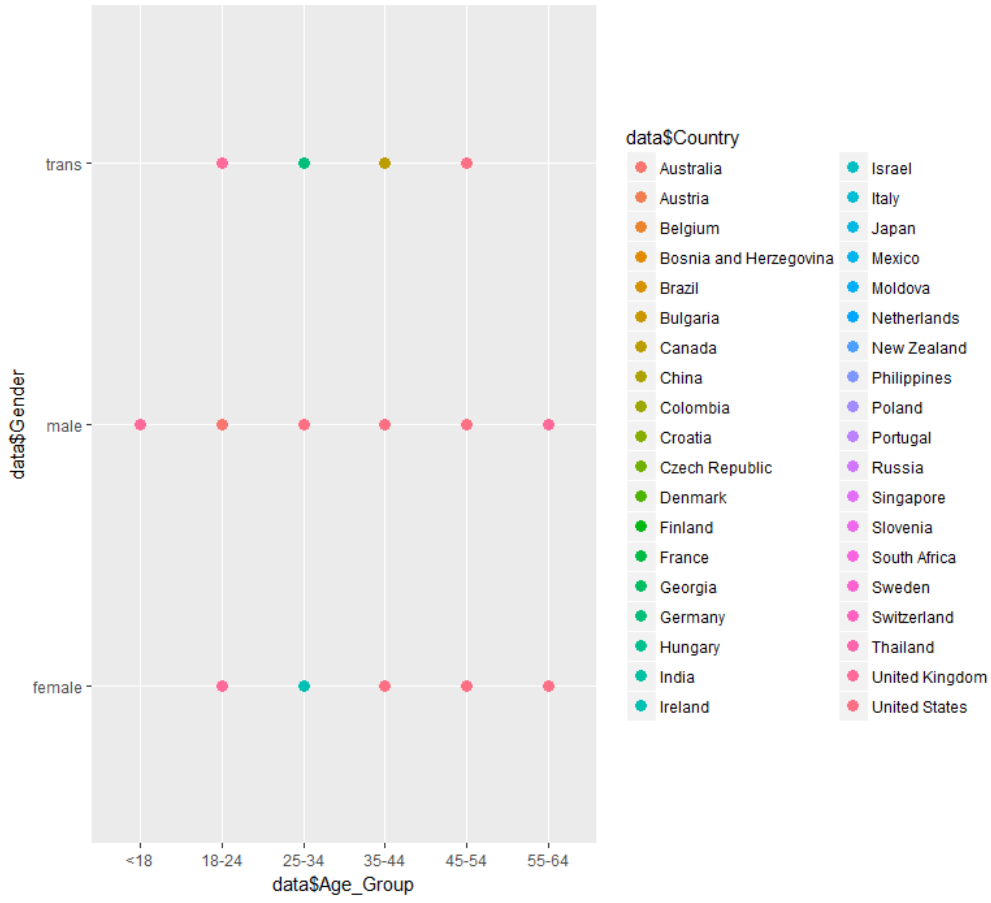
These two graphs basically show the number of employees who have undergone treatment.

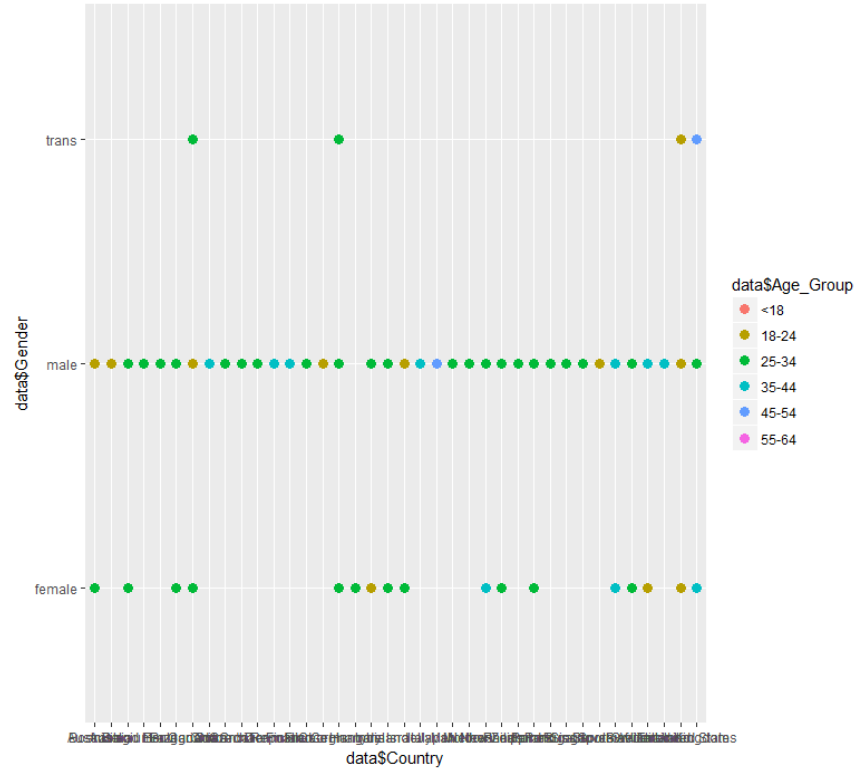


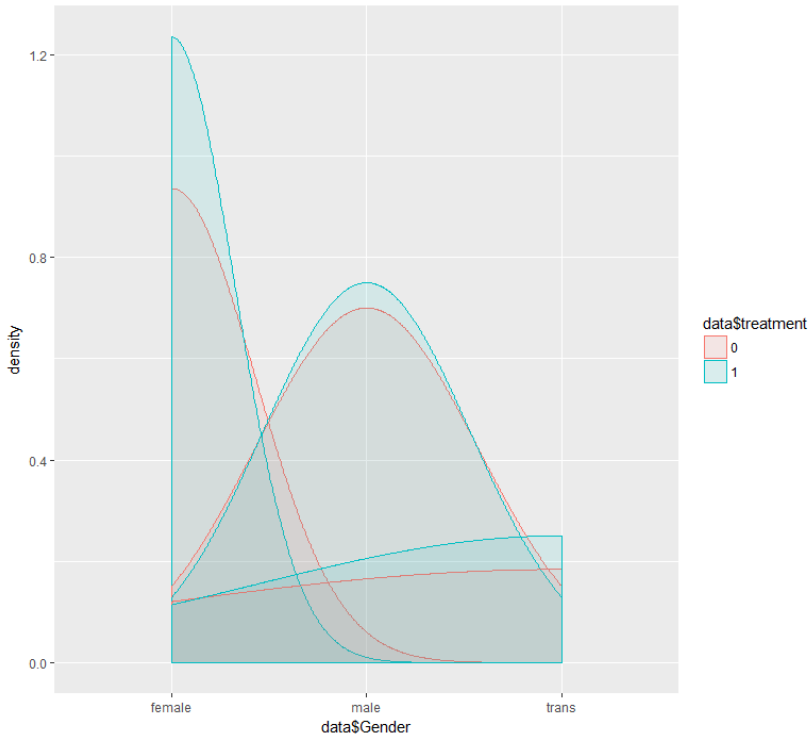
This graph shows that that people with history of mental illness in their families are likely to treat mental illness more than others. In other words they do not neglect it.



The graph shows the interference of the treatment in the employees' work which is there at times.


trans
male
female
data$Gender
<18
18-24
25-34
35-44
45-54
55-64
data$Age_Group
data$Country
Australia
Austria
Belgium
Bosnia and Herzegovina
Brazil
Bulgaria
Canada
China
Colombia
Croatia
Czech Republic
Denmark
Finland
France
Georgia
Germany
Hungary
India
Ireland
Israel
Italy
Japan
Mexico
Moldova
Netherlands
New Zealand
Philippines
Poland
Portugal
Russia
Singapore
Slovenia
South Africa
Sweden
Switzerland
Thailand
United Kingdom
United States

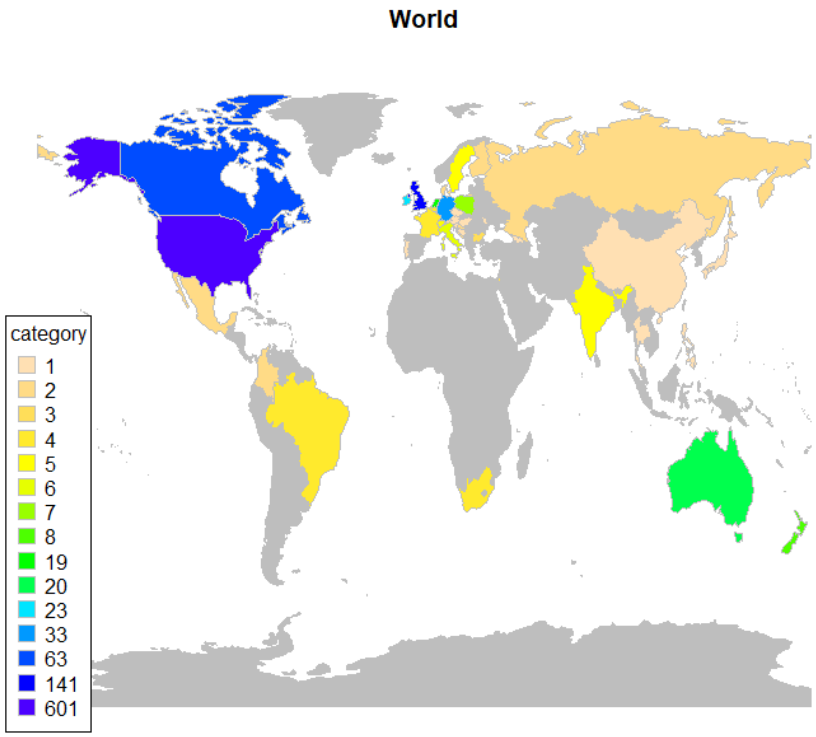

trans
male
female
data$Gender
data$Country
data$Age_Group
<18
18-24
25-34
35-44
45-54
55-64



The above three graphs potentially show the cleaned data with changed categories of gender and thus the gender wise people of different age group from different countries.

```
library(rworldmap)
> country_count <- data %>%
+     group_by(Country) %>%
+     dplyr::summarize(count = n())
> country_count
nrow(country_count)
[1] 38
> country_map <- joinCountryData2Map(country_count, joinCode="NAME", nameJoin
Column="Country")
> mapCountryData(country_map, nameColumnToPlot="count", mapTitle="World" ,cat
Method="categorical",
+                colourPalette = "topo", missingCountryCol="grey", aspect =0)
```

This heat map shows the effective subjects from the world on which the study was done.




```
fh_count <- data %>%
+     group_by(family_history) %>%
+     dplyr::summarize(count = n(), proportion = n()/nrow(data))
> fh_count
# A tibble: 2 x 3
  family_history count proportion
  <chr>          <int>      <dbl>
1 0                541      0.549
2 1                445      0.451
> treat_count <- data %>%
+     group_by(treatment) %>%
+     dplyr::summarize(count = n(), proportion = n()/nrow(data))
> treat_count
# A tibble: 2 x 3
  treatment count proportion
  <chr>     <int>      <dbl>
1 0           360      0.365
2 1           626      0.635
```

This shows that approximately 45% of the population has family history of mental illness and 63% of the population is seeking treatment for mental illness.

c) I had to do a couple of things to cleanse my data. The gender category in my data had duplicate data in different forms as visible from my first plot of Answer 2 – a. I changed the categories to Male, Female and Trans. To plot a boxplot for Age I excluded the outliners. For some categories I changed the "Yes" –"No" to 1 and 0 to make my life easier.

```
library(dplyr)
library("magrittr")
data$Gender %<>% tolower
> male_str <- c("male", "m", "male-ish", "maile", "mal", "male (cis)",
"make", "male ", "man","msle", "mail", "malr","cis man", "cis male")
> trans_str <- c("trans-female", "something kinda male?",
"queer/she/they", "non-binary","nah", "all", "enby", "fluid",
"genderqueer", "androgyne", "agender", "male leaning androgynous", "guy (-
ish) ^_^", "trans woman", "neuter", "female (trans)", "queer", "ostensibly
male, unsure what that really means" )
> female_str <- c("cis female", "f", "female", "woman",  "femake", "female
","cis-female/femme", "female (cis)", "femail")
> data$Gender <- sapply(as.vector(data$Gender), function(x) if(x %in%
male_str) "male" else x )
> data$Gender <- sapply(as.vector(data$Gender), function(x) if(x %in% fema
le_str) "female" else x )
> data$Gender <- sapply(as.vector(data$Gender), function(x) if(x %in% tran
s_str) "trans" else x )
> data %<>% filter(Gender != "a little about you")
data %<>% filter(Gender != "guy (-ish) ^_^")
> data %<>% filter(Gender != "p")
```