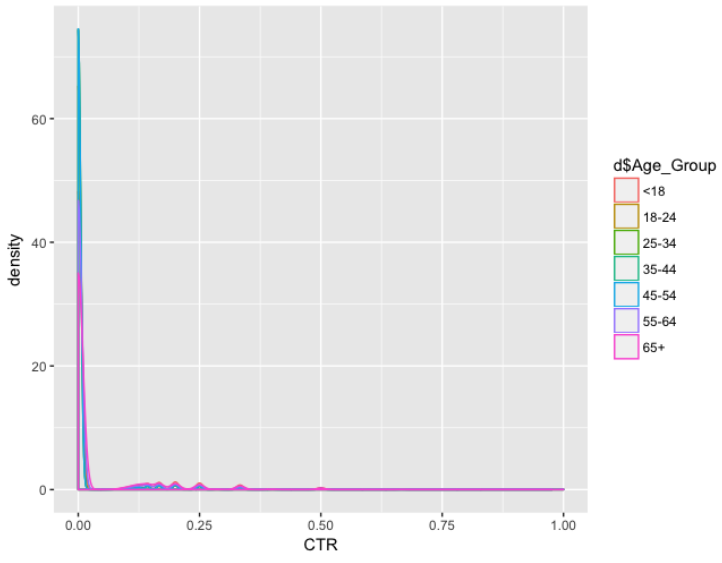
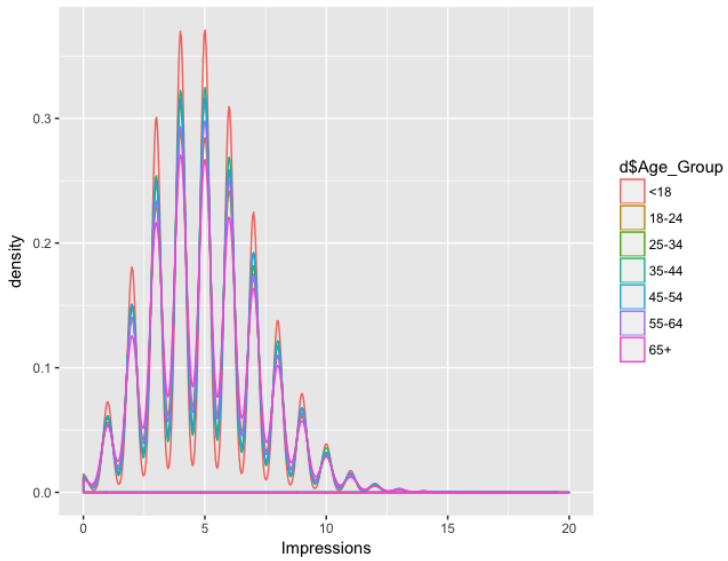
Answer 1

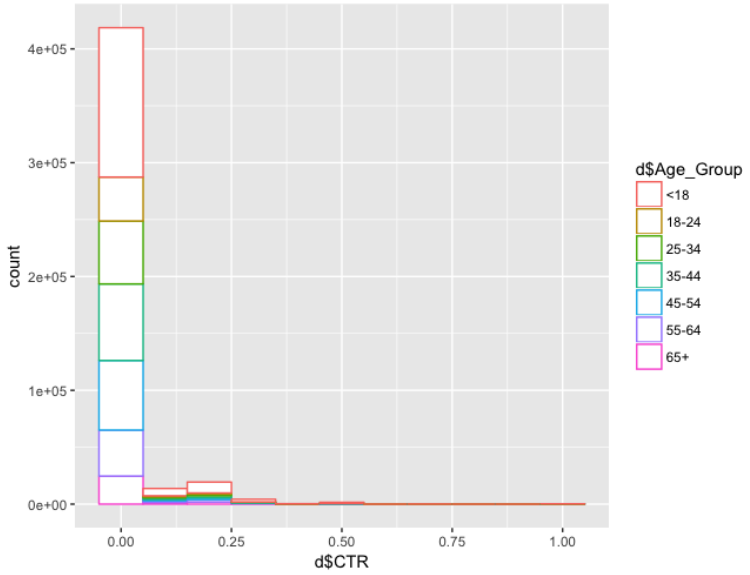
a)
```
> setwd("/Users/saurabhthakrani/Desktop/737/all_nyt")
> data=lapply(dir(),read.csv)
> totalData=do.call("rbind",data)
> attach(totalData)
> totalData$Age_Group= cut(Age, breaks = c(0,18,25,35,45,55,65,116), labels =
c('<18','18-24','25-34','35-44','45-54','55-64','65+'), right = FALSE)
> View(totalData)
```

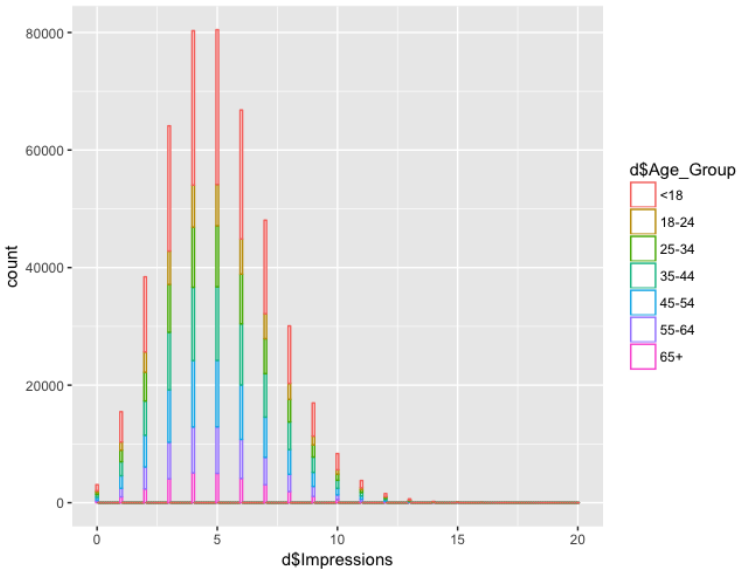
	Age	Gender	Impressions	Clicks	Signed_In	Age_Group
1	36	0	3	0	1	35-44
2	73	1	3	0	1	65+
3	30	0	3	0	1	25-34
4	49	1	3	0	1	45-54
5	47	1	11	0	1	45-54
6	47	0	11	1	1	45-54
7	0	0	7	1	0	<18
8	46	0	5	0	1	45-54
9	16	0	3	0	1	<18
10	52	0	4	0	1	45-54
11	0	0	8	1	0	<18

b) (i) The distributions of number impressions and CTR for 6 age categories is as follows. The .R file containing all the codes is attached in the mail.







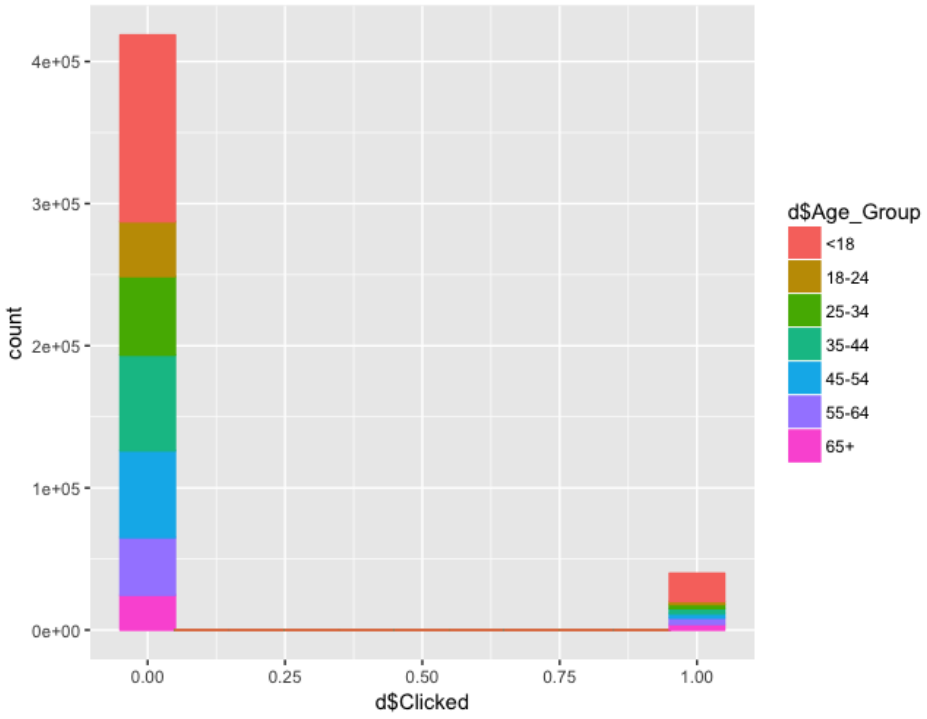


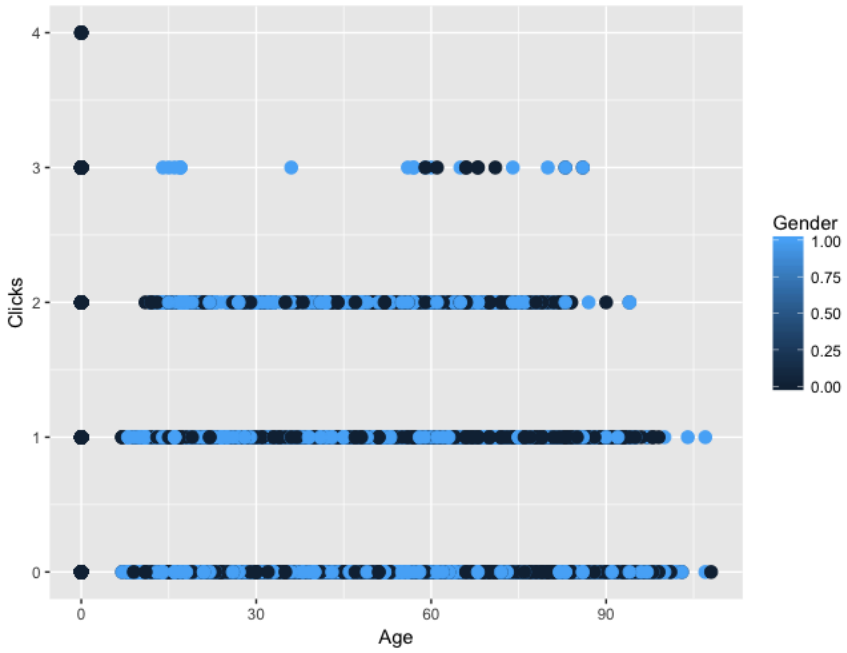
(ii) Categorization based on click behavior(either clicked or not).

```
> summary(Clicks)
> d$Clicked=ifelse(Clicks==0,0,1)
```

	Age	Gender	Impressions	Clicks	Signed_In	Age_Group	CTR	Clicked
1	36	0	3	0	1	35-44	0.00000000	0
2	73	1	3	0	1	65+	0.00000000	0
3	30	0	3	0	1	25-34	0.00000000	0
4	49	1	3	0	1	45-54	0.00000000	0
5	47	1	11	0	1	45-54	0.00000000	0
6	47	0	11	1	1	45-54	0.09090909	1
7	0	0	7	1	0	<18	0.14285714	1
8	46	0	5	0	1	45-54	0.00000000	0
9	16	0	3	0	1	<18	0.00000000	0
10	52	0	4	0	1	45-54	0.00000000	0
11	0	0	8	1	0	<18	0.12500000	1
12	21	0	3	0	1	18-24	0.00000000	0
13	0	0	4	0	0	<18	0.00000000	0
14	57	0	6	0	1	55-64	0.00000000	0

(iii) More visualizations on the data.





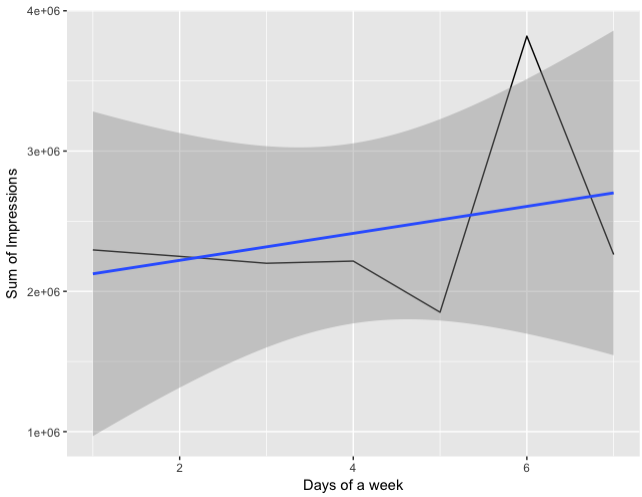
c) Measurements/Statistics that summarize the data.

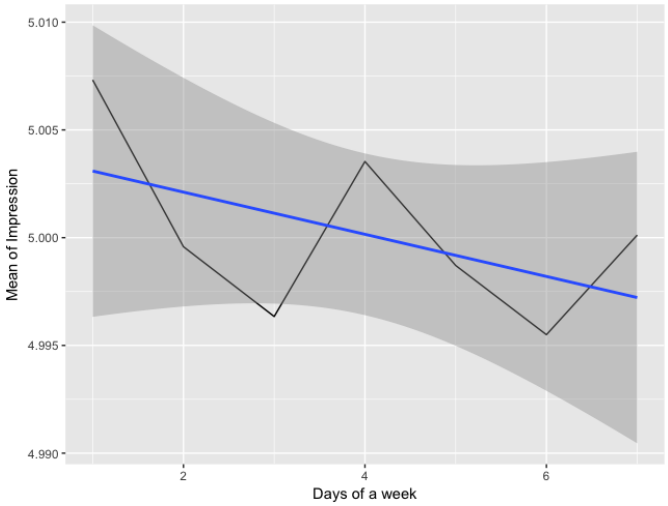
```
> summary(d)
      Age              Gender       Impressions         Clicks          Signed_In
 Min.   :  0.00   Min.   :0.000   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:  0.00   1st Qu.:0.000   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000
 Median : 31.00   Median :0.000   Median : 5.000   Median :0.00000   Median :1.0000
 Mean   : 29.48   Mean   :0.367   Mean   : 5.007   Mean   :0.09259   Mean   :0.7009
 3rd Qu.: 48.00   3rd Qu.:1.000   3rd Qu.: 6.000   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :108.00   Max.   :1.000   Max.   :20.000   Max.   :4.00000   Max.   :1.0000

 Age_Group           CTR              Clicked
 <18  :150934   Min.   :0.00000   Min.   :0.0000
 18-24: 40694   1st Qu.:0.00000   1st Qu.:0.0000
 25-34: 58174   Median :0.00000   Median :0.0000
 35-44: 70860   Mean   :0.01835   Mean   :0.0869
 45-54: 64288   3rd Qu.:0.00000   3rd Qu.:0.0000
 55-64: 44738   Max.   :1.00000   Max.   :1.0000
 65+  : 28753
```

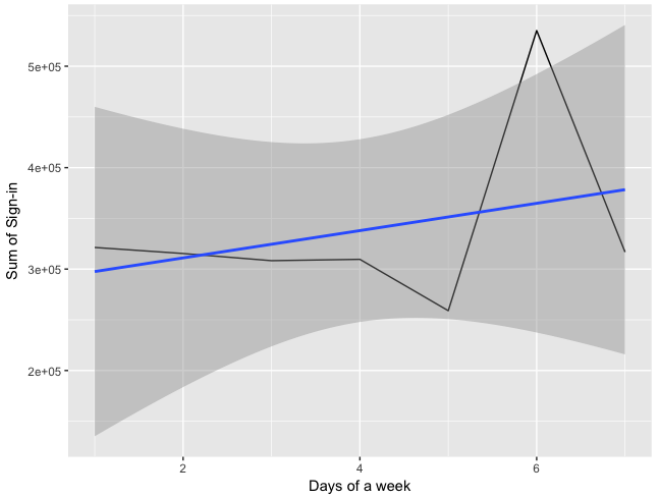
Here, we can see Minimum and maximum values for each variable. Also, we can see values for first and third quantiles. Mean and median can also be seen.

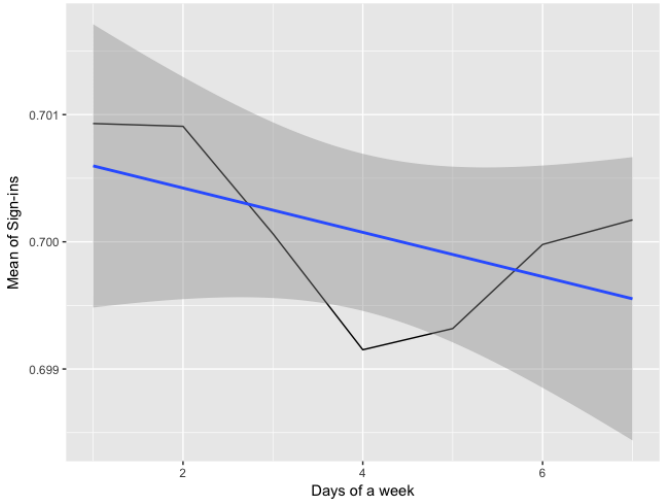
I think tracking the sum and mean for the 3 variables namely Impressions, Clicks and Signed-In will be of use and hence I am analyzing these fields for over a week.

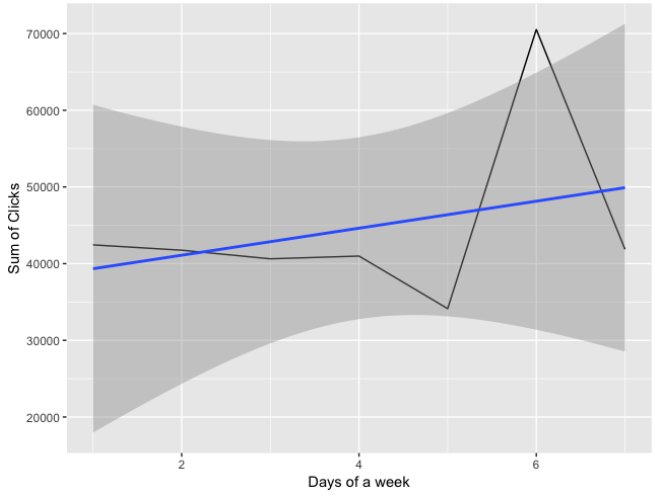
```
> setwd("/Users/saurabhthakrani/Desktop/737/allnyt7")
> data7=lapply(dir(),read.csv)
> totaldata7=do.call("rbind",data7)
> attach(totaldata7)
> impression_day <- aggregate(Impressions, by=list(Day),sum)
> View(impression_day)
> ggplot(impression_day, aes(Group.1, x)) +   geom_line() +xlab("Days of a week") + ylab("Sum of
Impressions") +geom_smooth(method = lm)
> clicks_day <- aggregate(Clicks, by=list(Day),sum)
> ggplot(clicks_day, aes(Group.1, x)) +   geom_line() +xlab("Days of a week") + ylab("Sum of Clicks")
+geom_smooth(method = lm)
> View(clicks_day)
> signin_day <- aggregate(Signed_In, by=list(Day),sum)
> ggplot(signin_day, aes(Group.1, x)) +   geom_line() +xlab("Days of a week") + ylab("Sum of Sign-in")
+geom_smooth(method = lm)
> View(signin_day)
> impression_mean_day <- aggregate(Impressions, by=list(Day),mean)
> ggplot(impression_mean_day, aes(Group.1, x)) +    geom_line() +xlab("Days of a week") + ylab("Mean of
Impression") +geom_smooth(method = lm)
> Clicks_mean_day <- aggregate(Clicks, by=list(Day),mean)
> ggplot(Clicks_mean_day, aes(Group.1, x)) +    geom_line() +xlab("Days of a week") + ylab("Mean of
Clicks") +geom_smooth(method = lm)
> Signin_mean_day <- aggregate(Signed_In, by=list(Day),mean)
> ggplot(Signin_mean_day, aes(Group.1, x)) +   geom_line() +xlab("Days of a week") + ylab("Mean of Sign-
ins") +geom_smooth(method = lm)
```

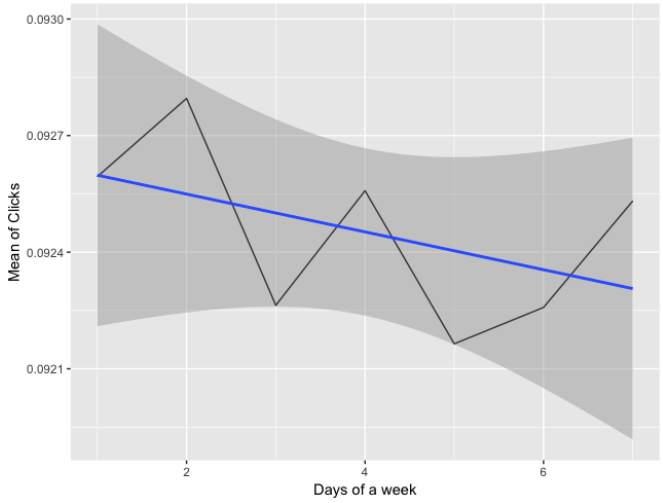












d) From the graphs we can see that the sum of impressions, clicks and sign-ins are highest on the 6th day i.e. Saturday. The reason could be that people tend to browse online sites more on holidays. The highest mean is rather on different days for than the sum.

Answer 2

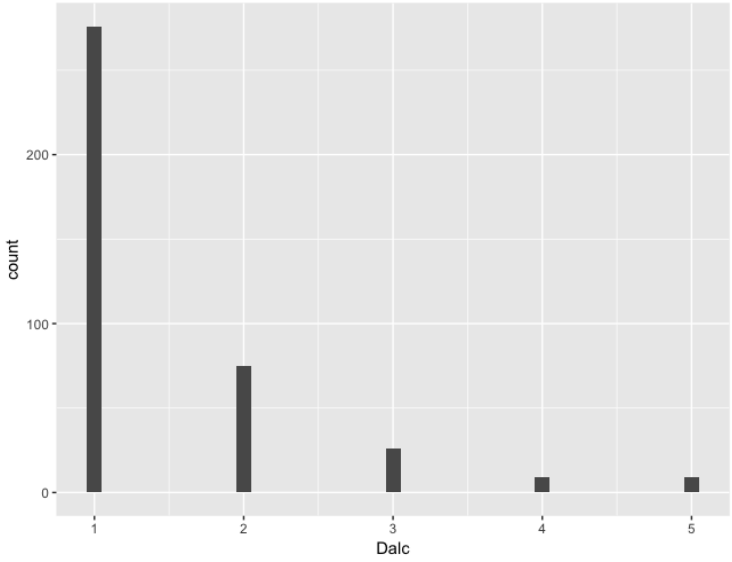
The dataset which I am using represents the data about the students in a math course of two schools who are aged between 15 years to 22 years. The data consists of 34 variables which depicts the personality of the student, family background, alcohol consumption levels, grades obtained in class, time spent on different activities and their health. There are some variables which have values from 1 to 5 where 1 means “very low” and 5 means “very high”. The grades are given on a scale of 0 to 20.

Here, are the details of the variables which I have used to perform visualizations on this data.

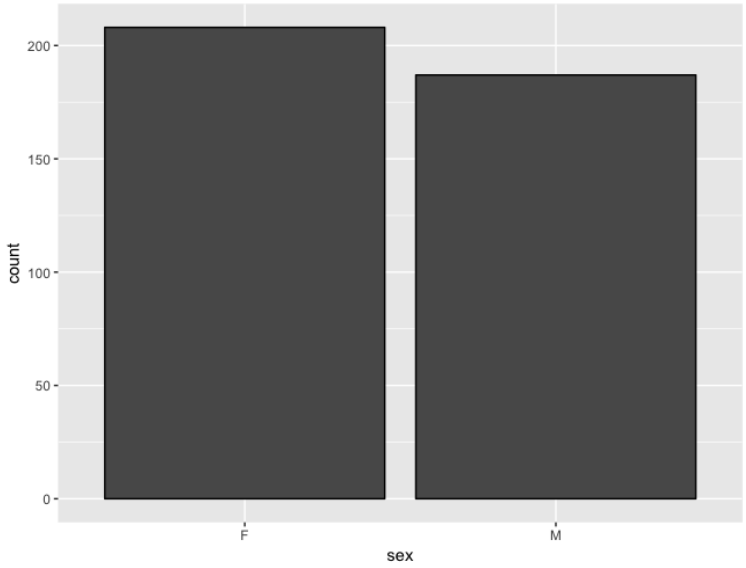
G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)
sex - student's sex (binary: 'F' - female or 'M' - male)
age - student's age (numeric: from 15 to 22)
internet - Internet access at home (binary: yes or no)
romantic - with a romantic relationship (binary: yes or no)
Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)

To perform the visualizations, I had to do some changes in data. In the data, three grades are given, and I calculated the mean of these three grades and added it in a new column.
Also, I have calculated the mean of average grade depending on two variables -“Romantic” and “Internet”.

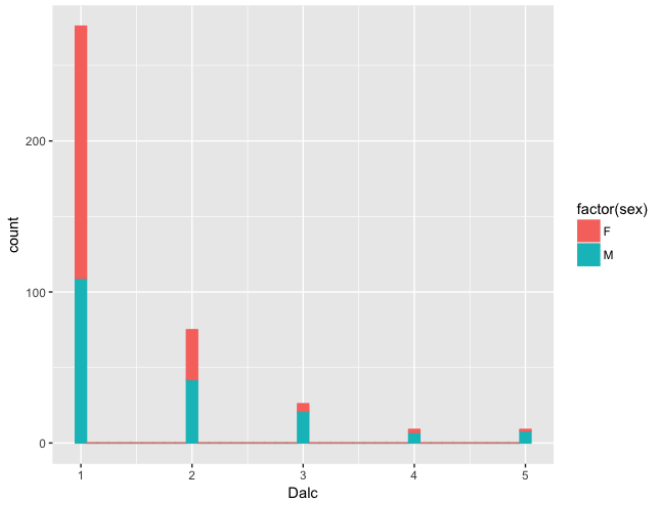
Below are the visualizations and their interpretation. The R code for these visualizations will be attached as an R script file in the mail.

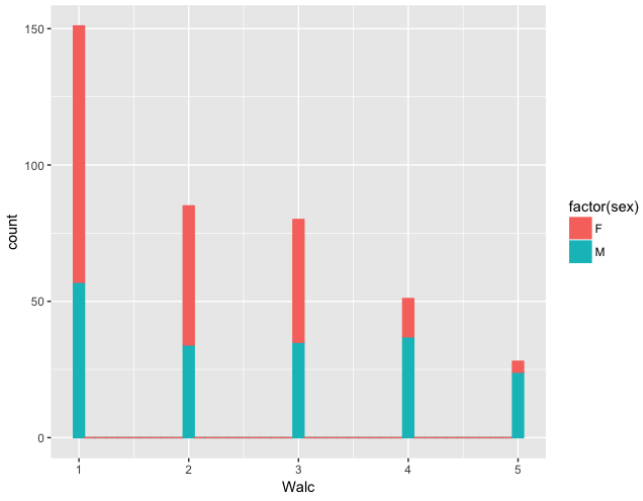


This graph shows that there are more no. of students who take less amount of alcohol on a regular basis than those who take high amount of alcohol.



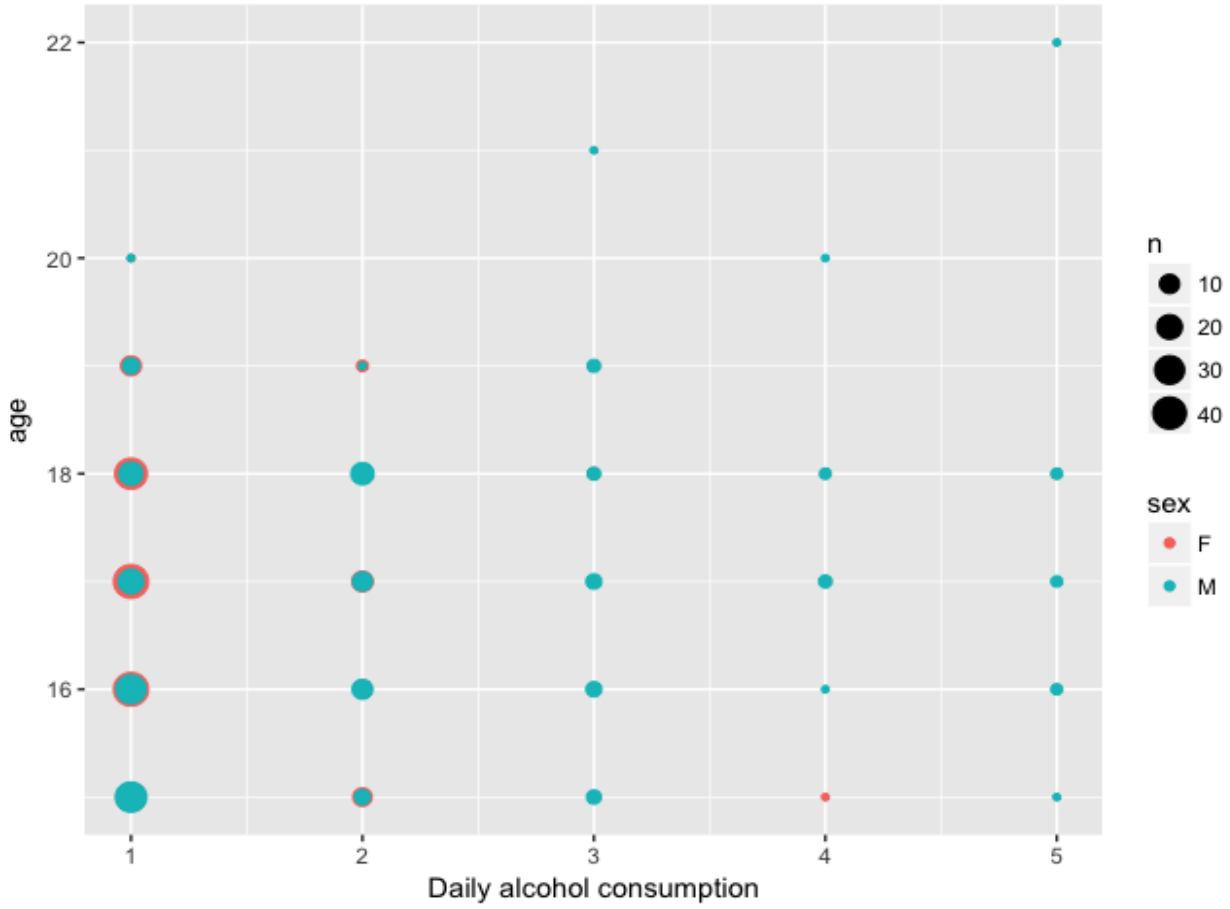
This plot shows the sex ratio of students in the dataset. It can be seen that there are more female students (around 25) than male students.



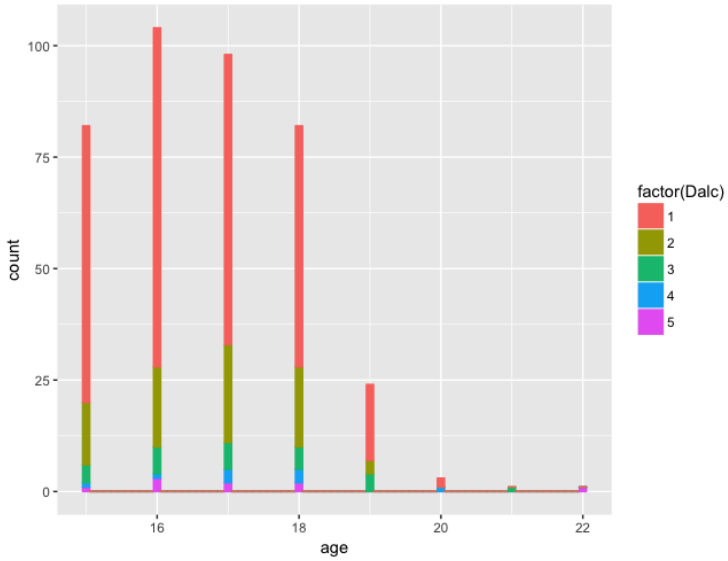


These two graphs show that:

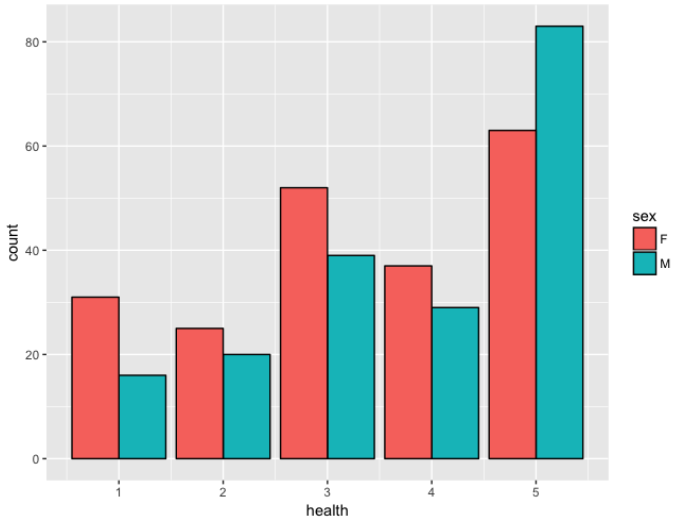
a. On daily basis, more females than males consume alcohol on a low level and very less people (male or female) consume alcohol on a high level.
b. On weekends, more females consume alcohol on low to medium level but more males consume alcohol on a high level.

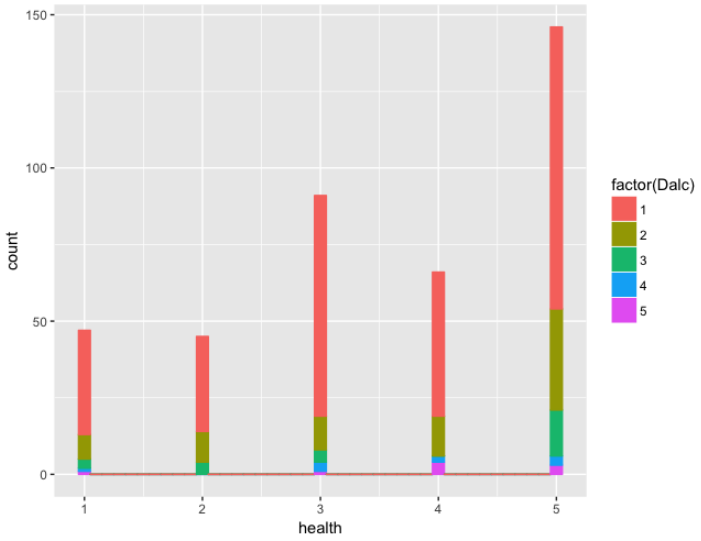


This graph shows that as compared to males, more females of age 16, 17, 18 and 19 consume alcohol on a low level. Apart from that, males of age 15,16,17 and 18 consume alcohol on a higher level than females.



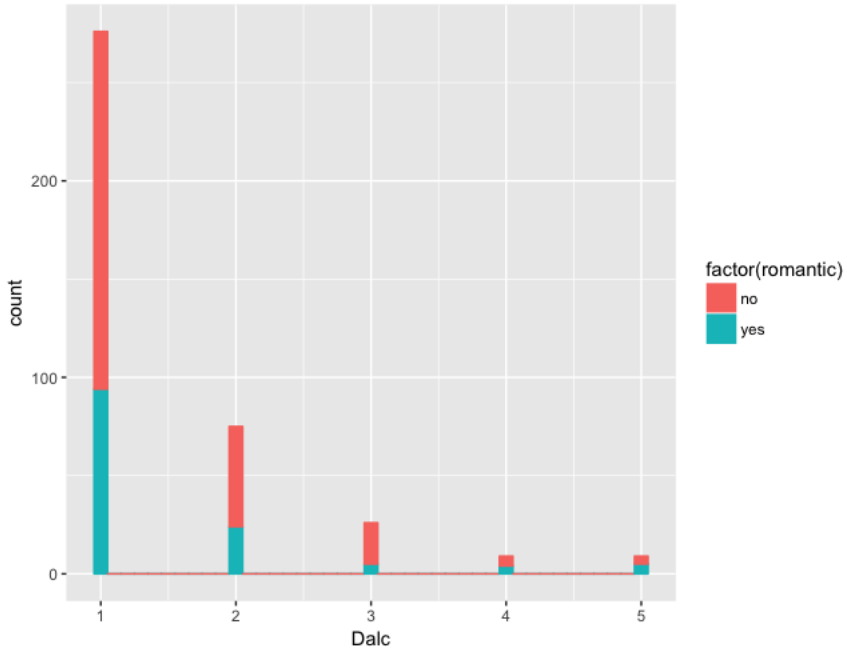
This graph shows that, including males and females, for almost every age there are more no. of people who consume alcohol on a lower level on daily basis than those who consume on a high level.



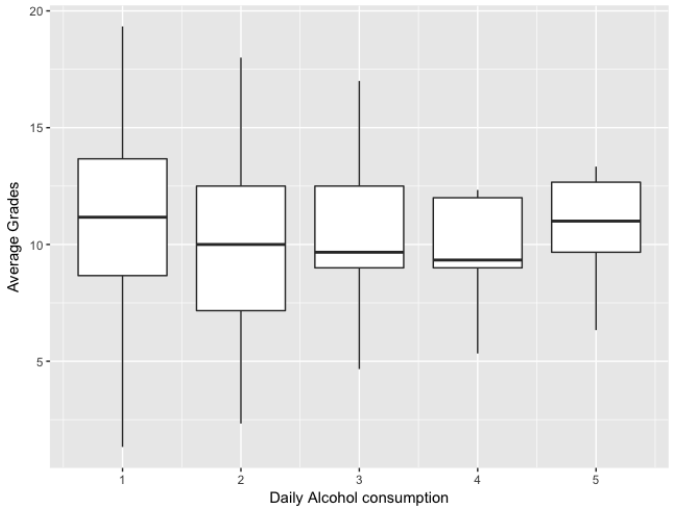


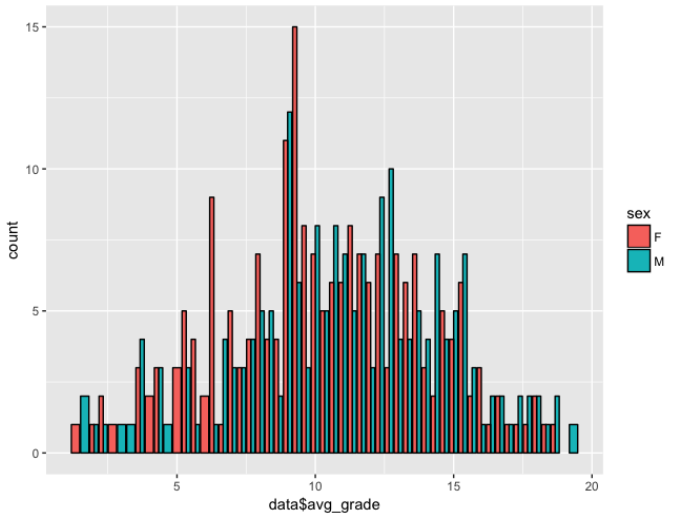
The above graphs show that-

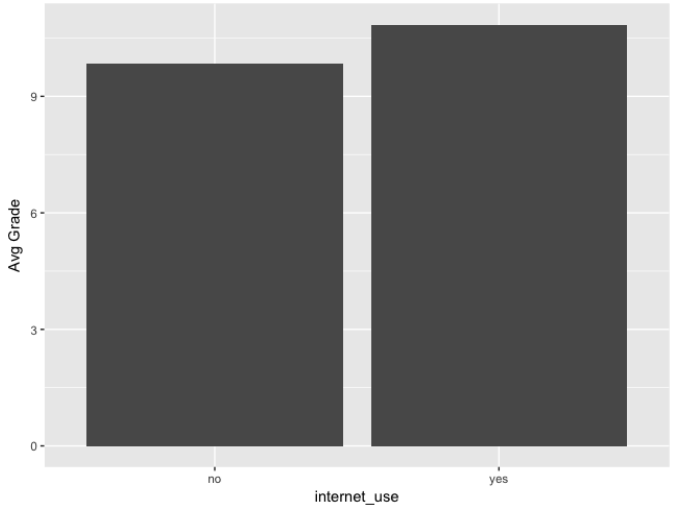
a. There are more females who find themselves to have poor health and more males who find themselves to be very healthy.
b. Also, when we see the alcohol consumption on the basis of health, we can see that people who consume alcohol on a high level also find themselves to be very healthy. People who feel they are not healthy consume alcohol on a low to moderate level.

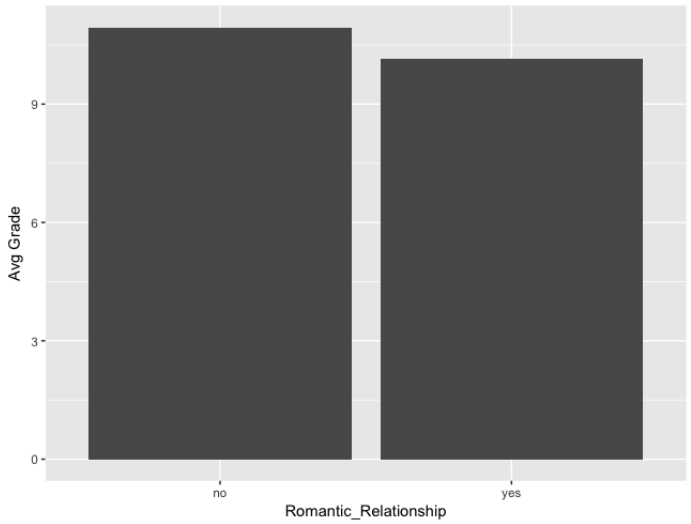


This plot shows that on every level, the consumers of alcohol consists of less people who are in a romantic relationship.









The above four graphs show the effect on average grade of a student depending on internet usage, romantic relationship, daily alcohol consumption and gender.

a. The median of average grade is highest for people who consume alcohol on a low level. Yet there is not much difference in medians for other levels.
b. The second graph shows the distribution of grades with respect to gender.
c. From the third graph, we can see that the average grade for people who use internet is more than those who do not.
d. It can be seen that the average grade for people who are in a romantic relationship is less than those who are not.

```
> t.test(data$avg_grade~data$romantic, data=data)

        Welch Two Sample t-test

data:  data$avg_grade by data$romantic
t = 2.0439, df = 261.25, p-value = 0.04197
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02943714 1.57875504
sample estimates:
 mean in group no mean in group yes
         10.94804          10.14394

> t.test(data$avg_grade~data$internet, data=data)

        Welch Two Sample t-test

data:  data$avg_grade by data$internet
t = -2.1027, df = 95.524, p-value = 0.03812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.97458708 -0.05682123
sample estimates:
 mean in group no mean in group yes
         9.833333         10.849037
```

On performing the t-test on the variables-internet and romantic for average grades, we find that the p-value is less than alpha in both the cases and therefore, these variables might be used for predictions.