Lab2Q4

Loading required libraries

```
library(dplyr)
library(ggplot2)
```

Reading data

```
data = read.csv("June 10-July 12, 2015 - Gaming, Jobs and Broadband - CSV.csv")
```

Removing the variables having more than 65% missing values

```
dt = data[, colMeans(is.na(data)) <= .65]
```

Missing value imputation

We will replace the remaining missing values using Mode imputation

```
Mode <- function(x, na.rm = FALSE) {
    if(na.rm){
        x = x[!is.na(x)]
    }

    ux <- unique(x)
    return(ux[which.max(tabulate(match(x, ux)))])
}

vec = c()
for(n in names(dt))
{
    s = sum(is.na(dt[,n]))
    if(s >0)
    {
        vec = c(vec,n)
    }
}

for(var in vec)
{
    dt[is.na(dt[,var]),var] <- Mode(dt[,var],na.rm = T)
}
```

Converting selected variables to factor

```
conames = c("int_date","age","zipcode","weight","standwt","ï..psraid")
dt[,!colnames(dt) %in% conames] = lapply(dt[,!colnames(dt) %in% conames], as.factor)
```

How would you rate your community as a place to live?

Overall, how would you rate your community as a place to live? Would you say it is...

1 Excellent

2 Good

3 Only fair, OR

4 Poor?

8 (VOL.) Don't know

9 (VOL.) Refused

Creating data frame with useful variables

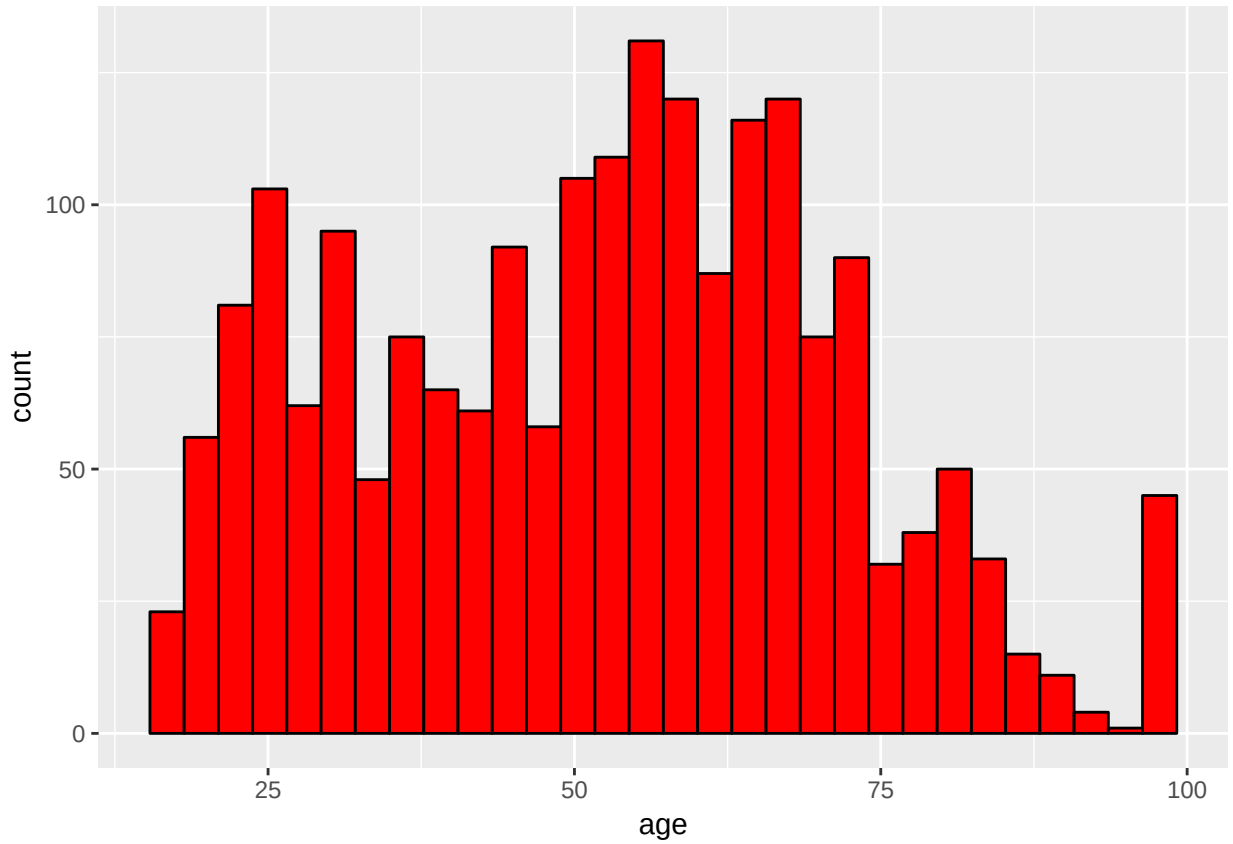
```
q1 = select(dt,sample,lang,usr,cregion,state,form,sex,age,race,q1,marital,par,educ2,inc)
```

We will look at couple of graphs and variables to answer the above question

Plot 1 : Age Distribution

```
ggplot(q1, aes(x = age)) + geom_histogram(color = "black", fill = "red")
```

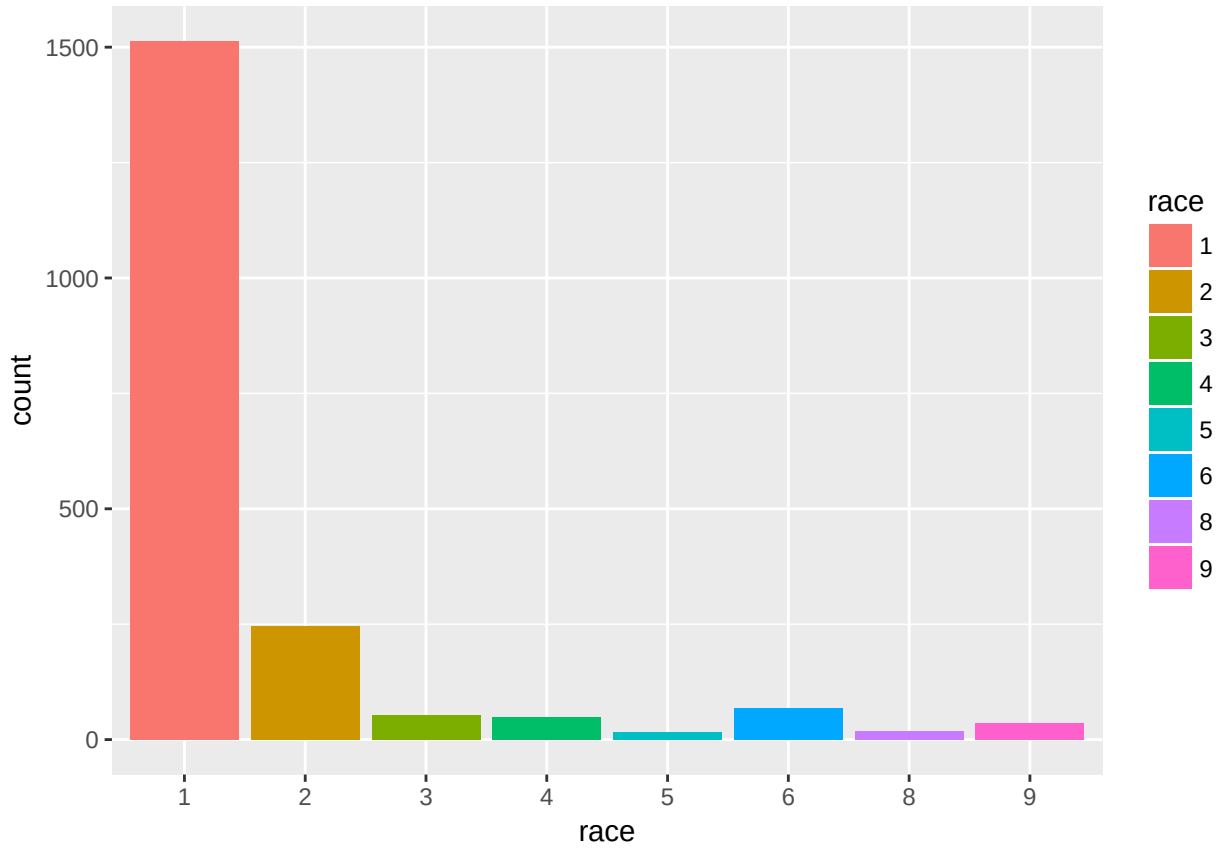
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the above plot, we can observe that most of the people belonging to age group 50-70 participated in survey.

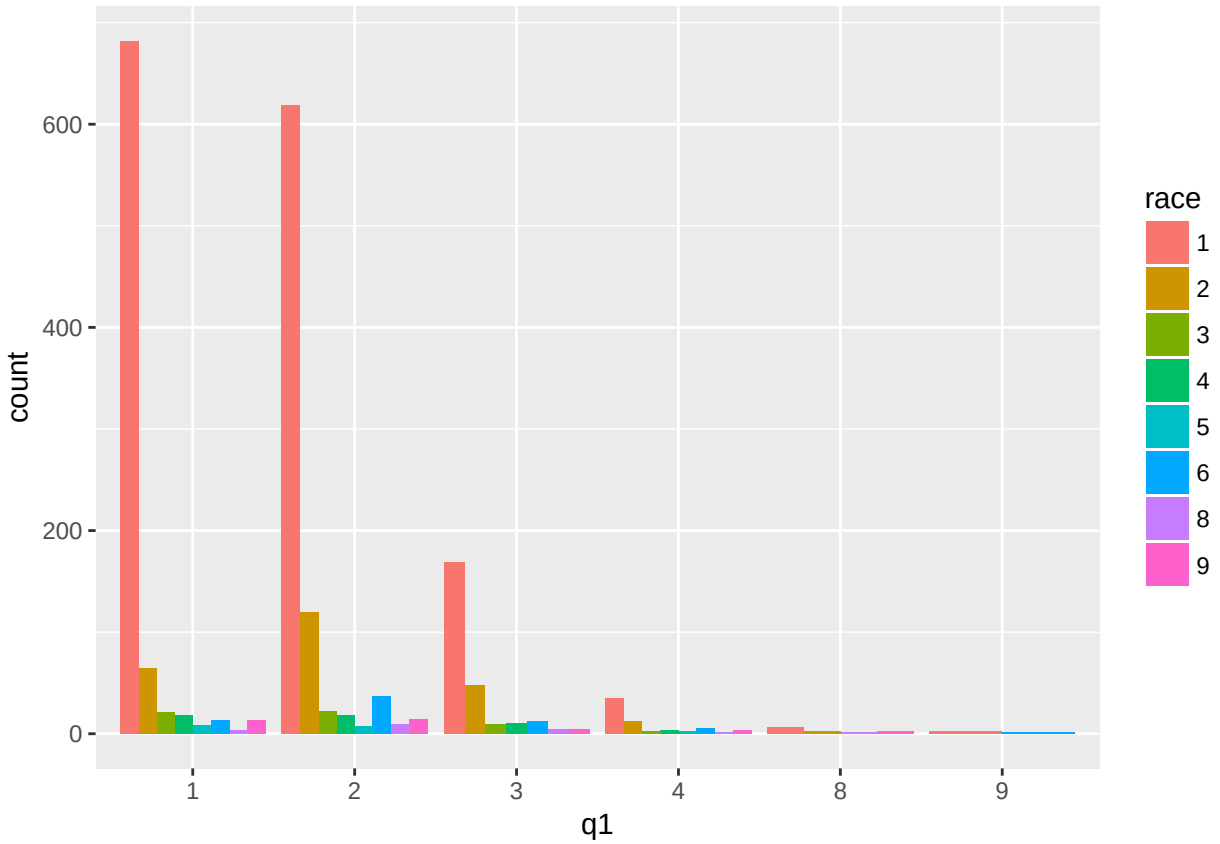
Plot 2 : q1 vs Race

```
ggplot(q1,aes(race,fill = race)) +geom_bar()
```



From the frequency curve, we can see that most surveys are from people belonging to race group 1, i.e, White

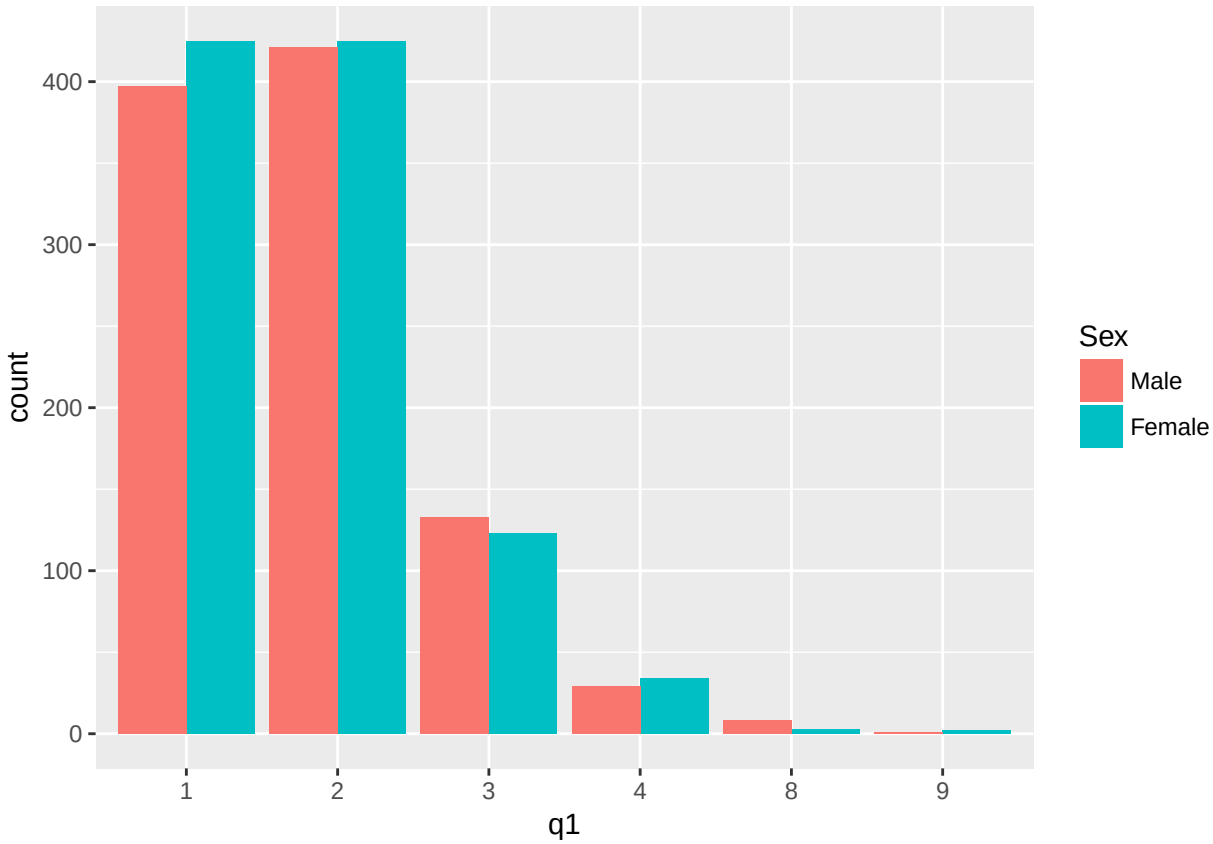
```
ggplot(q1,aes(q1,fill = race)) + geom_bar(position = "dodge")
```



From the above curve we can conclude that ppl belonging to "White"" race have voted the most for Excellent/Good quality of life,while the minority race such as Native American/American Indian have voted the least for Excellent/Good quality of life.

Plot 3 : q1 vs Sex

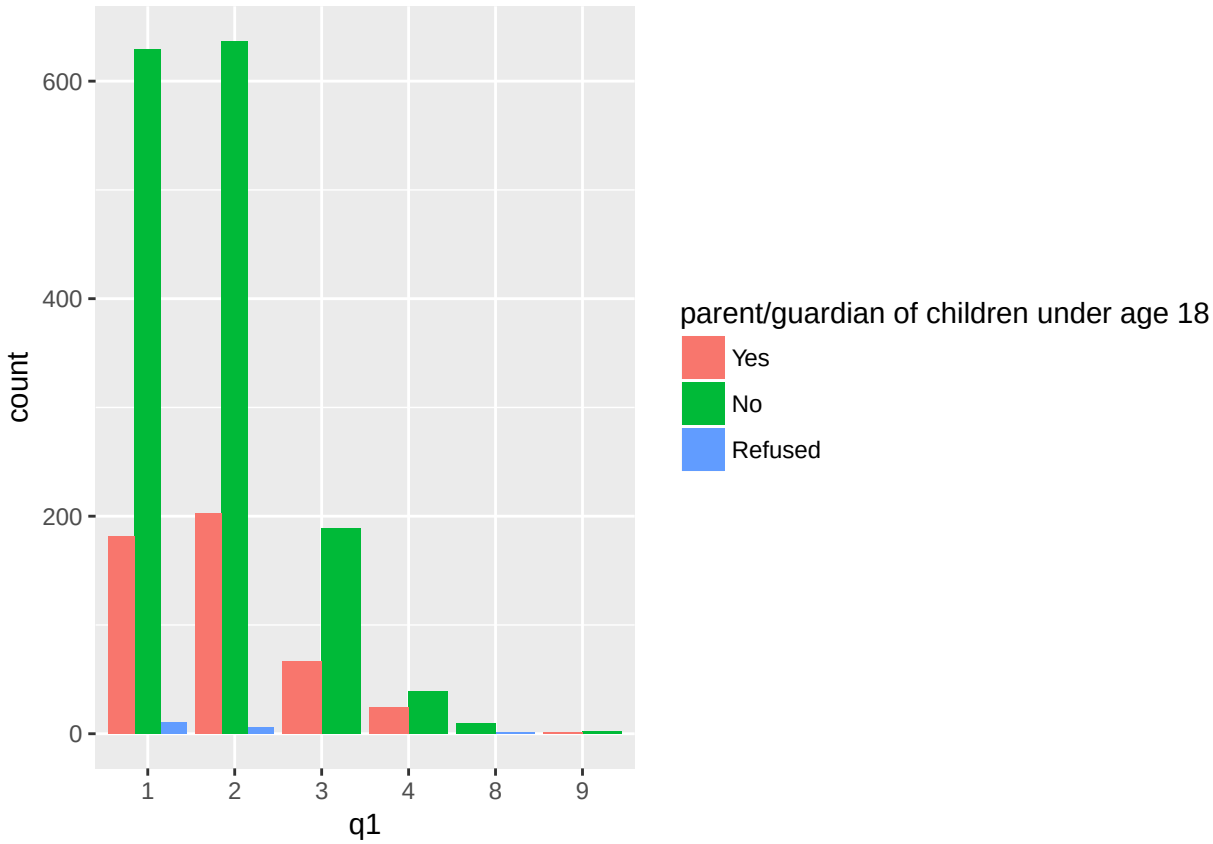
```
ggplot(q1,aes(q1,fill = sex)) + geom_bar(position = "dodge") + scale_fill_discrete(name="Sex",labels=c(
```



From the above plot,we can observe that more number of females have voted for "Excellent" quality while almost equal number of males and females have voted for "Good" quality.

Plot 4 : q1 vs PAR

```
ggplot(q1, aes(x = q1, fill = par)) + geom_bar(position = "dodge") + scale_fill_discrete(name = "parent,
```



Survey states that people or guardian who doesn't have children under age 18 have rated more to "Excellent/Good" community to live in.

Q2 Part(a) : Do you ever play video games on a computer, TV, game console, or portable device like a cell phone?

1 Yes

2 No

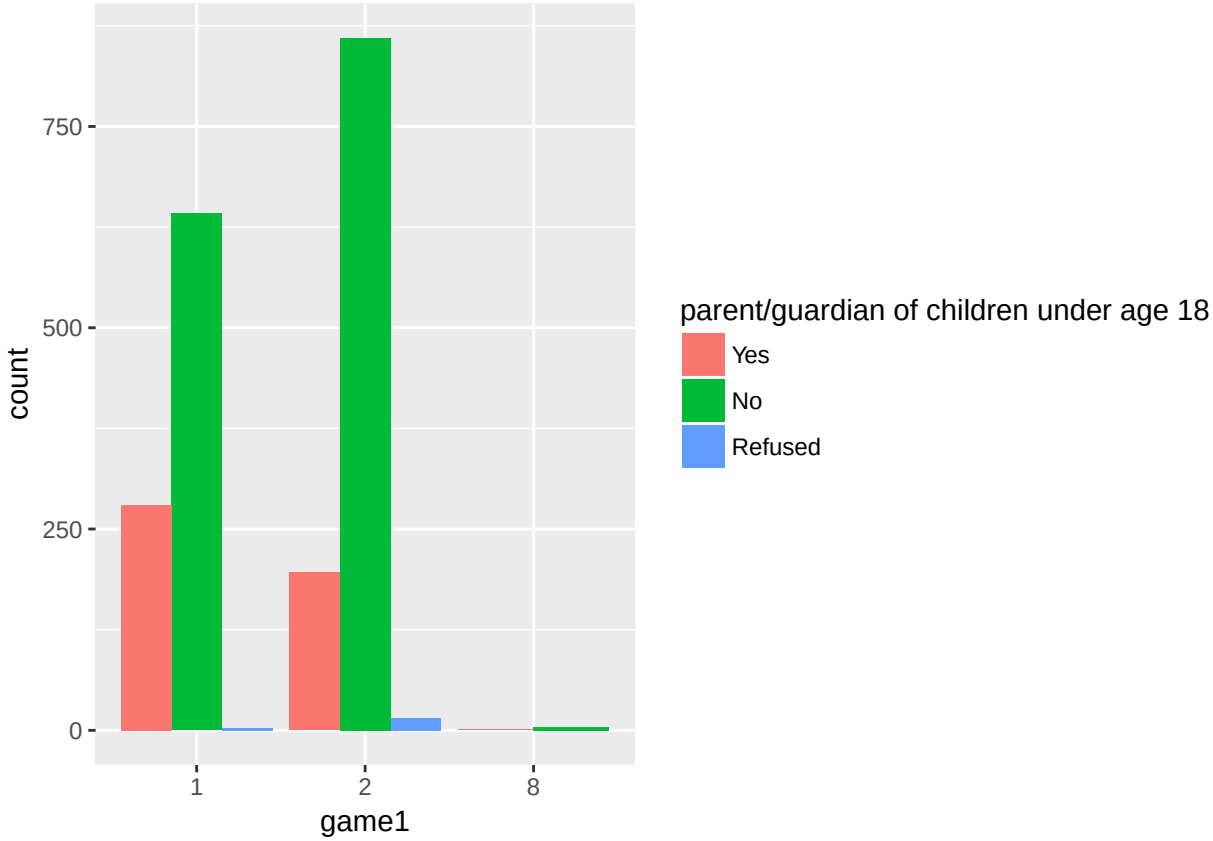
8 (VOL.) Don't know

9 (VOL.) Refused

```
Q2 <- select(dt,sample:q1,game1:game4,age,race,marital,par,educ2,inc)
Q2 <- select(Q2, -(sample:form))
Q2 <- select(Q2, -q1,-marital)
levels(Q2$game1)[4] = "8"
```

Plot 1: game1 vs Parent

```
ggplot(Q2, aes(x = game1, fill = par)) + geom_bar(position = "dodge") + scale_fill_discrete(name = "par
```

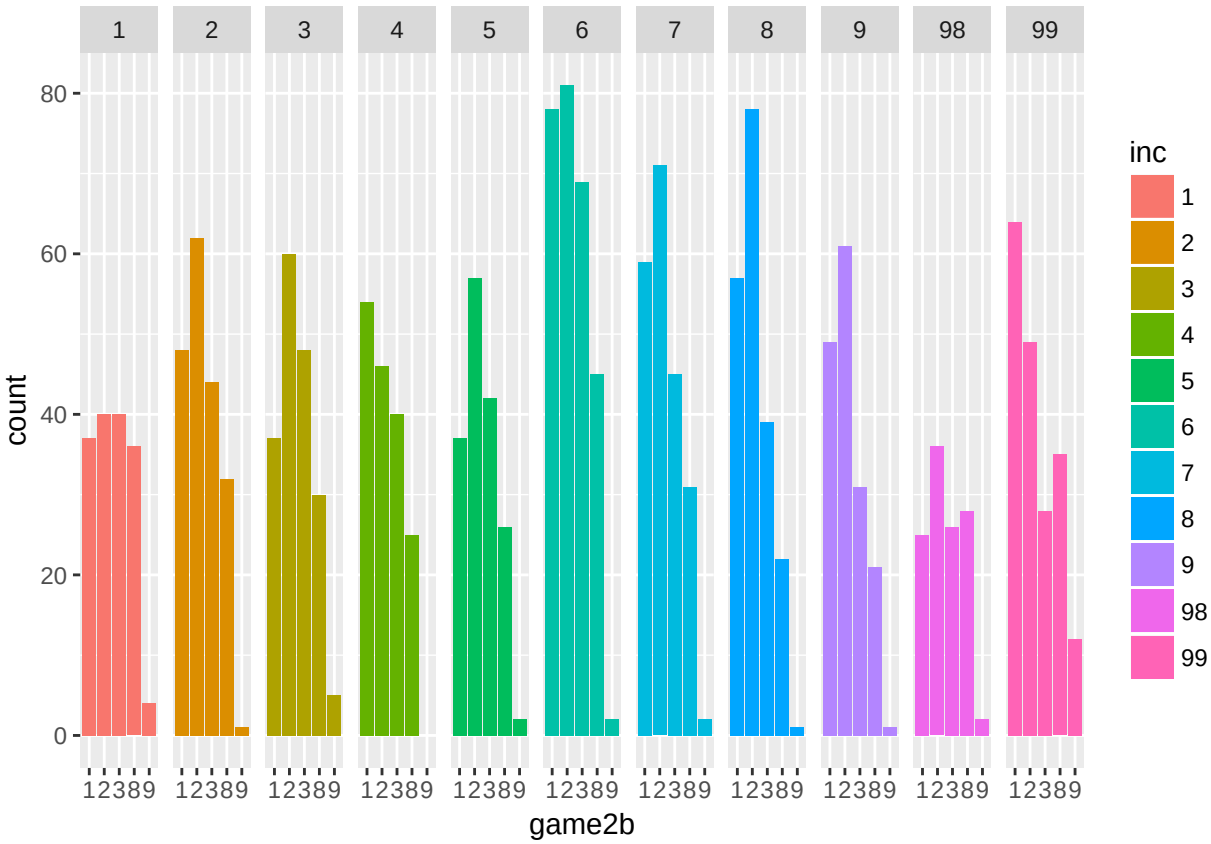


Survey states that people or guardian who doesn't have children under age 18 does not play video games.

Q2 Part(b) : Video games are a waste of time

Plot 2: game2b vs income

```
ggplot(Q2,aes(x = game2b, fill = inc)) + geom_bar() + facet_grid(~inc)
```

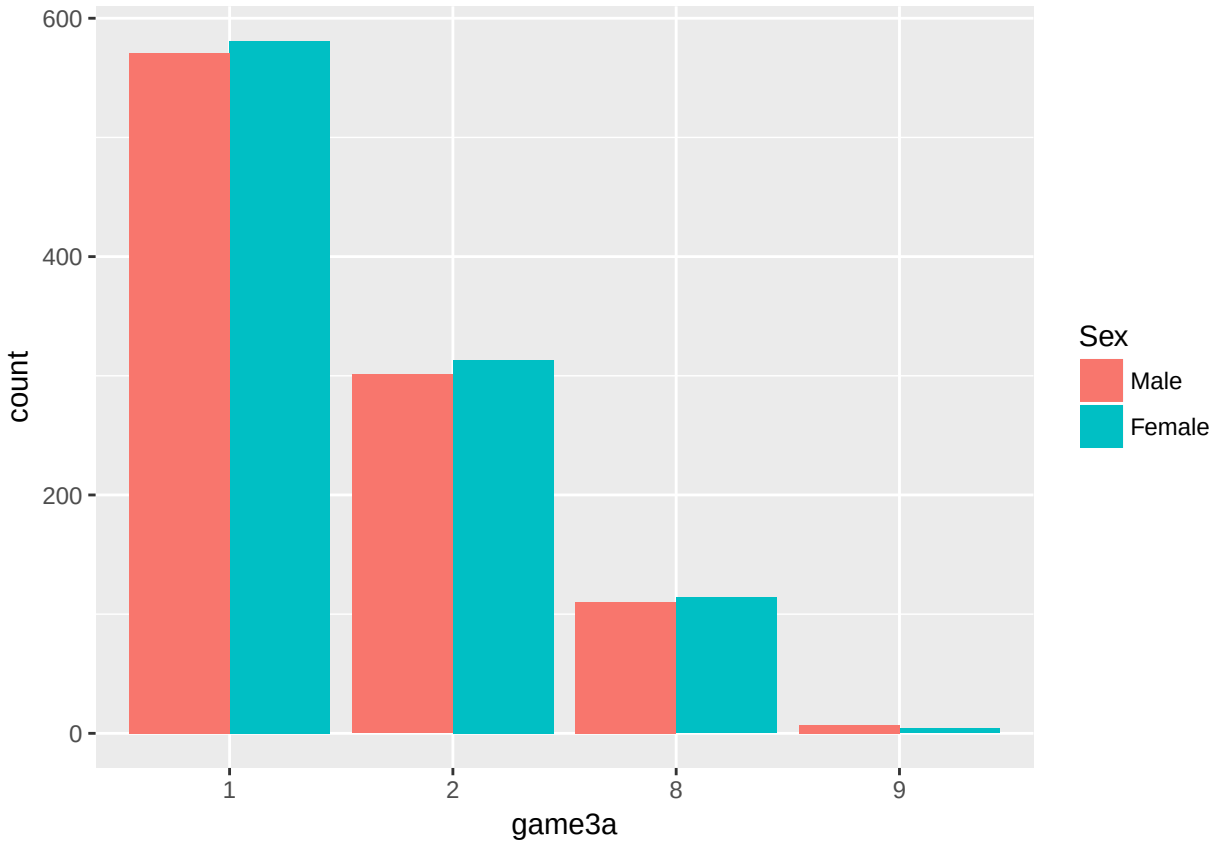


Most of the people believe that video games are waste of time is true for some of the games but not others,whereas people earning(30k-40k) believes that it is true for all games.

Q2 Part(c) : Most people who play video games are men

Plot 3: game3a vs sex

```
ggplot(Q2, aes(x = game3a, fill = sex)) + geom_bar(position = "dodge") + scale_fill_discrete(name = "Se
```

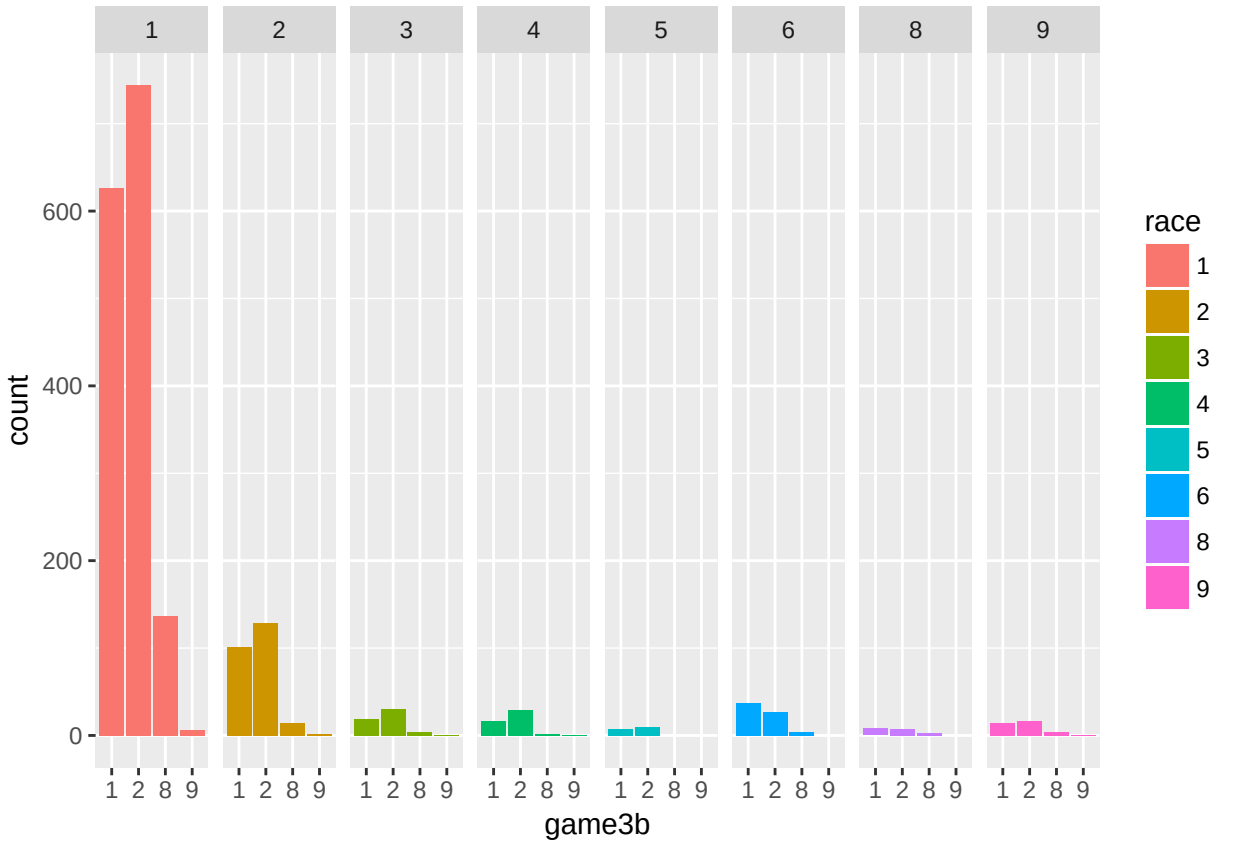


Almost everyone has agreed that most people who play video games are men.

Q2 Part(d) : People who play violent video games are more likely to be violent themselves

Plot 4 : game3b vs race

```
ggplot(Q2, aes(x = game3b, fill = race)) + geom_bar() + facet_grid(~race)
```



Survey states that a person from any race disagrees with the fact that b. People who play violent video games are more likely to be violent themselves.

Q3 About how often do you use the internet?

About how often do you use the internet? [READ] {Modified Teens Relationships}

1 Almost constantly

2 Several times a day

3 About once a day

4 Several times a week, OR

5 Less often?

8 (VOL.) Don't know

9 (VOL.) Refused

Creating dataframe with useful variables

```
Q3 = select(dt,sample,lang,usr,cregion,state,sex,par,educ2,race,inc,age,emplnw,intfreq)
```

We will merge levels of some factors that do not provide much information

```
levels(Q3$intfreq)[7] = "8"
levels(Q3$intfreq)
```

```
## [1] "1" "2" "3" "4" "5" "8"
```

```
levels(Q3$inc)[11] = "98"
```

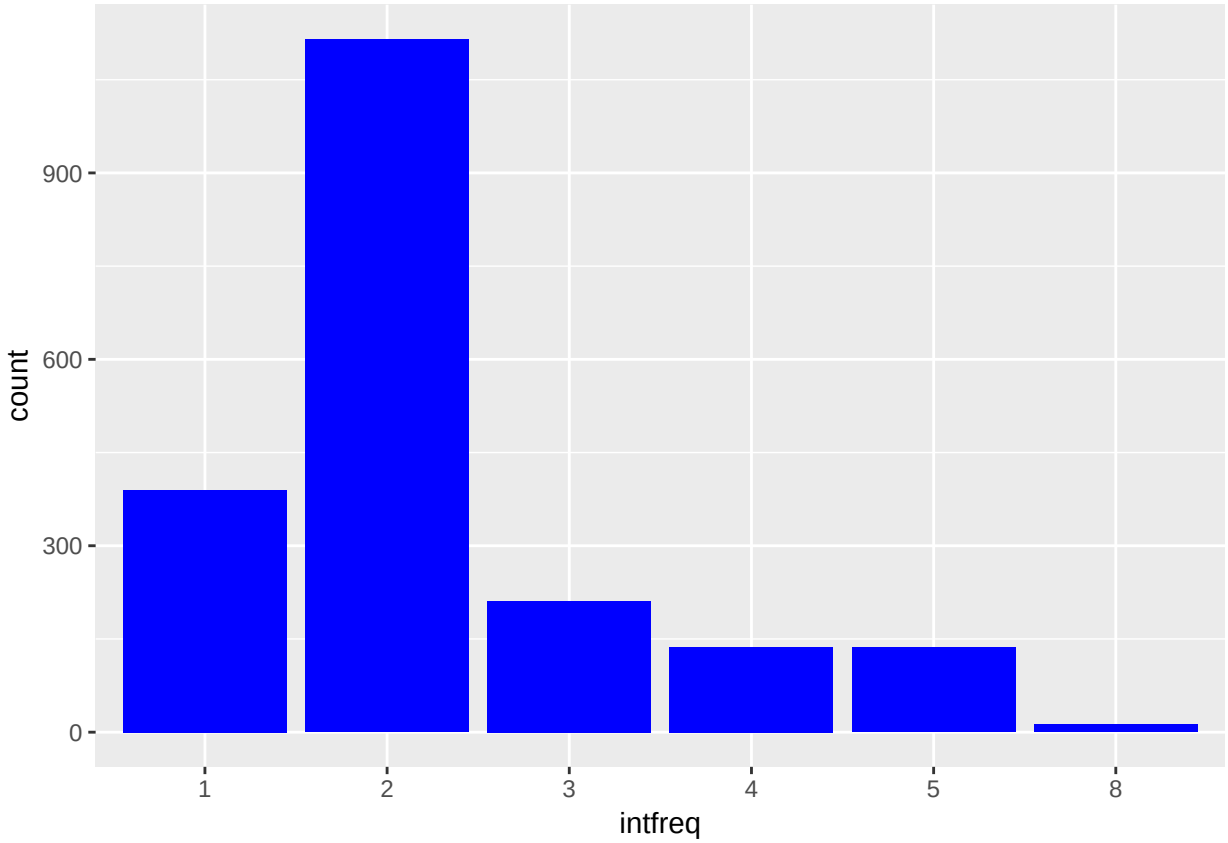
We will be looking at couple of plots to draw inferences and relationship between internet frequency and other factors

Plot 1 : intfreq

```
table(Q3$intfreq)
```

```
##
##    1    2    3    4    5    8
##  390 1115  211  137  136   12
```

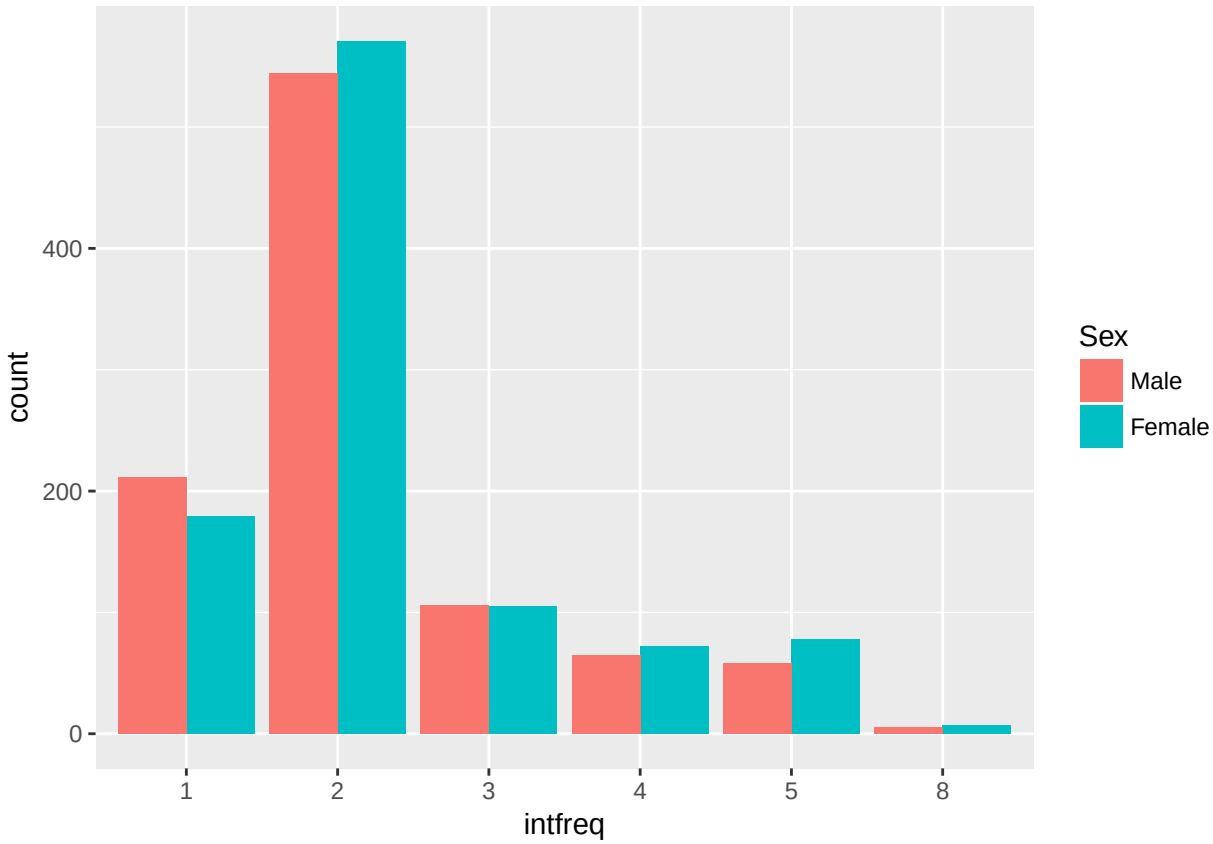
```
ggplot(Q3,aes(intfreq)) + geom_bar(fill = "blue")
```



The above frequency tells us that majority of people voted for option 2 i.e. “Several times a day”.

Plot 2 : intfreq vs sex

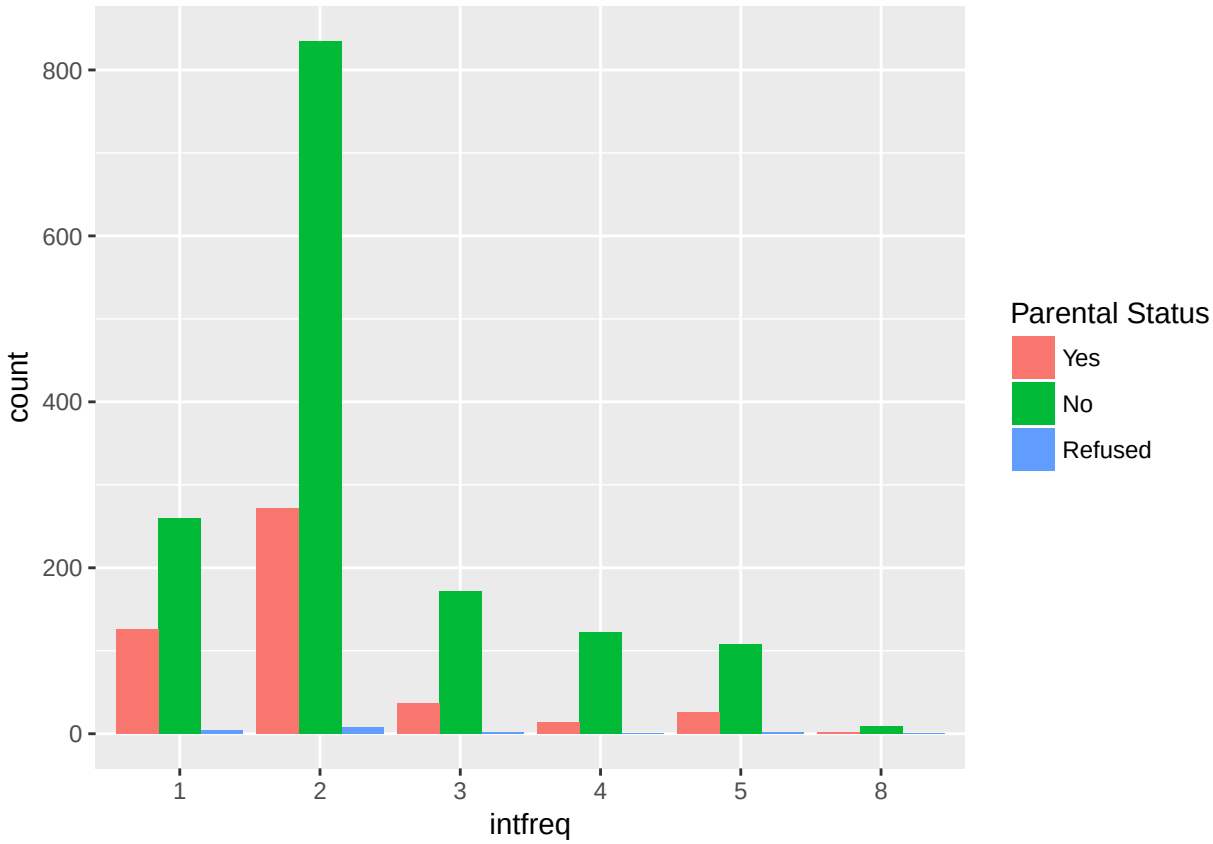
```
ggplot(Q3,aes(intfreq,fill = sex)) + geom_bar(position = "dodge") + scale_fill_discrete(name="Sex",label
```



From the above plot, we can see that males tend to spend more time on internet while there are more number of female users than male users.

Plot 3 : intfreq vs parental

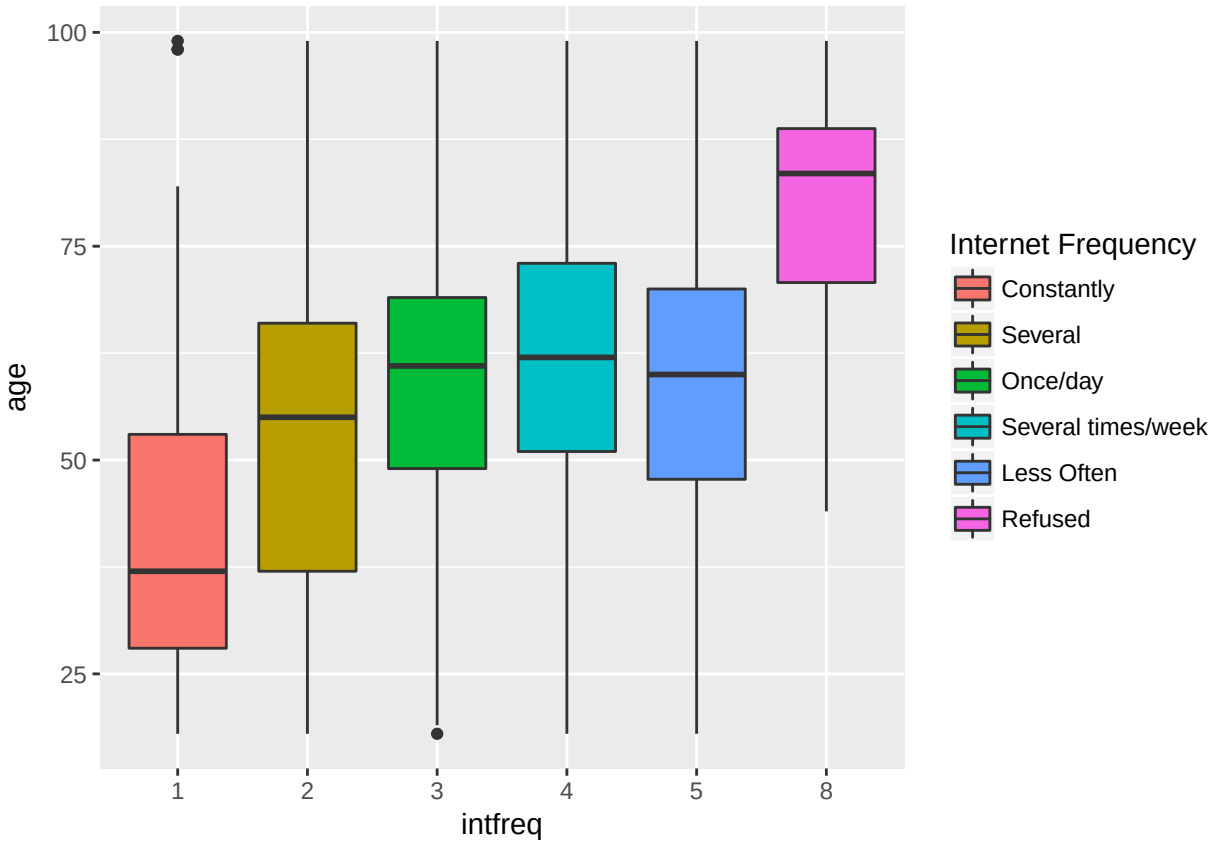
```
ggplot(Q3,aes(intfreq,fill = par)) + geom_bar(position = "dodge") +scale_fill_discrete(name="Parental S
```



From the above graph, we can observe that the people who are parent of under 18 year age tend to use internet comparatively less.

Plot 4 :intfreq vs age

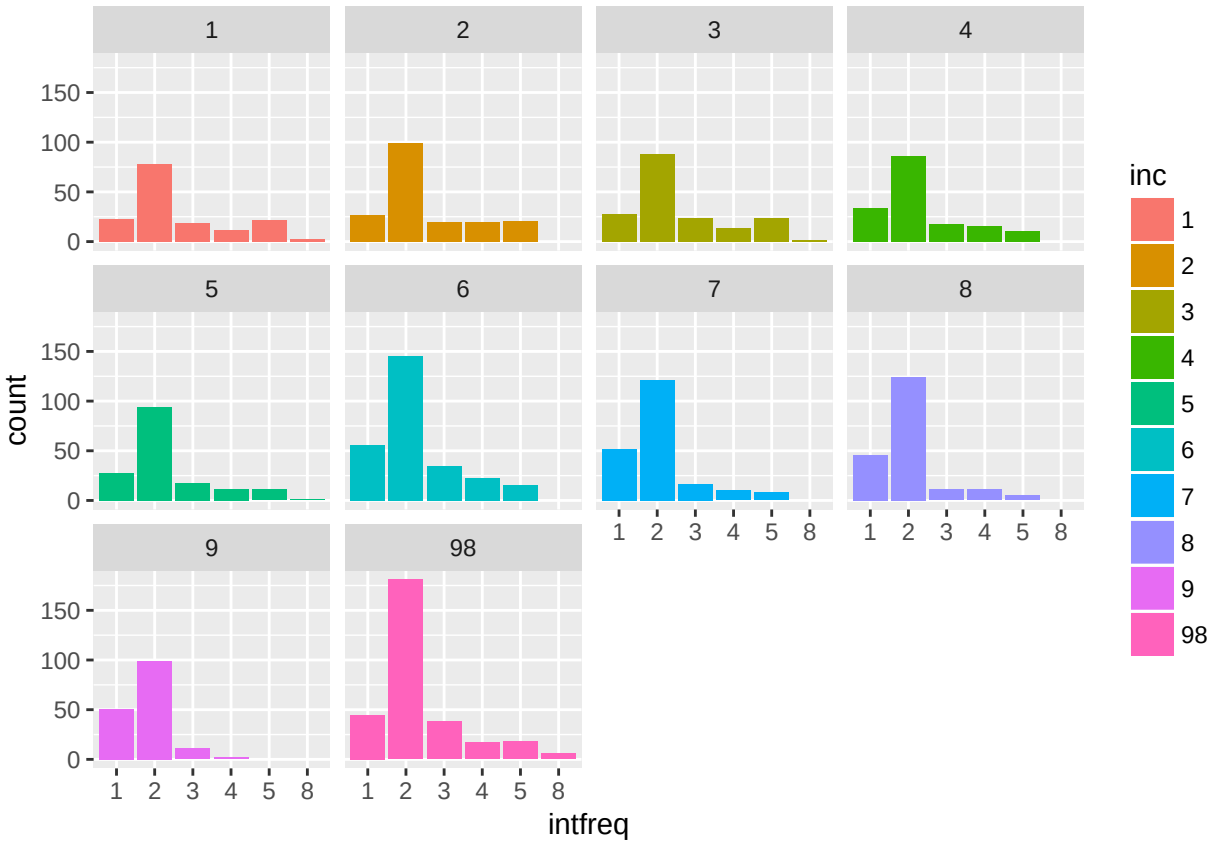
```
ggplot(Q3,aes(y=age,x=intfreq)) + geom_boxplot(aes(fill = intfreq)) + scale_fill_discrete(name="Interne
```



From the above plot,we can conclude that people belonging to younger age group(<40) tend to use internet more oftenly that others.

Plot 5 : intfreq vs income

```
levels(Q3$inc)[11] = "98"
ggplot(Q3,aes(x = intfreq,fill = inc)) + geom_bar() + facet_wrap(~inc)
```



By observing the above plots carefully,we came to know that as income level of people increases they tend to spend more time on internet.