Skill-5: Application of Graphics in R

Yutika

Q.1) The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles(1973–74 models). [, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders [, 3] disp Displacement (cu.in.) [, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (1000 lbs) [, 7] qsec 1/4 mile time [, 8] vs Engine (0 = V-shaped, 1 = straight) [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears [,11] carb Number of carburetors Perform Exploratory Data analysis for the above data sets and Comment on your findings.

A.1)

First, we install all the necessary packages:

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(dplyr)
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages ------------------------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v stringr 1.4.0
## v tidyr   1.0.2     v forcats 0.4.0
## v readr   1.3.1

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## -- Conflicts ---------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(explore)
```

```
## Warning: package 'explore' was built under R version 3.6.3
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.3
```

```
## corrplot 0.84 loaded
```

Now we install the dataset for "mtcars" and create another dataset "mtcars1" which will have the names of the cars as well

```r
data("mtcars")
mtcars1<-add_rownames(mtcars,"Carnames")
```

```
## Warning: Deprecated, use tibble::rownames_to_column() instead.
```

```r
mtcars1
```

```
## # A tibble: 32 x 12
##    Carnames      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <chr>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Mazda RX4    21       6  160    110  3.9   2.62  16.5     0     1     4     4
##  2 Mazda RX4 ~  21       6  160    110  3.9   2.88  17.0     0     1     4     4
##  3 Datsun 710   22.8     4  108     93  3.85  2.32  18.6     1     1     4     1
##  4 Hornet 4 D~  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
##  5 Hornet Spo~  18.7     8  360    175  3.15  3.44  17.0     0     0     3     2
##  6 Valiant      18.1     6  225    105  2.76  3.46  20.2     1     0     3     1
##  7 Duster 360   14.3     8  360    245  3.21  3.57  15.8     0     0     3     4
##  8 Merc 240D    24.4     4  147.    62  3.69  3.19  20       1     0     4     2
##  9 Merc 230     22.8     4  141.    95  3.92  3.15  22.9     1     0     4     2
## 10 Merc 280     19.2     6  168.   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

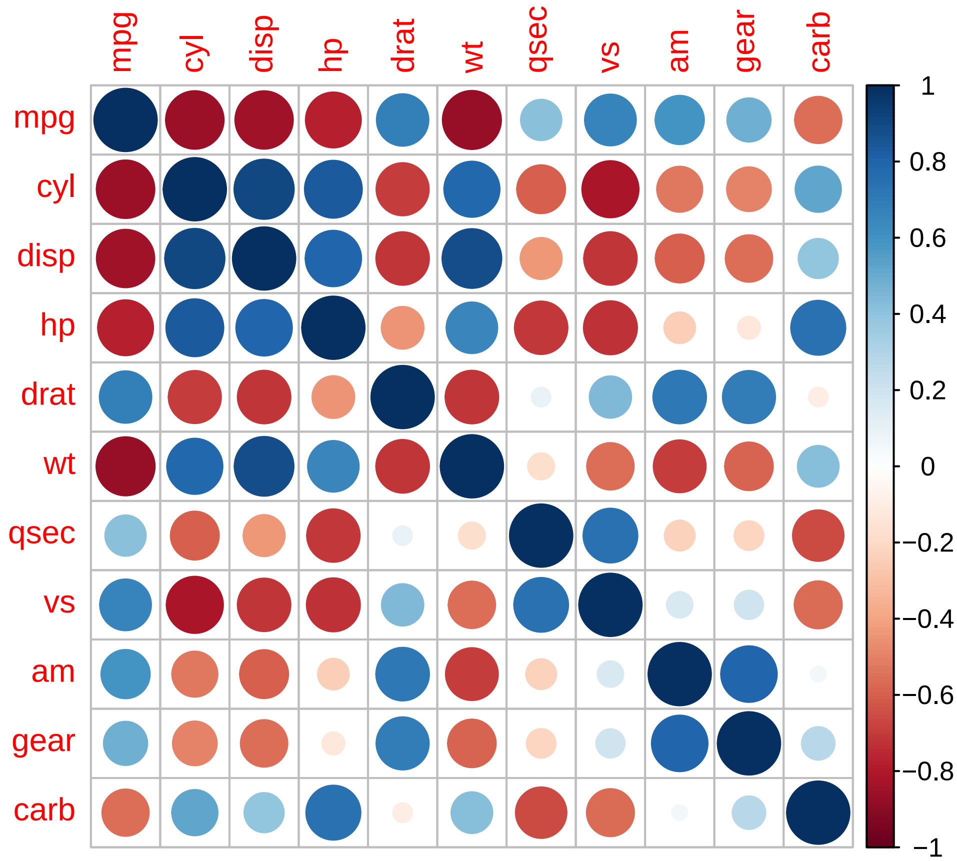
Here, we begin our EDA :- i) First, we find the summary statistics for our dataset :-

```r
summary(mtcars1)
```

```
##    Carnames              mpg             cyl             disp      
##  Length:32          Min.   :10.40   Min.   :4.000   Min.   : 71.1  
##  Class :character   1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8  
##  Mode  :character   Median :19.20   Median :6.000   Median :196.3  
##                     Mean   :20.09   Mean   :6.188   Mean   :230.7  
##                     3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0  
##                     Max.   :33.90   Max.   :8.000   Max.   :472.0  
##        hp             drat             wt             qsec      
##  Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50  
##  1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89  
##  Median :123.0   Median :3.695   Median :3.325   Median :17.71  
##  Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85  
##  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90  
##  Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90  
##        vs               am              gear            carb      
##  Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

ii) Then, we derive a correlation plot from our dataset :-

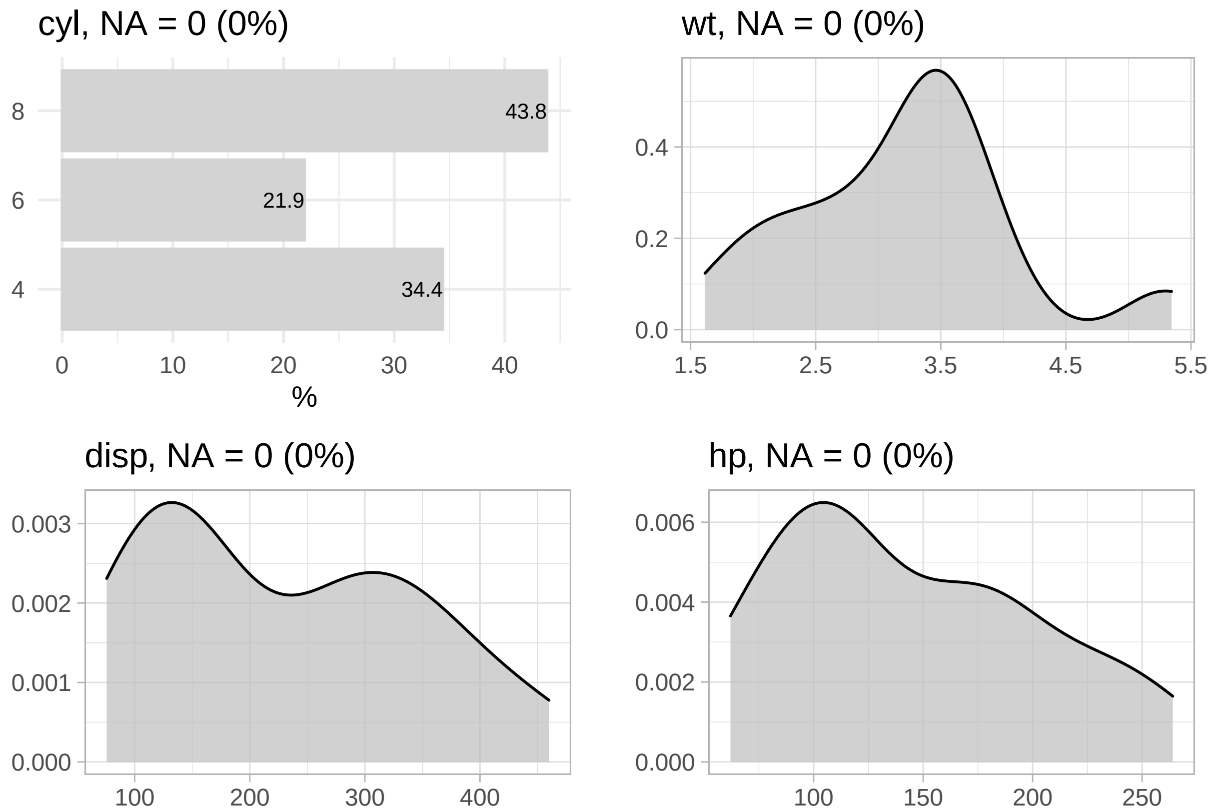
```r
M <- cor(mtcars)
corrplot(M, method = "circle")
```



Comment: From this correlation plot we can note the most obvious relation between any given variables. The darker the colours are, the more strongly correlated the variables are. Blue indicates a positive correlation whereas red indicates a negative correlation. Some variables such as disp-cyl, hp-cyl, hp-disp, wt-disp share a highly positive correlation. On the other hand- the variables: mpg-cyl, mpg-disp, mpg-wt, vs-cyl etc share a highly negative correlation.

iii) Finding the evident correlations between the following variables based on our the observations above :-
POSITIVE CORRELATIONS-

```r
mtcars1 %>%
  select(cyl,wt,disp,hp)%>%
  explore_all()
```



1)Disp, hp, wt and cyl :-

```
mtcars1 %>%
  filter(cyl == 6)%>%
  select("Carnames","hp","wt","cyl","disp")
```

```
## # A tibble: 7 x 5
##   Carnames          hp    wt   cyl  disp
##   <chr>          <dbl> <dbl> <dbl> <dbl>
## 1 Mazda RX4        110  2.62     6  160
## 2 Mazda RX4 Wag    110  2.88     6  160
## 3 Hornet 4 Drive   110  3.22     6  258
## 4 Valiant          105  3.46     6  225
## 5 Merc 280         123  3.44     6  168.
## 6 Merc 280C        123  3.44     6  168.
## 7 Ferrari Dino     175  2.77     6  145
```

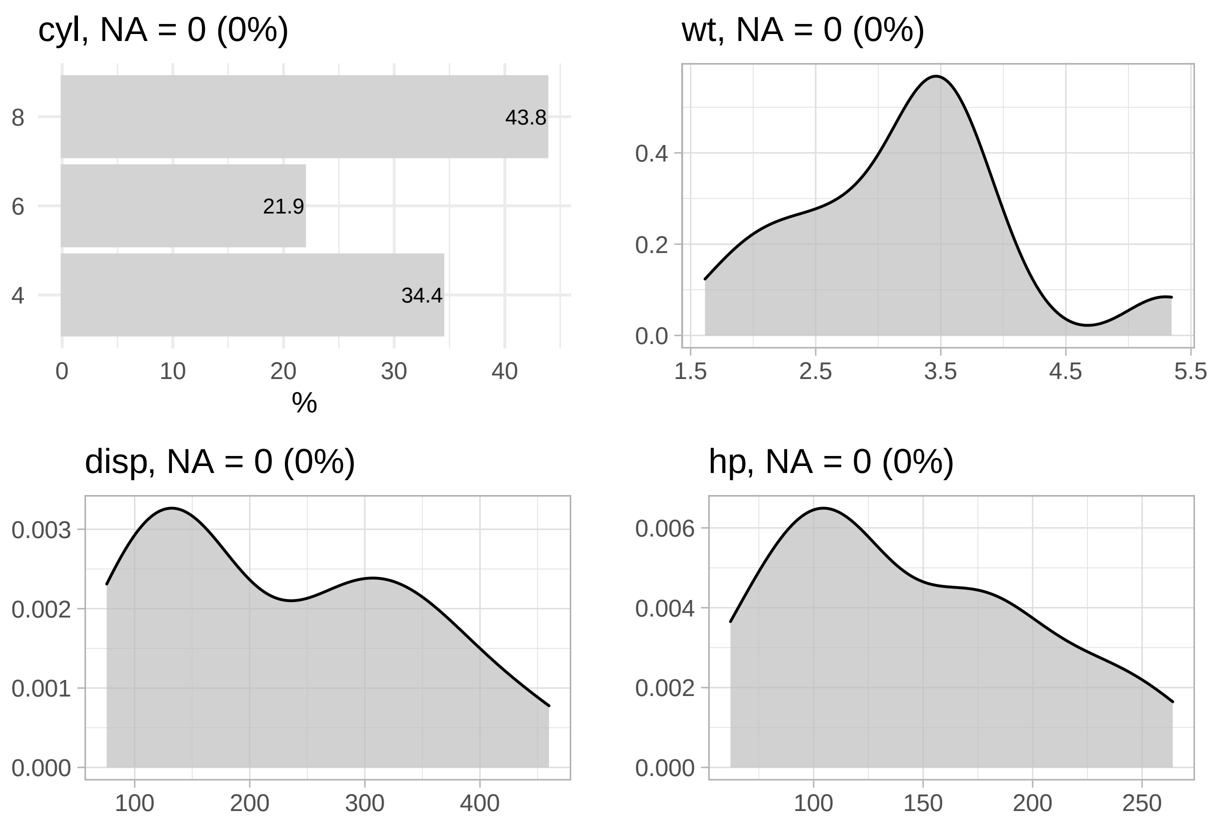
```
mtcars1 %>%
  filter(cyl == 8)%>%
  select("Carnames","hp","wt","cyl","disp")
```

```
## # A tibble: 14 x 5
##    Carnames                hp    wt   cyl  disp
##    <chr>                <dbl> <dbl> <dbl> <dbl>
##  1 Hornet Sportabout      175  3.44     8  360
##  2 Duster 360             245  3.57     8  360
##  3 Merc 450SE             180  4.07     8  276.
##  4 Merc 450SL             180  3.73     8  276.
```

```
##   5 Merc 450SLC           180  3.78      8  276.
##   6 Cadillac Fleetwood    205  5.25      8  472
##   7 Lincoln Continental   215  5.42      8  460
##   8 Chrysler Imperial     230  5.34      8  440
##   9 Dodge Challenger      150  3.52      8  318
##  10 AMC Javelin           150  3.44      8  304
##  11 Camaro Z28            245  3.84      8  350
##  12 Pontiac Firebird      175  3.84      8  400
##  13 Ford Pantera L        264  3.17      8  351
##  14 Maserati Bora         335  3.57      8  301
```

(Only the cars with 6 and 8 cylinders have been considered here)

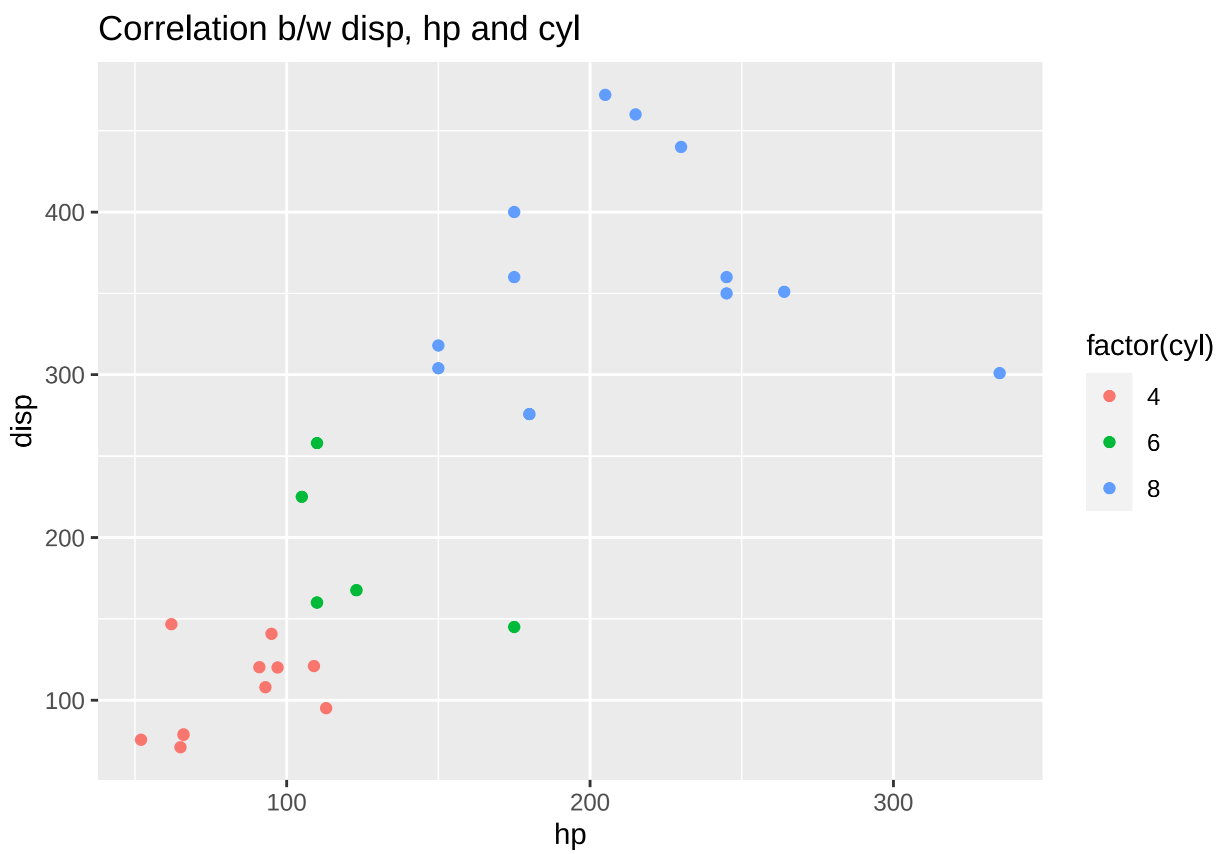
```r
mtcars1 %>%
  select(cyl,wt,disp,hp)%>%
  explore_all()
```




```r
library(ggplot2)
```

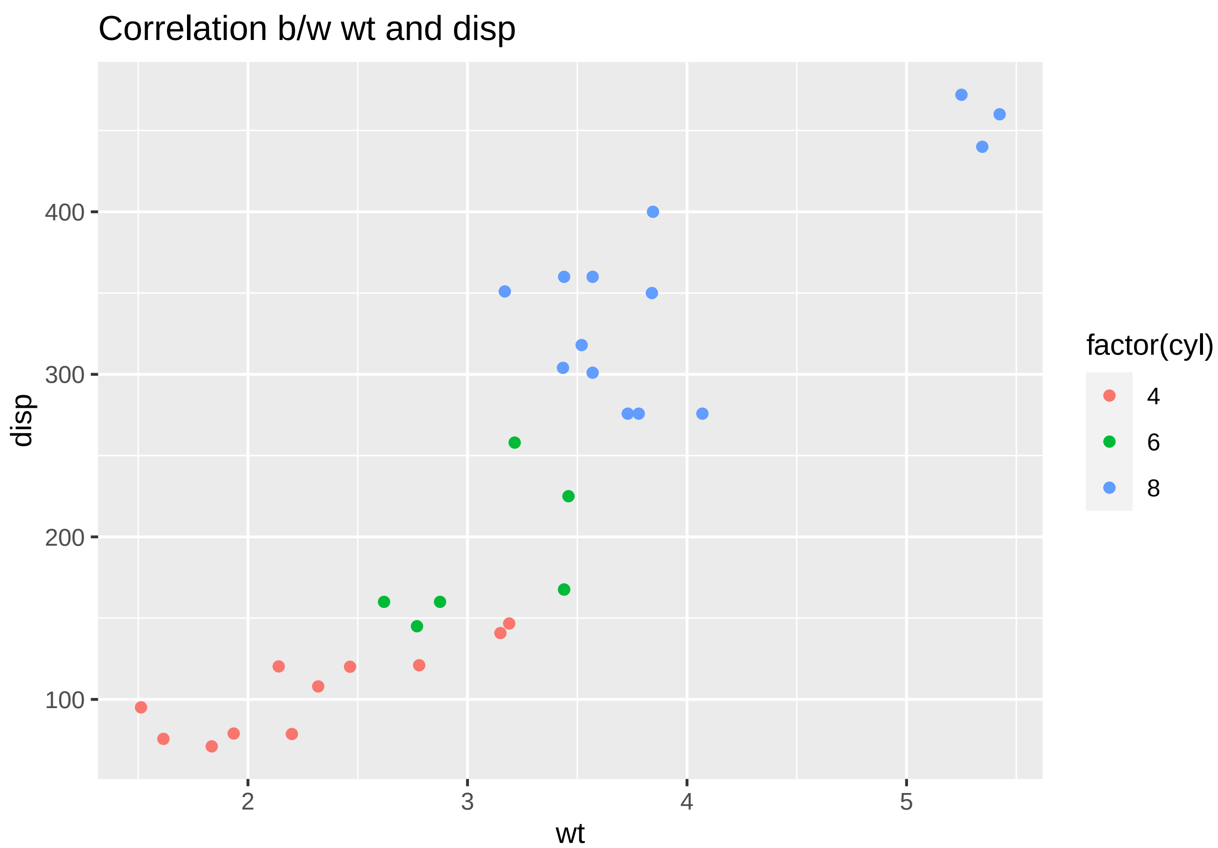
**a) Correlation : disp, hp; factor: cyl*

```r
ggplot(mtcars1, aes(x = hp , y = disp)) +
geom_point(aes(colour = factor(cyl))) + ggtitle ("Correlation b/w disp, hp and cyl")
```



b) Correlation: wt, disp; factor: cyl

```
ggplot(mtcars1, aes(x = wt , y = disp)) +
geom_point(aes(colour= factor(cyl))) + ggtitle ("Correlation b/w wt and disp")
```



Comment: With increase in the number of cylinders, there is a remarkable growth in the horsepower and hence the displacement.Weight (wt) and displacement(disp) also seem to have a positive correlation when factored by the no. of cylinders.

```
max(mtcars1$wt)
```

```
## [1] 5.424
```

```
min(mtcars1$wt)
```

```
## [1] 1.513
```

```
mtcars1%>%
  filter(wt==5.424)%>%
  select("Carnames","hp","wt","cyl","disp")
```

```
## # A tibble: 1 x 5
##   Carnames                hp    wt   cyl  disp
##   <chr>                <dbl> <dbl> <dbl> <dbl>
## 1 Lincoln Continental    215  5.42     8   460
```

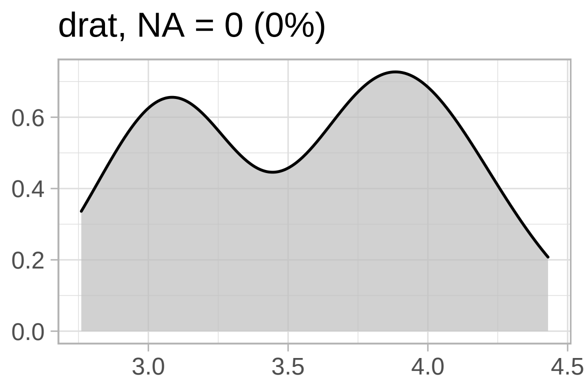
```
mtcars1%>%
  filter(wt==1.513)%>%
  select("Carnames","hp","wt","cyl","disp")
```

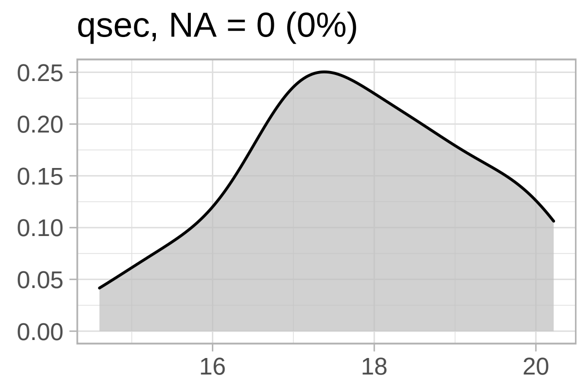
```
## # A tibble: 1 x 5
##   Carnames         hp    wt   cyl  disp
##   <chr>         <dbl> <dbl> <dbl> <dbl>
## 1 Lotus Europa    113  1.51     4  95.1
```

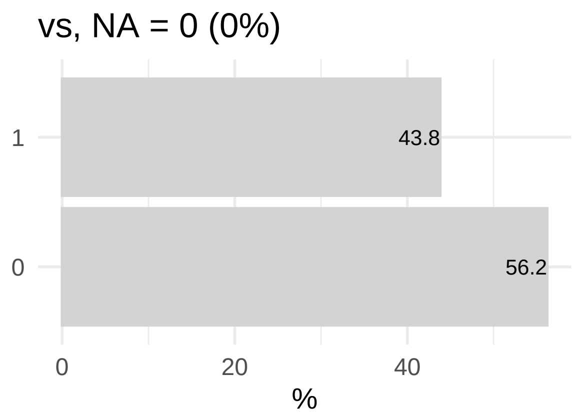
A car like the ***Lincoln Continental*** which requires 215 hp, has 8 cylinders weighs around 5,424 lbs and has a displacement of 460 cu.in. whereas the ***Lotus Europa*** which requires 113 hp, has 4 cylinders, weighs around 1,511 lbs and has a displacement of only about 95.1 cu.in. This justifies our assumptions about the correlation between their respective weights, no. of cylinders, displacement and horsepower.

2) QSEC, drat, vs :-

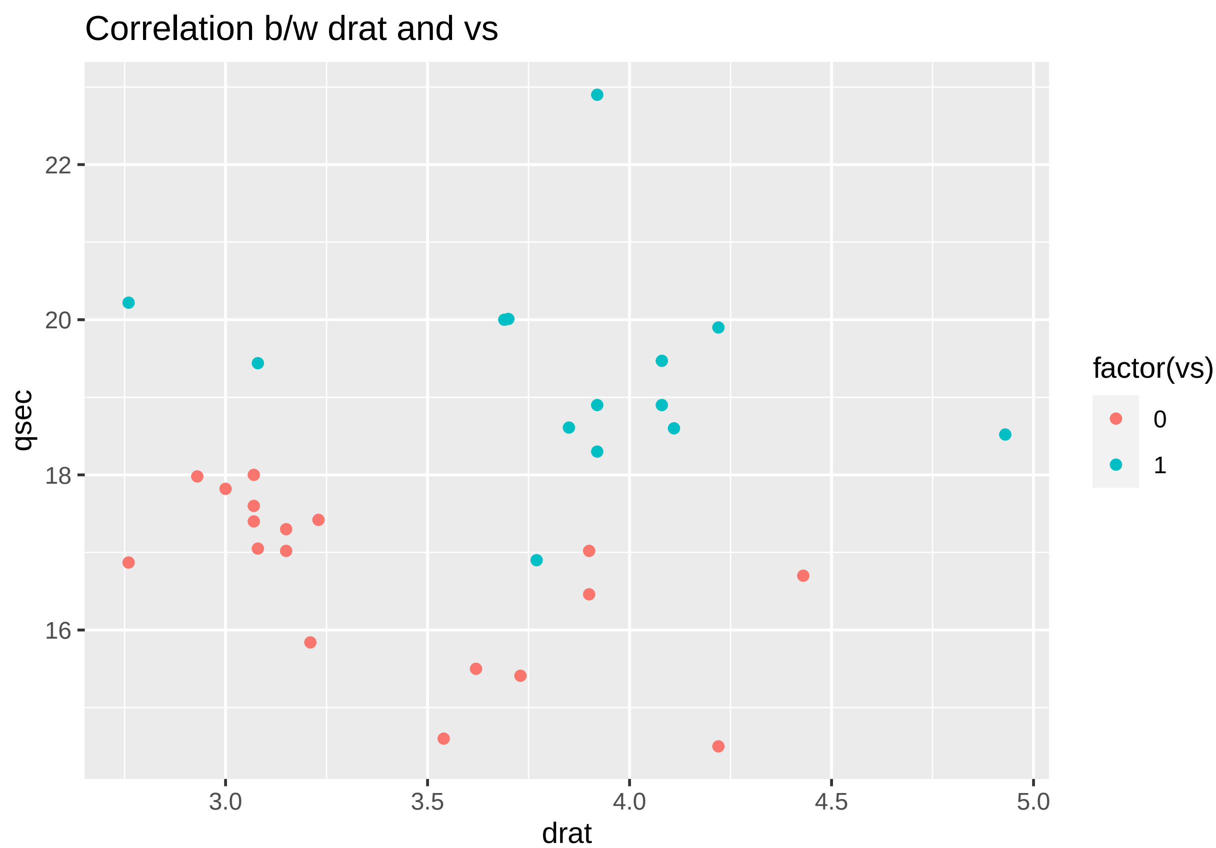
```
mtcars1 %>%
  select(drat,qsec,vs)%>%
  explore_all()
```





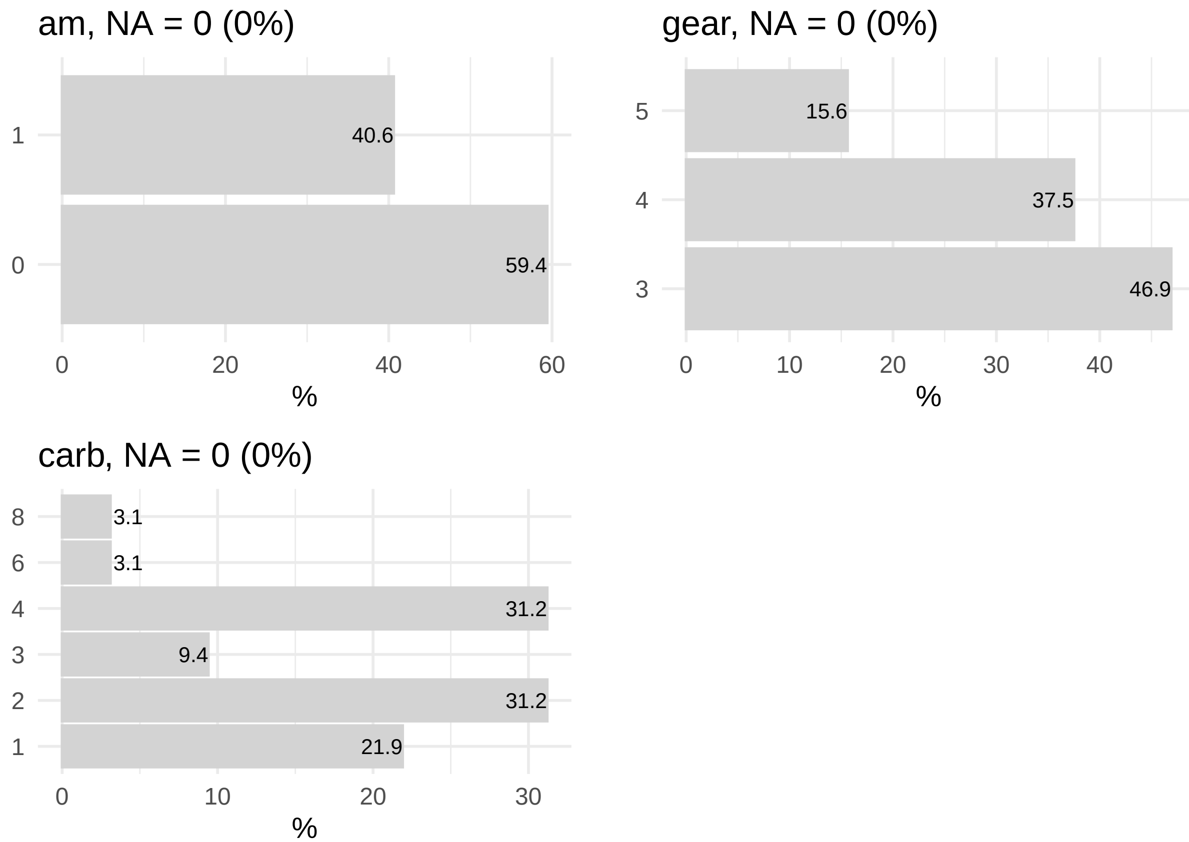



```
ggplot(mtcars1, aes(x = drat , y = qsec)) +
geom_point(aes(colour= factor(vs))) + ggtitle ("Correlation b/w drat and vs")
```

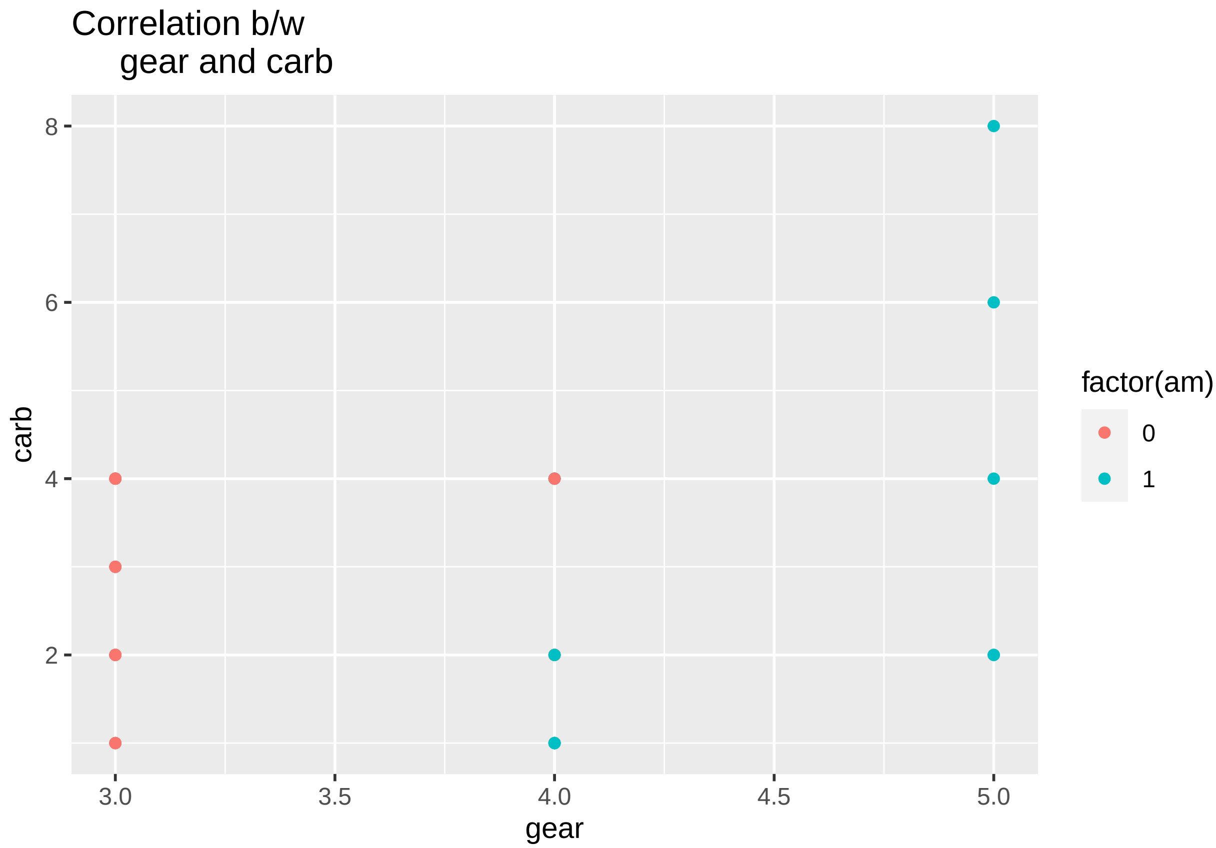


3) am, gear and carb :- Correlation : am, gear; factor: carb

```
mtcars1 %>%
  select(am,gear,carb)%>%
  explore_all()
```



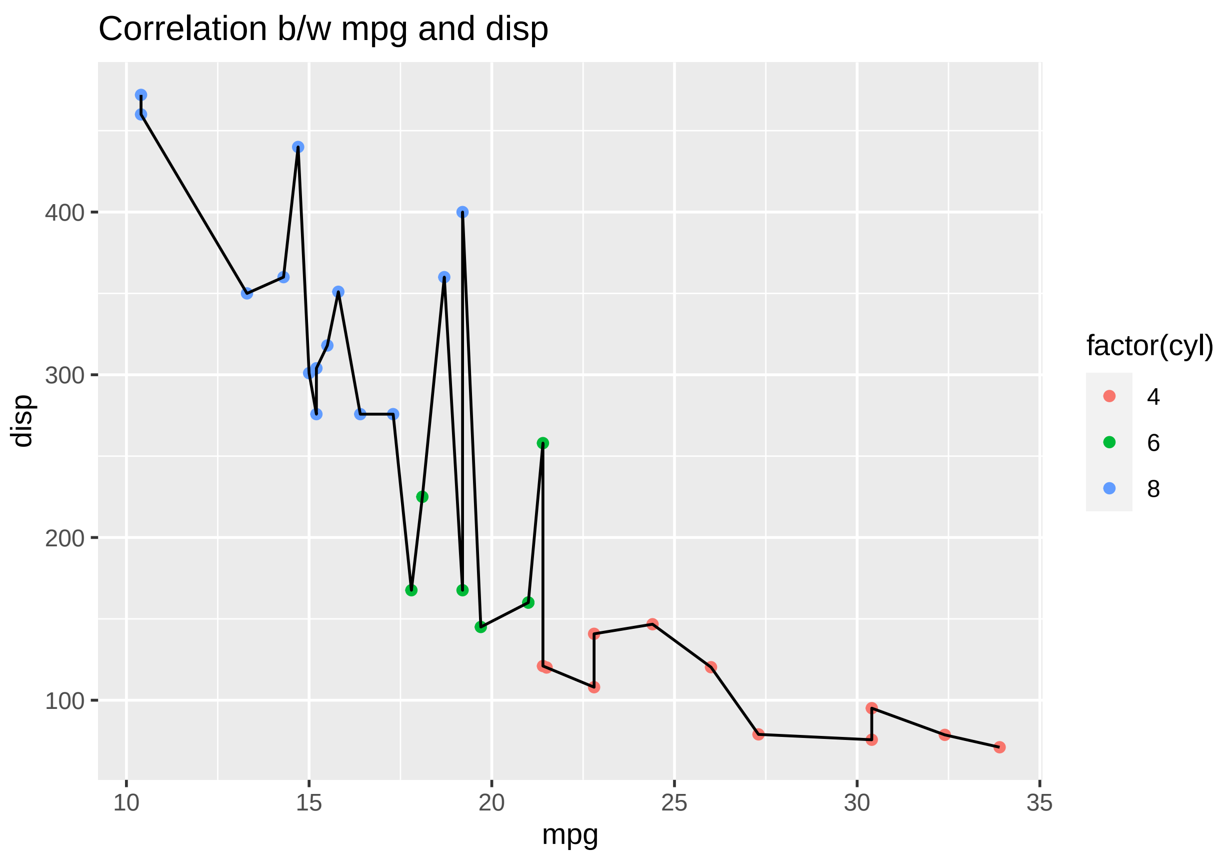

```
ggplot(mtcars1, aes(x = gear , y = carb)) +
geom_point(aes(colour= factor(am))) + ggtitle ("Correlation b/w
      gear and carb")
```



Comment: V-shaped engines (vs: 0) show less rear axle ratio and quarter mile time actions whereas the straight engines(vs: 0) tend to show more drat and qsec actions. Cars with a manual transmission have more no. of carburettors and forward gears. Likewise, cars with automatic transmission have lesser no. of carburettors and evidently, less no. of forward gears.

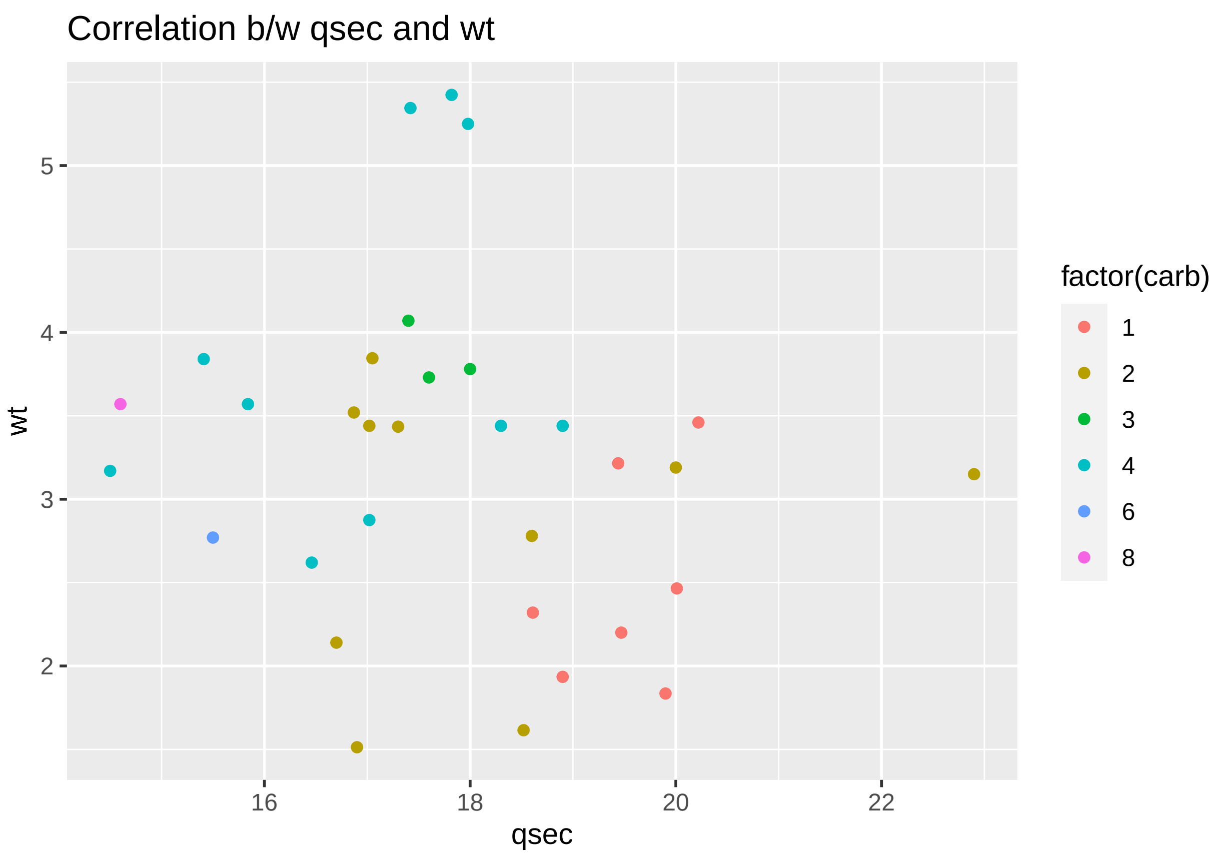
NEGATIVE CORRELATIONS - 1)Correlation between: mg, disp ; factor: cyl

```
ggplot(mtcars1, aes(x = mpg , y = disp)) +
geom_point(aes(colour= factor(cyl))) + geom_line()+ ggtitle ("Correlation b/w mpg and disp")
```



2)Qsec, wt; factor - carb

```
ggplot(mtcars1, aes(x = qsec , y = wt)) +
  geom_point(aes(colour= factor(carb))) + ggtitle ("Correlation b/w qsec and wt")
```



Comment: The more the number of cylinders, the lesser the mileage per gallon and consequently even lesser displacement.

```
max(mtcars1$disp)
```

```
## [1] 472
```

```
min(mtcars1$disp)
```

```
## [1] 71.1
```

```
mtcars1 %>%
  filter(disp==472)%>%
  select("Carnames","disp","mpg","cyl")
```

```
## # A tibble: 1 x 4
##   Carnames             disp   mpg   cyl
##   <chr>               <dbl> <dbl> <dbl>
## 1 Cadillac Fleetwood   472  10.4     8
```

```
mtcars1 %>%
  filter(disp==71.1)%>%
  select("Carnames","disp","mpg","cyl")
```

```
## # A tibble: 1 x 4
##   Carnames         disp   mpg   cyl
##   <chr>           <dbl> <dbl> <dbl>
## 1 Toyota Corolla   71.1  33.9     4
```

Toyota Corolla has a displacement of 71.1 cu.in. a mileage per gallon of 33.9 and has 4 cylinders whereas

Cadillac Fleetwood has a displacement of 472 cu.in. a mileage per gallon of 10.4 and has 8 cylinders. This implies that the no. of cylinders and displacement have negative or no correlation at all and that there appears to be no dependence of these factors on one another.

Q.2)The admission data of three popular colleges is given. Visualize the table using appropriate tools and comment on your findings.

Creation of a data frame for "Colleges".

```
college <- c("ABC","ABC","ABC","XYZ","XYZ","XYZ","PQR","PQR","PQR")
stream <- c("Arts","Commerce","Science","Arts","Commerce","Science","Arts","Commerce","Science")
male <-c(60,124,210,56,231,210,45,120,134)
female= c(60,128,220,67,231,230,45,130,166)
total = c(120,252,430,123,462,440,90,250,300)

df <- data.frame(college,stream,male,female,total)
df
```

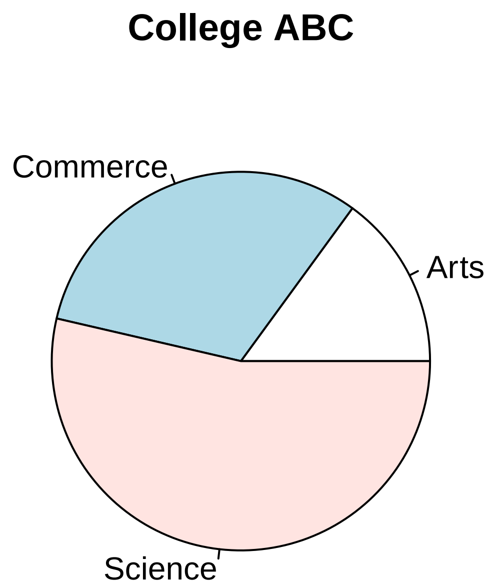
```
##   college   stream male female total
## 1     ABC     Arts   60     60   120
## 2     ABC Commerce  124    128   252
## 3     ABC  Science  210    220   430
## 4     XYZ     Arts   56     67   123
## 5     XYZ Commerce  231    231   462
## 6     XYZ  Science  210    230   440
## 7     PQR     Arts   45     45    90
## 8     PQR Commerce  120    130   250
## 9     PQR  Science  134    166   300
```

Visualization :- 1)Pie charts for strength per college :- COLLEGE-ABC :-

```
col_abc <- data.frame(subset(df, subset = college == "ABC"))
pa <- col_abc[1:3,c(2:5)]
pa
```

```
##     stream male female total
## 1     Arts   60     60   120
## 2 Commerce  124    128   252
## 3  Science  210    220   430
```

```
pie(pa$total, labels = pa$stream, main = "College ABC")
```

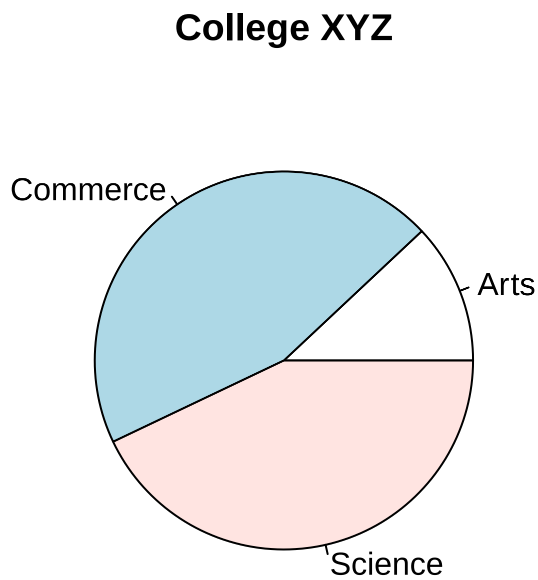


COLLEGE-XYZ :-

```
col_xyz <- data.frame(subset(df, subset = college == "XYZ"))
px <- col_xyz[1:3,c(2:5)]
px
```

```
##     stream male female total
## 4     Arts   56     67   123
## 5 Commerce  231    231   462
## 6  Science  210    230   440
```

```
pie(px$total, labels = px$stream, main = "College XYZ")
```

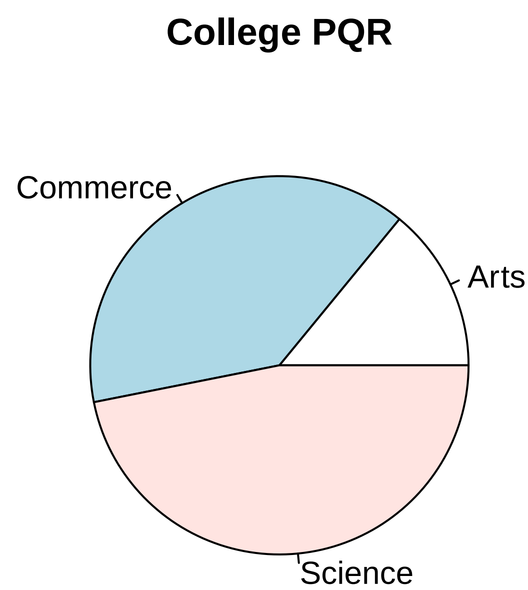


COLLEGE-PQR :-

```
col_pqr<- data.frame(subset(df, subset = college == "PQR"))
pp <- col_pqr[1:3,c(2:5)]
pp
```

```
##     stream male female total
## 7     Arts   45     45    90
## 8 Commerce  120    130   250
## 9  Science  134    166   300
```

```
pie(pp$total, labels = pp$stream, main = "College PQR")
```

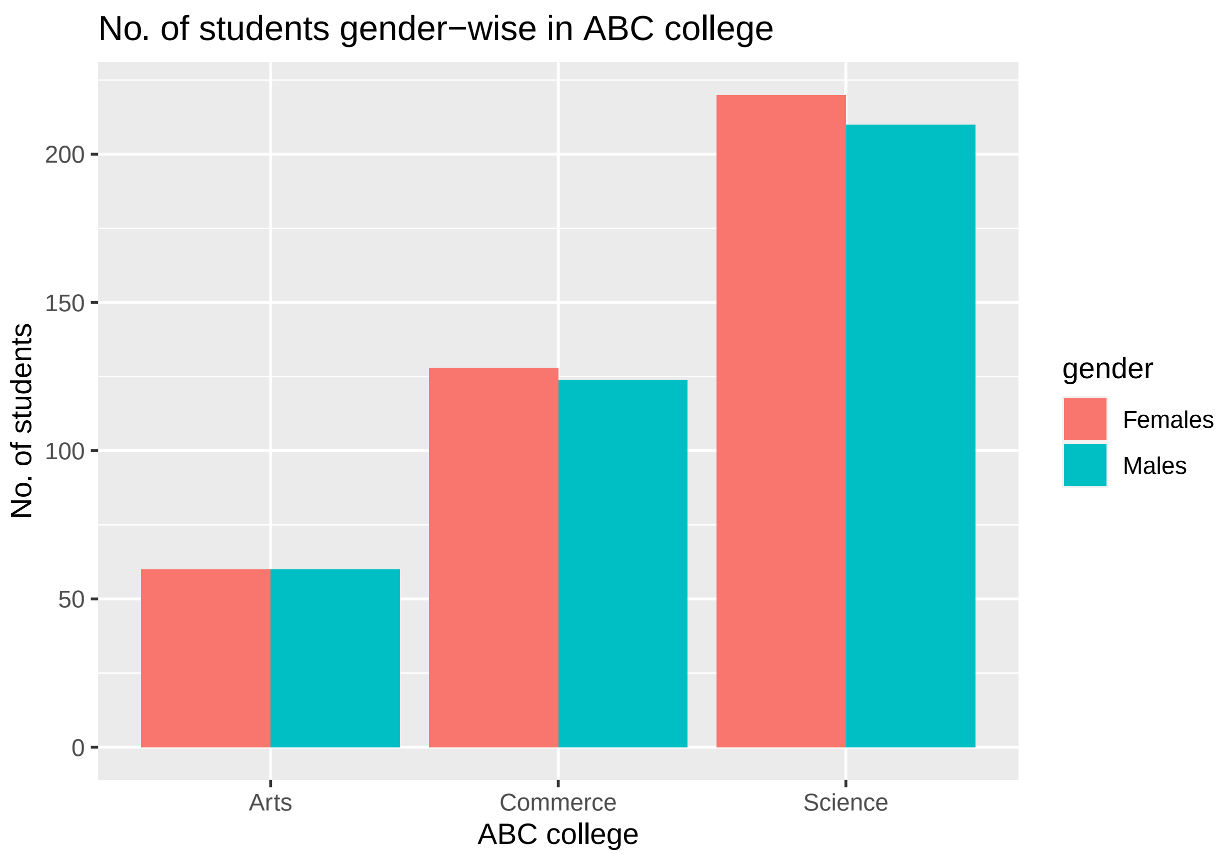


Comments : From the pie charts, it can be observed that very less no. of students have enrolled themselves in the Arts stream. Majority of the students have enrolled in either Commerce or Science.

2)Comparative bar plots for frequency of males and females per stream in colleges ABC, XYZ & PQR :-

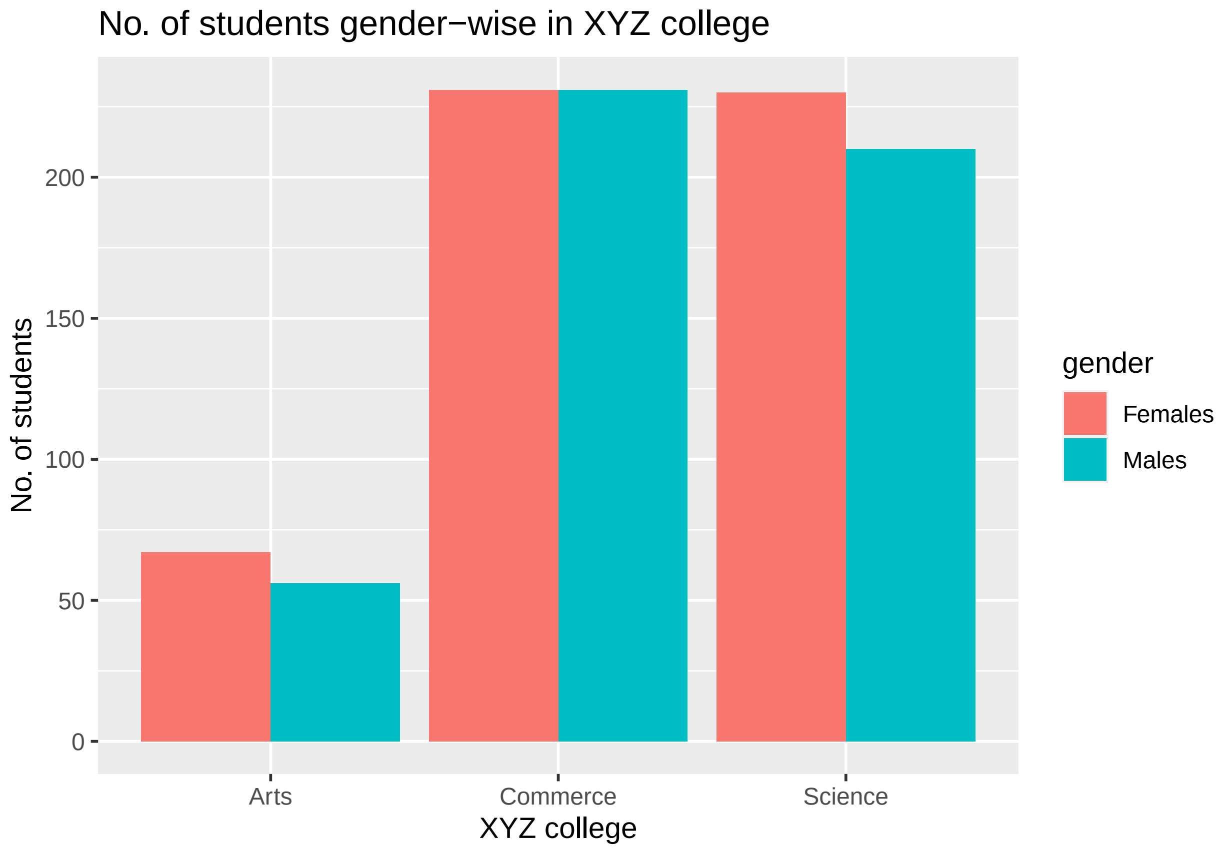
COLLEGE-ABC :-

```
col_abc <- data.frame(subset(df,subset = college == "ABC"))
x <- col_abc[1:3, 2:4]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$stream, freq)
ggplot(plot, aes(x = x.stream, y = freq, fill = gender)) + xlab("ABC college") + ylab("No. of students")
```



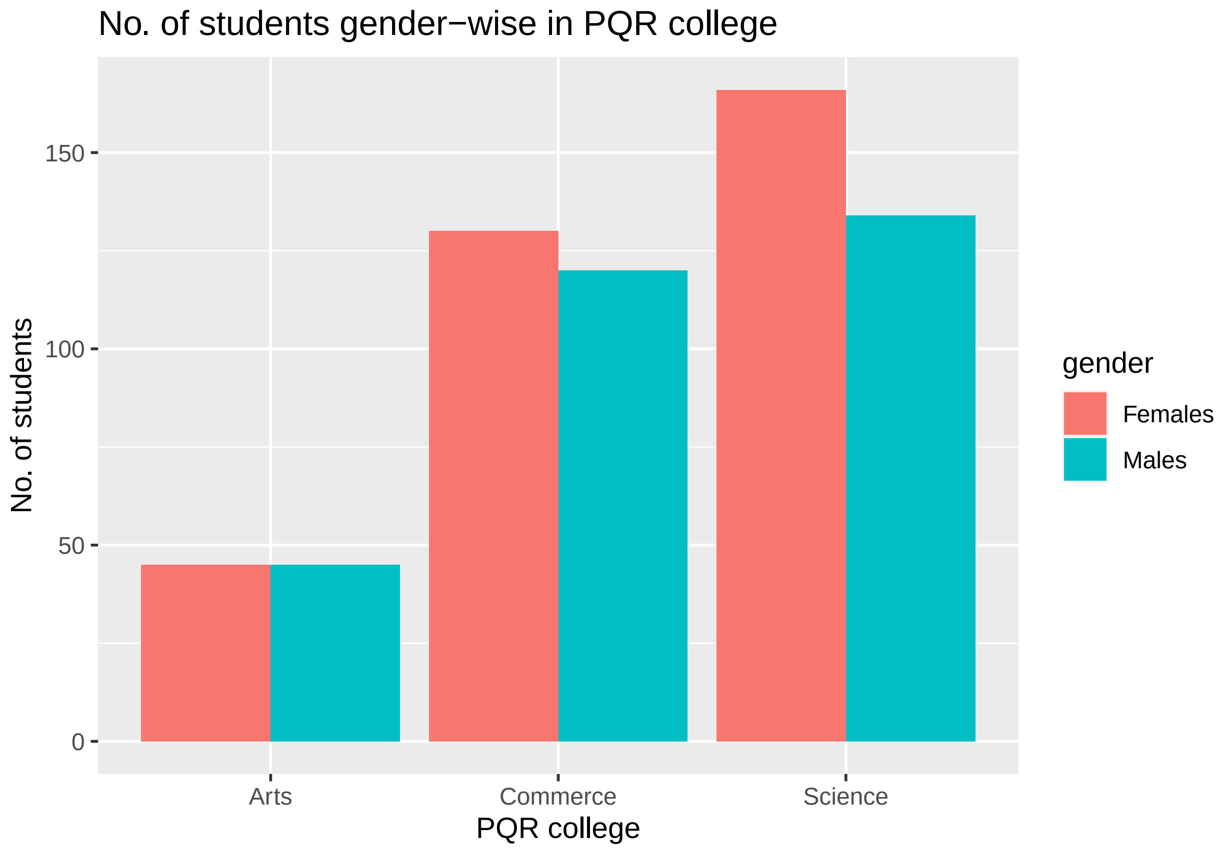
COLLEGE-XYZ :-

```
col_xyz <- data.frame(subset(df,subset = college == "XYZ"))
x <- col_xyz[1:3, 2:4]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$stream, freq)
ggplot(plot, aes(x = x.stream, y = freq, fill = gender)) + xlab("XYZ college") + ylab("No. of students"
```



COLLEGE-PQR :-

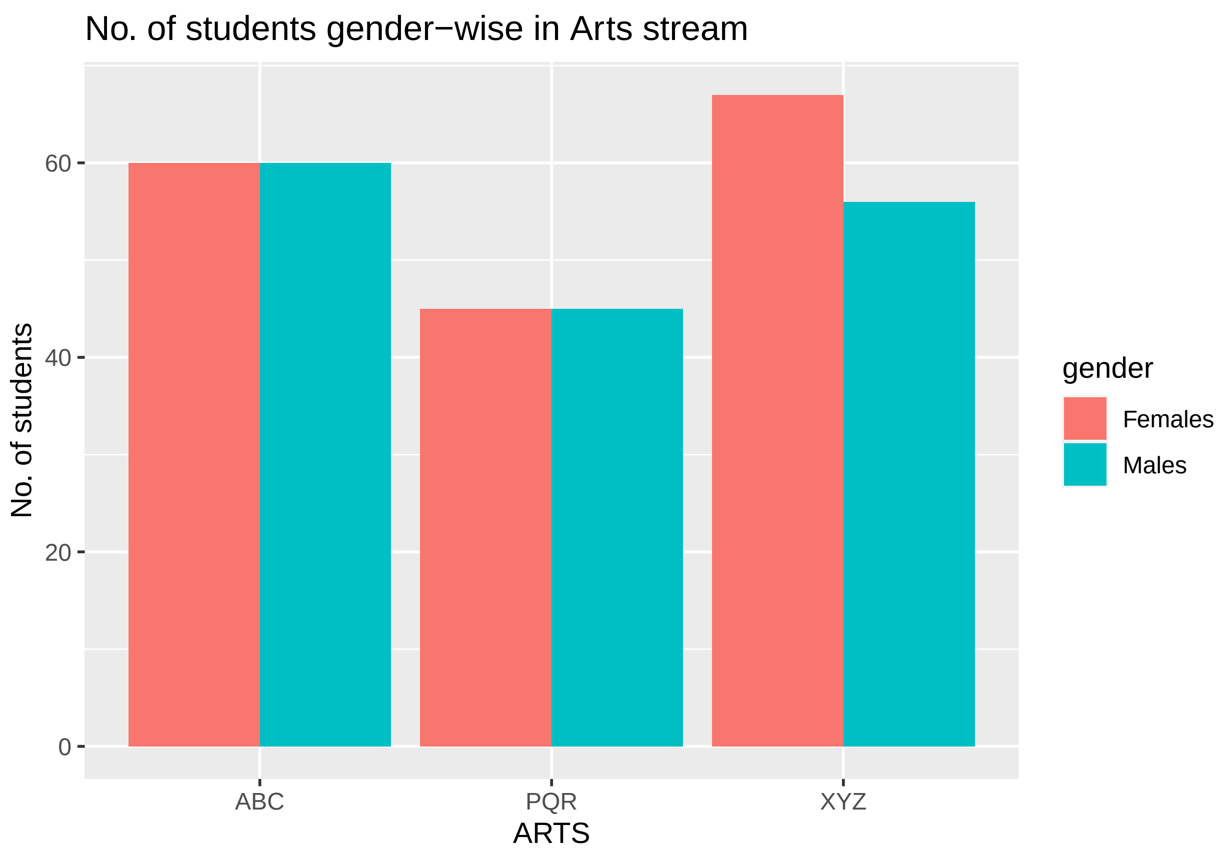
```
col_pqr <- data.frame(subset(df,subset = college == "PQR"))
x <- col_pqr[1:3, 2:4]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$stream, freq)
ggplot(plot, aes(x = x.stream, y = freq, fill = gender)) + xlab("PQR college") + ylab("No. of students")
```



3)Comparative bar plots for the various streams :-

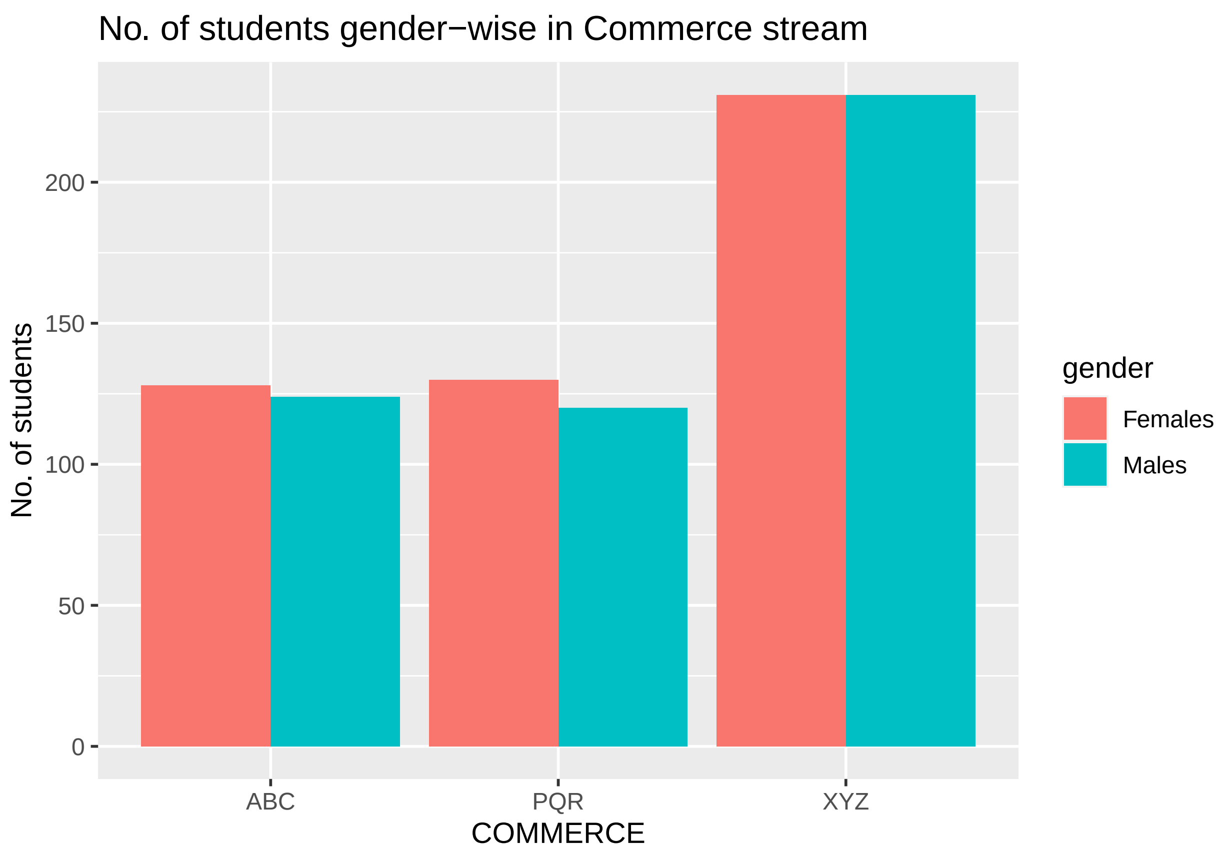
ARTS :-

```
a <- data.frame(subset(df,subset = stream == "Arts"))
x <- a[1:3, c(1,3,4)]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$college, freq)
ggplot(plot, aes(x = x.college, y = freq, fill = gender)) + xlab("ARTS") + ylab("No. of students") + ge
```



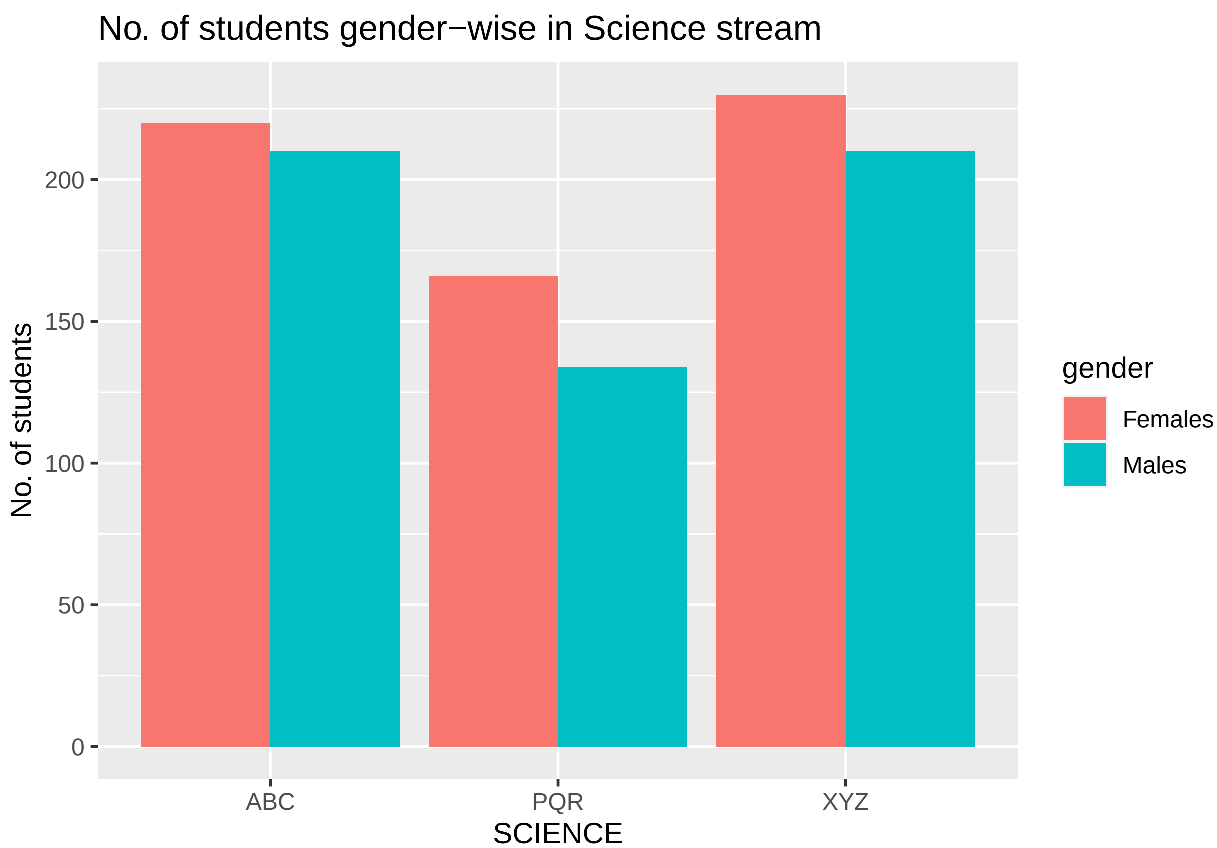
COMMERCE:-

```
c <- data.frame(subset(df,subset = stream == "Commerce"))
x <- c[1:3, c(1,3,4)]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$college, freq)
ggplot(plot, aes(x = x.college, y = freq, fill = gender)) + xlab("COMMERCE") + ylab("No. of students") +
```



SCIENCE:-

```
s <- data.frame(subset(df,subset = stream == "Science"))
x <- s[1:3, c(1,3,4)]
freq <- c(x$male, x$female)
plot <- data.frame(gender = rep(c("Males","Females"), each = 3), x$college, freq)
ggplot(plot, aes(x = x.college, y = freq, fill = gender)) + xlab("SCIENCE") + ylab("No. of students") +
```



Comments: It is observed that in all three colleges: ABC, XYZ and PQR, the number of females are more than or equal to the no. of males. It is also seen that in all three colleges the proportion of females in the Science and Commerce streams are particularly higher in comparison to males. Overall, there seem to be less students in college PQR for any given stream while many students seem to prefer college XYZ. The no. of students in ABC is neutral.