Applications of Graphics Using R - Surya Bhosale

Surya Bhosale

27/03/2020

Importing required packages:

```
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(plotrix)
```

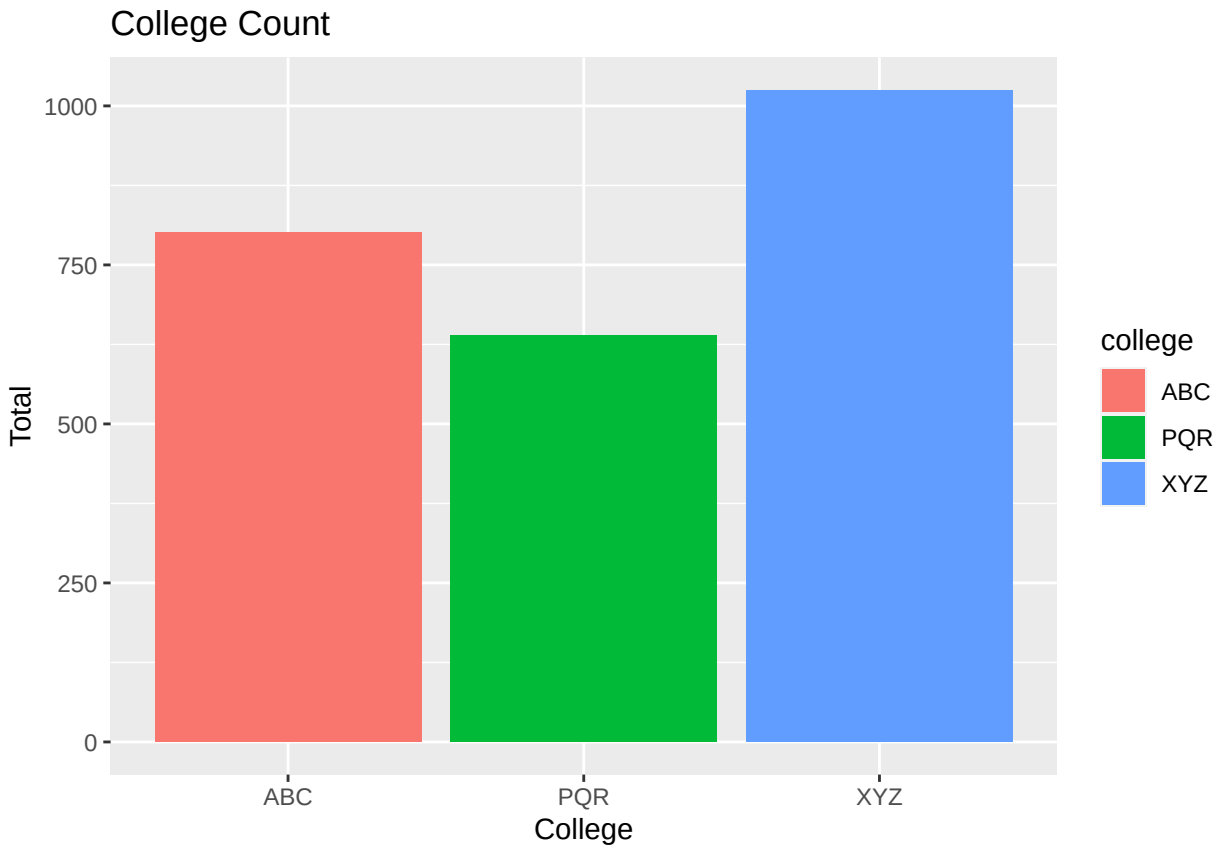
Creating the dataframe:

```
college = c("ABC", "ABC", "ABC", "XYZ", "XYZ", "XYZ", "PQR","PQR","PQR")
Stream = c("Arts", "Commerce", "Science","Arts", "Commerce", "Science","Arts", "Commerce", "Science")
datacol = data.frame(college, Stream)
male = c(60,124, 210,56,231,210,45,120,134)
female = c(60,128,220,67, 231,230,45,130,166)
datacol$male = male
datacol$female = female
total = datacol$male + datacol$female
datacol$total = total
datacol
```

```
##   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
```

Commencing visulatisation:

```
ggplot(datacol, aes(factor(college), total, fill = college)) + geom_col() + ggtitle("College Count") + x
```



The largest college with regards to total admissions is XYZ, and the smallest is PQR.

Data segmentation:

```
ABC = datacol[1:3,]
XYZ = datacol[4:6,]
PQR = datacol[7:9,]
```

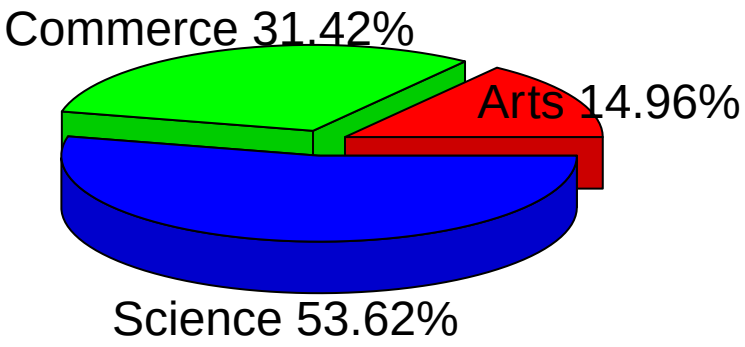
Analysis of ABC:

```
ABC
```

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

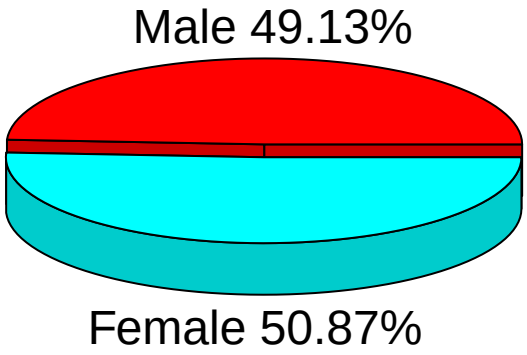
```
labels = ABC$Stream
pct = round(ABC$total/sum(ABC$total)*100, digits = 2)
labels = paste(labels, pct)
labels = paste(labels, "%", sep = "")
pie3D(ABC$total, labels = labels, explode = 0.1, main = "Proportion Of Total Student As Per Stream")
```

Proportion Of Total Student As Per Stream



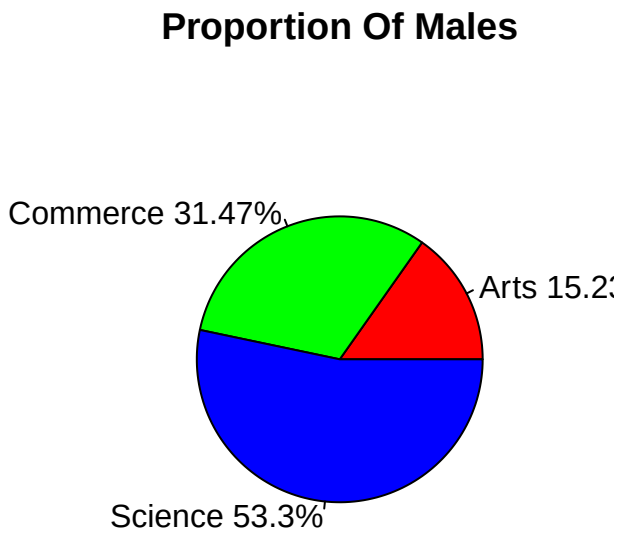

```
abc_pm = round((sum(ABC$male)/sum(ABC$total)*100), digits = 2)
abc_pf = 100 - abc_pm
p = c(abc_pm, abc_pf)
labels1 = paste(c("Male","Female"), p)
labels1 = paste(labels1, "%", sep = "")
pie3D(p, labels = labels1, col = rainbow(length(labels1)), explode = 0.05, main = "Proportion Of Males &
```

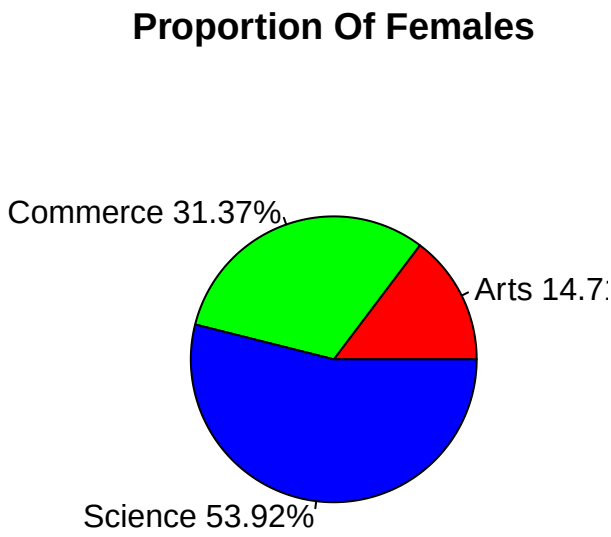
Proportion Of Males & Females In College ABC




```
par(mfrow=c(1,2))

pctmales = round(ABC$male/sum(ABC$male)*100, digits = 2)
labels3 = ABC$Stream
labels3 = paste(labels3, pctmales)
labels3 = paste(labels3, "%", sep = "")
pie(pctmales, labels = labels3, main = "Proportion Of Males", col = rainbow(length(labels3)))

pctfemales = round(ABC$female/sum(ABC$female)*100, digits = 2)
labels3 = ABC$Stream
labels3 = paste(labels3, pctfemales)
labels3 = paste(labels3, "%", sep = "")
```

```
pie(pctfemales, labels = labels3, main = "Proportion Of Females", col = rainbow(length(labels3)))
```



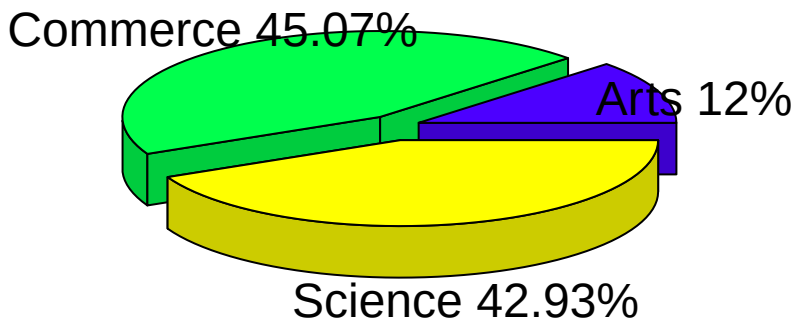


Analysis of XYZ:

```
XYZ
```

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

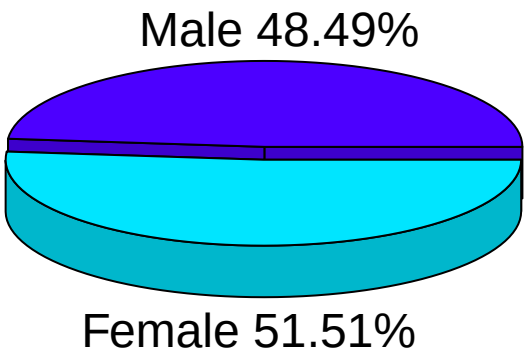
```
labelsa = XYZ$Stream
pcta = round(XYZ$total/sum(XYZ$total)*100, digits = 2)
labelsa = paste(labelsa, pcta)
labelsa = paste(labelsa, "%", sep = "")
pie3D(XYZ$total, labels = labelsa, explode = 0.1, main = "Proportion Of Total Student As Per Stream", c
```




```
xyz_pm = round((sum(XYZ$male)/sum(XYZ$total)*100), digits = 2)
xyz_pf = 100 - xyz_pm
pa = c(xyz_pm, xyz_pf)
labels1a = paste(c("Male","Female"), pa)
```

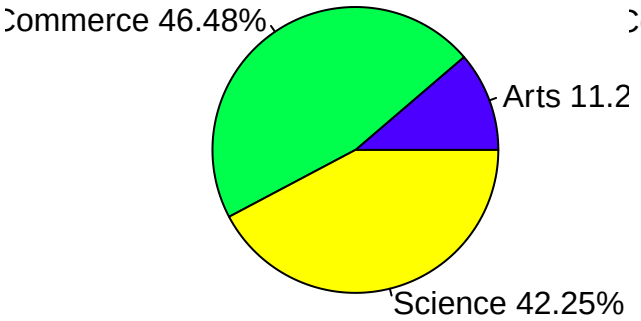
```
labels1a = paste(labels1a, "%", sep = "")
pie3D(pa, labels = labels1a, col = topo.colors(length(labels1)), explode = 0.05, main = "Proportion Of
```

Proportion Of Males & Females In College ABC

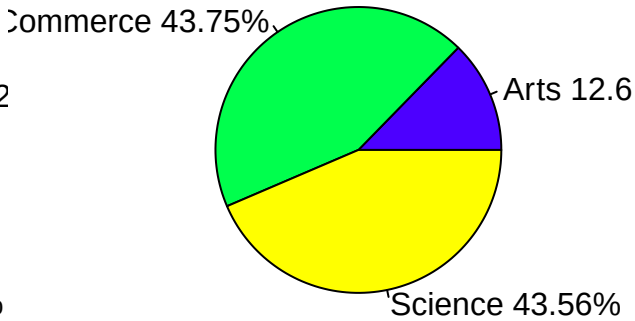



```
par(mfrow=c(1,2))

pctmalesxyz = round(XYZ$male/sum(XYZ$male)*100, digits = 2)
labels3a = XYZ$Stream
labels3a = paste(labels3a, pctmalesxyz)
labels3a = paste(labels3a, "%", sep = "")
pie(pctmalesxyz, labels = labels3a, main = "Proportion Of Males", col = topo.colors(length(labels3a)))

pctfemalesxyz = round(XYZ$female/sum(XYZ$female)*100, digits = 2)
labels3a = XYZ$Stream
labels3a = paste(labels3a, pctfemalesxyz)
labels3a = paste(labels3a, "%", sep = "")
pie(pctfemalesxyz, labels = labels3a, main = "Proportion Of Females", col = topo.colors(length(labels3)
```

Proportion Of Males



Proportion Of Females



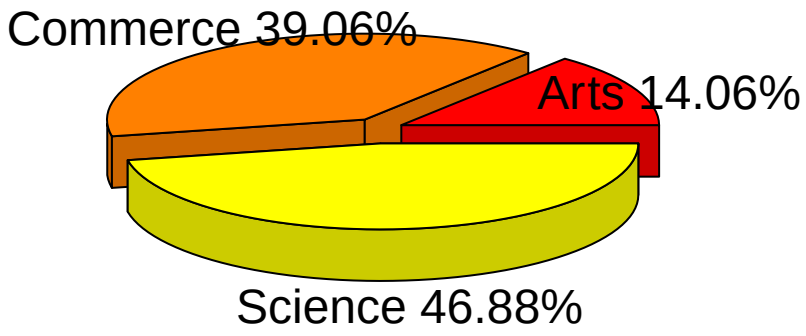
Analysis of PQR:

```
PQR
```

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

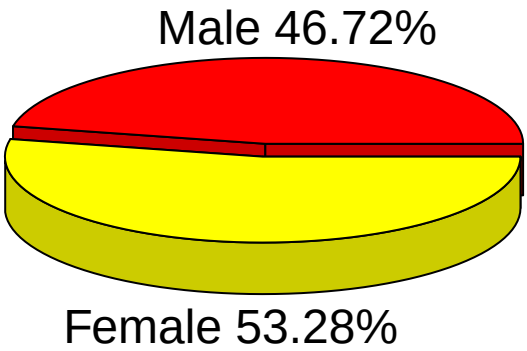
```
labelsb = PQR$Stream
pctb = round(PQR$total/sum(PQR$total)*100, digits = 2)
labelsb = paste(labelsb, pctb)
labelsb = paste(labelsb, "%", sep = "")
pie3D(PQR$total, labels = labelsb, explode = 0.1, main = "Proportion Of Total Student As Per Stream", c
```

Proportion Of Total Student As Per Stream

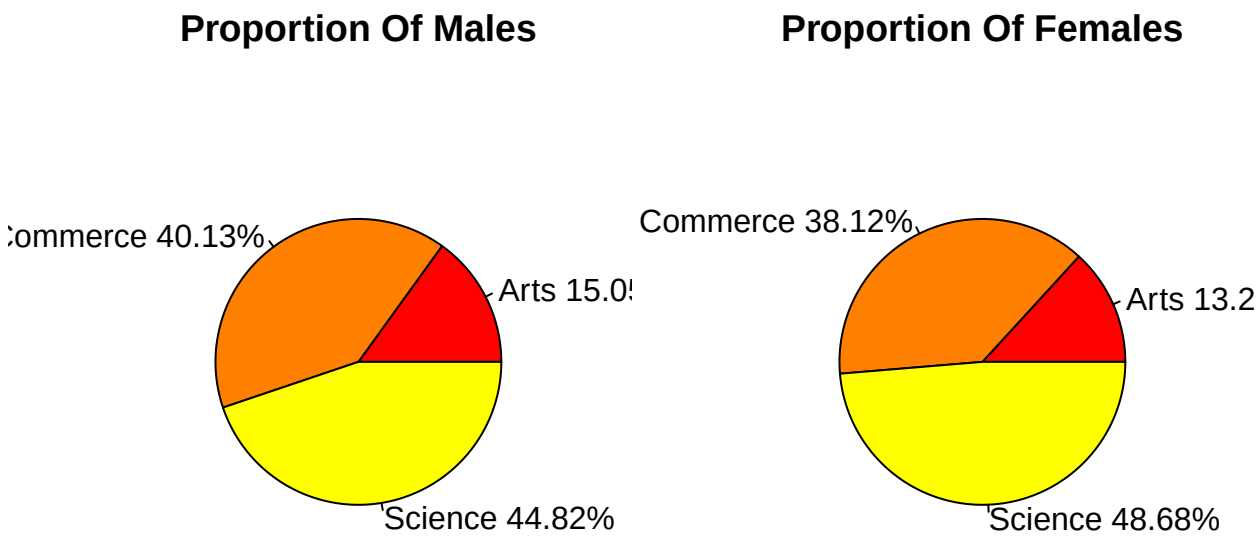



```
pqr_pm = round((sum(PQR$male)/sum(PQR$total)*100), digits = 2)
pqr_pf = 100 - pqr_pm
pb = c(pqr_pm, pqr_pf)
labels1b = paste(c("Male","Female"), pb)
labels1b = paste(labels1b, "%", sep = "")
pie3D(pb, labels = labels1b, col = heat.colors(length(labels1)), explode = 0.05, main = "Proportion Of
```

Proportion Of Males & Females In College ABC



```
par(mfrow=c(1,2))

pctmalespqr = round(PQR$male/sum(PQR$male)*100, digits = 2)
labels3b = PQR$Stream
labels3b = paste(labels3b, pctmalespqr)
labels3b = paste(labels3b, "%", sep = "")
pie(pctmalespqr, labels = labels3b, main = "Proportion Of Males", col = heat.colors(length(labels3)))

pctfemalespqr = round(PQR$female/sum(PQR$female)*100, digits = 2)
labels3b = PQR$Stream
labels3b = paste(labels3b, pctfemalespqr)
labels3b = paste(labels3b, "%", sep = "")
pie(pctfemalespqr, labels = labels3b, main = "Proportion Of Females", col = heat.colors(length(labels3b)
```



Comments on overall analysis:

Females dominate the student account in all three colleges and the difference between their proportions averages at 3.77% across all the colleges. The most popular stream is science with an average of 47.81% students wanting to puruse this stream. There is also a higher proportion of males in the commerce across all colleges, whilst science and arts is dominated by the females. The most popular college as percentage of all streams for science is ABC while by number of students its XYZ (but only by 10 students). Whereas XYZ dominates the commerce stream in both percentage of all streams and total students admitted.

Contingency table created as dentab:

```
Total = datacol%>%
  group_by(college, Stream)%>%
  summarise(total = total)
Arts = Total[1:3,3]
Commerce = Total[4:6,3]
Science = Total[7:9,3]
datacol1 = data_frame(Arts, Commerce, Science)
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

```
dentab = as.table(as.matrix(datacol1))
colnames(dentab) = c("ABC","PQR","XYZ")
dentab = t(dentab)
```

```r
colnames(dentab) = c("Arts","Commerce","Science")
dentab
```

```
##     Arts Commerce Science
## ABC  120      252     430
## PQR   90      250     300
## XYZ  123      462     440
```

Balloon plot of College and Stream:

```r
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

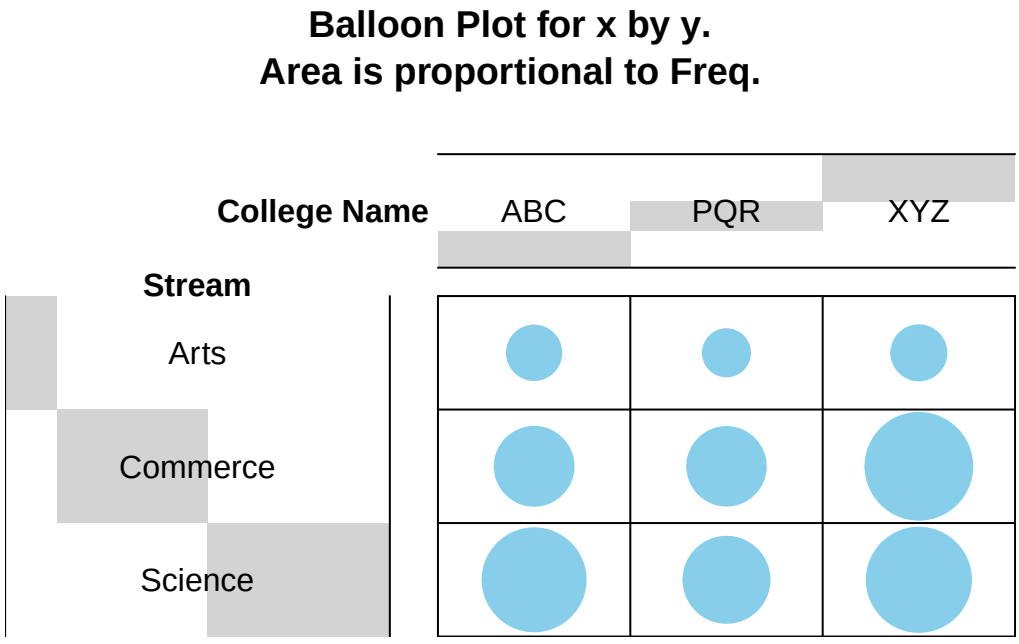
```
## The following object is masked from 'package:plotrix':
##
##     plotCI
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
dentab
```

```
##     Arts Commerce Science
## ABC  120      252     430
## PQR   90      250     300
## XYZ  123      462     440
```

```r
balloonplot(dentab, ylab = "Stream", xlab = "College Name", label = FALSE, show.margins = FALSE)
```



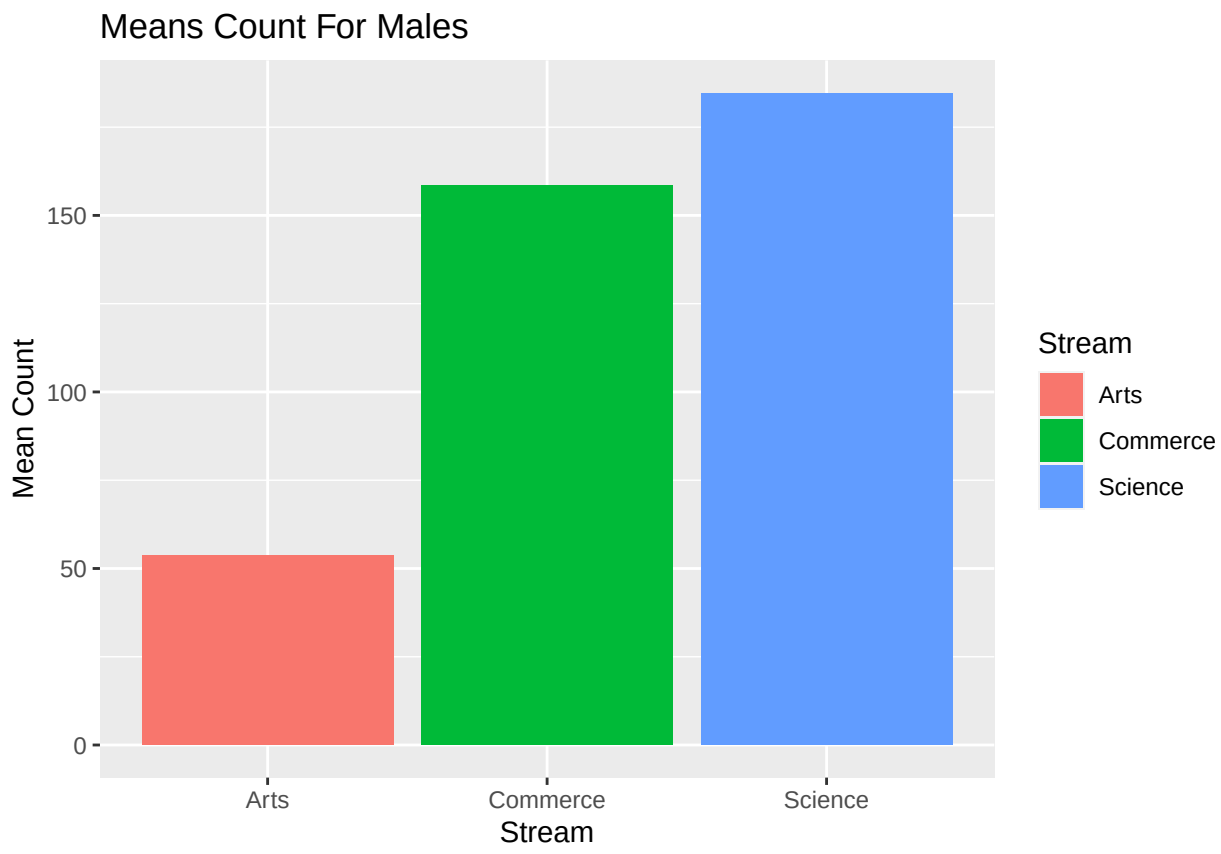
The ballon plots aids in easily identifying the magnitude of each stream against each of the colleges. We can clearly correspond the visual dynamics the cotengency table. The largest college by student population is XYZ, and its most popular stream is commerce with 462 students which differs from the overall most popular stream, science with overall admissions count of 1,170. The least sought after course is arts from college PQR with only 90 students admitted.

Means As Per Stream:

```
mean_male = datacol%>%
  group_by(Stream)%>%
  summarise(mean_male = mean(male))

mean_female = datacol%>%
  group_by(Stream)%>%
  summarise(mean_female = mean(female))

mean_male
```

```
## # A tibble: 3 x 2
##   Stream   mean_male
##   <fct>        <dbl>
## 1 Arts          53.7
## 2 Commerce     158.
## 3 Science      185.
```

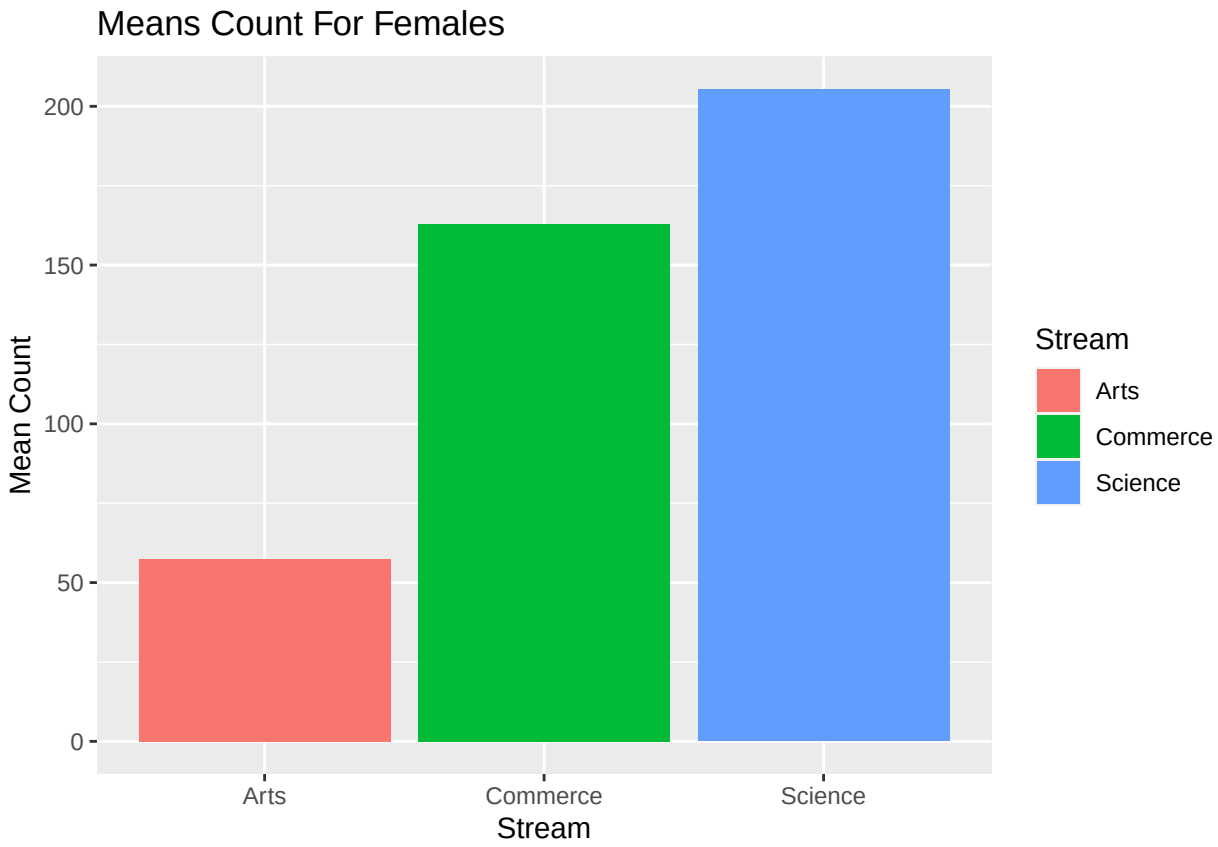
```
ggplot(mean_male, aes(factor(Stream), mean_male, fill = Stream))+ geom_col() + ggtitle("Means Count For
```




```
mean_female
```

```
## # A tibble: 3 x 2
##   Stream   mean_female
##   <fct>          <dbl>
## 1 Arts            57.3
## 2 Commerce       163
## 3 Science        205.
```

```
ggplot(mean_female, aes(factor(Stream), mean_female, fill = Stream))+ geom_col() + ggtitle("Means Count
```




```
diffmean = round(mean_female$mean_female - mean_male$mean_male)
diffmean
```

```
## [1]  4  5 21
```

We can see that there is an average of 54 males in arts, where as there are 158 male in commerce and 184 male students in science across all colleges. corresponding to the same sequence and average of 57 females in Arts, 163 females in commerce and 205 females in science. We can deduce that on average there are more females than males in all three streams across all three colleges. The smallest difference is in arts, where there are an average of 4 more girls than boys and the highest being in science where on an average, there are 21 more females than males.

Testing for association via chi-square test:

Null hypothesis (H0): the row and the column variables of the contingency table are independent.* **Alternative hypothesis (H1):** row and column variables are dependent

```
matden2 = matrix(as.integer(dentab), 3)
matden2
```

```
##      [,1] [,2] [,3]
## [1,]  120  252  430
## [2,]   90  250  300
## [3,]  123  462  440
```

```
chisq.test(matden2)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  matden2
## X-squared = 35.485, df = 4, p-value = 3.693e-07
```

Since the p-value = 0, the rows and columns of the contingency are satistically significantly association.