Exploratory Data Analysis

Srijit Mukherjee

23 June 2020

Types of Data Variables in R

- Quantitative
  - Continuous
  - Discrete
- Qualitative
  - Nominal
  - Ordinal

Datasets in R

If you type the following command, you will get to know all the data sets in R.

```
data()
library(DAAG)
```

```
Warning: package 'DAAG' was built under R version 3.6.3

Loading required package: lattice

Warning: package 'lattice' was built under R version 3.6.3
```

Understanding Data in

Let's see the top of the data matrix.

```
data = ais
head(data, n = 3)
```

```
   rcc wcc   hc   hg ferr   bmi   ssf pcBfat   lbm    ht   wt sex  sport
1 3.96 7.5 37.5 12.3   60 20.56 109.1  19.75 63.32 195.9 78.9   f B_Ball
2 4.41 8.3 38.2 12.7   68 20.67 102.8  21.30 58.55 189.7 74.4   f B_Ball
3 4.14 5.0 36.4 11.6   21 21.86 104.6  19.88 55.36 177.8 69.1   f B_Ball
```

Handling Data in R

Let's see the bottom of the data matrix.

```
tail(data, n = 3)
```

```
     rcc wcc   hc   hg ferr   bmi  ssf pcBfat lbm    ht   wt sex  sport
200 5.03 6.4 42.7 14.3  122 22.01 47.6   8.51  68 183.1 73.8   m Tennis
201 4.97 8.8 43.0 14.9  233 22.34 60.4  11.50  63 178.4 71.1   m Tennis
202 5.38 6.3 46.0 15.7   32 21.07 34.9   6.26  72 190.8 76.7   m Tennis
```

Handling Data in R

Let's see a concise summary of the data matrix.

```
str(data)
```

```
'data.frame':   202 obs. of  13 variables:
 $ rcc   : num  3.96 4.41 4.14 4.11 4.45 4.1 4.31 4.42 4.3 4.51 ...
 $ wcc   : num  7.5 8.3 5 5.3 6.8 4.4 5.3 5.7 8.9 4.4 ...
 $ hc    : num  37.5 38.2 36.4 37.3 41.5 37.4 39.6 39.9 41.1 41.6 ...
 $ hg    : num  12.3 12.7 11.6 12.6 14 12.5 12.8 13.2 13.5 12.7 ...
 $ ferr  : num  60 68 21 69 29 42 73 44 41 44 ...
 $ bmi   : num  20.6 20.7 21.9 21.9 19 ...
 $ ssf   : num  109.1 102.8 104.6 126.4 80.3 ...
 $ pcBfat: num  19.8 21.3 19.9 23.7 17.6 ...
 $ lbm   : num  63.3 58.5 55.4 57.2 53.2 ...
 $ ht    : num  196 190 178 185 185 ...
 $ wt    : num  78.9 74.4 69.1 74.9 64.6 63.7 75.2 62.3 66.5 62.9 ...
 $ sex   : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
 $ sport : Factor w/ 10 levels "B_Ball","Field",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Handling Data in R

```
class(data)
```

```
[1] "data.frame"
```

```
class(data$mpg)
```

```
[1] "NULL"
```

```
dim(data)
```

```
[1] 202  13
```

```
names(data)
```

```
 [1] "rcc"    "wcc"    "hc"     "hg"     "ferr"   "bmi"    "ssf"    "pcBfat"
 [9] "lbm"    "ht"     "wt"     "sex"    "sport" 
```

Handling Data in R

Exploratory Data Analysis

```
summary(data)
```

```
      rcc             wcc               hc              hg       
 Min.   :3.800   Min.   : 3.300   Min.   :35.90   Min.   :11.60  
 1st Qu.:4.372   1st Qu.: 5.900   1st Qu.:40.60   1st Qu.:13.50  
 Median :4.755   Median : 6.850   Median :43.50   Median :14.70  
 Mean   :4.719   Mean   : 7.109   Mean   :43.09   Mean   :14.57  
 3rd Qu.:5.030   3rd Qu.: 8.275   3rd Qu.:45.58   3rd Qu.:15.57  
 Max.   :6.720   Max.   :14.300   Max.   :59.70   Max.   :19.20  
                                                                 
      ferr             bmi             ssf             pcBfat      
 Min.   :  8.00   Min.   :16.75   Min.   : 28.00   Min.   : 5.630  
 1st Qu.: 41.25   1st Qu.:21.08   1st Qu.: 43.85   1st Qu.: 8.545  
 Median : 65.50   Median :22.72   Median : 58.60   Median :11.650  
 Mean   : 76.88   Mean   :22.96   Mean   : 69.02   Mean   :13.507  
 3rd Qu.: 97.00   3rd Qu.:24.46   3rd Qu.: 90.35   3rd Qu.:18.080
 Max.   :234.00   Max.   :34.42   Max.   :200.80   Max.   :35.520

      lbm              ht              wt          sex         sport
 Min.   : 34.36   Min.   :148.9   Min.   : 37.80   f:100   Row    :37
 1st Qu.: 54.67   1st Qu.:174.0   1st Qu.: 66.53   m:102   T_400m :29
 Median : 63.03   Median :179.7   Median : 74.40           B_Ball :25
 Mean   : 64.87   Mean   :180.1   Mean   : 75.01           Netball:23
 3rd Qu.: 74.75   3rd Qu.:186.2   3rd Qu.: 84.12           Swim   :22
 Max.   :106.00   Max.   :209.4   Max.   :123.20           Field  :19
                                                           (Other):47
```

Univariate Quantitative Data Analysis

Exploratory Data Analysis

```r
mean(data$hg)
```

```
[1] 14.56634
```

```r
var(data$hg)
```

```
[1] 1.856274
```

```r
sd(data$hg)
```

```
[1] 1.362451
```

```r
min(data$hg)
```

```
[1] 11.6
```

Univariate Quantitative Data Analysis

Exploratory Data Analysis

```r
max(data$hg)
```

```
[1] 19.2
```

```r
median(data$hg)
```

```
[1] 14.7
```

```r
quantile(data$hg)
```

```
    0%    25%    50%    75%   100%
11.600 13.500 14.700 15.575 19.200
```

```r
range(data$hg)
```

```
[1] 11.6 19.2
```

Univariate Quantitative Data Analysis

Summary of the Data

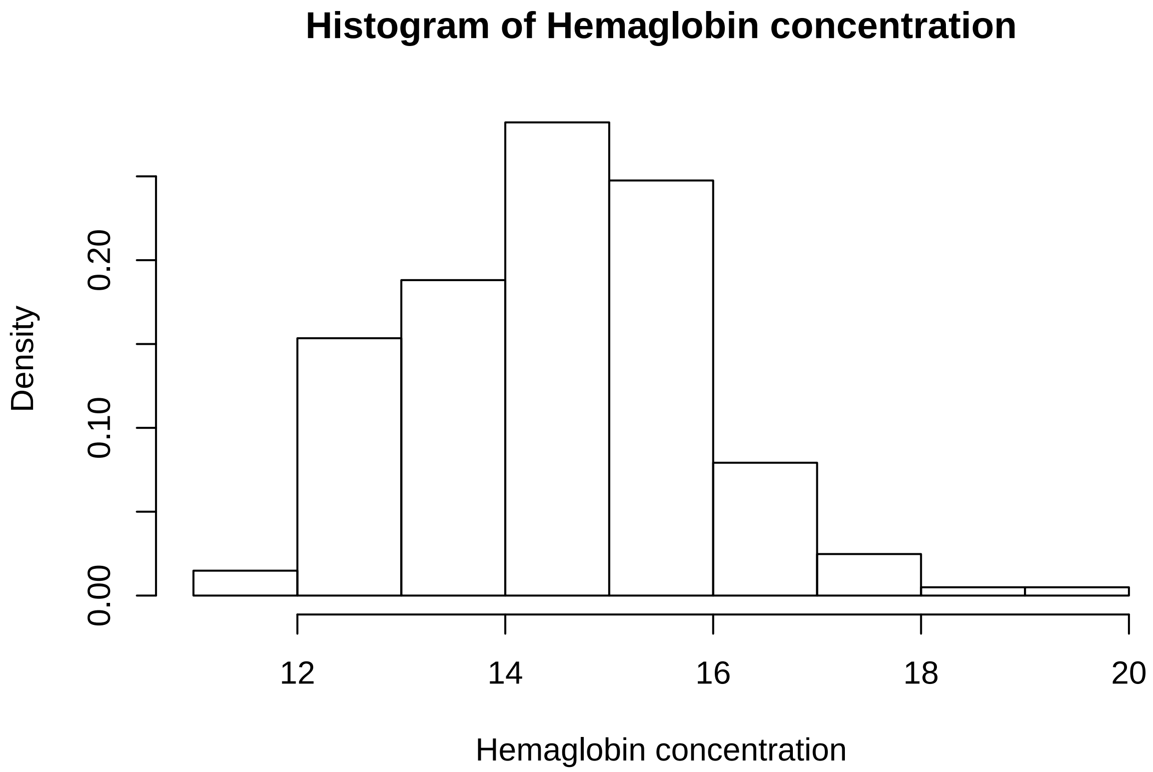
```r
summary(data$hg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.60   13.50   14.70   14.57   15.57   19.20
```

Univariate Quantitative Data Analysis

Histogram

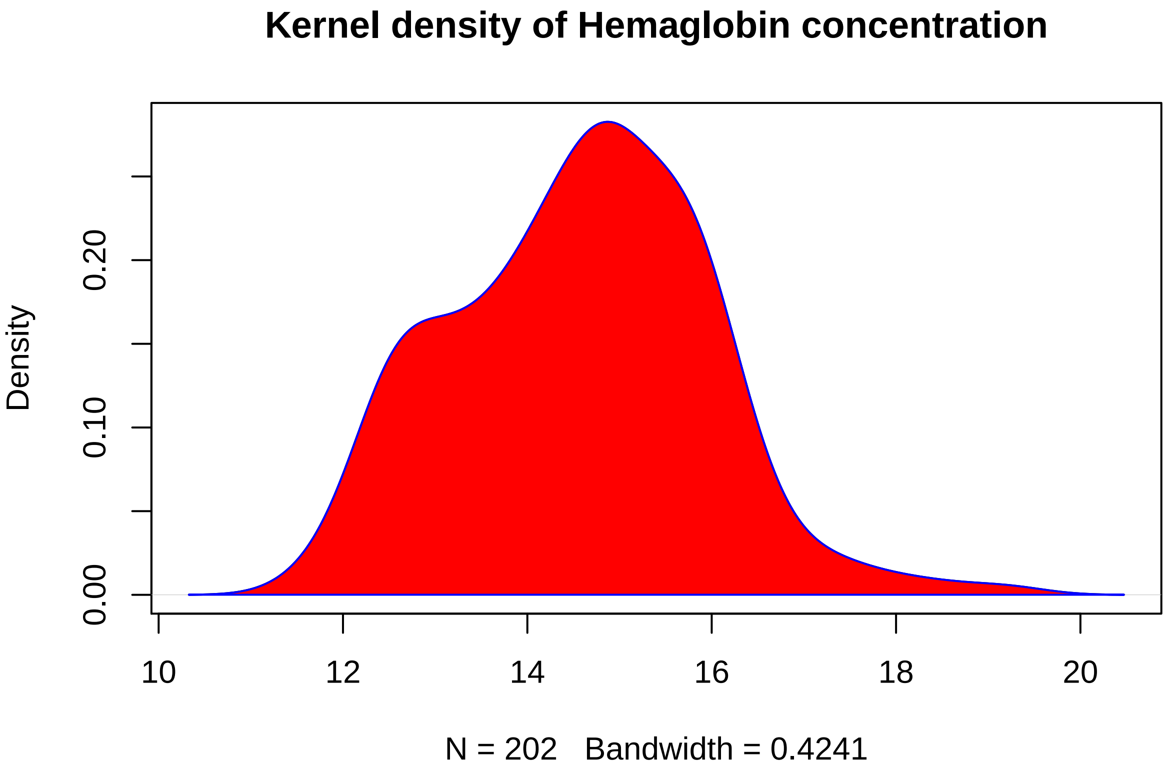
```
hist(data$hg, xlab = "Hemaglobin concentration", probability = TRUE, , main = "Histogram of Hemaglobin
```



Univariate Quantitative Data Analysis

Kernel Density

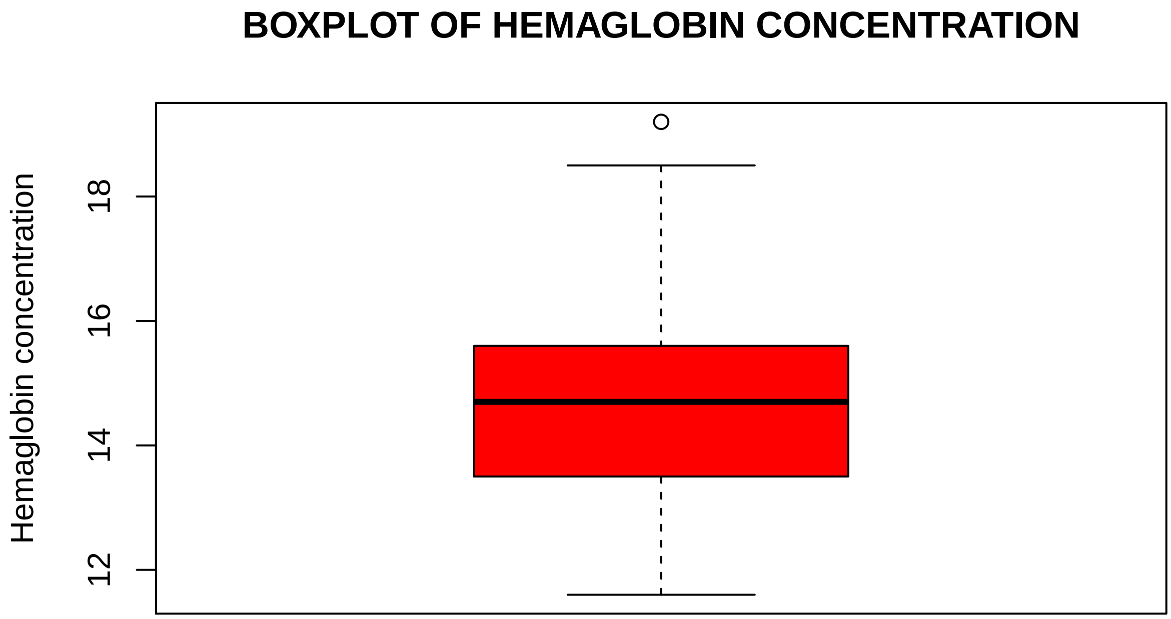
```
d <- density(data$hg)
plot(d, main = "Kernel density of Hemaglobin concentration")
polygon(d, col = "red", border = "blue")
```



Univariate Quantitative Data Analysis

Box Plot

```r
boxplot(data$hg,
        main = toupper("Boxplot of Hemaglobin concentration"),
        ylab = "Hemaglobin concentration",
        col = "red")
```



Univariate Qualitative Data Analysis

Frequency Distribution Table

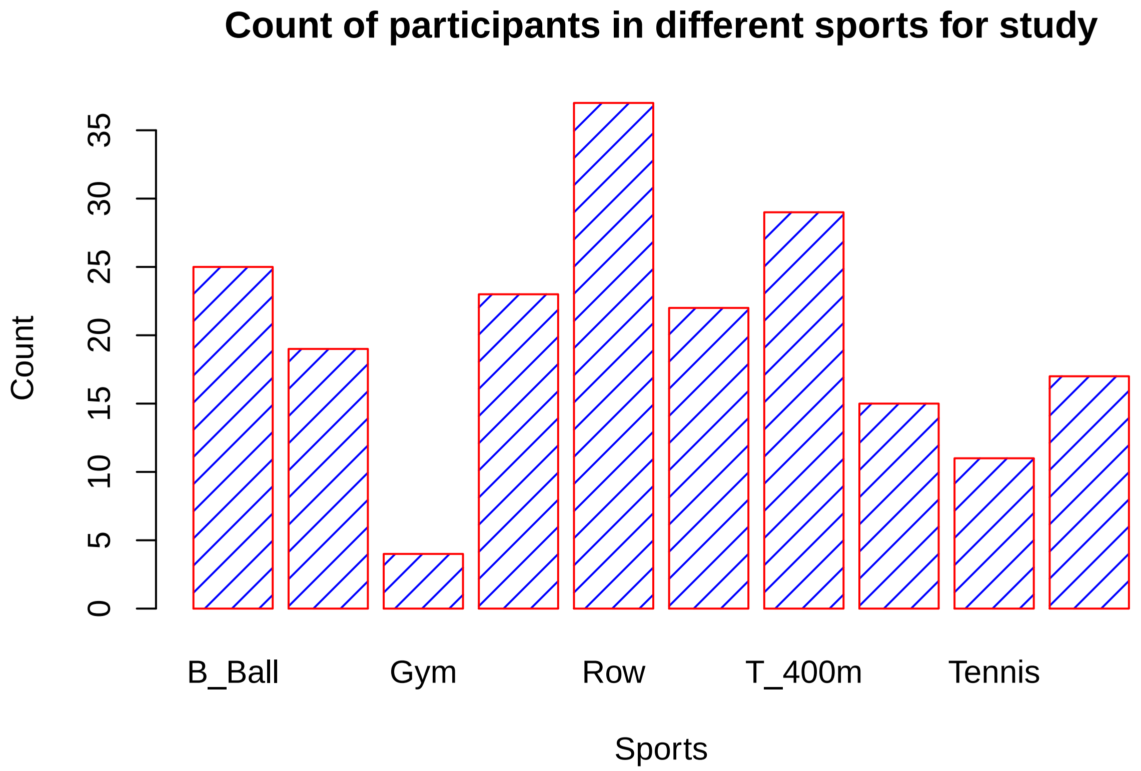
```
table(data$sport)
```

```
 B_Ball    Field      Gym Netball      Row     Swim  T_400m T_Sprnt  Tennis  W_Polo
     25       19        4      23       37       22      29      15      11      17
```

Univariate Qualitative Data Analysis

Vertical Bar Plot

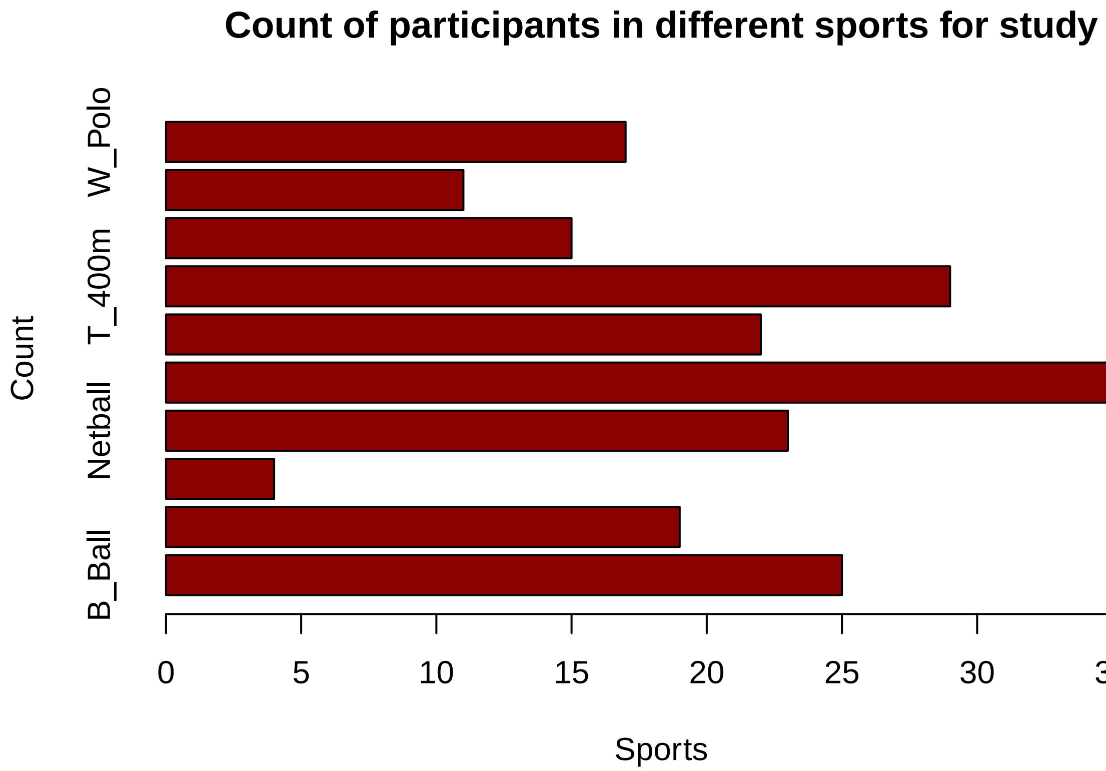
```
barplot(table(data$sport), main="Count of participants in different sports for study", xlab="Sports",yla
```



Univariate Qualitative Data Analysis

Horizontal Bar Plot

```
barplot(table(data$sport), main="Count of participants in different sports for study", xlab="Sports",yla
```

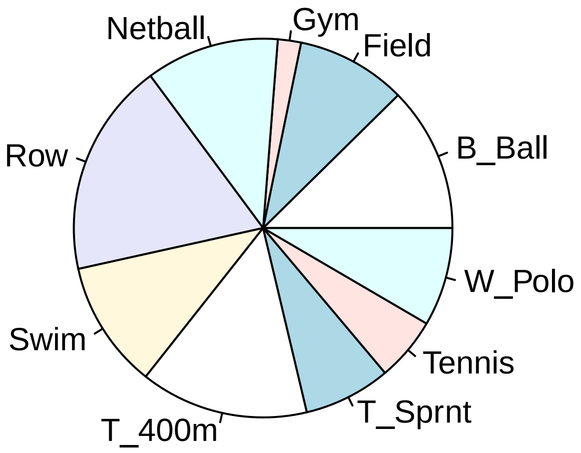


Univariate Qualitative Data Analysis

How to do probability bar plot?

Pie Chart

```
pie(table(data$sport), labels = levels(data$sport))
```



Multivariate Exploratory Data Analysis

Categorical vs Categorical

Contigency Table

```
sex_vs_sport = data[,12:13]
table(sex_vs_sport)
```

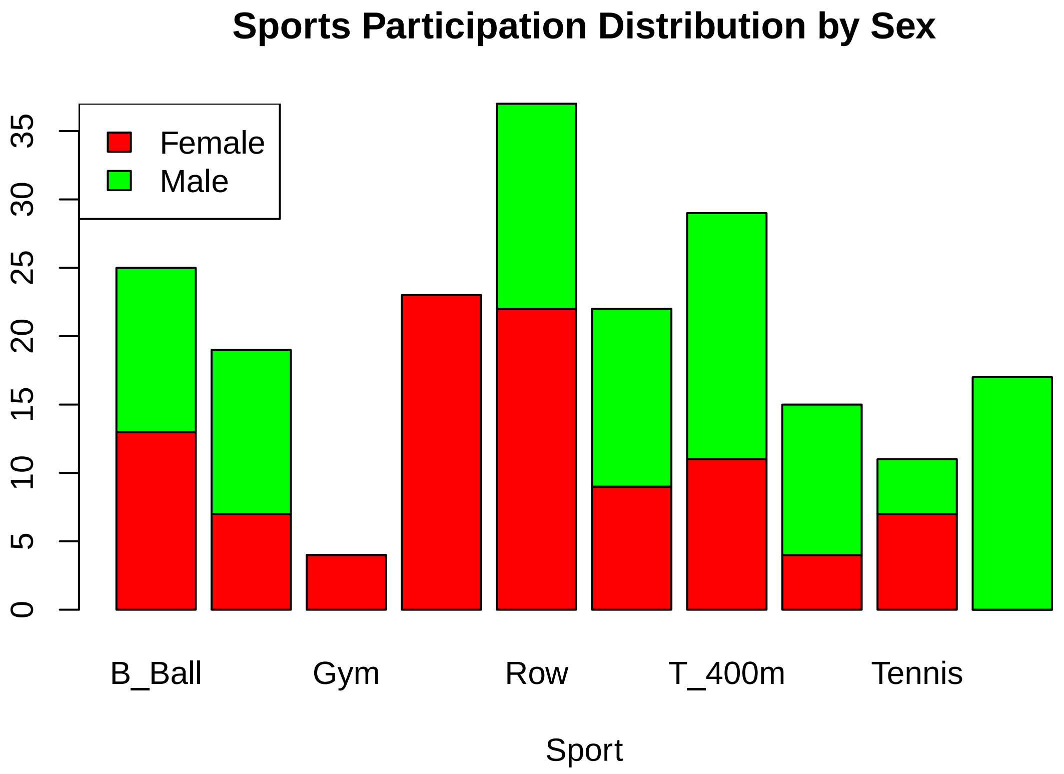
```
   sport
sex B_Ball Field Gym Netball Row Swim T_400m T_Sprnt Tennis W_Polo
  f     13     7   4      23  22    9     11       4      7      0
  m     12    12   0       0  15   13     18      11      4     17
```

```
#xtabs(~ sex + sport, sex_vs_sport)
```

Categorical vs Categorical

Bar Plots Vertical Comparison

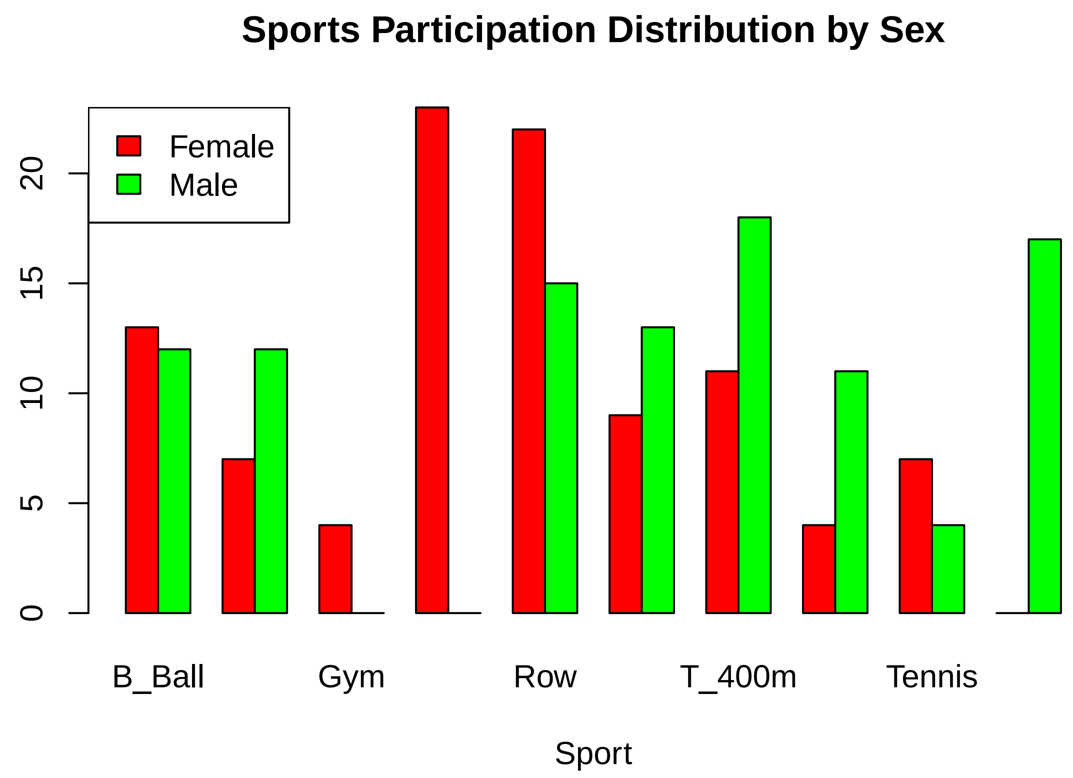
```
barplot(table(sex_vs_sport),
        main = "Sports Participation Distribution by Sex",
        xlab = "Sport",
        col = c("red","green")

)
legend("topleft",
       c("Female","Male"),
       fill = c("red","green")
)
```



Categorical vs Categorical

Bar Plot Beside Comparison

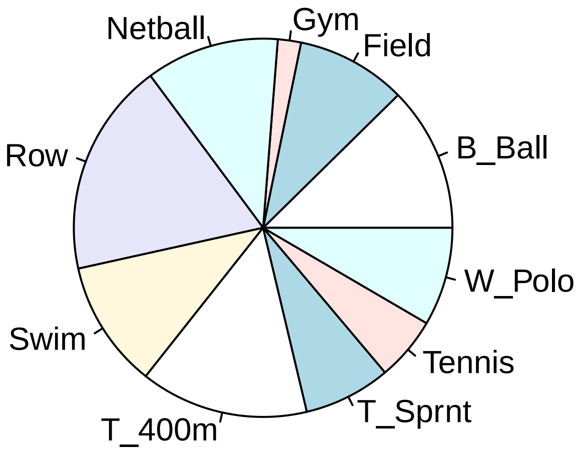
```
barplot(table(sex_vs_sport),
        main = "Sports Participation Distribution by Sex",
        xlab = "Sport",
        col = c("red","green"),
        beside =  TRUE
)
legend("topleft",
       c("Female","Male"),
       fill = c("red","green")
)
```



Categorical vs Categorical

Pie Chart

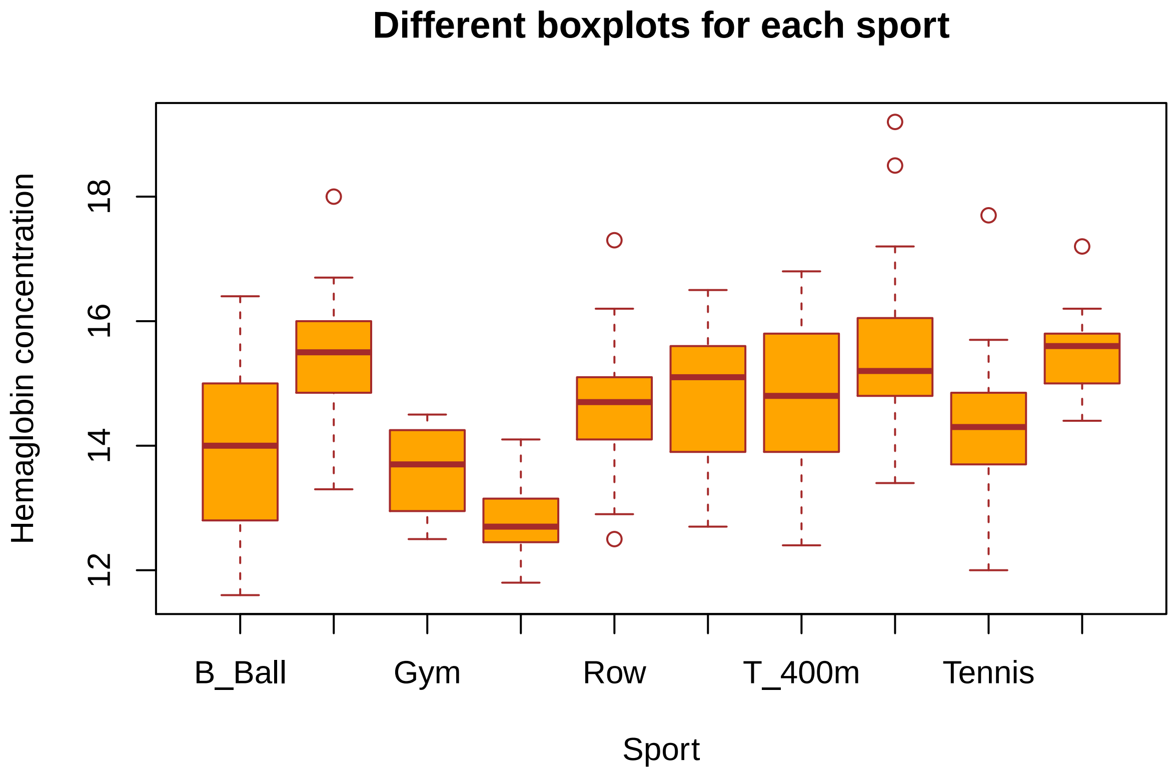
```
pie(table(data$sport), labels = levels(data$sport))
```



Continuous vs Categorical

Comparison of Box Plot

```
boxplot(hg~sport,
        data=data,
        main="Different boxplots for each sport",
        xlab="Sport",
        ylab="Hemaglobin concentration",
        col="orange",
        border="brown"
)
```



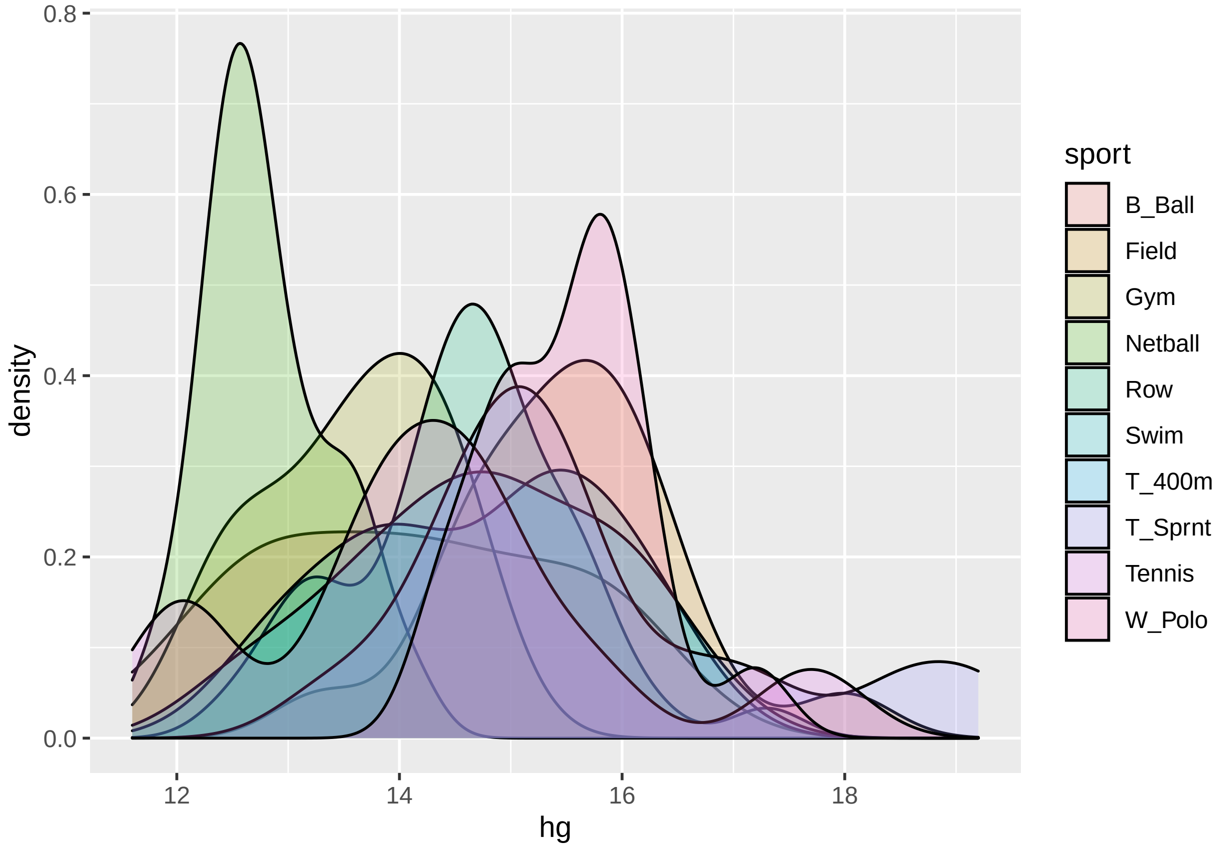
Continuous vs Categorical

Comparison of Histogram

```
library(ggplot2)
```

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

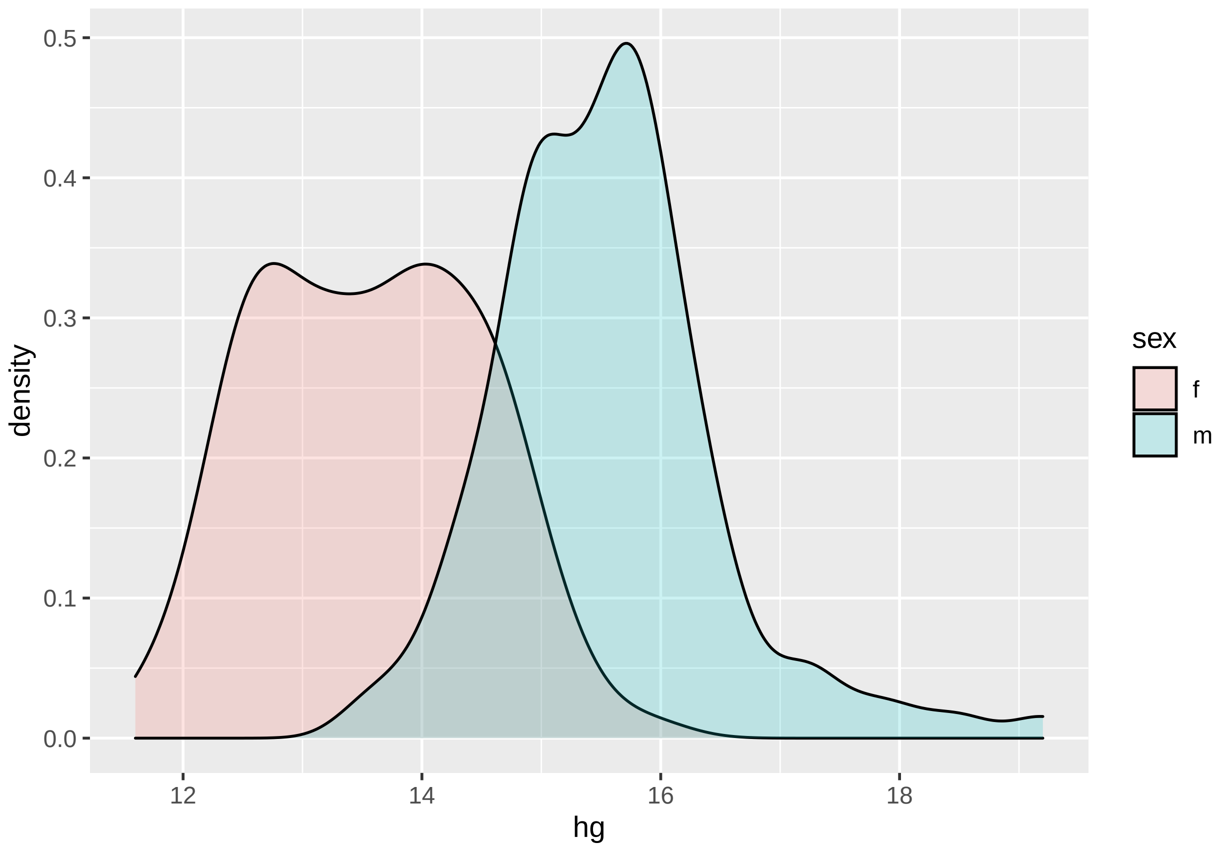
```
hg_vs_sport = data[,c(4,13)]
ggplot(hg_vs_sport, aes(hg, fill = sport)) + geom_density(alpha = 0.2)
```



Continuous vs Categorical

Comparison of Histogram

```
library(ggplot2)

hg_vs_sex = data[,c(4,12)]
ggplot(hg_vs_sex, aes(hg, fill = sex)) + geom_density(alpha = 0.2)
```



Continuous vs Categorical

Comparing Summary Data

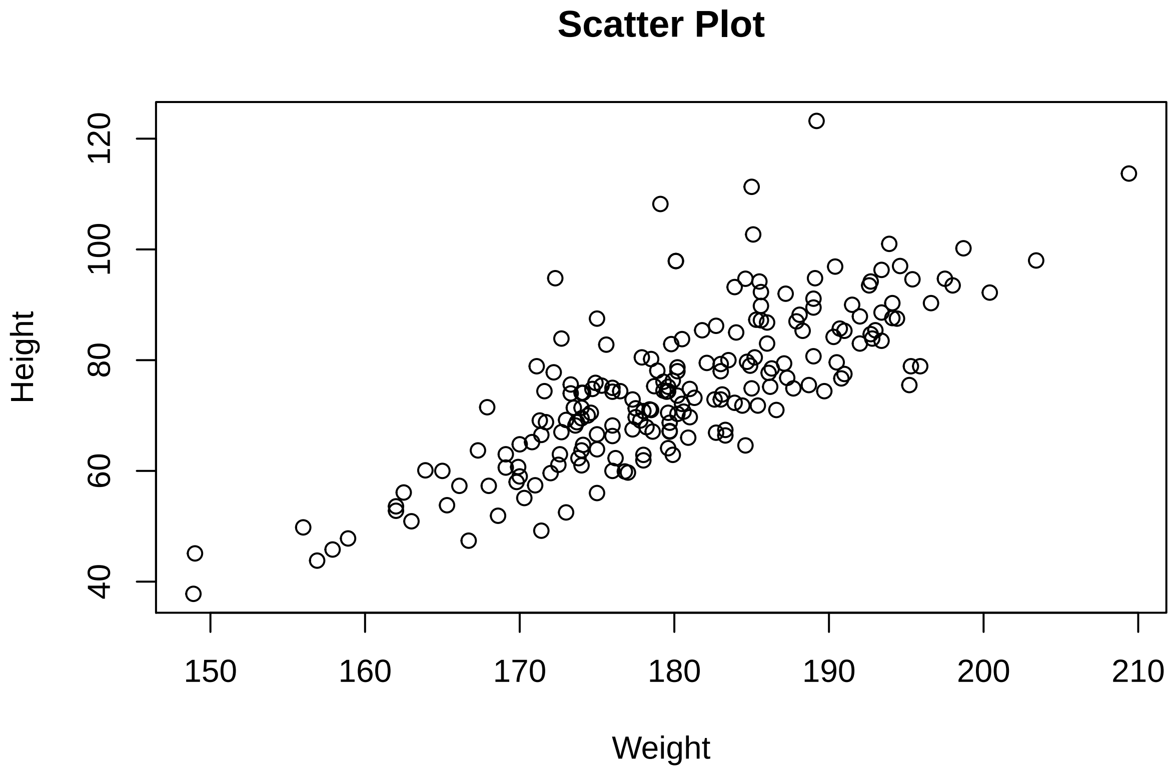
```
by(hg_vs_sex, hg_vs_sex$sex, summary)
```

```
hg_vs_sex$sex: f
       hg         sex    
 Min.   :11.60   f:100  
 1st Qu.:12.70   m:  0  
 Median :13.50          
 Mean   :13.56          
 3rd Qu.:14.30          
 Max.   :15.90          
------------------------------------------------------------ 
hg_vs_sex$sex: m
       hg         sex    
 Min.   :13.50   f:  0  
 1st Qu.:14.93   m:102  
 Median :15.50          
 Mean   :15.55          
 3rd Qu.:15.90          
 Max.   :19.20          
```

Continuous vs Continuous

Plot

```
plot(data$wt ~ data$ht , data,
            xlab="Weight", ylab="Height",
            main="Scatter Plot")
```



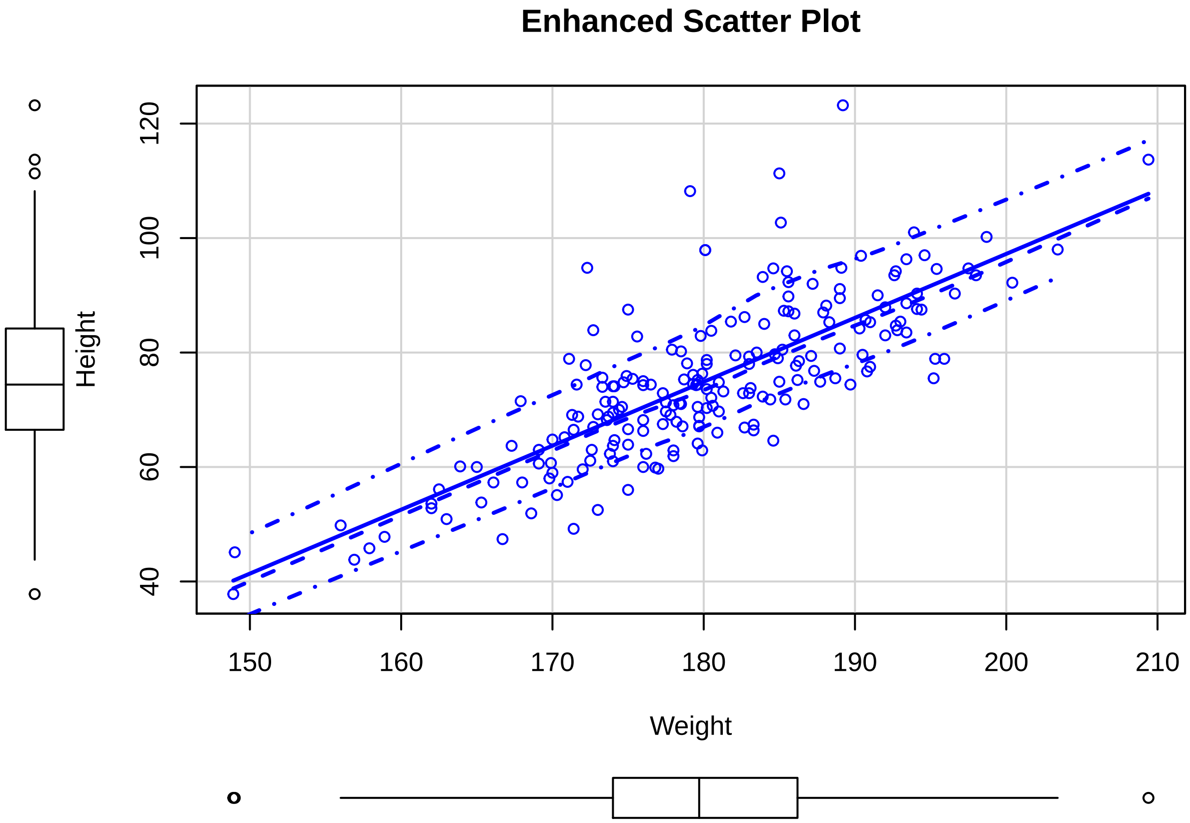
Continuous vs Continuous

Scatter Plot

```
library(car)
```

```
Warning: package 'car' was built under R version 3.6.3

Loading required package: carData

Warning: package 'carData' was built under R version 3.6.3


Attaching package: 'car'

The following object is masked from 'package:DAAG':

    vif
```

```
scatterplot(data$wt ~ data$ht , data,
            xlab="Weight", ylab="Height",
            main="Enhanced Scatter Plot")
```


Enhanced Scatter Plot
Height
Weight
40
60
80
100
120
150
160
170
180
190
200
210