Luis Alberto Alvarez Zavaleta

Exercise 1: Inspection of data

1. **Try the following commands.**

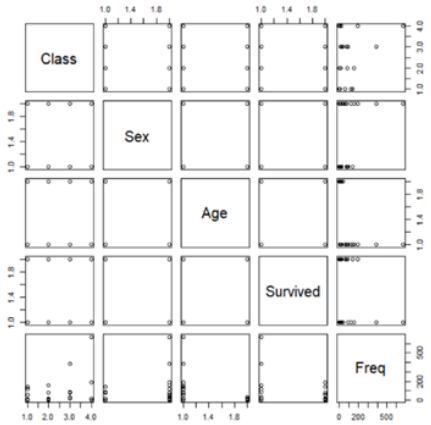
```
> head(titanic)
  X Class    Sex   Age Survived Freq
1 1   1st   Male Child       No    0
2 2   2nd   Male Child       No    0
3 3   3rd   Male Child       No   35
4 4  Crew   Male Child       No    0
5 5   1st Female Child       No    0
6 6   2nd Female Child       No    0
> summary(titanic)
       X            Class               Sex                Age
 Min.   : 1.00   Length:32          Length:32          Length:32
 1st Qu.: 8.75   Class :character   Class :character   Class :character
 Median :16.50   Mode  :character   Mode  :character   Mode  :character
 Mean   :16.50
 3rd Qu.:24.25
 Max.   :32.00
   Survived             Freq
 Length:32          Min.   :  0.00
 Class :character   1st Qu.:  0.75
 Mode  :character   Median : 13.50
                    Mean   : 68.78
                    3rd Qu.: 77.00
                    Max.   :670.00
> plot(titanic)
```

2. **Which variables are quantitative and which variables are categorical? How can we know it?**

We can know which types the variables are from the function summary, those who are Quantitative will display Min, Median, Mean. In contrast those that are Categorical will only display the length “32”and the type “character” of the variable.

The variables that are Quantitative are: X and Freq

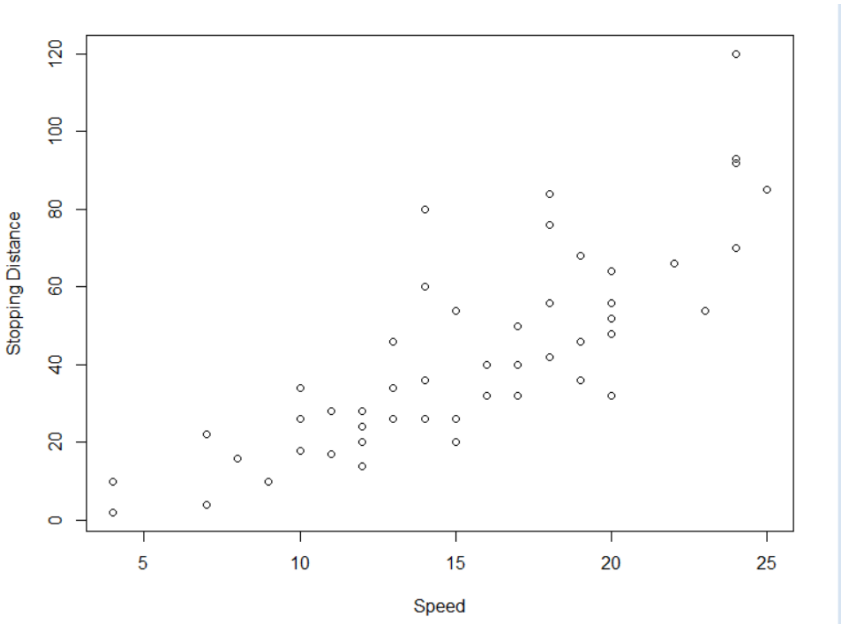
The variables that are Categorical: Class, Sex, Sage, Survived



Exercise 2: Working with basic graphics.

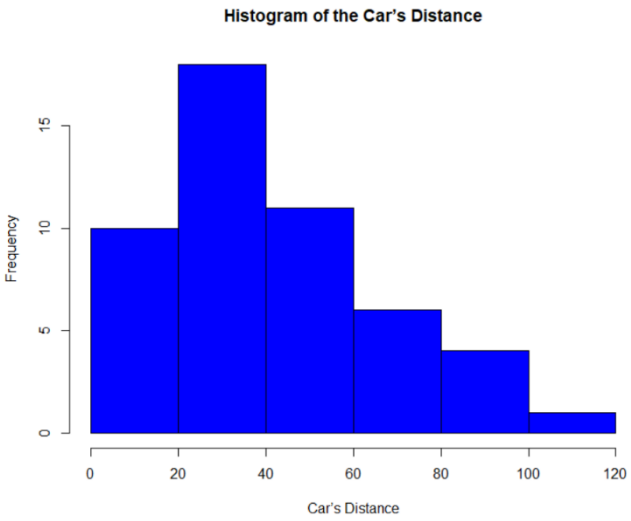
3. **Make a plot of the distance field in terms of the speed field (use the $ syntax).**

```
plot(cars$dist,cars$speed, ylab = "Stopping Distance", xlab = "Speed")
```



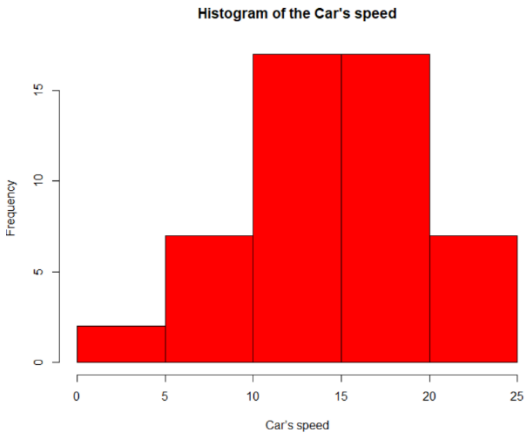
4. **Make a histogram of the distance variable.**

```
hist(cars$dist,xlab="Car's Distance",main="Histogram of the Car's Distance",col="blue")
```



5. **Make a histogram of the speed variable.**

```
hist(cars$speed,xlab="Car’s speed",main="Histogram of the Car's speed",col="red")
```



6. **Modify the previous plots to show the name of the variables (“speed” or “distance”) as the title of the axis.**

Already upload the images of the previous histograms and plot using the xlab and ylab parameter to change the name of the x axis variables to speed or distance.

Exercise 3: Transformations of variables and datasets

1. **Construct a new data frame with the above data.**

```
> cars2<- data.frame(speed=c(21,34),dist=c(47,87))
> cars2
  speed dist
1    21   47
2    34   87
```

2. **Add the constructed data frame to the cars data frame.**

cars <- rbind(cars, cars2)

3. **Sort the data in the resulting dataset by column speed(ascending)**

```
> cars <- cars[ order(cars$speed),]
> cars
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
```

You can use cars <- cars[order(-cars$speed),] to order it by descending order

Exercise 4: Data manipulation.

1. **Extract the first 2 rows of the data frame and print them to the console. What does the output look like?**

```
> rows<- head(airq,2)
> rows
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
```

2. **How many observations (i.e., rows) there are in this data frame?**

```
> nrow(airq)
[1] 153
```

3. **What is the value of Ozone in the 40th row?**

```
> airq$Ozone[40]
[1] 71
```

4. **How many missing values there are in the Ozone column of this data frame?**

```
> sum(is.na(airq$Ozone))
[1] 37
```

5. **What is the mean of the Ozone column in this dataset? Exclude missing values (coded as NA) from this calculation.**

```
> mean(airq$Ozone,na.rm=TRUE)
[1] 42.12931
```

6. **Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?**

```
> subairq<-subset(airq,airq$Ozone>31 & airq$Temp>90)
> subairq
    Ozone Solar.R Wind Temp Month Day
69     97     267  6.3   92     7   8
70     97     272  5.7   92     7   9
120    76     203  9.7   97     8  28
121   118     225  2.3   94     8  29
122    84     237  6.3   96     8  30
123    85     188  6.3   94     8  31
124    96     167  6.9   91     9   1
125    78     197  5.1   92     9   2
126    73     183  2.8   93     9   3
127    91     189  4.6   93     9   4
> mean(subairq$Solar.R)
[1] 212.8
```

Exercise 5

1. **Discretise the Ozone column into five bins ('bin1', 'bin2', ...) of equal width and a sixth bin ('binNA') for NA.**

```
breaks <- seq(min(airq$Ozone, na.rm = TRUE), max(airq$Ozone, na.rm = TRUE), length = 6)

labels <- c('bin1', 'bin2', 'bin3', 'bin4', 'bin5')

airq$Ozone_bin <- cut(airq$Ozone, breaks = breaks, labels = labels, include.lowest = TRUE)

levels(airq$Ozone_bin) <- c(levels(airq$Ozone_bin), 'binNA')

airq$Ozone_bin[is.na(airq$Ozone)] <- 'binNA'
```

```
> airq
    Ozone Solar.R Wind Temp Month Day Ozone_bin
1      41     190  7.4   67     5   1      bin2
2      36     118  8.0   72     5   2      bin2
3      12     149 12.6   74     5   3      bin1
4      18     313 11.5   62     5   4      bin1
5      NA      NA 14.3   56     5   5     binNA
6      28      NA 14.9   66     5   6      bin1
7      23     299  8.6   65     5   7      bin1
8      19      99 13.8   59     5   8      bin1
```

2. **Discretise the Solar column into four bins of equal size and a fifth bin for NA.**

```
breaks <- quantile (airq$Solar.R, probs = seq(0, 1, 0.25), na.rm = TRUE)

labels <- c ('bin1', 'bin2', 'bin3', 'bin4')

airq$Solar_bin <- cut (airq$Solar.R, breaks = breaks, labels = labels, include.lowest = TRUE)

levels(airq$Solar_bin) <- c(levels(airq$Solar_bin), 'binNA')

airq$Solar_bin[is.na (airq$Solar.R)] <- c('binNA')
```

```
> airq
    Ozone Solar.R Wind Temp Month Day Ozone_bin Solar_bin
1      41     190  7.4   67     5   1      bin2      bin2
2      36     118  8.0   72     5   2      bin2      bin2
3      12     149 12.6   74     5   3      bin1      bin2
4      18     313 11.5   62     5   4      bin1      bin4
5      NA      NA 14.3   56     5   5     binNA     binNA
6      28      NA 14.9   66     5   6      bin1     binNA
7      23     299  8.6   65     5   7      bin1      bin4
8      19      99 13.8   59     5   8      bin1      bin1
9       8      19 20.1   61     5   9      bin1      bin1
10     NA     194  8.6   69     5  10     binNA      bin2
11      7      NA  6.9   74     5  11      bin1     binNA
12     16     256  9.7   69     5  12      bin1      bin3
13     11     290  9.2   66     5  13      bin1      bin4
14     14     274 10.9   68     5  14      bin1      bin4
```

3. **Create a new column AbsDay from the columns Month and Day such that counts the number of days passed from Month=5 and Day=1.**

```
# Create a Date object for May 1st
may_first <- as.Date("2023-05-01")
# Create a Date object using 'Month' and 'Day'
Date <- as.Date(paste("2023", airq$Month, airq$Day, sep="-"))
# Calculate the difference in days from May 1st
airq$Days_since <- as.numeric(difftime(Date,  may_first, units = "days"))
```

```
> airq
    Ozone Solar.R Wind Temp Month Day Ozone_bin Solar_bin Days_since
1      41     190  7.4   67     5   1      bin4      bin2          0
2      36     118  8.0   72     5   2      bin3      bin2          1
3      12     149 12.6   74     5   3      bin1      bin2          2
4      18     313 11.5   62     5   4      bin2      bin4          3
5      NA      NA 14.3   56     5   5     binNA     binNA          4
6      28      NA 14.9   66     5   6      bin3     binNA          5
7      23     299  8.6   65     5   7      bin2      bin4          6
8      19      99 13.8   59     5   8      bin2      bin1          7
9       8      19 20.1   61     5   9      bin1      bin1          8
10     NA     194  8.6   69     5  10     binNA      bin2          9
11      7      NA  6.9   74     5  11      bin1     binNA         10
12     16     256  9.7   69     5  12      bin2      bin3         11
13     11     290  9.2   66     5  13      bin1      bin4         12
```

Exercise 6

1. **Numerise the class column, where Crew=4, 1st=3, 2nd=2 and 3rd=1.**

```
class_mapping <- c ("Crew" = 4, "1st" = 3, "2nd" = 2, "3rd" = 1)

titanic$Class <- sapply(titanic$Class, function(x) class_mapping[x])
```

```
> titanic
    X Class    Sex   Age Survived Freq
1   1     3   Male Child       No    0
2   2     2   Male Child       No    0
3   3     1   Male Child       No   35
4   4     4   Male Child       No    0
```

2. **Transform the titanic data frame into a new data frame (titanic2) with as many examples as passengers using the Freq column. In other words, there should be no rows for those for which Freq=0 and there should be 35 replicated rows for those with Freq=35.**

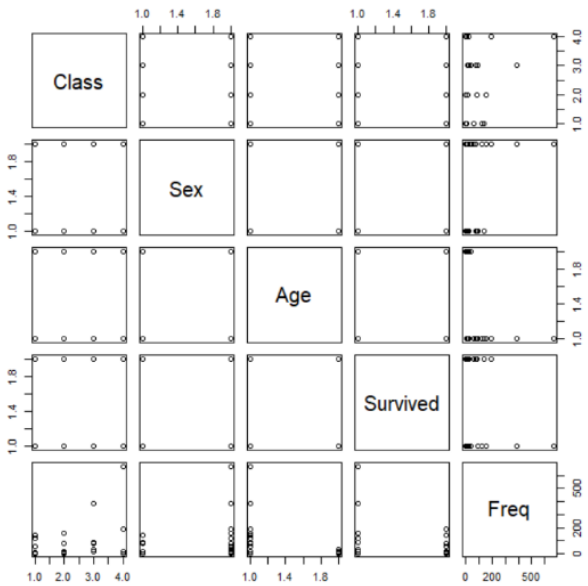
```
titanic2 <- titanic[titanic$Freq != 0,]

titanic2 <- titanic[rep(rownames(titanic), titanic$Freq),]
```

3. **Compare the plots of the original titanic data frame with the new one.**

They are the same, that's because the points in the plot of the other 2169 replicated elements are in the same spots as the original elements.

```
> nrow(titanic)
[1] 32
> nrow(titanic2)
[1] 2201
```

plot(titanic2)



Exercise 7

1. **Calculate a correlation matrix for the air dataset. Do you see a pair of attributes that are redundant?**

We could argue about Temperature and Month, or Temperature and Wind being correlated because they have a relative high correlation value 0.42 and 0.45. But it is uncertain that they are redundant.

```
> cor(airq)
         Ozone Solar.R       Wind       Temp        Month          Day
Ozone        1      NA         NA         NA           NA           NA
Solar.R     NA       1         NA         NA           NA           NA
Wind        NA      NA  1.0000000 -0.4579879 -0.178292579  0.027180903
Temp        NA      NA -0.4579879  1.0000000  0.420947252 -0.130593175
Month       NA      NA -0.1782926  0.4209473  1.000000000 -0.007961763
Day         NA      NA  0.0271809 -0.1305932 -0.007961763  1.000000000
```

2. **Calculate a correlation matrix for the car's dataset. Do you see a pair of attributes that are redundant?**

We clearly see that the is a correlation between the speed of the car and his stopping distance it could be a sign that one of the attributes could be redundant.

```
> cor(cars)
          speed      dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
```

3. **Using the data frame 'air', perform a simple random sampling of 50 examples.**

```
> airq[sample(nrow(airq), 50), ]
    Ozone Solar.R Wind Temp Month Day Ozone_bin Solar_bin
140    18     224 13.8   67     9  17      bin2      bin3
137     9      24 10.9   71     9  14      bin1      bin1
107    NA      64 11.5   79     8  15     binNA      bin1
55     NA     250  6.3   76     6  24     binNA      bin3
75     NA     291 14.9   91     7  14     binNA      bin4
35     NA     186  9.2   84     6   4     binNA      bin2
58     NA      47 10.3   73     6  27     binNA      bin1
26     NA     266 14.9   58     5  26     binNA      bin4
48     37     284 20.7   72     6  17      bin3      bin4
80     79     187  5.1   87     7  19      bin5      bin2
118    73     215  8.0   86     8  26      bin4      bin3
```

4. **Using the data frame 'air', perform a stratified random sampling of 5 examples of each month.**

```
stratified_data<- data.frame()

for (month in unique(airq$Month)) {

        subset_month<- subset(airq, Month == month)

        strat<-subset_month[sample(nrow(subset_month),5),]

        stratified_data <- rbind(stratified_data ,strat)

}
```

```
> stratified_data
    Ozone Solar.R Wind Temp Month Day Ozone_bin Solar_bin
14     14     274 10.9   68     5  14      bin1      bin4
26     NA     266 14.9   58     5  26     binNA      bin4
31     37     279  7.4   76     5  31      bin3      bin4
3      12     149 12.6   74     5   3      bin1      bin2
11      7      NA  6.9   74     5  11      bin1     binNA
59     NA      98 11.5   80     6  28     binNA      bin1
39     NA     273  6.9   87     6   8     binNA      bin4
57     NA     127  8.0   78     6  26     binNA      bin2
43     NA     250  9.2   92     6  12     binNA      bin3
51     13     137 10.3   76     6  20      bin1      bin2
84     NA     295 11.5   82     7  23     binNA      bin4
78     35     274 10.3   82     7  17      bin3      bin4
80     79     187  5.1   87     7  19      bin5      bin2
86    108     223  8.0   85     7  25      bin5      bin3
72     NA     139  8.6   82     7  11     binNA      bin2
108    22      71 10.3   77     8  16      bin2      bin1
102    NA     222  8.6   92     8  10     binNA      bin3
```