sample-pratice-age-and-sales

09/03/2022

Contents

statistics pratice

```
# create a variable
scores = c(75, 80, 85)

# Do statistics
scores * 2
```

```
## [1] 150 160 170
```

```
mean(scores)     # mean()
```

```
## [1] 80
```

```
median(scores)   # median()
```

```
## [1] 80
```

```
sd(scores)       # sd()
```

```
## [1] 5
```

Data Frame Pratice

Build a Data Frame

```
# Build a group of related vectors
Names  =  c("Mike  Hussey",  "Aaron  Finch",  "Brad  Hogg",  "Steve  Smith",
           "George  Bailey",  "Mitchell  Johnson",  "Shaun  Marsh",
           "Glenn  Maxwel l",  "Pat  Cummins",  "Mitchell  Starc",  "David  Warner")
Ages = c(39,28,44,25,32,33,31,26,22,25,28)
IPLSals = c(310,662,103,828,672,1340,455,1240,207,1030,1140)


# Combine different vectors
df = data.frame(Names, Ages,IPLSals)
df
```

```
##                Names Ages IPLSals
## 1       Mike  Hussey   39     310
## 2       Aaron  Finch   28     662
## 3         Brad  Hogg   44     103
## 4       Steve  Smith   25     828
## 5     George  Bailey   32     672
## 6  Mitchell  Johnson   33    1340
## 7       Shaun  Marsh   31     455
## 8    Glenn  Maxwel l   26    1240
## 9       Pat  Cummins   22     207
## 10   Mitchell  Starc   25    1030
## 11     David  Warner   28    1140
```

Retrive rows/columns from dataframe

```
df[1,]      # (row1,columns*)
```

```
##          Names Ages IPLSals
## 1 Mike  Hussey   39     310
```

```
df[,1]      # (row*,column1)
```

```
##  [1] "Mike  Hussey"      "Aaron  Finch"      "Brad  Hogg"
##  [4] "Steve  Smith"      "George  Bailey"    "Mitchell  Johnson"
##  [7] "Shaun  Marsh"      "Glenn  Maxwel l"   "Pat  Cummins"
## [10] "Mitchell  Starc"   "David  Warner"
```

```
df[2:3]     # (column2,column3)
```

```
##    Ages IPLSals
## 1    39     310
## 2    28     662
```

```
## 3    44     103
## 4    25     828
## 5    32     672
## 6    33    1340
## 7    31     455
## 8    26    1240
## 9    22     207
## 10   25    1030
## 11   28    1140
```

```
df[2:3,]    # (row2,row3)
```

```
##          Names Ages IPLSals
## 2 Aaron  Finch   28     662
## 3   Brad  Hogg   44     103
```

Determine statistics

```
mean(df[,2])
```

```
## [1] 30.27273
```

```
mean(df[,"Ages"])
```

```
## [1] 30.27273
```

```
mean(df$Ages)
```

```
## [1] 30.27273
```

```
var(df[,2]) # run help("var") for explaination
```

```
## [1] 42.81818
```

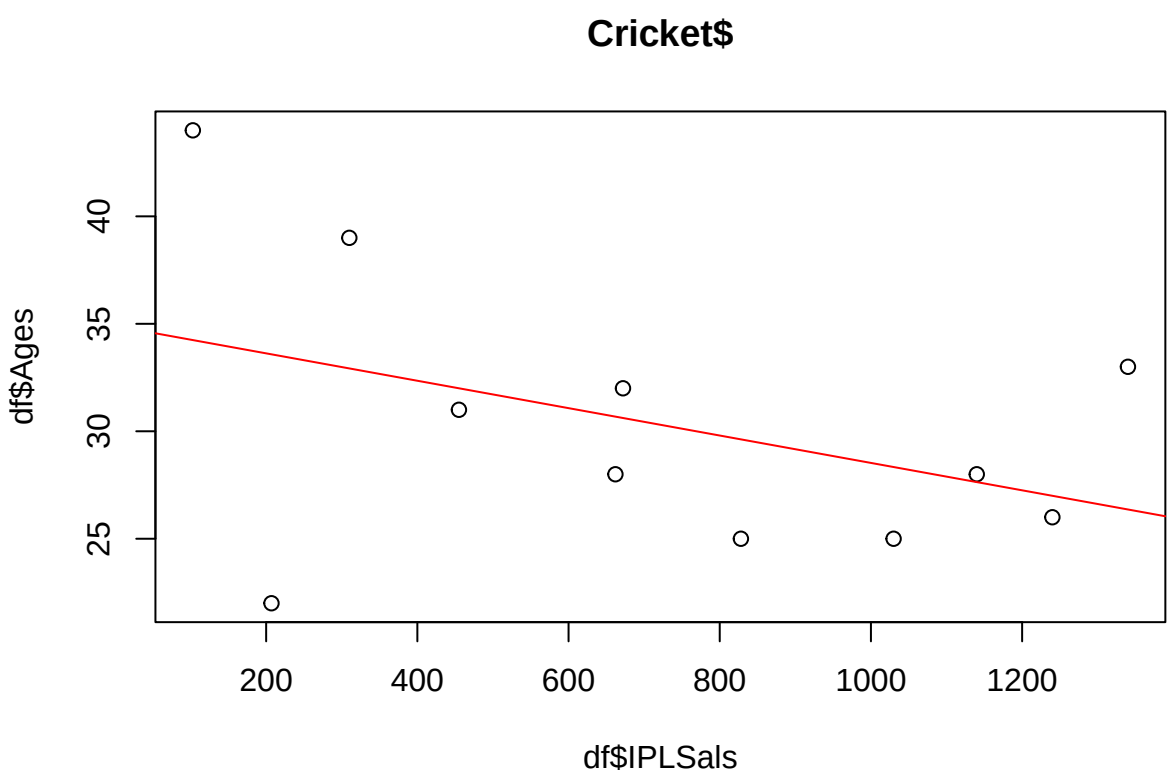
```
sd(df[,2])
```

```
## [1] 6.54356
```

Display Data

Predict Age

```
plot(df$Ages ~ df$IPLSals, main = "Cricket$")    #  Add a plot
abline(lm(df$Ages ~ df$IPLSals), col = "red")    #  Adding a linear regression
```



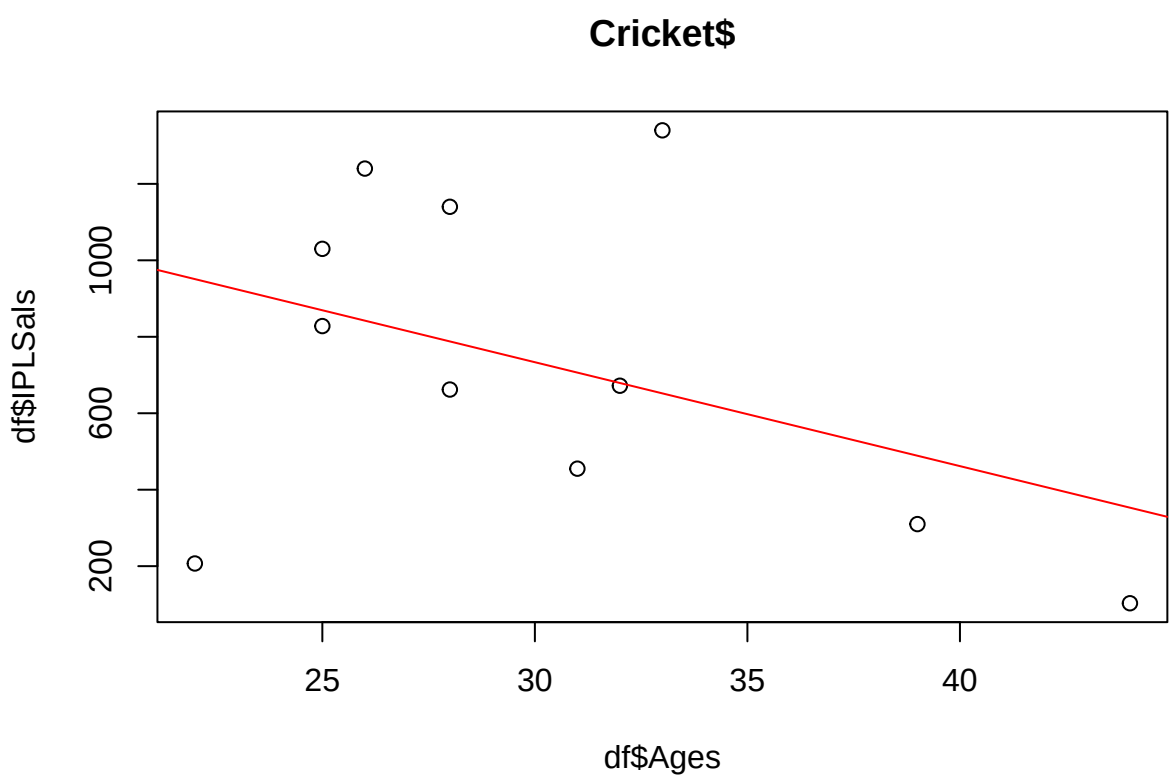

```
fit = lm(df$Ages ~ df$IPLSals)  # lm() Function to Fit Linear Models
fit
```

```
##
## Call:
## lm(formula = df$Ages ~ df$IPLSals)
##
## Coefficients:
## (Intercept)   df$IPLSals
##   34.899143    -0.006372
```

NOTES
Ages: as response variable
IPLSales: as a predictor
I.e. predict a player's age from his IPL cricket salary
So the regression equation is: **estimated age = -0.006372 x salary + 34.899143**

Predict Salary

```
plot(df$IPLSals ~ df$Ages, main = "Cricket$")   # Plot it
abline(lm(df$IPLSals ~ df$Ages), col="red")
```



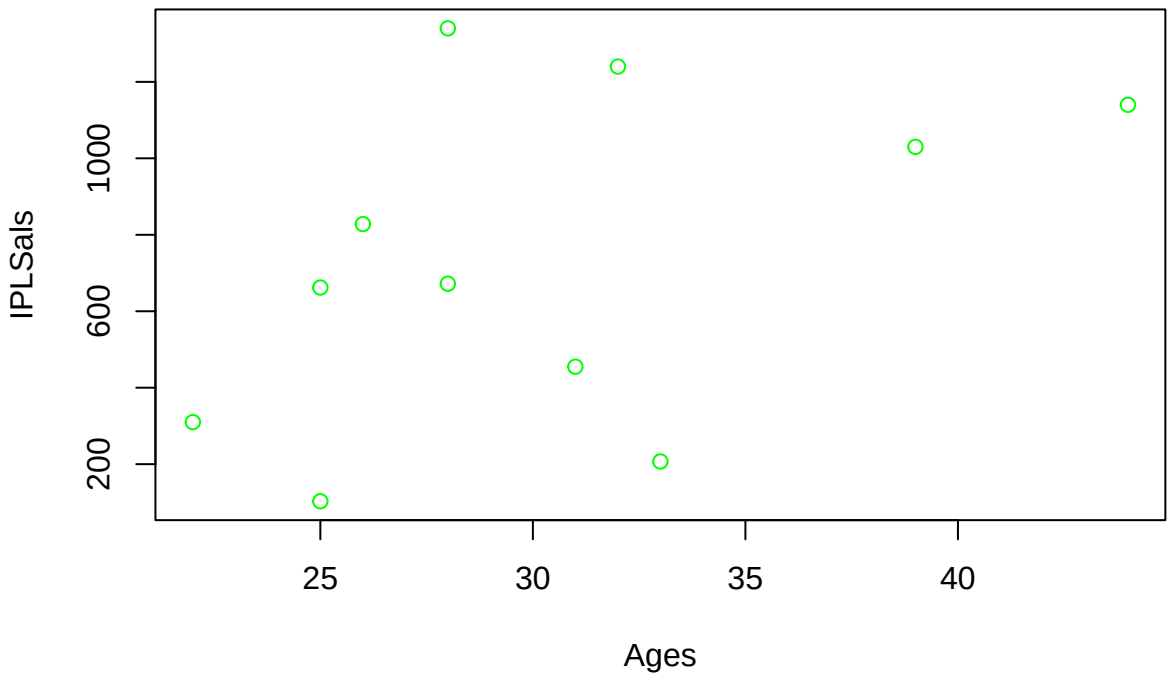

```
fit2 = lm(df$IPLSals ~ df$Ages)     # lm() Function to Fit Linear Models
fit2
```

```
##
## Call:
## lm(formula = df$IPLSals ~ df$Ages)
##
## Coefficients:
## (Intercept)      df$Ages
##     1548.85       -27.18
```

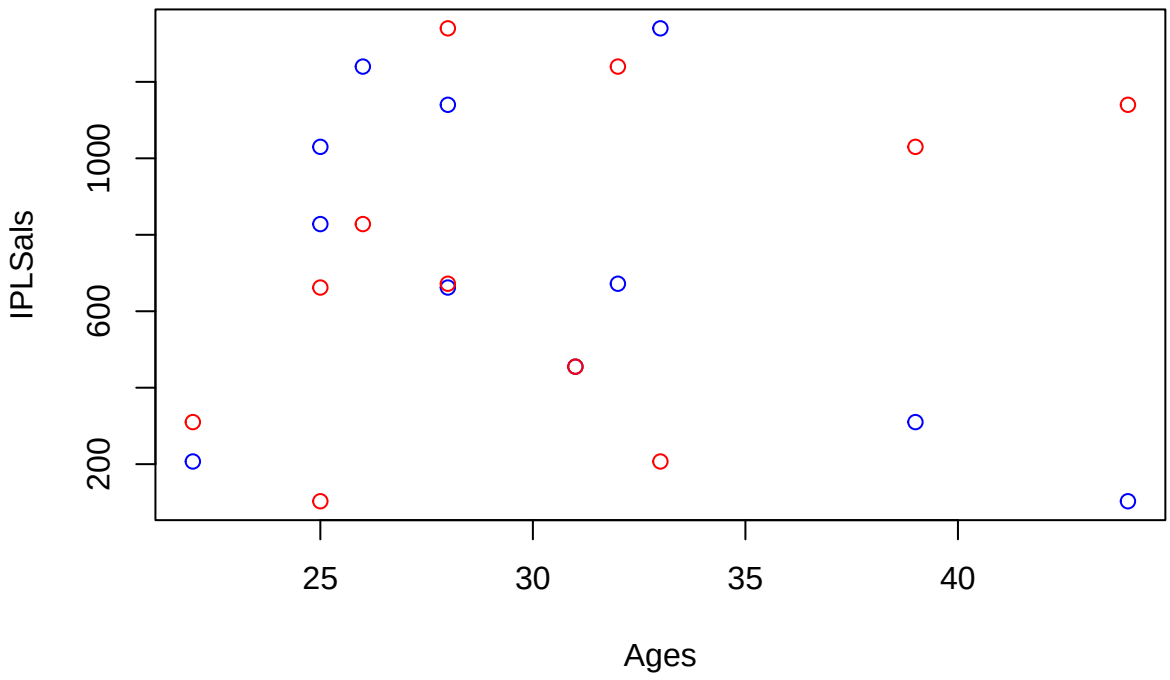
NOTES
IPLSales: as response variable
Ages: as a predictor
I.e. predict a player's IPL cricket salary from his age.
So the regression equation is: **estimated salary = -27.18 x age + 1548.85**

explore other plots

```
sorted = sort(Ages)
plot(sorted, IPLSals, xlab = "Ages", col = "green" )
```

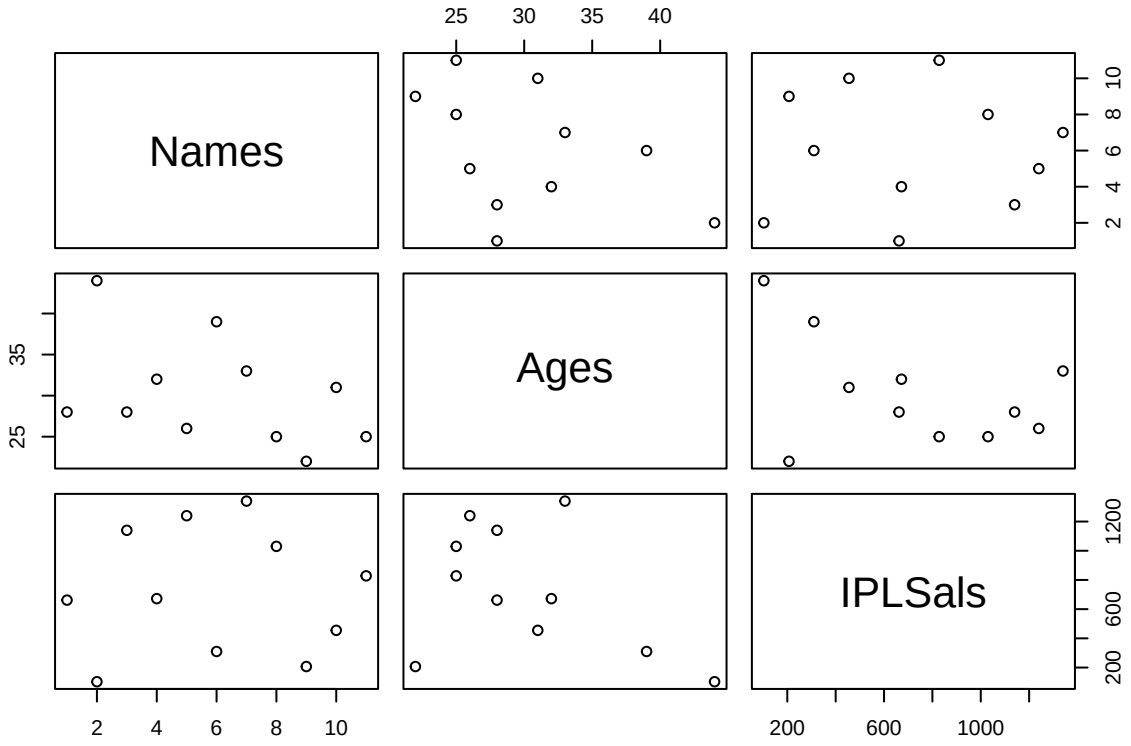



```
plot(Ages, IPLSals, col = "blue" )
points(sorted, IPLSals, col = "red" )                    # add points to a plot
```



explore plot of whole dataframe

```
plot(df)
```


Names
Ages
IPLSals
25
30
35
40
2
4
6
8
10
25
35
2
4
6
8
10
200
600
1000
200
600
1200