R assignment

2023-02-22

```
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(lattice)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2
## --

## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.3.0     v forcats 1.0.0
## v readr   2.1.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following object is masked from 'package:purrr':
##
##     compact
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(readr)

#import the data
df <- read.csv('Seagal box office.csv')

#BASIC_INSIGHTS
summary(df)
```

```
##       Year          Film             Box.Office        Budget
##  Min.   :1988   Length:47          Min.   : 0.00   Min.   : 0.00
##  1st Qu.:2002   Class :character   1st Qu.: 0.00   1st Qu.: 8.50
##  Median :2006   Mode  :character   Median : 0.00   Median :10.00
##  Mean   :2005                      Mean   :11.19   Mean   :16.64
##  3rd Qu.:2011                      3rd Qu.:15.50   3rd Qu.:15.00
##  Max.   :2016                      Max.   :83.00   Max.   :60.00
```

```
glimpse(df)
```

```
## Rows: 47
## Columns: 4
## $ Year       <int> 1988, 1990, 1990, 1991, 1992, 1994, 1995, 1996, 1996, 1997,~
## $ Film       <chr> "Above the Law", "Hard to Kill", "Marked for Death", "Out f~
## $ Box.Office <int> 19, 47, 46, 39, 83, 39, 50, 68, 20, 16, 1, 51, 1, 15, 1, 1,~
## $ Budget     <int> 8, 10, 12, 14, 35, 50, 60, 55, 45, 60, 25, 50, 7, 13, 17, 1~
```

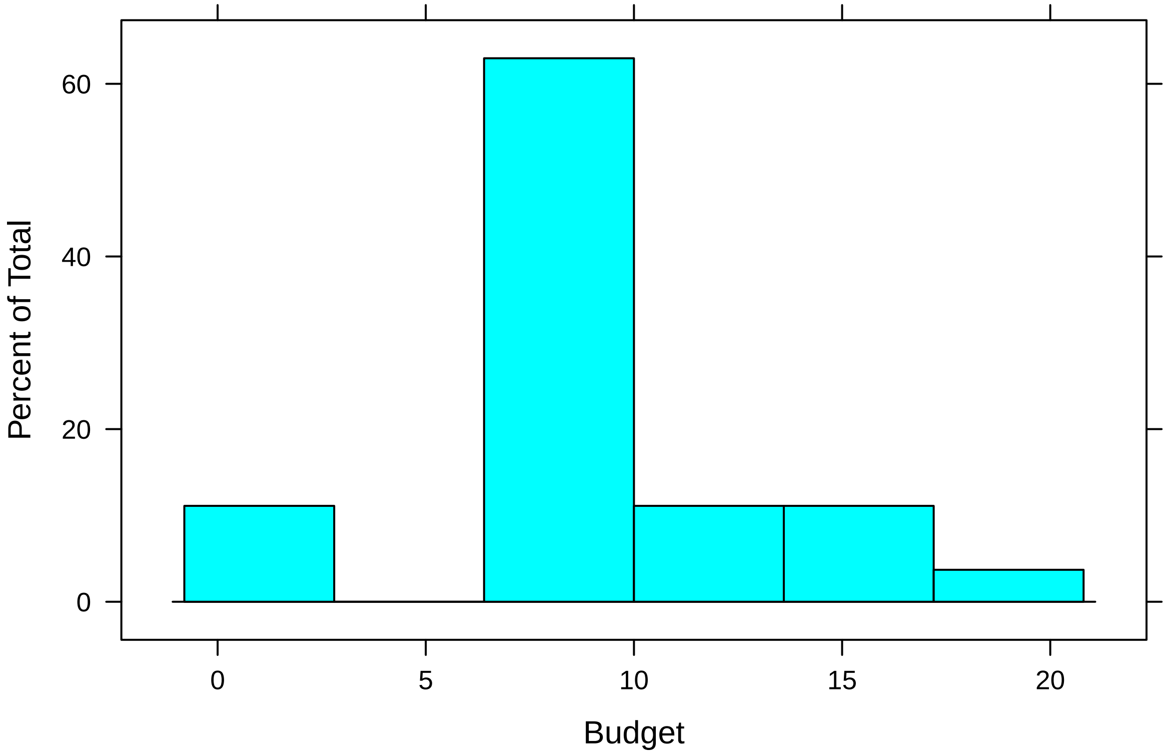
```
#DISTINCT_DATA

colSums(is.na(df))
```

```
##       Year       Film Box.Office     Budget
##          0          0          0          0
```

```
df2=subset(df,Box.Office==0)
summary(df2)
```

```
##       Year          Film             Box.Office     Budget
##  Min.   :2004   Length:27          Min.   :0    Min.   : 0.000
##  1st Qu.:2006   Class :character   1st Qu.:0    1st Qu.: 8.000
##  Median :2009   Mode  :character   Median :0    Median :10.000
##  Mean   :2011                      Mean   :0    Mean   : 9.519
##  3rd Qu.:2016                      3rd Qu.:0    3rd Qu.:11.000
##  Max.   :2016                      Max.   :0    Max.   :20.000
```

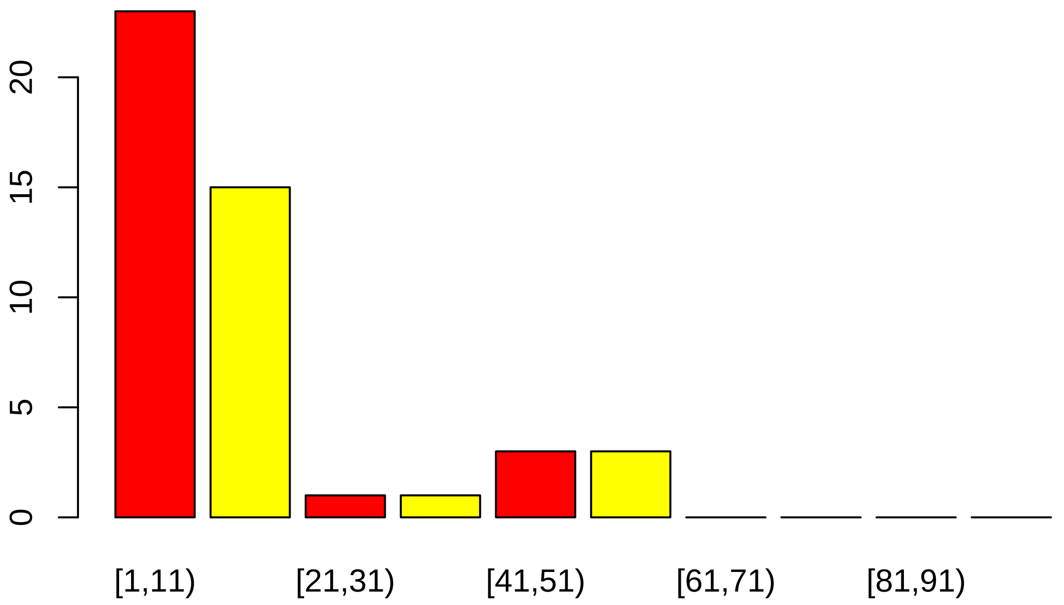
```
histogram(~Budget,data = df2)
```



```
#EXPLORATORY DATA ANALYSIS
#BAR_PLOT
Budget = pull(df,Budget)
Budget_1=cut(Budget,breaks=seq(1,101,by=10),right=FALSE)
table(Budget_1)
```

```
## Budget_1
##   [1,11)  [11,21)  [21,31)  [31,41)  [41,51)  [51,61)  [61,71)  [71,81)
##       23       15        1        1        3        3        0        0
##  [81,91) [91,101)
##        0        0
```

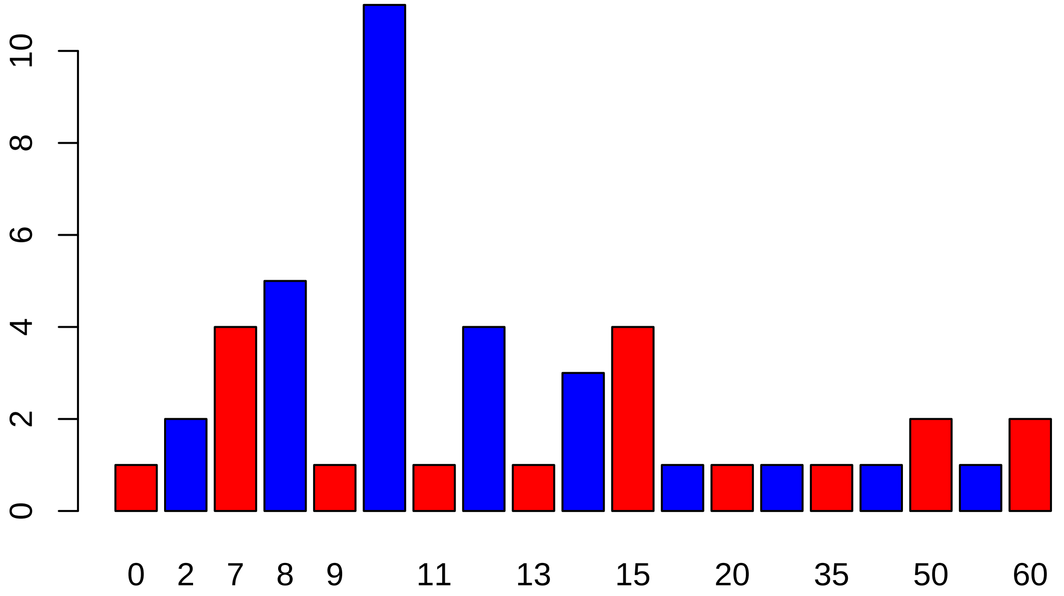
```
barplot(table(Budget_1),col=c("red","yellow"))
```



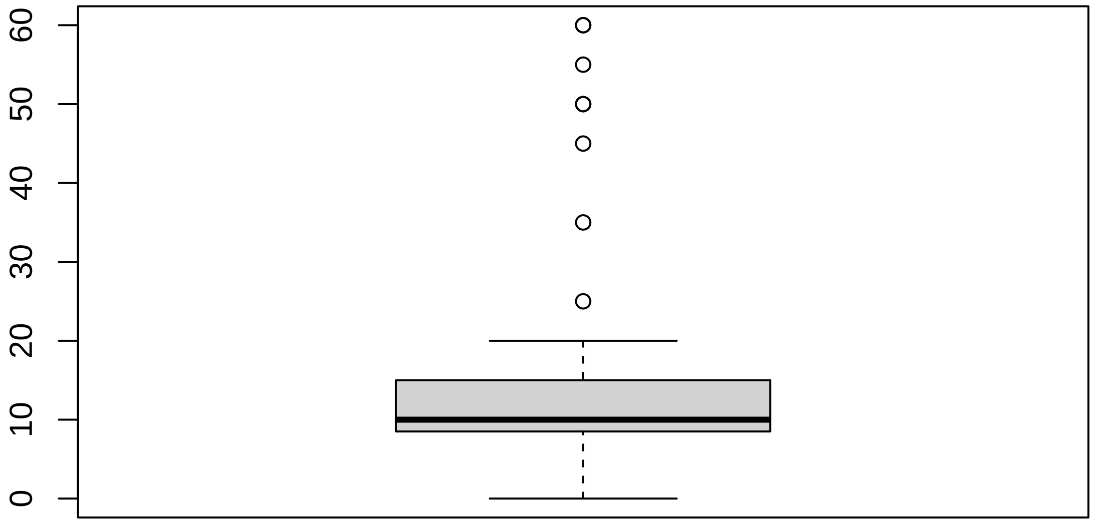
```
table(Budget)
```

```
## Budget
##  0  2  7  8  9 10 11 12 13 14 15 17 20 25 35 45 50 55 60
##  1  2  4  5  1 11  1  4  1  3  4  1  1  1  1  1  2  1  2
```

```
barplot(table(Budget),col=c("red","blue"))
```



```
#BOXPLOT
boxplot(Budget)
boxplot(x=df$Budget,y=df$Box.Office)
```



```
#SUBSETTING INTERQUARTILE DATA OF BUDGET
df3 =filter(df,Budget>=20 & Budget<=60)
df3
```

```
##   Year                           Film Box.Office Budget
## 1 1992                    Under Siege         83     35
## 2 1994               On Deadly Ground         39     50
## 3 1995 Under Siege 2: Dark Territory         50     60
## 4 1996             Executive Decision         68     55
## 5 1996                The Glimmer Man         20     45
## 6 1997                Fire Down Below         16     60
## 7 1998                    The Patriot          1     25
## 8 2001                    Exit Wounds         51     50
## 9 2004                   Out of Reach          0     20
```

```
glimpse(df3)
```

```
## Rows: 9
## Columns: 4
## $ Year       <int> 1992, 1994, 1995, 1996, 1996, 1997, 1998, 2001, 2004
## $ Film       <chr> "Under Siege", "On Deadly Ground", "Under Siege 2: Dark Ter~
## $ Box.Office <int> 83, 39, 50, 68, 20, 16, 1, 51, 0
## $ Budget     <int> 35, 50, 60, 55, 45, 60, 25, 50, 20
```

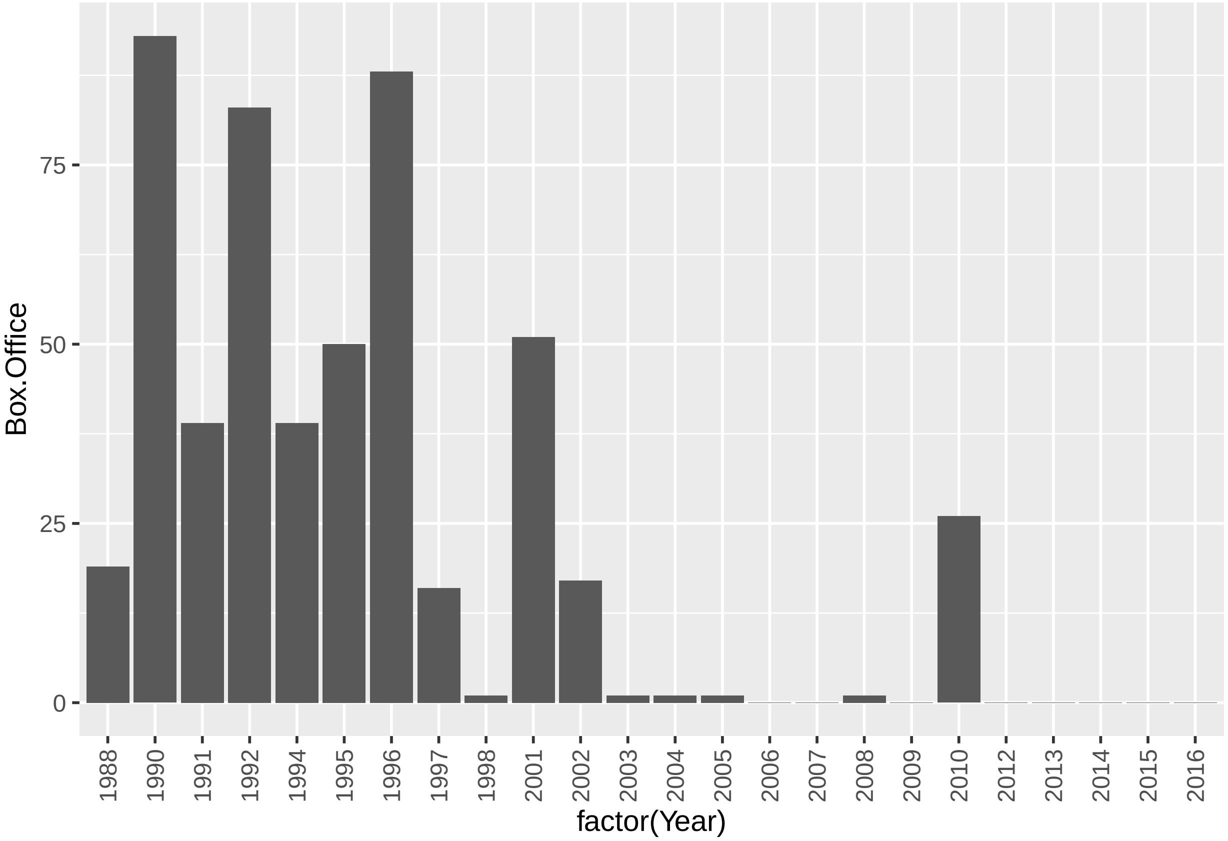
```
summary(df3)
```

```
##       Year          Film             Box.Office       Budget
##  Min.   :1992   Length:9           Min.   : 0.00   Min.   :20.00
##  1st Qu.:1995   Class :character   1st Qu.:16.00   1st Qu.:35.00
##  Median :1996   Mode  :character   Median :39.00   Median :50.00
##  Mean   :1997                      Mean   :36.44   Mean   :44.44
##  3rd Qu.:1998                      3rd Qu.:51.00   3rd Qu.:55.00
##  Max.   :2004                      Max.   :83.00   Max.   :60.00
```

```
#BAR CHART OF TWO ATTRIBUTES
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

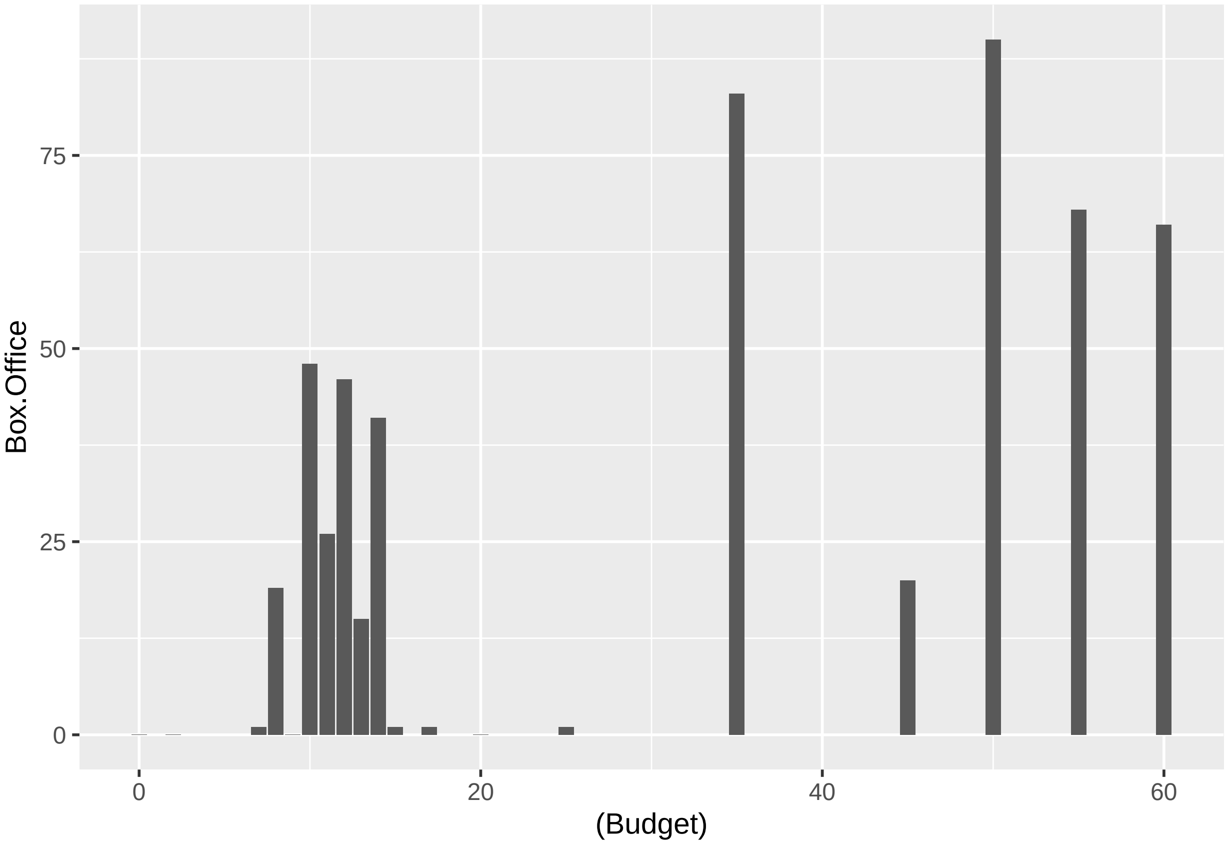
```
library(ggplot2)
ggplot(df, aes(x = factor(Year), y = Box.Office)) +
  geom_bar(stat = "Identity")+ scale_x_discrete(guide = guide_axis(angle = 90))
```



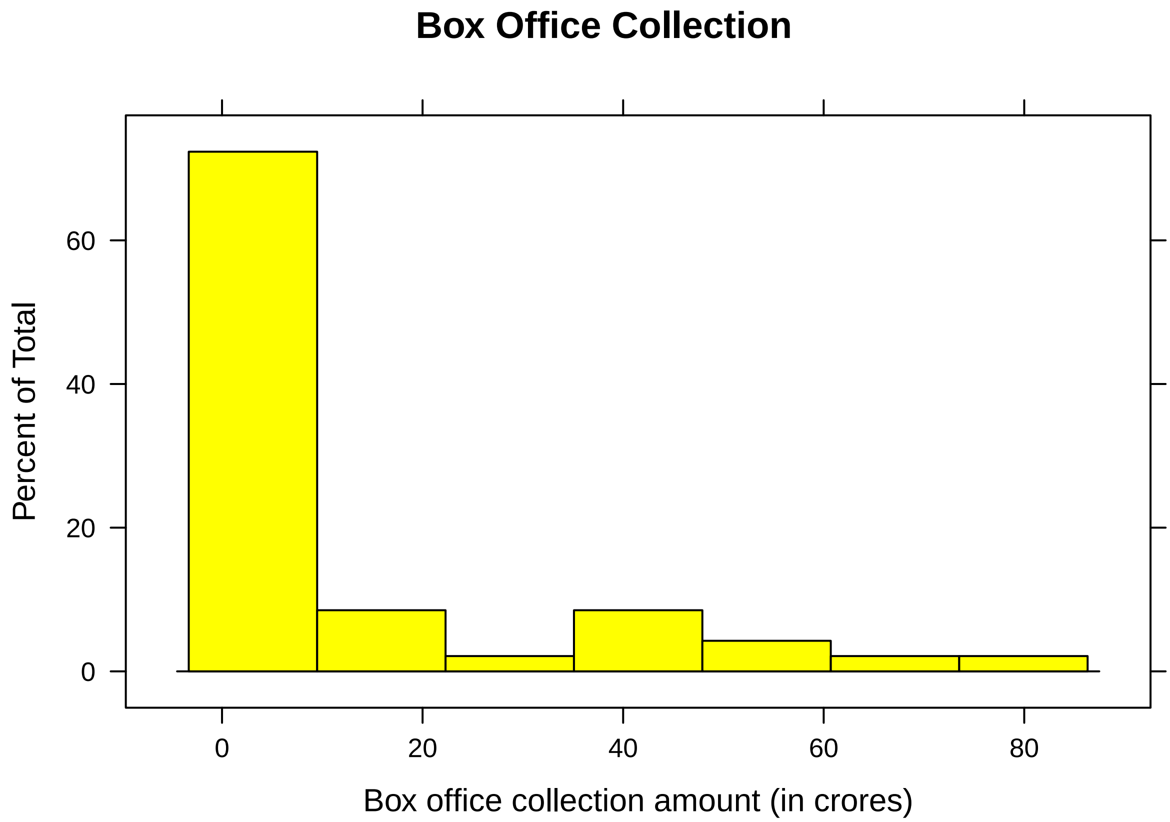
```
dfy=subset(df,Year==1990)
dfy
```

```
##   Year              Film Box.Office Budget
## 2 1990      Hard to Kill         47     10
## 3 1990 Marked for Death         46     12
```

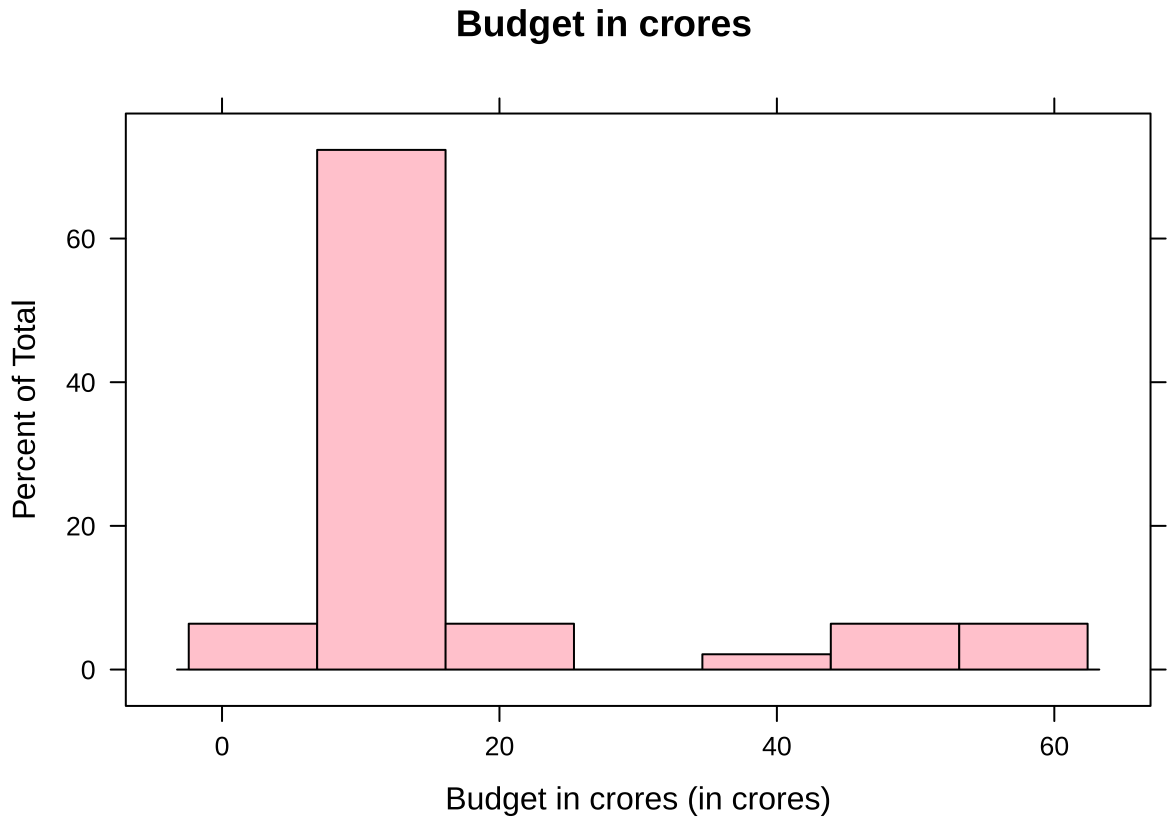
```
ggplot(df, aes(x = (Budget), y = Box.Office)) +
  geom_bar(stat = "Identity")
```



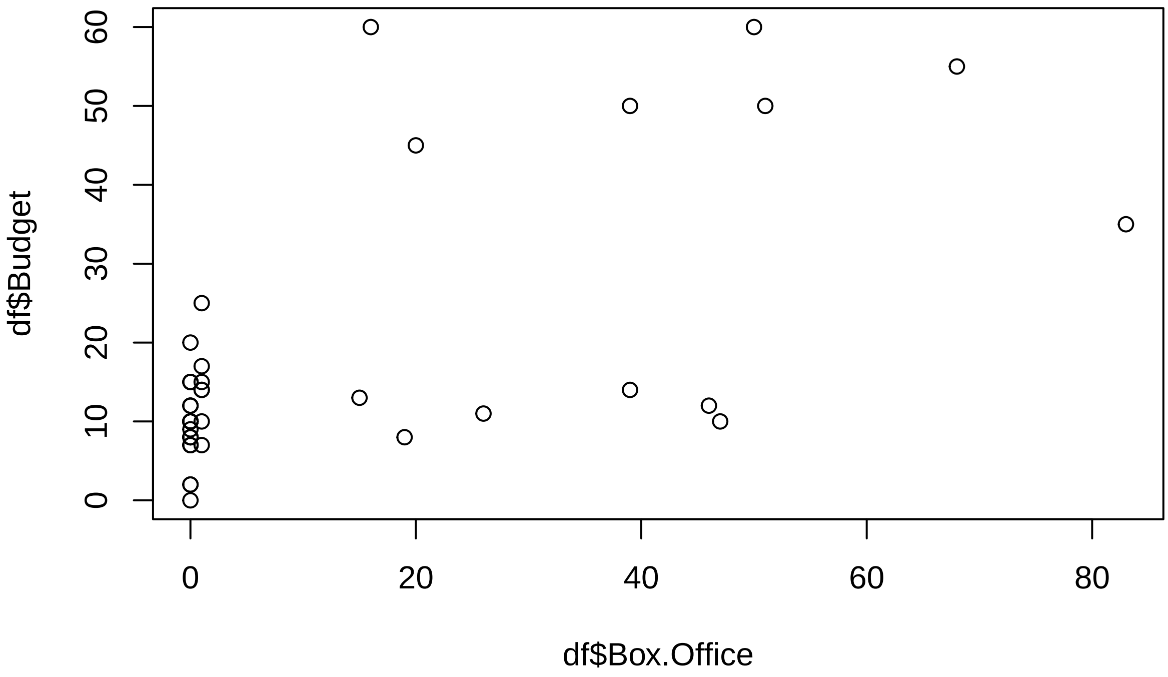
```
#HISTOGRAM
histogram(~df$`Box.Office`,col='yellow',main='Box Office Collection',xlab='Box office collection amount
```



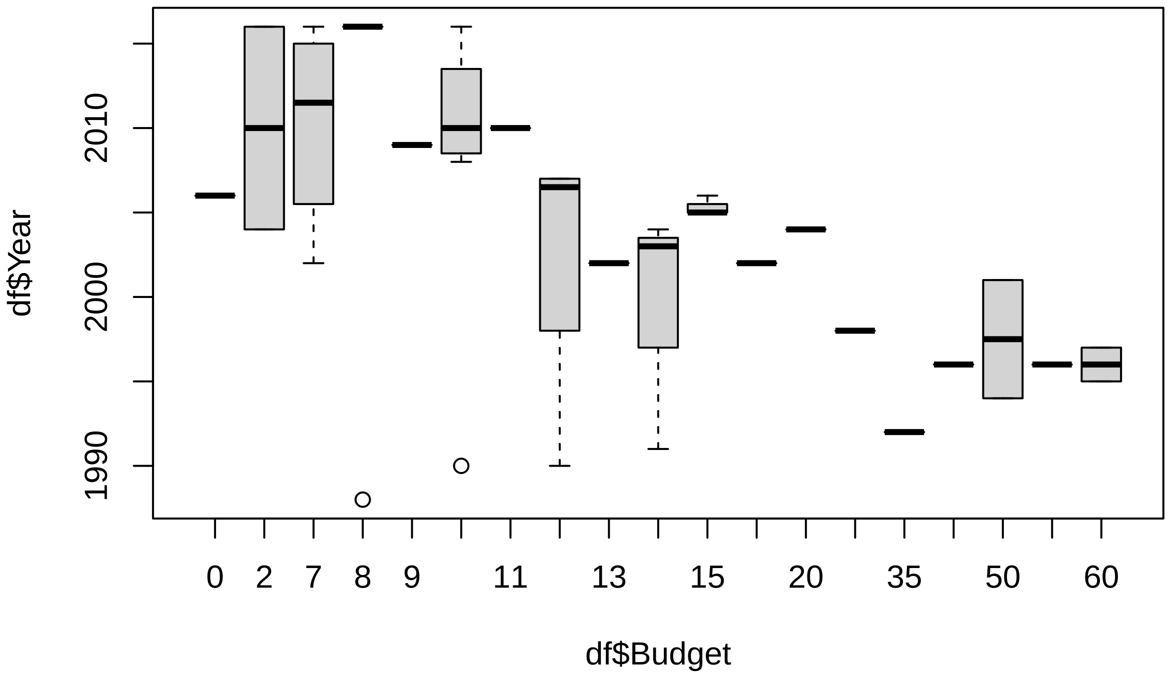
```
histogram(~df$`Budget`,col='pink',main='Budget in crores',xlab='Budget in crores (in crores)')
```



```
#SCATTER_PLOT
plot(x=df$Box.Office,y=df$Budget)
```



```
#BOXPLOT
boxplot(df$Year~df$Budget)
```


df$Year
df$Budget