STAT 40001/STAT50001 Statistical Computing Fall 2024

Lab-5

1) The *Duncan* data frame has 45 rows and 4 columns. Data on the prestige and other characteristics of 45 U. S. occupations in 1950. The data is in the library `car`.

a) Access the data.

b) Print first five observations of the data set.

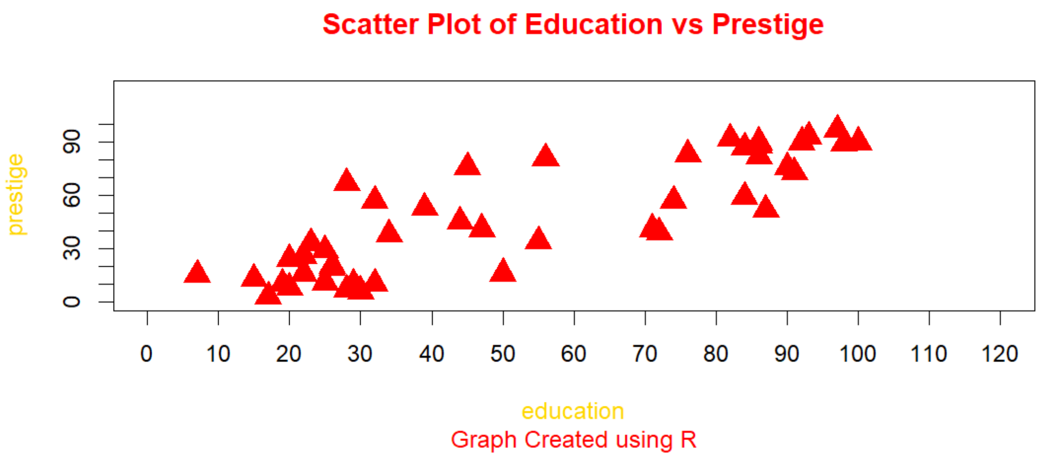
c) Use scatterplot to display the prestige scores according to the education level.

d) Change the color, title, labels etc. and save it.

```
> cat("The data Duncan is available in car package. So, the car package was installed
from the packages using the below command")
The data Duncan is available in car package. So, the car package was installed from th
e packages using the below command.
> install.packages("car")
> library(car)
Loading required package: carData
> data("Duncan")

> head(Duncan,5)
           type income education prestige
accountant prof     62        86       82
pilot      prof     72        76       83
architect  prof     75        92       90
author     prof     55        90       76
chemist    prof     64        86       90

> attach(Duncan)
> plot(education,prestige, pch = 17,col = "red",main = "Scatter Plot of Education vs P
restige", sub = "Graph Created using R", cex = 2,col.main = "red",col.sub = "red",col.
lab = "gold",xlim = c(0,120),ylim = c(0,120),axes = F)
> axis(1,at = seq(0,120,10))
> axis(2,at = seq(0,100,10))
> box()
```



2) The Davis data in the `car` package contains data on the measured and reported heights and weights of 200 men and women engaged in regular exercise.

a) Access the data.

b) A few of the data values are missing and are marked as “NA”. Clean the data by deleting the missing values.

c) How many individuals do you have with complete information?

```
> cat("The data Davis is available in car package. So, the car package was installed f
rom the packages using the below command")
The data Davis is available in car package. So, the car package was installed from the
packages using the below command
> install.packages("car")
> library(car)
Loading required package: carData
> data(Davis)

> head(Davis)
  sex weight height repwt repht
1   M     77    182    77   180
2   F     58    161    51   159
3   F     53    161    54   158
4   M     68    177    70   175
5   F     59    157    59   155
6   M     76    170    76   165

> cat("Dimension Before Cleaning")
Dimension Before Cleaning
> dim(Davis)
[1] 200   5

> Clean = na.omit(Davis)
> cat("Dimension After Cleaning")
Dimension After Cleaning
> dim(Clean)
[1] 181   5

> cat("So, after removing the rows with NA values in them, the #rows in the data reduc
es from 200 to 181, which indicates there are 19 such rows with NA vAlues")
So, after removing the rows with NA values in them, the #rows in the data reduces from
200 to 181, which indicates there are 19 such rows with NA vAlues
```

3) Access the data set “Elections” from ‘mdsr” package and extract the variable names included in the dataset.

```
> cat("Installing mdsr package, to access Elections data")
Installing mdsr package, to access Elections data
> install.packages("mdsr")
> library(mdsr)
> data("Elections")

> head(Elections, 6)
# A tibble: 6 x 13
   Ward Precinct `Registered Voters at 7am` `Voters Registering at Polls`
  <int>    <dbl>                      <int>                         <int>
1     1        1                       1878                            25
2     1        2                       2769                            43
3     1        3                       2337                            40
4     1        4                       2139                            24
5     1        5                       1875                            31
6     1        6                       2258                            69
```

```
# ℹ 9 more variables: `Voters Registering by Absentee` <int>,
#   `Total Registrations` <int>, `Voters at Polls` <int>,
#   `Absentee Voters` <int>, `Total Ballots Cast` <int>, `Total Turnout` <dbl>,
#   `Percentage Absentee` <dbl>, `% Registered to Total (Election Day)` <dbl>,
#   `Spoiled Ballots` <int>
> dim(Elections)
[1] 117  13
```

```
> names(Elections)
 [1] "Ward"                                "Precinct"
 [3] "Registered Voters at 7am"            "Voters Registering at Polls"
 [5] "Voters Registering by Absentee"      "Total Registrations"
 [7] "Voters at Polls"                     "Absentee Voters"
 [9] "Total Ballots Cast"                  "Total Turnout"
[11] "Percentage Absentee"                 "% Registered to Total (Election Day)"
[13] "Spoiled Ballots"
```

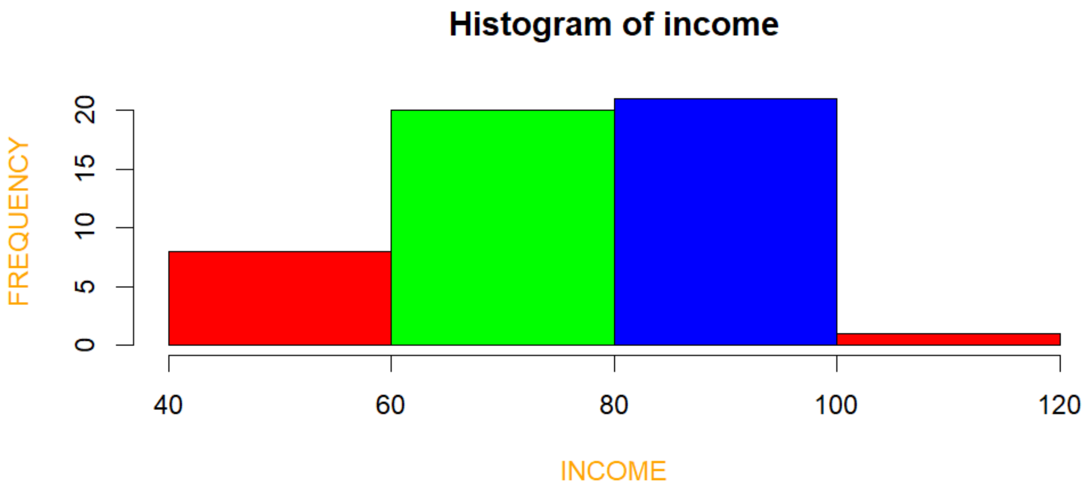
4) Link below provides a list of datasets related to economics (Data are from: principles of Econometrics)

http://www.principlesofeconometrics.com/poe4/poe4stata.htm

a) Import dataset entitled “savings” in R.
b) What is the dimension of the data?
c) Draw a histogram of the data related to the income. Please make sure to change the color, provide the title, labels etc.

```
> cat("Installing haven package for accessing the stata data")
Installing haven package for accessing the stata data
> install.packages("haven")
> library(haven)
> url <- "http://www.principlesofeconometrics.com/poe4/data/stata/savings.dta"
> Q4 = read_dta(url)
> head(Q4)
# A tibble: 6 × 3
  savings income avgincome
    <dbl>  <dbl>     <dbl>
1    2.41   83.8      65.9
2    2.47   68.1      64.6
3    4.59   84.2      71.7
4    3.89   84.0      64.6
5    3.82   52.3      60.7
6    5.35   97.0      79.5
> dim(Q4)
[1] 50  3
> attach(Q4)
> hist(income, col = rainbow(3),breaks = 4, xlab = "INCOME", ylab = "FREQUENCY",col.la
b = "orange")
```

PTO



Histogram of income
FREQUENCY
0
5
10
15
20
40
60
80
100
120
INCOME