Alexandria University
Faculty of Computer and Data Science
Department : Data Science
Course Title: Data Science 2023-2024

Project Name

In

Introduction to Data Science

Course Code: 02-24-00104

Members Names and Role

Name	ID	Role
1. Ziad Elkhateeb	22010100	
2.Qamar Mosallam	22011438	
3.Menna Osama	22010399	
4.Gehad Yousry	22011437	
5.		
6.		

1. Introduction:

In this section you should describe the idea of the project and its objective, the inputs and outputs, the used dataset and its parameters.

- Description & objective:

  the dataset is about (Employee Promotion Data) based on some features (our variables) to make the HR in company decide if the employee deserve the promotion or not.

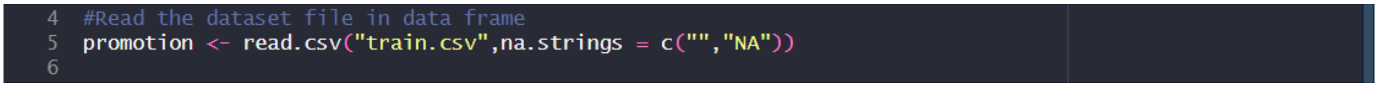
- Features (Inputs & Outputs):
  - employee_id: Unique ID for employee (input)
  - department: Department of employee (input)
  - region: Region of employment (unordered) (input))
  - education: Education Level (input)
  - gender: Gender of Employee (input)
  - recruitment_channel: Channel of recruitment for employee (input)
  - no_ of_ trainings: no of other trainings completed in previous year on soft skills, technical skills etc. (input)
  - age: Age of Employee (input)
  - previous_ year_ rating: Employee Rating for the previous year (input)
  - length_ of_ service: Length of service in years (input)
  - awards_ won?: if awards won during previous year then 1 else 0 (input)
  - avg_ training_ score: Average score in current training evaluations (input)
  - is_promoted: (Target) Recommended for promotion (output)

- Inspiration:

  Predict whether a potential promotee at checkpoint in the test set will be promoted or not after the evaluation process.

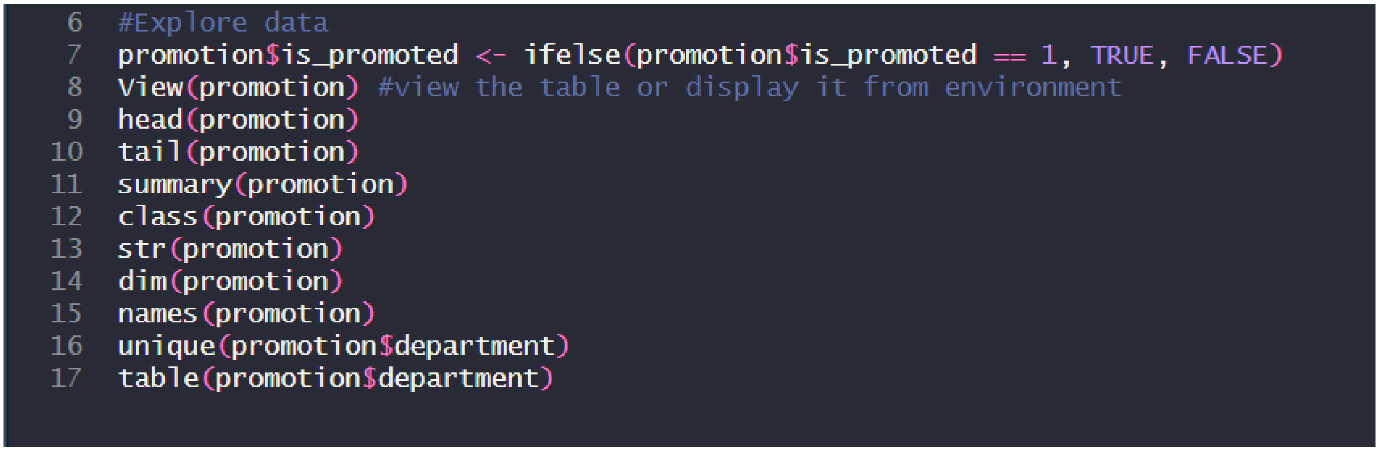
2. Methodologies used:

In this section you should explain your project steps in details, write the name of your project methodologies or techniques used and how and why you use them.



```r
#Read the dataset file in data frame
promotion <- read.csv("train.csv",na.strings = c("","NA"))

```

we used read.csv () to read dataset & use the na.strings() argument to replace the "" with "NA" to do the right statistics in cleaning and exploration.

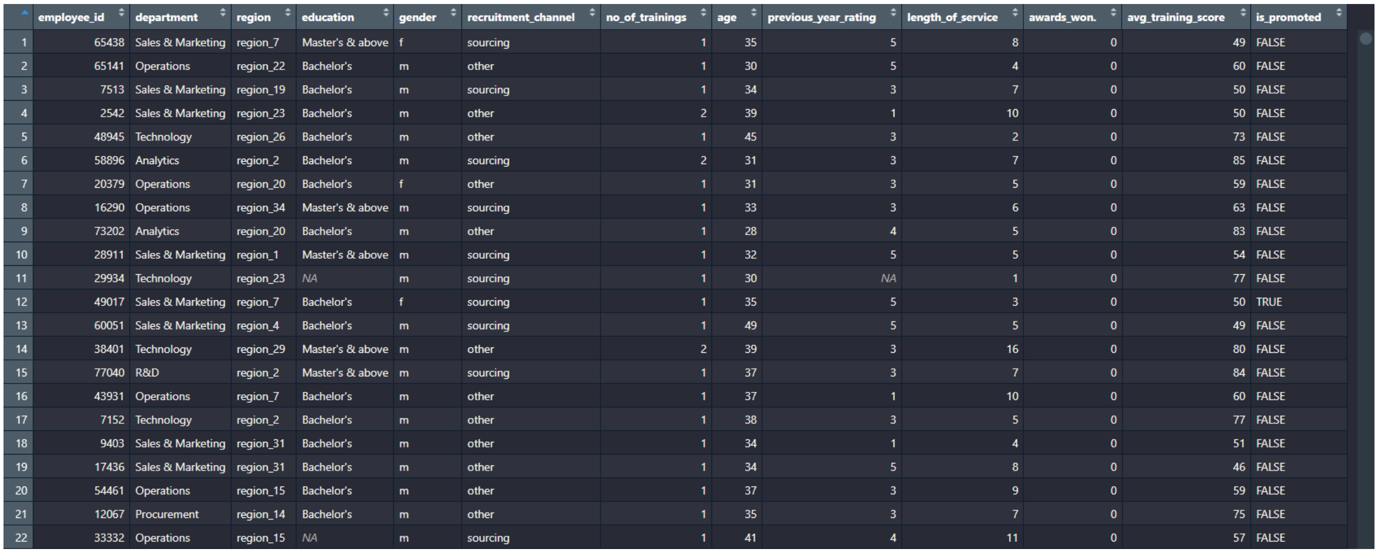


```r
#Explore data
promotion$is_promoted <- ifelse(promotion$is_promoted == 1, TRUE, FALSE)
View(promotion) #view the table or display it from environment
head(promotion)
tail(promotion)
summary(promotion)
class(promotion)
str(promotion)
dim(promotion)
names(promotion)
unique(promotion$department)
table(promotion$department)
```

1-ifelse(promotion$variable): used to convert the 1s with TRUE and 0s with FALSE in data frame and make is_promoted column is logical

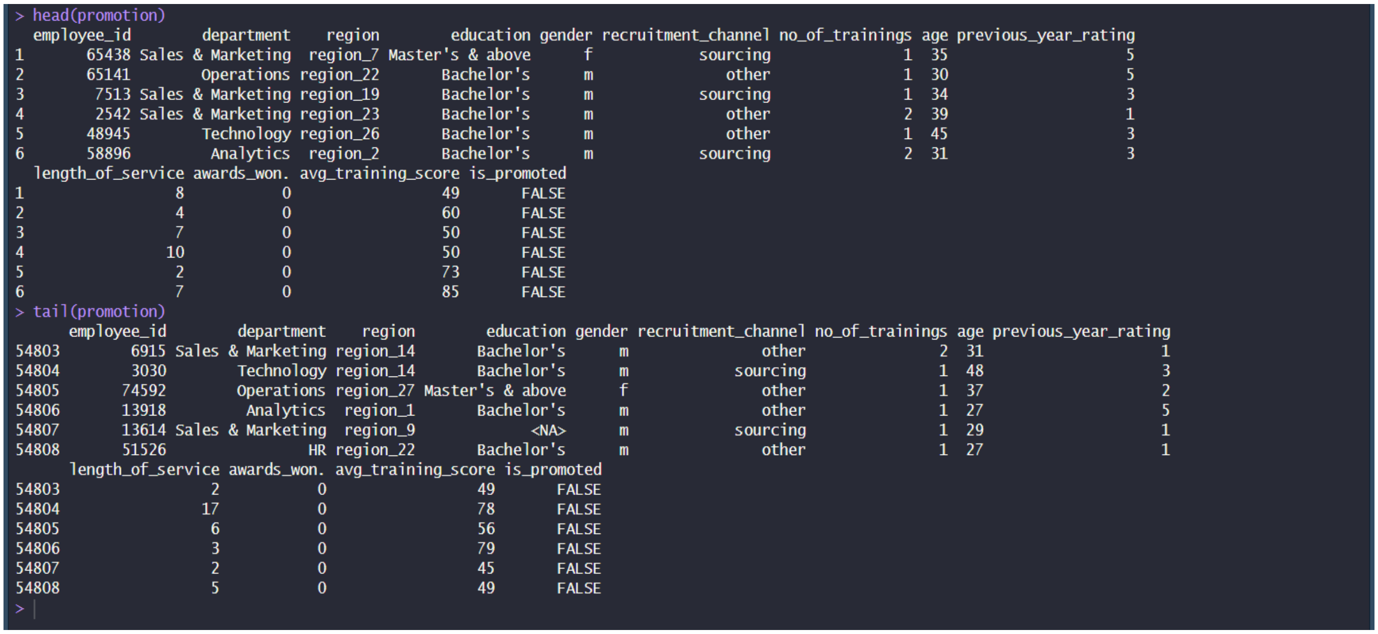
2-promotion$is_promoted <- as.logical(promotion$is_promoted) : used to convert the data type of values in is_promoted (column) to logical data type

View(promotion): display the data frame in table.



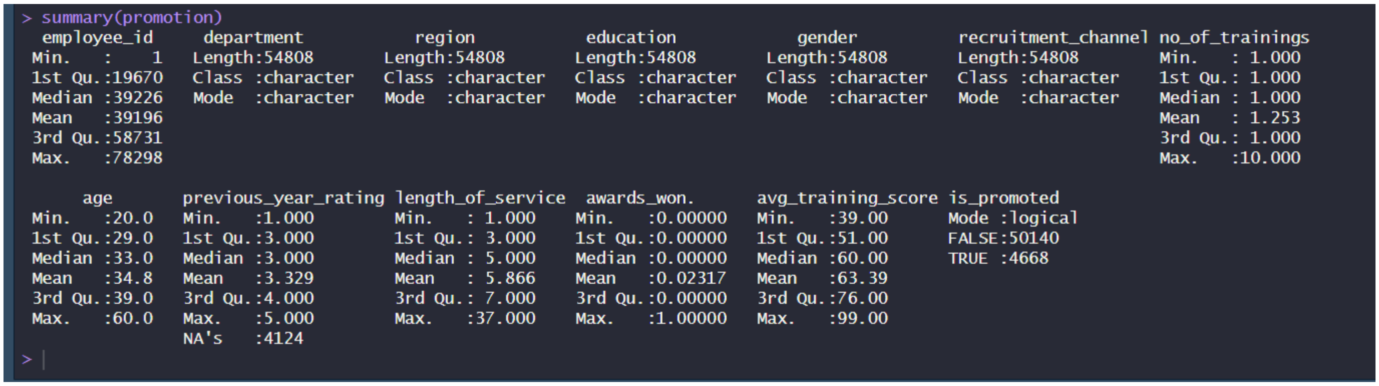
	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won.	avg_training_score	is_promoted
1	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5	8	0	49	FALSE
2	65141	Operations	region_22	Bachelor's	m	other	1	30	5	4	0	60	FALSE
3	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3	7	0	50	FALSE
4	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1	10	0	50	FALSE
5	48945	Technology	region_26	Bachelor's	m	other	1	45	3	2	0	73	FALSE
6	58896	Analytics	region_2	Bachelor's	m	sourcing	2	31	3	7	0	85	FALSE
7	20379	Operations	region_20	Bachelor's	f	other	1	31	3	5	0	59	FALSE
8	16290	Operations	region_34	Master's & above	m	sourcing	1	33	3	6	0	63	FALSE
9	73202	Analytics	region_20	Bachelor's	m	other	1	28	4	5	0	83	FALSE
10	28911	Sales & Marketing	region_1	Master's & above	m	sourcing	1	32	5	5	0	54	FALSE
11	29934	Technology	region_23	NA	m	sourcing	1	30	NA	1	0	77	FALSE
12	49017	Sales & Marketing	region_7	Bachelor's	f	sourcing	1	35	5	3	0	50	TRUE
13	60051	Sales & Marketing	region_4	Bachelor's	m	sourcing	1	49	5	5	0	49	FALSE
14	38401	Technology	region_29	Master's & above	m	other	2	39	3	16	0	80	FALSE
15	77040	R&D	region_2	Master's & above	m	sourcing	1	37	3	7	0	84	FALSE
16	43931	Operations	region_7	Bachelor's	m	other	1	37	1	10	0	60	FALSE
17	7152	Technology	region_2	Bachelor's	m	other	1	38	3	5	0	77	FALSE
18	9403	Sales & Marketing	region_31	Bachelor's	m	other	1	34	1	4	0	51	FALSE
19	17436	Sales & Marketing	region_31	Bachelor's	m	other	1	34	5	8	0	46	FALSE
20	54461	Operations	region_15	Bachelor's	m	other	1	37	3	9	0	59	FALSE
21	12067	Procurement	region_14	Bachelor's	m	other	1	35	3	7	0	75	FALSE
22	33332	Operations	region_15	NA	m	sourcing	1	41	4	11	0	57	FALSE

head(promotion), tail(promotion): display the first or last few rows of your dataset, respectively.



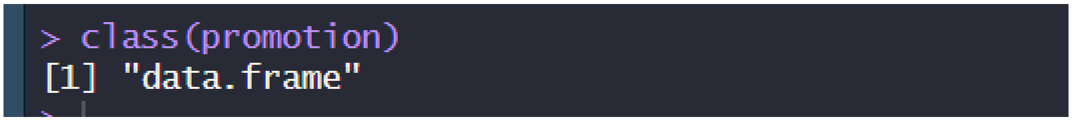
```
> head(promotion)
  employee_id        department    region        education gender recruitment_channel no_of_trainings age previous_year_rating
1       65438 Sales & Marketing  region_7 Master's & above      f            sourcing               1  35                    5
2       65141        Operations region_22       Bachelor's      m               other               1  30                    5
3        7513 Sales & Marketing region_19       Bachelor's      m            sourcing               1  34                    3
4        2542 Sales & Marketing region_23       Bachelor's      m               other               2  39                    1
5       48945        Technology region_26       Bachelor's      m               other               1  45                    3
6       58896         Analytics  region_2       Bachelor's      m            sourcing               2  31                    3
  length_of_service awards_won. avg_training_score is_promoted
1                 8           0                 49       FALSE
2                 4           0                 60       FALSE
3                 7           0                 50       FALSE
4                10           0                 50       FALSE
5                 2           0                 73       FALSE
6                 7           0                 85       FALSE
> tail(promotion)
      employee_id        department    region        education gender recruitment_channel no_of_trainings age previous_year_rating
54803        6915 Sales & Marketing region_14       Bachelor's      m               other               2  31                    1
54804        3030        Technology region_14       Bachelor's      m            sourcing               1  48                    3
54805       74592        Operations region_27 Master's & above      f               other               1  37                    2
54806       13918         Analytics  region_1       Bachelor's      m               other               1  27                    5
54807       13614 Sales & Marketing  region_9             <NA>      m            sourcing               1  29                    1
54808       51526                HR region_22       Bachelor's      m               other               1  27                    1
      length_of_service awards_won. avg_training_score is_promoted
54803                 2           0                 49       FALSE
54804                17           0                 78       FALSE
54805                 6           0                 56       FALSE
54806                 3           0                 79       FALSE
54807                 2           0                 45       FALSE
54808                 5           0                 49       FALSE
> 
```

summary(promotion): provides a summary of the central tendency, dispersion, and shape of the distribution of a dataset's variable**s.**



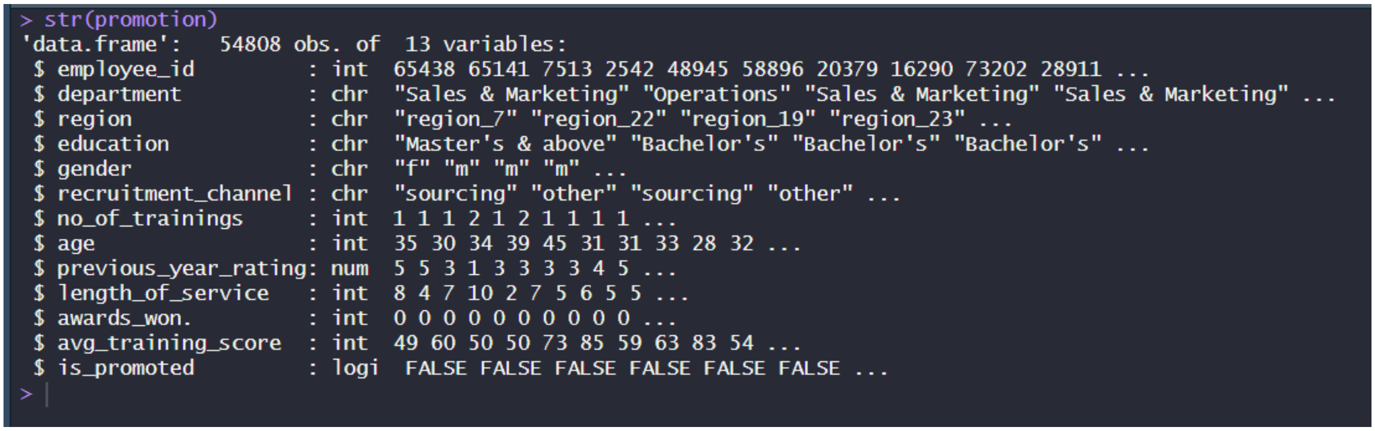
```
> summary(promotion)
  employee_id     department           region           education            gender          recruitment_channel no_of_trainings
 Min.   :    1   Length:54808       Length:54808       Length:54808       Length:54808       Length:54808        Min.   : 1.000 
 1st Qu.:19670   Class :character   Class :character   Class :character   Class :character   Class :character    1st Qu.: 1.000 
 Median :39226   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character    Median : 1.000 
 Mean   :39196                                                                                                   Mean   : 1.253 
 3rd Qu.:58731                                                                                                   3rd Qu.: 1.000 
 Max.   :78298                                                                                                   Max.   :10.000 
                                                                                                                                
      age        previous_year_rating length_of_service  awards_won.      avg_training_score is_promoted    
 Min.   :20.0   Min.   :1.000        Min.   : 1.000    Min.   :0.00000   Min.   :39.00      Mode :logical  
 1st Qu.:29.0   1st Qu.:3.000        1st Qu.: 3.000    1st Qu.:0.00000   1st Qu.:51.00      FALSE:50140    
 Median :33.0   Median :3.000        Median : 5.000    Median :0.00000   Median :60.00      TRUE :4668     
 Mean   :34.8   Mean   :3.329        Mean   : 5.866    Mean   :0.02317   Mean   :63.39                     
 3rd Qu.:39.0   3rd Qu.:4.000        3rd Qu.: 7.000    3rd Qu.:0.00000   3rd Qu.:76.00                     
 Max.   :60.0   Max.   :5.000        Max.   :37.000    Max.   :1.00000   Max.   :99.00                     
                NA's   :4124                                                                               
> 
```

Class(promotion): show the data type of our variable.



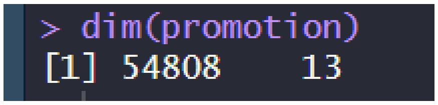
```
> class(promotion)
[1] "data.frame"
```

str(promotion): provides information about the structure of the dataset, including the data types of each variable and total of observations and variables.



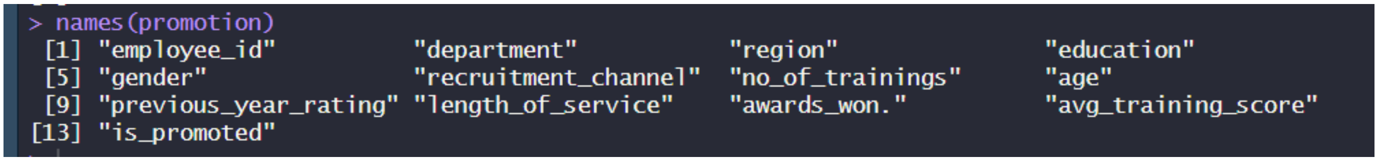
```
> str(promotion)
'data.frame':	54808 obs. of  13 variables:
 $ employee_id         : int  65438 65141 7513 2542 48945 58896 20379 16290 73202 28911 ...
 $ department          : chr  "Sales & Marketing" "Operations" "Sales & Marketing" "Sales & Marketing" ...
 $ region              : chr  "region_7" "region_22" "region_19" "region_23" ...
 $ education           : chr  "Master's & above" "Bachelor's" "Bachelor's" "Bachelor's" ...
 $ gender              : chr  "f" "m" "m" "m" ...
 $ recruitment_channel : chr  "sourcing" "other" "sourcing" "other" ...
 $ no_of_trainings     : int  1 1 1 2 1 2 1 1 1 1 ...
 $ age                 : int  35 30 34 39 45 31 31 33 28 32 ...
 $ previous_year_rating: num  5 5 3 1 3 3 3 3 4 5 ...
 $ length_of_service   : int  8 4 7 10 2 7 5 6 5 5 ...
 $ awards_won.         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ avg_training_score  : int  49 60 50 50 73 85 59 63 83 54 ...
 $ is_promoted         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
> 
```

dim(promotion): used to get the dimensions (number of rows and columns) of the dataset.



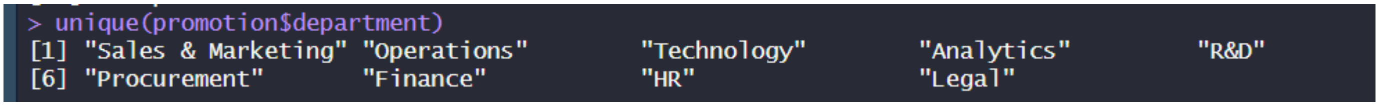
```
> dim(promotion)
[1] 54808    13
```

names(promotion): use to display the variable names(columns) in the dataset.



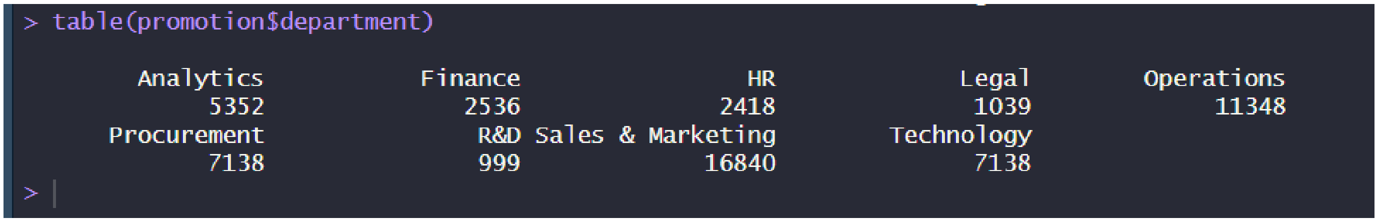
```
> names(promotion)
 [1] "employee_id"          "department"          "region"               "education"
 [5] "gender"               "recruitment_channel" "no_of_trainings"      "age"
 [9] "previous_year_rating" "length_of_service"   "awards_won."          "avg_training_score"
[13] "is_promoted"
```

unique(promotion$variable): display all possible values in specified variable, it be useful with categorical data.



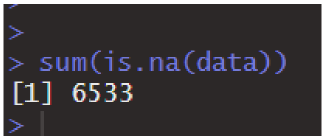
```
> unique(promotion$department)
[1] "Sales & Marketing" "Operations"        "Technology"        "Analytics"         "R&D"
[6] "Procurement"       "Finance"           "HR"                "Legal"
```

table(promotion$variable): display frequency of each value in specified variable.



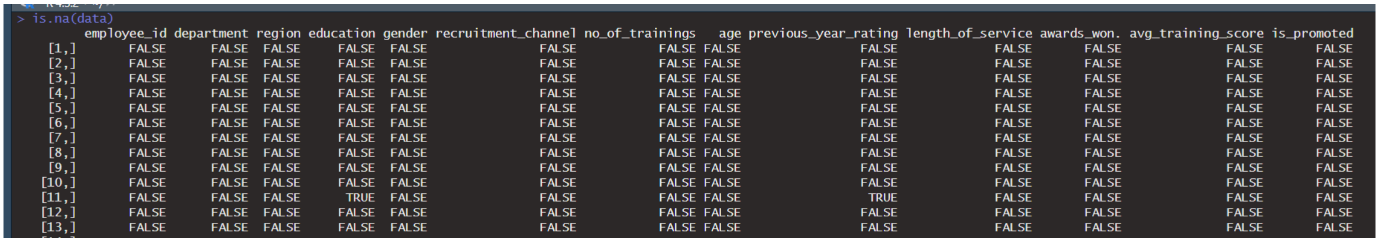
```
> table(promotion$department)

        Analytics           Finance                HR             Legal        Operations 
             5352              2536              2418              1039             11348 
      Procurement               R&D Sales & Marketing        Technology 
             7138               999             16840              7138 
> 
```

sum(is. na(data) : How many NA values are in data file?



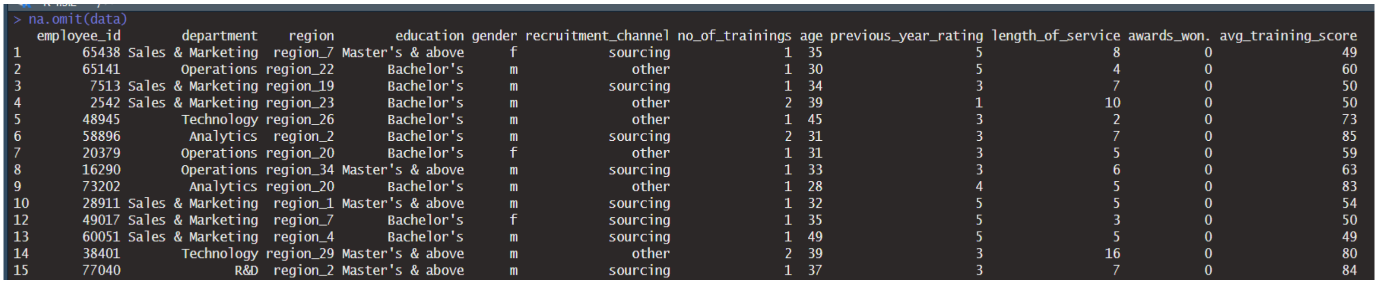
```
>
> sum(is.na(data))
[1] 6533
>
```

is.na (data): are there any NA values in data file, if so which row?



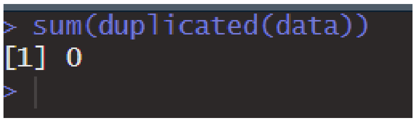
```
> is.na(data)
      employee_id department region education gender recruitment_channel no_of_trainings   age previous_year_rating length_of_service awards_won. avg_training_score is_promoted
 [1,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [2,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [3,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [4,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [5,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [6,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [7,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [8,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
 [9,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
[10,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
[11,]       FALSE      FALSE  FALSE      TRUE  FALSE               FALSE           FALSE FALSE                 TRUE             FALSE       FALSE              FALSE       FALSE
[12,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
[13,]       FALSE      FALSE  FALSE     FALSE  FALSE               FALSE           FALSE FALSE                FALSE             FALSE       FALSE              FALSE       FALSE
```

cleaned_data <- na.omit(promotion) : Handle the missing value with deleting their rows



```
> na.omit(data)
   employee_id        department    region        education gender recruitment_channel no_of_trainings age previous_year_rating length_of_service awards_won. avg_training_score
1        65438 Sales & Marketing  region_7 Master's & above      f            sourcing               1  35                    5                 8           0                 49
2        65141        Operations region_22       Bachelor's      m               other               1  30                    5                 4           0                 60
3         7513 Sales & Marketing region_19       Bachelor's      m            sourcing               1  34                    3                 7           0                 50
4         2542 Sales & Marketing region_23       Bachelor's      m               other               2  39                    1                10           0                 50
5        48945        Technology region_26       Bachelor's      m               other               1  45                    3                 2           0                 73
6        58896         Analytics  region_2       Bachelor's      m            sourcing               2  31                    3                 7           0                 85
7        20379        Operations region_20       Bachelor's      f               other               1  31                    3                 5           0                 59
8        16290        Operations region_34 Master's & above      m            sourcing               1  33                    3                 6           0                 63
9        73202         Analytics region_20       Bachelor's      m               other               1  28                    4                 5           0                 83
10       28911 Sales & Marketing  region_1 Master's & above      m            sourcing               1  32                    5                 5           0                 54
12       49017 Sales & Marketing  region_7       Bachelor's      f            sourcing               1  35                    5                 3           0                 50
13       60051 Sales & Marketing  region_4       Bachelor's      m            sourcing               1  49                    5                 5           0                 49
14       38401        Technology region_29 Master's & above      m               other               2  39                    3                16           0                 80
15       77040               R&D  region_2 Master's & above      m            sourcing               1  37                    3                 7           0                 84
```

sum(duplicated()) : how many duplicated rows are in your data?



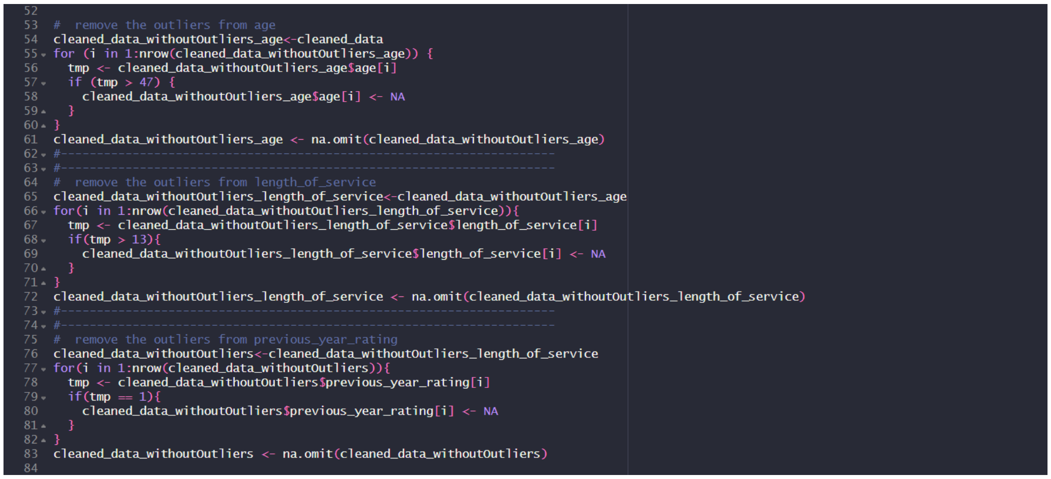
```
> sum(duplicated(data))
[1] 0
>
```

-there is not duplicated rows in data file then we need not to use distinct(data)

Now we need to remove the Outliers :

we apply the boxplot for each one , after we get the Max value and the Min value → we write aloop that check if the column`s value consider outliers then =NA

THEN we apply na.omit()method：



```r

#  remove the outliers from age
cleaned_data_withoutOutliers_age<-cleaned_data
for (i in 1:nrow(cleaned_data_withoutOutliers_age)) {
  tmp <- cleaned_data_withoutOutliers_age$age[i]
  if (tmp > 47) {
    cleaned_data_withoutOutliers_age$age[i] <- NA
  }
}
cleaned_data_withoutOutliers_age <- na.omit(cleaned_data_withoutOutliers_age)
#-------------------------------------------------------------------
#-------------------------------------------------------------------
#  remove the outliers from length_of_service
cleaned_data_withoutOutliers_length_of_service<-cleaned_data_withoutOutliers_age
for(i in 1:nrow(cleaned_data_withoutOutliers_length_of_service)){
  tmp <- cleaned_data_withoutOutliers_length_of_service$length_of_service[i]
  if(tmp > 13){
    cleaned_data_withoutOutliers_length_of_service$length_of_service[i] <- NA
  }
}
cleaned_data_withoutOutliers_length_of_service <- na.omit(cleaned_data_withoutOutliers_length_of_service)
#-------------------------------------------------------------------
#-------------------------------------------------------------------
#  remove the outliers from previous_year_rating
cleaned_data_withoutOutliers<-cleaned_data_withoutOutliers_length_of_service
for(i in 1:nrow(cleaned_data_withoutOutliers)){
  tmp <- cleaned_data_withoutOutliers$previous_year_rating[i]
  if(tmp == 1){
    cleaned_data_withoutOutliers$previous_year_rating[i] <- NA
  }
}
cleaned_data_withoutOutliers <- na.omit(cleaned_data_withoutOutliers)

```

“cleaned _data_withoutOutliers” This is the data that we will use in the rest of the code, as it is free of outliers and empty values

3. Challenges in the dataset:

In this section you should write the difficulties and challenges you face while working on your dataset.

- First challenge we met when we import the dataset the variables with data type <character> if the any cell was null it was like that "" not NA so it not considered null value and that obviously will give us misinformation in exploration & cleaning

- The second challenge is the outliers...how to make the data free of outliers. So we resorted to creating a boxplot on the data (free of empty values and repetitions) and studying the outliers for each column and deleting them from all the data. After several attempts, we succeeded in getting rid of the outliers.

- The third challenge is how to represent the k-means clusters graphically. After research, we came up with a method that helps us represent the clusters graphically (by changing the required columns, the result changes).

- Final challenge is the UI It took us a lot of time and more effort to search for the names of the appropriate buttons and how to use each of them, and try to apply them in a way that suits the project, and try to link the original code (the names of its variables) with the server and UI.
  This was the first time we used UI. Unfortunately, we were not able to enjoy the experience because time was limited and running out quickly.

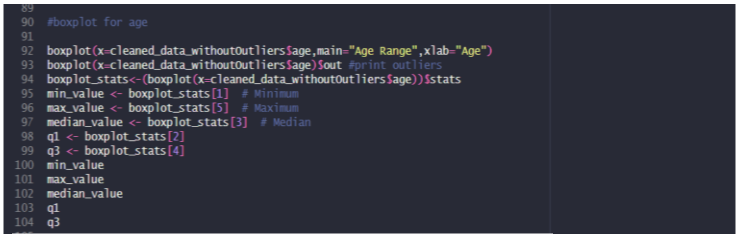
4. Interpretations of the results

In this section you should write the results, its explanation and show the plotted graphs.

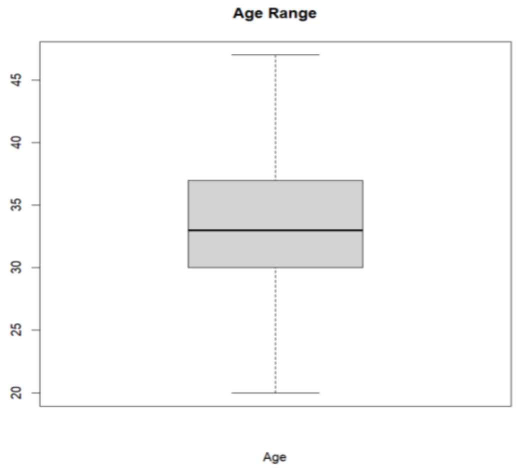
Visulasation :

- **Box Plot :**

For Age :

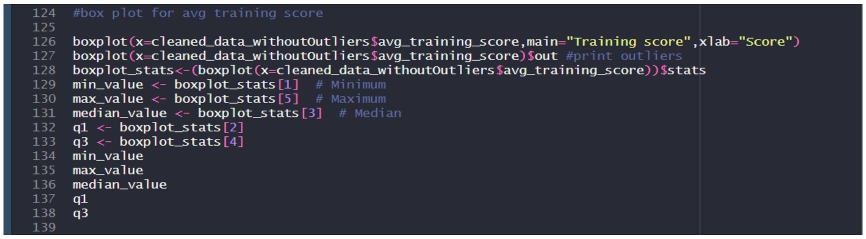


```
#boxplot for age

boxplot(x=cleaned_data_withoutOutliers$age,main="Age Range",xlab="Age")
boxplot(x=cleaned_data_withoutOutliers$age)$out #print outliers
boxplot_stats<-(boxplot(x=cleaned_data_withoutOutliers$age))$stats
min_value <- boxplot_stats[1]  # Minimum
max_value <- boxplot_stats[5]  # Maximum
median_value <- boxplot_stats[3]  # Median
q1 <- boxplot_stats[2]
q3 <- boxplot_stats[4]
min_value
max_value
median_value
q1
q3
```

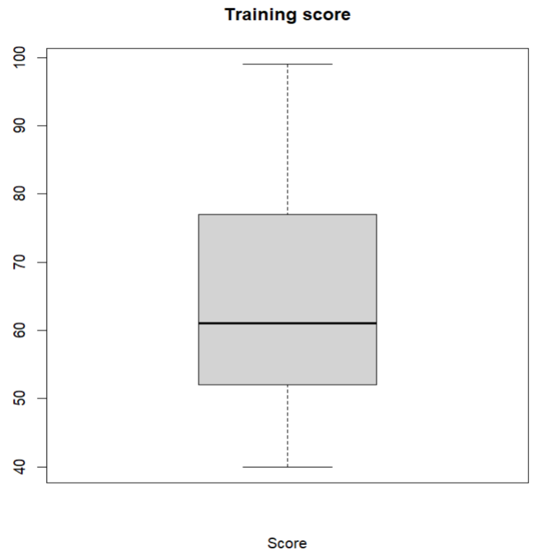


```
> min_value
[1] 20
> max_value
[1] 47
> median_value
[1] 33
> q1
[1] 30
> q3
[1] 37
```

For Training Score :

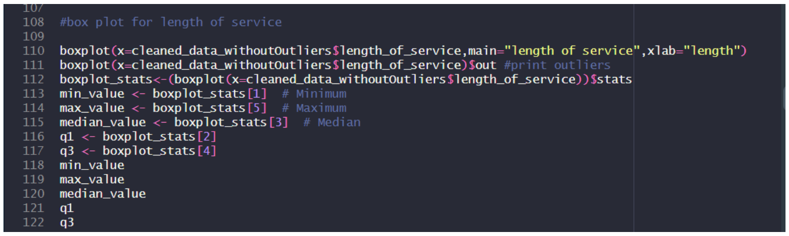


```r
#box plot for avg training score

boxplot(x=cleaned_data_withoutOutliers$avg_training_score,main="Training score",xlab="Score")
boxplot(x=cleaned_data_withoutOutliers$avg_training_score)$out #print outliers
boxplot_stats<-(boxplot(x=cleaned_data_withoutOutliers$avg_training_score))$stats
min_value <- boxplot_stats[1]  # Minimum
max_value <- boxplot_stats[5]  # Maximum
median_value <- boxplot_stats[3]  # Median
q1 <- boxplot_stats[2]
q3 <- boxplot_stats[4]
min_value
max_value
median_value
q1
q3

```

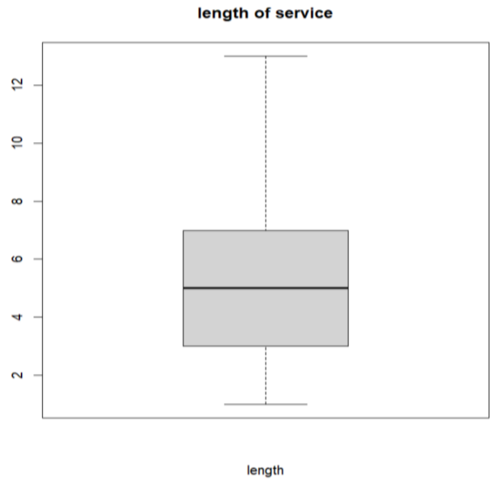


```
> min_value
[1] 40
> max_value
[1] 99
> median_value
[1] 61
> q1
[1] 52
> q3
[1] 77
```

For Length Of Service :



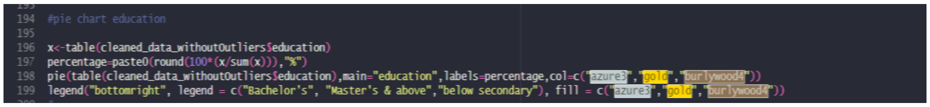
```
#box plot for length of service

boxplot(x=cleaned_data_withoutOutliers$length_of_service,main="length of service",xlab="length")
boxplot(x=cleaned_data_withoutOutliers$length_of_service)$out #print outliers
boxplot_stats<-(boxplot(x=cleaned_data_withoutOutliers$length_of_service))$stats
min_value <- boxplot_stats[1]  # Minimum
max_value <- boxplot_stats[5]  # Maximum
median_value <- boxplot_stats[3]  # Median
q1 <- boxplot_stats[2]
q3 <- boxplot_stats[4]
min_value
max_value
median_value
q1
q3
```



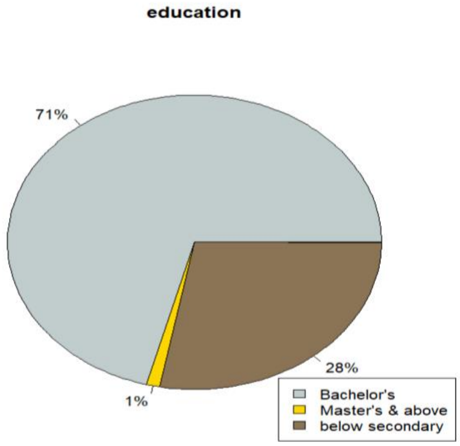
```
> min_value
[1] 1
> max_value
[1] 13
> median_value
[1] 5
> q1
[1] 3
> q3
[1] 7
```

• Pie chart :

For Education :

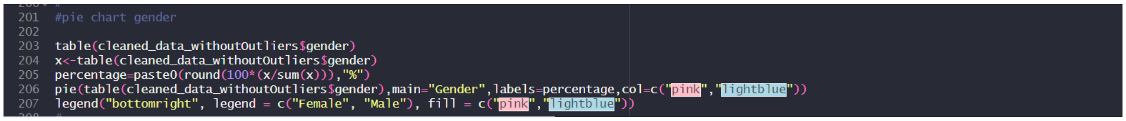


```r
#pie chart education

x<-table(cleaned_data_withoutOutliers$education)
percentage=paste0(round(100*(x/sum(x))),"%")
pie(table(cleaned_data_withoutOutliers$education),main="education",labels=percentage,col=c("azure3","gold","burlywood4"))
legend("bottomright", legend = c("Bachelor's", "Master's & above","below secondary"), fill = c("azure3","gold","burlywood4"))
```

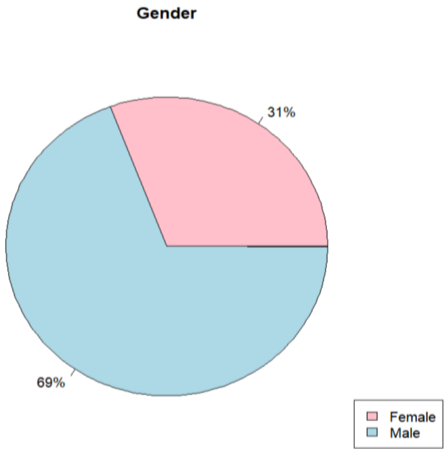


- The percentage of Bachelor’s is the biggest
- The percentage of below secondary is the lowest

For Gender :

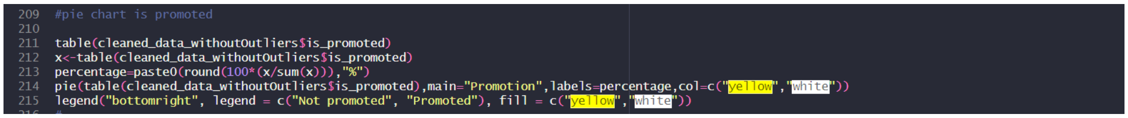


```r
#pie chart gender

table(cleaned_data_withoutOutliers$gender)
x<-table(cleaned_data_withoutOutliers$gender)
percentage=paste0(round(100*(x/sum(x))),"%")
pie(table(cleaned_data_withoutOutliers$gender),main="Gender",labels=percentage,col=c("pink","lightblue"))
legend("bottomright", legend = c("Female", "Male"), fill = c("pink","lightblue"))
```

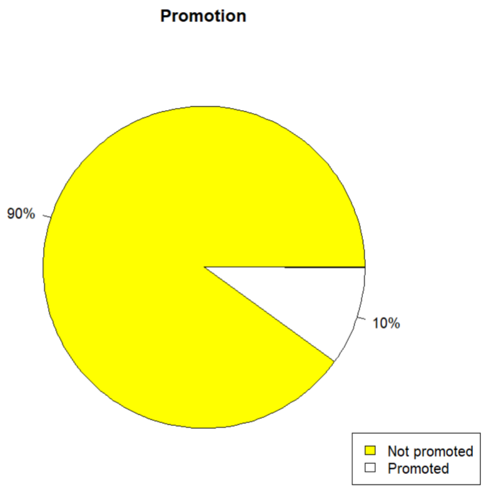


- The percentage of male is bigger than the percentage of female

For promotion :

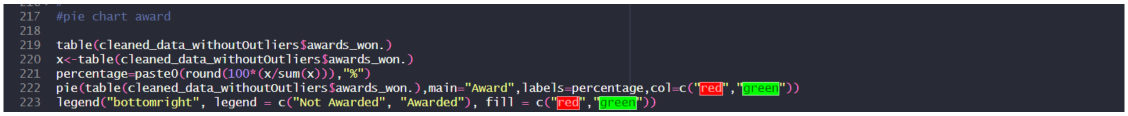


```r
#pie chart is promoted

table(cleaned_data_withoutOutliers$is_promoted)
x<-table(cleaned_data_withoutOutliers$is_promoted)
percentage=paste0(round(100*(x/sum(x))),"%")
pie(table(cleaned_data_withoutOutliers$is_promoted),main="Promotion",labels=percentage,col=c("yellow","white"))
legend("bottomright", legend = c("Not promoted", "Promoted"), fill = c("yellow","white"))
```

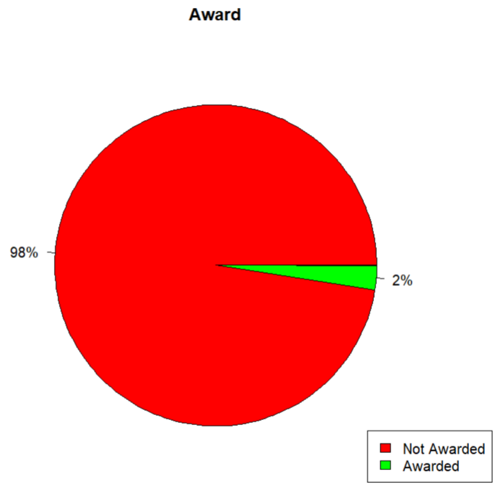


- The   percentage of being promoted is lower than the percentage of not being promoted

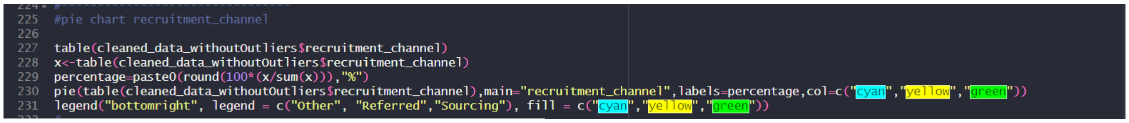
For award :



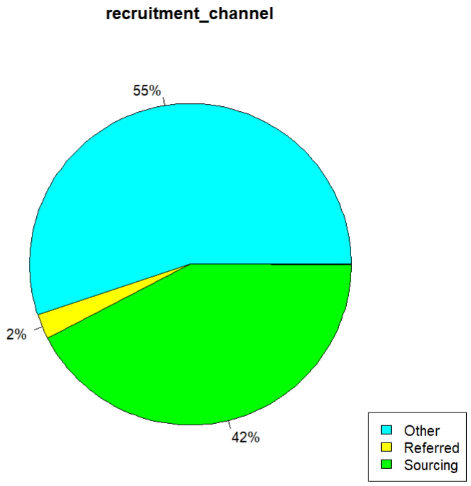
```r
#pie chart award

table(cleaned_data_withoutOutliers$awards_won.)
x<-table(cleaned_data_withoutOutliers$awards_won.)
percentage=paste0(round(100*(x/sum(x))),"%")
pie(table(cleaned_data_withoutOutliers$awards_won.),main="Award",labels=percentage,col=c("red","green"))
legend("bottomright", legend = c("Not Awarded", "Awarded"), fill = c("red","green"))
```



For recruitment_channel :

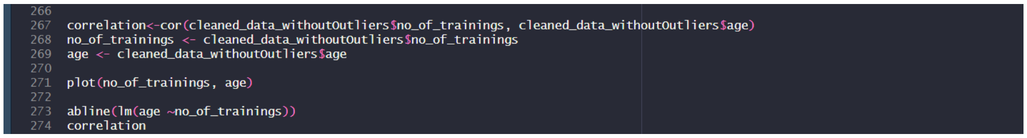


```
#pie chart recruitment_channel

table(cleaned_data_withoutOutliers$recruitment_channel)
x<-table(cleaned_data_withoutOutliers$recruitment_channel)
percentage=paste0(round(100*(x/sum(x))),"%")
pie(table(cleaned_data_withoutOutliers$recruitment_channel),main="recruitment_channel",labels=percentage,col=c("cyan","yellow","green"))
legend("bottomright", legend = c("Other", "Referred","Sourcing"), fill = c("cyan","yellow","green"))
```

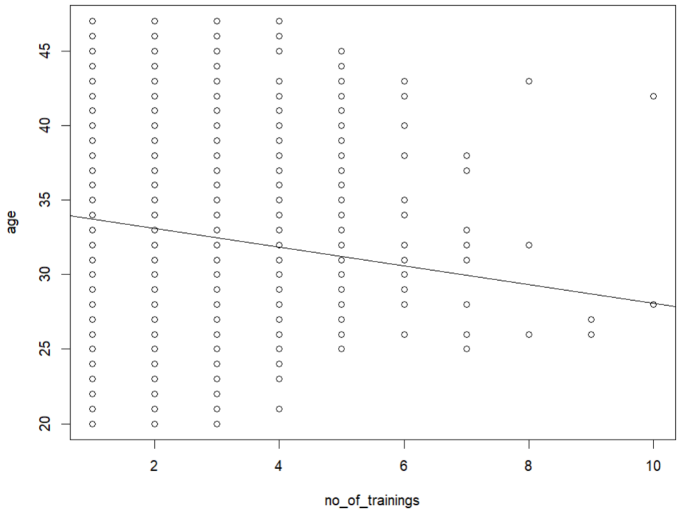


- ## Scatter Plot :

Between number of trainings and age :

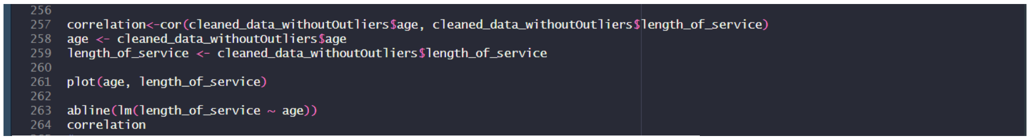


```
correlation<-cor(cleaned_data_withoutOutliers$no_of_trainings, cleaned_data_withoutOutliers$age)
no_of_trainings <- cleaned_data_withoutOutliers$no_of_trainings
age <- cleaned_data_withoutOutliers$age

plot(no_of_trainings, age)

abline(lm(age ~no_of_trainings))
correlation
```

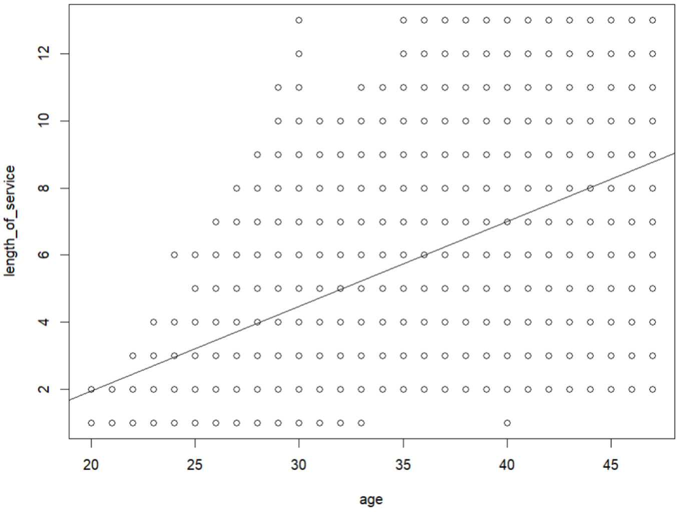


- From this graph we notice that number of trainings and age have a negative correlation

Between age and length of service :



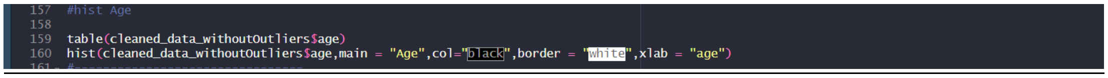
```
correlation<-cor(cleaned_data_withoutOutliers$age, cleaned_data_withoutOutliers$length_of_service)
age <- cleaned_data_withoutOutliers$age
length_of_service <- cleaned_data_withoutOutliers$length_of_service

plot(age, length_of_service)

abline(lm(length_of_service ~ age))
correlation
```



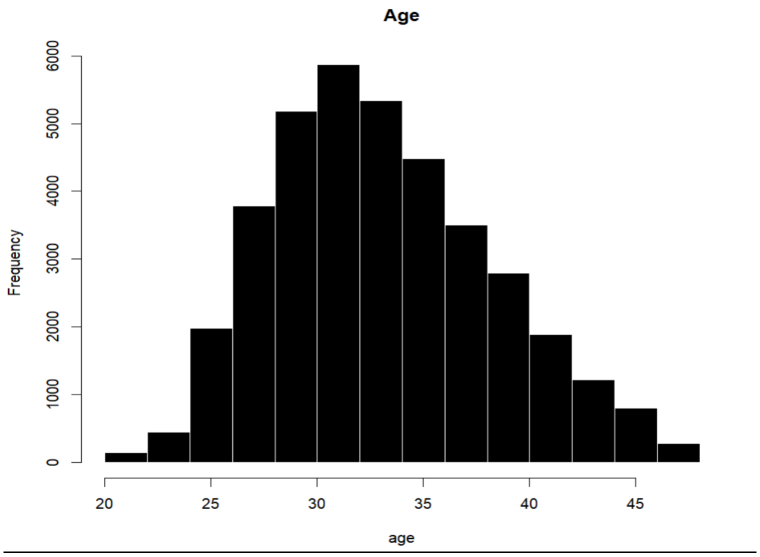
- From this graph we notice that number of trainigs and age have a positive correlation

Histogram :

For age :

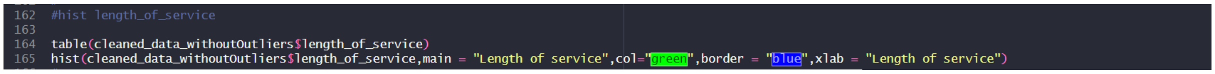


```
#hist Age

table(cleaned_data_withoutOutliers$age)
hist(cleaned_data_withoutOutliers$age,main = "Age",col="black",border = "white",xlab = "age")
```

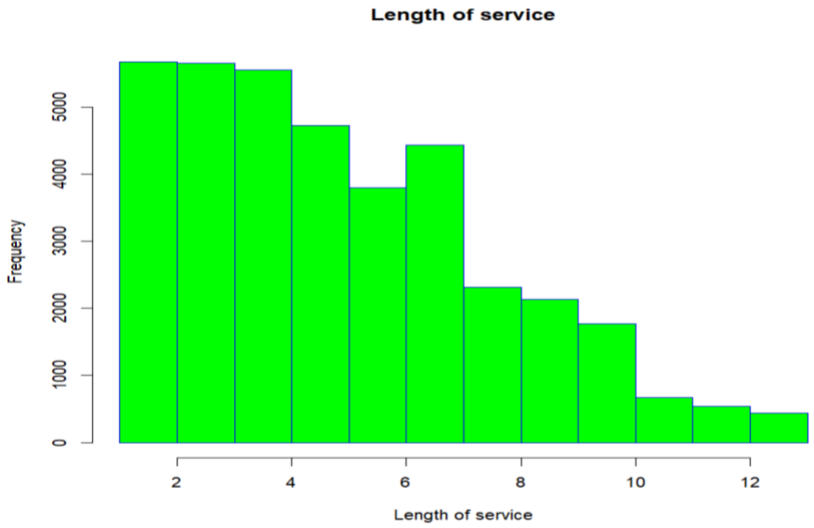


- From this histogram we notice that the most frequent age is between 30 and 32 years

For length of service :

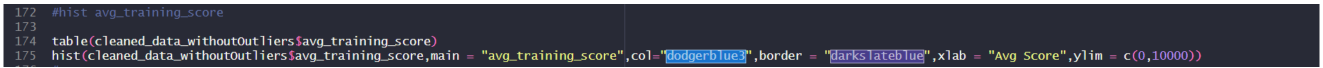


```
#hist length_of_service

table(cleaned_data_withoutOutliers$length_of_service)
hist(cleaned_data_withoutOutliers$length_of_service,main = "Length of service",col="green",border = "blue",xlab = "Length of service")
```

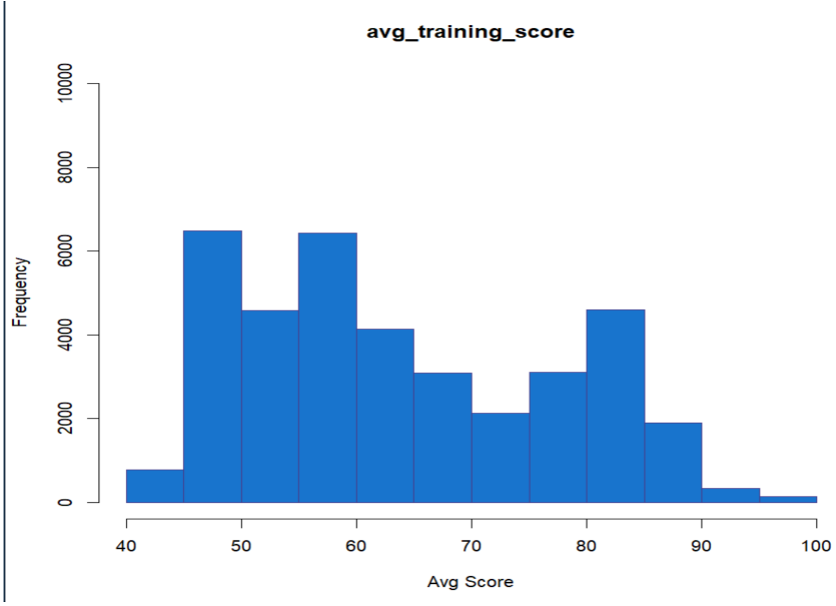


- From this histogram we notice that the most frequent length of service is between 0 and 4 years

For average training score :

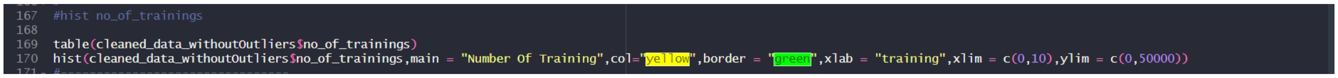


```
#hist avg_training_score

table(cleaned_data_withoutOutliers$avg_training_score)
hist(cleaned_data_withoutOutliers$avg_training_score,main = "avg_training_score",col="dodgerblue3",border = "darkslateblue",xlab = "Avg Score",ylim = c(0,10000))
```

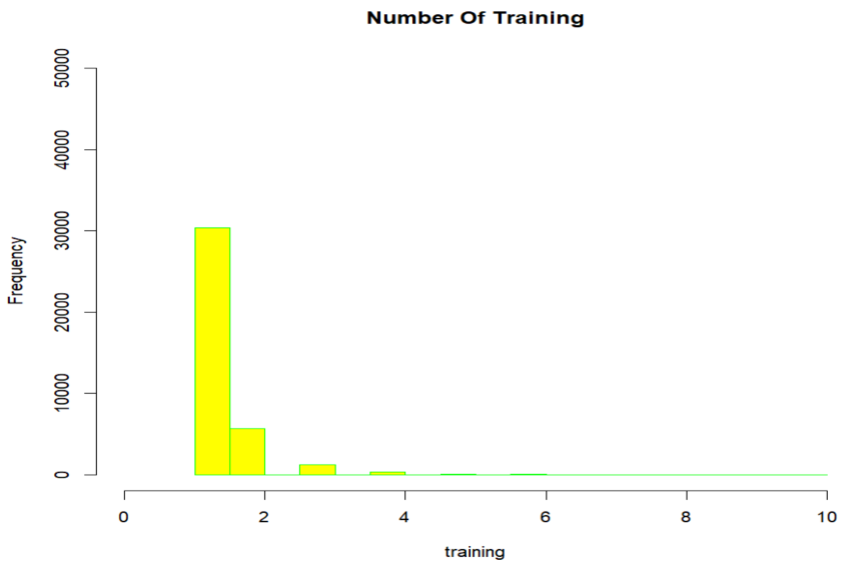


- From this histogram we notice that the most frequent average training score is between 45 and 50 , 55 and 60

For no_of_trainings :



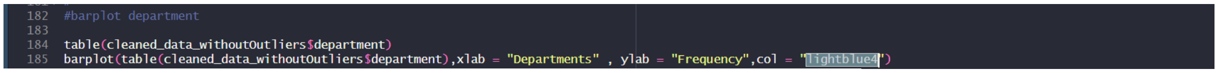
```
#hist no_of_trainings

table(cleaned_data_withoutOutliers$no_of_trainings)
hist(cleaned_data_withoutOutliers$no_of_trainings,main = "Number Of Training",col="yellow",border = "green",xlab = "training",xlim = c(0,10),ylim = c(0,50000))
```



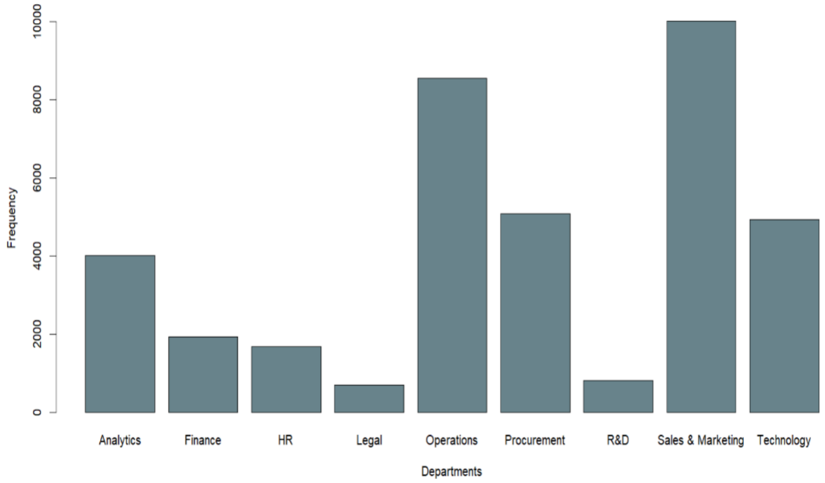
- From this histogram we notice that the most frequent number of training is 1

Bar Plot :

For department :

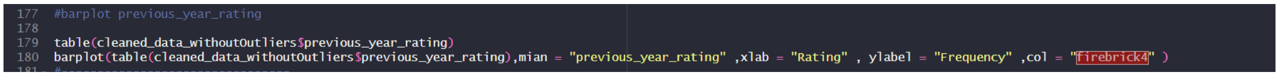


```
#barplot department

table(cleaned_data_withoutOutliers$department)
barplot(table(cleaned_data_withoutOutliers$department),xlab = "Departments" , ylab = "Frequency",col = "lightblue4")
```

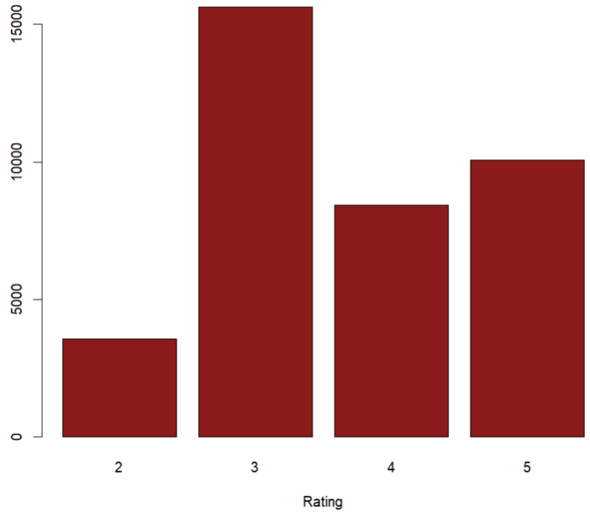


- from this bar plot we notice that the most department that has employees is sales & marketing dapartment and the least one is legal dapartment

For previous year rating :

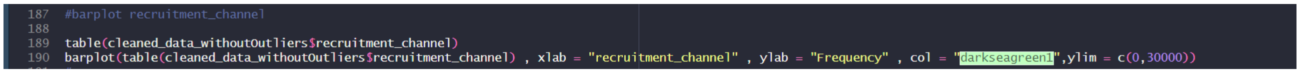


```
#barplot previous_year_rating

table(cleaned_data_withoutOutliers$previous_year_rating)
barplot(table(cleaned_data_withoutOutliers$previous_year_rating),mian = "previous_year_rating" ,xlab = "Rating" , ylabel = "Frequency" ,col = "firebrick4" )
```

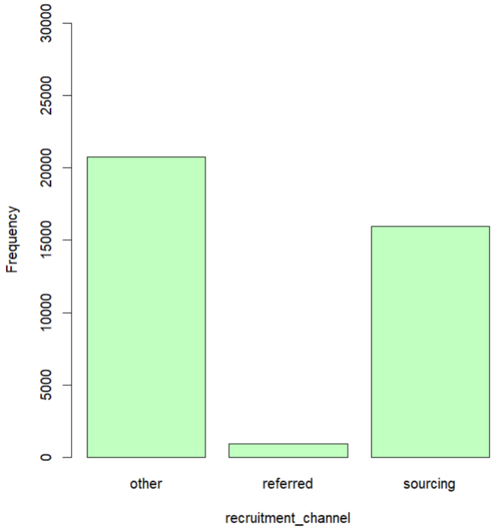


- from this bar plot we notice that the most employees have a previous year rating equal to 3

For recruitment channel :



```
#barplot recruitment_channel

table(cleaned_data_withoutOutliers$recruitment_channel)
barplot(table(cleaned_data_withoutOutliers$recruitment_channel) , xlab = "recruitment_channel" , ylab = "Frequency" , col = "darkseagreen1",ylim = c(0,30000))
```

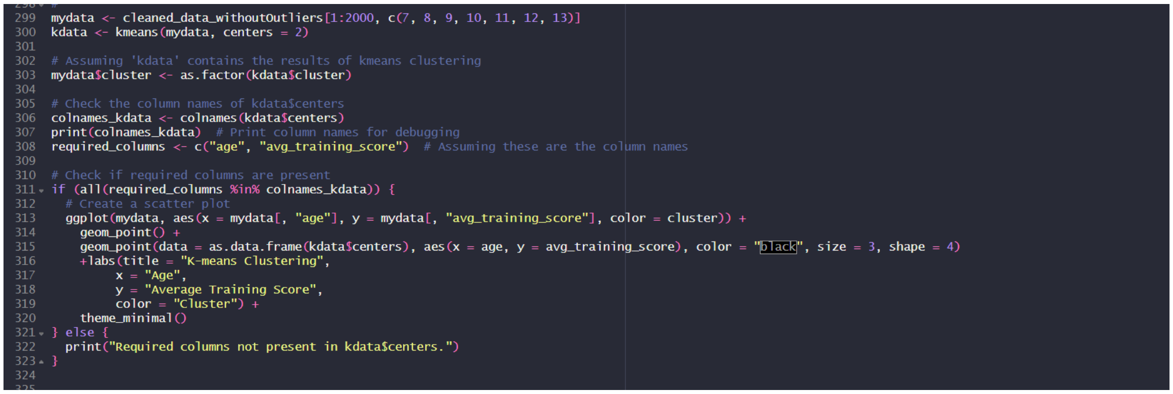


- Most of the employees came from an “other” recruitment channel

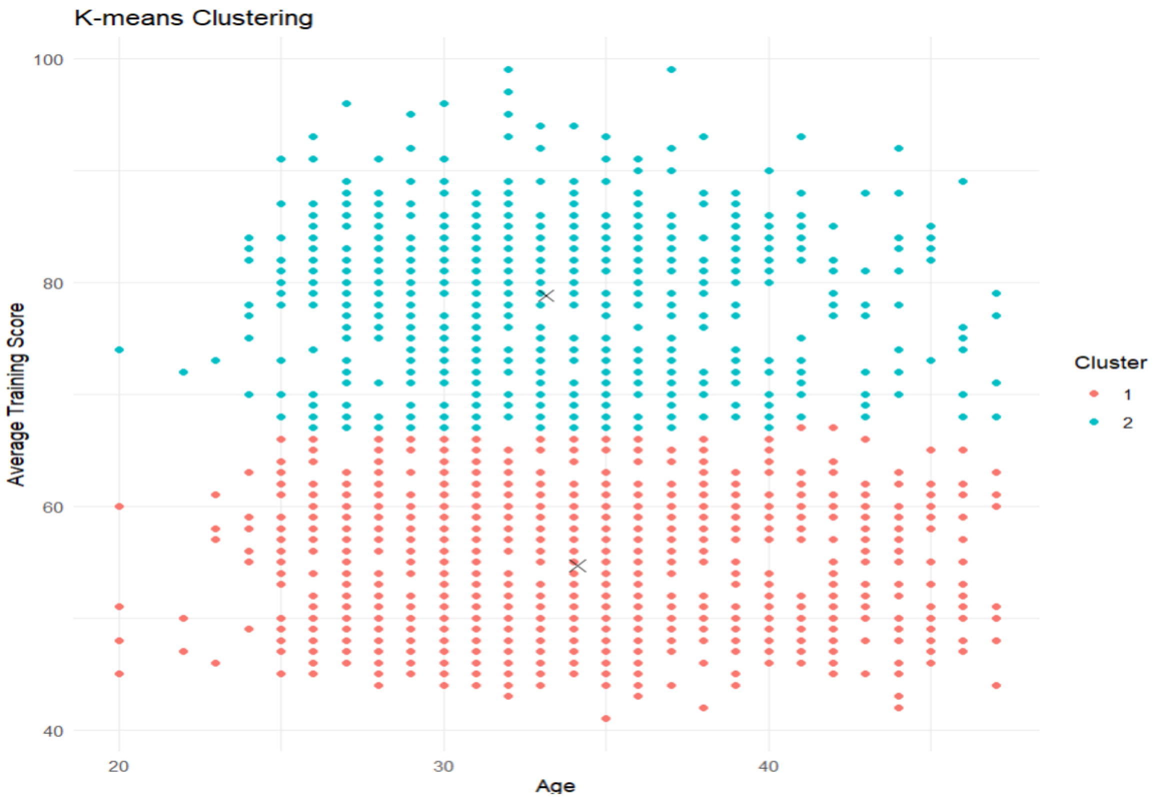
Analytics of data:

K-means Clustering:

Purpose: K-means clustering is an unsupervised machine learning algorithm used for clustering or grouping similar data points into k clusters. The goal is to partition the data into clusters in a way that data points within the same cluster are more similar to each other than to those in other clusters.

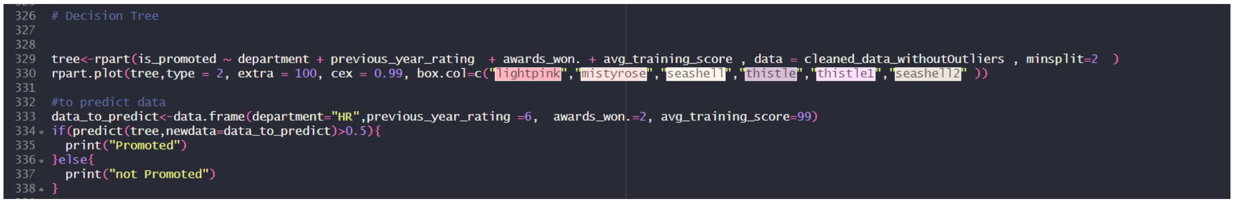


```r
mydata <- cleaned_data_withoutOutliers[1:2000, c(7, 8, 9, 10, 11, 12, 13)]
kdata <- kmeans(mydata, centers = 2)

# Assuming 'kdata' contains the results of kmeans clustering
mydata$cluster <- as.factor(kdata$cluster)

# Check the column names of kdata$centers
colnames_kdata <- colnames(kdata$centers)
print(colnames_kdata)  # Print column names for debugging
required_columns <- c("age", "avg_training_score")  # Assuming these are the column names

# Check if required columns are present
if (all(required_columns %in% colnames_kdata)) {
  # Create a scatter plot
  ggplot(mydata, aes(x = mydata[, "age"], y = mydata[, "avg_training_score"], color = cluster)) +
    geom_point() +
    geom_point(data = as.data.frame(kdata$centers), aes(x = age, y = avg_training_score), color = "black", size = 3, shape = 4)
    +labs(title = "K-means Clustering",
         x = "Age",
         y = "Average Training Score",
         color = "Cluster") +
    theme_minimal()
} else {
  print("Required columns not present in kdata$centers.")
}
```

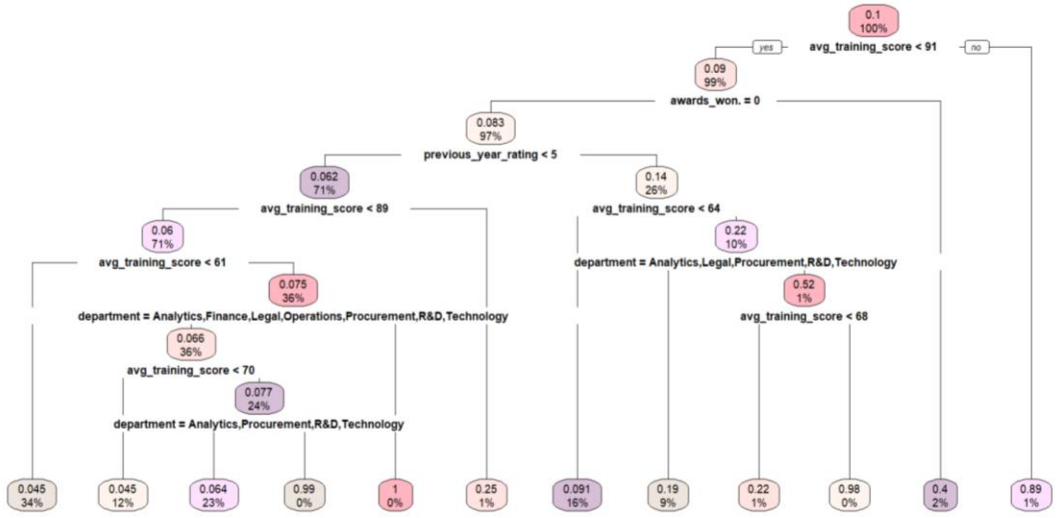


Decision Trees:

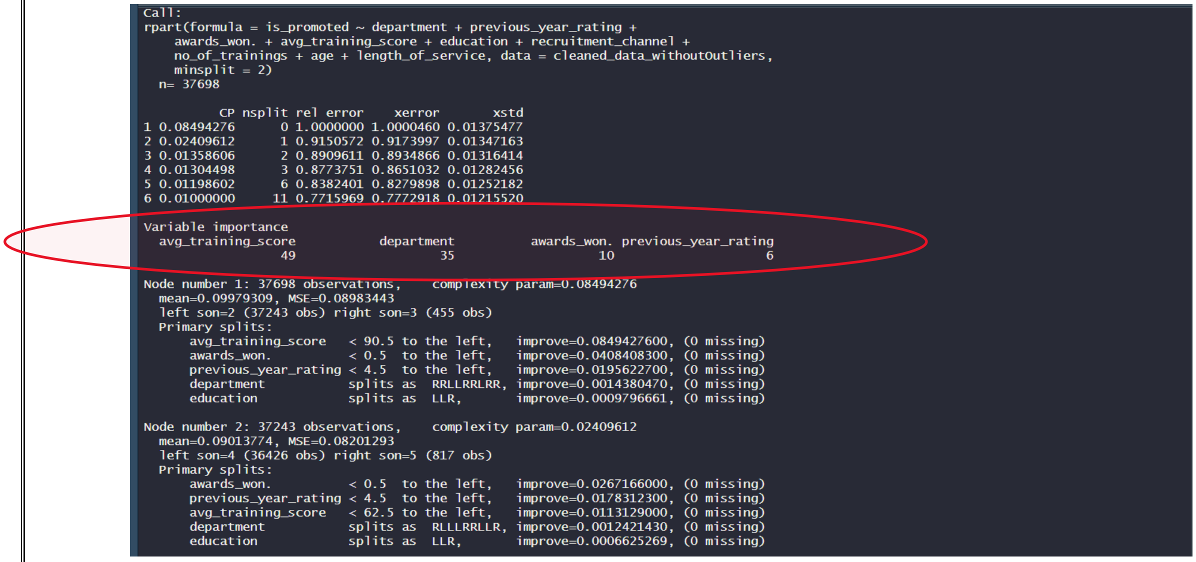
Purpose: Decision trees are a supervised machine learning algorithm used for both classification and regression tasks. They recursively split the data based on features to create a tree-like structure that makes decisions at each node.



```r
# Decision Tree


tree<-rpart(is_promoted ~ department + previous_year_rating  + awards_won. + avg_training_score , data = cleaned_data_withoutOutliers , minsplit=2  )
rpart.plot(tree,type = 2, extra = 100, cex = 0.99, box.col=c("lightpink","mistyrose","seashell","thistle","thistle1","seashell2" ))

#to predict data
data_to_predict<-data.frame(department="HR",previous_year_rating =6,  awards_won.=2, avg_training_score=99)
if(predict(tree,newdata=data_to_predict)>0.5){
  print("Promoted")
}else{
  print("not Promoted")
}
```



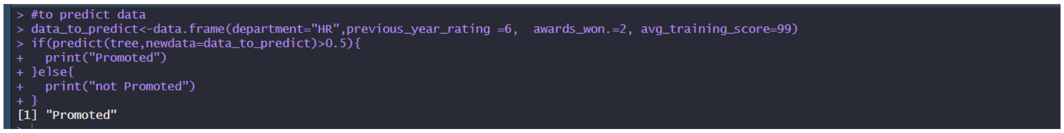
When we apply summary(tree) Method we got this :



```
Call:
rpart(formula = is_promoted ~ department + previous_year_rating +
    awards_won. + avg_training_score + education + recruitment_channel +
    no_of_trainings + age + length_of_service, data = cleaned_data_withoutOutliers,
    minsplit = 2)
  n= 37698

          CP nsplit rel error    xerror       xstd
1 0.08494276      0 1.0000000 1.0000460 0.01375477
2 0.02409612      1 0.9150572 0.9173997 0.01347163
3 0.01358606      2 0.8909611 0.8934866 0.01316414
4 0.01304498      3 0.8773751 0.8651032 0.01282456
5 0.01198602      6 0.8382401 0.8279898 0.01252182
6 0.01000000     11 0.7715969 0.7772918 0.01215520

Variable importance
  avg_training_score           department          awards_won. previous_year_rating
                  49                   35                   10                    6

Node number 1: 37698 observations,    complexity param=0.08494276
  mean=0.09979309, MSE=0.08983443
  left son=2 (37243 obs) right son=3 (455 obs)
  Primary splits:
      avg_training_score   < 90.5 to the left,   improve=0.0849427600, (0 missing)
      awards_won.          < 0.5  to the left,   improve=0.0408408300, (0 missing)
      previous_year_rating < 4.5  to the left,   improve=0.0195622700, (0 missing)
      department           splits as  RRLLRRLRR, improve=0.0014380470, (0 missing)
      education            splits as  LLR,       improve=0.0009796661, (0 missing)

Node number 2: 37243 observations,    complexity param=0.02409612
  mean=0.09013774, MSE=0.08201293
  left son=4 (36426 obs) right son=5 (817 obs)
  Primary splits:
      awards_won.          < 0.5  to the left,   improve=0.0267166000, (0 missing)
      previous_year_rating < 4.5  to the left,   improve=0.0178312300, (0 missing)
      avg_training_score   < 62.5 to the left,   improve=0.0113129000, (0 missing)
      department           splits as  RLLLRRLLR, improve=0.0012421430, (0 missing)
      education            splits as  LLR,       improve=0.0006625269, (0 missing)
```

Therefore, we modified the decision tree and remove the columns that had no effect on the upgrade decision

Example on data prediction :

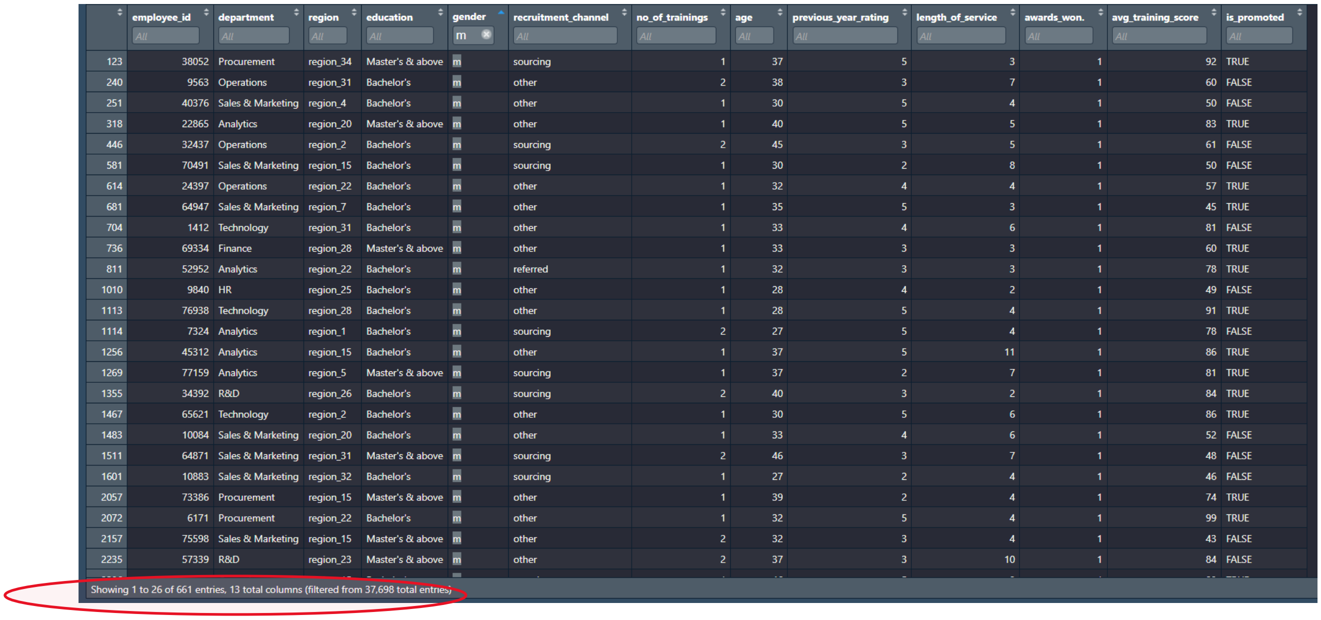


```
> #to predict data
> data_to_predict<-data.frame(department="HR",previous_year_rating =6,  awards_won.=2, avg_training_score=99)
> if(predict(tree,newdata=data_to_predict)>0.5){
+   print("Promoted")
+ }else{
+   print("not Promoted")
+ }
[1] "Promoted"
```

5. Conclusion:

After trying to find some statistics, we found that

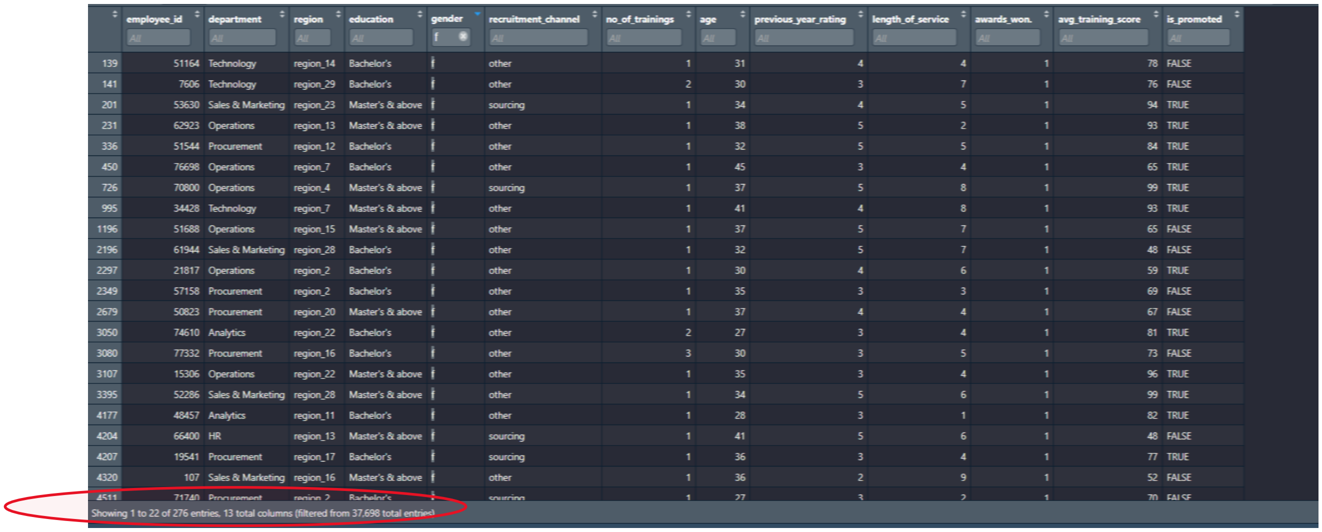
-The number of men receiving a reward is 661



	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won.	avg_training_score	is_promoted
	All	All	All	All	m	All	All	All	All	All	All	All	All
123	38052	Procurement	region_34	Master's & above	m	sourcing	1	37	5	3	1	92	TRUE
240	9563	Operations	region_31	Bachelor's	m	other	2	38	3	7	1	60	FALSE
251	40376	Sales & Marketing	region_4	Bachelor's	m	other	1	30	5	4	1	50	FALSE
318	22865	Analytics	region_20	Master's & above	m	other	1	40	5	5	1	83	TRUE
446	32437	Operations	region_2	Bachelor's	m	sourcing	2	45	3	5	1	61	FALSE
581	70491	Sales & Marketing	region_15	Bachelor's	m	sourcing	1	30	2	8	1	50	FALSE
614	24397	Operations	region_22	Bachelor's	m	other	1	32	4	4	1	57	TRUE
681	64947	Sales & Marketing	region_7	Bachelor's	m	other	1	35	5	3	1	45	TRUE
704	1412	Technology	region_31	Bachelor's	m	other	1	33	4	6	1	81	FALSE
736	69334	Finance	region_28	Master's & above	m	other	1	33	3	3	1	60	TRUE
811	52952	Analytics	region_22	Bachelor's	m	referred	1	32	3	3	1	78	TRUE
1010	9840	HR	region_25	Bachelor's	m	other	1	28	4	2	1	49	FALSE
1113	76938	Technology	region_28	Bachelor's	m	other	1	28	5	4	1	91	TRUE
1114	7324	Analytics	region_1	Bachelor's	m	sourcing	2	27	5	4	1	78	FALSE
1256	45312	Analytics	region_15	Bachelor's	m	other	1	37	5	11	1	86	TRUE
1269	77159	Analytics	region_5	Master's & above	m	sourcing	1	37	2	7	1	81	TRUE
1355	34392	R&D	region_26	Bachelor's	m	sourcing	2	40	3	2	1	84	TRUE
1467	65621	Technology	region_2	Bachelor's	m	other	1	30	5	6	1	86	TRUE
1483	10084	Sales & Marketing	region_20	Bachelor's	m	other	1	33	4	6	1	52	FALSE
1511	64871	Sales & Marketing	region_31	Master's & above	m	sourcing	2	46	3	7	1	48	FALSE
1601	10883	Sales & Marketing	region_32	Bachelor's	m	sourcing	1	27	2	4	1	46	FALSE
2057	73386	Procurement	region_15	Master's & above	m	other	1	39	2	4	1	74	TRUE
2072	6171	Procurement	region_22	Bachelor's	m	other	1	32	5	4	1	99	TRUE
2157	75598	Sales & Marketing	region_15	Master's & above	m	other	2	32	3	4	1	43	FALSE
2235	57339	R&D	region_23	Master's & above	m	other	2	37	3	10	1	84	FALSE

Showing 1 to 26 of 661 entries, 13 total columns (filtered from 37,698 total entries)

And The number of women receiving a reward is 276

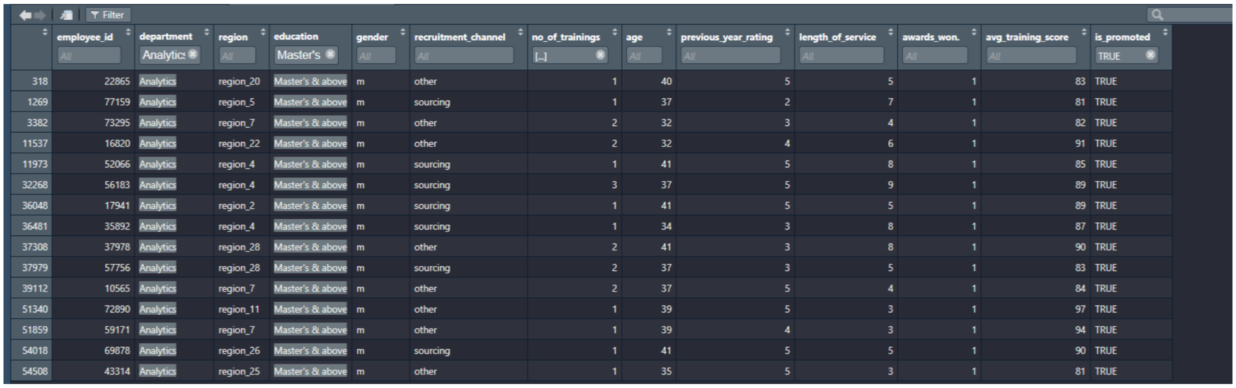


	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won.	avg_training_score	is_promoted
	All	All	All	All	f	All	All	All	All	All	All	All	All
139	51164	Technology	region_14	Bachelor's	f	other	1	31	4	4	1	78	FALSE
141	7606	Technology	region_29	Bachelor's	f	other	2	30	3	7	1	76	FALSE
201	53630	Sales & Marketing	region_23	Master's & above	f	sourcing	1	34	4	5	1	94	TRUE
231	62923	Operations	region_13	Master's & above	f	other	1	38	5	2	1	93	TRUE
336	51544	Procurement	region_12	Bachelor's	f	other	1	32	5	5	1	84	TRUE
450	76698	Operations	region_7	Bachelor's	f	other	1	45	3	4	1	65	TRUE
726	70800	Operations	region_4	Master's & above	f	sourcing	1	37	5	8	1	99	TRUE
995	34428	Technology	region_7	Master's & above	f	other	1	41	4	8	1	93	TRUE
1196	51688	Operations	region_15	Master's & above	f	other	1	37	5	7	1	65	FALSE
2196	61944	Sales & Marketing	region_28	Bachelor's	f	other	1	32	5	7	1	48	FALSE
2297	21817	Operations	region_2	Bachelor's	f	other	1	30	4	6	1	59	TRUE
2349	57158	Procurement	region_2	Bachelor's	f	other	1	35	3	3	1	69	FALSE
2679	50823	Procurement	region_20	Master's & above	f	other	1	37	4	4	1	67	FALSE
3050	74610	Analytics	region_22	Bachelor's	f	other	2	27	3	4	1	81	TRUE
3080	77332	Procurement	region_16	Bachelor's	f	other	3	30	3	5	1	73	FALSE
3107	15306	Operations	region_22	Master's & above	f	other	1	35	3	4	1	96	TRUE
3395	52286	Sales & Marketing	region_28	Master's & above	f	other	1	34	5	6	1	99	TRUE
4177	48457	Analytics	region_11	Bachelor's	f	other	1	28	3	1	1	82	TRUE
4204	66400	HR	region_13	Master's & above	f	sourcing	1	41	5	6	1	48	FALSE
4207	19541	Procurement	region_17	Bachelor's	f	sourcing	1	36	3	4	1	77	TRUE
4320	107	Sales & Marketing	region_16	Master's & above	f	other	1	36	2	9	1	52	FALSE
4511	71740	Procurement	region_2	Bachelor's	f	sourcing	1	27	3	2	1	70	FALSE

Showing 1 to 22 of 276 entries, 13 total columns (filtered from 37,698 total entries)

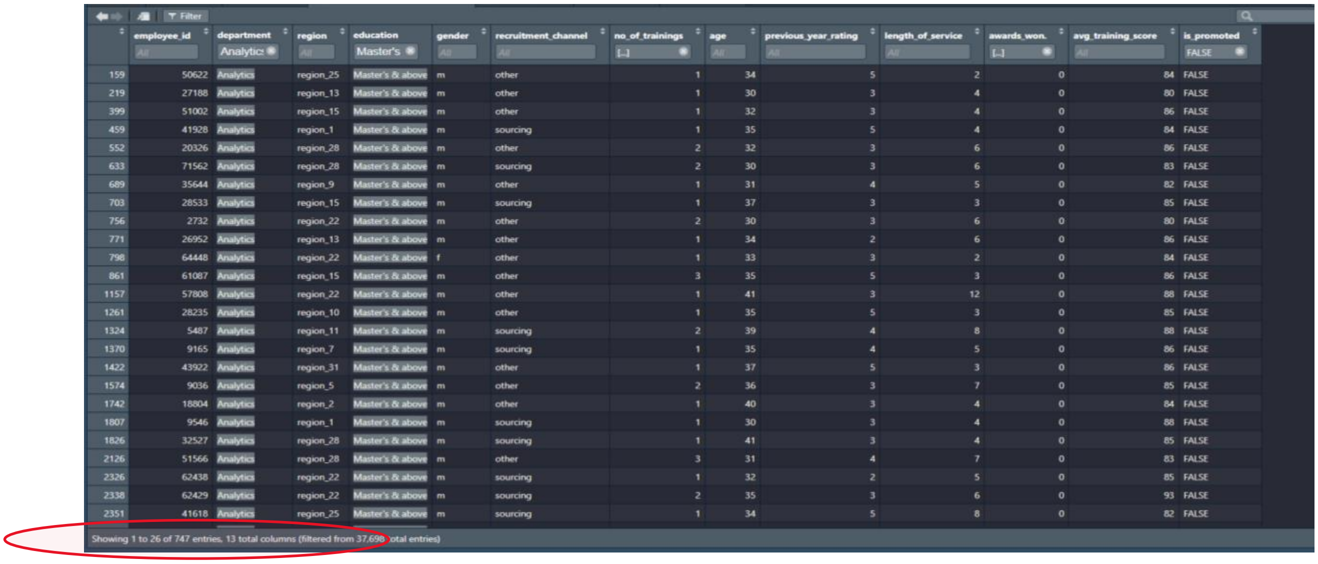
We conclude this that the number of males have more prizes from female awards and the male number is generally in our data greater than females (this is expected)

- The number of people working in the Analytics Department, whose education is master`s & above, and who have won awards and promotions is 15



	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won.	avg_training_score	is_promoted
	All	Analytics	All	Master's	All	All	[...]	All	All	All	All	All	TRUE
318	22865	Analytics	region_20	Master's & above	m	other	1	40	5	5	1	83	TRUE
1269	77159	Analytics	region_5	Master's & above	m	sourcing	1	37	2	7	1	81	TRUE
3382	73295	Analytics	region_7	Master's & above	m	other	2	32	3	4	1	82	TRUE
11537	16820	Analytics	region_22	Master's & above	m	other	2	32	4	6	1	91	TRUE
11973	52066	Analytics	region_4	Master's & above	m	sourcing	1	41	5	8	1	85	TRUE
32268	56183	Analytics	region_4	Master's & above	m	sourcing	3	37	5	9	1	89	TRUE
36048	17941	Analytics	region_2	Master's & above	m	sourcing	1	41	5	5	1	89	TRUE
36481	35892	Analytics	region_4	Master's & above	m	sourcing	1	34	3	8	1	87	TRUE
37308	37978	Analytics	region_28	Master's & above	m	other	2	41	3	8	1	90	TRUE
37979	57756	Analytics	region_28	Master's & above	m	sourcing	2	37	3	5	1	83	TRUE
39112	10565	Analytics	region_7	Master's & above	m	other	2	37	5	4	1	84	TRUE
51340	72890	Analytics	region_11	Master's & above	m	other	1	39	5	3	1	97	TRUE
51859	59171	Analytics	region_7	Master's & above	m	other	1	39	4	3	1	94	TRUE
54018	69878	Analytics	region_26	Master's & above	m	sourcing	1	41	5	5	1	90	TRUE
54508	43314	Analytics	region_25	Master's & above	m	other	1	35	5	3	1	81	TRUE

And the number of people who did not win rewards or promotions is 747



	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won.	avg_training_score	is_promoted
	All	Analytics	All	Master's	All	All	[...]	All	All	All	[...]	All	FALSE
159	50622	Analytics	region_25	Master's & above	m	other	1	34	5	2	0	84	FALSE
219	27188	Analytics	region_13	Master's & above	m	other	1	30	3	4	0	80	FALSE
399	51002	Analytics	region_15	Master's & above	m	other	1	32	3	4	0	86	FALSE
459	41928	Analytics	region_1	Master's & above	m	sourcing	1	35	5	4	0	84	FALSE
552	20326	Analytics	region_28	Master's & above	m	other	2	32	3	6	0	86	FALSE
633	71562	Analytics	region_28	Master's & above	m	sourcing	2	30	3	6	0	83	FALSE
689	35644	Analytics	region_9	Master's & above	m	other	1	31	4	5	0	82	FALSE
703	28533	Analytics	region_15	Master's & above	m	sourcing	1	37	3	3	0	85	FALSE
756	2732	Analytics	region_22	Master's & above	m	other	2	30	3	6	0	80	FALSE
771	26952	Analytics	region_13	Master's & above	m	other	1	34	2	6	0	86	FALSE
798	64448	Analytics	region_22	Master's & above	f	other	1	33	3	2	0	84	FALSE
861	61087	Analytics	region_15	Master's & above	m	other	3	35	5	3	0	86	FALSE
1157	57808	Analytics	region_22	Master's & above	m	other	1	41	3	12	0	88	FALSE
1261	28235	Analytics	region_10	Master's & above	m	other	1	35	5	3	0	85	FALSE
1324	5487	Analytics	region_11	Master's & above	m	sourcing	2	39	4	8	0	88	FALSE
1370	9165	Analytics	region_7	Master's & above	m	sourcing	1	35	4	5	0	86	FALSE
1422	43922	Analytics	region_31	Master's & above	m	other	1	37	5	3	0	86	FALSE
1574	9036	Analytics	region_5	Master's & above	m	other	2	36	3	7	0	85	FALSE
1742	18804	Analytics	region_2	Master's & above	m	other	1	40	3	4	0	84	FALSE
1807	9546	Analytics	region_1	Master's & above	m	sourcing	1	30	3	4	0	88	FALSE
1826	32527	Analytics	region_28	Master's & above	m	sourcing	1	41	3	4	0	85	FALSE
2126	51566	Analytics	region_28	Master's & above	m	other	3	31	4	7	0	83	FALSE
2326	62438	Analytics	region_22	Master's & above	m	sourcing	1	32	2	5	0	85	FALSE
2338	62429	Analytics	region_22	Master's & above	m	sourcing	2	35	3	6	0	93	FALSE
2351	41618	Analytics	region_25	Master's & above	m	sourcing	1	34	5	8	0	82	FALSE

Showing 1 to 26 of 747 entries, 13 total columns (filtered from 37,698 total entries)

There are many, many statistics that can be generated regarding employees (such as employees of the same gender who share the same region, and many other examples) from Method {View(cleaned_data_withoutOutliers)} Then click on the filter icon and choose the desired filtering method.

For more Details About code: https://github.com/ZizoElkhateeb/DataScience_Project