# EMPLOYEE ATTRITION ANALYSIS

Employees are the backbone of any organization. An organization's performance depends on the quality of employees and how they can retain them. Employee attrition can be for voluntary or involuntary reasons. The reasons are through natural means like retirement, or it can be through resignation, termination of contract. It costs precious time and money and can result in a loss of staff morale. This could also tarnish a Company's reputation. It is important for any organization to monitor their employee attrition rate and understand why employees are leaving if they want to avoid negative repercussions.

Challenges faced by an Organization due to Employee Attrition:

- Cost in training new employees
- Loss of experienced employees
- Employee Productivity
- Organization's profit

We are using the IBM HR Analytics Employee Attrition & Performance dataset for this project. This data set has 1471 rows and 35 columns providing various information on employee's personal and employment details which will help us in studying the behavioral pattern of the employees and predicting the employee churn rate.

Therefore, Organizations like IBM has to study the behavior of employee attrition to stabilize their work culture, in turn, decrease the loss of employee.

Questions to solve from data :

- What is the likelihood of an active employee leaving the company?
- What are the key factors of an employee's attrition?
- If there is any gender bias in the organization, which gender has the higher rate of attrition?
- How are the variables correlated?
- How is the distribution of attrition variable?

# DATA DICTIONARY :

| Attribute Name | Description |
|---|---|
| Age | Age of Employee |
| Attrition | Employee leaving the company - (0=no,1=yes) |
| Business Travel | Employee traveling level - (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely) |
| Daily Rate | Salary level of Employee |
| Department | Department of Employee - (1=HR, 2=R&D, 3=Sales) |
| Distance From Home | The distance from work to home |
| Education | Employee's education level - (1= Below College, 2= College, 3= Bachelor, 4= Master, 5= Doctor) |
| Education Field | Employee's field of Education - (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL) |
| Employee Count | Count of Employee |
| Employee Number | Employee ID |
| Environment Satisfaction | Satisfaction with Environment - (1=Low, 2=Medium ,3=High, 4=Very,5= High) |
| Gender | Employee Gender - (1=FEMALE, 2=MALE) |
| Hourly Rate | Employee Hourly Salary |
| Job Involvement | Job involvement- (1=Low ,2=Medium, 3=High, 4=Very High) |
| Job Level | Employee's level of job |
| Job Role | Employee's Job Role - (1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE) |
| Job Satisfaction | Employees Job Satisfaction- (1=Low, 2=Medium, 3=High, 4=Very High) |

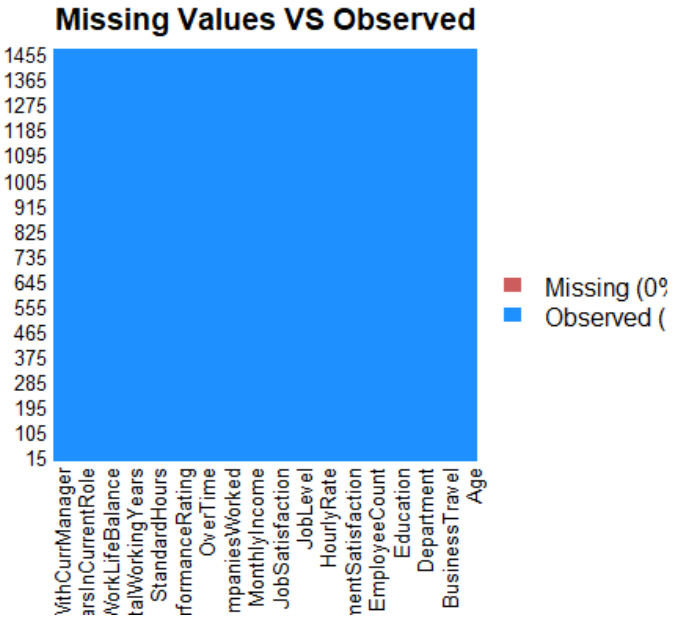
| Attribute Name | Description |
|---|---|
| Martial Status | Employee's Martial Status- (1=DIVORCED, 2=MARRIED, 3=SINGLE) |
| Monthly Rate | Employee's Monthly salary |
| NumCompanies Worked | No. of companies worked at |
| Over 18 | Is employee 18 years old? (1=Yes, 2=No) |
| Overtime | Does employee work overtime? (1=No, 2=Yes) |
| Percent Salary Hike | Employee's percentage increase in salary |
| Performance Rating | Employee's performance rating(1=Low, 2=Good, 3=Excellent, 4=Outstanding) |
| Relations Satisfaction | Employee's relations satisfaction(1=Low, 2=Medium, 3=High, 4=Very High) |
| Standard Hours | Employee's Standard working hours |
| Stock Options Level | Employee's stock options |
| Total Working Years | Employee total years worked |
| Training Times Last Year | Employee training hours |
| Work Life Balance | Employee's time spent between work and outside (1=Bad, 2=Good,3=Better, 4=Best) |
| Years at company | Employee's total nuber of years at company |
| Years in Current Role | Employee's number of years in current role |
| Years since Last Promotion | Number of years since last promotion |
| Years with Current Manager | Number of years spent with current manager |

# Steps :

1. Data Cleaning and Formatting
2. EDA
3. Principal Component Analysis
4. Cluster Analysis
5. Factor Analysis
6. Multiple Regression
7. Logistic Regression
8. Discriminant Analysis

# Data Cleaning and Formatting :

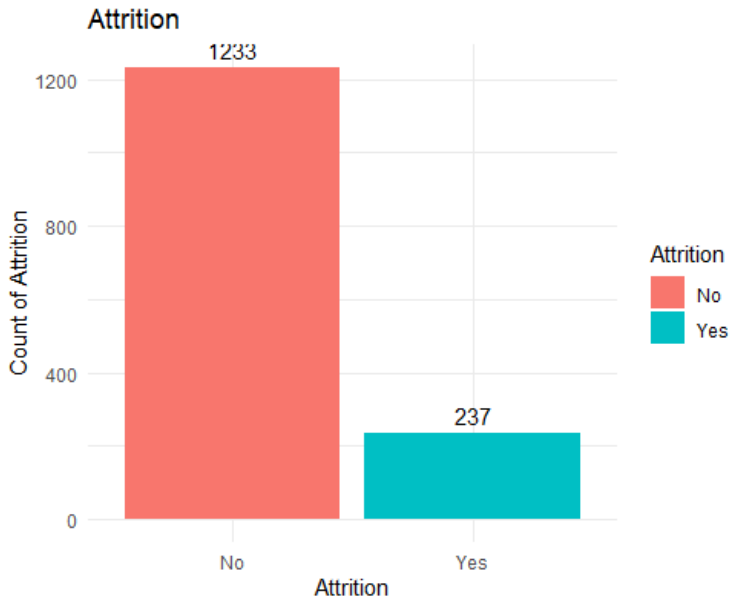
- Changed the name of the Column Age.
- Checked for null values

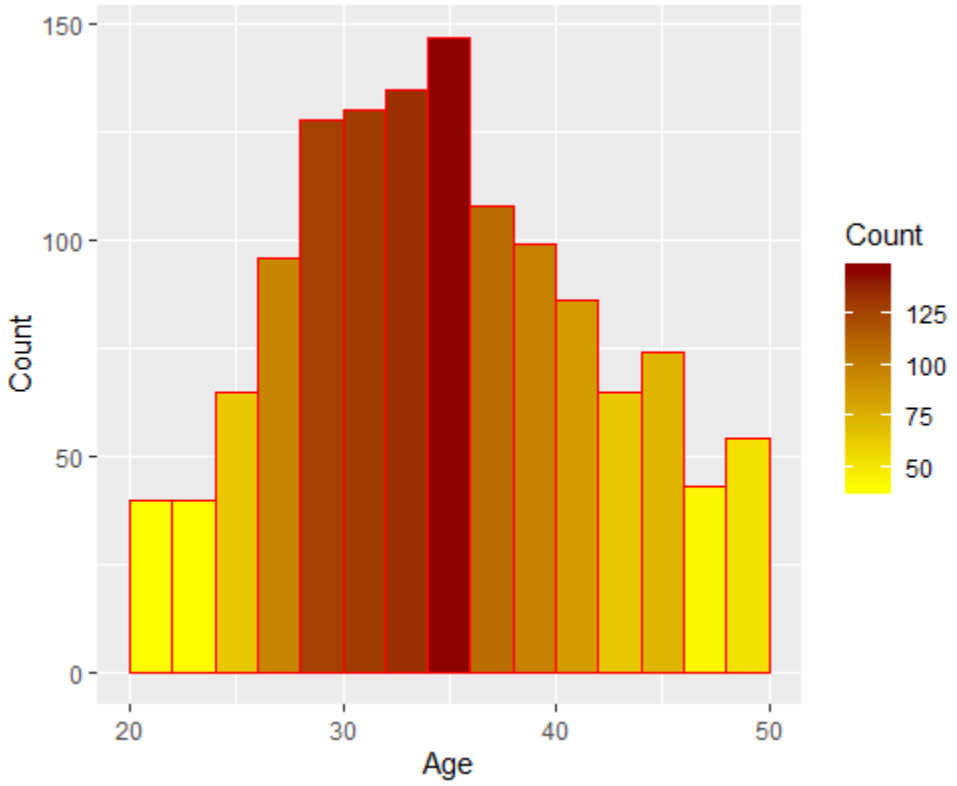



- Removed the redundant Columns
    a. Employee Number
    b. Standard Hours
    c. Over 18
    d. Employee Count

# Exploratory Data Analysis :

We have visualized our dataset by plotting graphs to see how the independent variables relate to the dependent variable.


Attrition
1233
237
1200
800
400
0
Count of Attrition
No
Yes
Attrition
Attrition
No
Yes


150
100
50
0
Count
20
30
40
50
Age
Count
125
100
75
50

## Histogram and QQ plots for Numeric columns :

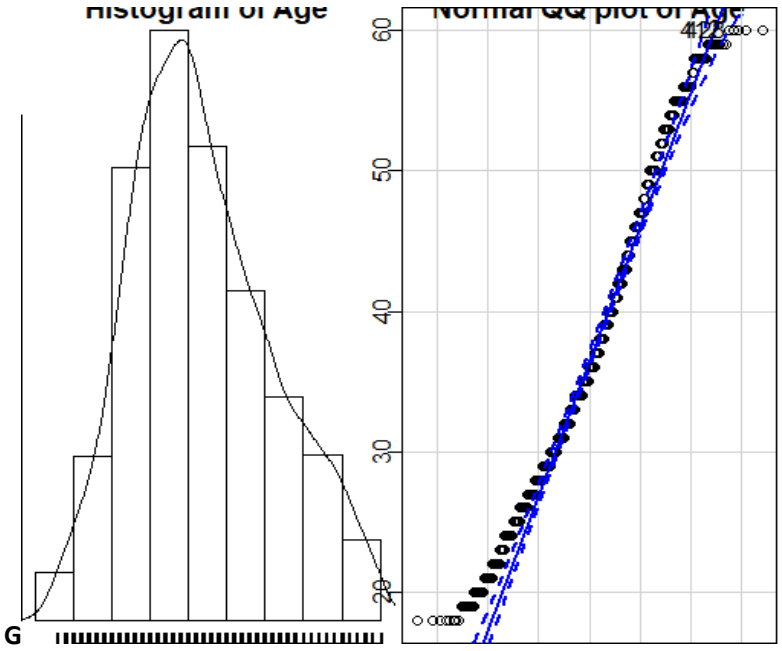



**Fig : Age**

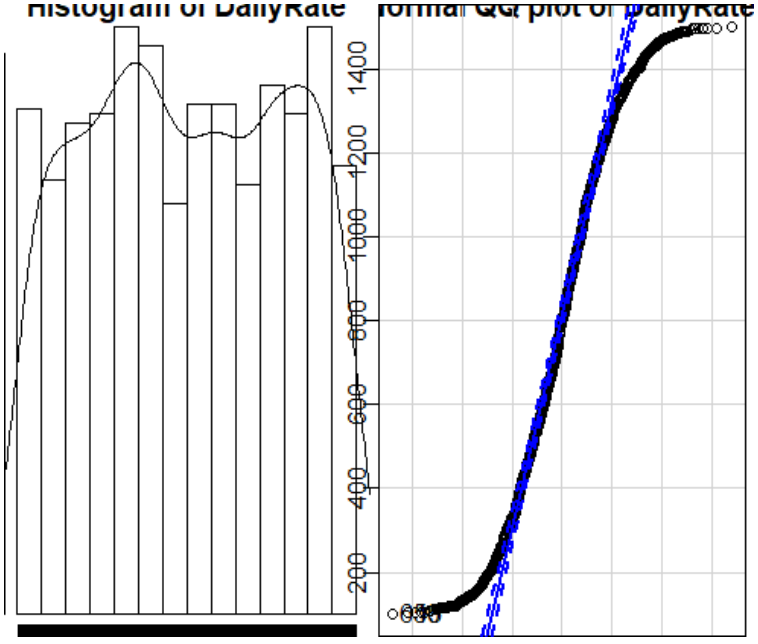


**Fig : Daily Rate**

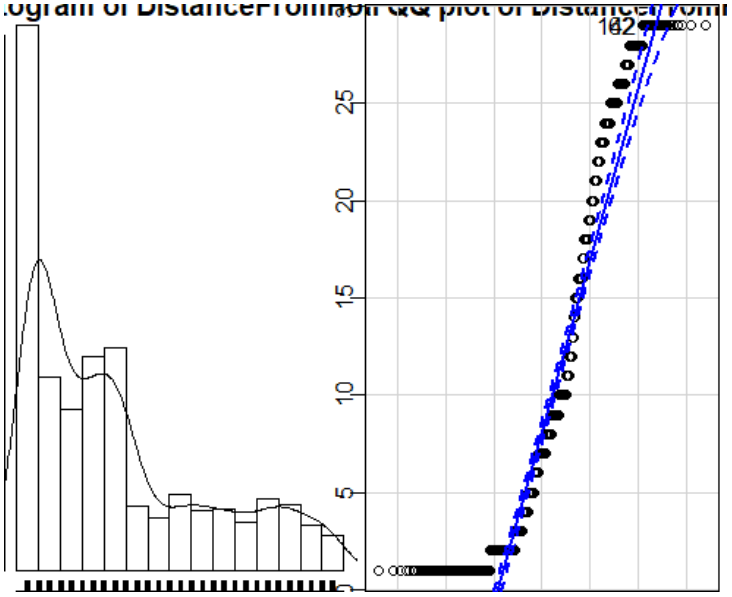



**Fig : Distance From Home**

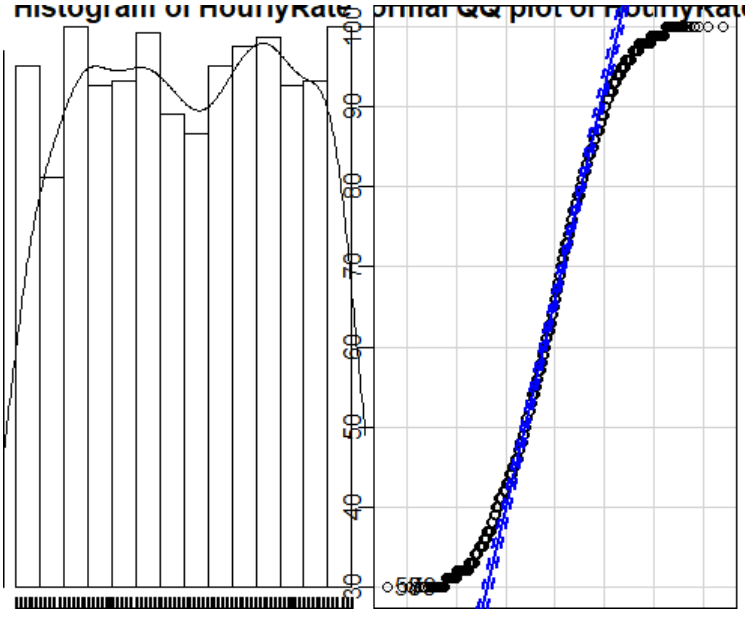



**Fig : Hourly Rate**

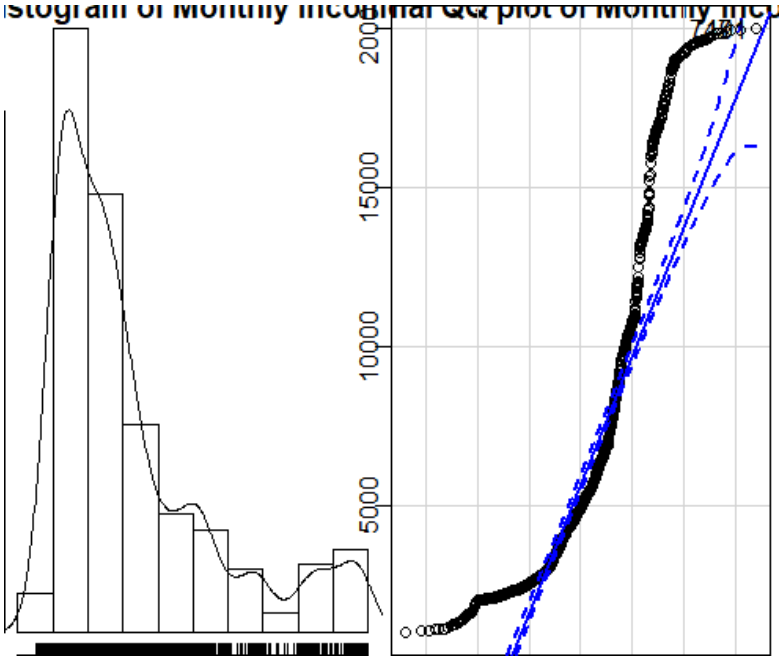



**Fig : Monthly Income**

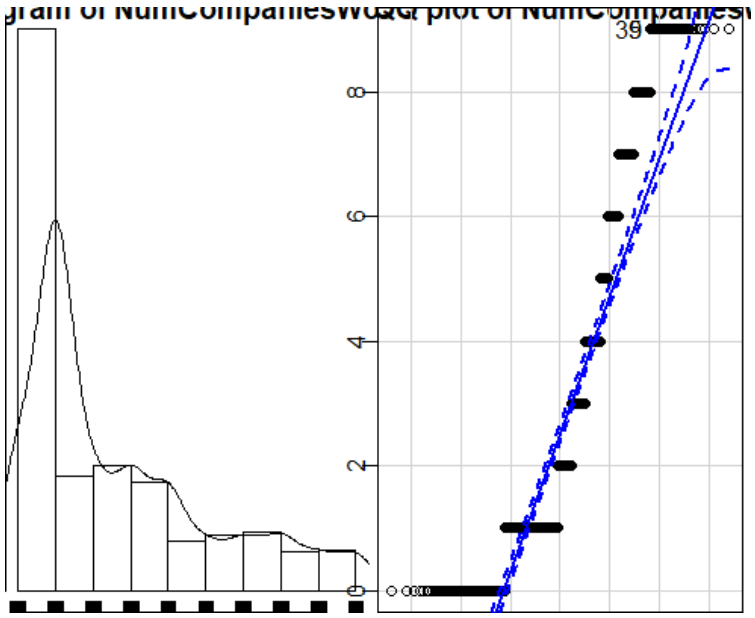



**Fig : Number of Companies worked**

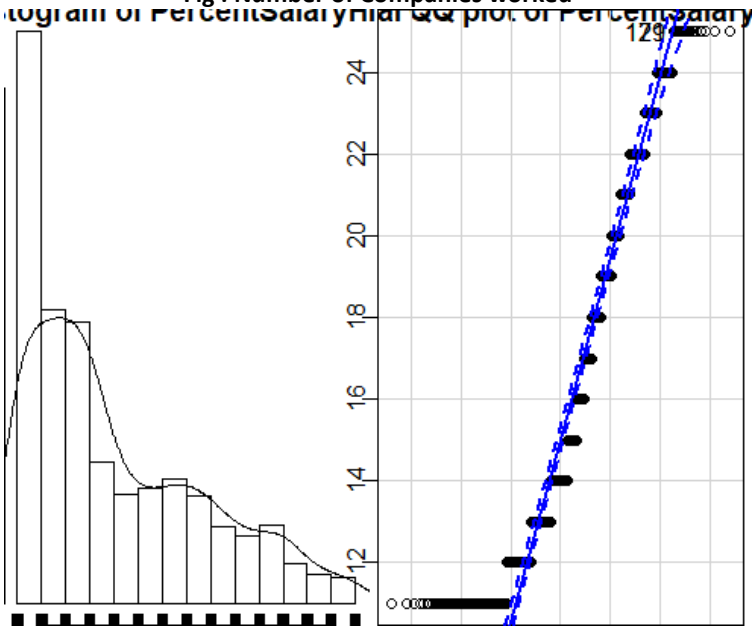



**Fig : Percent Salary Hike**

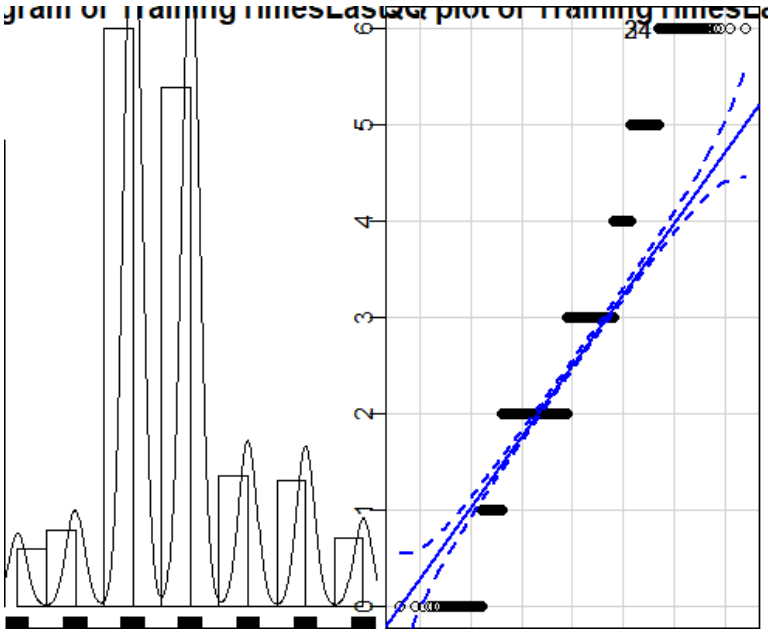



**Fig : Training times since last year**

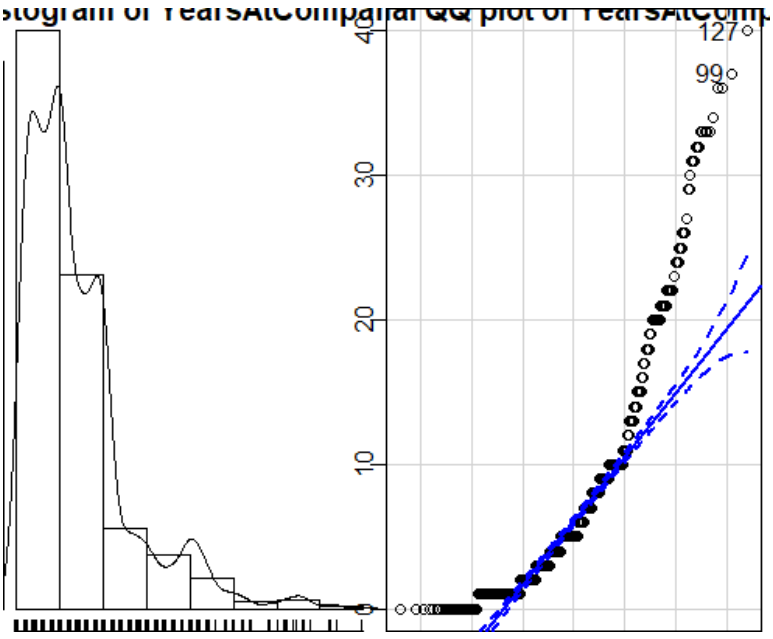



**Fig : Years at Company**

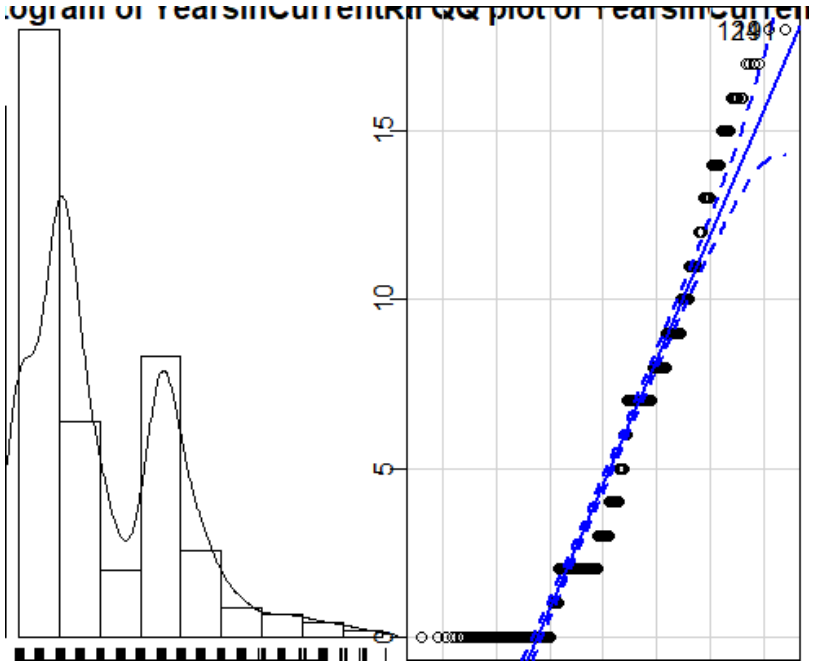



**Fig : Years in Current Role**

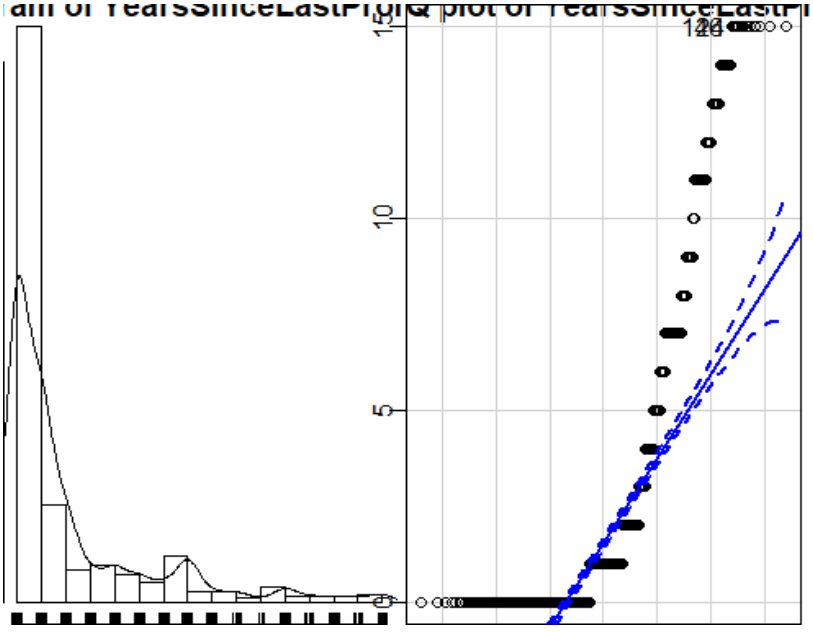



**Fig : Years since last Promotion**

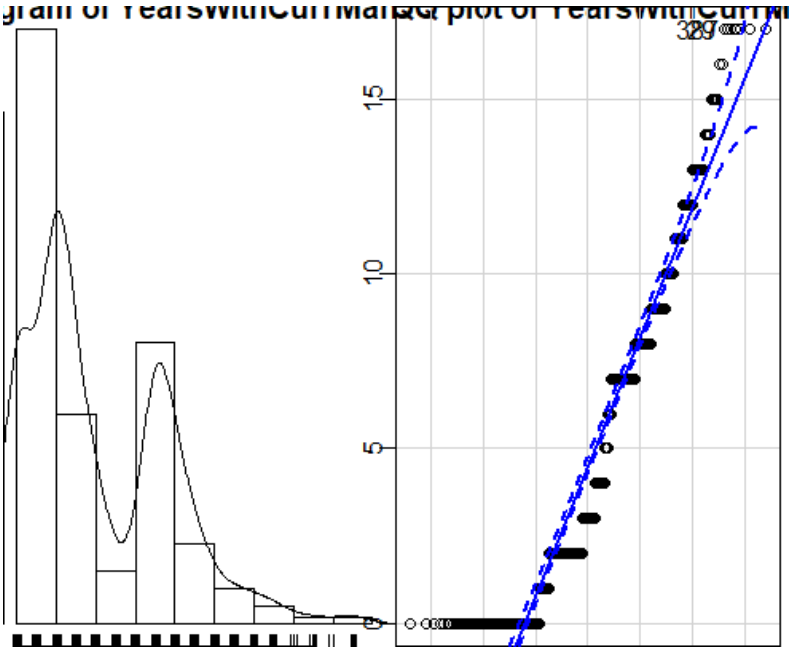



**Fig : Years with current Manager**

**Boxplot distributions for our numeric columns :**

- The dashed line shows the mean and the dark center line shows the median
- Difference between these two lines depict the deviation from the central limit theorem

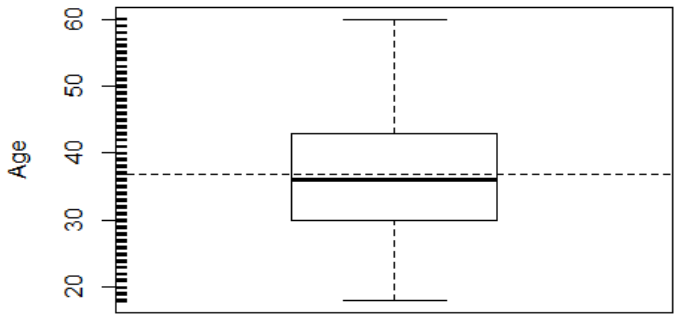



**Fig : Age**

**Plotting the Age with 3 lines for mean, median and mean+std**

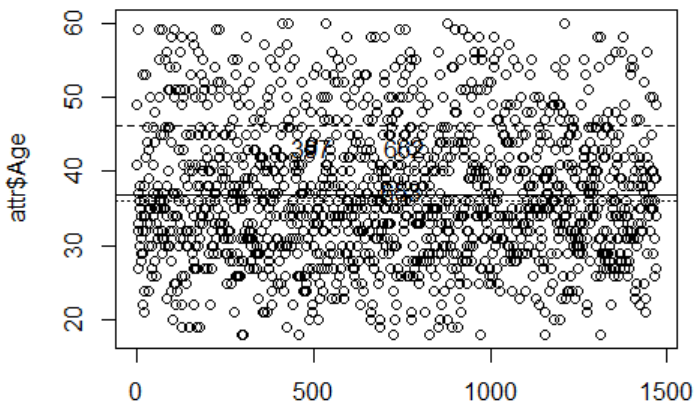



**Fig : Mean, Median and SD for Age**

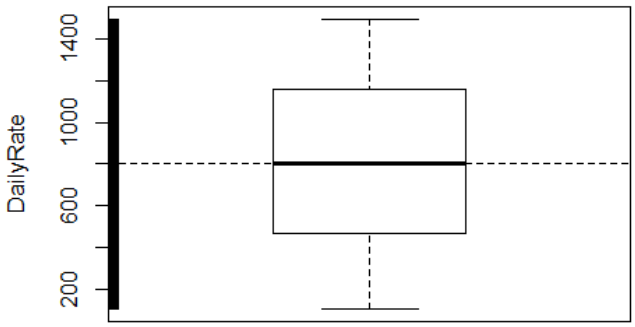



**Fig : Daily Rate**

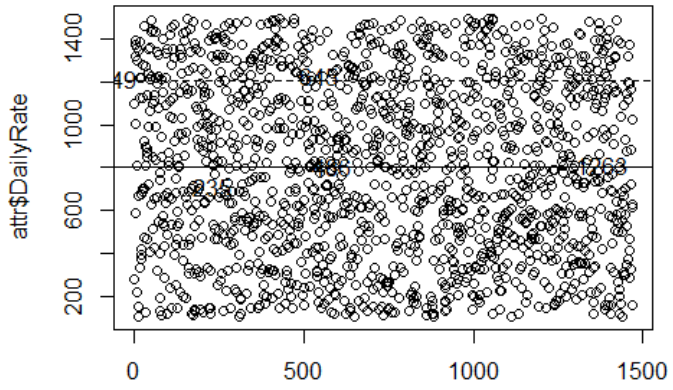



**Fig : Mean, Median and SD for Daily Rate**

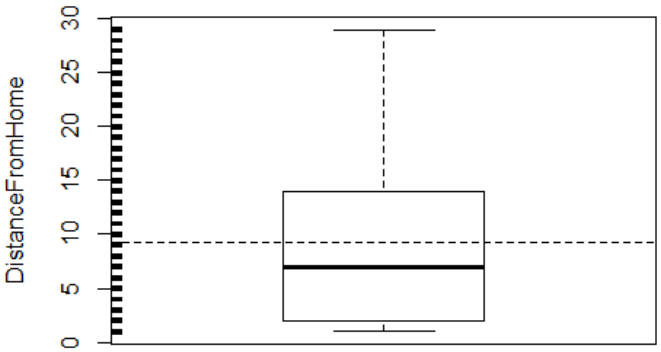



**Fig : Distance from Home**

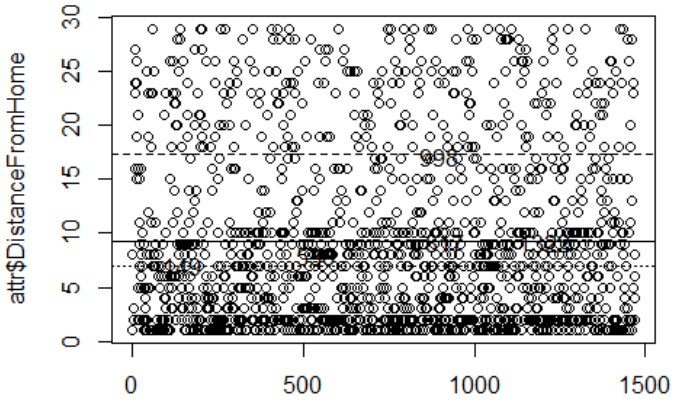



**Fig : Mean, Median and SD for Distance from Home**

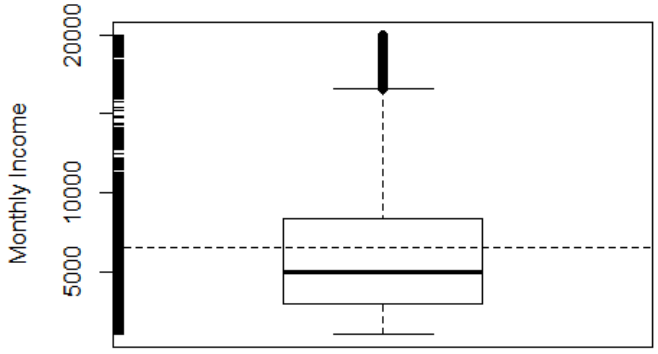



**Fig : Monthly Income**

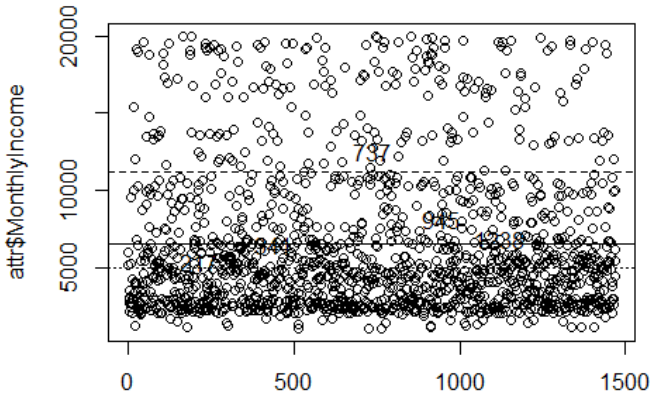



**Fig : Mean, Median and SD for Monthly Income**

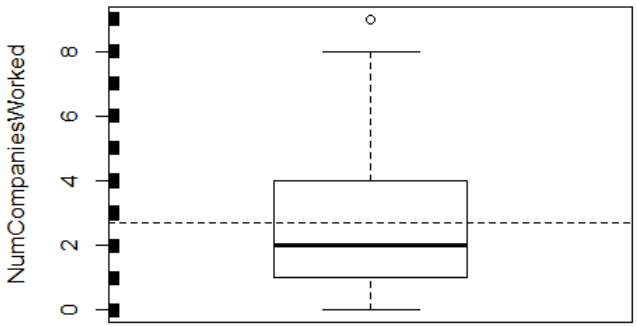



**Fig : Number of Companies worked**

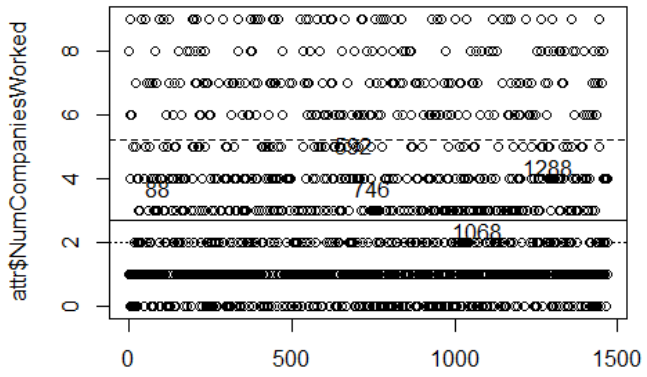



**Fig : Mean, Median and SD for Number of Companies worked**

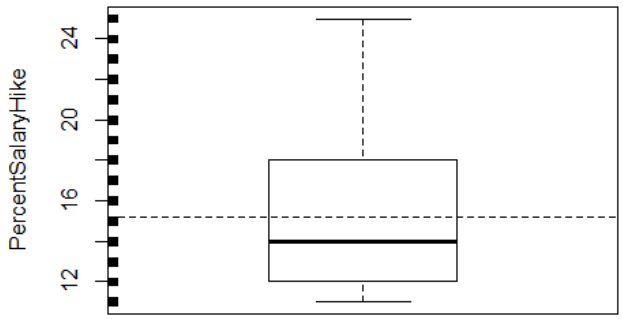



**Fig : percent salary hike**

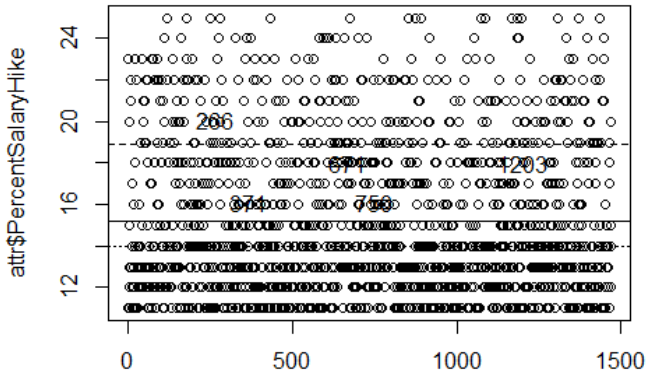



**Fig : Mean, Median and SD for percent salary hike**

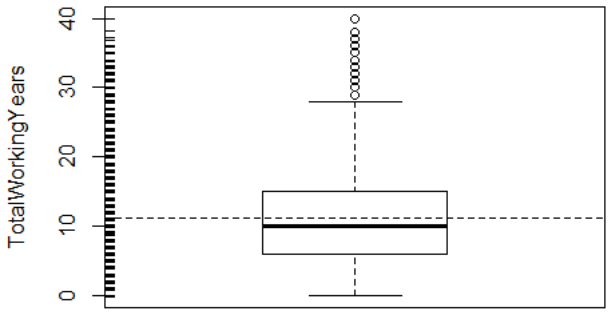



**Fig : Total working years**

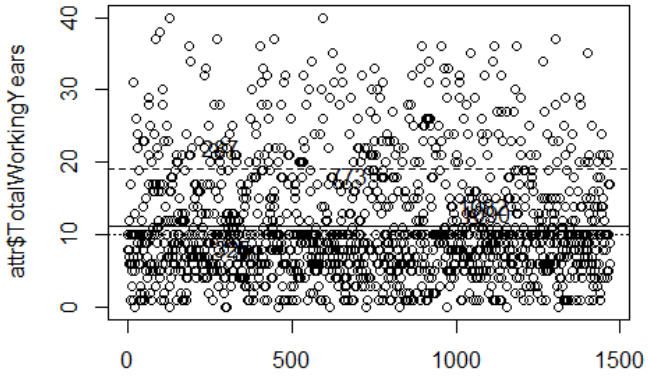



**Fig : Mean, Median and SD for Total working years**

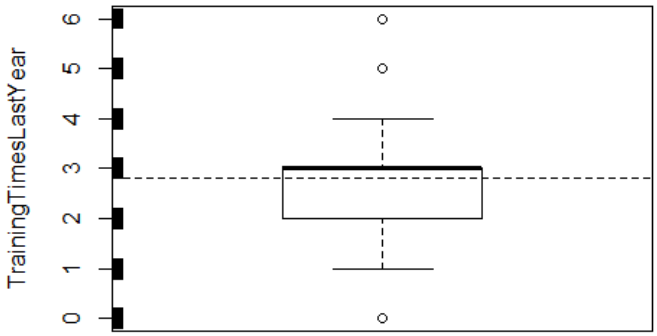



**Fig : Training times last year**

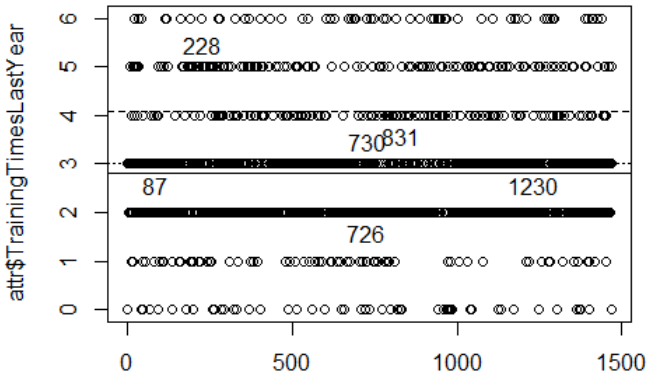



**Fig : Mean, Median and SD for Training times last year**

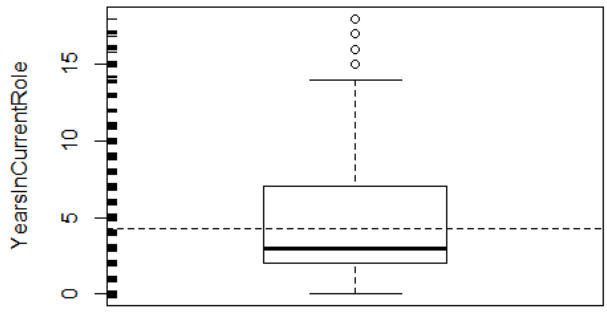



**Fig : Years in Current role**

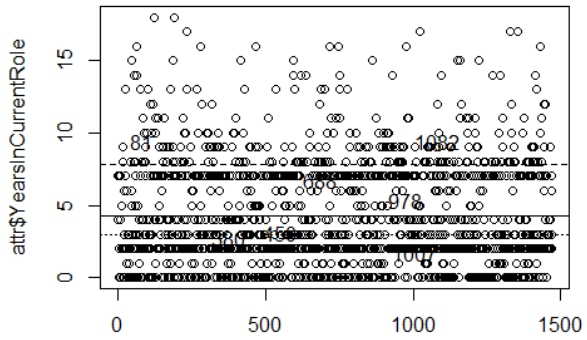



**Fig : Mean, Median and SD for Years in Current role**

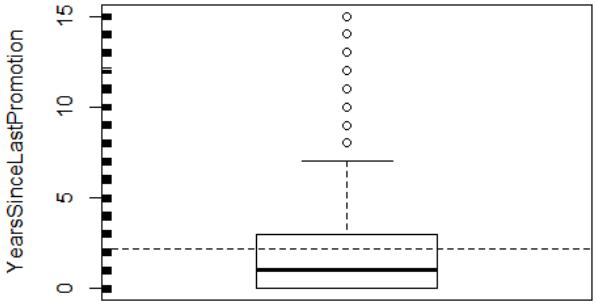



**Fig : Years since last promotion**

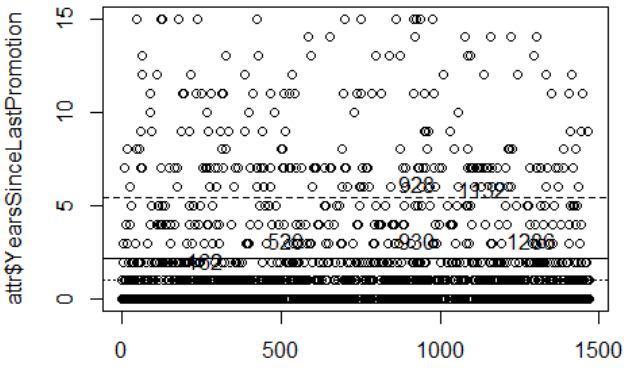



**Fig : Mean, Median and SD for Years since last promotion**

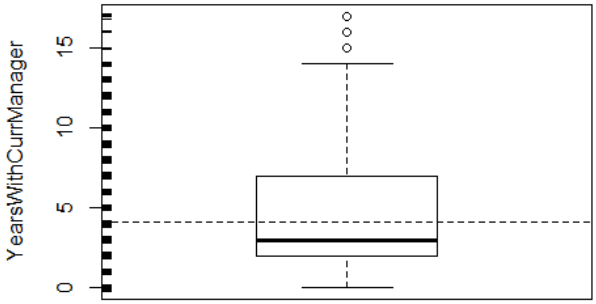



**Fig : Years with current manager**

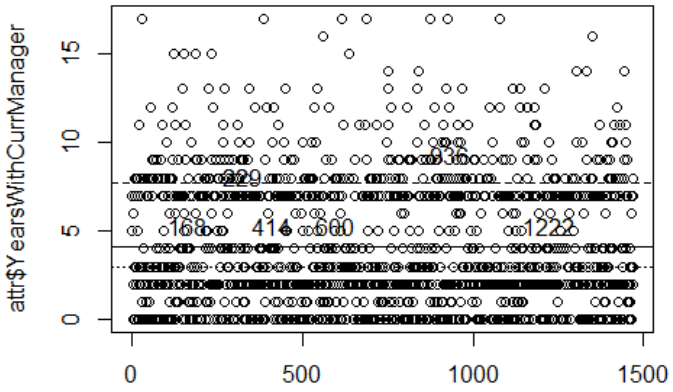
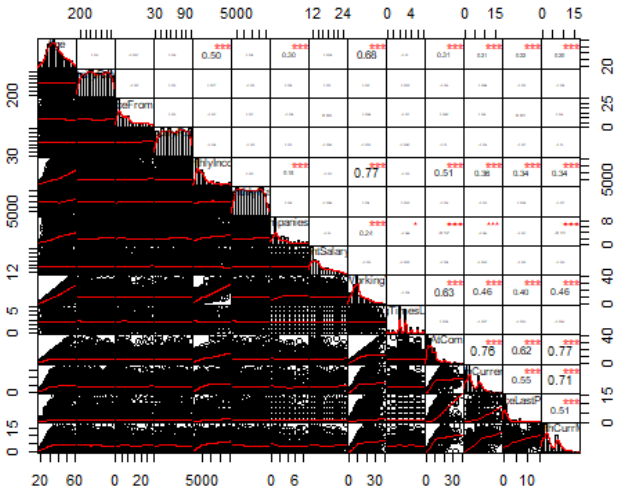


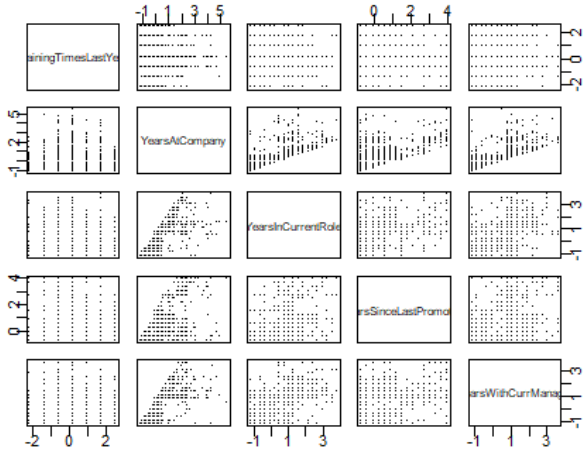

**Fig : Mean, Median and SD for Years with current manager**

## Correlation Plot :

Plotting a correlation and a covariance plot for our numerical columns.




**Correlation**




**Covariance**

## T- Test :

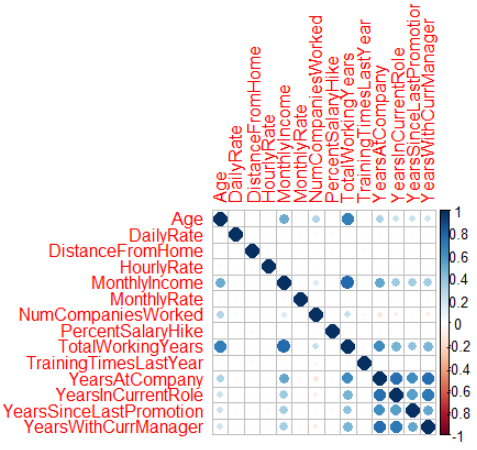
```
#T-Test
#Null Hypothesis - The two means are equal
#Alternate Hypothesis - Difference in the two means is not zero
#pvalue >= 0.05, accept null hypothesis
#Or
#else accept the alternate hypothesis

#Univariate mean comparison using t test
> #Monthly Income and Attrition
> with(data=attr,t.test(attr$MonthlyIncome[attr$Attrition=="Yes"],attr$MonthlyInc
ome[attr$Attrition=="No"],var.equal=TRUE))

Two Sample t-test

data:  attr$MonthlyIncome[attr$Attrition == "Yes"] and attr$MonthlyIncome[attr$At
trition == "No"]
t = -6.2039, df = 1468, p-value = 7.147e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2692.446 -1398.847
sample estimates:
mean of x mean of y
4787.093  6832.740

> #HourlyRate and Attrition
> with(data=attr,t.test(attr$HourlyRate[attr$Attrition=="Yes"],attr$HourlyRate[at
tr$Attrition=="No"],var.equal=TRUE))

Two Sample t-test

data:  attr$HourlyRate[attr$Attrition == "Yes"] and attr$HourlyRate[attr$Attritio
n == "No"]
t = -0.26229, df = 1468, p-value = 0.7931
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
-3.207565  2.450946
sample estimates:
mean of x mean of y
65.57384  65.95215

> #Daily Rate and Attrition
> with(data=attr,t.test(attr$DailyRate[attr$Attrition=="Yes"],attr$DailyRate[attr
$Attrition=="No"],var.equal=TRUE))

Two Sample t-test

data:  attr$DailyRate[attr$Attrition == "Yes"] and attr$DailyRate[attr$Attrition
== "No"]
t = -2.1741, df = 1468, p-value = 0.02986
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-118.209251   -6.073932
sample estimates:
mean of x mean of y
750.3629  812.5045

> #Age and Attrition
> with(data=attr,t.test(attr$Age[attr$Attrition=="Yes"],attr$Age[attr$Attrition==
"No"],var.equal=TRUE))

Two Sample t-test

data:  attr$Age[attr$Attrition == "Yes"] and attr$Age[attr$Attrition == "No"]
t = -6.1787, df = 1468, p-value = 8.356e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.208825 -2.698450
sample estimates:
mean of x mean of y
33.60759  37.56123

> #DistanceFromHome and Attrition
> with(data = attr,t.test(attr$DistanceFromHome[attr$Attrition=="Yes"],attr$Age[a
ttr$Attrition=="No"],var.equal = TRUE))

Two Sample t-test

data:  attr$DistanceFromHome[attr$Attrition == "Yes"] and attr$Age[attr$Attrition
 == "No"]
t = -43.048, df = 1468, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-28.15538 -25.70126
sample estimates:
mean of x mean of y
10.63291  37.56123

> #Monthly Income and Gender
> with(data = attr,t.test(attr$MonthlyIncome[attr$Gender=="Male"],attr$MonthlyInc
ome[attr$Gender=="Female"],var.equal = TRUE))

Two Sample t-test

data:  attr$MonthlyIncome[attr$Gender == "Male"] and attr$MonthlyIncome[attr$Gend
er == "Female"]
t = -1.2213, df = 1468, p-value = 0.2222
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-797.6470  185.5303
```

```
sample estimates:
mean of x mean of y
6380.508  6686.566

> #DistanceFromHome and Gender
> with(data = attr,t.test(attr$DistanceFromHome[attr$Gender=="Male"],attr$Distanc
eFromHome[attr$Gender=="Female"],var.equal = TRUE))

Two Sample t-test

data:  attr$DistanceFromHome[attr$Gender == "Male"] and attr$DistanceFromHome[att
r$Gender == "Female"]
t = -0.070902, df = 1468, p-value = 0.9435
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.8775316  0.8163071
sample estimates:
mean of x mean of y
9.180272  9.210884

> #Monthly Income and gender
> t2testgender <- hotelling.test(attr$MonthlyIncome + attr$DistanceFromHome ~ att
r$Gender, data=attr)
> cat("T2 statistic =",t2testgender$stat[[1]],"\n")
T2 statistic = 1.499903
> print(t2testgender)
Test stat:  0.74944
Numerator df:  2
Denominator df:  1467
P-value:  0.4728

> #Monthly Income and Attrition
> t2testattr <- hotelling.test(attr$MonthlyIncome + attr$DistanceFromHome ~ attr$
Attrition, data=attr)
> cat("T2 statistic =",t2testattr$stat[[1]],"\n")
T2 statistic = 47.28597
> print(t2testattr)
Test stat:  23.627
Numerator df:  2
Denominator df:  1467
P-value:  7.957e-11
```

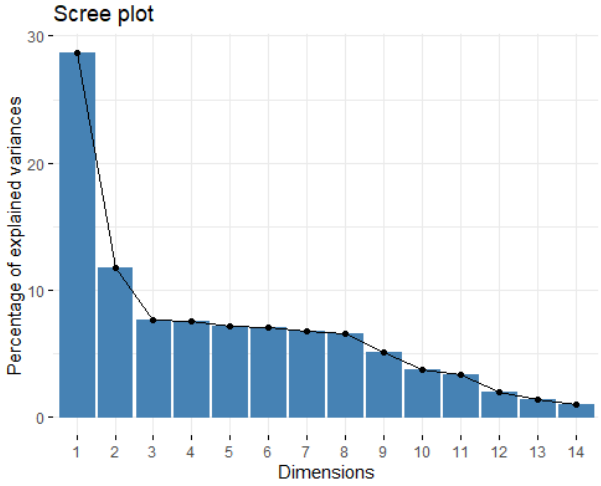
# Principal Component Analysis :

**Plotting the correlation.**



```
$rotation
PC1           PC2           PC3           PC4           PC5           PC6
Age                   0.280157344 -0.472170158  0.003362193  0.004488409 -0.03
9563410 -0.058709821
DailyRate            -0.006815197 -0.077962430 -0.207301367 -0.609569867 -0.21
1568990 -0.130624253
DistanceFromHome      0.004812032  0.041564987 -0.664884791  0.306131593  0.04
8941659 -0.176841356
HourlyRate           -0.011288550 -0.062668026 -0.352147686 -0.255816205  0.60
2292088 -0.481631551
MonthlyIncome         0.360622909 -0.290395305  0.052415102  0.025332267 -0.03
4941693 -0.033266121
MonthlyRate           0.001123298 -0.086158010  0.020312197  0.664085954 -0.10
1166486 -0.371270681
NumCompaniesWorked    0.030991906 -0.560133264  0.005628265 -0.041875610  0.01
7785645  0.101255602
PercentSalaryHike    -0.015351368  0.004618486 -0.465841883 -0.055689609 -0.69
8726672 -0.008298517
TotalWorkingYears     0.415285665 -0.318115831  0.009368263  0.007027664 -0.02
4159198 -0.025255659
TrainingTimesLastYear -0.010993402  0.092457674  0.409028173 -0.138279489 -0.29
3982017 -0.751634233
YearsAtCompany        0.443443529  0.213079968  0.002115638 -0.010571214  0.02
4921329 -0.001937722
YearsInCurrentRole    0.391353065  0.279423881 -0.048111956 -0.038785223 -0.00
4927194  0.014570562
YearsSinceLastPromotion  0.344322397  0.198658357  0.003993040  0.027659809  0.01
9935007  0.018688744
YearsWithCurrManager  0.386171187  0.295138965 -0.031745944 -0.034459502  0.02
1898300  0.028658936
PC7           PC8           PC9          PC10         PC11          PC12
Age                  -0.098196914 -0.05927715 -0.183114693  0.005033984 -0.743
67068 -0.0415507268
DailyRate             0.715405171 -0.02770642 -0.028707475  0.040304455 -0.019
80752  0.0404456766
DistanceFromHome      0.031447533 -0.65217193  0.037737577  0.002338630  0.029
27699  0.0034205705
HourlyRate           -0.221010405  0.40142111 -0.004675476  0.018009772  0.037
25997 -0.0040743265
MonthlyIncome        -0.012272736 -0.03685912 -0.377381332  0.104651321  0.617
75910 -0.0474998229
MonthlyRate           0.482943083  0.40448871  0.056690883 -0.044889268 -0.016
81584  0.0237091682
NumCompaniesWorked   -0.032989593 -0.03355765  0.775796629 -0.129586743  0.196
87866  0.0210288451
PercentSalaryHike    -0.376210309  0.38335261  0.012190972  0.019568502  0.040
82482  0.0143172942
TotalWorkingYears    -0.029511945 -0.04398227 -0.196663458 -0.038585533  0.083
01594  0.0611091374
TrainingTimesLastYear -0.217564575 -0.29622601  0.130785998 -0.017811234  0.029
54240 -0.0002892709
YearsAtCompany        0.005335572  0.01862614 -0.001551392 -0.104225054  0.037
62072  0.0795277828
YearsInCurrentRole    0.062086964  0.05420752  0.201595025 -0.271683842 -0.044
58982 -0.7658067069
YearsSinceLastPromotion  0.022129234  0.03850513  0.306725567  0.845951303 -0.080
53456  0.0759463597
YearsWithCurrManager  0.011525930  0.04176204  0.161413516 -0.407140185 -0.065
76770  0.6251855631
PC13          PC14
Age                   0.1893016403  0.237072230
DailyRate            -0.0099081253  0.018837870
DistanceFromHome      0.0057709912  0.011991386
HourlyRate            0.0049037331 -0.003314012
```

```
MonthlyIncome             0.4041044054  0.279006032
MonthlyRate              -0.0083381972  0.009367738
NumCompaniesWorked       -0.0369554196  0.107234673
PercentSalaryHike        -0.0183445446  0.010585240
TotalWorkingYears        -0.4158888971 -0.705989494
TrainingTimesLastYear     0.0008431661 -0.012085362
YearsAtCompany           -0.6494086309  0.562584645
YearsInCurrentRole        0.1948014145 -0.130643692
YearsSinceLastPromotion   0.0977011821 -0.083447864
YearsWithCurrManager      0.3959144832 -0.121010943
```




# Cluster Analysis :

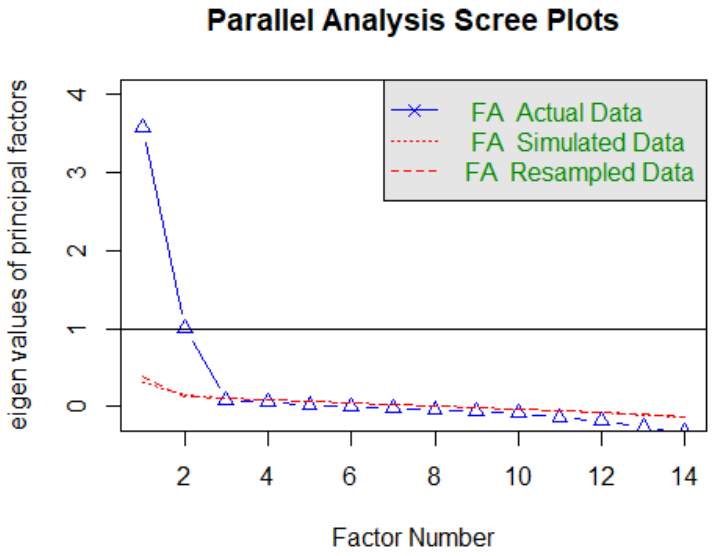
Formed 6 clusters using K-means clustering.

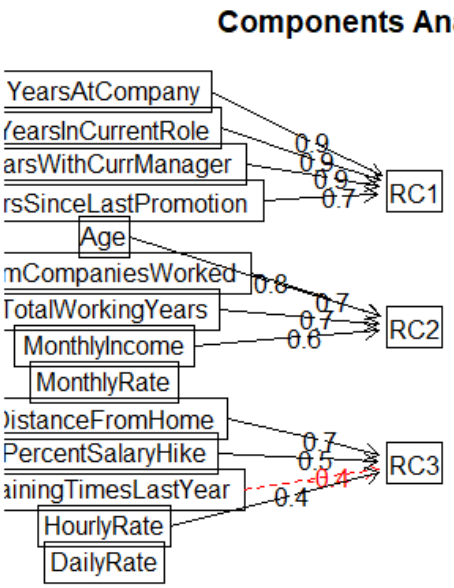
```
# K-means, k=2, 3, 4, 5, 6
>
> # Centers (k's) are numbers thus, 10 random sets are chosen
>
>
>
> (kmeans2_attr_std <- kmeans(attr_std,2,nstart = 10))
K-means clustering with 2 clusters of sizes 988, 482

Cluster means:
Age    DailyRate DistanceFromHome  HourlyRate MonthlyIncome  MonthlyRate NumCompa
niesWorked PercentSalaryHike
1 -0.2736309 -0.003308239     -0.003147081  0.02800718    -0.4005389 -0.000543936
        0.004100873         0.0171433
2  0.5608866  0.006781204      0.006450863 -0.05740890     0.8210216  0.001114956
       -0.008405938        -0.0351402
TotalWorkingYears TrainingTimesLastYear YearsAtCompany YearsInCurrentRole YearsSi
nceLastPromotion YearsWithCurrManager
1        -0.4455643         -0.0005714331     -0.4814244         -0.4920861
        -0.4116207           -0.4699297
2         0.9133144          0.0011713194      0.9868202          1.0086744
         0.8437371           0.9632583

Clustering vector:
[1] 1 2 1 1 1 1 1 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1
 1 1 1 1 2 2 2 2 1 2 1 1 1 1 1 1 2 2 1
[59] 1 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 1 2 2 2 1 2 2 2 1
1 1 2 1 1 1 1 2 1 2 2 1 1 1 2 1 2 1 1 1
```

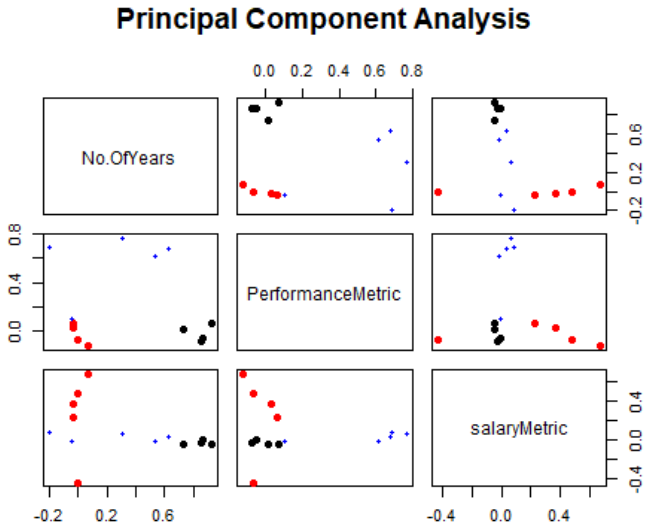
```
[117] 2 2 1 2 1 1 1 2 2 1 2 1 1 2 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2 1 1 2 1 2 2
 2 1 1 1 2 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 2
[175] 1 1 1 1 2 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2
 1 2 1 1 1 1 2 1 2 1 2 2 1 1 2 2 2 1 1 2
[233] 1 2 1 2 1 2 1 1 1 1 1 2 2 1 1 2 1 1 1 2 1 1 1 1 2 2 1 1 1 2 1 2 1 1 1 1 2 2
 2 2 1 1 1 1 2 1 1 2 2 2 1 2 1 2 1 1 1 1
[291] 2 1 1 1 1 2 1 2 1 1 2 1 1 2 2 2 2 2 1 1 1 2 1 2 2 1 2 2 1 2 1 1 1 1 2 2 2 1
 1 2 1 1 1 1 2 1 1 1 1 1 1 2 2 1 2 1 1 1
[349] 1 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1
 2 1 1 1 2 1 2 1 2 1 1 1 2 1 2 2 1 2 1 1
[407] 2 1 2 1 1 2 2 1 1 1 1 2 1 1 1 1 1 2 1 2 2 2 1 2 1 1 1 1 2 2 1 1 1 1 1 1 2 1
 2 2 2 2 2 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1
[465] 1 1 2 2 2 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1
 2 2 1 1 2 1 2 2 2 1 1 1 2 1 1 1 1 1 1 1
[523] 1 2 2 1 2 1 1 2 1 2 2 2 2 2 1 1 2 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 2 1
 1 2 2 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1
[581] 1 1 1 1 2 1 1 1 2 1 2 1 2 1 1 2 1 1 1 1 2 1 1 1 2 1 1 2 2 1 2 2 1 1 1 1 2 1
 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1
[639] 1 1 1 2 1 1 1 1 1 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1
 1 2 1 1 1 2 1 1 1 1 2 2 1 1 2 1 1 2 1 2
[697] 2 1 1 2 1 2 2 1 2 1 1 2 1 1 2 1 1 1 2 1 2 1 2 1 1 2 1 2 1 1 1 1 2 2 2 1 1 1
 1 1 2 1 2 1 1 1 1 2 1 2 2 1 1 2 2 2 2 2
[755] 1 2 2 2 2 2 1 2 1 1 1 1 1 2 1 2 2 1 2 2 1 1 1 2 1 2 1 1 1 2 2 1 1 1 2 1 2
 2 1 1 1 1 1 2 1 1 1 1 2 2 2 2 1 1 2 1
[813] 2 2 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 2 2 1 1
 1 2 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 2 1 2
[871] 2 1 1 1 2 2 1 1 1 2 1 1 2 2 1 1 2 1 2 2 2 2 1 1 2 1 2 1 2 1 2 1 1 1 1 2 1 2
 2 1 1 1 1 2 2 1 2 1 2 2 2 1 2 2 1 1 2 2
[929] 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 2 2 2 1 2 1 1 2 2 1 1 2 2 2 1 1 2 1 1 2 2 2 1
 2 1 2 2 1 2 1 1 1 2 2 1 2 2 1 1 1 2 1 1
[987] 1 1 1 1 1 1 1 1 2 2 1 1 1 2
[ reached getOption("max.print") -- omitted 470 entries ]

Within cluster sum of squares by cluster:
[1] 9563.508 7087.951
(between_SS / total_SS =  19.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "b
etweenss"    "size"         "iter"
[9] "ifault"
> # Computing the percentage of variation accounted for. Two clusters
>
> perc.var.2 <- round(100*(1 - kmeans2_attr_std$betweenss/kmeans2_attr_std$totss)
,1)
>
> names(perc.var.2) <- "Perc. 2 clus"
>
> perc.var.2
Perc. 2 clus
81
(kmeans3_attr_std <- kmeans(attr_std,3,nstart = 10))
K-means clustering with 3 clusters of sizes 367, 705, 398
perc.var.3 <- round(100*(1 - kmeans3_attr_std$betweenss/kmeans3_attr_std$totss),1
)
>
> names(perc.var.3) <- "Perc. 3 clus"
>
> perc.var.3
Perc. 3 clus
74.5
(kmeans4_attr_std  <- kmeans(attr_std,4,nstart = 10))
K-means clustering with 4 clusters of sizes 370, 560, 156, 384
```

```
perc.var.4 <- round(100*(1 - kmeans4_attr_std$betweenss/kmeans4_attr_std$totss),1
)
>
> names(perc.var.4) <- "Perc. 4 clus"
>
> perc.var.4
Perc. 4 clus
68.6
> # Computing the percentage of variation accounted for. Five clusters
>
> (kmeans5_attr_std  <- kmeans(attr_std,5,nstart = 10))
K-means clustering with 5 clusters of sizes 362, 121, 464, 385, 138
> perc.var.5 <- round(100*(1 - kmeans5_attr_std$betweenss/kmeans5_attr_std$totss)
,1)
>
> names(perc.var.5) <- "Perc. 5 clus"
>
> perc.var.5
Perc. 5 clus
65.3
> # Computing the percentage of variation accounted for. Six clusters
> (kmeans6_attr_std  <- kmeans(attr_std,6,nstart = 10))
K-means clustering with 6 clusters of sizes 315, 204, 117, 367, 133, 334
> perc.var.6 <- round(100*(1 - kmeans6_attr_std$betweenss/kmeans6_attr_std$totss)
,1)
>
> names(perc.var.6) <- "Perc. 6 clus"
>
> perc.var.6
Perc. 6 clus
62.8
```

## Factor Analysis :






# Multiple Regression :

```
> fit9<-  lm(Attrition~Age+DistanceFromHome+MonthlyIncome+NumCompaniesWorked+Year
sInCurrentRole+YearsSinceLastPromotion,data=attr)
> summary(fit9)

Call:
lm(formula = Attrition ~ Age + DistanceFromHome + MonthlyIncome +
NumCompaniesWorked + YearsInCurrentRole + YearsSinceLastPromotion,
data = attr)

Residuals:
```

```
   Min       1Q   Median       3Q      Max
-0.39822 -0.20509 -0.13451 -0.03268  1.16520

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.374e+00  4.105e-02  33.474  < 2e-16 ***
Age                     -5.374e-03  1.221e-03  -4.400 1.16e-05 ***
DistanceFromHome         3.658e-03  1.147e-03   3.189 0.001460 **
MonthlyIncome           -6.871e-06  2.420e-06  -2.839 0.004592 **
NumCompaniesWorked       1.311e-02  3.963e-03   3.308 0.000962 ***
YearsInCurrentRole      -1.586e-02  3.175e-03  -4.995 6.59e-07 ***
YearsSinceLastPromotion  1.305e-02  3.514e-03   3.713 0.000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3561 on 1463 degrees of freedom
Multiple R-squared:  0.06676,       Adjusted R-squared:  0.06293
F-statistic: 17.44 on 6 and 1463 DF,  p-value: < 2.2e-16
```

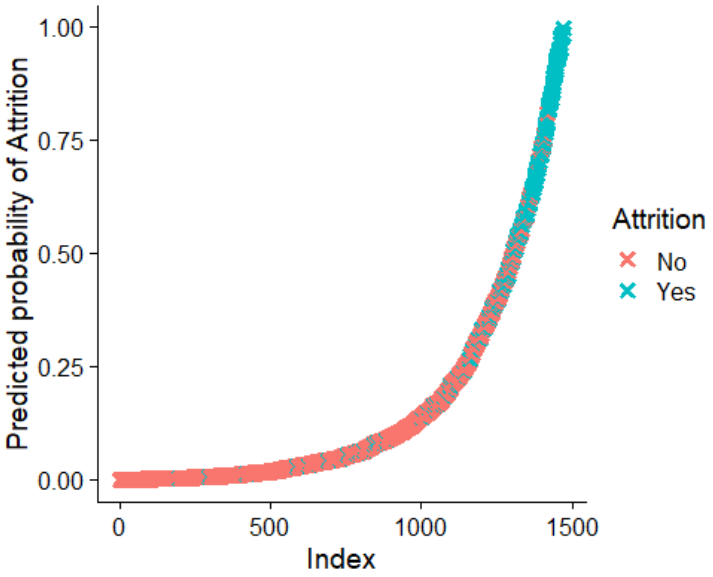
**Logistic Regression :**

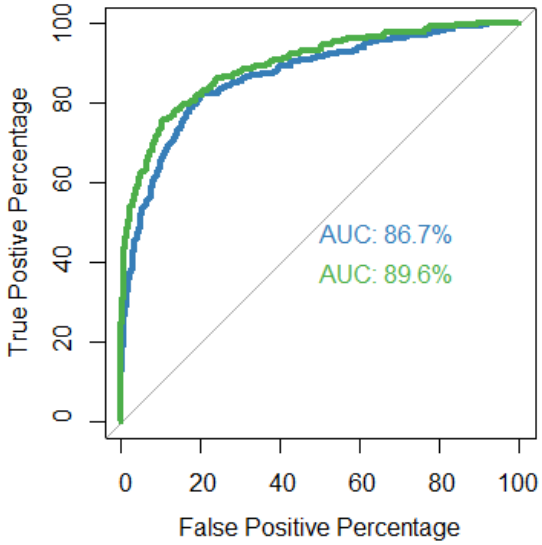
```
logistic_simple <- glm(Attrition~BusinessTravel+Department+Education+EducationField
+EnvironmentSatisfaction+Gender+JobInvolvement+JobLevel+JobRole+JobSatisfaction+Mar
italStatus+OverTime+PerformanceRating+RelationshipSatisfaction+StockOptionLevel+Wor
kLifeBalance, data=attr, family="binomial")
> summary(logistic_simple)

Call:
glm(formula = Attrition ~ BusinessTravel + Department + Education +
    EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +
    JobLevel + JobRole + JobSatisfaction + MaritalStatus + OverTime +
    PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
    WorkLifeBalance, family = "binomial", data = attr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0543  -0.4857  -0.2643  -0.1006   3.3248

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -10.18051  396.43314  -0.026 0.979512
BusinessTravelTravel_Frequently  1.68023    0.40059   4.194 2.74e-05 ***
BusinessTravelTravel_Rarely      0.79631    0.37253   2.138 0.032553 *
DepartmentResearch & Development 13.01387  396.43231   0.033 0.973812
DepartmentSales                  12.09486  396.43291   0.031 0.975661
Education2                        0.23495    0.32119   0.731 0.464480
Education3                        0.24902    0.28611   0.870 0.384107
Education4                        0.17917    0.30845   0.581 0.561317
Education5                        0.02805    0.60790   0.046 0.963197
EducationFieldLife Sciences      -1.38073    0.81317  -1.698 0.089515 .
EducationFieldMarketing          -0.93860    0.85554  -1.097 0.272601
EducationFieldMedical            -1.42088    0.81631  -1.741 0.081749 .
EducationFieldOther              -1.44764    0.88090  -1.643 0.100306
EducationFieldTechnical Degree   -0.48734    0.82724  -0.589 0.555781
EnvironmentSatisfaction2         -0.89578    0.27248  -3.287 0.001011 **
EnvironmentSatisfaction3         -1.02358    0.24634  -4.155 3.25e-05 ***
EnvironmentSatisfaction4         -1.20090    0.25132  -4.778 1.77e-06 ***
GenderMale                        0.35123    0.18329   1.916 0.055333 .
JobInvolvement2                  -1.18453    0.34828  -3.401 0.000671 ***
JobInvolvement3                  -1.52167    0.32887  -4.627 3.71e-06 ***
JobInvolvement4                  -2.12428    0.45688  -4.650 3.33e-06 ***
JobLevel2                        -2.10261    0.42564  -4.940 7.82e-07 ***
```

```
JobLevel3                          -1.24154    0.48889  -2.540 0.011101 *
JobLevel4                          -2.03631    0.74788  -2.723 0.006474 **
JobLevel5                          -0.37731    1.07673  -0.350 0.726025
JobRoleHuman Resources             12.92989  396.43242   0.033 0.973981
JobRoleLaboratory Technician        0.34152    0.57103   0.598 0.549785
JobRoleManager                     -0.53701    1.01974  -0.527 0.598461
JobRoleManufacturing Director       0.16390    0.52217   0.314 0.753604
JobRoleResearch Director           -2.10097    1.03707  -2.026 0.042779 *
JobRoleResearch Scientist          -0.61304    0.58677  -1.045 0.296133
JobRoleSales Executive              2.13192    1.22890   1.735 0.082772 .
JobRoleSales Representative         1.71998    1.31031   1.313 0.189302
JobSatisfaction2                   -0.49575    0.26833  -1.848 0.064666 .
JobSatisfaction3                   -0.51239    0.23642  -2.167 0.030211 *
JobSatisfaction4                   -1.18826    0.25387  -4.681 2.86e-06 ***
MaritalStatusMarried                0.24536    0.27093   0.906 0.365129
MaritalStatusSingle                 0.49710    0.38298   1.298 0.194299
OverTimeYes                         1.95894    0.19152  10.228  < 2e-16 ***
PerformanceRating4                 -0.15678    0.25507  -0.615 0.538778
RelationshipSatisfaction2          -0.67013    0.27822  -2.409 0.016010 *
RelationshipSatisfaction3          -0.77295    0.24948  -3.098 0.001947 **
RelationshipSatisfaction4          -0.75381    0.24903  -3.027 0.002470 **
StockOptionLevel1                  -1.09215    0.30122  -3.626 0.000288 ***
StockOptionLevel2                  -0.97594    0.43405  -2.248 0.024546 *
StockOptionLevel3                  -0.36374    0.43968  -0.827 0.408075
WorkLifeBalance2                   -0.98196    0.36058  -2.723 0.006464 **
WorkLifeBalance3                   -1.38617    0.33593  -4.126 3.69e-05 ***
WorkLifeBalance4                   -0.88866    0.41005  -2.167 0.030217 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.58  on 1469  degrees of freedom
Residual deviance:  876.06  on 1421  degrees of freedom
AIC: 974.06

Number of Fisher Scoring iterations: 14


logistic <- glm(Attrition~Age+BusinessTravel+DailyRate+Department+DistanceFromHome+
Education+EducationField+EnvironmentSatisfaction+Gender+HourlyRate+JobInvolvement+J
obLevel+JobRole+JobSatisfaction+MaritalStatus+MonthlyIncome+MonthlyRate+NumCompanie
sWorked+OverTime+PercentSalaryHike+PerformanceRating+RelationshipSatisfaction+Stock
OptionLevel+TotalWorkingYears+TrainingTimesLastYear+WorkLifeBalance+YearsAtCompany+
YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager, data=attr, family=
"binomial")
> summary(logistic)

Call:
glm(formula = Attrition ~ Age + BusinessTravel + DailyRate +
    Department + DistanceFromHome + Education + EducationField +
    EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement +
    JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome +
    MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
    PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, family = "binomial", data = attr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8204  -0.4400  -0.1959  -0.0546   3.5997

Coefficients:
```

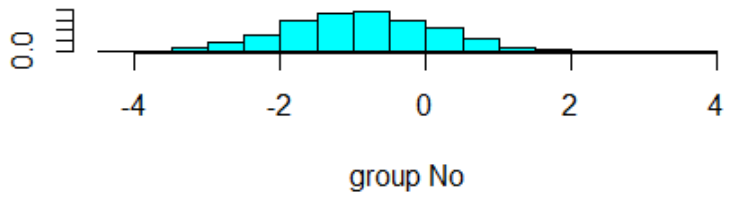
| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.055e+01 | 5.865e+02 | -0.018 | 0.985654 | |
| Age | -3.070e-02 | 1.438e-02 | -2.135 | 0.032798 | * |
| BusinessTravelTravel_Frequently | 2.155e+00 | 4.466e-01 | 4.826 | 1.39e-06 | *** |
| BusinessTravelTravel_Rarely | 1.138e+00 | 4.118e-01 | 2.763 | 0.005736 | ** |
| DailyRate | -4.401e-04 | 2.349e-04 | -1.873 | 0.061004 | . |
| DepartmentResearch & Development | 1.439e+01 | 5.865e+02 | 0.025 | 0.980426 | |
| DepartmentSales | 1.353e+01 | 5.865e+02 | 0.023 | 0.981599 | |
| DistanceFromHome | 5.509e-02 | 1.160e-02 | 4.749 | 2.04e-06 | *** |
| Education2 | 2.637e-01 | 3.489e-01 | 0.756 | 0.449809 | |
| Education3 | 2.433e-01 | 3.062e-01 | 0.795 | 0.426774 | |
| Education4 | 2.652e-01 | 3.344e-01 | 0.793 | 0.427686 | |
| Education5 | 1.776e-01 | 6.356e-01 | 0.280 | 0.779855 | |
| EducationFieldLife Sciences | -1.159e+00 | 8.860e-01 | -1.308 | 0.190930 | |
| EducationFieldMarketing | -6.167e-01 | 9.312e-01 | -0.662 | 0.507849 | |
| EducationFieldMedical | -1.154e+00 | 8.845e-01 | -1.304 | 0.192064 | |
| EducationFieldOther | -1.063e+00 | 9.563e-01 | -1.111 | 0.266434 | |
| EducationFieldTechnical Degree | 1.004e-02 | 8.993e-01 | 0.011 | 0.991089 | |
| EnvironmentSatisfaction2 | -1.078e+00 | 2.939e-01 | -3.669 | 0.000243 | *** |
| EnvironmentSatisfaction3 | -1.210e+00 | 2.702e-01 | -4.480 | 7.48e-06 | *** |
| EnvironmentSatisfaction4 | -1.437e+00 | 2.739e-01 | -5.248 | 1.54e-07 | *** |
| GenderMale | 4.455e-01 | 1.964e-01 | 2.268 | 0.023331 | * |
| HourlyRate | 4.348e-03 | 4.763e-03 | 0.913 | 0.361358 | |
| JobInvolvement2 | -1.261e+00 | 3.773e-01 | -3.344 | 0.000827 | *** |
| JobInvolvement3 | -1.563e+00 | 3.558e-01 | -4.394 | 1.11e-05 | *** |
| JobInvolvement4 | -2.185e+00 | 4.914e-01 | -4.447 | 8.72e-06 | *** |
| JobLevel2 | -1.574e+00 | 4.815e-01 | -3.269 | 0.001078 | ** |
| JobLevel3 | 1.525e-01 | 7.397e-01 | 0.206 | 0.836709 | |
| JobLevel4 | -7.741e-01 | 1.279e+00 | -0.605 | 0.545064 | |
| JobLevel5 | 2.324e+00 | 1.675e+00 | 1.388 | 0.165252 | |
| JobRoleHuman Resources | 1.470e+01 | 5.865e+02 | 0.025 | 0.979999 | |
| JobRoleLaboratory Technician | 6.024e-01 | 6.078e-01 | 0.991 | 0.321687 | |
| JobRoleManager | -2.077e-01 | 1.117e+00 | -0.186 | 0.852479 | |
| JobRoleManufacturing Director | 4.526e-01 | 5.637e-01 | 0.803 | 0.422078 | |
| JobRoleResearch Director | -1.925e+00 | 1.179e+00 | -1.633 | 0.102438 | |
| JobRoleResearch Scientist | -5.351e-01 | 6.301e-01 | -0.849 | 0.395796 | |
| JobRoleSales Executive | 2.196e+00 | 1.290e+00 | 1.702 | 0.088682 | . |
| JobRoleSales Representative | 1.936e+00 | 1.371e+00 | 1.412 | 0.157955 | |
| JobSatisfaction2 | -6.562e-01 | 2.882e-01 | -2.277 | 0.022785 | * |
| JobSatisfaction3 | -6.401e-01 | 2.559e-01 | -2.501 | 0.012370 | * |
| JobSatisfaction4 | -1.287e+00 | 2.734e-01 | -4.710 | 2.48e-06 | *** |
| MaritalStatusMarried | 3.095e-01 | 2.903e-01 | 1.066 | 0.286405 | |
| MaritalStatusSingle | 6.250e-01 | 4.151e-01 | 1.506 | 0.132129 | |
| MonthlyIncome | -1.295e-04 | 9.565e-05 | -1.354 | 0.175878 | |
| MonthlyRate | 1.001e-05 | 1.325e-05 | 0.756 | 0.449743 | |
| NumCompaniesWorked | 2.116e-01 | 4.152e-02 | 5.097 | 3.44e-07 | *** |
| OverTimeYes | 2.192e+00 | 2.120e-01 | 10.343 | < 2e-16 | *** |
| PercentSalaryHike | -2.061e-02 | 4.133e-02 | -0.499 | 0.618027 | |
| PerformanceRating4 | 1.099e-01 | 4.242e-01 | 0.259 | 0.795545 | |
| RelationshipSatisfaction2 | -9.523e-01 | 3.048e-01 | -3.124 | 0.001784 | ** |
| RelationshipSatisfaction3 | -1.008e+00 | 2.704e-01 | -3.729 | 0.000192 | *** |
| RelationshipSatisfaction4 | -1.001e+00 | 2.690e-01 | -3.721 | 0.000198 | *** |
| StockOptionLevel1 | -1.150e+00 | 3.275e-01 | -3.511 | 0.000446 | *** |
| StockOptionLevel2 | -1.107e+00 | 4.553e-01 | -2.433 | 0.014993 | * |
| StockOptionLevel3 | -3.613e-01 | 4.869e-01 | -0.742 | 0.458012 | |
| TotalWorkingYears | -5.933e-02 | 3.107e-02 | -1.910 | 0.056155 | . |
| TrainingTimesLastYear | -1.924e-01 | 7.649e-02 | -2.515 | 0.011901 | * |
| WorkLifeBalance2 | -9.456e-01 | 3.913e-01 | -2.416 | 0.015682 | * |
| WorkLifeBalance3 | -1.475e+00 | 3.678e-01 | -4.009 | 6.11e-05 | *** |
| WorkLifeBalance4 | -1.064e+00 | 4.452e-01 | -2.391 | 0.016812 | * |
| YearsAtCompany | 1.013e-01 | 4.301e-02 | 2.356 | 0.018469 | * |
| YearsInCurrentRole | -1.429e-01 | 5.187e-02 | -2.755 | 0.005873 | ** |
| YearsSinceLastPromotion | 1.726e-01 | 4.605e-02 | 3.749 | 0.000178 | *** |
| YearsWithCurrManager | -1.541e-01 | 5.069e-02 | -3.040 | 0.002369 | ** |

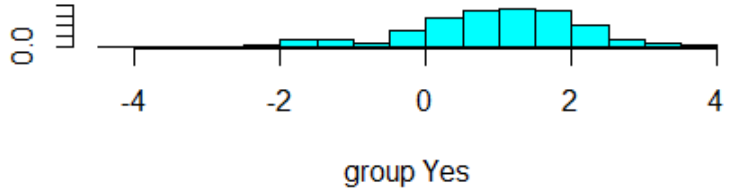
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.6  on 1469  degrees of freedom
Residual deviance:  775.3  on 1407  degrees of freedom
AIC: 901.3

Number of Fisher Scoring iterations: 15
```

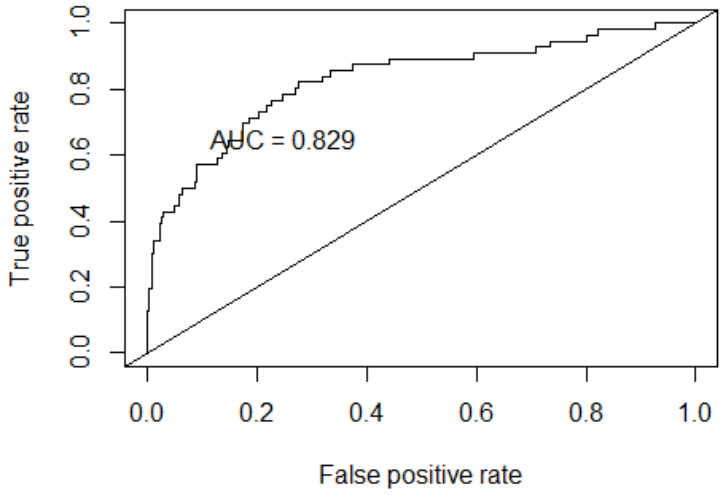


# Multiple Discriminant Analysis :



# CONCLUSION :

The **logistic regression model** is the best model for this dataset as this model has an AUC of **89.6%.** The AUC provides an aggregate measure of performance across all possible classification thresholds. This means that the accuracy of the predictions of this model is **89.6%**.