Human Resource Analytics using R

Aman Mahajan

PROBLEM STATEMENT: Companies face the problem that their human resources on whom the company have invested time and money to train them, leave the company voluntarily. It is important for the management and stakeholders to know the variables responsible for employees quitting jobs and also have a prediction that which employees will be quitting their jobs in future.

Goal: To predict whether an employee will stay or leave the company within the next year.

Dataset: Humanresources Dataset from Kaggle.com with 11111 observations and 8 variables. This is a historical data giving us the information who did and who did not leave the company within the last year. In this dataset, we will be predicting the variable “vol_leave” (0 = stay, 1 = leave) using the other variables.

Now, installing all the required datasets.

```
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.4.3
```

```
library(RColorBrewer)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.4
```

```
## Loading required package: rpart
```

```
library(ellipse)
```

```
## Warning: package 'ellipse' was built under R version 3.4.4
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:ellipse':
##
##     ellipse
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.4.3
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
## The following object is masked from 'package:rpart':
##
##     solder
```

```
## The following object is masked from 'package:plyr':
##
##     ozone
```

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.4.3
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

Calling the dataset:

```r
humanresources <- read.csv("D:/humanresource.csv")
str(humanresources)
```

```
## 'data.frame':    11111 obs. of  8 variables:
##  $ role     : Factor w/ 5 levels "CEO","Director",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ perf     : int  3 3 1 2 3 1 2 3 2 1 ...
##  $ area     : Factor w/ 5 levels "Accounting","Finance",..: 5 3 2 5 3 4 1 2 5 3 ..
.
##  $ sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1 1 1 ...
##  $ id       : int  1 32 76 69 28 77 70 103 71 25 ...
##  $ age      : num  62 53.4 53.5 49.2 49.8 ...
##  $ salary   : num  1000000 258935 189828 207492 188205 ...
##  $ vol_leave: int  0 0 1 0 0 0 0 0 1 0 ...
```

```
summary(humanresources)
```

```
##       role             perf               area            sex      
##  CEO     :    1   Min.   :1.000   Accounting:1609   Female:6068  
##  Director:  100   1st Qu.:2.000   Finance   :1677   Male  :5043  
##  Ind     :10000   Median :2.000   Marketing :2258                
##  Manager : 1000   Mean   :2.198   Other     :2198                
##  VP      :   10   3rd Qu.:3.000   Sales     :3369                
##                   Max.   :3.000                                  
##        id             age            salary          vol_leave     
##  Min.   :    1   Min.   :22.02   Min.   :  42168   Min.   :0.0000  
##  1st Qu.: 2778   1st Qu.:24.07   1st Qu.:  57081   1st Qu.:0.0000  
##  Median : 5556   Median :25.70   Median :  60798   Median :0.0000  
##  Mean   : 5556   Mean   :27.79   Mean   :  65358   Mean   :0.3812  
##  3rd Qu.: 8334   3rd Qu.:28.49   3rd Qu.:  64945   3rd Qu.:1.0000  
##  Max.   :11111   Max.   :62.00   Max.   :1000000   Max.   :1.0000  
```

As we see, there 5 kinds of roles in the dataset, namely, CEO, Director, Ind, Manager and VP. But since, CEO and VP fall in a seperate segment than the other job roles, we will not include them in our model. Therefore, now calling the data again and summarizing it.

```
humanresources = filter(humanresources, humanresources$role == "Ind" |
humanresources$role == "Manager" | humanresources$role == "Director")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

```
humanresources$role <- factor(humanresources$role)
summary(humanresources)
```

```
##        role           perf                area          sex
##  Director:  100   Min.   :1.000   Accounting:1607   Female:6064
##  Ind      :10000   1st Qu.:2.000   Finance    :1676   Male  :5036
##  Manager : 1000   Median :2.000   Marketing :2255
##                    Mean   :2.198   Other      :2197
##                    3rd Qu.:3.000   Sales      :3365
##                    Max.   :3.000
##        id              age             salary          vol_leave
##  Min.   :   12   Min.   :22.02   Min.   : 42168   Min.   :0.0000
##  1st Qu.: 2787   1st Qu.:24.07   1st Qu.: 57080   1st Qu.:0.0000
##  Median : 5562   Median :25.70   Median : 60788   Median :0.0000
##  Mean   : 5562   Mean   :27.77   Mean   : 64860   Mean   :0.3815
##  3rd Qu.: 8336   3rd Qu.:28.48   3rd Qu.: 64928   3rd Qu.:1.0000
##  Max.   :11111   Max.   :61.67   Max.   :311131   Max.   :1.0000
```

Since, the response output variable consist of two groups i.e. (0, 1), comparing it with other columns would be much easier if we use an aggregate function.

1. Performance v/s Voluntarily Leaving

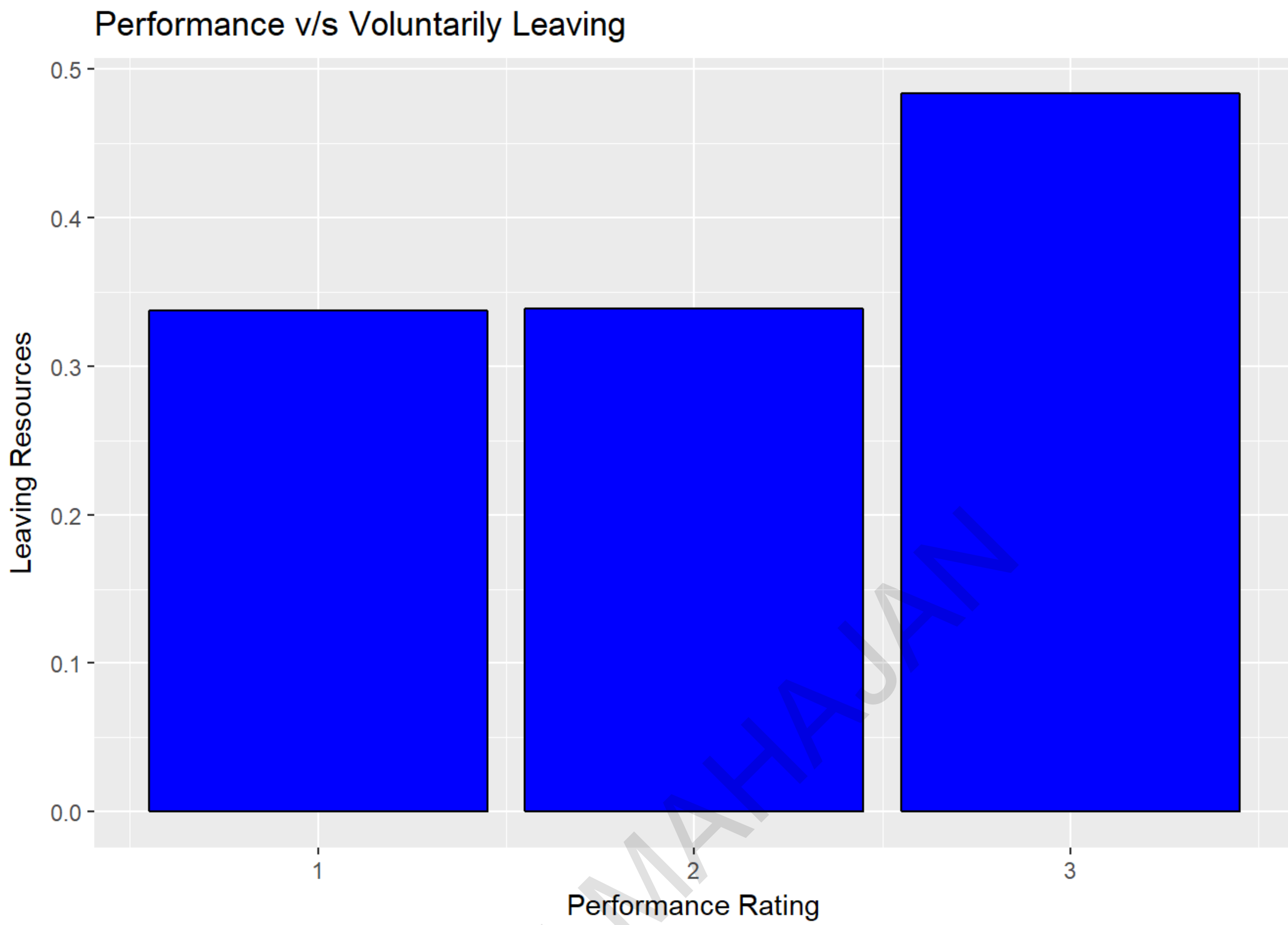
```
agg_perf = aggregate(vol_leave ~ perf, data = humanresources, mean)
agg_perf
```

perf <int>	vol_leave <dbl>
1	0.3375112
2	0.3383831
3	0.4831122

3 rows

Analysis:

```
ggplot(agg_perf, aes(x = perf, y = vol_leave)) + geom_bar(stat =
"identity", fill = 'blue', colour = 'black') + ggtitle("Performance v/s Voluntarily L
eaving") + labs(y = "Leaving Resources", x =
"Performance Rating")
```



Inference:

The histogram plot shows that the employees with higher performance rating are more likely to leave the company.

2. Sex v/s Voluntarily Leaving

```
agg_sex = aggregate(vol_leave ~ sex, data = humanresources, mean)
agg_sex
```

sex <fctr>	vol_leave <dbl>
Female	0.4673483
Male	0.2781970

2 rows

Analysis:

```
ggplot(agg_sex, aes(x = sex, y = vol_leave)) + geom_bar(stat = "identity",
fill = 'red', colour = 'black') + ggtitle("Sex v/s Voluntary Leaving") + labs(y = "Le
aving Resources", x = "Sex")
```

Inference:

The plot shows that the female employees are more likely to leave their jobs as compared to their male counterparts.

3. Business Area v/s Voluntarily Leaving

```
agg_area = aggregate(vol_leave ~ area, data = humanresources, mean)
agg_area
```

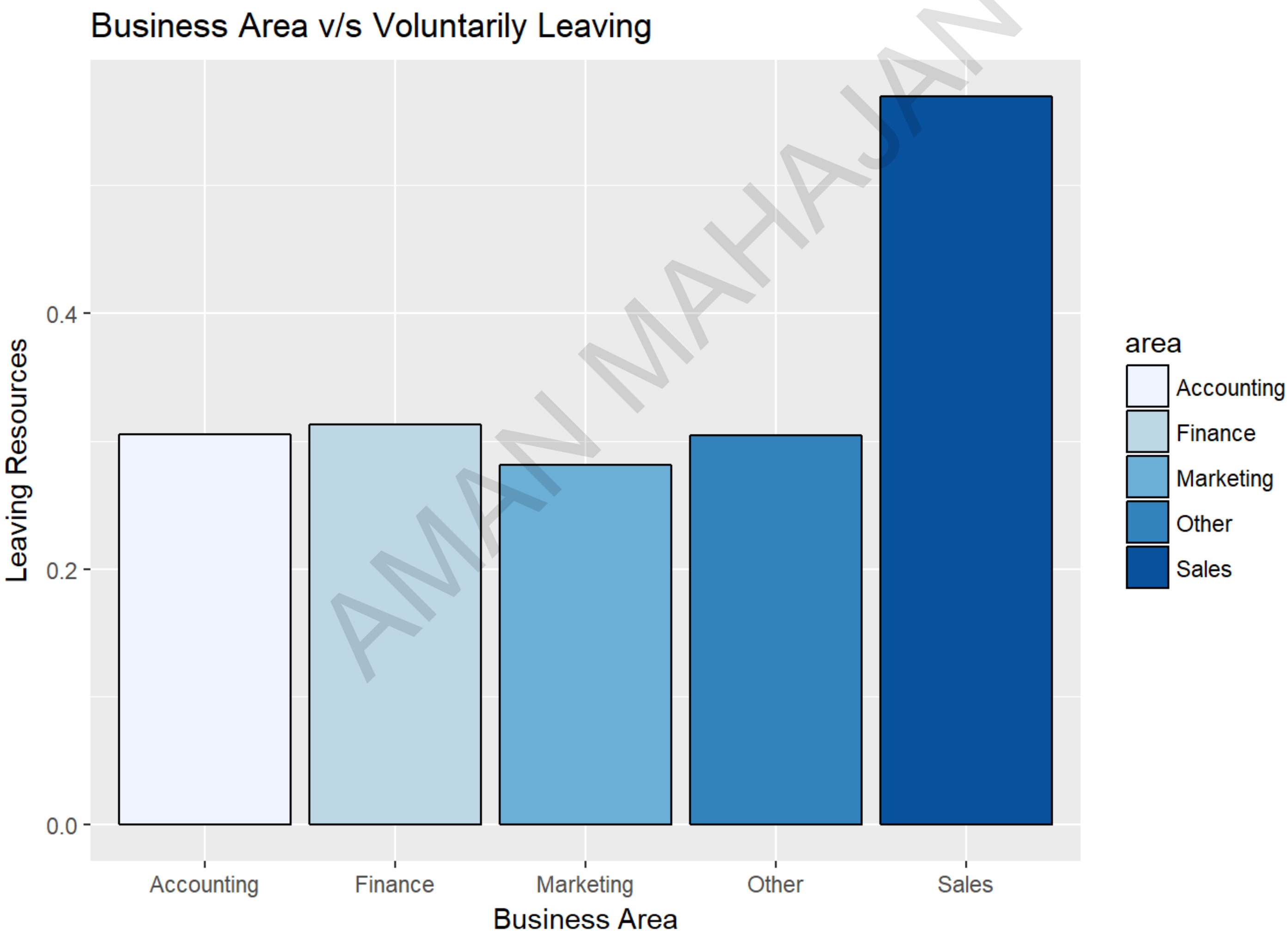
area <fctr>	**vol_leave** <dbl>
Accounting	0.3055383

Finance	0.3126492
Marketing	0.2815965
Other	0.3040510
Sales	0.5696880

5 rows

Analysis:

```
ggplot(agg_area, aes(x = area, y = vol_leave, fill = area)) + geom_bar(stat =
"identity", colour = "black") + scale_fill_brewer() + ggtitle("Business Area v/s Volu
ntarily Leaving") + labs(y = "Leaving Resources", x = "Business Area")
```



Inference:

Employees from Sales department are most likely to leave their k=jobs as compared to other Business Areas in the company.

4. Role v/s Voluntarily Leaving

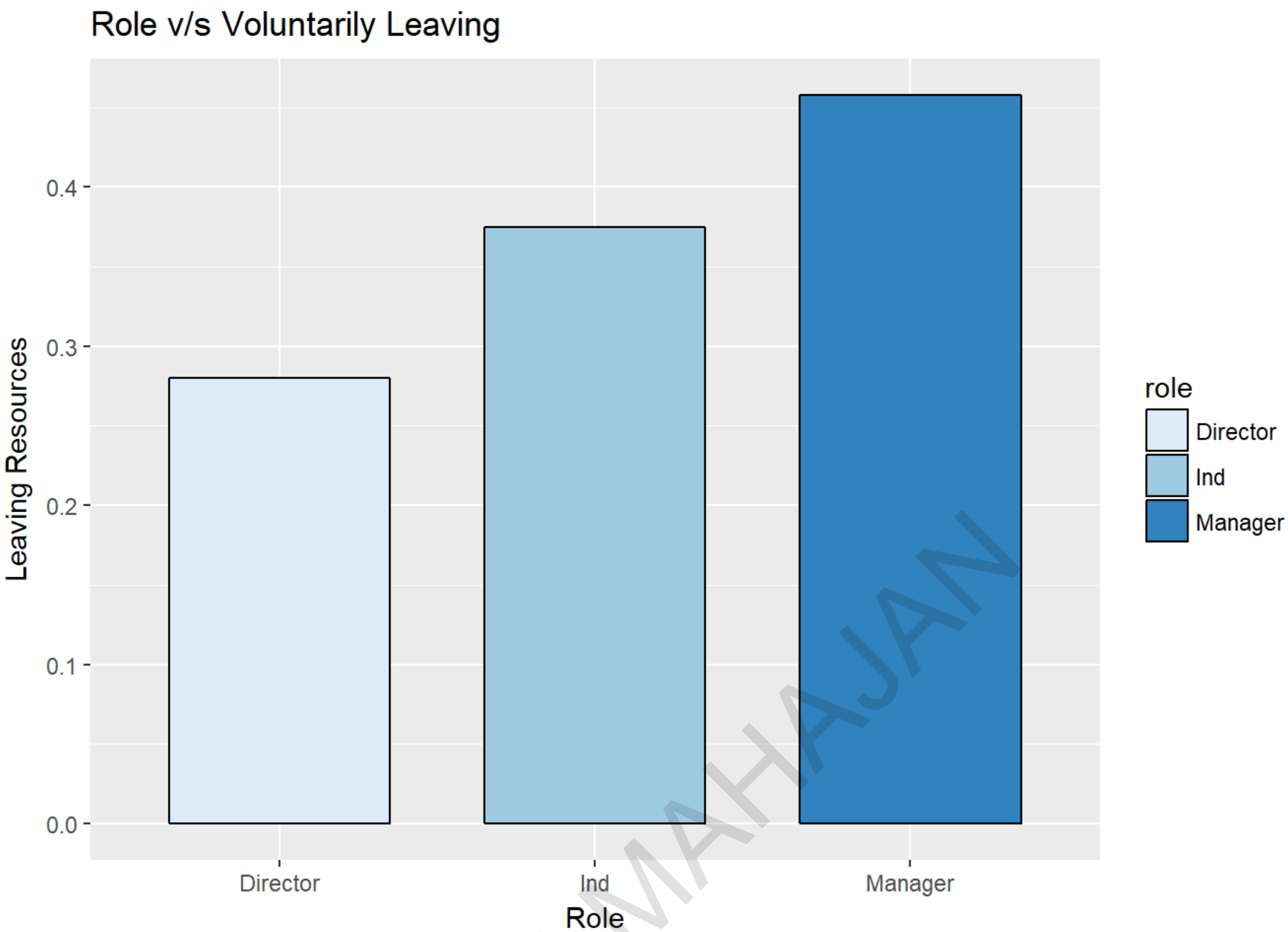
```
agg_role = aggregate(vol_leave ~ role, data = humanresources, mean)
agg_role
```

role <fctr>	vol_leave <dbl>
Director	0.2800
Ind	0.3749
Manager	0.4580

3 rows

Analysis:

```
ggplot(agg_role, aes(x = role, y = vol_leave, fill = role)) + geom_bar(stat =
"identity", width = .7, colour = 'black') + scale_fill_brewer() + ggtitle("Role v/s V
oluntarily Leaving") + labs (y = "Leaving Resources", x
= "Role")
```

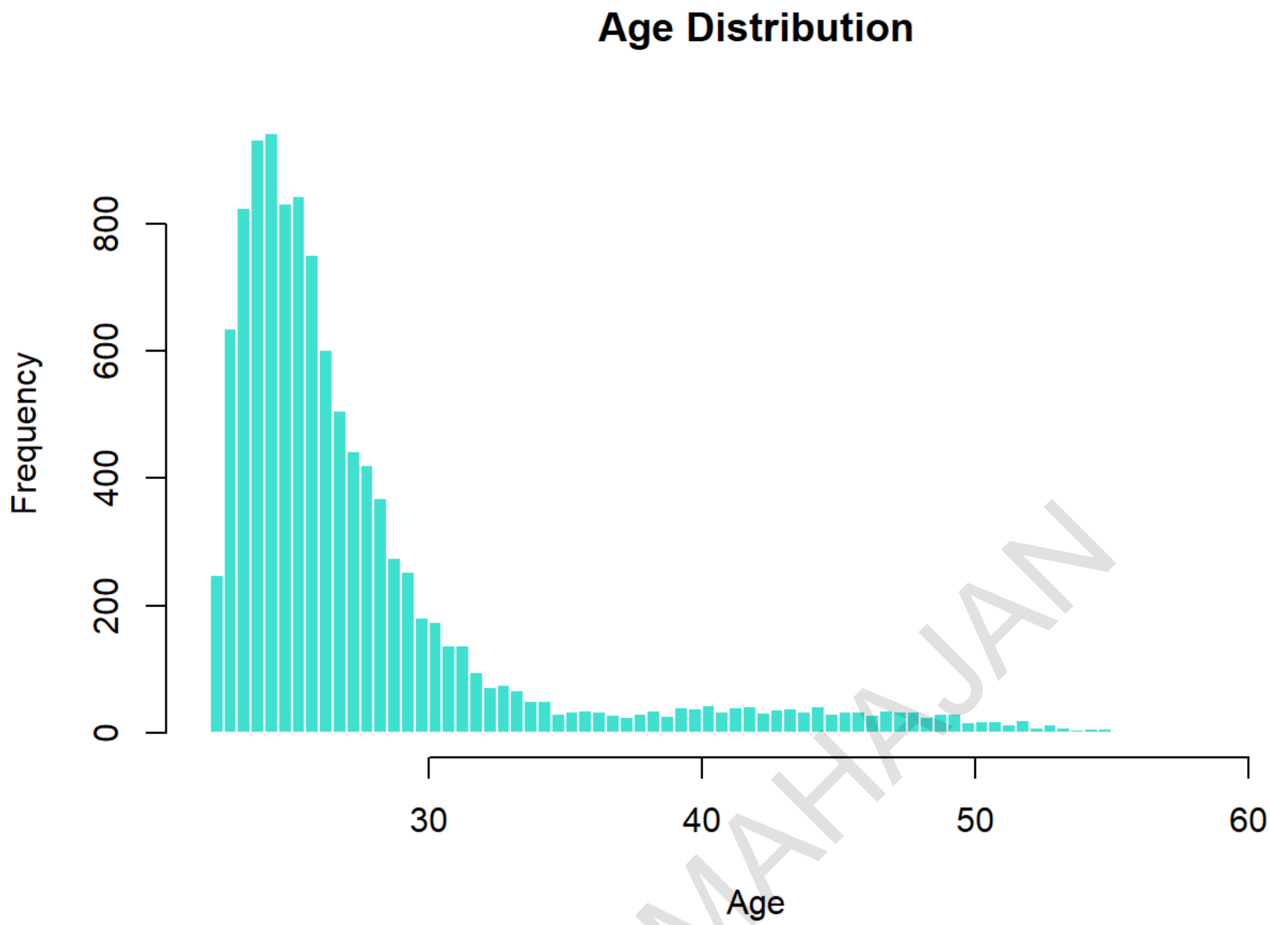


Inference:

Managers are most likely to leave their jobs whilst Directors are least likely to leave their jobs.

4. Age v/s Voluntarily Leaving

```
hist(humanresources$age, breaks = 100, main = "Age Distribution", border = F,
xlab = "Age", col = 'turquoise')
```




```
quantile(humanresources$age, probs = seq(0,1,.1))
```

```
##        0%      10%      20%      30%      40%      50%      60%      70%
## 22.02289 23.14094 23.76757 24.36880 25.01564 25.69533 26.55048 27.73737
##      80%      90%     100%
## 29.51513 35.70077 61.67132
```

Inference:

90% of the employees fall in the age bracket of 22 years to 36 years. This categorization looks skewed.

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.4.3
```

```
skewness(humanresources$age)
```

```
## [1] 2.2669
```

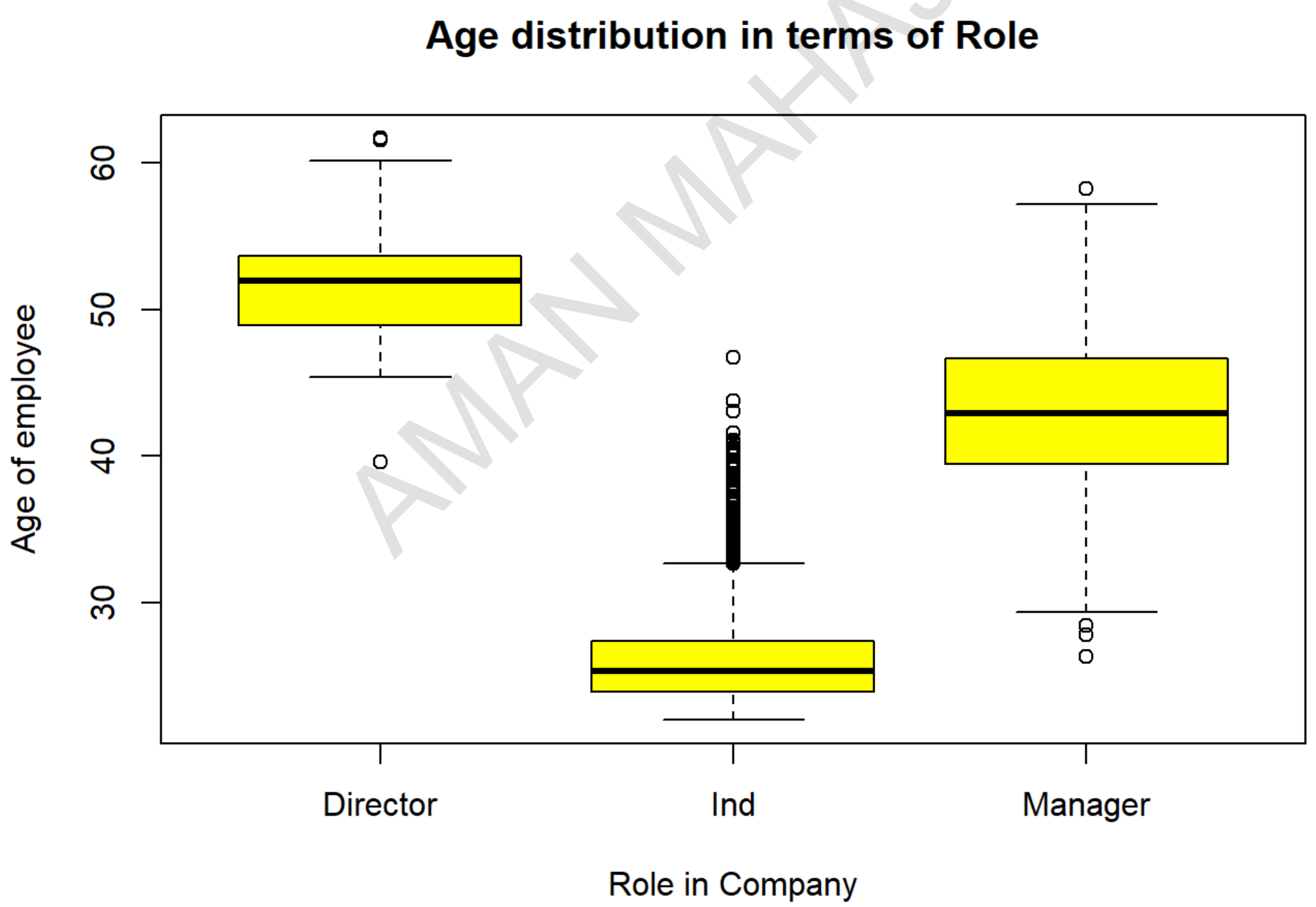
We see that the age distribution is Positive/Right Skewed which implies that the Mean is less than the Median. Therefore, taking the log of the age variable.

```
humanresources$log_age = log(humanresources$age)
summary(humanresources$log_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.092   3.181   3.246   3.304   3.349   4.122
```

Let us categorize the age distribution further in terms of employee roles in the company.

```
boxplot(age ~ role, data = humanresources, col = 'yellow', xlab = 'Role in Company',
ylab = 'Age of employee', main = 'Age distribution in terms of Role')
```

The above box plot shows that there is a relatioship between employee role in company to his/her age. Directors fall in the higher age bracket while the Ind employees fall in the lower to mid age bracket.

Let us now aggregate the age variable to see the relation with employee leaving.

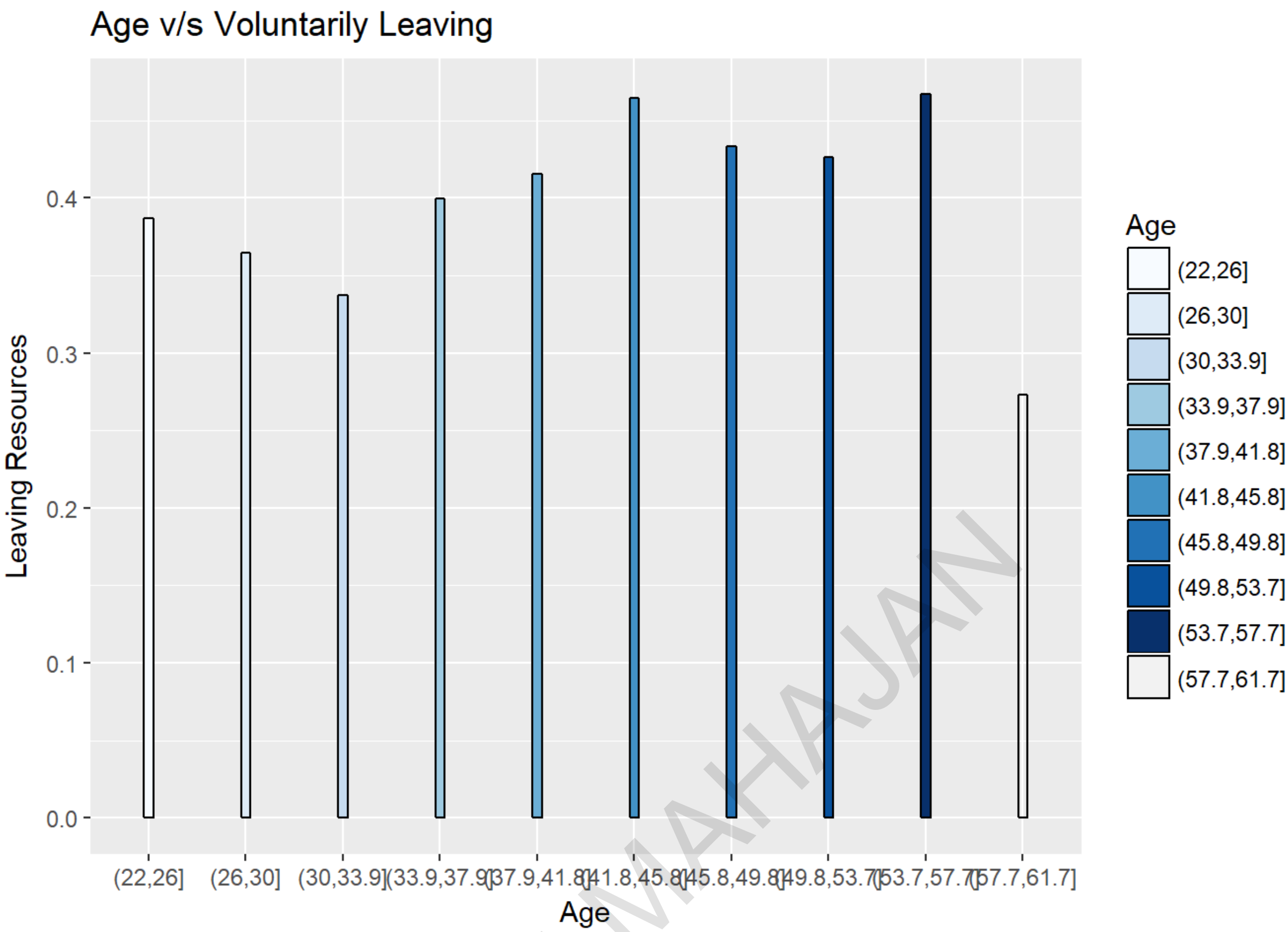
```
agg_age = aggregate(x = humanresources$vol_leave, by = list(cut(humanresources$age, 1
0)), mean)
agg_age
```

Group.1 <fctr>	x <dbl>
(22,26]	0.3866177
(26,30]	0.3645902
(30,33.9]	0.3374536
(33.9,37.9]	0.3992806
(37.9,41.8]	0.4155405
(41.8,45.8]	0.4640288
(45.8,49.8]	0.4333333
(49.8,53.7]	0.4260870
(53.7,57.7]	0.4666667
(57.7,61.7]	0.2727273

1-10 of 10 rows


```
names(agg_age) = c("Age", "Probability")
ggplot(agg_age, aes(x = Age, y = Probability, fill = Age)) + geom_bar(stat =
"identity", width = .1, colour = 'black') + scale_fill_brewer() +
ggtitle("Age v/s Voluntarily Leaving") + labs(y = "Leaving Resources", x = "Age")
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for pale
tte Blues is 9
## Returning the palette you asked for with that many colors
```



The above plot shows that the employees with age from 42 to 57 are most likely to leave the jobs as compared to employees with age 22 to 41. And employees with age over 57 are least likely to leave the job, since that is usually the CEO and Director job role.

5. Analyzing Salary variable

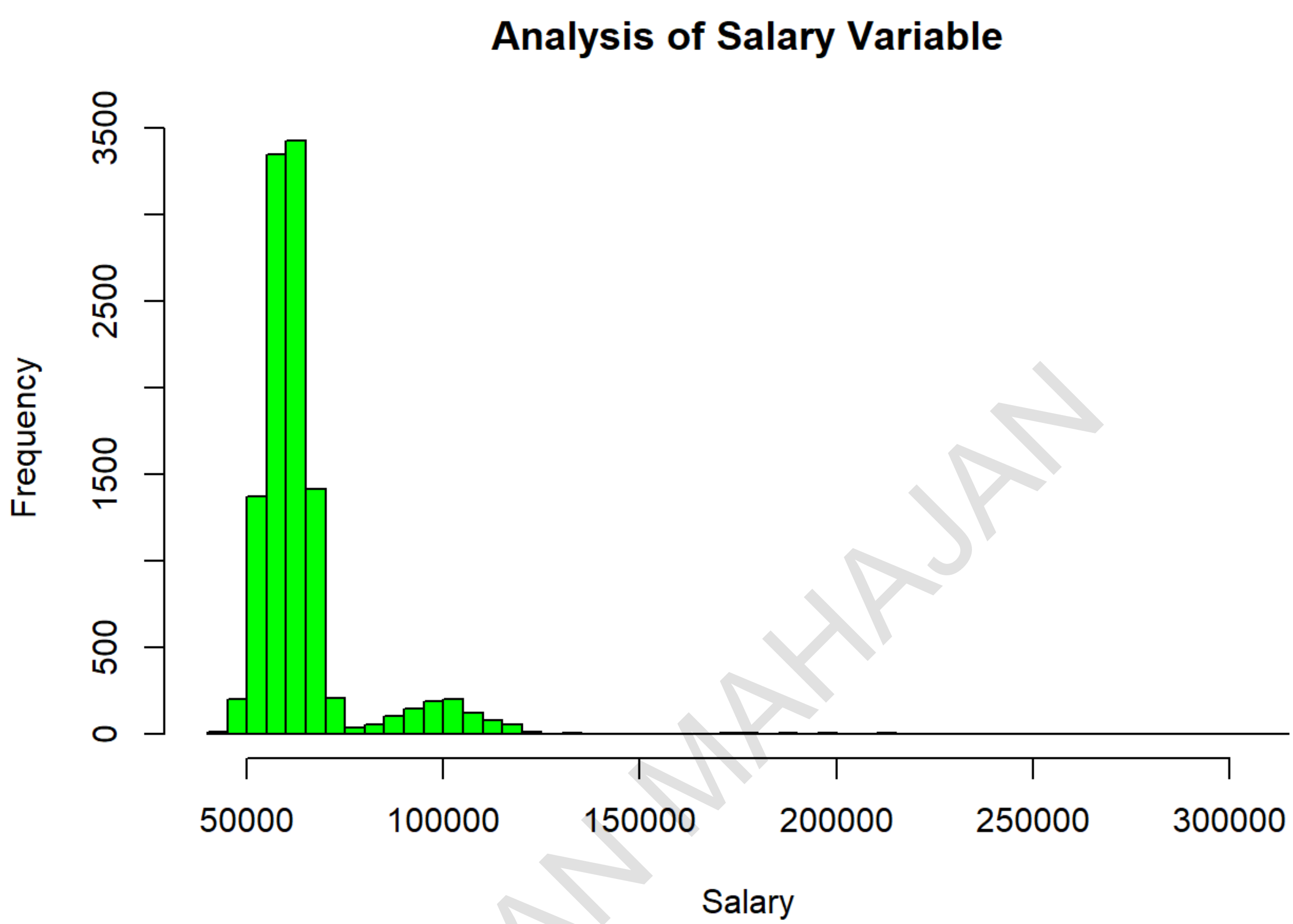
```
summary(humanresources$salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   42168   57080   60788   64860   64928  311131
```

```
quantile(humanresources$salary, probs = seq(0,1,.2))
```

```
##        0%       20%       40%       60%       80%      100% 
##  42168.22  56189.17  59385.03  62307.14  66151.43 311130.51
```

```
hist(humanresources$salary, breaks = 50, col = 'green', main = "Analysis of Salary Va
riable", xlab = "Salary")
```



Inference:

The median salary of employees is $60788. But maximum (80%) employees are earning till $66173.65

DATA MODELLING: Firstly, we need to split our data into a training set and test set. Two thirds of data is dedicated to training dataset and one third is dedicated to testing dataset

```
set.seed(42)
split_data = sample.split(humanresources$vol_leave, 2/3)
train = humanresources[split_data,]
test = humanresources[!split_data,]
```

a. LOGISTIC REGRESSION We have a classification proble, with outcomes being ‘Staying’ or ‘Leaving’ predicted through the significant variables. So, we use logistic regression to fit the model.

```
test_mean = mean(test$vol_leave)
train_mean = mean(train$vol_leave)
print(c(test_mean, train_mean))
```

```
## [1] 0.3816216 0.3814865
```

Fitting the model using genralized linear model (GLM)

```
fit = glm(vol_leave ~ role + perf + area + sex + log_age + salary, data
= humanresources, family = 'binomial')
summary(fit)
```

```
##
## Call:
## glm(formula = vol_leave ~ role + perf + area + sex + log_age +
##     salary, family = "binomial", data = humanresources)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4737  -0.9123  -0.6068   1.0906   3.2238
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.290e+01  1.100e+00  11.725  < 2e-16 ***
## roleInd       -8.146e+00  5.573e-01 -14.617  < 2e-16 ***
## roleManager   -4.865e+00  4.327e-01 -11.242  < 2e-16 ***
## perf           4.931e-01  3.598e-02  13.703  < 2e-16 ***
## areaFinance    3.517e-02  7.920e-02   0.444 0.657003
## areaMarketing -9.517e-02  7.490e-02  -1.271 0.203862
## areaOther     -9.540e-05  7.471e-02  -0.001 0.998981
## areaSales      1.239e+00  6.799e-02  18.230  < 2e-16 ***
## sexMale       -9.435e-01  4.374e-02 -21.571  < 2e-16 ***
## log_age       -7.516e-01  2.037e-01  -3.689 0.000225 ***
## salary        -6.515e-05  3.723e-06 -17.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14759  on 11099  degrees of freedom
## Residual deviance: 13004  on 11089  degrees of freedom
## AIC: 13026
##
## Number of Fisher Scoring iterations: 4
```

Now, checking the p-value for all the independent variables, we see that areaFinance, areaMarketing, areaOther are insignaficant factors as the p-value is greater than 0.05

No, we will analyze the deviance to test the differneces between two or more means through ANOVA (Analysis of Variance) test using Chi-square method.

```
anova(fit, test = "Chisq")
```

	Df <int>	Deviance <dbl>	Resid. Df <int>	Resid. Dev <dbl>	Pr(>Chi) <dbl>
NULL	NA	NA	11099	14758.76	NA
role	2	30.69441	11097	14728.06	2.161692e-07
perf	1	161.13871	11096	14566.92	6.380562e-37
area	4	735.01843	11092	13831.91	9.103811e-158
sex	1	466.68696	11091	13365.22	1.685450e-103
log_age	1	11.20500	11090	13354.01	8.157716e-04
salary	1	350.08268	11089	13003.93	4.065693e-78

7 rows

Deviance is a measure of goodness of fit for a model. The differnece between the null deviance and residual deviance along with the low values of p shows all the significant variables.

Now, analyzing the predictive ability of our model through Confusion Matrix

```
pred_model = predict(fit, test, type = 'response')
pred_model = ifelse(pred_model > 0.5,1,0)
MCE = mean(pred_model != test$vol_leave)

table(actual = test$vol_leave, prediction = pred_model)
```

```
##       prediction
## actual    0    1
##      0 1919  369
##      1  780  632
```

Calculating the Accuracy of our Logistic Regression model:

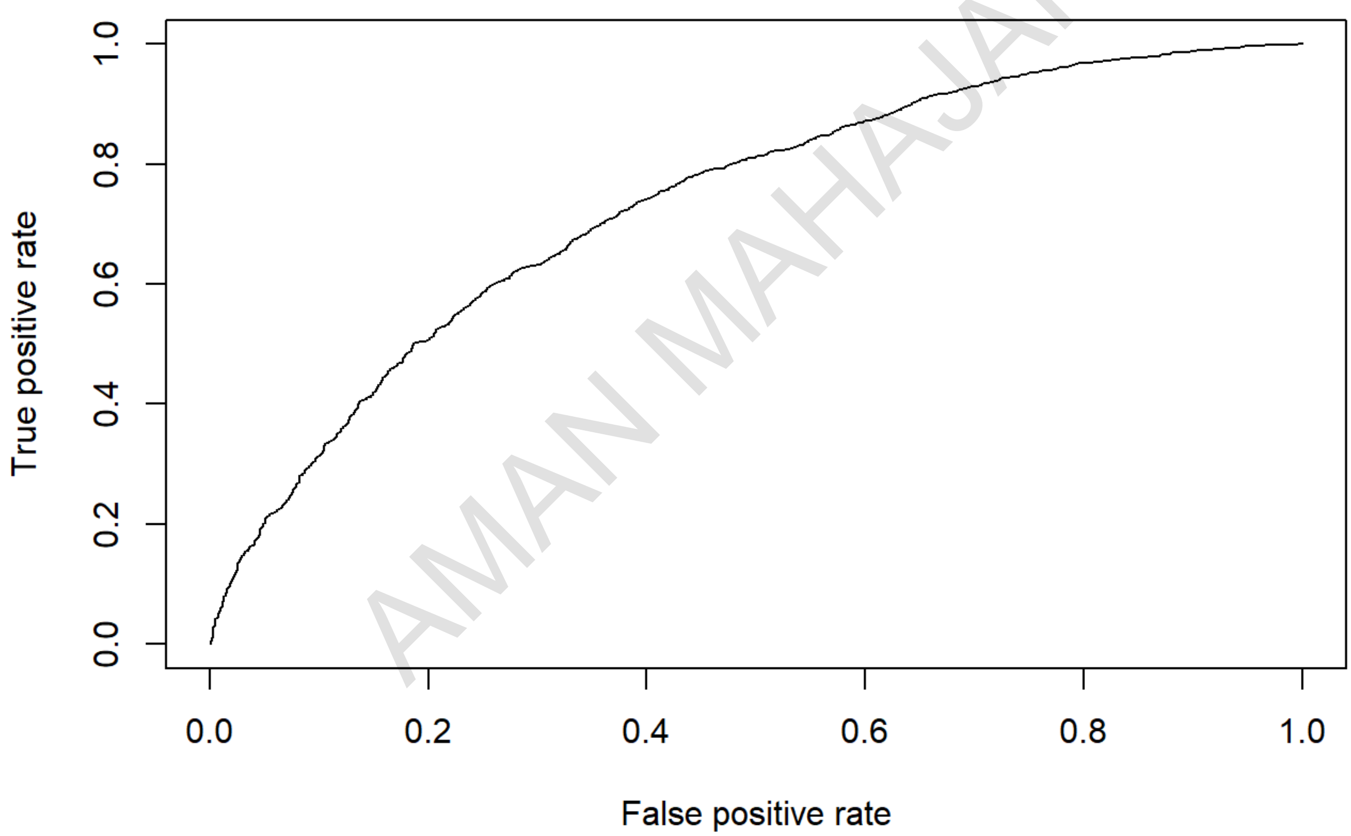
```
print(paste('Accuracy', 1 - MCE))
```

```
## [1] "Accuracy 0.689459459459459"
```

ACCURACY of our model is 68.94%

Lastly, we are going to plot the ROC curve and calculate the AUC (area under the curve) which are typical performance measurements for a binary classifier.

```
plot1 = predict(fit, test, type = "response")
plot2 = prediction(plot1, test$vol_leave)
plot3 = performance(plot2, measure = "tpr", x.measure = "fpr")
plot(plot3)
```



Calculating the Area Under the Curve (AUC)

```
AUC = performance(plot2, measure = "auc")
AUC = AUC@y.values[[1]]
AUC
```

```
## [1] 0.7326298
```

Based on the rule of thumb, a model has a good predictive ability if the AUC is closer to 1. As per our analysis, the AUC of 0.73 is closer to 1, therefore, the logistic regression model has a good predictive ability.

MODEL 2 b) DECISION TREES

Let us start by fitting the model

```
set.seed(42)
dt = rpart(vol_leave ~ role + perf + age + sex + area + salary,
data = train, method = "class")
dt
```

```
## n= 7400
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 7400 2823 0 (0.6185135 0.3814865)
##    2) area=Accounting,Finance,Marketing,Other 5188 1544 0 (0.7023901 0.2976099) *
##    3) area=Sales 2212  933 1 (0.4217902 0.5782098)
##      6) sex=Male 1015  479 0 (0.5280788 0.4719212)
##       12) perf< 2.5 682  281 0 (0.5879765 0.4120235) *
##       13) perf>=2.5 333  135 1 (0.4054054 0.5945946) *
##      7) sex=Female 1197  397 1 (0.3316625 0.6683375) *
```

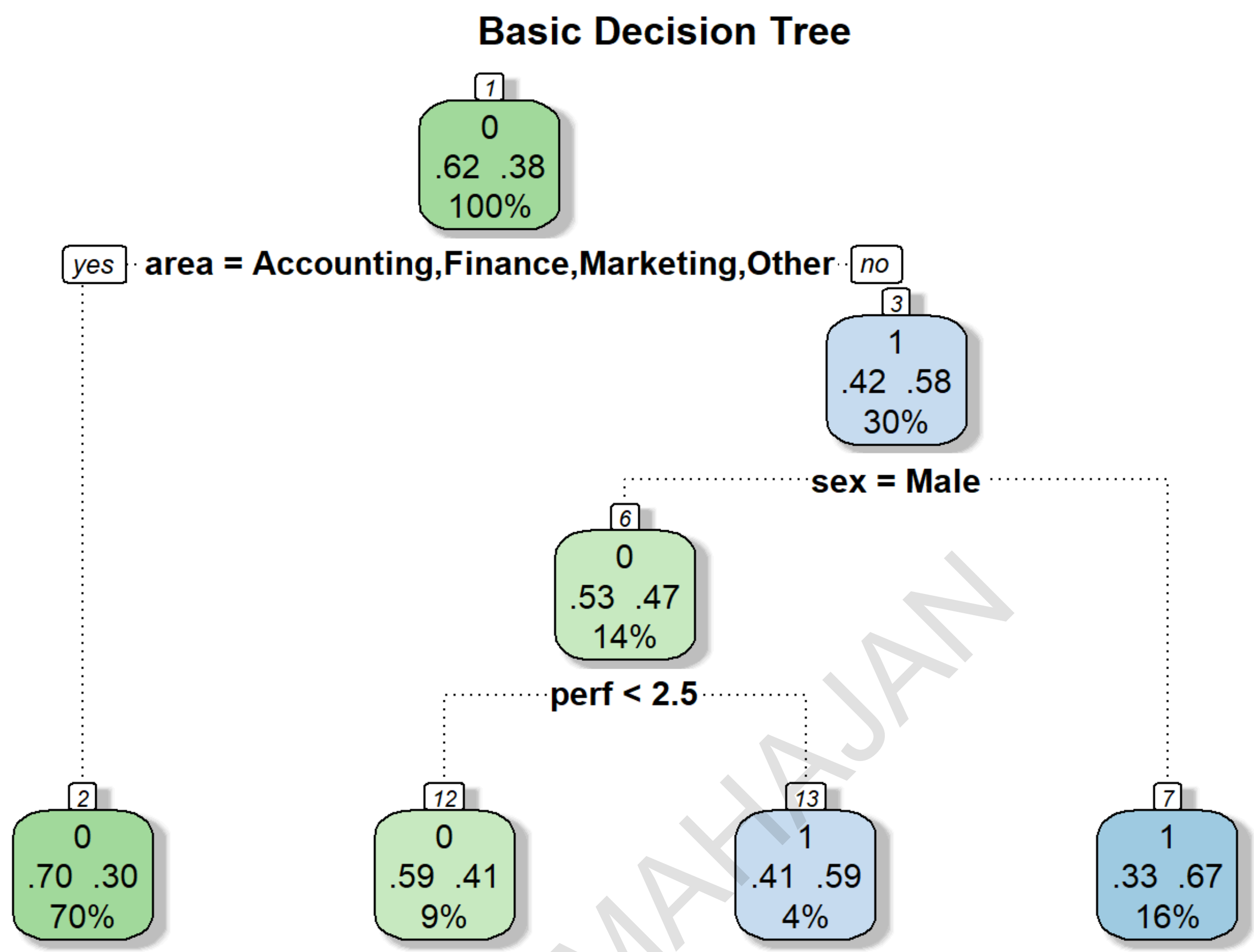
Plotting the Decision Tree:

```
library(rattle)
```

```
## Warning: package 'rattle' was built under R version 3.4.4
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
par(mar = c(5,4,1,2))
fancyRpartPlot(dt, sub = NULL, main = "Basic Decision Tree")
```



Analysis: The first node means the root. Here, 62% of those in our training data have 0 (Stay) for the response variable and 38% have a 1 (Leave). Below that, we see our first decision node. In the event that our workers are in the Accounting, Finance, Marketing, or Other regions, then we say 'yes' and take the left branch. On the off chance that the answer is 'no' (i.e. they are in Sales), then we take the right branch. After the left branch, we see that it ends into a solitary node for all of those who are not in Sales. For all of these people, the most common response is '0' (Stay), with 70% employee who will stay in the company and only 30% in this bucket will leave the company. The '70%' reported in the bottom of the node tells us that this single bucket accounts for 70% of the total sample we are modeling. On following the right branch, we see that the most well-known reaction is '1' for the employee who will leave the company. Moreover, the node is likewise letting us know 42% of employees in this bucket will stay while 58% will leave. Proceeding with the right branch, if the worker is male, we say 'yes' and go to the left side. On the off chance that the worker is female, we go right. For females, we wind up in a terminating node that has a dominant response of 1 (33% - Stay and 67% - Leave). This ending node represents 16% of the aggregate populace. For male, we further go down to performance variable. If the performance is less than 2.5 we go left else we go right. For performance less than 2.5, we wind up in a terminating node that has a dominant response of 0 (59% - Stay and 41% - Leave). This ending node represents 16%. For performance greater than 2.5, we wind up in a terminating node that has a dominant response of 1 (33% - Stay and 67% - Leave). This ending node represents 4%.

Now, analyzing the pradictive ability of the model using Confusion Matrix:

```
pred_dt = predict(dt, test, type = 'class')

cm_dt = table(actual = test$vol_leave, prediction = pred_dt)

cm_dt
```

```
##         prediction
## actual    0    1
##       0 2006  282
##       1  930  482
```

Calculating Accuracy for the model:

```
accuracy = sum(diag(cm_dt))/sum(cm_dt)
accuracy
```

```
## [1] 0.6724324
```

Accuracy of the Decison Tree model is 67.24%

CONCLUSION: Logistic Regression is better than decision tree in predicting the output response variable to predict whether the employee will Stay in the Company or Leave the Company in future.