HR-Analytics.R

Shraddha Somani

```
library(plyr)
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##
## 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)
library(caTools) # for sample splitting
library(RColorBrewer)
library(rattle)

## Warning: package 'rattle' was built under R version 3.3.3

## Rattle: A free graphical interface for data mining with R.
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

library(rpart.plot)

## Loading required package: rpart

library(ellipse)
library(car)

##
## Attaching package: 'car'

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

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

library(faraway)

##
## 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

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.3.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.3.3

##
## Attaching package: 'gplots'

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

# PROBLEM STATEMENT
# The specific goal here is to predict whether an employee will stay or
voluntarily leave within the next year. In the present data, this means
predicting the variable "vol_leave" (0 = stay, 1 = leave) using the other
columns of data. You can think of this data as historical data which tells us
who did and who did not leave within the last year.

# DATA
HRAnalytics <- read.csv("C:/Users/Shraddha Somani/Desktop/humanresource.csv")
str(HRAnalytics)

## '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 ...
```

```
head(HRAnalytics)
```

```
##       role perf      area    sex id      age    salary vol_leave
## 1      CEO    3     Sales   Male  1 62.00000 1000000.0         0
## 2 Director    3 Marketing   Male 32 53.35897  258934.7         0
## 3 Director    1   Finance   Male 76 53.48636  189828.4         1
## 4 Director    2     Sales   Male 69 49.16571  207492.3         0
## 5 Director    3 Marketing   Male 28 49.77968  188204.7         0
## 6 Director    1     Other Female 77 39.59079  194836.6         0
```

```
summary(HRAnalytics)
```

```
##        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
```

```
# Analysis:
# - The summary information lets us know that we have 5 fundamental roles:
CEO, Director, Individual Contributors, Manager and VP.
# - Since CEOs and VPs encounter an altogether different labor market than
the Directors, Managers, and Individuals, incorporating them in our modeling
doesn't bode well.
# - Resetting the data
HRAnalytics = filter(HRAnalytics, HRAnalytics$role == "Ind" |
HRAnalytics$role == "Manager" | HRAnalytics$role == "Director")
HRAnalytics$role <- factor(HRAnalytics$role)
summary(HRAnalytics)
```

```
##        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
```

```r
# VISUALIZATION
# As the response output variable consist of two groups (0, 1), comparing it
with other columns would be much easier if we use aggregate along with the
mean function.

# (a) Performance v/s Voluntarily Leaving
performance_agg = aggregate(vol_leave ~ perf, data = HRAnalytics, mean)
performance_agg
```

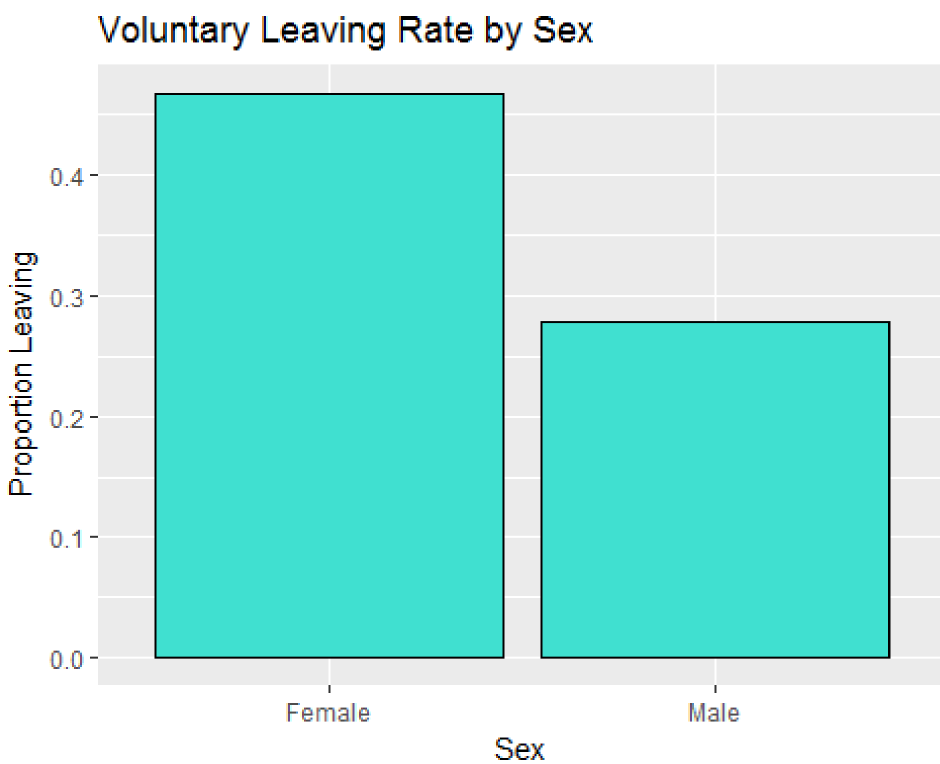
```
##   perf vol_leave
## 1    1 0.3375112
## 2    2 0.3383831
## 3    3 0.4831122
```

```r
ggplot(performance_agg, aes(x = perf, y = vol_leave)) + geom_bar(stat =
"identity", fill = 'turquoise', colour = 'black') + ggtitle("Voluntary
Leaving Rate by Performance Rating") + labs(y = "Proportion Leaving", x =
"Performance Rating")
```

```r
# Analysis:
# - Employees with performance rating 3 are likely to leave the company next year

# (b) Sex v/s Voluntarily Leaving
sex_agg = aggregate(vol_leave ~ sex, data = HRAnalytics, mean)
sex_agg
```

```
##      sex vol_leave
## 1 Female 0.4673483
## 2   Male 0.2781970
```

```r
ggplot(sex_agg, aes(x = sex, y = vol_leave)) + geom_bar(stat = "identity", fill = 'turquoise', colour = 'black') + ggtitle("Voluntary Leaving Rate by Sex") + labs(y = "Proportion Leaving", x = "Sex")
```

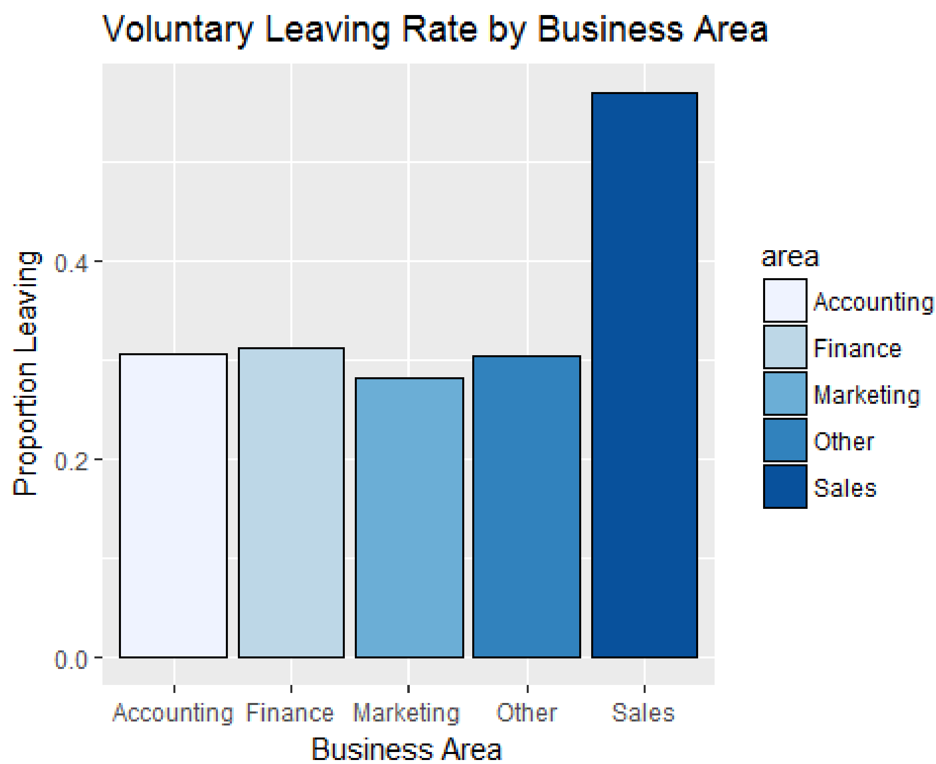



```r
# Analysis:
# - Female attrition rate is higher than the males in the entire organization
# - Females are more prone to voluntarily leaving the company

# (c) Business Area v/s Voluntarily Leaving
area_agg = aggregate(vol_leave ~ area, data = HRAnalytics, mean)
area_agg
```

```
##         area vol_leave
## 1 Accounting 0.3055383
## 2    Finance 0.3126492
```

```
## 3  Marketing 0.2815965
## 4      Other 0.3040510
## 5      Sales 0.5696880
```

```
ggplot(area_agg, aes(x = area, y = vol_leave, fill = area)) + geom_bar(stat =
"identity", colour = "black") + scale_fill_brewer() + ggtitle("Voluntary
Leaving Rate by Business Area") + labs(y = "Proportion Leaving", x =
"Business Area")
```

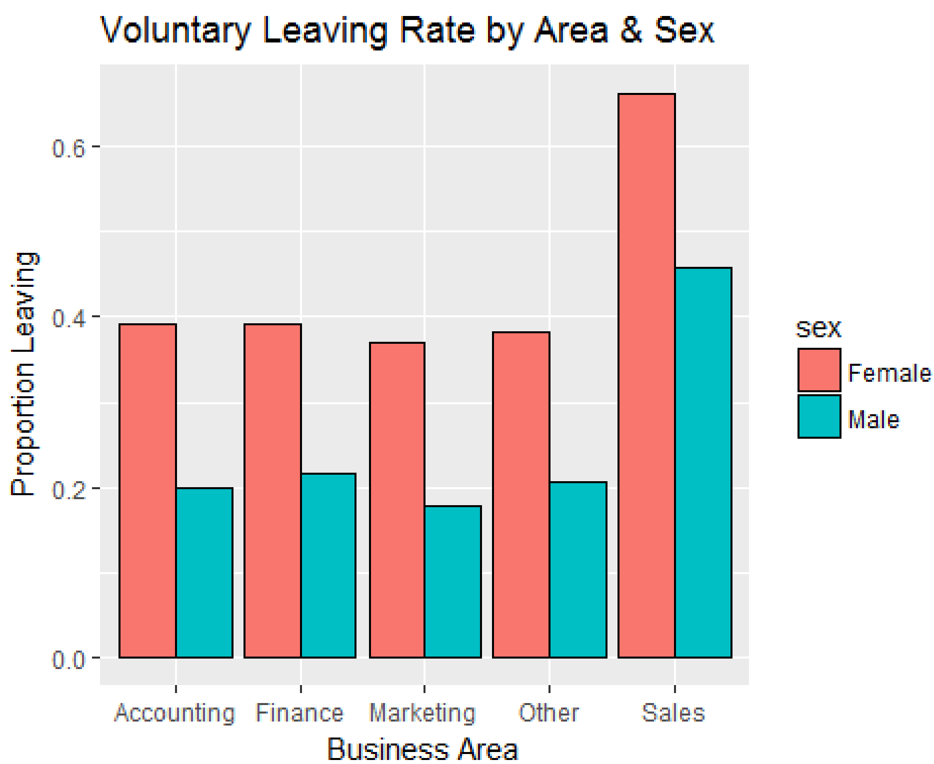



```
# Analysis:
# - People working in Sales department are much more likely to leave the job
reason being:
#   - Most sales jobs are paid less and mundane
#   - No fixed working hours
#   - Work timing extends to late nights as well.

# (d) Business Area and Gender v/s Voluntarily Leaving
area_sex_agg = aggregate(vol_leave ~ area + sex, data = HRAnalytics, mean)
area_sex_agg
```

```
##          area    sex vol_leave
## 1  Accounting Female 0.3923337
## 2     Finance Female 0.3923497
## 3   Marketing Female 0.3691550
## 4       Other Female 0.3828383
## 5       Sales Female 0.6624795
## 6  Accounting   Male 0.1986111
```

```
## 7      Finance   Male 0.2168200
## 8    Marketing   Male 0.1785714
## 9        Other   Male 0.2071066
## 10       Sales   Male 0.4589309
```

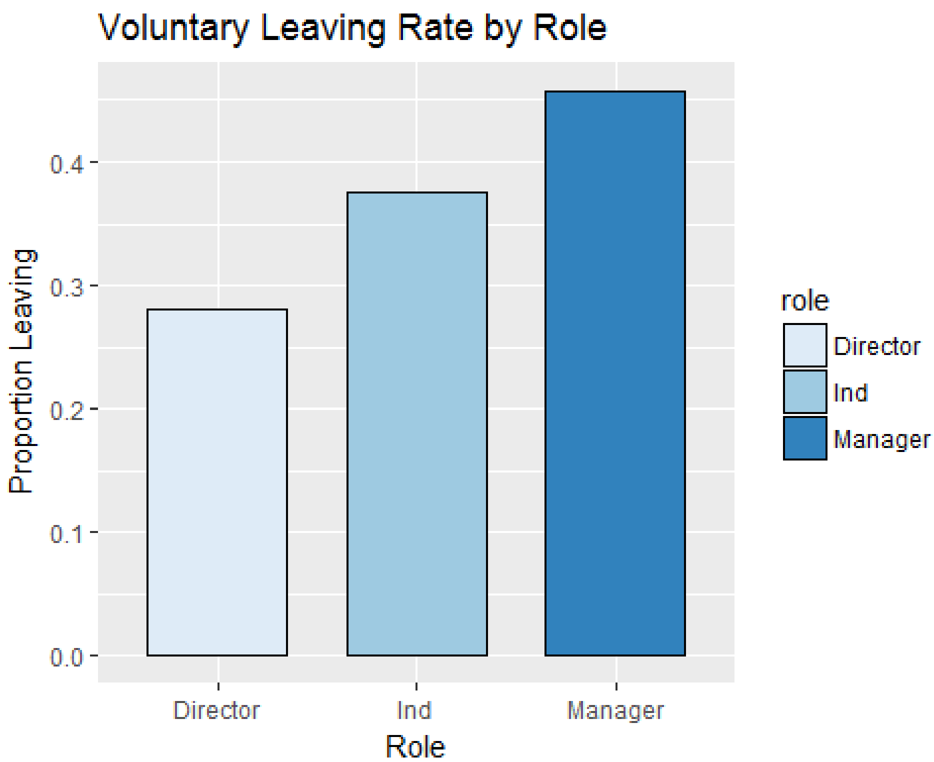
```
ggplot(area_sex_agg, aes(x = area, y = vol_leave)) + geom_bar(aes(fill =
sex), stat = "identity", colour = "black", position = position_dodge()) +
ggtitle("Voluntary Leaving Rate by Area & Sex") + labs(y = "Proportion
Leaving", x = "Business Area")
```



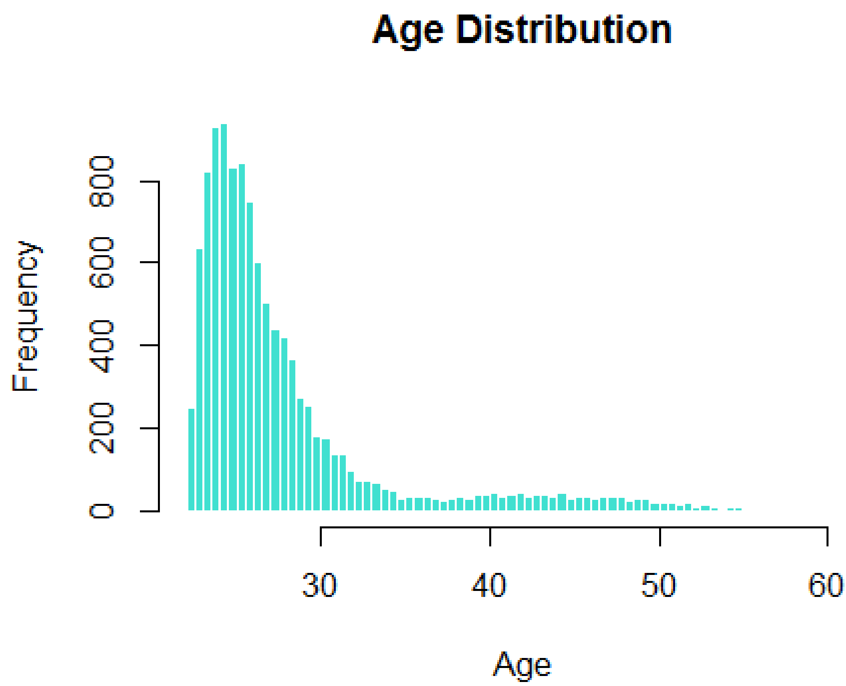

```
# Analysis:
# - Voluntary termination is higher in females.
# - Under sales department, all employees are nearly unhappy

# (e)   Role v/s Voluntarily Leaving
role_agg = aggregate(vol_leave ~ role, data = HRAnalytics, mean)
role_agg
```

```
##         role vol_leave
## 1 Director    0.2800
## 2      Ind    0.3749
## 3  Manager    0.4580
```

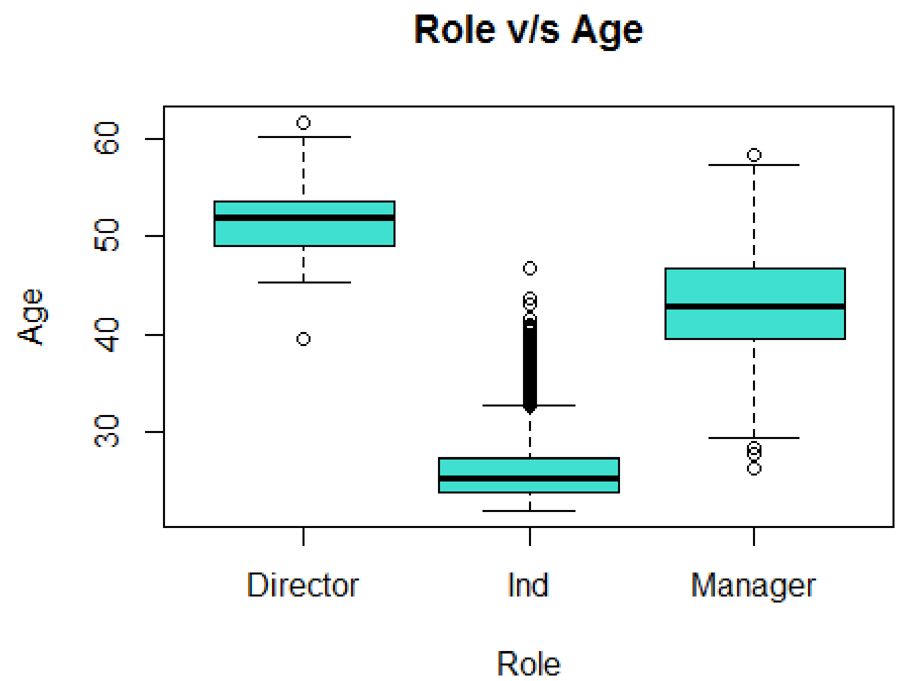
```
ggplot(role_agg, aes(x = role, y = vol_leave, fill = role)) + geom_bar(stat =
"identity", width = .7, colour = 'black') + scale_fill_brewer() +
ggtitle("Voluntary Leaving Rate by Role") + labs (y = "Proportion Leaving", x
= "Role")
```




```r
# Analysis:
# - Managers have higher attrition rate
# - Directors have a longer run at a company.

# (f) Analyzing the age of the employee
hist(HRAnalytics$age, breaks = 100, main = "Age Distribution", border = F,
xlab = "Age", col = 'turquoise')
```

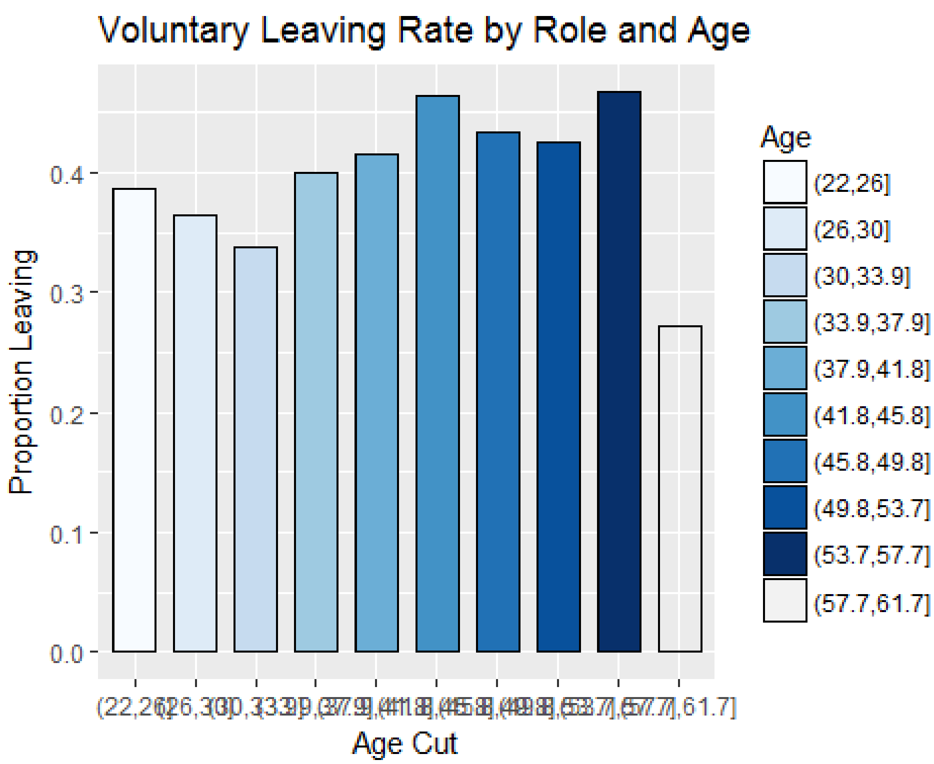



```
quantile(HRAnalytics$age)

##       0%      25%      50%      75%     100%
## 22.02289 24.07050 25.69533 28.48035 61.67132

# Analysis:
# - Skewness is present here with half of our workforce somewhere around 22
and 26 years old.
# - However there are three distinct levels: people, supervisors and
executives. It will be more informative to see how those ages breakdown when
we take that into account. Therefore box plots have been utilized for this
purpose.

boxplot(age ~ role, data = HRAnalytics, col = 'turquoise', xlab = 'Role',
ylab = 'Age', main = 'Role v/s Age')
```



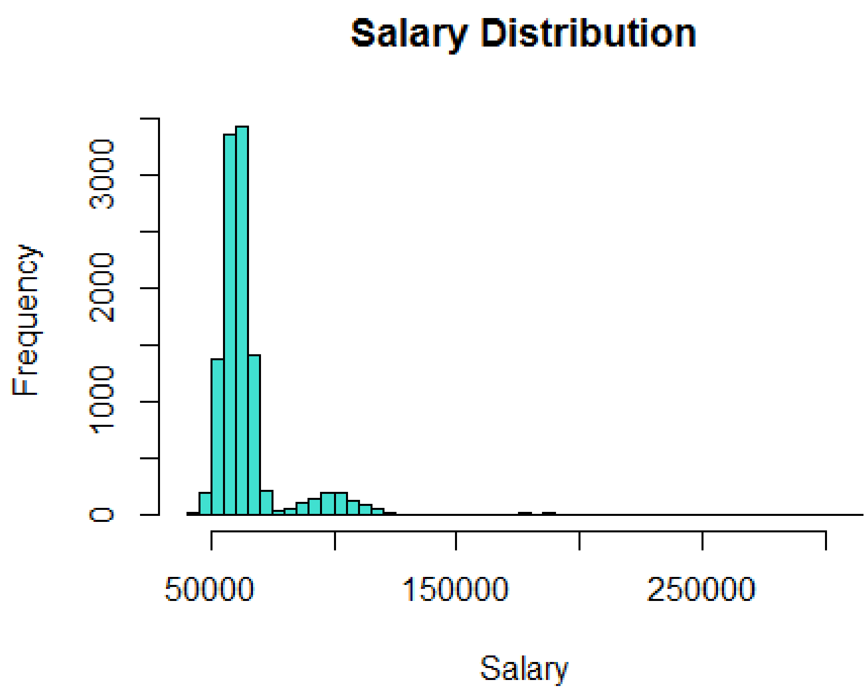

```
# Analysis:
# - There is a solid relationship between role and age
# - Since the variable age is skewed, we will take the log of age while
fitting a model.

HRAnalytics$log_age = log(HRAnalytics$age)
summary(HRAnalytics$log_age)

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

# Segmenting age variable even further to get proper insights
age_agg = aggregate(x = HRAnalytics$vol_leave, by = list(cut(HRAnalytics$age,
10)), mean)
age_agg

##        Group.1         x
## 1      (22,26] 0.3866177
## 2      (26,30] 0.3645902
## 3    (30,33.9] 0.3374536
## 4  (33.9,37.9] 0.3992806
## 5  (37.9,41.8] 0.4155405
## 6  (41.8,45.8] 0.4640288
## 7  (45.8,49.8] 0.4333333
## 8  (49.8,53.7] 0.4260870
## 9  (53.7,57.7] 0.4666667
## 10 (57.7,61.7] 0.2727273
```

```r
names(age_agg) = c("Age", "Probability")
ggplot(age_agg, aes(x = Age, y = Probability, fill = Age)) + geom_bar(stat =
"identity", width = .7, colour = 'black') + scale_fill_brewer() +
ggtitle("Voluntary Leaving Rate by Role and Age") + labs(y = "Proportion
Leaving", x = "Age Cut")

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

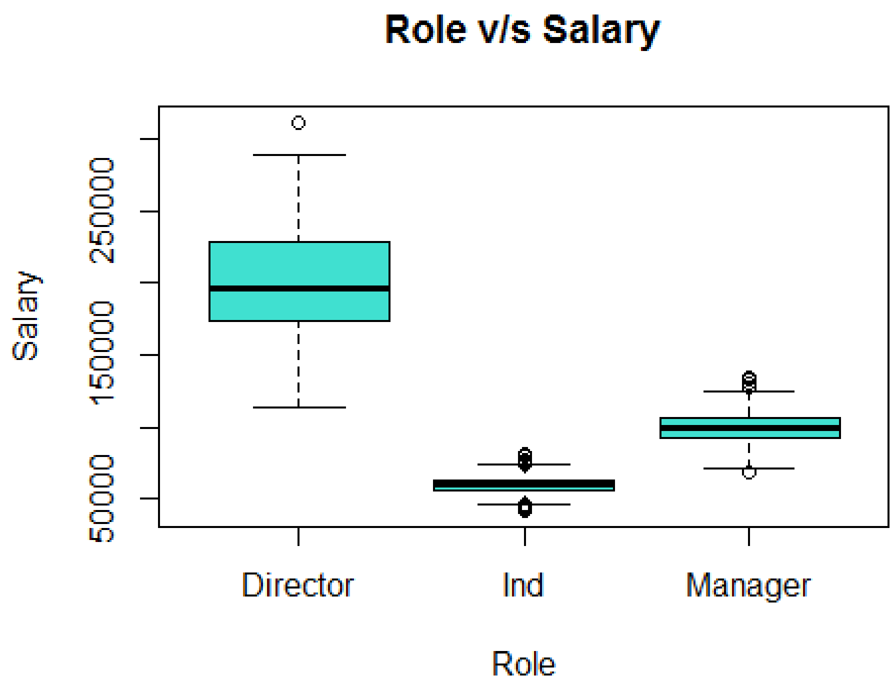



```r
# Analysis:
# - This shows that People within 34-54 age group terminate the company more
likely than the people within 22-34 who might be individual employees.
# - Age group of 54- 62 is at Director level and the attrition is least in
that age group.

# (g) Analyzing the salary pattern
summary(HRAnalytics$salary)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   42170   57080   60790   64860   64930  311100

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

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

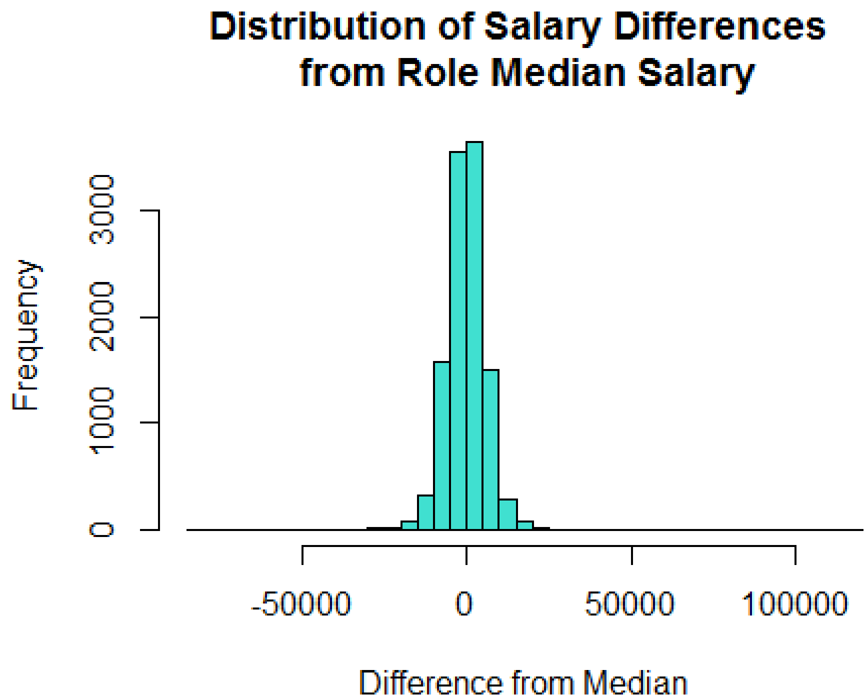
```r
hist(HRAnalytics$salary, breaks = 50, col = 'turquoise', main = "Salary
Distribution", xlab = "Salary")
```




```r
# Analysis:
# - The median salary is 60800, with the max being 1000000 and the min being
42170.
# - Salary variable is highly skewed with almost 80% of the people earning
till $66173.65.
# - Segmenting salary division based on role.

salary_agg = aggregate(salary ~ role, data = HRAnalytics, median)
salary_agg

##       role    salary
## 1 Director 195598.67
## 2      Ind  60102.17
## 3  Manager  99545.18

# Plot
boxplot(salary ~ role, data = HRAnalytics, col = 'turquoise', xlab = 'Role',
ylab = 'Salary', main = 'Role v/s Salary')
```




```r
# Creating normalized variable based on median values obtained above
names(salary_agg)[2] = "role_mean_salary"
HRAnalytics = merge(HRAnalytics, salary_agg, by = "role")
HRAnalytics$salary_diff = HRAnalytics$salary - HRAnalytics$role_mean_salary

# Analyzing normalized salary
hist(HRAnalytics$salary_diff, breaks = 50, main = "Distribution of Salary
Differences \n from Role Median Salary", col = 'turquoise', xlab =
"Difference from Median")
```




```
# DATA MODELING
# Before we start creating models, we need to split our data into a training
set and a test set. We will utilize two-thirds of the data for training and
model development and one third of the data for testing the models.
# We set the random seed to a particular number so we can simply replicate
our outcomes.
set.seed(42) # setting the random seed for replication
sample = sample.split(HRAnalytics$vol_leave, 2/3)
train = HRAnalytics[sample,]
test = HRAnalytics[!sample,]

# We will be using two techniques,
# (a) Logistic Regression
# (b) Decision Tree
# Logistic regression builds a condition that as a result predicts the
probability of a two-class result (staying or leaving) utilizing the chosen
indicators. Each of the indicators are connected with a "significance""
pointer that lets you know whether the indicator is helpful or not.
# On contrast, decision trees work by utilizing the indicators to part the
data into buckets using a set of decision rules.

# (a) LOGESTIC REGRESSION
test_mean = mean(test$vol_leave)
train_mean = mean(train$vol_leave)
print(c(test_mean, train_mean))

## [1] 0.3816216 0.3814865
```

```
# (1) Fit the model
fit = glm(vol_leave ~ role + perf + area + sex + log_age + salary_diff, data
= HRAnalytics, family = 'binomial')
summary(fit)

##
## Call:
## glm(formula = vol_leave ~ role + perf + area + sex + log_age +
##     salary_diff, family = "binomial", data = HRAnalytics)
##
## 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)    1.581e-01  8.676e-01   0.182 0.855451
## roleInd        6.819e-01  3.456e-01   1.973 0.048495 *
## roleManager    1.393e+00  3.249e-01   4.289  1.8e-05 ***
## 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_diff   -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

# Analysis:
# - First of all, we can see that areaFinance, areaMarketing and areaOther is
not statistically significant.
# - As for the statistically significant variables, salary, areaSales and
perf has the lowest p-value suggesting a strong association of these variable
with the probability of leaving the company.
# - Now we can run the anova() function on the model to analyze the table of
deviance

# (2) Chi Square Test
anova(fit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vol_leave
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       11099      14759
## role         2    30.69     11097      14728 2.162e-07 ***
## perf         1   161.14     11096      14567 < 2.2e-16 ***
## area         4   735.02     11092      13832 < 2.2e-16 ***
## sex          1   466.69     11091      13365 < 2.2e-16 ***
## log_age      1    11.21     11090      13354 0.0008158 ***
## salary_diff  1   350.08     11089      13004 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Analysis:
# - The difference between the null deviance and the residual deviance shows
how our model is doing against the null model (a model with only the
intercept). The wider this gap, the better
# - A smaller p-value here indicates that all the variables in the model are
significant

# (3) Assessing the predictive ability of the model
fitted.results = predict(fit, test, type = 'response')
fitted.results = ifelse(fitted.results > 0.5,1,0)
misClasificError = mean(fitted.results != test$vol_leave)
# Confusion Matrix
table(actual = test$vol_leave, prediction = fitted.results)
```

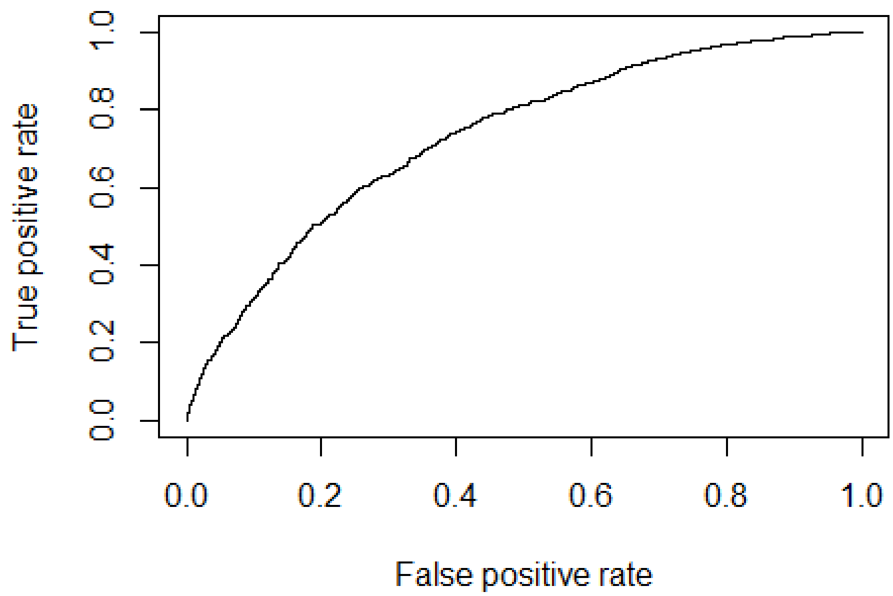
```
##       prediction
## actual    0    1
##      0 1919  369
##      1  780  632
```

```
# Accuracy
print(paste('Accuracy', 1 - misClasificError))
```

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

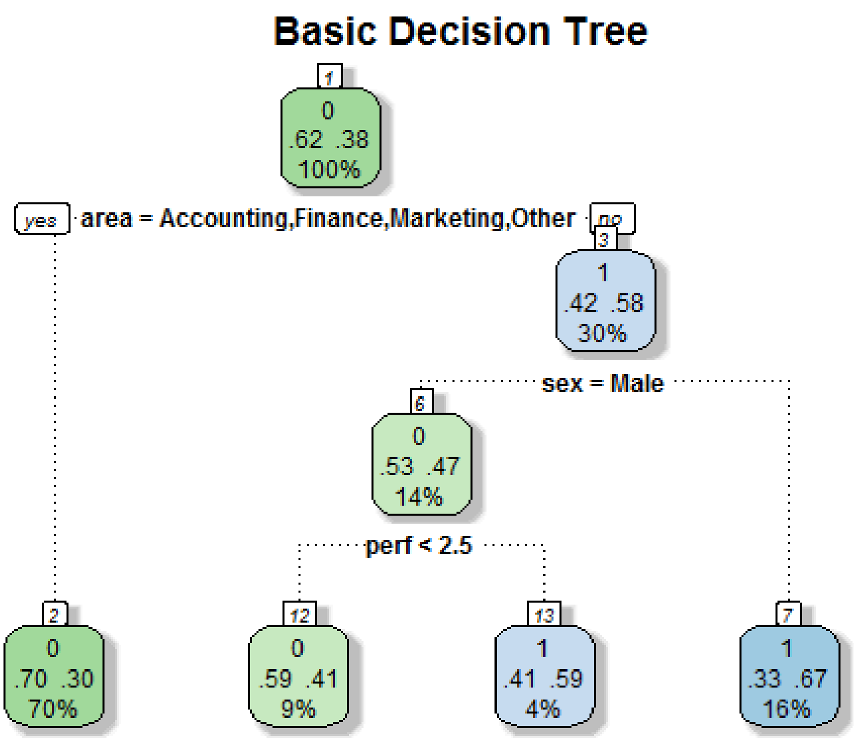
```
# Analysis:
# - The accuracy of our model is 68%

# (4) ROC Curve & AUC
# As a last step, 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.
```

```r
p = predict(fit, test, type = "response")
pr = prediction(p, test$vol_leave)
prf = performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




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

## [1] 0.7326298

# Analysis:
# - The ROC is a curve generated by plotting the true positive rate (TPR)
against the false positive rate (FPR) at various threshold settings.
# - The AUC is the area under the ROC curve.
# - As a rule of thumb, a model with good predictive ability should have an
AUC closer to 1 (1 is ideal) than to 0.5.
# - Based on the value of AUC for our dataset, we can say that it has good
predictive ability.

# (b) DECISION TREE
# We have already divided our dataset into training and testing. So we
proceed further by making the decision tree
# (1) Fit the model
set.seed(42)
decision_fit = rpart(vol_leave ~ role + perf + age + sex + area + salary,
```

```
data = train, method = "class")
decision_fit

## 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) *

# Plot the tree
par(mar = c(5,4,1,2))
fancyRpartPlot(decision_fit, sub = NULL, main = "Basic Decision Tree")
```




```
# Analysis:
# - The first node is alluded to as the root. The '0' alludes to the dominate
case. 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
```

```r
'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. Think of this node like a bucket 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 is further, 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% of the aggregate populace.
# - 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% of the aggregate populace.

# (2) Assessing the predictive ability of the model
t_pred = predict(decision_fit, test, type = 'class')
# Confusion Matrix
confMat = table(actual = test$vol_leave, prediction = t_pred)
confMat

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

# Accuracy
accuracy = sum(diag(confMat))/sum(confMat)
accuracy

## [1] 0.6724324

# CONCLUSION
# - Logistic regression is better than decision tree in predicting the output response variable.
# - To play more important and vital part in the organization, the HR function needs to move past beyond mere reporting to precise expectation.
# - Rather than simply creating receptive reports, it needs to grasp advanced
```

analytics and predictive techniques that bolster key organizational objectives.