IBM_HR_ANALYTICS

Aravind

March 1, 2018

```
ibm_data <- read.csv("IBM_HR_Attrition.csv")
str(ibm_data)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41       Yes      Travel_Rarely      1102                  Sales
## 2     49        No Travel_Frequently       279 Research & Development
## 3     37       Yes      Travel_Rarely      1373 Research & Development
## 4     33        No Travel_Frequently      1392 Research & Development
## 5     27        No      Travel_Rarely       591 Research & Development
## 6     32        No Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4        Single          5993
## 2    Research Scientist               2       Married          5130
## 3 Laboratory Technician               3        Single          2090
## 4    Research Scientist               3       Married          2909
## 5 Laboratory Technician               2       Married          3468
## 6 Laboratory Technician               4        Single          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y      Yes                11
## 2       24907                  1      Y       No                23
## 3        2396                  6      Y      Yes                15
## 4       23159                  1      Y      Yes                11
## 5       16632                  9      Y       No                12
## 6       11864                  0      Y       No                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

Check for missing values

```r
colSums(is.na(ibm_data))
```

```
##                   ï..Age                Attrition           BusinessTravel
##                        0                        0                        0
##                DailyRate               Department         DistanceFromHome
##                        0                        0                        0
##                Education           EducationField            EmployeeCount
##                        0                        0                        0
##           EmployeeNumber  EnvironmentSatisfaction                   Gender
##                        0                        0                        0
##               HourlyRate           JobInvolvement                 JobLevel
##                        0                        0                        0
##                  JobRole          JobSatisfaction            MaritalStatus
##                        0                        0                        0
##            MonthlyIncome              MonthlyRate       NumCompaniesWorked
##                        0                        0                        0
##                   Over18                 OverTime        PercentSalaryHike
##                        0                        0                        0
##        PerformanceRating RelationshipSatisfaction            StandardHours
##                        0                        0                        0
##         StockOptionLevel        TotalWorkingYears    TrainingTimesLastYear
##                        0                        0                        0
##          WorkLifeBalance           YearsAtCompany       YearsInCurrentRole
##                        0                        0                        0
##  YearsSinceLastPromotion     YearsWithCurrManager
##                        0                        0
```

```r
table(ibm_data$Attrition)
```

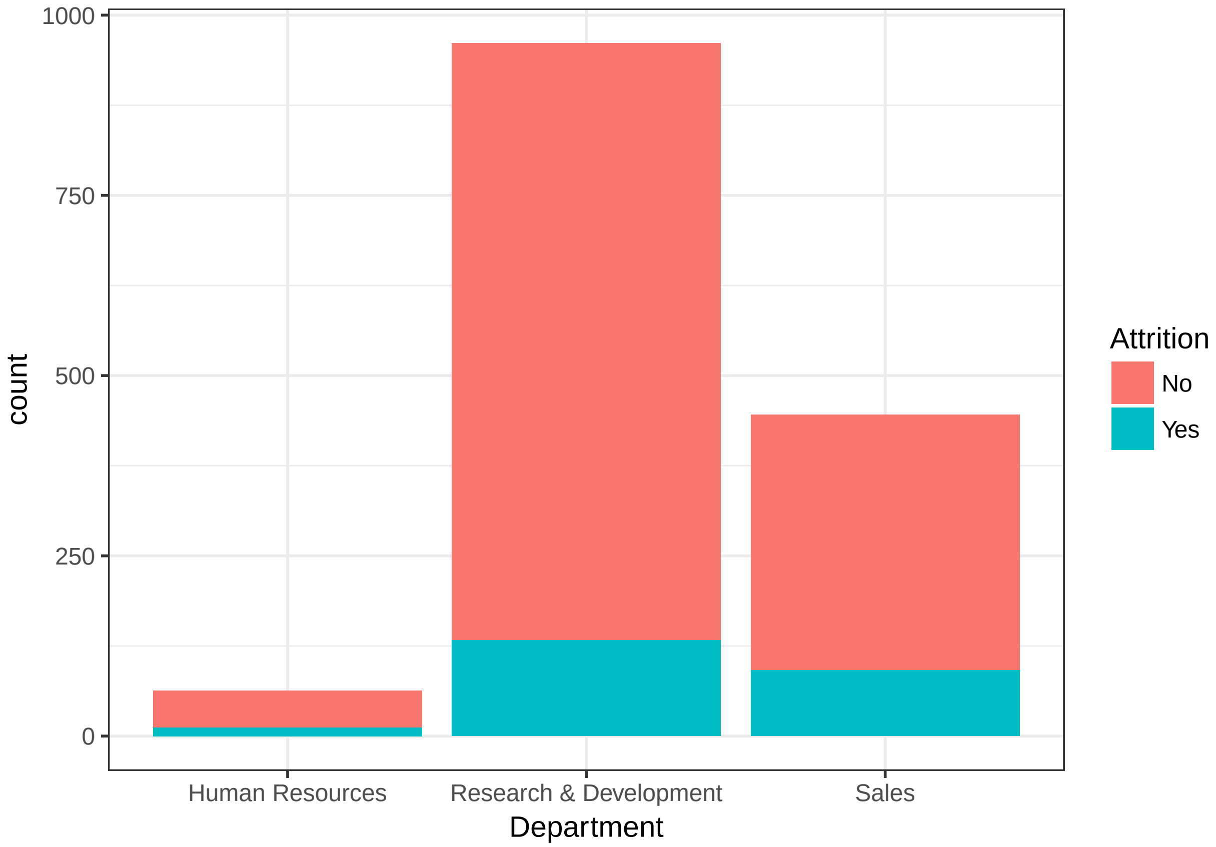
```
## 
##   No  Yes 
## 1233  237
```

Visualization

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

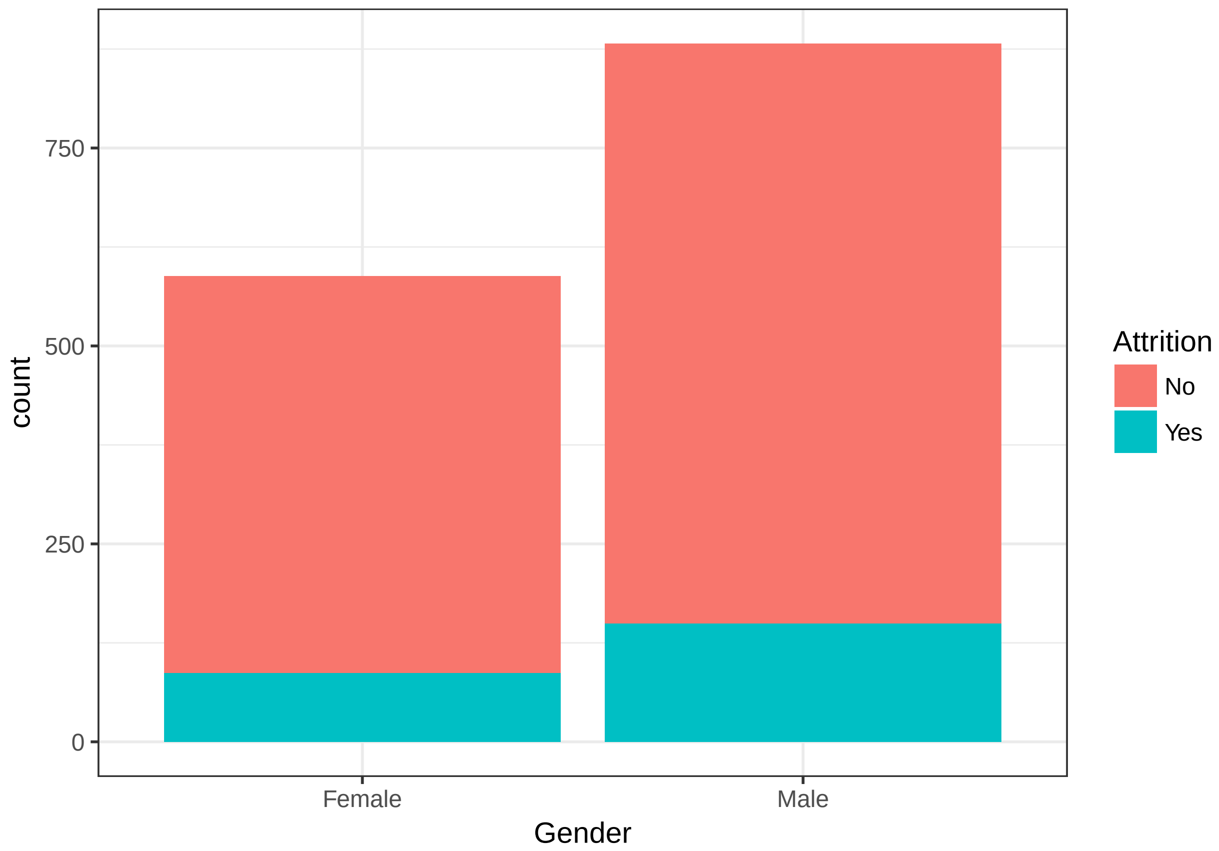
```r
ggplot(ibm_data, aes(x=Department, fill=Attrition))+geom_bar() + theme_bw()
```



HR department has the least count. R&D department has more Attrition = No(High proportion of no)

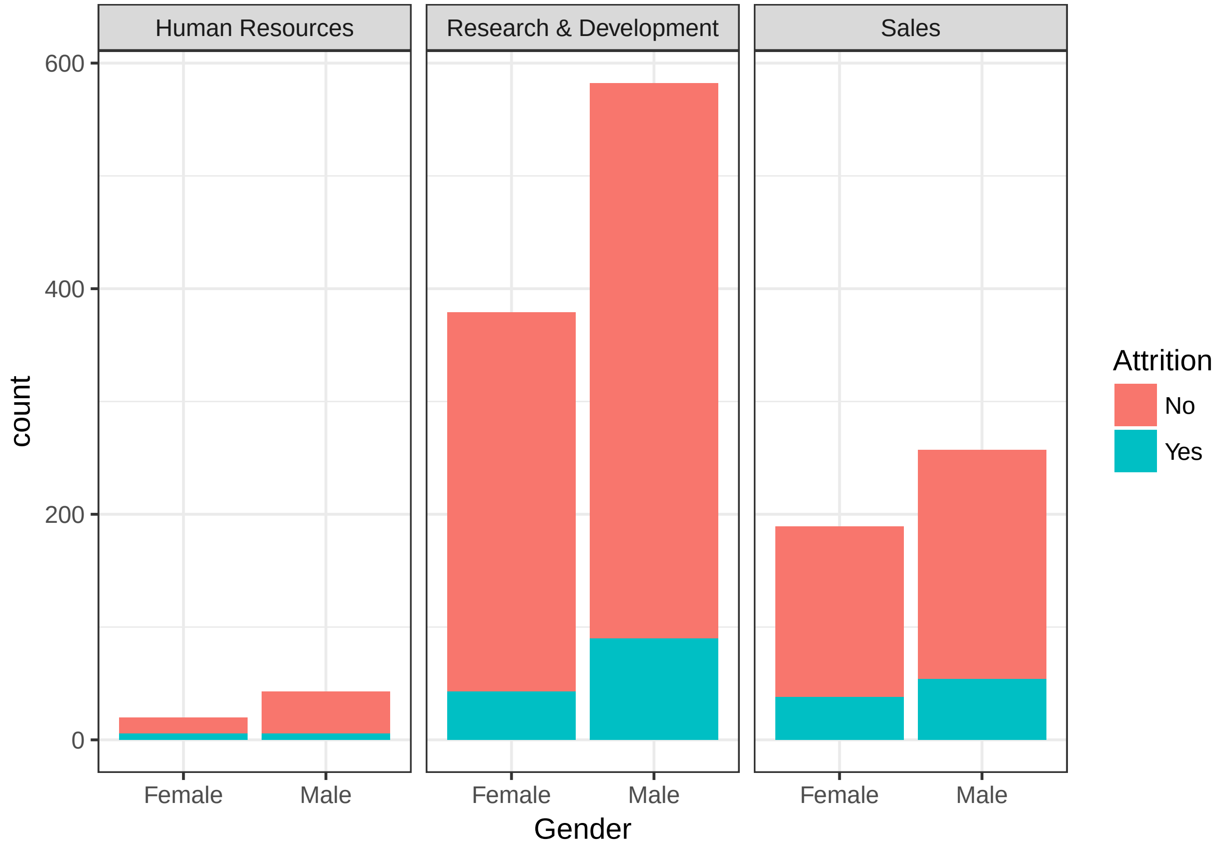
Lets see Gender and Attrition

```
ggplot(ibm_data, aes(x=Gender, fill=Attrition))+geom_bar() + theme_bw()
```

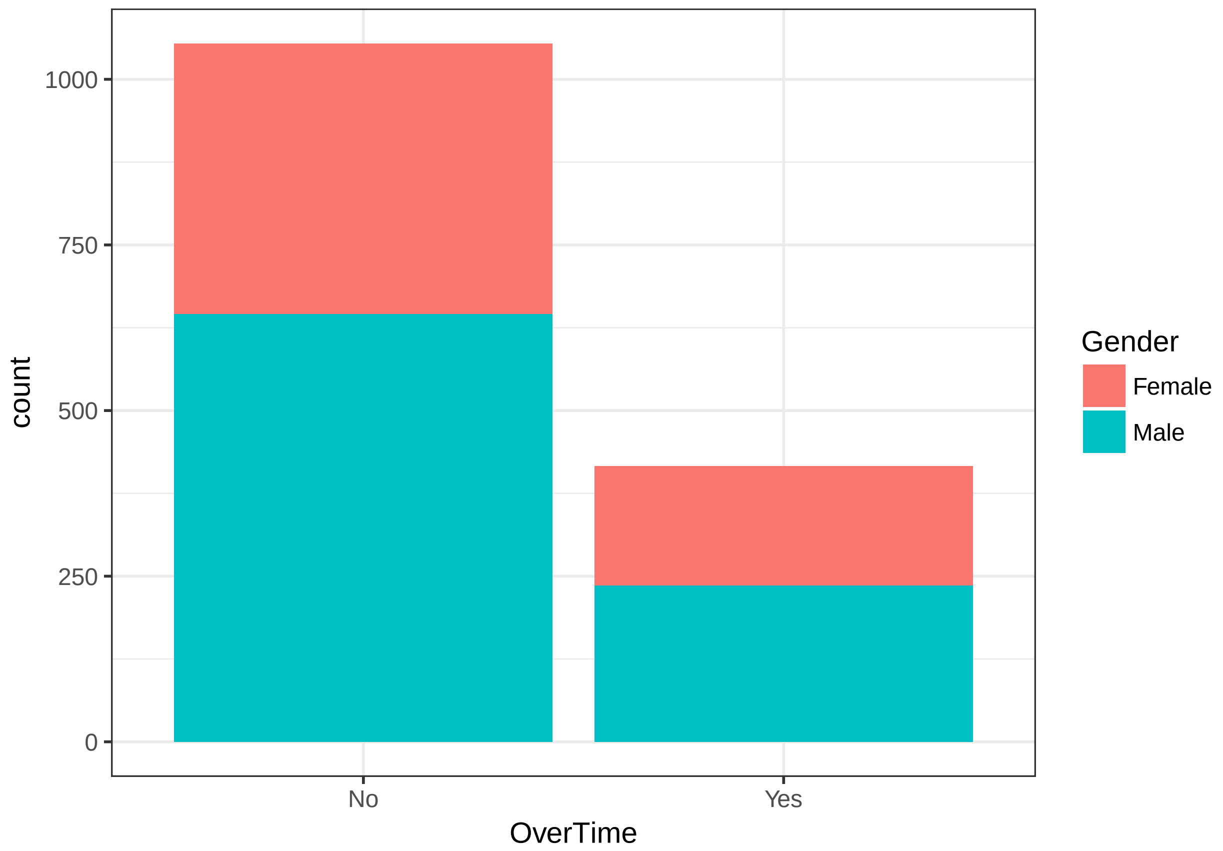


Gender, Attrition from each department

```
ggplot(ibm_data, aes(x=Gender, fill=Attrition))+geom_bar() + theme_bw() + facet_wrap(~Department)
```



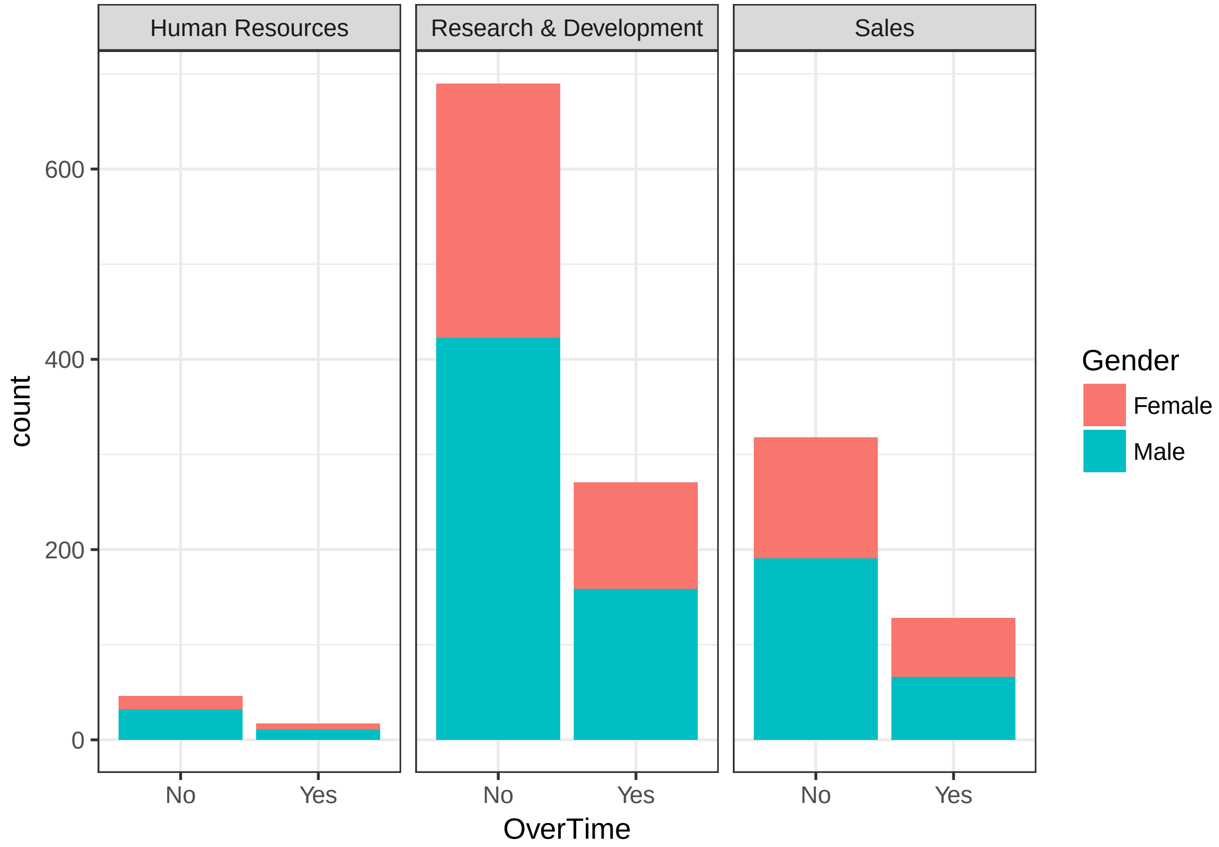
```
ggplot(ibm_data, aes(x=OverTime, fill=Gender))+geom_bar() + theme_bw()
```



Females do work overtime. There is quite a lot.

Department wise.

```
ggplot(ibm_data, aes(x=OverTime, fill=Gender))+geom_bar() + theme_bw() + facet_wrap(~Department)
```



Attrition = Yes - 1, No - 0

Use dummy_cols{fastDummies} for label encoding. I have done manually.

```
ibm_data$Attrition = as.integer(ibm_data$Attrition)
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41         2     Travel_Rarely      1102                  Sales
## 2     49         1 Travel_Frequently       279 Research & Development
## 3     37         2     Travel_Rarely      1373 Research & Development
## 4     33         1 Travel_Frequently      1392 Research & Development
## 5     27         1     Travel_Rarely       591 Research & Development
## 6     32         1 Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4        Single          5993
## 2    Research Scientist               2       Married          5130
## 3 Laboratory Technician               3        Single          2090
## 4    Research Scientist               3       Married          2909
## 5 Laboratory Technician               2       Married          3468
## 6 Laboratory Technician               4        Single          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y      Yes                11
## 2       24907                  1      Y       No                23
## 3        2396                  6      Y      Yes                15
## 4       23159                  1      Y      Yes                11
## 5       16632                  9      Y       No                12
## 6       11864                  0      Y       No                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

```
ibm_data$Attrition[ibm_data$Attrition == 1] <- 0
ibm_data$Attrition[ibm_data$Attrition == 2] <- 1
```

```
ibm_data$Attrition = as.factor(ibm_data$Attrition)
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41         1     Travel_Rarely      1102                  Sales
## 2     49         0 Travel_Frequently       279 Research & Development
## 3     37         1     Travel_Rarely      1373 Research & Development
## 4     33         0 Travel_Frequently      1392 Research & Development
## 5     27         0     Travel_Rarely       591 Research & Development
## 6     32         0 Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4        Single          5993
## 2    Research Scientist               2       Married          5130
## 3 Laboratory Technician               3        Single          2090
## 4    Research Scientist               3       Married          2909
## 5 Laboratory Technician               2       Married          3468
## 6 Laboratory Technician               4        Single          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y      Yes                11
## 2       24907                  1      Y       No                23
## 3        2396                  6      Y      Yes                15
## 4       23159                  1      Y      Yes                11
## 5       16632                  9      Y       No                12
## 6       11864                  0      Y       No                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                     5
## 2                     7
## 3                     0
## 4                     0
## 5                     2
## 6                     6
```

Turning numeric variables into factors

```r
ibm_data$Education <- as.factor(ibm_data$Education)
ibm_data$EnvironmentSatisfaction <- as.factor(ibm_data$EnvironmentSatisfaction)
ibm_data$JobInvolvement <- as.factor(ibm_data$JobInvolvement)
ibm_data$JobSatisfaction <- as.factor(ibm_data$JobSatisfaction)
ibm_data$PerformanceRating <- as.factor(ibm_data$PerformanceRating)
ibm_data$RelationshipSatisfaction <- as.factor(ibm_data$RelationshipSatisfaction)
ibm_data$WorkLifeBalance <- as.factor(ibm_data$WorkLifeBalance)
str(ibm_data)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```r
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41         1     Travel_Rarely      1102                  Sales
## 2     49         0 Travel_Frequently       279 Research & Development
## 3     37         1     Travel_Rarely      1373 Research & Development
## 4     33         0 Travel_Frequently      1392 Research & Development
## 5     27         0     Travel_Rarely       591 Research & Development
## 6     32         0 Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4        Single          5993
## 2    Research Scientist               2       Married          5130
## 3 Laboratory Technician               3        Single          2090
## 4    Research Scientist               3       Married          2909
## 5 Laboratory Technician               2       Married          3468
## 6 Laboratory Technician               4        Single          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y      Yes                11
## 2       24907                  1      Y       No                23
## 3        2396                  6      Y      Yes                15
## 4       23159                  1      Y      Yes                11
## 5       16632                  9      Y       No                12
## 6       11864                  0      Y       No                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

Gender

```r
ibm_data$Gender <- as.integer(ibm_data$Gender)
ibm_data$MaritalStatus <- as.integer(ibm_data$MaritalStatus)
ibm_data$OverTime <- as.integer(ibm_data$OverTime)
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41         1     Travel_Rarely      1102                  Sales
## 2     49         0 Travel_Frequently       279 Research & Development
## 3     37         1     Travel_Rarely      1373 Research & Development
## 4     33         0 Travel_Frequently      1392 Research & Development
## 5     27         0     Travel_Rarely       591 Research & Development
## 6     32         0 Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2      1         94              3        2
## 2                       3      2         61              2        2
## 3                       4      2         92              2        1
## 4                       4      1         56              3        1
## 5                       1      2         40              3        1
## 6                       4      2         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4             3          5993
## 2    Research Scientist               2             2          5130
## 3 Laboratory Technician               3             3          2090
## 4    Research Scientist               3             2          2909
## 5 Laboratory Technician               2             2          3468
## 6 Laboratory Technician               4             3          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y        2                11
## 2       24907                  1      Y        1                23
## 3        2396                  6      Y        2                15
## 4       23159                  1      Y        2                11
## 5       16632                  9      Y        1                12
## 6       11864                  0      Y        1                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

```
ibm_data$Gender <- as.factor(ibm_data$Gender)
ibm_data$MaritalStatus <- as.factor(ibm_data$MaritalStatus)
ibm_data$OverTime <- as.factor(ibm_data$OverTime)
head(ibm_data)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate             Department
## 1     41         1     Travel_Rarely      1102                  Sales
## 2     49         0 Travel_Frequently       279 Research & Development
## 3     37         1     Travel_Rarely      1373 Research & Development
## 4     33         0 Travel_Frequently      1392 Research & Development
## 5     27         0     Travel_Rarely       591 Research & Development
## 6     32         0 Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2      1         94              3        2
## 2                       3      2         61              2        2
## 3                       4      2         92              2        1
## 4                       4      1         56              3        1
## 5                       1      2         40              3        1
## 6                       4      2         79              3        1
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4             3          5993
## 2    Research Scientist               2             2          5130
## 3 Laboratory Technician               3             3          2090
## 4    Research Scientist               3             2          2909
## 5 Laboratory Technician               2             2          3468
## 6 Laboratory Technician               4             3          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y        2                11
## 2       24907                  1      Y        1                23
## 3        2396                  6      Y        2                15
## 4       23159                  1      Y        2                11
## 5       16632                  9      Y        1                12
## 6       11864                  0      Y        1                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

Train Test Split

```
require(caret)
```

```
## Loading required package: caret
## Loading required package: lattice
```

```
set.seed(1)
ind = createDataPartition(ibm_data$Attrition, p=0.80, list = F)
train = ibm_data[ind,]
test = ibm_data[-ind,]
```

Logistic Model

```
model <- glm(Attrition ~DailyRate+EnvironmentSatisfaction+JobInvolvement+RelationshipSatisfaction , dat
summary(model)
```

```
##
## Call:
## glm(formula = Attrition ~ DailyRate + EnvironmentSatisfaction +
##     JobInvolvement + RelationshipSatisfaction, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4632  -0.5991  -0.4977  -0.4105   2.4402
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.9345089  0.3660083   2.553  0.01067 *
## DailyRate                -0.0004937  0.0002033  -2.428  0.01518 *
## EnvironmentSatisfaction2  -0.6771158  0.2430255  -2.786  0.00533 **
## EnvironmentSatisfaction3  -0.9069009  0.2216922  -4.091 4.30e-05 ***
## EnvironmentSatisfaction4  -0.9430097  0.2228605  -4.231 2.32e-05 ***
## JobInvolvement2          -0.9906657  0.3019406  -3.281  0.00103 **
## JobInvolvement3          -1.2549518  0.2829040  -4.436 9.17e-06 ***
## JobInvolvement4          -1.8239490  0.4117666  -4.430 9.44e-06 ***
## RelationshipSatisfaction2 -0.4745642  0.2454090  -1.934  0.05314 .
## RelationshipSatisfaction3 -0.5148690  0.2244683  -2.294  0.02181 *
## RelationshipSatisfaction4 -0.5869789  0.2314739  -2.536  0.01122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1040.54  on 1176  degrees of freedom
## Residual deviance:  978.95  on 1166  degrees of freedom
## AIC: 1000.9
##
## Number of Fisher Scoring iterations: 5
```

Prediction

```
prediction <- predict(model, newdata = test, type = 'response')
head(prediction)
```

```
##         11         13         14         15         31         34
## 0.14139655 0.22474550 0.10183574 0.18415052 0.09359310 0.09795004
```

```
prediction <- ifelse(prediction > 0.5,1,0)
head(prediction)
```

```
## 11 13 14 15 31 34
##  0  0  0  0  0  0
```

```
tab = table(predicted = prediction, original = test$Attrition)
tab
```

```
##          original
## predicted   0   1
##         0 245  46
##         1   1   1
```

```
print(sum(diag(tab))/sum(tab))
```

```
## [1] 0.8395904
```

Lets try random forest

```
require(randomForest)
```

```
## Loading required package: randomForest
```

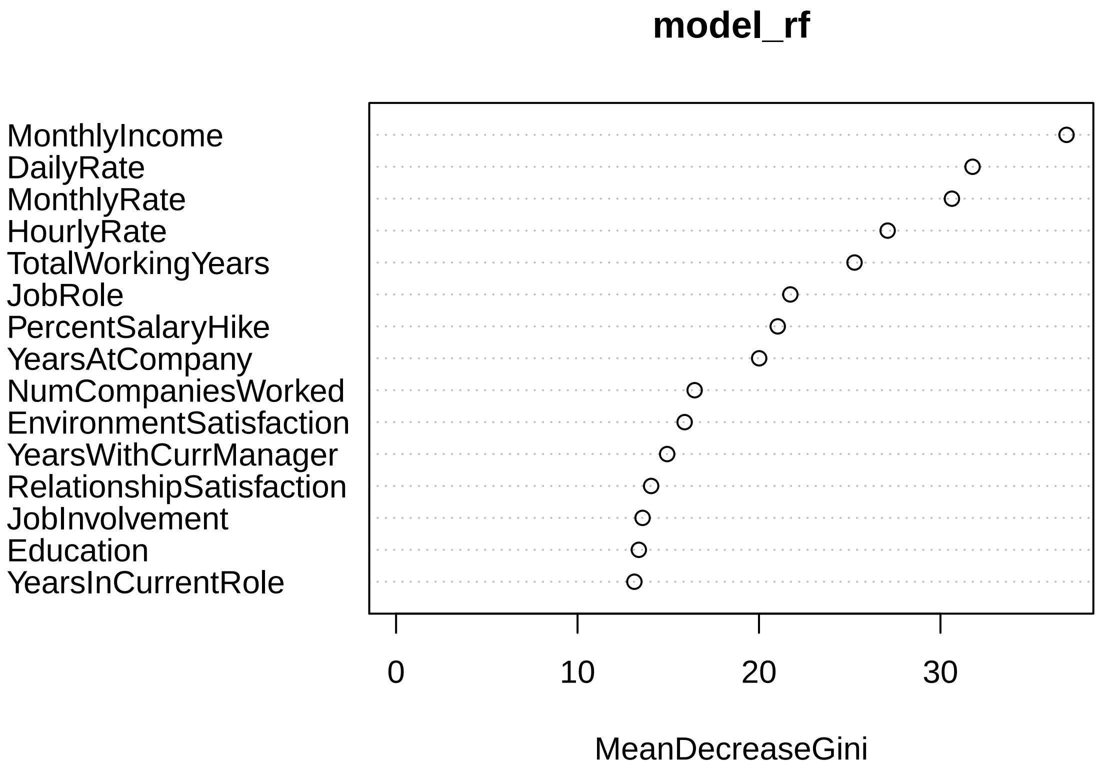
```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
model_rf <- randomForest(Attrition ~DailyRate+EnvironmentSatisfaction+JobInvolvement+
                           RelationshipSatisfaction+Education+MonthlyIncome+MonthlyRate+
                           PercentSalaryHike+TotalWorkingYears+YearsAtCompany+
                           YearsInCurrentRole+YearsWithCurrManager+NumCompaniesWorked+
                           JobRole+HourlyRate,
                           data=train)
varImpPlot(model_rf)
```



Prediction

```
prediction <- predict(model_rf, newdata = test)
head(prediction)
```

```
## 11 13 14 15 31 34
##  0  0  0  0  0  0
## Levels: 0 1
```

Accuracy

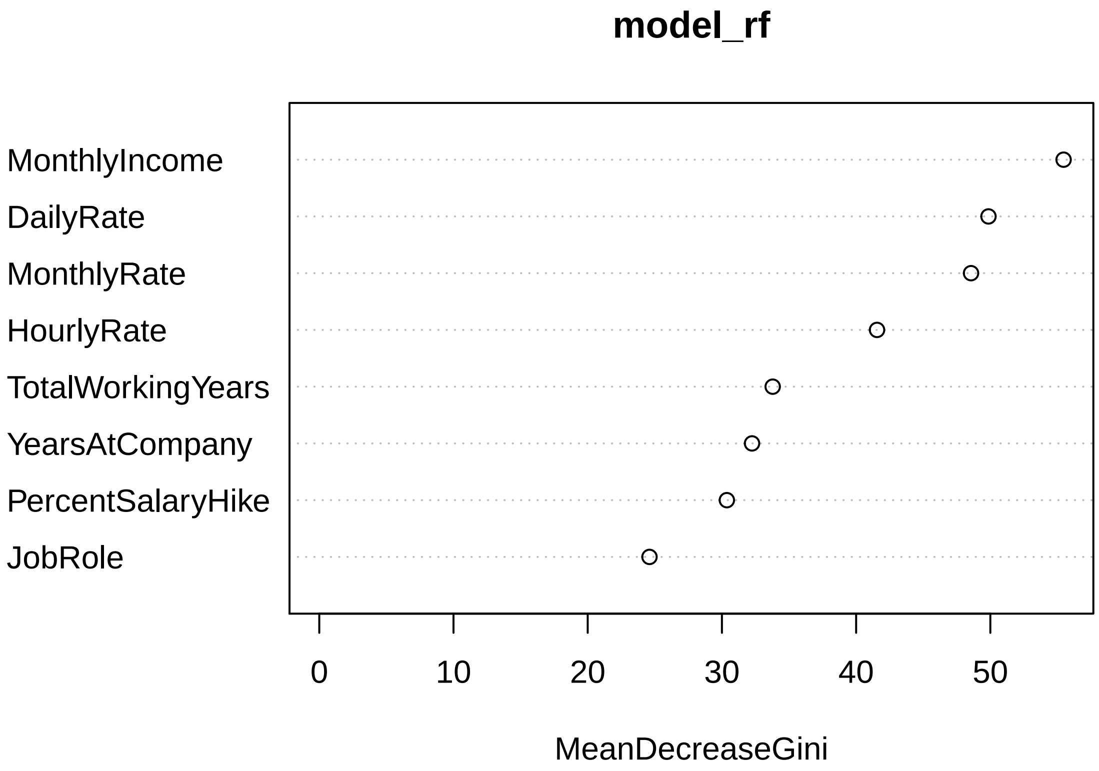
```
library(caret)
pl1 <- data.frame(original = test, predicted = prediction)
confusionMatrix(table(pl1$original.Attrition,pl1$predicted))
```

```
## Confusion Matrix and Statistics
##
##
##       0   1
##   0 246   0
##   1  42   5
##
```

```
##                 Accuracy : 0.8567
##                   95% CI : (0.8112, 0.8947)
##      No Information Rate : 0.9829
##      P-Value [Acc > NIR] : 1
## 
##                    Kappa : 0.1666
##   Mcnemar's Test P-Value : 2.509e-10
## 
##              Sensitivity : 0.8542
##              Specificity : 1.0000
##           Pos Pred Value : 1.0000
##           Neg Pred Value : 0.1064
##               Prevalence : 0.9829
##           Detection Rate : 0.8396
##     Detection Prevalence : 0.8396
##        Balanced Accuracy : 0.9271
## 
##         'Positive' Class : 0
## 
```

Drop some variables and see how accuracy changes.

```
require(randomForest)
model_rf <- randomForest(Attrition ~DailyRate+
                           MonthlyIncome+MonthlyRate+PercentSalaryHike+TotalWorkingYears+
                           YearsAtCompany+
                           JobRole+HourlyRate,
                         data=train)
varImpPlot(model_rf)
```



Prediction

```
prediction <- predict(model_rf, newdata = test)
head(prediction)
```

```
## 11 13 14 15 31 34
##  0  0  0  0  0  0
## Levels: 0 1
```

accuracy

```
library(caret)
pl1 <- data.frame(original = test, predicted = prediction)
confusionMatrix(table(pl1$original.Attrition,pl1$predicted))
```

```
## Confusion Matrix and Statistics
##
##
##       0   1
##   0 244   2
##   1  43   4
##
```

```
##                Accuracy : 0.8464
##                  95% CI : (0.7999, 0.8857)
##     No Information Rate : 0.9795
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1189
##  Mcnemar's Test P-Value : 2.479e-09
##
##             Sensitivity : 0.85017
##             Specificity : 0.66667
##          Pos Pred Value : 0.99187
##          Neg Pred Value : 0.08511
##              Prevalence : 0.97952
##          Detection Rate : 0.83276
##    Detection Prevalence : 0.83959
##       Balanced Accuracy : 0.75842
##
##        'Positive' Class : 0
##
```

Therefore, not much difference in accuracy.