Project 2

James Weaver

2023-12-22

```
# import libraries
## install.packages("naniar")
## install.packages("ggpubr")
## install.packages("dplyr")
## install.packages("ggplot2")
## install.packages("tibble")
## install.packages("tidyr")
## install.packages("GGally")
## install.packages("corrplot")
## install.packages("caret")
```

```
# upload libraries
#library(dplyr)
#library(ggplot2)
#library(tibble)
#library(tidyr)
#library(ggpubr)
#library(naniar)
#library(GGally)
#library(corrplot)
# library(caret)
```

```
# upload data
data <- read.csv("CaseStudy2-data.csv")
head(data)
```

```
##   ID Age Attrition    BusinessTravel DailyRate              Department
## 1  1  32        No     Travel_Rarely       117                   Sales
## 2  2  40        No     Travel_Rarely      1308 Research & Development
## 3  3  35        No Travel_Frequently       200 Research & Development
## 4  4  32        No     Travel_Rarely       801                   Sales
## 5  5  24        No Travel_Frequently       567 Research & Development
## 6  6  27        No Travel_Frequently       294 Research & Development
##   DistanceFromHome Education   EducationField EmployeeCount EmployeeNumber
## 1               13         4    Life Sciences             1            859
## 2               14         3          Medical             1           1128
## 3               18         2    Life Sciences             1           1412
## 4                1         4        Marketing             1           2016
## 5                2         1 Technical Degree             1           1646
## 6               10         2    Life Sciences             1            733
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
```

```
## 1                      2   Male         73              3        2
## 2                      3   Male         44              2        5
## 3                      3   Male         60              3        3
## 4                      3 Female         48              3        3
## 5                      1 Female         32              3        1
## 6                      4   Male         32              3        3
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1        Sales Executive              4      Divorced          4403
## 2     Research Director              3        Single         19626
## 3 Manufacturing Director              4        Single          9362
## 4        Sales Executive              4       Married         10422
## 5     Research Scientist              4        Single          3760
## 6 Manufacturing Director              1      Divorced          8793
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1        9250                  2      Y       No                11
## 2       17544                  1      Y       No                14
## 3       19944                  2      Y       No                11
## 4       24032                  1      Y       No                19
## 5       17218                  1      Y      Yes                13
## 6        4809                  1      Y       No                21
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                         3            80                1
## 2                 3                         1            80                0
## 3                 3                         3            80                0
## 4                 3                         3            80                2
## 5                 3                         3            80                0
## 6                 4                         3            80                2
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     3               2              5
## 2                21                     2               4             20
## 3                10                     2               3              2
## 4                14                     3               3             14
## 5                 6                     2               3              6
## 6                 9                     4               2              9
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  2                       0                    3
## 2                  7                       4                    9
## 3                  2                       2                    2
## 4                 10                       5                    7
## 5                  3                       1                    3
## 6                  7                       1                    7
```

```
str(data)
```

```
## 'data.frame':    870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : chr  "No" "No" "No" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Travel_Frequently" "Travel_Rarely"
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Sales"
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : chr  "Life Sciences" "Medical" "Life Sciences" "Marketing" ...
```

```
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : chr  "Male" "Male" "Male" "Female" ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : int  3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel                : int  2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : chr  "Sales Executive" "Research Director" "Manufacturing Director" "Sal
##  $ JobSatisfaction         : int  4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : chr  "Divorced" "Single" "Single" "Married" ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
##  $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
##  $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr  "No" "No" "No" "No" ...
##  $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
##  $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
##  $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
##  $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
##  $ WorkLifeBalance         : int  2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
##  $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
##  $ YearsSinceLastPromotion : int  0 4 2 5 1 1 0 0 0 7 ...
##  $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```
# Change all char variables to factors
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- data %>%
  mutate_if(is.character, as.factor)

str(data)
```

```
## 'data.frame':    870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
```

```
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 3 2 2 2 3 3 2 ...
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 2 3 6 2 4 2 2 6 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 1 1 2 ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : int  3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel                : int  2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1
##  $ JobSatisfaction         : int  4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 1 3 3 2 3 1 2 1 2 2 ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
##  $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
##  $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
##  $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
##  $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
##  $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
##  $ WorkLifeBalance         : int  2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
##  $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
##  $ YearsSinceLastPromotion : int  0 4 2 5 1 1 0 0 0 7 ...
##  $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```r
set.seed(314)

obv <- nrow(data) # number of obervations
shuff_obv <- sample(obv) # shuffled obv index
data_shuff <- data[shuff_obv,] # shuffled data
split <- round(obv*0.80)

train <- data_shuff[1:split,] #  train
test <- data_shuff[(split+1):obv,] # test

table(train$Attrition)
```

```
##
##  No Yes
## 586 110
```

Classification

```r
str(train)
```

```
## 'data.frame':    696 obs. of  36 variables:
```

```
##  $ ID                        : int  334 384 60 387 419 423 770 246 831 620 ...
##  $ Age                       : int  27 23 25 35 37 31 45 29 45 35 ...
##  $ Attrition                 : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 1 1 1 1 ...
##  $ BusinessTravel            : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 2 3 3 1 3
##  $ DailyRate                 : int  608 541 599 944 977 249 1050 1086 248 727 ...
##  $ Department                : Factor w/ 3 levels "Human Resources",..: 2 3 3 3 2 3 3 2 2 2 ...
##  $ DistanceFromHome          : int  1 2 24 1 1 6 9 7 23 3 ...
##  $ Education                 : int  2 1 1 3 3 4 4 1 2 3 ...
##  $ EducationField            : Factor w/ 6 levels "Human Resources",..: 2 6 2 3 2 2 2 4 2 2 ...
##  $ EmployeeCount             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber            : int  725 113 1273 314 1196 163 1117 912 1002 704 ...
##  $ EnvironmentSatisfaction : int  3 3 3 3 4 2 2 1 4 3 ...
##  $ Gender                    : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 1 2 2 ...
##  $ HourlyRate                : int  68 62 73 92 56 76 65 62 42 41 ...
##  $ JobInvolvement            : int  3 3 1 3 2 1 2 2 3 2 ...
##  $ JobLevel                  : int  3 1 1 3 2 2 2 1 2 1 ...
##  $ JobRole                   : Factor w/ 9 levels "Healthcare Representative",..: 5 9 9 8 5 8 8 3 3 3
##  $ JobSatisfaction           : int  1 1 4 3 4 3 3 4 1 3 ...
##  $ MaritalStatus             : Factor w/ 3 levels "Divorced","Married",..: 2 1 3 3 2 2 2 1 2 2 ...
##  $ MonthlyIncome             : int  7412 2322 1118 8789 6474 6172 5593 2532 3633 1281 ...
##  $ MonthlyRate               : int  6009 9518 8040 9096 9961 20739 17970 6054 14039 16900 ...
##  $ NumCompaniesWorked        : int  1 3 1 1 1 4 1 6 1 1 ...
##  $ Over18                    : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                  : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 1 1 2 1 ...
##  $ PercentSalaryHike         : int  11 13 14 14 13 18 13 14 15 18 ...
##  $ PerformanceRating         : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  4 3 4 1 2 2 4 3 3 3 ...
##  $ StandardHours             : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel          : int  0 1 0 0 1 0 1 3 1 2 ...
##  $ TotalWorkingYears         : int  9 3 1 10 14 12 15 8 9 1 ...
##  $ TrainingTimesLastYear     : int  3 3 4 3 2 3 2 5 2 3 ...
##  $ WorkLifeBalance           : int  3 3 3 4 2 2 3 3 3 3 ...
##  $ YearsAtCompany            : int  9 0 1 10 14 7 15 4 9 1 ...
##  $ YearsInCurrentRole        : int  7 0 0 7 8 7 10 3 8 0 ...
##  $ YearsSinceLastPromotion : int  0 0 1 0 3 7 4 0 0 0 ...
##  $ YearsWithCurrManager      : int  7 0 0 8 11 7 12 3 8 0 ...
```

Checking For Multicollinairity

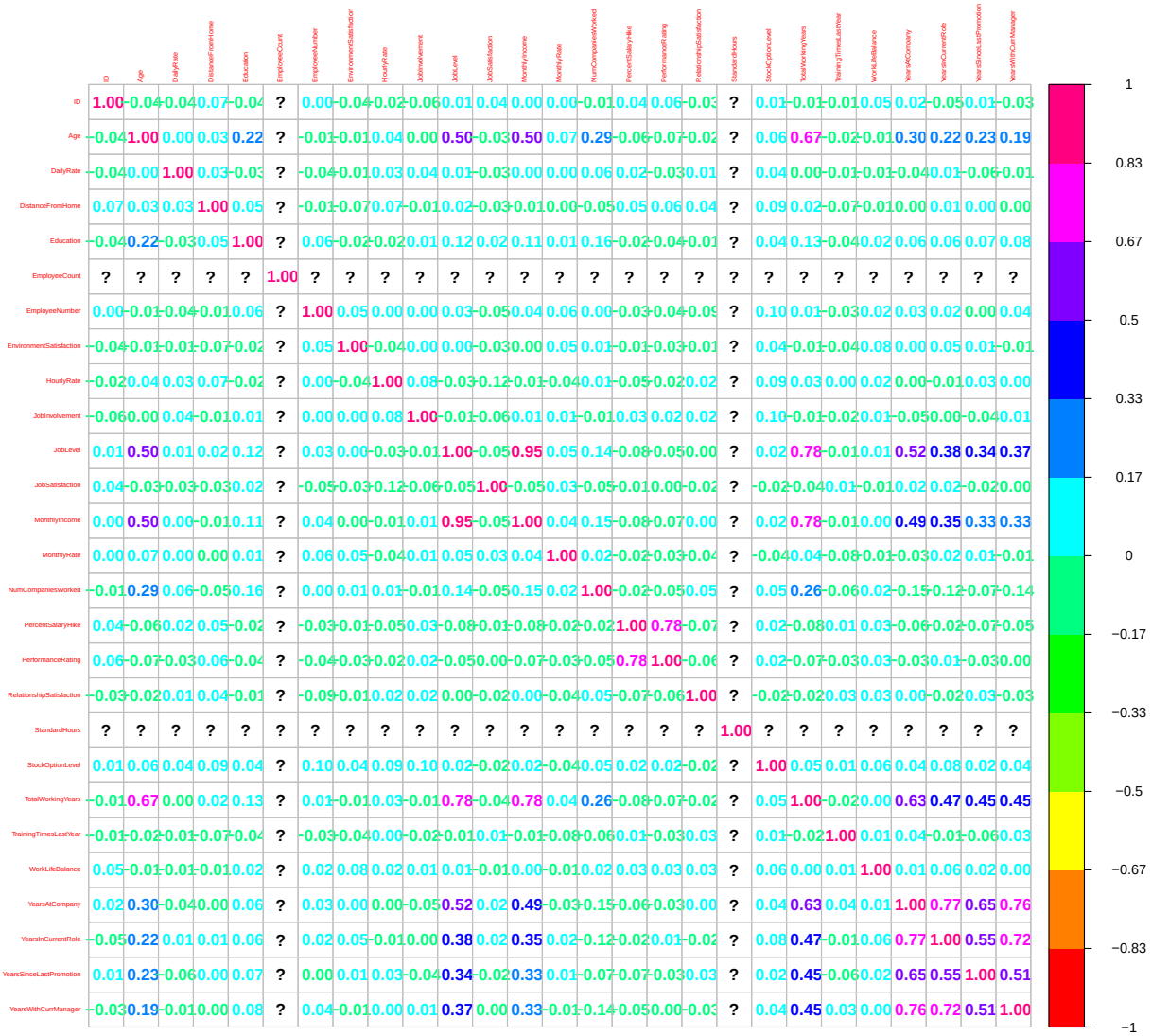
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
num_train <- train %>%
  select_if(~!is.character(.x) & !is.factor(.x))

M <- cor(num_train)
```

```
## Warning in cor(num_train): the standard deviation is zero
```

```
corrplot(M, method="number",col = rainbow(12), tl.cex = 0.4)
```



Eliminating High Corrlated variables

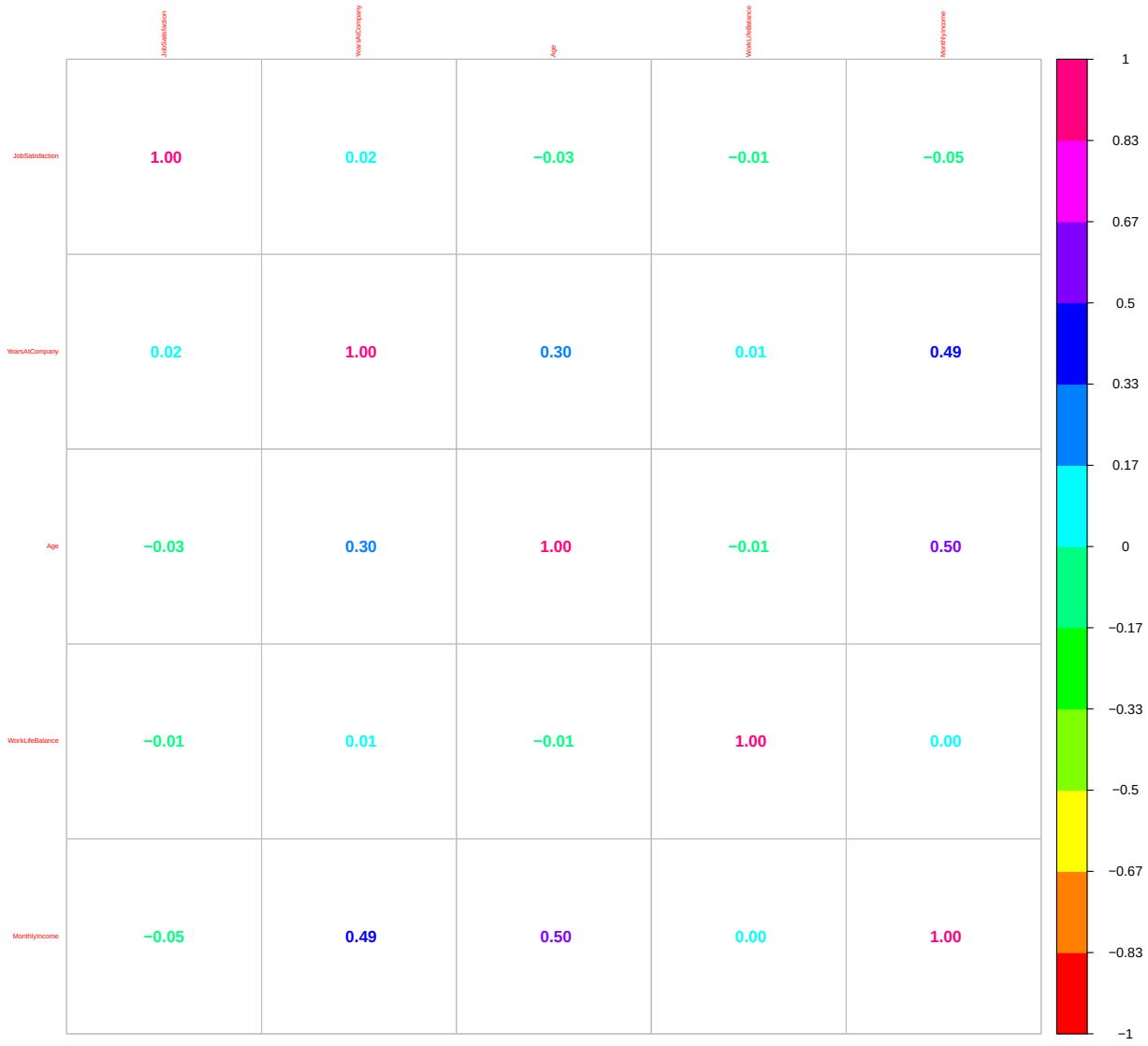
```
train_top6 <- train %>%
              select(JobSatisfaction,YearsAtCompany,Age,WorkLifeBalance,MonthlyIncome,JobRole,Attrition)

head(train_top6)
```

```
##     JobSatisfaction YearsAtCompany Age WorkLifeBalance MonthlyIncome
## 334               1              9  27               3          7412
## 384               1              0  23               3          2322
## 60                4              1  25               3          1118
## 387               3             10  35               4          8789
## 419               4             14  37               2          6474
## 423               3              7  31               2          6172
```

```
##                     JobRole Attrition
## 334 Manufacturing Director        No
## 384   Sales Representative        No
## 60    Sales Representative       Yes
## 387        Sales Executive        No
## 419 Manufacturing Director        No
## 423        Sales Executive       Yes
```

Check the corr for the numeric variables

```
num_train <- train_top6 %>%
  select_if(~!is.character(.x) & !is.factor(.x))

M <- cor(num_train)
corrplot(M, method="number",col = rainbow(12), tl.cex = 0.4)
```



There is no corr thats greater than abs of 55

```
library(GGally)
```

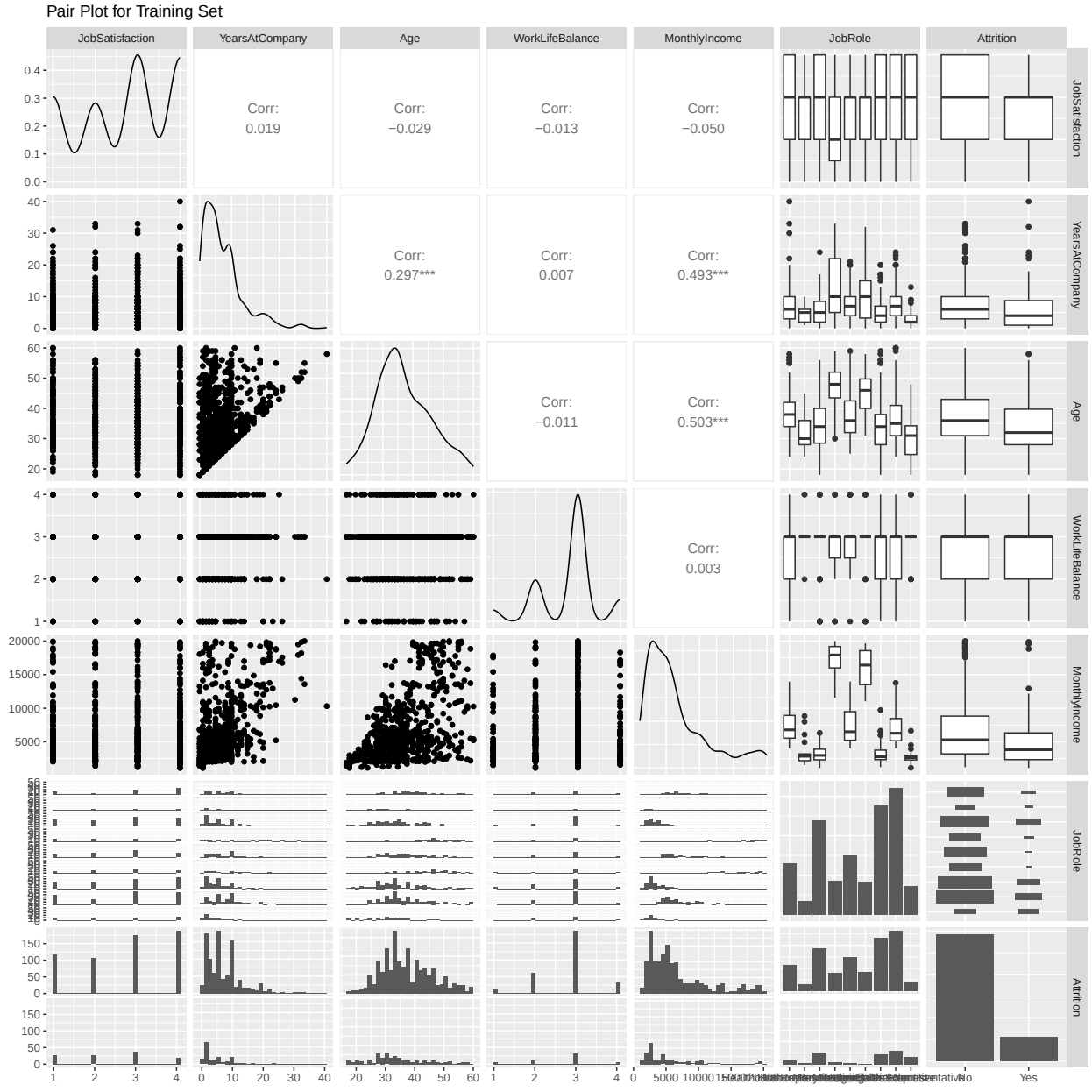
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(train_top6, title = "Pair Plot for Training Set")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Down sampling

```
library(caret)
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(314)
train_top6_ds <- downSample(x = train_top6[, -ncol(train_top6)],
                            y = train_top6$Attrition,
                            yname = "Attrition")

table(train_top6_ds$Attrition)
```

```
##
##  No Yes
## 110 110
```

```
colnames(train_top6_ds)
```

```
## [1] "JobSatisfaction" "YearsAtCompany"  "Age"             "WorkLifeBalance"
## [5] "MonthlyIncome"   "JobRole"         "Attrition"
```

select best 3 factors

```
train_top3_ds <- train_top6_ds %>%
  select(JobSatisfaction,WorkLifeBalance,YearsAtCompany,Attrition)

head(train_top3_ds)
```

```
##   JobSatisfaction WorkLifeBalance YearsAtCompany Attrition
## 1               3               3              8        No
## 2               2               3              4        No
## 3               2               3              8        No
## 4               3               3             14        No
## 5               3               2              1        No
## 6               3               3             21        No
```

```
test_top3 <- test %>%
  select(JobSatisfaction,WorkLifeBalance,YearsAtCompany,Attrition)

head(test_top3)
```

```
##     JobSatisfaction WorkLifeBalance YearsAtCompany Attrition
## 97                2               4             13        No
## 517               2               3              1       Yes
## 134               2               3              5        No
## 282               1               3             18        No
## 439               3               3              3       Yes
## 182               1               3              7        No
```

now model making knn

```
set.seed(314)


#  summaryFunction = f1Summary
ctrl <- trainControl(method = "LOOCV")


knn.model <- train(Attrition ~ .,
                   data = train_top3_ds,
                   method = 'knn',
                   trControl = ctrl,
                   tuneGrid = data.frame(k = c(1:100)))

best.knn.model <- knn3(Attrition ~ .,
                       data = train_top3_ds,
                       k = knn.model$bestTune$k)

predictions1 <- predict(best.knn.model, test_top3, type = "class")
cm <- confusionMatrix(predictions1, test_top3$Attrition)
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  106   9
##        Yes  38  21
##
##                Accuracy : 0.7299
##                  95% CI : (0.6575, 0.7943)
##     No Information Rate : 0.8276
##     P-Value [Acc > NIR] : 0.9995
##
##                   Kappa : 0.3154
##
##  Mcnemar's Test P-Value : 4.423e-05
##
##             Sensitivity : 0.7361
##             Specificity : 0.7000
##          Pos Pred Value : 0.9217
##          Neg Pred Value : 0.3559
##              Prevalence : 0.8276
##          Detection Rate : 0.6092
##    Detection Prevalence : 0.6609
##       Balanced Accuracy : 0.7181
##
##        'Positive' Class : No
##
```

```
knn.model$bestTune$k
```

```
## [1] 34
```

Naitve Bayes

```
set.seed(314)
library(naivebayes)
```

```
## naivebayes 0.9.7 loaded
```

```
#  summaryFunction = f1Summary
ctrl <- trainControl(method = "LOOCV")

#tg <- expand.grid(usekernel =c(T,F),
                  #adjust = seq(0.5,2,by=0.1),
                  #laplace  = c(0,1))


best.nb.model <- train(Attrition ~ .,
                   data = train_top3_ds,
                   method = 'naive_bayes',
                   trControl = ctrl)


predictions2 <- predict(best.nb.model, test_top3, type = "raw")
cm <- confusionMatrix(predictions2, test_top3$Attrition)
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  89   8
##        Yes 55  22
##
##                Accuracy : 0.6379
##                  95% CI : (0.5618, 0.7093)
##     No Information Rate : 0.8276
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2169
##
##  Mcnemar's Test P-Value : 6.814e-09
##
##             Sensitivity : 0.6181
##             Specificity : 0.7333
##          Pos Pred Value : 0.9175
##          Neg Pred Value : 0.2857
##              Prevalence : 0.8276
##          Detection Rate : 0.5115
##    Detection Prevalence : 0.5575
##       Balanced Accuracy : 0.6757
##
##        'Positive' Class : No
##
```

compair models

```
cm_nb <- confusionMatrix(predictions2, test_top3$Attrition)
cm_knn <- confusionMatrix(predictions1, test_top3$Attrition)

#  accuracy
accuracy_nb <- cm_nb$overall['Accuracy']
accuracy_knn <- cm_knn$overall['Accuracy']

sensitivity_nb <- cm_nb$byClass['Sensitivity']
specificity_nb <- cm_nb$byClass['Specificity']

sensitivity_knn <- cm_knn$byClass['Sensitivity']
specificity_knn <- cm_knn$byClass['Specificity']

#  F1 Score
f1_nb <- (cm_nb$byClass['Precision'] * cm_nb$byClass['Recall']) / (cm_nb$byClass['Precision'] + cm_nb$by

f1_knn <- (cm_knn$byClass['Precision'] * cm_knn$byClass['Recall']) / (cm_knn$byClass['Precision'] + cm_k

cat("Naive Bayes - Sensitivity:", sensitivity_nb, "Specificity:", specificity_nb, "\n")
```

```
## Naive Bayes - Sensitivity: 0.6180556 Specificity: 0.7333333
```

```
cat("KNN - Sensitivity:", sensitivity_knn, "Specificity:", specificity_knn, "\n\n\n\n\n")
```

```
## KNN - Sensitivity: 0.7361111 Specificity: 0.7
```

```
cat("Naive Bayes - Accuracy:", accuracy_nb, "F1 Score:", f1_nb, "\n")
```

```
## Naive Bayes - Accuracy: 0.637931 F1 Score: 0.7385892
```

```
cat("KNN - Accuracy:", accuracy_knn, "F1 Score:", f1_knn, "\n")
```

```
## KNN - Accuracy: 0.7298851 F1 Score: 0.8185328
```

unlabeled data

```
class_data <- read.csv("CaseStudy2CompSet No Attrition.csv")
class_data_top3 <- class_data %>%
  select(ID,JobSatisfaction,WorkLifeBalance,YearsAtCompany)

head(class_data_top3)
```

```
##     ID JobSatisfaction WorkLifeBalance YearsAtCompany
## 1 1171               3               2             10
## 2 1172               3               3              5
## 3 1173               3               2              1
## 4 1174               4               3              5
## 5 1175               3               3             10
## 6 1176               1               2             13
```

```
row.names(class_data_top3) <- class_data_top3$ID
class_data_top3$ID <- NULL

head(class_data_top3)
```

```
##      JobSatisfaction WorkLifeBalance YearsAtCompany
## 1171               3               2             10
## 1172               3               3              5
## 1173               3               2              1
## 1174               4               3              5
## 1175               3               3             10
## 1176               1               2             13
```

```
final_predictions <- predict(best.knn.model, class_data_top3, type = "class")
df <- data.frame(ID = rownames(class_data_top3), Attrition = final_predictions )
```

```
write.csv(df, "Case2PredictionsWeaver Attrition.csv", row.names = FALSE)
```

Regression

```
head(data)
```

```
##   ID Age Attrition    BusinessTravel DailyRate              Department
## 1  1  32        No     Travel_Rarely       117                   Sales
## 2  2  40        No     Travel_Rarely      1308 Research & Development
## 3  3  35        No Travel_Frequently       200 Research & Development
## 4  4  32        No     Travel_Rarely       801                   Sales
## 5  5  24        No Travel_Frequently       567 Research & Development
## 6  6  27        No Travel_Frequently       294 Research & Development
##   DistanceFromHome Education   EducationField EmployeeCount EmployeeNumber
## 1               13         4    Life Sciences             1            859
## 2               14         3          Medical             1           1128
## 3               18         2    Life Sciences             1           1412
## 4                1         4        Marketing             1           2016
## 5                2         1 Technical Degree             1           1646
## 6               10         2    Life Sciences             1            733
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2   Male         73              3        2
## 2                       3   Male         44              2        5
## 3                       3   Male         60              3        3
## 4                       3 Female         48              3        3
## 5                       1 Female         32              3        1
## 6                       4   Male         32              3        3
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1        Sales Executive               4      Divorced          4403
## 2     Research Director               3        Single         19626
## 3 Manufacturing Director               4        Single          9362
## 4        Sales Executive               4       Married         10422
## 5     Research Scientist               4        Single          3760
## 6 Manufacturing Director               1      Divorced          8793
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1        9250                 2      Y       No               11
```

```
## 2       17544                  1      Y       No                14
## 3       19944                  2      Y       No                11
## 4       24032                  1      Y       No                19
## 5       17218                  1      Y      Yes                13
## 6        4809                  1      Y       No                21
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                        3            80                1
## 2                 3                        1            80                0
## 3                 3                        3            80                0
## 4                 3                        3            80                2
## 5                 3                        3            80                0
## 6                 4                        3            80                2
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     3               2              5
## 2                21                     2               4             20
## 3                10                     2               3              2
## 4                14                     3               3             14
## 5                 6                     2               3              6
## 6                 9                     4               2              9
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  2                       0                    3
## 2                  7                       4                    9
## 3                  2                       2                    2
## 4                 10                       5                    7
## 5                  3                       1                    3
## 6                  7                       1                    7
```

```r
set.seed(314)

obv <- nrow(data) # number of obervations
shuff_obv <- sample(obv) # shuffled obv index
data_shuff <- data[shuff_obv,] # shuffled data
split <- round(obv*0.80)

train <- data_shuff[1:split,] #  train
test <- data_shuff[(split+1):obv,] # test
```

```r
# train
train_top10 <- train%>%
  select(JobLevel,TotalWorkingYears,JobRole,DistanceFromHome,Education,YearsAtCompany,YearsInCurrentRol

head(train_top10)
```

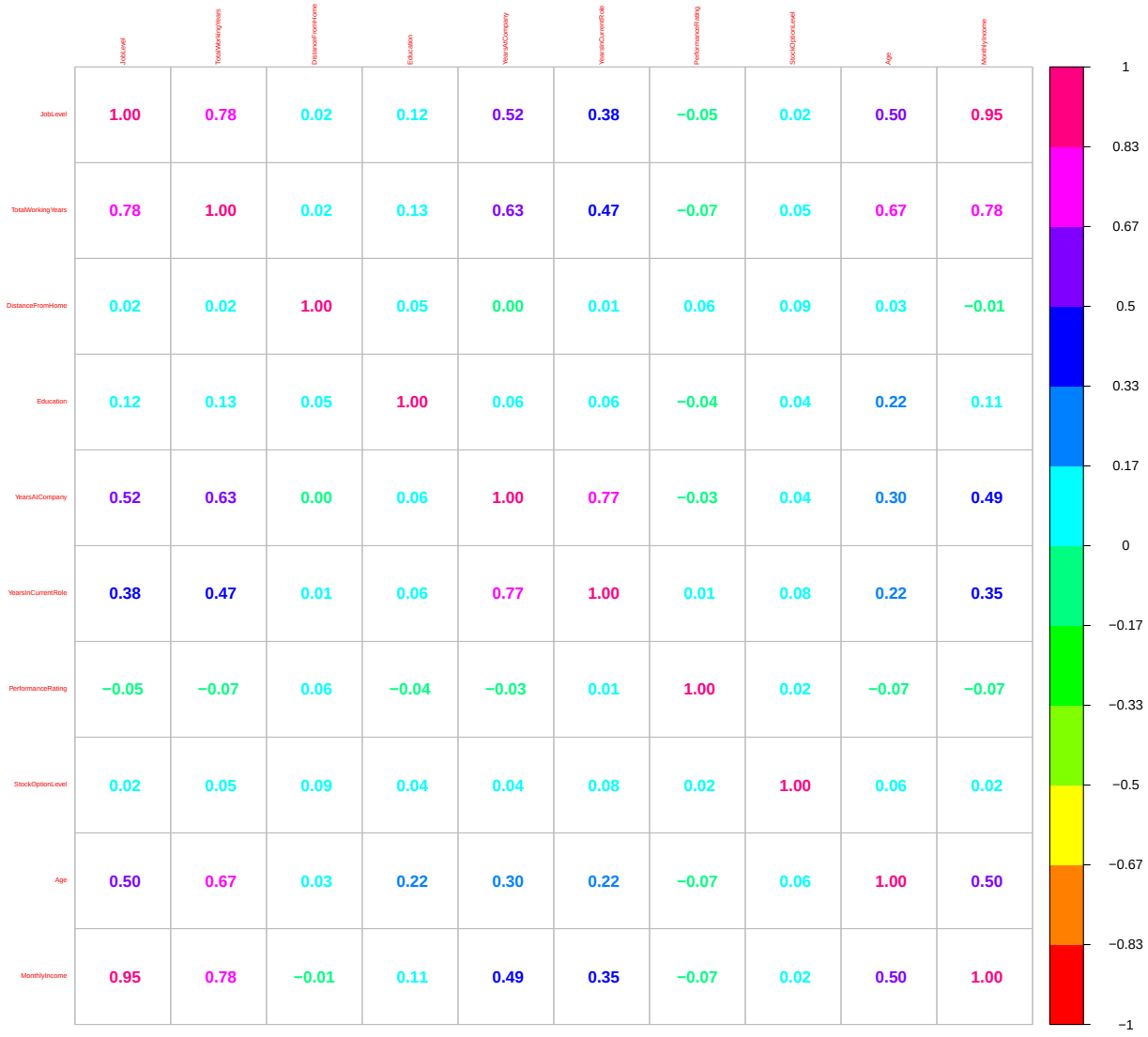
```
##     JobLevel TotalWorkingYears               JobRole DistanceFromHome
## 334        3                 9 Manufacturing Director                1
## 384        1                 3   Sales Representative                2
## 60         1                 1   Sales Representative               24
## 387        3                10        Sales Executive                1
## 419        2                14 Manufacturing Director                1
## 423        2                12        Sales Executive                6
##     Education YearsAtCompany YearsInCurrentRole PerformanceRating
## 334         2              9                  7                 3
## 384         1              0                  0                 3
```

```
## 60          1              1                  0                 3
## 387         3             10                  7                 3
## 419         3             14                  8                 3
## 423         4              7                  7                 3
##     StockOptionLevel Age MonthlyIncome
## 334                0  27          7412
## 384                1  23          2322
## 60                 0  25          1118
## 387                0  35          8789
## 419                1  37          6474
## 423                0  31          6172
```

```
str(train_top10)
```

```
## 'data.frame':    696 obs. of  11 variables:
##  $ JobLevel          : int  3 1 1 3 2 2 2 1 2 1 ...
##  $ TotalWorkingYears : int  9 3 1 10 14 12 15 8 9 1 ...
##  $ JobRole           : Factor w/ 9 levels "Healthcare Representative",..: 5 9 9 8 5 8 8 3 3 3 ...
##  $ DistanceFromHome  : int  1 2 24 1 1 6 9 7 23 3 ...
##  $ Education         : int  2 1 1 3 3 4 4 1 2 3 ...
##  $ YearsAtCompany    : int  9 0 1 10 14 7 15 4 9 1 ...
##  $ YearsInCurrentRole: int  7 0 0 7 8 7 10 3 8 0 ...
##  $ PerformanceRating : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ StockOptionLevel  : int  0 1 0 0 1 0 1 3 1 2 ...
##  $ Age               : int  27 23 25 35 37 31 45 29 45 35 ...
##  $ MonthlyIncome     : int  7412 2322 1118 8789 6474 6172 5593 2532 3633 1281 ...
```

```
num_train <- train_top10 %>%
  select_if(~!is.character(.x) & !is.factor(.x))

M <- cor(num_train)
corrplot(M, method="number",col = rainbow(12), tl.cex = 0.4)
```



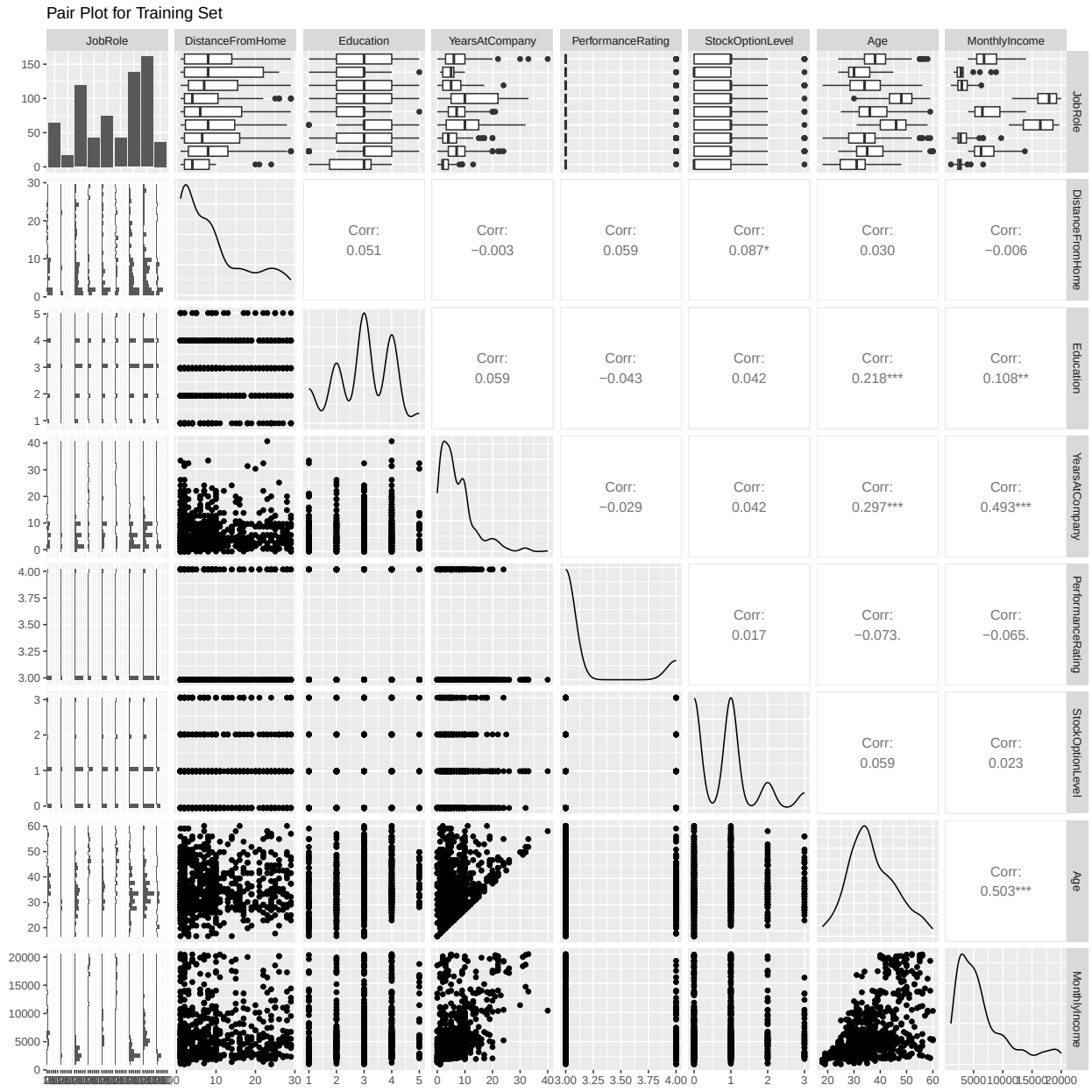

```
# train
train_top7 <- train_top10%>%
  select(JobRole,DistanceFromHome,Education,YearsAtCompany,PerformanceRating, StockOptionLevel, Age, Mo

head(train_top7)
```

```
##                     JobRole DistanceFromHome Education YearsAtCompany
## 334 Manufacturing Director                 1         2              9
## 384   Sales Representative                 2         1              0
## 60    Sales Representative                24         1              1
## 387         Sales Executive                1         3             10
## 419 Manufacturing Director                 1         3             14
## 423         Sales Executive                6         4              7
```

```
##     PerformanceRating StockOptionLevel Age MonthlyIncome
## 334                 3                0  27          7412
## 384                 3                1  23          2322
## 60                  3                0  25          1118
## 387                 3                0  35          8789
## 419                 3                1  37          6474
## 423                 3                0  31          6172
```

```
ggpairs(train_top7, title = "Pair Plot for Training Set")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
train_top3 <- train_top7%>%
  select(JobRole,YearsAtCompany,Age,MonthlyIncome)

head(train_top3)
```

```
##                     JobRole YearsAtCompany Age MonthlyIncome
## 334 Manufacturing Director              9  27          7412
## 384    Sales Representative              0  23          2322
## 60     Sales Representative              1  25          1118
## 387         Sales Executive             10  35          8789
## 419 Manufacturing Director             14  37          6474
## 423         Sales Executive              7  31          6172
```

```r
test_top3 <- test%>%
  select(JobRole,YearsAtCompany,Age,MonthlyIncome)

head(test_top3)
```

```
##                    JobRole YearsAtCompany Age MonthlyIncome
## 97                 Manager             13  46         16606
## 517   Sales Representative              1  25          2413
## 134 Manufacturing Director              5  39          5295
## 282  Laboratory Technician             18  41          4721
## 439   Sales Representative              3  50          2683
## 182         Sales Executive              7  38          6893
```

Regression

```r
set.seed(314)


ctrl <- trainControl(method = "LOOCV")


best.lm.model <- train(MonthlyIncome ~ .,
                   data = test_top3,
                   method = 'lm',
                   trControl = ctrl)


predictions <- predict(best.lm.model, test_top3[,-4])
```

```r
results <- postResample(predictions, test_top3$MonthlyIncome)
results
```

```
##         RMSE     Rsquared          MAE
## 1772.3572817    0.8281887 1351.3877132
```

```r
summary(best.lm.model$finalModel)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4043.6 -1226.6  -139.4   796.8  6853.6
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4302.78     879.16   4.894 2.36e-06 ***
## `JobRoleHuman Resources`       -4283.97     819.60  -5.227 5.22e-07 ***
## `JobRoleLaboratory Technician` -3796.13     657.87  -5.770 3.88e-08 ***
## JobRoleManager                  8128.80     870.85   9.334  < 2e-16 ***
```

```
## `JobRoleManufacturing Director`   461.47     768.92   0.600 0.549239
## `JobRoleResearch Director`       7372.83     830.63   8.876 1.19e-15 ***
## `JobRoleResearch Scientist`     -3629.60     658.97  -5.508 1.39e-07 ***
## `JobRoleSales Executive`         -528.71     635.71  -0.832 0.406803
## `JobRoleSales Representative`   -3887.82     760.69  -5.111 8.89e-07 ***
## YearsAtCompany                    94.17      30.17   3.121 0.002130 **
## Age                               61.80      16.71   3.698 0.000296 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1831 on 163 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:  0.8176
## F-statistic: 78.57 on 10 and 163 DF,  p-value: < 2.2e-16
```

```
library(readxl)

reg_data <- read_xlsx("CaseStudy2CompSet No Salary.xlsx")
reg_data_top3 <- class_data %>%
  select(ID,JobRole,YearsAtCompany,Age,MonthlyIncome)

row.names(reg_data_top3) <- reg_data_top3$ID
reg_data_top3$ID <- NULL


final_predictions <- predict(best.lm.model, reg_data_top3)
df <- data.frame(ID = rownames(reg_data_top3), MonthlyI = final_predictions )
```

```
write.csv(df, "Case2PredictionsWeaver Salary.csv", row.names = FALSE)
```