100634271\_3.R

USER

2022-11-27

# ANALYSIS OF HR EMPLOYEE ATTRITION DATASET  
  
# load the necessary libraries  
library(tidyverse)  
  
# Force R not to use scientific notations  
options(scipen = 99)  
  
# Load the wine quality dataset  
# It is the first data statistical analyses will be performed on  
  
hr\_raw <- read.csv(  
 "C:\\Users\\USER\\Desktop\\100634271\\HR Employee Attrition.csv")  
  
str(hr\_raw)

## 'data.frame': 1470 obs. of 35 variables:  
## $ Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ Attrition : chr "Yes" "No" "Yes" "No" ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" "Travel\_Frequently" ...  
## $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...  
## $ 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 : chr "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...  
## $ 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 : chr "Female" "Male" "Male" "Female" ...  
## $ 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 : chr "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...  
## $ JobSatisfaction : int 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : chr "Single" "Married" "Single" "Married" ...  
## $ 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 : chr "Y" "Y" "Y" "Y" ...  
## $ OverTime : chr "Yes" "No" "Yes" "Yes" ...  
## $ 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 ...

# Dataset Description  
# The HR Employee Attrition Dataset has 35 variables and 1470 observations  
# The variables are of different data types  
# Some are numeric in nature, while some others are of character data types.  
  
# Not all the variables in the data will be used for analysis in this section.  
  
# The focus of this section is logistic regression analysis  
# The relationship between the dependent variable: Attrition and some   
# independent variables will be examined in this section.  
  
# select the columns needed for analysis  
  
hr <- hr\_raw %>%   
 select(c(Age,Attrition,DailyRate,DistanceFromHome,Education,  
 EnvironmentSatisfaction,Gender,  
 HourlyRate,MonthlyIncome,MonthlyRate,  
 OverTime,TotalWorkingYears,YearsAtCompany))  
  
# Encode Attrition column which is a categorical variable  
hr$Attrition <- factor(hr$Attrition, levels = c("Yes","No"),labels = c(1,0))  
  
  
# Perform summary statistics  
summary(hr)

## Age Attrition DailyRate DistanceFromHome Education   
## Min. :18.00 1: 237 Min. : 102.0 Min. : 1.000 Min. :1.000   
## 1st Qu.:30.00 0:1233 1st Qu.: 465.0 1st Qu.: 2.000 1st Qu.:2.000   
## Median :36.00 Median : 802.0 Median : 7.000 Median :3.000   
## Mean :36.92 Mean : 802.5 Mean : 9.193 Mean :2.913   
## 3rd Qu.:43.00 3rd Qu.:1157.0 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :60.00 Max. :1499.0 Max. :29.000 Max. :5.000   
## EnvironmentSatisfaction Gender HourlyRate MonthlyIncome   
## Min. :1.000 Length:1470 Min. : 30.00 Min. : 1009   
## 1st Qu.:2.000 Class :character 1st Qu.: 48.00 1st Qu.: 2911   
## Median :3.000 Mode :character Median : 66.00 Median : 4919   
## Mean :2.722 Mean : 65.89 Mean : 6503   
## 3rd Qu.:4.000 3rd Qu.: 83.75 3rd Qu.: 8379   
## Max. :4.000 Max. :100.00 Max. :19999   
## MonthlyRate OverTime TotalWorkingYears YearsAtCompany   
## Min. : 2094 Length:1470 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 8047 Class :character 1st Qu.: 6.00 1st Qu.: 3.000   
## Median :14236 Mode :character Median :10.00 Median : 5.000   
## Mean :14313 Mean :11.28 Mean : 7.008   
## 3rd Qu.:20462 3rd Qu.:15.00 3rd Qu.: 9.000   
## Max. :26999 Max. :40.00 Max. :40.000

logmodel <- glm(Attrition~ Age+DailyRate+DistanceFromHome+Education+  
 EnvironmentSatisfaction+HourlyRate+MonthlyIncome+  
 TotalWorkingYears+YearsAtCompany,  
 data=hr, family = "binomial")  
  
summary(logmodel)

##   
## Call:  
## glm(formula = Attrition ~ Age + DailyRate + DistanceFromHome +   
## Education + EnvironmentSatisfaction + HourlyRate + MonthlyIncome +   
## TotalWorkingYears + YearsAtCompany, family = "binomial",   
## data = hr)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1097 0.2994 0.5100 0.6399 1.1959   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.79696607 0.49078354 -1.624 0.10440   
## Age 0.02668543 0.01153956 2.313 0.02075 \*   
## DailyRate 0.00035844 0.00018249 1.964 0.04951 \*   
## DistanceFromHome -0.02685055 0.00866870 -3.097 0.00195 \*\*   
## Education -0.00886051 0.07336071 -0.121 0.90387   
## EnvironmentSatisfaction 0.26189556 0.06655292 3.935 0.0000831 \*\*\*  
## HourlyRate 0.00142686 0.00362861 0.393 0.69415   
## MonthlyIncome 0.00006402 0.00002995 2.138 0.03254 \*   
## TotalWorkingYears 0.01228659 0.02232106 0.550 0.58201   
## YearsAtCompany 0.04024889 0.02081258 1.934 0.05313 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.6 on 1469 degrees of freedom  
## Residual deviance: 1204.9 on 1460 degrees of freedom  
## AIC: 1224.9  
##   
## Number of Fisher Scoring iterations: 5

# INDEPENDENT SAMPLES T-TEST  
  
# Using the HR Employee Attrition Dataset, we aim to test if there is a   
# significant difference in the monthly income earned by male employees  
# and the monthly income earned by female employees  
  
  
# Let's first subset the data  
  
male\_income<- hr$MonthlyIncome[hr$Gender=="Male"]  
female\_income<- hr$MonthlyIncome[hr$Gender=="Female"]  
  
# Conduct the T-test  
test\_res1<- t.test(male\_income,female\_income)  
  
test\_res1

##   
## Welch Two Sample t-test  
##   
## data: male\_income and female\_income  
## t = -1.2223, df = 1261.5, p-value = 0.2218  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -797.3111 185.1943  
## sample estimates:  
## mean of x mean of y   
## 6380.508 6686.566

# Using the HR Employee Attrition Dataset, we aim to also test if there is a   
# significant difference in the age of the employees who attrit and   
# those who do not attrit  
# Let's first subset the data  
  
yes\_attrit\_age<- hr$Age[hr$Attrition=="1"]  
no\_attrit\_age<- hr$Age[hr$Attrition=="0"]  
  
  
# Conduct the T-test  
test\_res2<- t.test(yes\_attrit\_age,no\_attrit\_age)  
  
test\_res2

##   
## Welch Two Sample t-test  
##   
## data: yes\_attrit\_age and no\_attrit\_age  
## t = -5.828, df = 316.93, p-value = 0.0000000138  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -5.288346 -2.618930  
## sample estimates:  
## mean of x mean of y   
## 33.60759 37.56123

# Using the HR Employee Attrition Dataset, we aim to also test if there is a   
# significant difference in the years worked at a company   
# and whether they do overnight or not.  
# Let's first subset the data  
  
yes\_overtime\_years\_at\_company<- hr$YearsAtCompany[hr$OverTime=="Yes"]  
no\_overtime\_years\_at\_company<- hr$YearsAtCompany[hr$OverTime=="No"]  
  
# Conduct the T-test  
test\_res3<- t.test(yes\_overtime\_years\_at\_company,no\_overtime\_years\_at\_company)  
  
test\_res3

##   
## Welch Two Sample t-test  
##   
## data: yes\_overtime\_years\_at\_company and no\_overtime\_years\_at\_company  
## t = -0.43256, df = 709.49, p-value = 0.6655  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.8801244 0.5623240  
## sample estimates:  
## mean of x mean of y   
## 6.894231 7.053131