sriram\_G01230629\_13.R

sumanth

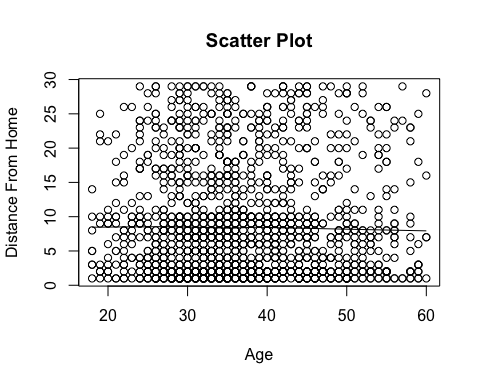
2019-11-17

###Regression and Clustering

###Students Name: Sriram Sai Sumanth  
###GNumber: G01230629  
  
  
rm(list=ls())  
  
library(ggplot2)  
  
data <- read.csv("./EmployeeAttrition.csv")  
  
#1.a Show the scatter plot with relationship curve between TotalWorkingYears and MonthlyIncome. Briefly explain your observation in the plot   
scatter.smooth(x = data$TotalWorkingYears,y = data$MonthlyIncome,main = "Scatter Plot",xlab = "Total Working Years",ylab = "Monthly Income")



#Here we can see that with increase in TotalWorkingYears the MonthlyIncome of the employee is also increased.  
  
#1.b Show the scatter plot with relationship curve between Age and DistanceFromHome. Briefly explain your observation in the plot  
scatter.smooth(x = data$Age,y = data$DistanceFromHome,main = "Scatter Plot",xlab = "Age",ylab = "Distance From Home")



#Here we can see that the Age and DistanceFromHome negligibly related.  
  
#1.c Calculate Correlation for (a) and (b) and explain the values to support your answer in (a) and (b)  
cor(x = data$TotalWorkingYears,y = data$MonthlyIncome,method = ("pearson"))

## [1] 0.7728932

cor.test(x = data$TotalWorkingYears,y = data$MonthlyIncome,method = ('pearson'))

##   
## Pearson's product-moment correlation  
##   
## data: data$TotalWorkingYears and data$MonthlyIncome  
## t = 46.669, df = 1468, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7514606 0.7926965  
## sample estimates:  
## cor   
## 0.7728932

cor(x = data$Age,y = data$DistanceFromHome,method = ("pearson"))

## [1] -0.00168612

cor.test(x = data$Age,y = data$DistanceFromHome,method=("pearson"))

##   
## Pearson's product-moment correlation  
##   
## data: data$Age and data$DistanceFromHome  
## t = -0.064603, df = 1468, p-value = 0.9485  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.05280902 0.04944560  
## sample estimates:  
## cor   
## -0.00168612

#When the correlation coefficient value is close to zero there is no or negligible relationship  
#between the attributes. Therefore the correlation coefficient of a is 0.7728932 showing that the attributes are related,   
#whereas the correlation coefficient of b is -0.00168612 showing that the attribues are negligibly related.   
  
#1.d Using Linear Regression, find details of the relationship between TotalWorkingYears and   
#MonthlyIncome. Explain results in terms of p-value at 95% confidence interval and determine   
#whether the relationship is significant or not  
a <- data$TotalWorkingYears  
b <- data$MonthlyIncome  
LinearMod <- lm(a~b)  
print(LinearMod)

##   
## Call:  
## lm(formula = a ~ b)  
##   
## Coefficients:  
## (Intercept) b   
## 2.973066 0.001277

summary(LinearMod)

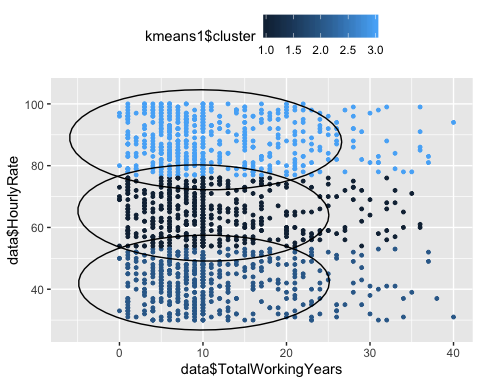
##   
## Call:  
## lm(formula = a ~ b)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2552 -3.5622 -0.7767 2.1552 24.7534   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.973e+00 2.197e-01 13.53 <2e-16 \*\*\*  
## b 1.277e-03 2.737e-05 46.67 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.939 on 1468 degrees of freedom  
## Multiple R-squared: 0.5974, Adjusted R-squared: 0.5971   
## F-statistic: 2178 on 1 and 1468 DF, p-value: < 2.2e-16

#Since the p-value(2.2e-16) is less than the significant level(0.05) both the attributes TotalWorkingYears and MonthlyIncome  
#are positively related.  
  
## 2. Clustering   
library(cluster)   
cl <- data.frame(data$TotalWorkingYears,data$HourlyRate)  
set.seed(100)  
kmeans1 <- kmeans(cl,3,iter.max=300,nstart = 10)  
  
plot(cl,col=kmeans1$cluster)



#visualize 3-clusters  
library(ggplot2)  
ggplot(cl)+  
 geom\_point(aes(x=data$TotalWorkingYears,y=data$HourlyRate,color=kmeans1$cluster),size=1)+  
 stat\_ellipse(aes(x=data$TotalWorkingYears,y=data$HourlyRate,group=kmeans1$cluster,type="norm"))+  
 theme(legend.position='top')

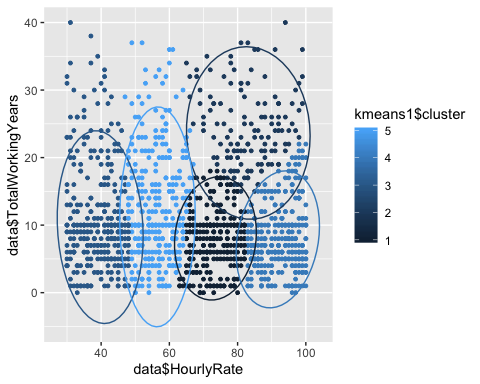
## Warning: Ignoring unknown aesthetics: type



#The employees are grouped with respect to total working years into 3 clusters, but the 3-clustering is not good enough to cluster the employees.  
#visualize 5 clusters  
kmeans1 <- kmeans(cl,5,iter.max=300,nstart = 10)  
plot(cl,col=kmeans1$cluster)



ggplot(cl)+  
 geom\_point(aes(y=data$TotalWorkingYears,x=data$HourlyRate,color=kmeans1$cluster),size=1)+  
 stat\_ellipse(aes(y=data$TotalWorkingYears,x=data$HourlyRate,group=kmeans1$cluster,color=kmeans1$cluster))



#The employees are grouped with respect to total working years into 5 clusters, making it a good division.