**EMPLOYEE ATTRITION**

**Submitted By :**

HEMAMALINI MADHANGURU

SRIHITHA REDDY SIVANNAGARI

Guided By : Dr.Shilpa Balan

**Objective:**

This project aims on analyzing the employee attrition level for a particular profession based on features such as job involvement, total number of working years, job satisfaction to retrieve information for the betterment of hiring and maintenance of employees.

**Dataset URL’s and Format:**

The dataset was obtained from the source: <http://www.scan-support.com/help/sample-data-sets>

Format: CSV

**Project Description:**

The dataset has 36 columns, 1471 rows in which attributes such as Employee Number, Department, Monthly Income Rate, Age will help to analyse about job role, type, salary etc. Additional features such as Education, Years of Working, Working Hours will help to analyse performance rate, job with longer work period etc.

Employee attrition study aims at the thought of varied causes and effects of losing employees and to produce an indepth analysis on the necessities of staff. Attrition could consult with the gradual reduction of the scale of employees by not commutating personnel lost through resignation (Batty Dorance Jeen, Jan-March 2014)

High wearing down rates bring about heightening enlistment and preparing expenses and parcel of time required in new worker change in accordance with the workplace and in this way upgrade their confidence (Chen, Ying-Chang, June 2010)

Working fulfillment can decrease nonattendance and worker turnover, it likewise can lessen the rate of mishaps. In any case, in the more drawn out term, unfriendly impacts, for example, loss of prepared workers, undiscovered profitability, and brought down resolve frequently convert into lower monetary benefits than foreseen. (Cascio, 2002). Salary Scale is also known for the most common cause of the employee turnover rate being so high. Employees are in search of jobs, which pay well. If the company, which they are working in, does not offer good and reasonable salary, they tend to hunt for jobs that pay them considerably well. The prospect of getting higher pay elsewhere is one of the most obvious contributors to turnover. This practice can be regularly observed at all levels of the economic ladder, from executives and generously paid professionals in high-stress positions to entry-level workers in relatively undemanding jobs

**Data Cleaning**

**1) Missing Values**

The dataset which we have taken does not have any missing values. Hence, we did not use the replace command to remove null values

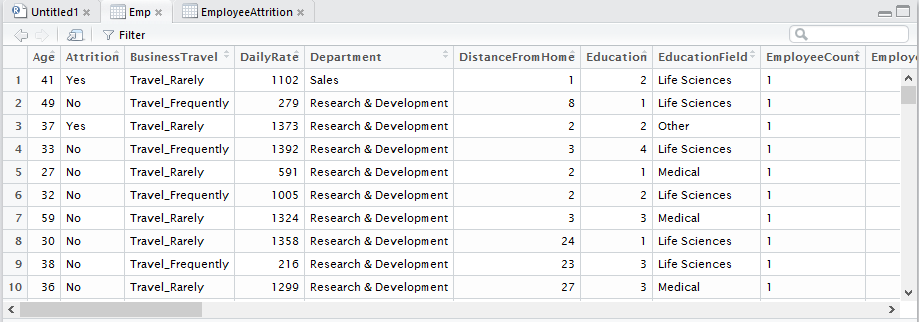
> setwd("~/HEMASIRI")

> getwd()

[1] "C:/Users/srihi/OneDrive/Documents/HEMASIRI"

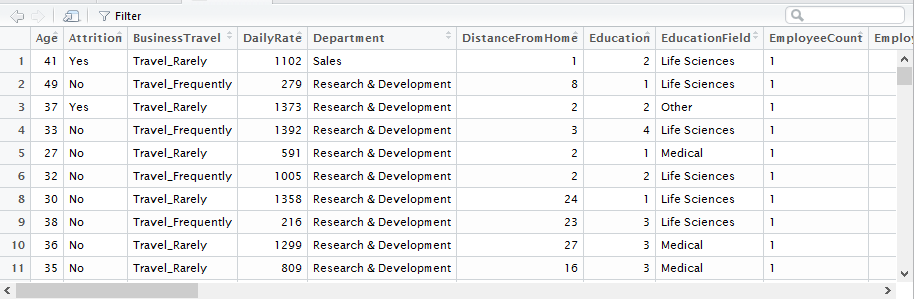
> Emp<-read.csv("Employee.csv", header = T, sep = ",")

> View(Emp)



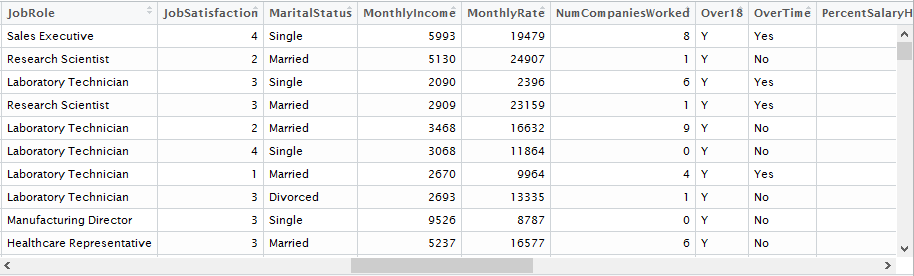
**2)Splitting column:**

The data set which we have taken doesn’t have any date column. So, we did not split any column using separate command nor combined any column using unite command



**3)Deleting Irrelevant Columns:**

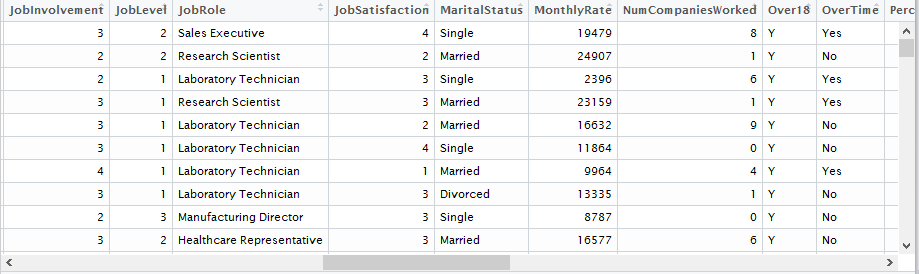
The data set has columns which have data regarding salary. The data set has hourly rate, daily rate, monthly rate and monthly income. The columns monthly rate and monthly income holds same information and is irrelevant



We have cleaned the data by deleting the column using the below command:

> EmployeeAttrition<-Emp[,-19]

> View(EmployeeAttrition)



4)REMOVING REDUNDANT DATA

We did not have any redundant data is our dataset.

|  |
| --- |
| >setwd("C:/CSULA/MSIS/spring 2017/5270")  > getwd()  [1] "C:/CSULA/MSIS/spring 2017/5270"    To show how to remove redundancy we have duplicated the values for few records and deleting them  using the following command :  > correct<-read.csv("incorrect.csv", header = T, sep = ",")  > View(correct)  >correct<-correct1[-7,]  >correct1<-correct[-7,]  > View(correct1)    ANALYSIS AND VISUALIZATIONS  **1) What age group does the highest over time among all the other working age groups?**  In our data set we have various age groups under working. The above visualization provides us an  insight of the highest age group that does works over time. It can easily be understood that the age  group which is highly involved in working overtime of all others is 30 to 40 age group in our data set.  We have used a histogram for better comprehension of the comparison. We initially assigned two  columns overtime and age to a data frame and then applied “split” for the histogram  C:\Users\Hema\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1.png  **2)Which job role has highest raise in the salary?**  In this visualization the percentage of salary hike of every category of job has  been analysed. This is an important factor in any department and hence we have  analysed this category. The pie chart depicts that professionals such as “Manager”,  “Research Scientist” have the highest salary hikes. With this analysis employees  in the future can enhance their job roles also choose the job accordingly if they  focus mainly on salary    3D Visualization Of Pie Chart  The same pie chart has been depicted in three dimensional view  C:\Users\Hema\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.png  R CODES  PIE  Script :  grouping <- group\_by(Employee,JobRole) %>% summarise(average\_hike = round(mean(PercentSalaryHike),2))  d=data.frame(grouping)  hike<-d[,c('JobRole','average\_hike')]  print(hike)  p <- plot\_ly(hike, labels = ~JobRole, values = ~average\_hike, type = 'pie',  textposition = 'inside',  textinfo = 'label+percent',  insidetextfont = list(color = '#FFFFFF'),  hoverinfo = 'text',  text = ~paste('$', average\_hike, 'is the average'),  marker = list(colors = colors,  line = list(color = '#FFFFFF', width = 1)),  showlegend = FALSE) %>%  layout(title = 'Average',  xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),  yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))  print(p)  Console Code :  install.packages("dplyr")  library(dplyr)  install.packages("plotly")  library(plotly)  source("piechart.R")  2)Which age group does more over-time  dataframe1 <- data.frame(Employee$OverTime,Employee$ï..Age)  mylist <- list(split(dataframe1,df\_1$Employee.OverTime))  list <- data.frame(sapply(mylist,'[','Yes'))  colors <- c("violetred","violetred1","violetred2","violetred3")  Age\_Range<-list$Yes.Employee.ï..Age  hist(Age\_Range,breaks=c(20,30,40,50,60),main="Age Groups Working Over Time)",col=colors)  1)Compare the performance rating and working hours for all departments and find whether department  with more working hours has highest performance?  > setwd("~/HEMASIRI")  > getwd()  [1] "C:/Users/srihi/OneDrive/Documents/HEMASIRI"  > Emp<-read.csv("Employee.csv", header = T, sep = ",")  > View(Emp)  > emp\_dataframe<-data.frame(Employee)  > department<-emp\_df$Department  > performance\_rating<-emp\_df$PerformanceRating  > TotalWorkingYears<-emp\_df$TotalWorkingYears  > Compare\_df<-data.frame(performance\_rating, TotalWorkingYears)  > install.packages("ggplot2")  > library(ggplot2)  > install.packages("reshape2")  > library(reshape2)  > cmp\_long<-melt(Compare\_df,id.vars = "department")  Colors<-”orange”,”yellow”,”green”  >ggplot(cmp\_long,aes(x=variable,y=value,fill=factor(department),col=”colors”))+geom\_bar  (stat="identity",position="dodge")+scale\_fill\_discrete(name="Department",breaks=c(1, 2, 3),  labels=c("Sales", "Human Resources", "Research & Development"))+xlab("Comparison Based  on Department")+ylab("Range") |