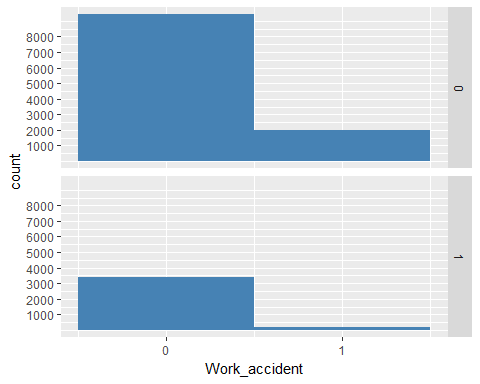
eda.R

Mrinal

Mon Jul 24 19:23:17 2017

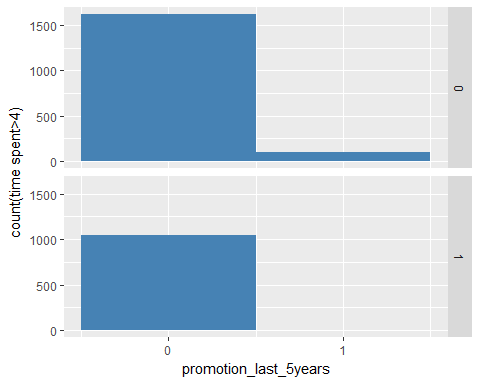
HR\_Data=read.csv("E:\\R\_Lectures\\dr. mohit\\PROJECT\\HR\_Data.csv")  
  
library(ggplot2)  
  
#Impact of work accident on people leaving their jobs.  
ggplot(HR\_Data,aes(Work\_accident))+geom\_histogram(fill="steelblue",binwidth = 1)+facet\_grid(left~.)+scale\_x\_continuous(breaks = c(0,1))+scale\_y\_continuous(breaks = c(1000,2000,3000,4000,5000,6000,7000,8000))



#From the second histogram we can quite clearly see that majority of people leaving didn't have a work accident. To take a decisive call of its impact on leaving we will find out the proportion of people who had work accident in subcategories made on the basis of whether they have left or not.  
#Proportion of people who had work accident in subcategories made on the basis of whether they have left or not.  
table(HR\_Data$Work\_accident,HR\_Data$left,dnn=c("Work\_accident","left"))

## left  
## Work\_accident 0 1  
## 0 9428 3402  
## 1 2000 169

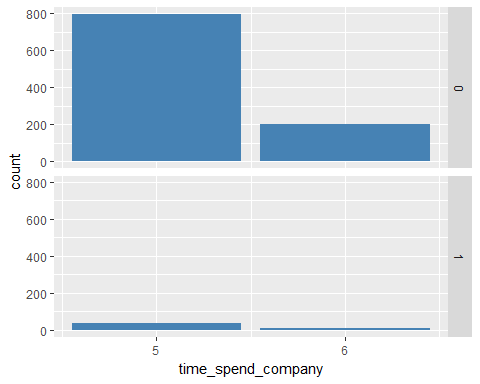
#From this table we can quite clearly see that proportion of people having work accident is significantly low in people who have left as compared to those who haven't. So it is quite clear that it is not as considerable a factor.  
  
#Impact of not getting promotion on people leaving their jobs. Now this has to be for those who have worked above significant time and according to the data provided this is maintained for whether someone has been promoted in last five yaers or not.  
ggplot(HR\_Data[which(HR\_Data$time\_spend\_company>=5),],aes(promotion\_last\_5years))+geom\_histogram(fill="steelblue",binwidth = 1)+facet\_grid(left~.)+scale\_x\_continuous(breaks = c(0,1))+ylab("count(time spent>4)")



#From the second hisogram it is quite clear that almost all of the senior employes who have left are those who weren't promoted in last 5 years.So now we will try to look at the proportion of people who were promoted in the subcategories made on the basis of whether they have left or not.  
#proportion of people who were promoted in the subcategories made on the basis of whether they have left or not  
data=HR\_Data[which(HR\_Data$time\_spend\_company>4),]  
table(data$promotion\_last\_5years,data$left,dnn=c("promotion","left"))

## left  
## promotion 0 1  
## 0 1618 1041  
## 1 95 1

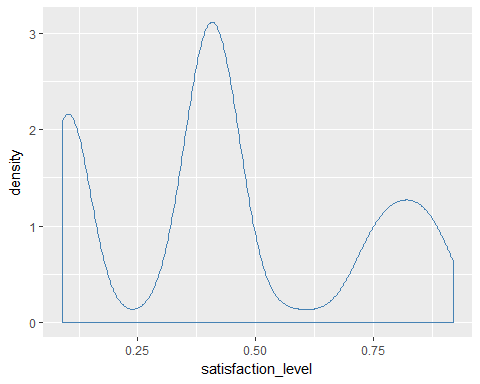
#We can quite clearly see that proportion of people who have been promoted is significantly low in people who have left as compared to who haven't. This shows that there is quite a possibility of it being a significant factor.  
  
#Combined impact of getting promotion and time spent in the company for those who have worked for more than 4 years.  
ggplot(HR\_Data[which(HR\_Data$left==1 & HR\_Data$time\_spend\_company>=5), ],aes(time\_spend\_company))+geom\_bar(fill="steelblue")+facet\_grid(Work\_accident~.)+scale\_x\_continuous(breaks = c(5,6,7,8,9,10))



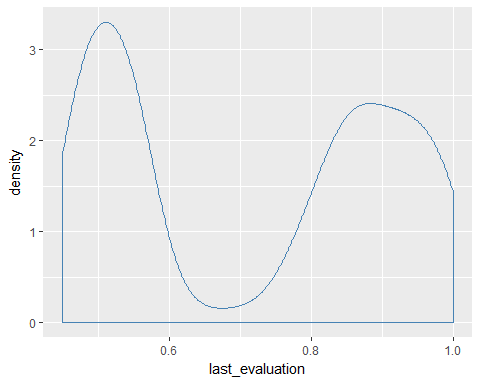
#We can see that the count of people who were not promoted and have left is less for those who have spent 7 years in company as compared to those who have spent 6 years and there after that is for those who have spent more than 7 years this count is zero. But to be more certain about this we will find out the proportion of people left without promotion for the categorise made on the basis of their time spent in the company.  
#proportion of people left without promotion for the categories made on the basis of their time spent in the company.  
data=HR\_Data[which(HR\_Data$left==1 & HR\_Data$time\_spend\_company>=5), ]  
table(data$promotion\_last\_5years, data$time\_spend\_compan,dnn=c("promotion","time\_spent"))

## time\_spent  
## promotion 5 6  
## 0 832 209  
## 1 1 0

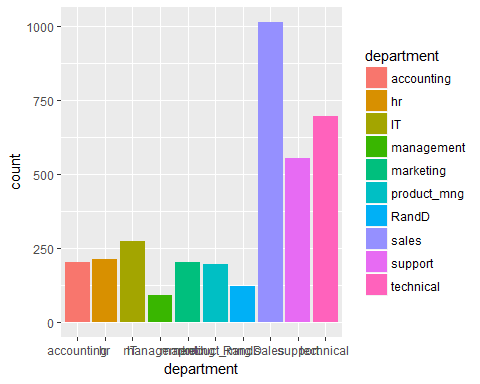
#we can pretty clearly see that almost all of the people who have have left after working 5 years or more are those which have not been promoted except one but still ther are people who have not been promoted and still working without promotion. So there combined effect can be significant while modelling.  
  
#Seprate density plots of people who have left and those who haven't left on the basis of their respective satisfaction level.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(satisfaction\_level))+geom\_density(color="steelblue")



#Seprate density plots of people who have left and those who haven't left on the basis of how they have been evaluated last time.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(last\_evaluation))+geom\_density(color="steelblue")



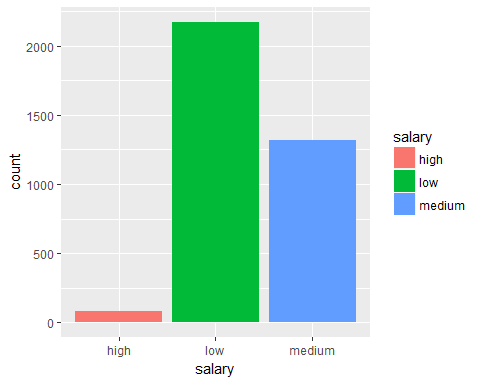
#Count of people leaving of different departments.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(department,fill=department))+geom\_bar()



#By this we can quite clearly see that amount of people leaving from sales department is max but we also have to look at the proportion of people leaving from each department.  
#proportion of people leaving.  
table(HR\_Data$department,HR\_Data$left,dnn=c('department','left'))

## left  
## department 0 1  
## accounting 563 204  
## hr 524 215  
## IT 954 273  
## management 539 91  
## marketing 655 203  
## product\_mng 704 198  
## RandD 666 121  
## sales 3126 1014  
## support 1674 555  
## technical 2023 697

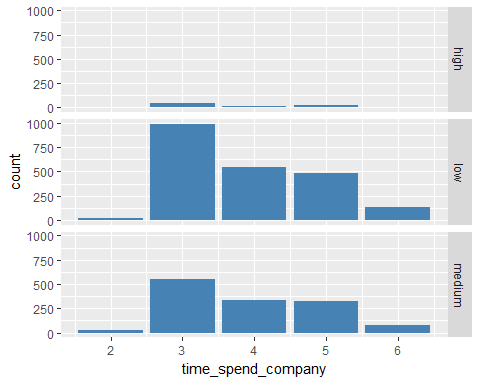
#This pretty well tells that department has the role to play as we can see that ratio of people left from each deparment to the people who didn't is varying.   
  
#count of people leaving with high, medium and low salary.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(salary,fill=salary))+geom\_bar()



#Clearly the salary category has the role to play since count of people leaving is highest in low salary category and is lowest in high salary category.  
#proportion of people leaving from each salary category.  
table(HR\_Data$salary,HR\_Data$left,dnn=c('salary','left'))

## left  
## salary 0 1  
## high 1155 82  
## low 5144 2172  
## medium 5129 1317

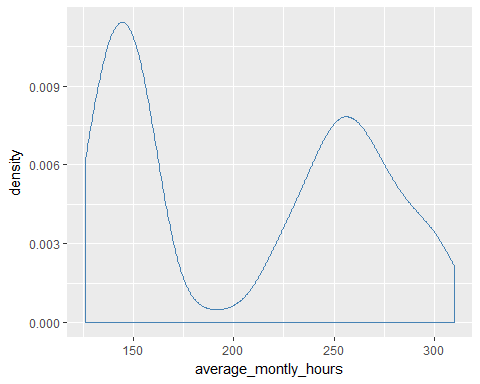
#Table pretty clearly tells that salary is significant as the proportion of people who have left with low salary is the hisghest and those left with the high salary is the lowest.  
  
#Impact of time spent in company and salary on the decision of people leaving the job.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(time\_spend\_company))+geom\_bar(fill="steelblue")+facet\_grid(salary~.)+scale\_x\_continuous(breaks = c(2,3,4,5,6,7,8,9,10))



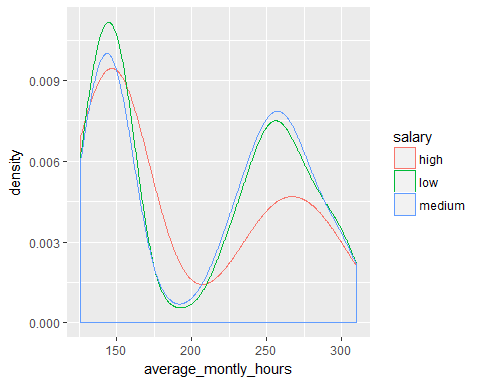
#proportion of people leaving those who have spent the same amount of time in the company.  
table(HR\_Data$salary,HR\_Data$left,HR\_Data$time\_spend\_company, dnn=c('salary','left','time spent'))

## , , time spent = 2  
##   
## left  
## salary 0 1  
## high 303 0  
## low 1505 22  
## medium 1383 31  
##   
## , , time spent = 3  
##   
## left  
## salary 0 1  
## high 474 46  
## low 2219 986  
## medium 2164 554  
##   
## , , time spent = 4  
##   
## left  
## salary 0 1  
## high 157 16  
## low 759 541  
## medium 751 333  
##   
## , , time spent = 5  
##   
## left  
## salary 0 1  
## high 46 20  
## low 311 488  
## medium 283 325  
##   
## , , time spent = 6  
##   
## left  
## salary 0 1  
## high 55 0  
## low 198 135  
## medium 256 74  
##   
## , , time spent = 7  
##   
## left  
## salary 0 1  
## high 38 0  
## low 36 0  
## medium 114 0  
##   
## , , time spent = 8  
##   
## left  
## salary 0 1  
## high 18 0  
## low 60 0  
## medium 84 0  
##   
## , , time spent = 10  
##   
## left  
## salary 0 1  
## high 64 0  
## low 56 0  
## medium 94 0

#These two factors have the combined effect as we can quite clearly see that the ratio of people leaving to that of who have stayed for almost all categories of salary is increasing with inccrease in the number of years spend in the company and there after the count of people leaving is 0 as no one has left the company after working for 6 years.  
  
#Density plot of people leaving on the basis of their average monthly working hours.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(average\_montly\_hours))+geom\_density(color="steelblue")



#Seprate density plots of people of different salary category leaving on the basis of their average monthly working hours.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(average\_montly\_hours,color=salary))+geom\_density()



#The count of people leaving in each salary category follows the similar trend so we will check the proportion of people leaving in each of the subcategories made on the basis of average monthly hours:  
 #average\_monthly\_hours<175,175<average\_monthly\_hours<225,225<average\_monthly\_hours  
#proportion of people leaving in each of the subcategories.  
data1=HR\_Data[which(HR\_Data$average\_montly\_hours<=175),]  
table(data1$left)

##   
## 0 1   
## 3962 1614

data2=HR\_Data[which(HR\_Data$average\_montly\_hours>175 & HR\_Data$average\_montly\_hours>225),]  
table(data2$left)

##   
## 0 1   
## 3735 1767

data3=HR\_Data[which(HR\_Data$average\_montly\_hours>=225),]  
table(data3$left)

##   
## 0 1   
## 3817 1780

#We can roughly say that the proportion of people leaving is continuously decreasing with increse in average monthly working hours. So this can be significant factor in modelling.  
#combined effect of salary and average monthly working hours.  
data1=HR\_Data[which(HR\_Data$average\_montly\_hours<=175),]  
table(data1$left,data1$salary)

##   
## high low medium  
## 0 385 1784 1793  
## 1 50 1003 561

data2=HR\_Data[which(HR\_Data$average\_montly\_hours>175 & HR\_Data$average\_montly\_hours>225),]  
table(data2$left,data2$salary)

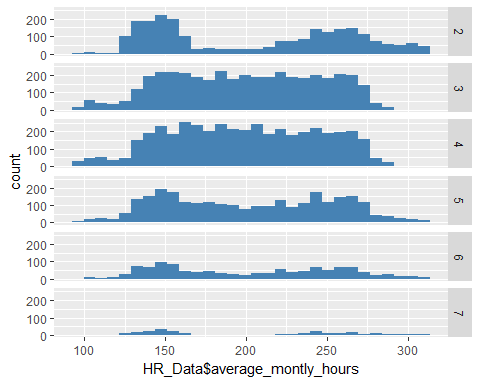
##   
## high low medium  
## 0 393 1632 1710  
## 1 31 1058 678

data3=HR\_Data[which(HR\_Data$average\_montly\_hours>=225),]  
table(data3$left,data3$salary)

##   
## high low medium  
## 0 406 1668 1743  
## 1 31 1069 680

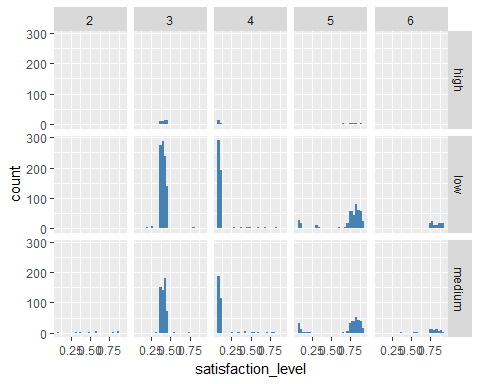
#We can quite clearly see that proportion of people leaving in each category of average monthly working hours is varying in different salary categories. So these can have the combined effect in modelling.  
  
#Seprate histogrmas for people with different number of projects on the basis of their average monthly working hours.  
ggplot(HR\_Data,aes(HR\_Data$average\_montly\_hours))+geom\_histogram(fill="steelblue")+facet\_grid(number\_project~.)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



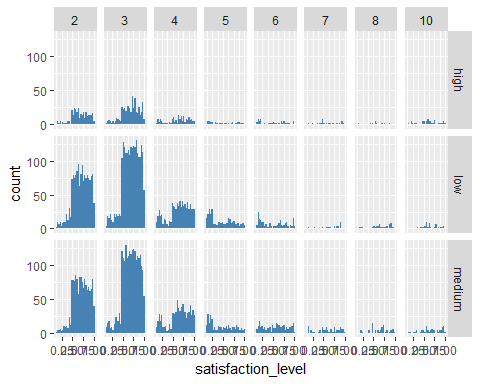
#Histograms quite clearly show that there is no corelation between average monthly working hours and number of projects.  
  
#Plotting seprate histograms on the basis of satisfaction level for each salary category, time spent in the company & whtehter they have left or not.  
ggplot(HR\_Data[which(HR\_Data$left==1),],aes(satisfaction\_level))+geom\_histogram(fill="steelblue")+facet\_grid(salary~time\_spend\_company)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



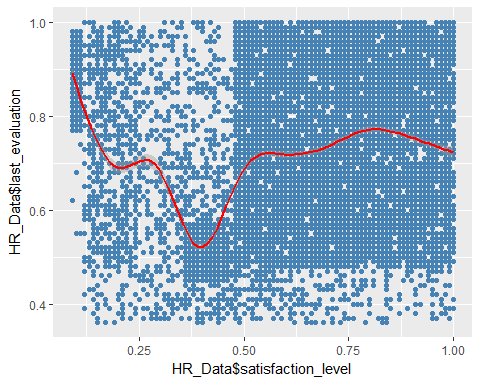
ggplot(HR\_Data[which(HR\_Data$left==0),],aes(satisfaction\_level))+geom\_histogram(fill="steelblue")+facet\_grid(salary~time\_spend\_company)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#We can quite clearly see that initially for first 2 years satisfacation level is not that significant in each of the salary category. But for those who have spend 3 & 4 years we can quite clearly see that count of poeple leving with satisfaction level less than 0.5 is more than those who have left this difference in count is comparitively less in high salary category. And for those who have spent 5 & 6 years count of people leaving with high satisfaction level is more in medium and low salary category as compared to those who have stayed. No one has left the company after working more than 6 years and in high salary category no one has left after working more than 5 years. This shows that there can be interaction between these three terms which can be significant while modelling.   
  
#Scatter plot between last evaluation and satisfaction level.  
ggplot(HR\_Data,aes(HR\_Data$satisfaction\_level,HR\_Data$last\_evaluation))+geom\_point(color="steelblue")+geom\_smooth(color="red")

## `geom\_smooth()` using method = 'gam'



#Plot shows no clear cut trend.Hencce, no corelation between the used two attributes.