ABSTRACT

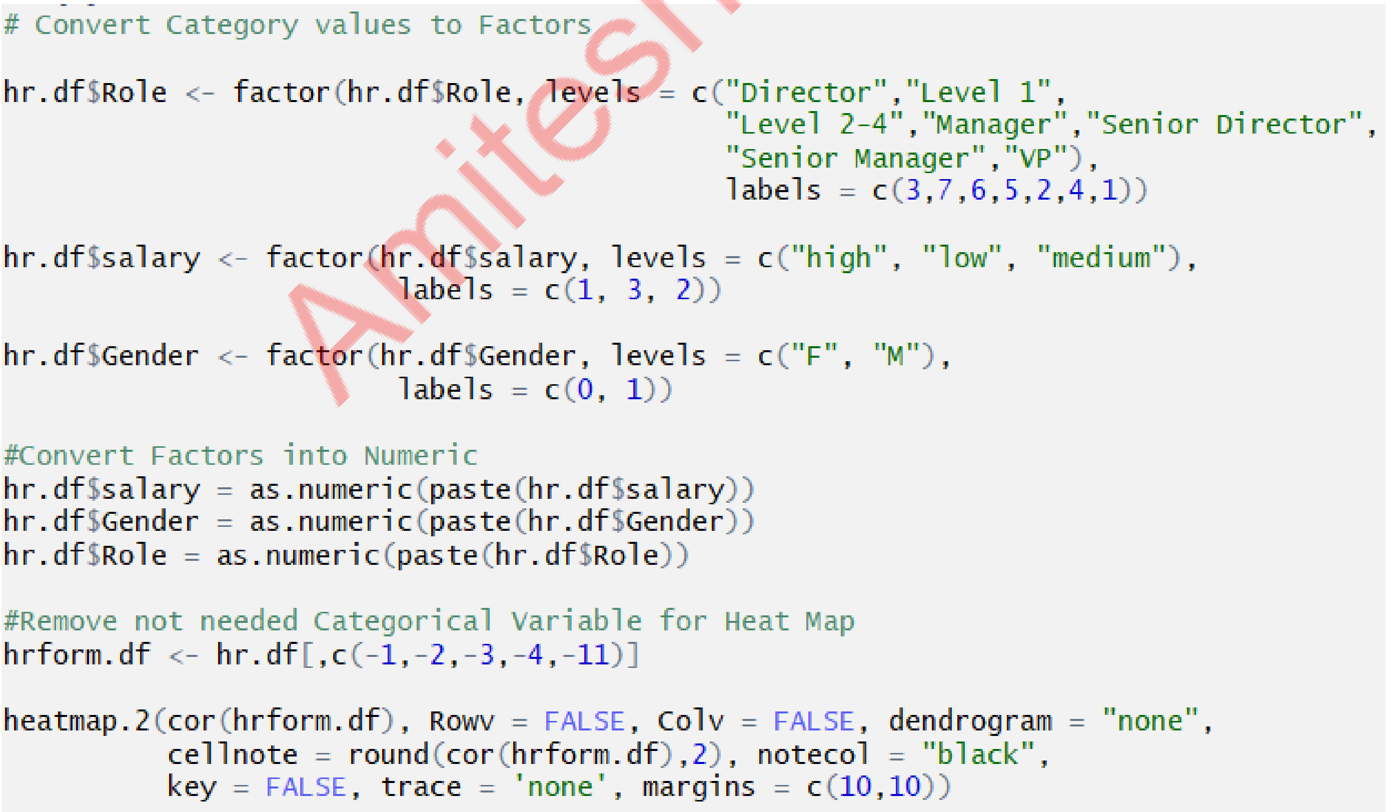
Data set

Our data set represents 14,999 employees and is composed of both currently employed and people who have already left the company with 30 variables defining the best possible way to answer the below questions and insights.

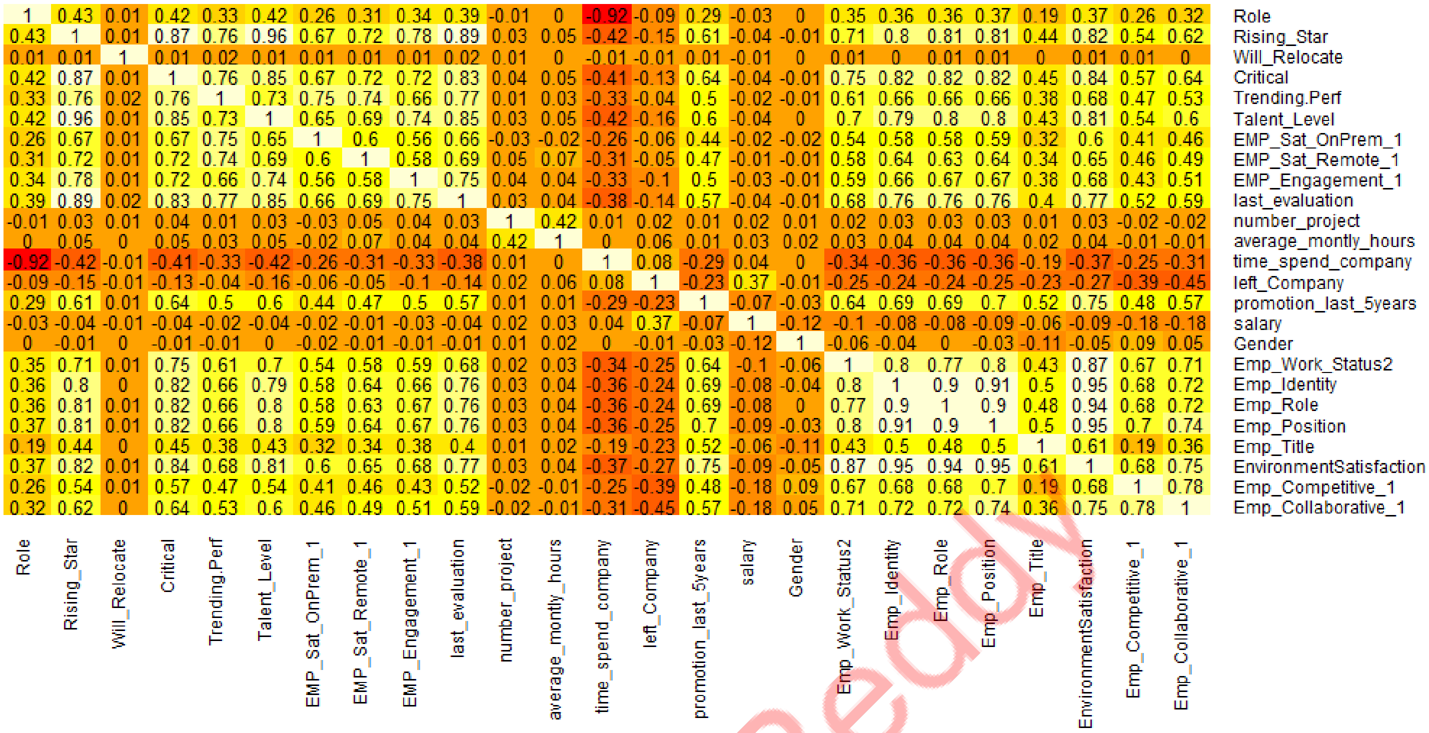
Initially after loading the dataset, we saw 20+ variables that had no significance for any of our analysis model and hence we decided to discard them. It is always recommended to run some basic checks and see if there are missing values or any unusual patterns amongst other things (in most data sets Kaggle gives you clean data). Right from the very first correlation that we ran, we were clear about incorporating few changes to the dataset. We compared the Kaggle dataset with the IBM HR analytics dataset and included a field called Employee_satisfaction from the latter and merged it with the existing file to create a new variable with the same name, representing an average of four other parameters from the file.

To improve the correlation significance between various predictors, we made changes against few variables. (Role, Rising_Star, Left_Company, promotion_last_5years, Salary, Emp_Satisfaction)

Correlation matric before and after making changes to the dataset



```
# Convert Category values to Factors

hr.df$Role <- factor(hr.df$Role, levels = c("Director","Level 1",
                                            "Level 2-4","Manager","Senior Director",
                                            "Senior Manager","VP"),
                                            labels = c(3,7,6,5,2,4,1))

hr.df$salary <- factor(hr.df$salary, levels = c("high", "low", "medium"),
                       labels = c(1, 3, 2))

hr.df$Gender <- factor(hr.df$Gender, levels = c("F", "M"),
                       labels = c(0, 1))

#Convert Factors into Numeric
hr.df$salary = as.numeric(paste(hr.df$salary))
hr.df$Gender = as.numeric(paste(hr.df$Gender))
hr.df$Role = as.numeric(paste(hr.df$Role))

#Remove not needed Categorical Variable for Heat Map
hrform.df <- hr.df[,c(-1,-2,-3,-4,-11)]

heatmap.2(cor(hrform.df), Rowv = FALSE, Colv = FALSE, dendrogram = "none",
          cellnote = round(cor(hrform.df),2), notecol = "black",
          key = FALSE, trace = 'none', margins = c(10,10))
```



	Role	Rising_Star	Will_Relocate	Critical	Trending.Perf	Talent_Level	EMP_Sat_OnPrem_1	EMP_Sat_Remote_1	EMP_Engagement_1	last_evaluation	number_project	average_montly_hours	time_spend_company	left_Company	promotion_last_5years	salary	Gender	Emp_Work_Status2	Emp_Identity	Emp_Role	Emp_Position	Emp_Title	EnvironmentSatisfaction	Emp_Competitive_1	Emp_Collaborative_1
Role	1	0.43	0.01	0.42	0.33	0.42	0.26	0.31	0.34	0.39	-0.01	0	-0.92	-0.09	0.29	-0.03	0	0.35	0.36	0.36	0.37	0.19	0.37	0.26	0.32
Rising_Star	0.43	1	0.01	0.87	0.76	0.96	0.67	0.72	0.78	0.89	0.03	0.05	-0.42	-0.15	0.61	-0.04	-0.01	0.71	0.8	0.81	0.81	0.44	0.82	0.54	0.62
Will_Relocate	0.01	0.01	1	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0	-0.01	-0.01	0.01	-0.01	0	0.01	0	0.01	0.01	0	0.01	0.01	0
Critical	0.42	0.87	0.01	1	0.76	0.85	0.67	0.72	0.72	0.83	0.04	0.05	-0.41	-0.13	0.64	-0.04	-0.01	0.75	0.82	0.82	0.82	0.45	0.84	0.57	0.64
Trending.Perf	0.33	0.76	0.02	0.76	1	0.73	0.75	0.74	0.66	0.77	0.01	0.03	-0.33	-0.04	0.5	-0.02	-0.01	0.61	0.66	0.66	0.66	0.38	0.68	0.47	0.53
Talent_Level	0.42	0.96	0.01	0.85	0.73	1	0.65	0.69	0.74	0.85	0.03	0.05	-0.42	-0.16	0.6	-0.04	0	0.7	0.79	0.8	0.8	0.43	0.81	0.54	0.6
EMP_Sat_OnPrem_1	0.26	0.67	0.01	0.67	0.75	0.65	1	0.6	0.56	0.66	-0.03	-0.02	-0.26	-0.06	0.44	-0.02	-0.02	0.54	0.58	0.58	0.59	0.32	0.6	0.41	0.46
EMP_Sat_Remote_1	0.31	0.72	0.01	0.72	0.74	0.69	0.6	1	0.58	0.69	0.05	0.07	-0.31	-0.05	0.47	-0.01	-0.01	0.58	0.64	0.63	0.64	0.34	0.65	0.46	0.49
EMP_Engagement_1	0.34	0.78	0.01	0.72	0.66	0.74	0.56	0.58	1	0.75	0.04	0.04	-0.33	-0.1	0.5	-0.03	-0.01	0.59	0.66	0.67	0.67	0.38	0.68	0.43	0.51
last_evaluation	0.39	0.89	0.02	0.83	0.77	0.85	0.66	0.69	0.75	1	0.03	0.04	-0.38	-0.14	0.57	-0.04	-0.01	0.68	0.76	0.76	0.76	0.4	0.77	0.52	0.59
number_project	-0.01	0.03	0.01	0.04	0.01	0.03	-0.03	0.05	0.04	0.03	1	0.42	0.01	0.02	0.01	0.02	0.01	0.02	0.03	0.03	0.03	0.01	0.03	-0.02	-0.02
average_montly_hours	0	0.05	0	0.05	0.03	0.05	-0.02	0.07	0.04	0.04	0.42	1	0	0.06	0.01	0.03	0.02	0.03	0.04	0.04	0.04	0.02	0.04	-0.01	-0.01
time_spend_company	-0.92	-0.42	-0.01	-0.41	-0.33	-0.42	-0.26	-0.31	-0.33	-0.38	0.01	0	1	0.08	-0.29	0.04	0	-0.34	-0.36	-0.36	-0.36	-0.19	-0.37	-0.25	-0.31
left_Company	-0.09	-0.15	-0.01	-0.13	-0.04	-0.16	-0.06	-0.05	-0.1	-0.14	0.02	0.06	0.08	1	-0.23	0.37	-0.01	-0.25	-0.24	-0.24	-0.25	-0.23	-0.27	-0.39	-0.45
promotion_last_5years	0.29	0.61	0.01	0.64	0.5	0.6	0.44	0.47	0.5	0.57	0.01	0.01	-0.29	-0.23	1	-0.07	-0.03	0.64	0.69	0.69	0.7	0.52	0.75	0.48	0.57
salary	-0.03	-0.04	-0.01	-0.04	-0.02	-0.04	-0.02	-0.01	-0.03	-0.04	0.02	0.03	0.04	0.37	-0.07	1	-0.12	-0.1	-0.08	-0.08	-0.09	-0.06	-0.09	-0.18	-0.18
Gender	0	-0.01	0	-0.01	-0.01	0	-0.02	-0.01	-0.01	-0.01	0.01	0.02	0	-0.01	-0.03	-0.12	1	-0.06	-0.04	0	-0.03	-0.11	-0.05	0.09	0.05
Emp_Work_Status2	0.35	0.71	0.01	0.75	0.61	0.7	0.54	0.58	0.59	0.68	0.02	0.03	-0.34	-0.25	0.64	-0.1	-0.06	1	0.8	0.77	0.8	0.43	0.87	0.67	0.71
Emp_Identity	0.36	0.8	0	0.82	0.66	0.79	0.58	0.64	0.66	0.76	0.03	0.04	-0.36	-0.24	0.69	-0.08	-0.04	0.8	1	0.9	0.91	0.5	0.95	0.68	0.72
Emp_Role	0.36	0.81	0.01	0.82	0.66	0.8	0.58	0.63	0.67	0.76	0.03	0.04	-0.36	-0.24	0.69	-0.08	0	0.77	0.9	1	0.9	0.48	0.94	0.68	0.72
Emp_Position	0.37	0.81	0.01	0.82	0.66	0.8	0.59	0.64	0.67	0.76	0.03	0.04	-0.36	-0.25	0.7	-0.09	-0.03	0.8	0.91	0.9	1	0.5	0.95	0.7	0.74
Emp_Title	0.19	0.44	0	0.45	0.38	0.43	0.32	0.34	0.38	0.4	0.01	0.02	-0.19	-0.23	0.52	-0.06	-0.11	0.43	0.5	0.48	0.5	1	0.61	0.19	0.36
EnvironmentSatisfaction	0.37	0.82	0.01	0.84	0.68	0.81	0.6	0.65	0.68	0.77	0.03	0.04	-0.37	-0.27	0.75	-0.09	-0.05	0.87	0.95	0.94	0.95	0.61	1	0.68	0.75
Emp_Competitive_1	0.26	0.54	0.01	0.57	0.47	0.54	0.41	0.46	0.43	0.52	-0.02	-0.01	-0.25	-0.39	0.48	-0.18	0.09	0.67	0.68	0.68	0.7	0.19	0.68	1	0.78
Emp_Collaborative_1	0.32	0.62	0	0.64	0.53	0.6	0.46	0.49	0.51	0.59	-0.02	-0.01	-0.31	-0.45	0.57	-0.18	0.05	0.71	0.72	0.72	0.74	0.36	0.75	0.78	1

OBJECTIVES

The main objectives that we had set out before working on the dataset were :

- Identify the primary reasons for employees leaving both low and high performance
- Why do good employees leave?
- Likelihood of a promotion
- What factors increase job satisfaction
- Relationship between time_spend_company and other variables
- Which employee will leave next?

DATA EXPLORATION

Read the HR Dataset

```
hr.df <- read.csv("HR.csv", header = TRUE)
```

Dataset Details

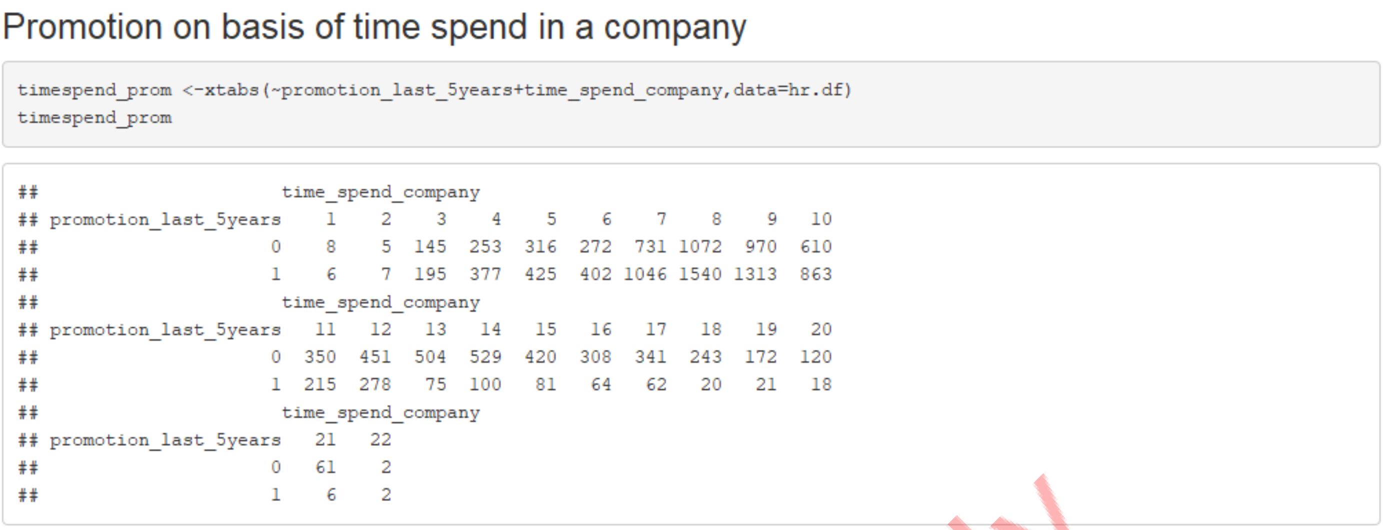
```
dim(hr.df)
```

```
## [1] 14999    30
```

Describe Dataset

```
summary(hr.df)
```

```
##        ID              Name                  Department        GEO
##  Min.   :    1    AARON   :     1    Finance         :1983   UK       :1772
##  1st Qu.: 3750    ABAD    :     1    Human Resources:1785   France   :1699
##  Median : 7500    ABADIE  :     1    IT              :3485   Korea    :1685
##  Mean   : 7500    ABARCA  :     1    Operations      :2500   Japan    :1669
##  3rd Qu.:11250    ABATE   :     1    Sales           :2500   China    :1667
##  Max.   :14999    (Other):14993    Support         : 247   Colombia:1659
##                   NA's    :     1    Warehouse       :2499   (Other) :4848
##                Role        Rising_Star     Will_Relocate       Critical
##  Director         : 660   Min.   :1.000   Min.   :0.0000   Min.   :0.000
##  Level 1          :3270   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.000
##  Level 2-4        :6889   Median :4.000   Median :0.0000   Median :1.000
##  Manager          :2420   Mean   :3.511   Mean   :0.4998   Mean   :0.682
##  Senior Director: 330   3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Senior Manager :1326   Max.   :5.000   Max.   :1.0000   Max.   :1.000
##  VP               : 104
##  Trending.Perf     Talent_Level     Percent_Remote    EMP_Sat_OnPrem_1
##  Min.   : 1.000   Min.   : 1.000   Min.   :0.4000   Min.   : 0.000
##  1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.:0.4000   1st Qu.: 5.000
##  Median : 8.000   Median : 7.000   Median :0.8000   Median : 7.000
##  Mean   : 7.171   Mean   : 6.451   Mean   :0.6173   Mean   : 6.615
##  3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.:0.8000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :10.000   Max.   :1.0000   Max.   :10.000
##
##  EMP_Sat_Remote_1 EMP_Engagement_1 last_evaluation  number_project
##  Min.   : 1.000   Min.   :1.000    Min.   : 3.000   Min.   :2.000
##  1st Qu.: 6.000   1st Qu.:2.000    1st Qu.: 5.000   1st Qu.:3.000
##  Median : 8.000   Median :3.000    Median : 7.000   Median :4.000
##  Mean   : 7.273   Mean   :2.997    Mean   : 7.017   Mean   :3.803
##  3rd Qu.: 9.000   3rd Qu.:4.000    3rd Qu.: 9.000   3rd Qu.:5.000
##  Max.   :10.000   Max.   :5.000    Max.   :10.000   Max.   :7.000
##
##  average_montly_hours time_spend_company  left_Company
##  Min.   : 40          Min.   : 1.000     Min.   :0.0000
##  1st Qu.:156          1st Qu.: 7.000     1st Qu.:0.0000
##  Median :200          Median : 9.000     Median :0.0000
##  Mean   :201          Mean   : 9.616     Mean   :0.3062  
##  3rd Qu.:245          3rd Qu.:12.000     3rd Qu.:1.0000  
##  Max.   :310          Max.   :22.000     Max.   :1.0000  
##                                                          
##  promotion_last_5years    salary     Gender   Emp_Work_Status2
##  Min.   :0.0000        high  :1668   F:7596   Min.   : 1.00   
##  1st Qu.:0.0000        low   :6857   M:7403   1st Qu.: 4.00   
##  Median :0.0000        medium:6474            Median : 7.00   
##  Mean   :0.4744                               Mean   : 6.41   
##  3rd Qu.:1.0000                               3rd Qu.: 9.00   
##  Max.   :1.0000                               Max.   :10.00   
##                                                               
##   Emp_Identity       Emp_Role       Emp_Position      Emp_Title     
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000  
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000  
##  Median : 7.000   Median : 7.000   Median : 7.000   Median : 3.000  
##  Mean   : 6.143   Mean   : 6.143   Mean   : 6.067   Mean   : 3.287  
##  3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 5.000  
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000  
##                                                                     
##  Emp_Satisfaction Emp_Competitive_1 Emp_Collaborative_1
##  Min.   : 1.000   Min.   : 1.000    Min.   : 1.000     
##  1st Qu.: 3.000   1st Qu.: 2.000    1st Qu.: 3.000     
##  Median : 7.000   Median : 6.000    Median : 7.000     
##  Mean   : 5.608   Mean   : 4.998    Mean   : 5.938     
##  3rd Qu.: 8.000   3rd Qu.: 8.000    3rd Qu.: 9.000     
##  Max.   :10.000   Max.   :10.000    Max.   :10.000
```

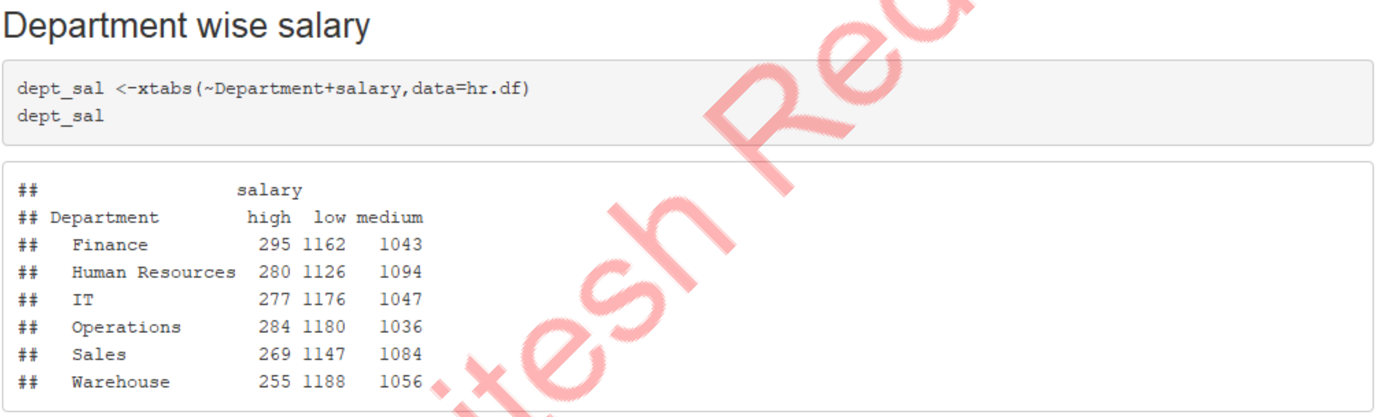


Promotion on basis of time spend in a company

```
timespend_prom <-xtabs(~promotion_last_5years+time_spend_company,data=hr.df)
timespend_prom
```

```
##                       time_spend_company
## promotion_last_5years    1    2    3    4    5    6    7    8    9   10
##                     0    8    5  145  253  316  272  731 1072  970  610
##                     1    6    7  195  377  425  402 1046 1540 1313  863
##                       time_spend_company
## promotion_last_5years   11   12   13   14   15   16   17   18   19   20
##                     0  350  451  504  529  420  308  341  243  172  120
##                     1  215  278   75  100   81   64   62   20   21   18
##                       time_spend_company
## promotion_last_5years   21   22
##                     0   61    2
##                     1    6    2
```

Employees who have been in the company for 7-9 years have been awarded the most number of promotions in the last 5 years and as the number of years spent at the company increases, the number of promotions decreases.

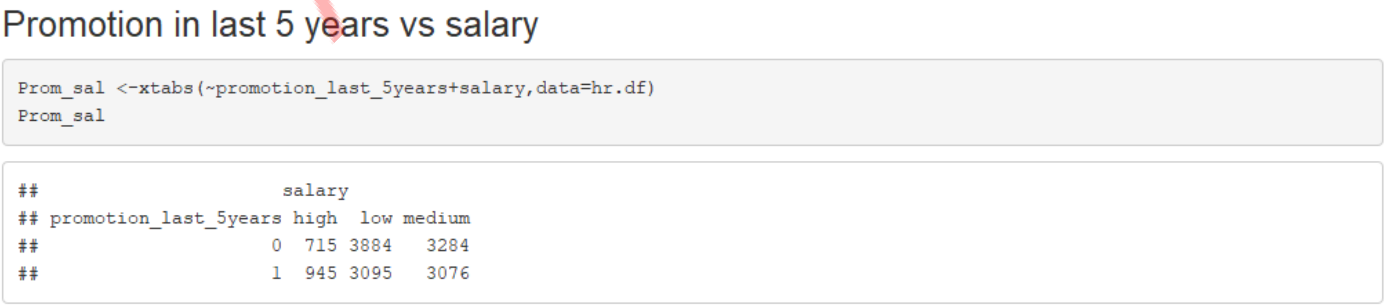


Department wise salary

```
dept_sal <-xtabs(~Department+salary,data=hr.df)
dept_sal
```

```
##                  salary
## Department        high  low medium
##   Finance          295 1162   1043
##   Human Resources  280 1126   1094
##   IT               277 1176   1047
##   Operations       284 1180   1036
##   Sales            269 1147   1084
##   Warehouse        255 1188   1056
```

The finance department has the highest number of high-wage workers whereas the warehouse department has the highest number of low-wage workers.



Promotion in last 5 years vs salary

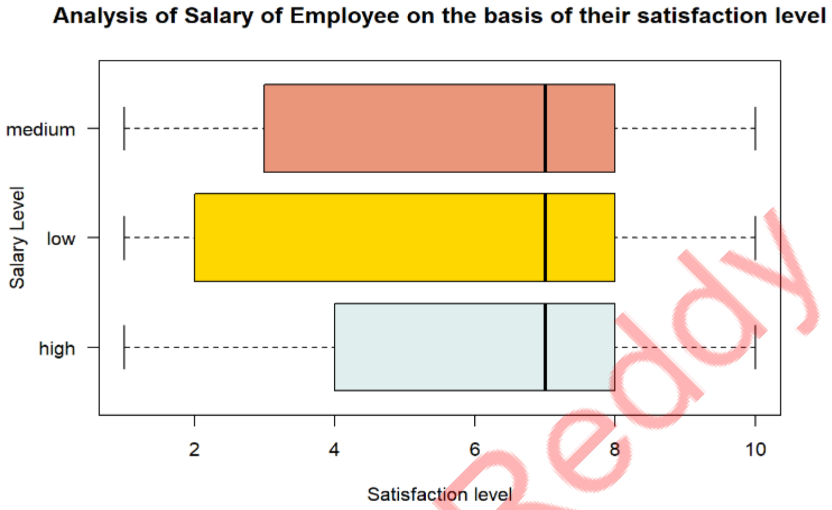
```
Prom_sal <-xtabs(~promotion_last_5years+salary,data=hr.df)
Prom_sal
```

```
##                       salary
## promotion_last_5years high  low medium
##                     0  715 3884   3284
##                     1  945 3095   3076
```

Employees getting the maximum promotions in the last 5 years have had a low to medium increase in their salary, with very few of them promoted with a high wage

Box Plot describing relationship between Salary and Emp_Satisfaction

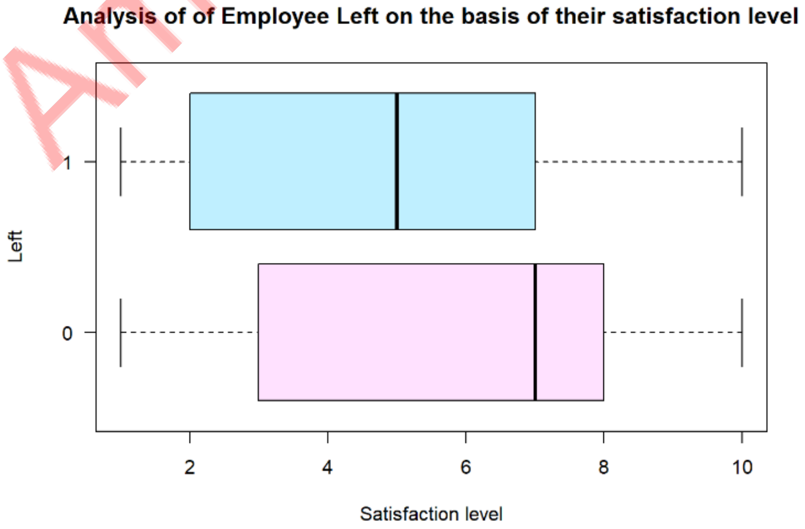
```
boxplot(Emp_Satisfaction ~salary,data=hr.df, horizontal=TRUE,
          ylab="Salary Level", xlab="Satisfaction level", las=1,
          main="Analysis of Salary of Employee on the basis of their satisfaction level",
          col=c("azure2","gold","darksalmon")
          )
```



Employees in the higher wage category have more satisfaction levels than lower wage level employees.

Box Plot describing relationship between Left_company and Emp_Satisfaction

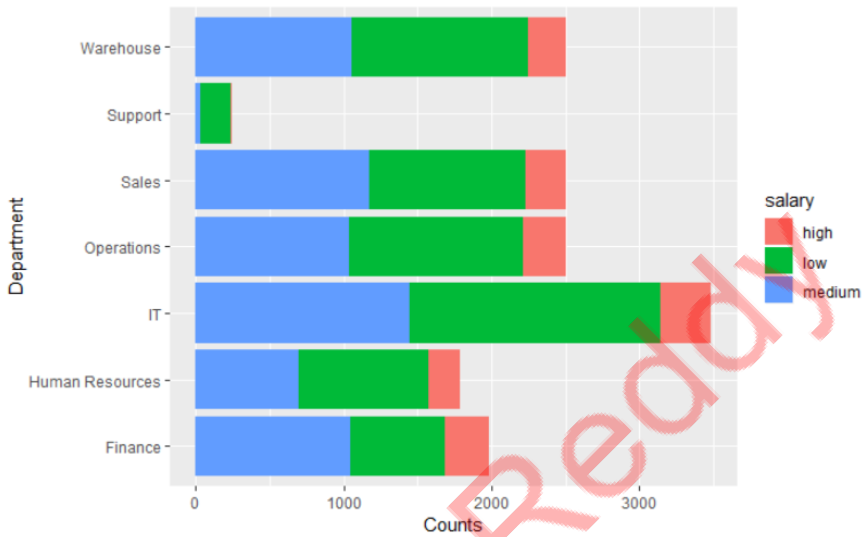
```
boxplot(Emp_Satisfaction ~left_Company, data=hr.df, horizontal=TRUE,
          ylab="Left", xlab="Satisfaction level", las=1,
          main="Analysis of of Employee Left on the basis of their satisfaction level",
          col=c("thistle1","lightblue1")
          )
```



As it can be seen , employees with lower satisfaction levels tend to leave the company.

Barplot to ascertain the salaries of employees by their department using GGPLOT

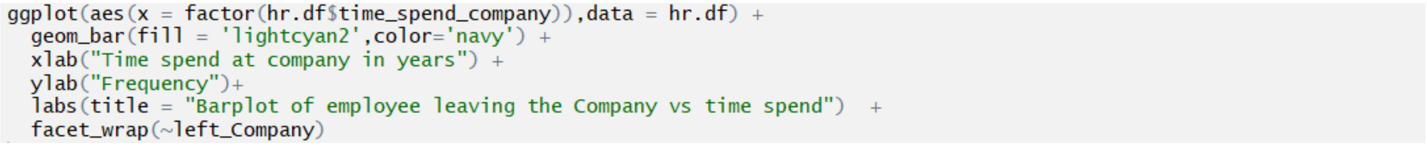
```
ggplot(aes(x = Department),data = hr.df ) +
  geom_bar(aes(fill = salary))  +
  xlab('Department') +
  ylab('Counts') +
  coord_flip()
```



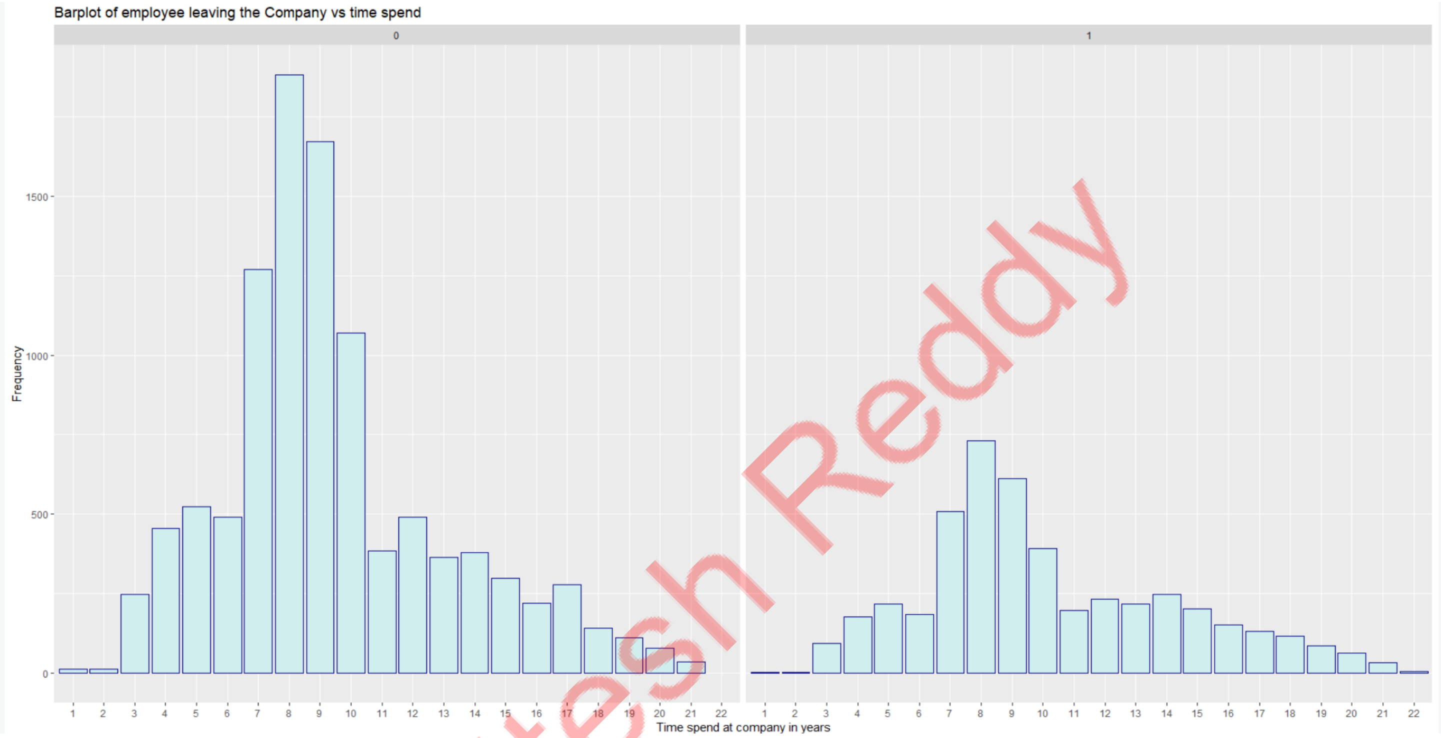
Interpretation

- IT department, having the maximum employees working in shows considerable variability in term of salary distribution.
- Sales, Operation and Warehouse departments have a similar trend in terms of salary distribution.
- Support dept, having the least count of employees working in have majority of the employees in the low salary bracket giving us more insights about potentially being the crowd about to leave the company or not performing well.

Barplot of employees leaving/not-leaving the company vs time spend using GGPLOT



```
ggplot(aes(x = factor(hr.df$time_spend_company)),data = hr.df) +
  geom_bar(fill = 'lightcyan2',color='navy') +
  xlab("Time spend at company in years") +
  ylab("Frequency")+
  labs(title = "Barplot of employee leaving the Company vs time spend")  +
  facet_wrap(~left_Company)
```

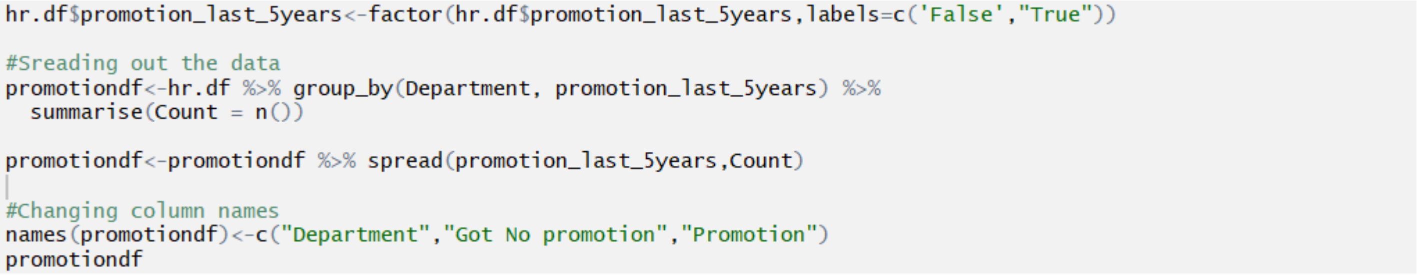


Interpretation

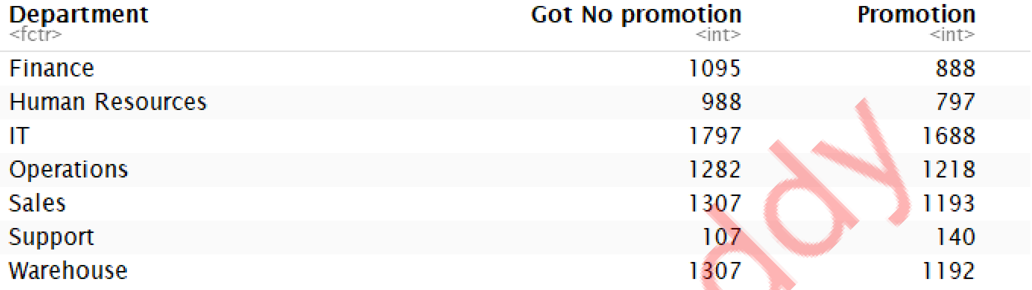
- From the second plot above that represents the employees having left the company, it is evident that employees tend to leave a company after spending 7-10 years with average being 8 years
- Very less number of employees leave the company within the first 2 years of joining
- There are employees who after spending 11-15 years leave the company, something we will figure out in the next chart

- From the first plot, we see majority of current employees have spent 7-10 years in the company with tough fight between employees having spent 8 years. This bracket might have intense competition in terms of promotion and salary as there are more employees
- Very few employees are in the 20-22 years category that says they belong to the higher bands within the company
- Company might have reduced its recruiting in the past 2 years as shown above with less number of employees having spent 2 years

Table showing department wise promotion

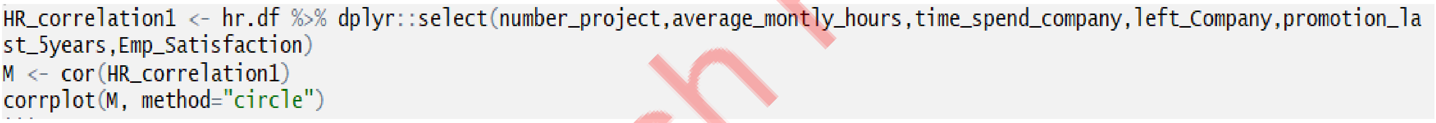


```
hr.df$promotion_last_5years<-factor(hr.df$promotion_last_5years,labels=c('False',"True"))

#Sreading out the data
promotiondf<-hr.df %>% group_by(Department, promotion_last_5years) %>%
  summarise(Count = n())

promotiondf<-promotiondf %>% spread(promotion_last_5years,Count)

#Changing column names
names(promotiondf)<-c("Department","Got No promotion","Promotion")
promotiondf
```



Department <fctr>	Got No promotion <int>	Promotion <int>
Finance	1095	888
Human Resources	988	797
IT	1797	1688
Operations	1282	1218
Sales	1307	1193
Support	107	140
Warehouse	1307	1192

Correlation showing the important factors on which employee satisfaction depends on :



```
HR_correlation1 <- hr.df %>% dplyr::select(number_project,average_montly_hours,time_spend_company,left_Company,promotion_la
st_5years,Emp_Satisfaction)
M <- cor(HR_correlation1)
corrplot(M, method="circle")
```

Interpretation

Employee_Satisfaction has a very positive correlation with promotion_received in last 5 years which directly gives us more insights for such employees to stay longer in a company.

Also , the satisfaction levels depend on Emp_Collaborative_1 which describes how collaborative an employee thinks his coworkers are. If an employee has a good relationship with their coworkers , then their satisfaction levels are also high.

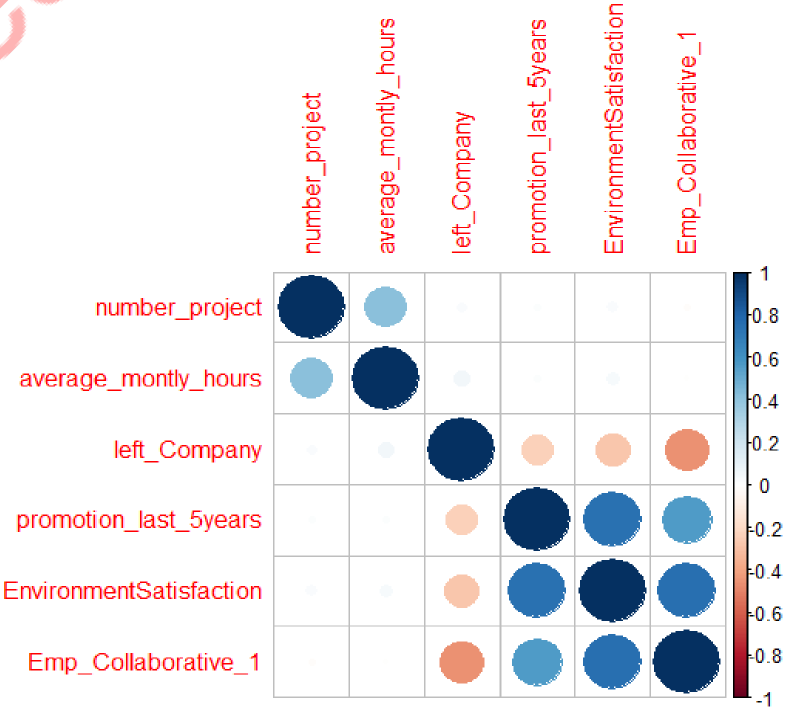
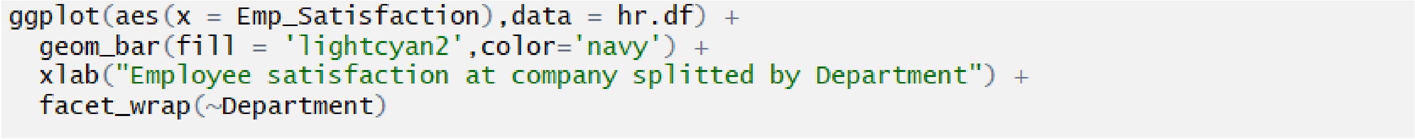
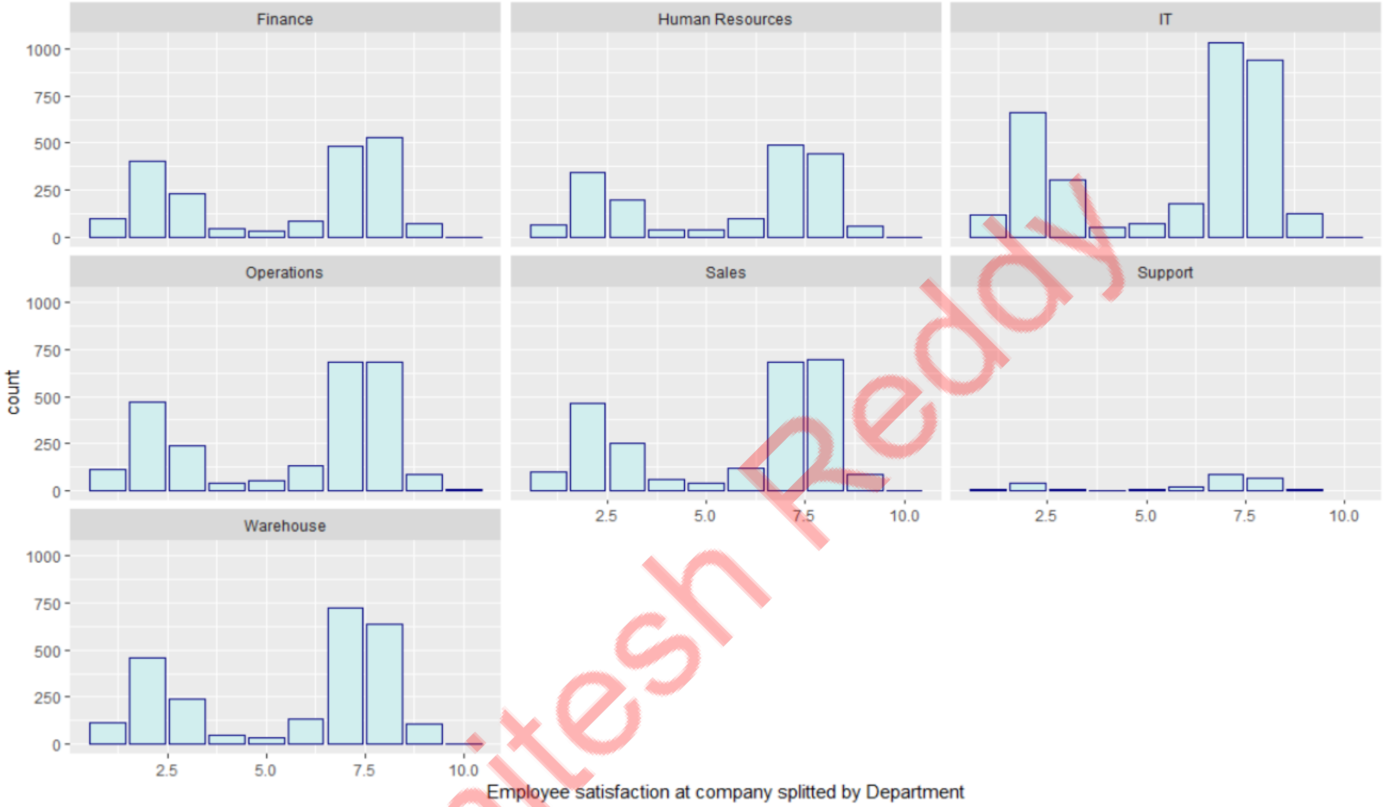


Table showing department wise Employee_Satisfaction



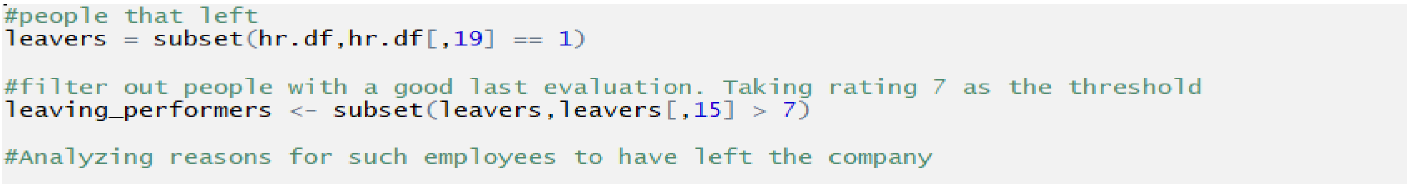
```
ggplot(aes(x = Emp_Satisfaction),data = hr.df) +
  geom_bar(fill = 'lightcyan2',color='navy') +
  xlab("Employee satisfaction at company splitted by Department") +
  facet_wrap(~Department)
```



Interpretation

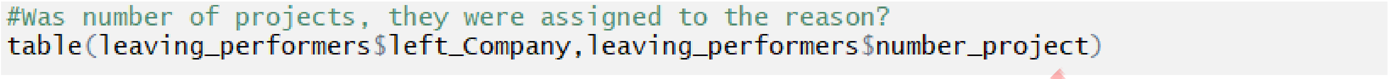
- IT department has got the most number of employees falling in both the categories(Satisfied and not satisfied) giving us takeaway that a high number of employees aren't happy with their work.
- We see a bimodal barplot for across departments telling us that employees are either not satisfied; with average between 2-4 and employees satisfied with average being 7-8.
- Very less employees are highly satisfied across the departments.

WHY GOOD EMPLOYEES LEAVE?

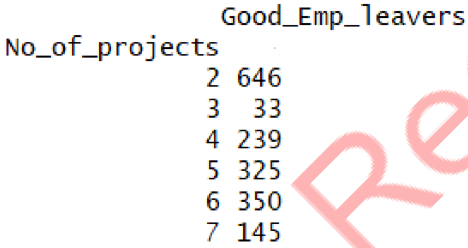


```
#people that left
leavers = subset(hr.df,hr.df[,19] == 1)

#filter out people with a good last evaluation. Taking rating 7 as the threshold
leaving_performers <- subset(leavers,leavers[,15] > 7)

#Analyzing reasons for such employees to have left the company
```

Are the number of projects employees assigned to the reason?



```
#Was number of projects, they were assigned to the reason?
table(leaving_performers$left_Company,leaving_performers$number_project)
```

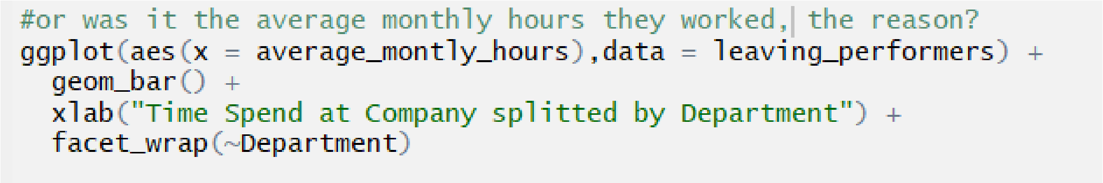


```
               Good_Emp_leavers
No_of_projects
             2 646
             3  33
             4 239
             5 325
             6 350
             7 145
```

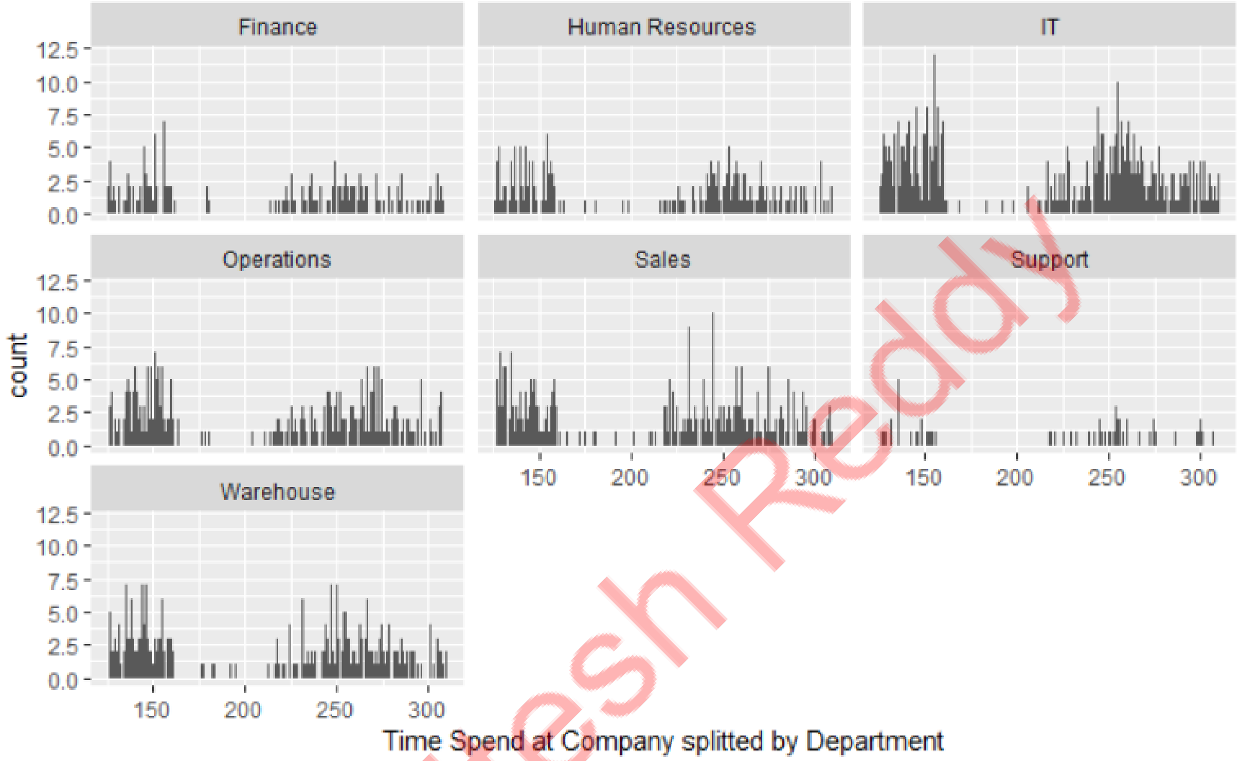
Interpretation

- The data says a lot, high number of projects assigned can be a determinant factor and closely related to leave the company.
- Imagine someone handling 7 projects at a time. Let us see more stats below to conclude

Or the average monthly hours they work for across projects?



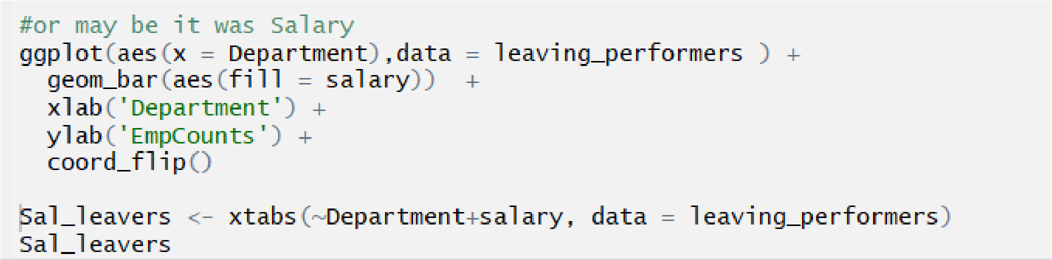
```
#or was it the average monthly hours they worked, the reason?
ggplot(aes(x = average_montly_hours),data = leaving_performers) +
  geom_bar() +
  xlab("Time Spend at Company splitted by Department") +
  facet_wrap(~Department)
```



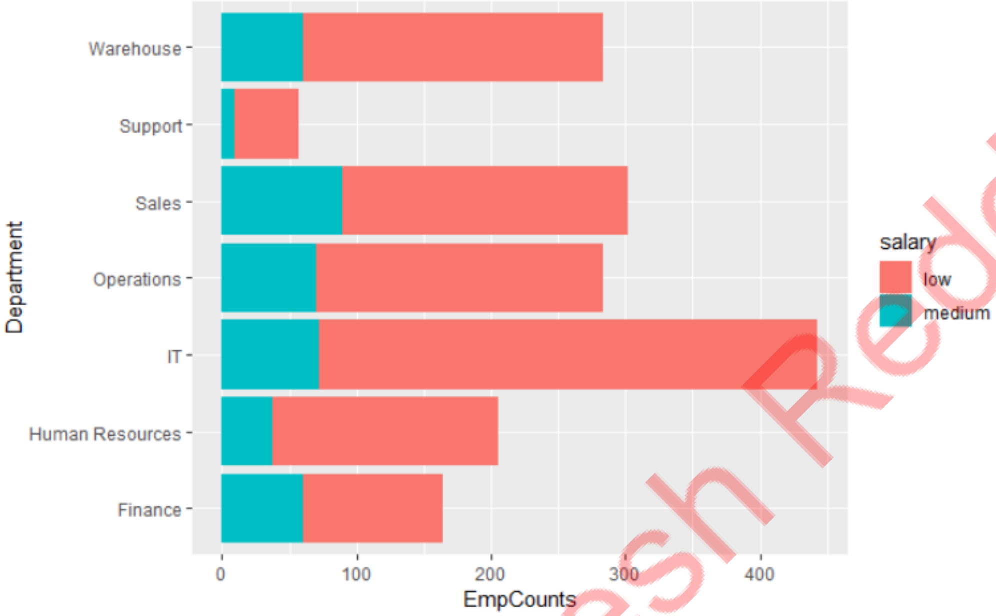
Interpretation

- Average monthly hours are the highest for multiple departments as shown above.
- In terms of the number of employees, IT department has the maximum count of employees working for more than 250 hours, suggesting a certain kind of load they have working across multiple projects as we have seen in the previous chart.

Probably salary could reveal more?

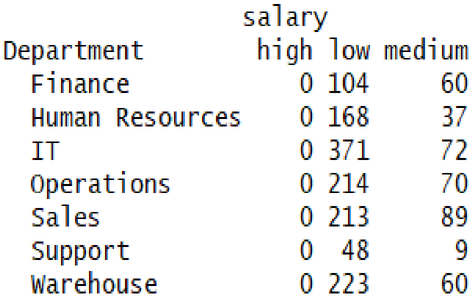


```
#or may be it was Salary
ggplot(aes(x = Department),data = leaving_performers ) +
  geom_bar(aes(fill = salary))  +
  xlab('Department') +
  ylab('EmpCounts') +
  coord_flip()

Sal_leavers <- xtabs(~Department+salary, data = leaving_performers)
Sal_leavers
```



Interpretation

- Salary gives us a final picture in concluding that last evaluation or a promotion gives no major boost in terms of financial satisfaction for any employee, also clearly seen from the table and chart above.
- Not a single employee having left got a high salary package despite having an excellent performance review.



```
                 salary
Department        high low medium
  Finance            0 104     60
  Human Resources    0 168     37
  IT                 0 371     72
  Operations         0 214     70
  Sales              0 213     89
  Support            0  48      9
  Warehouse          0 223     60
```

Conclusion is that these employees are highly valuable assets that should not be lost.

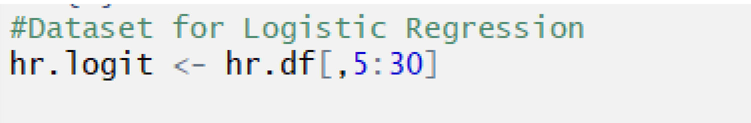
MODEL ANALYSIS

After running descriptive diagnostics on the data, we move on to predictive analytics. In this section we aim to answer the questions that will help the management to mitigate the attrition rate of employees. This analysis is important in the sense that it assists HR personnel to analyze the factors that drive employees out of the organization and to take proactive actions in retaining employees.

Question:

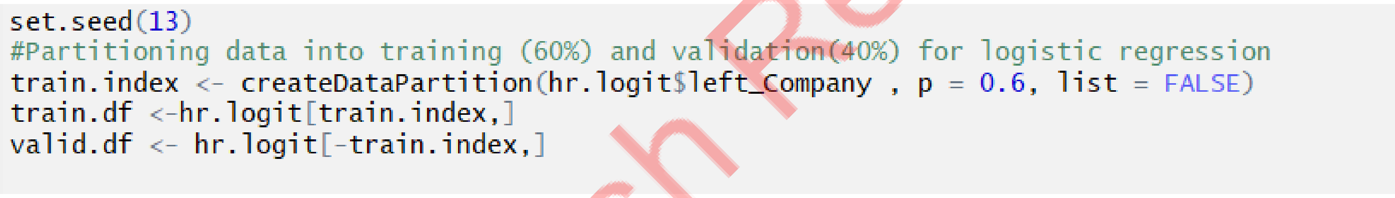
A) Will the employee leave the company?

We make a model using logistic regression to predict if the employee will leave the company. We run the algorithm after excluding the “Name”, “Department” and “Geographical location”.

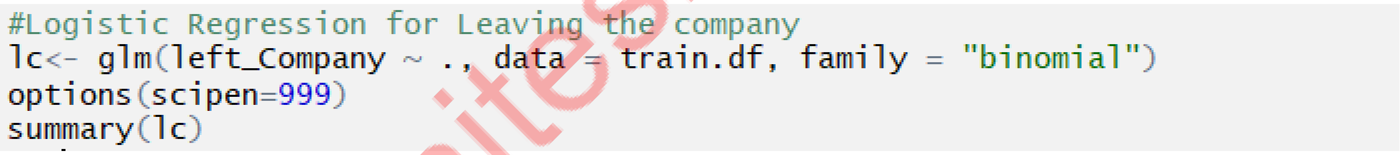


```
#Dataset for Logistic Regression
hr.logit <- hr.df[,5:30]
```

The model is trained on test data that comprises 60% of the total data and validated on the rest.

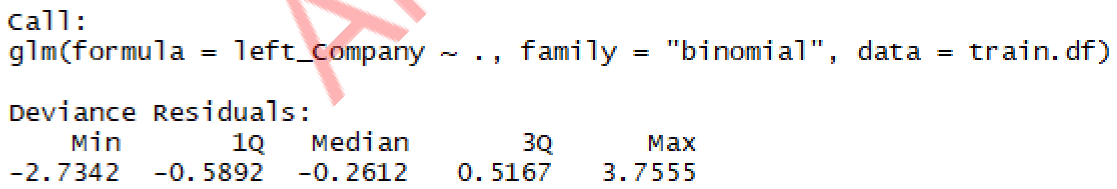


```
set.seed(13)
#Partitioning data into training (60%) and validation(40%) for logistic regression
train.index <- createDataPartition(hr.logit$left_Company , p = 0.6, list = FALSE)
train.df <-hr.logit[train.index,]
valid.df <- hr.logit[-train.index,]
```



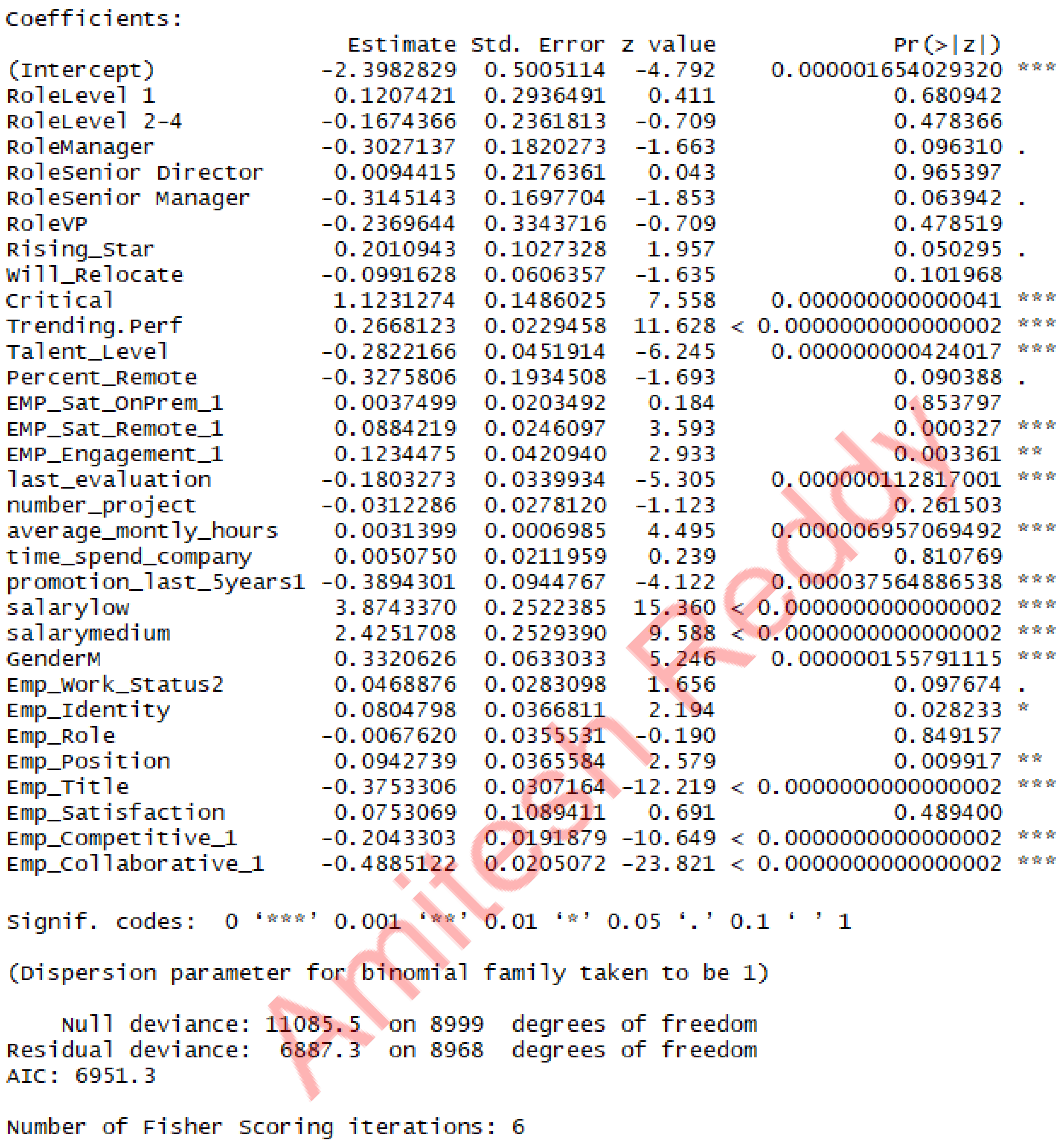
```
#Logistic Regression for Leaving the company
lc<- glm(left_Company ~ ., data = train.df, family = "binomial")
options(scipen=999)
summary(lc)
```

Output:



```
Call:
glm(formula = left_Company ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7342  -0.5892  -0.2612   0.5167   3.7555
```

Deviance residuals is the measure of how far the line of regression is from the actual point. A perfect fit of the given point equates to 0 as the log (1) is zero. However, this never occurs.



```
Coefficients:
                         Estimate Std. Error z value              Pr(>|z|)    
(Intercept)            -2.3982829  0.5005114  -4.792    0.000001654029320 ***
RoleLevel 1             0.1207421  0.2936491   0.411             0.680942    
RoleLevel 2-4          -0.1674366  0.2361813  -0.709             0.478366    
RoleManager            -0.3027137  0.1820273  -1.663             0.096310 .  
RoleSenior Director     0.0094415  0.2176361   0.043             0.965397    
RoleSenior Manager     -0.3145143  0.1697704  -1.853             0.063942 .  
RoleVP                 -0.2369644  0.3343716  -0.709             0.478519    
Rising_Star             0.2010943  0.1027328   1.957             0.050295 .  
Will_Relocate          -0.0991628  0.0606357  -1.635             0.101968    
Critical                1.1231274  0.1486025   7.558    0.000000000000041 ***
Trending.Perf           0.2668123  0.0229458  11.628 < 0.0000000000000002 ***
Talent_Level           -0.2822166  0.0451914  -6.245    0.000000000424017 ***
Percent_Remote         -0.3275806  0.1934508  -1.693             0.090388 .  
EMP_Sat_OnPrem_1        0.0037499  0.0203492   0.184             0.853797    
EMP_Sat_Remote_1        0.0884219  0.0246097   3.593             0.000327 ***
EMP_Engagement_1        0.1234475  0.0420940   2.933             0.003361 ** 
last_evaluation        -0.1803273  0.0339934  -5.305    0.000000112817001 ***
number_project         -0.0312286  0.0278120  -1.123             0.261503    
average_montly_hours    0.0031399  0.0006985   4.495    0.000006957069492 ***
time_spend_company      0.0050750  0.0211959   0.239             0.810769    
promotion_last_5years1 -0.3894301  0.0944767  -4.122    0.000037564886538 ***
salarylow               3.8743370  0.2522385  15.360 < 0.0000000000000002 ***
salarymedium            2.4251708  0.2529390   9.588 < 0.0000000000000002 ***
GenderM                 0.3320626  0.0633033   5.246    0.000000155791115 ***
Emp_Work_Status2        0.0468876  0.0283098   1.656             0.097674 .  
Emp_Identity            0.0804798  0.0366811   2.194             0.028233 *  
Emp_Role               -0.0067620  0.0355531  -0.190             0.849157    
Emp_Position            0.0942739  0.0365584   2.579             0.009917 ** 
Emp_Title              -0.3753306  0.0307164 -12.219 < 0.0000000000000002 ***
Emp_Satisfaction        0.0753069  0.1089411   0.691             0.489400    
Emp_Competitive_1      -0.2043303  0.0191879 -10.649 < 0.0000000000000002 ***
Emp_Collaborative_1    -0.4885122  0.0205072 -23.821 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11085.5  on 8999  degrees of freedom
Residual deviance:  6887.3  on 8968  degrees of freedom
AIC: 6951.3

Number of Fisher Scoring iterations: 6
```

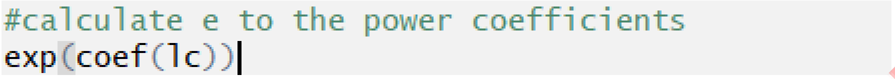
Interpretation:

Three stars indicate an extremely low P value (approximately 0), it signifies that probability of a dependent variable occurring in a certain way in accordance with the corresponding dependent variable is very low. This suggest that there is relationship between two variables in a way that independent variable largely effects the outcome of the dependent variable.

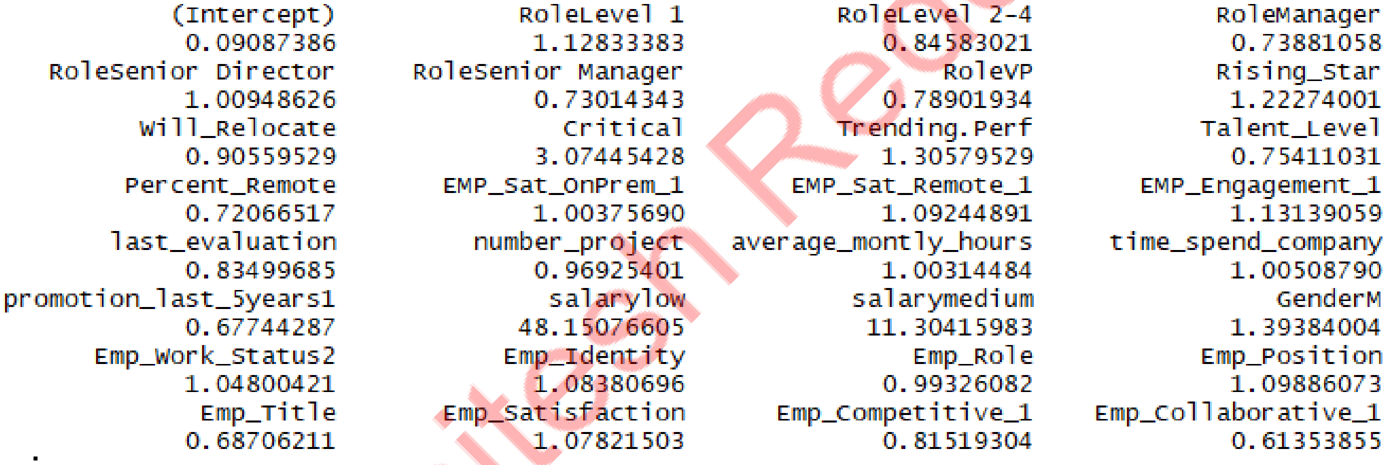
The predictors with two and three stars can be deemed important for predicting if the employee will leave the company.

Let’s go ahead and try to interpret how the coefficient estimate of “Critical” can be interpreted. The dependent variable here is “Left_Company” with “0” as still in the company and “1” as left the company. The independent variable “Critical” has “0” as not critical to the organization and “1” as critical to the organization. “0” comes first numerically for both the variables, the sequence is important in deciding the sign of coefficient estimates. The positive estimate 1.21 of “Critical” indicates that when the critical value is “0” it proves as a driving factor for the employee to leave the company resulting in “1” of the variable “Left_Company” and when the critical value is “1” it motivates the employee to stay resulting in “0” for variable “Left_Company”.

Based on the above summary and P-values of coefficient estimates it can be concluded that following predictors are important in deciding whether the employee will or will not leave the company. “Critical”, “Trending.perf”, “Talent Level”, “EMP_Sat_Remote_1”, “EMP_Engagement_1”, “last_evaluation”, “average_montly_hours”, “promotion_last_5years1”, “salarylow”, “salarymedium”, “GenderM”, “Emp_Position”, “Emp_Title”, “Emp_Competitive_1” and “Emp_Collaborative_1”

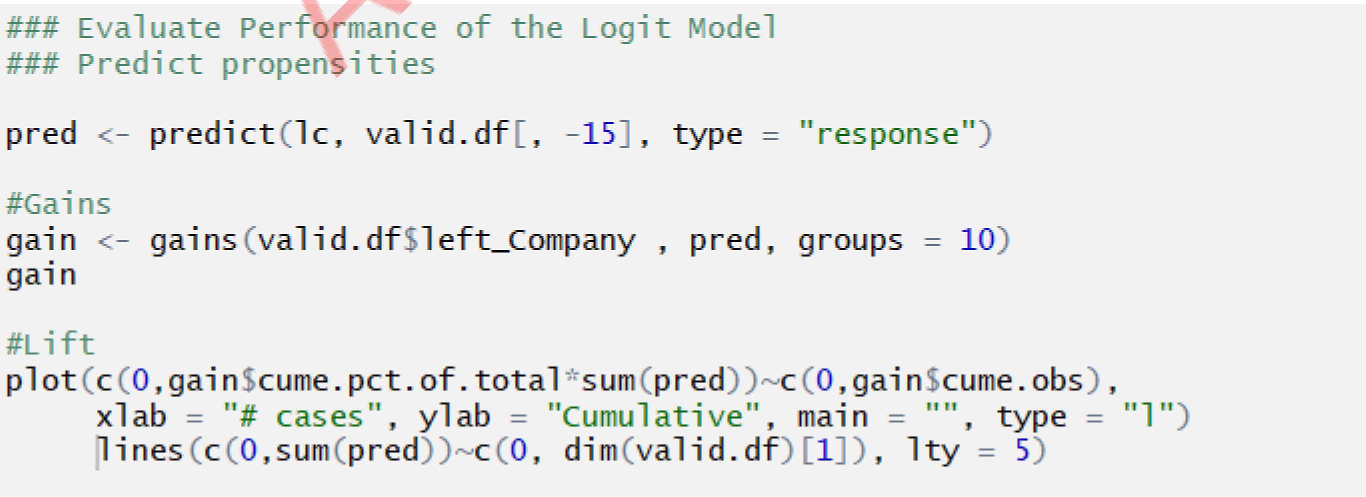


```
#calculate e to the power coefficients
exp(coef(lc))
```

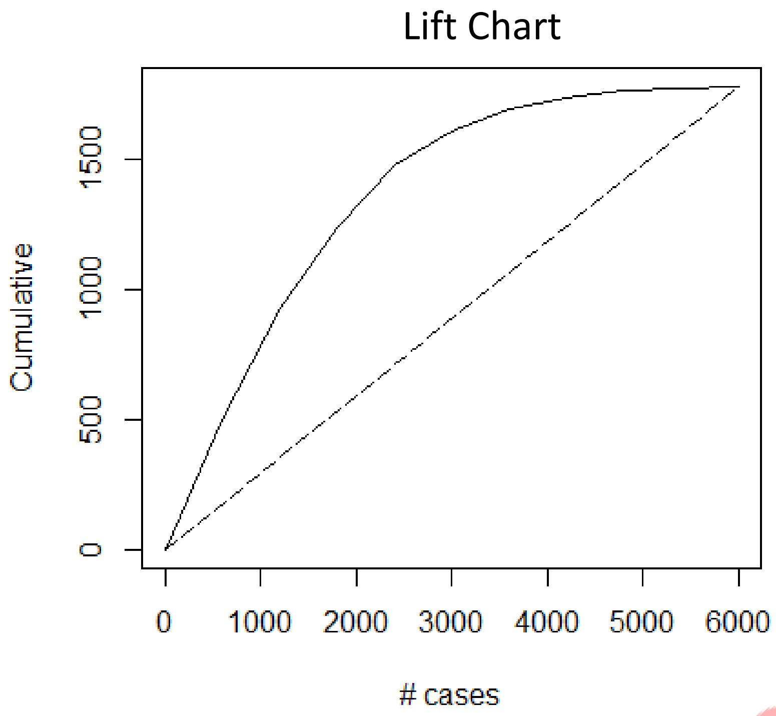


```
          (Intercept)           RoleLevel 1         RoleLevel 2-4           RoleManager
           0.09087386            1.12833383            0.84583021            0.73881058
 RoleSenior Director   RoleSenior Manager               RoleVP           Rising_Star
           1.00948626            0.73014343            0.78901934            1.22274001
        Will_Relocate              Critical         Trending.Perf          Talent_Level
           0.90559529            3.07445428            1.30579529            0.75411031
       Percent_Remote        EMP_Sat_OnPrem_1      EMP_Sat_Remote_1      EMP_Engagement_1
           0.72066517            1.00375690            1.09244891            1.13139059
      last_evaluation        number_project  average_montly_hours    time_spend_company
           0.83499685            0.96925401            1.00314484            1.00508790
promotion_last_5years1             salarylow          salarymedium               GenderM
           0.67744287           48.15076605           11.30415983            1.39384004
      Emp_Work_Status2          Emp_Identity              Emp_Role          Emp_Position
           1.04800421            1.08380696            0.99326082            1.09886073
            Emp_Title      Emp_Satisfaction      Emp_Competitive_1   Emp_Collaborative_1
           0.68706211            1.07821503            0.81519304            0.61353855
  .
```

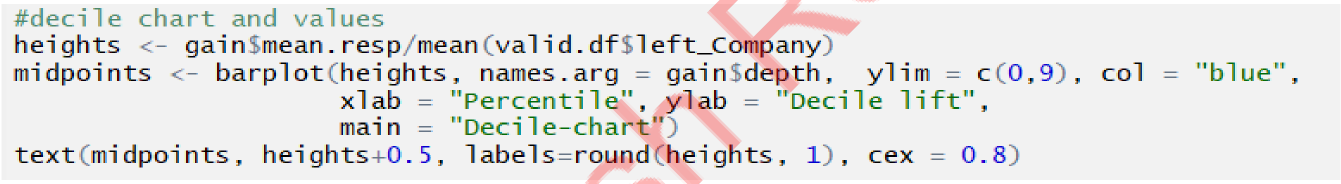
From the above values it is evident that Low salary has the highest impact on employees leaving the company followed by medium salary and criticalness.



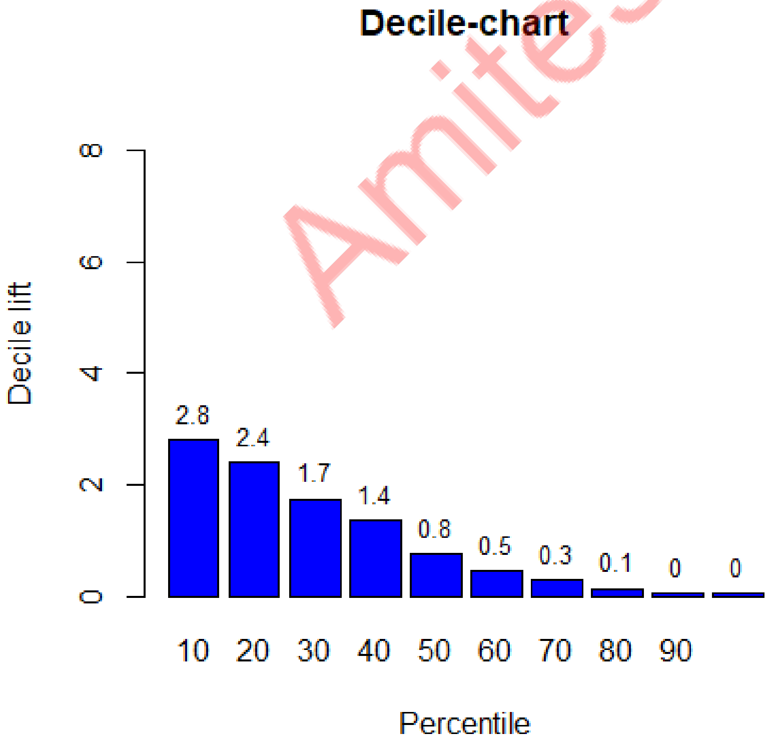
```
### Evaluate Performance of the Logit Model
### Predict propensities

pred <- predict(lc, valid.df[, -15], type = "response")

#Gains
gain <- gains(valid.df$left_Company , pred, groups = 10)
gain

#Lift
plot(c(0,gain$cume.pct.of.total*sum(pred))~c(0,gain$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
     lines(c(0,sum(pred))~c(0, dim(valid.df)[1]), lty = 5)
```



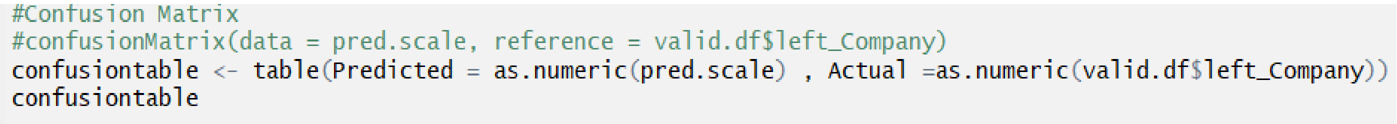
As seen from the above lift chart, it is evident that the model curve has more area under it compared to the naïve rule represented by the straig ht line.



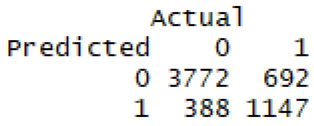
```
#decile chart and values
heights <- gain$mean.resp/mean(valid.df$left_Company)
midpoints <- barplot(heights, names.arg = gain$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)
```



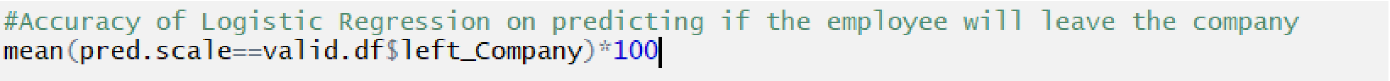
- Decile chart follows an ideal structure representing maximum variation cover ed in initial deciles.
- First 5 deciles cover 90% of the variation.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 2.8 time better than the one with Naïve rule.



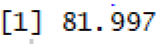
```
#Confusion Matrix
#confusionMatrix(data = pred.scale, reference = valid.df$left_Company)
confusiontable <- table(Predicted = as.numeric(pred.scale) , Actual =as.numeric(valid.df$left_Company))
confusiontable
```



```
         Actual
Predicted    0    1
        0 3772  692
        1  388 1147
```



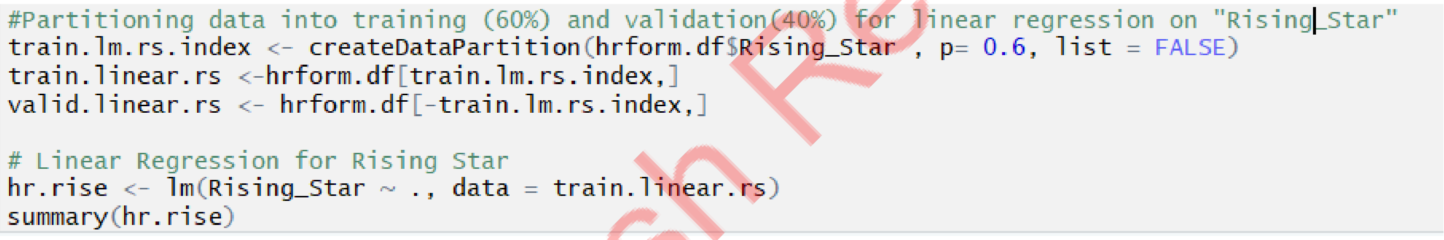
```
#Accuracy of Logistic Regression on predicting if the employee will leave the company
mean(pred.scale==valid.df$left_Company)*100
```



```
[1] 81.997
```

B) What is the likelihood of Employee getting a promotion?

We run the linear regression algorithm on non-categorical variables keeping "Rising_Star" as the dependent variable. The model is trained on test data that comprises 60% of the total data and validated on the rest.



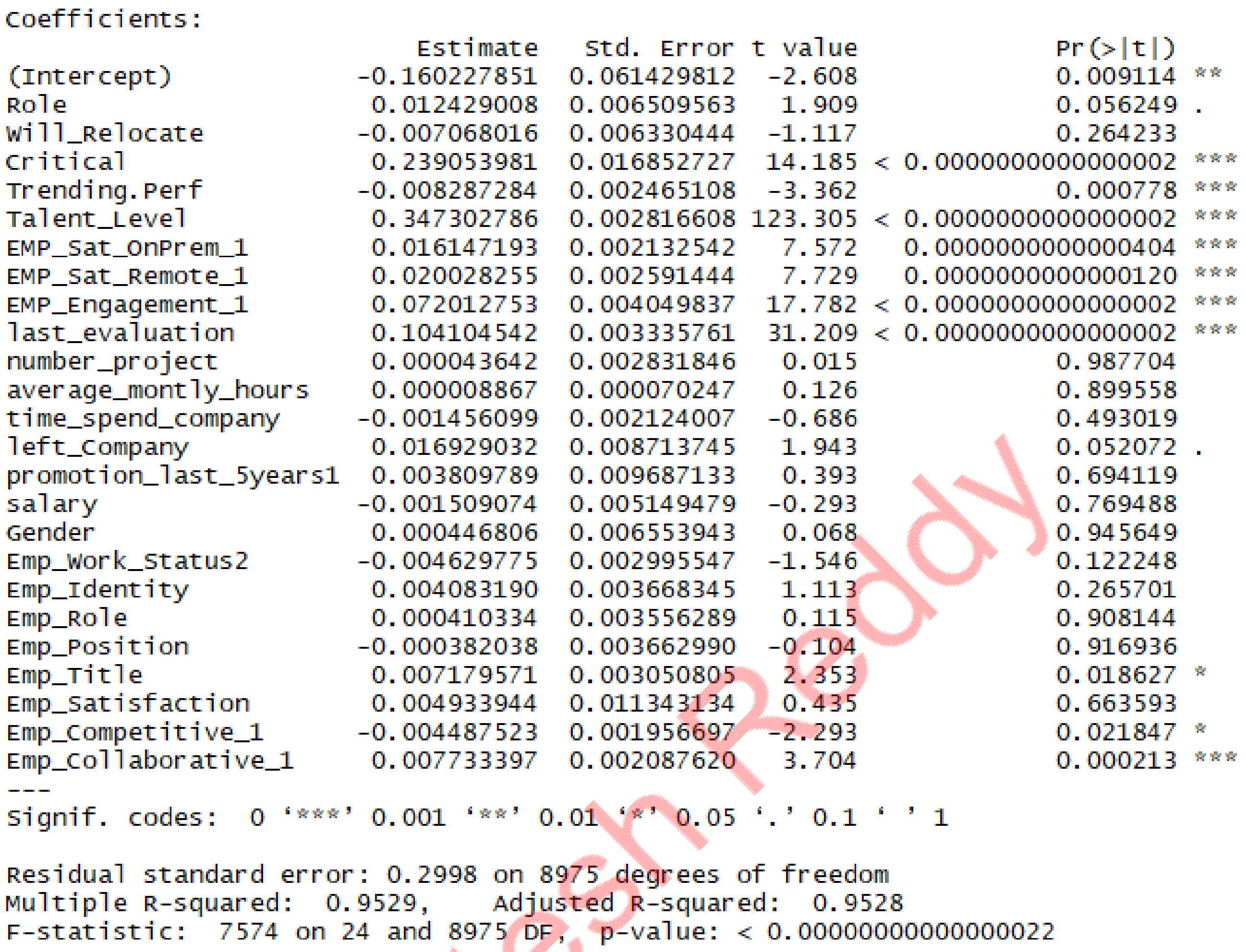
```
#Partitioning data into training (60%) and validation(40%) for linear regression on "Rising_Star"
train.lm.rs.index <- createDataPartition(hrform.df$Rising_Star , p= 0.6, list = FALSE)
train.linear.rs <-hrform.df[train.lm.rs.index,]
valid.linear.rs <- hrform.df[-train.lm.rs.index,]

# Linear Regression for Rising Star
hr.rise <- lm(Rising_Star ~ ., data = train.linear.rs)
summary(hr.rise)
```

The significant coefficients (P Value two and three stars) for Rising_Star are:

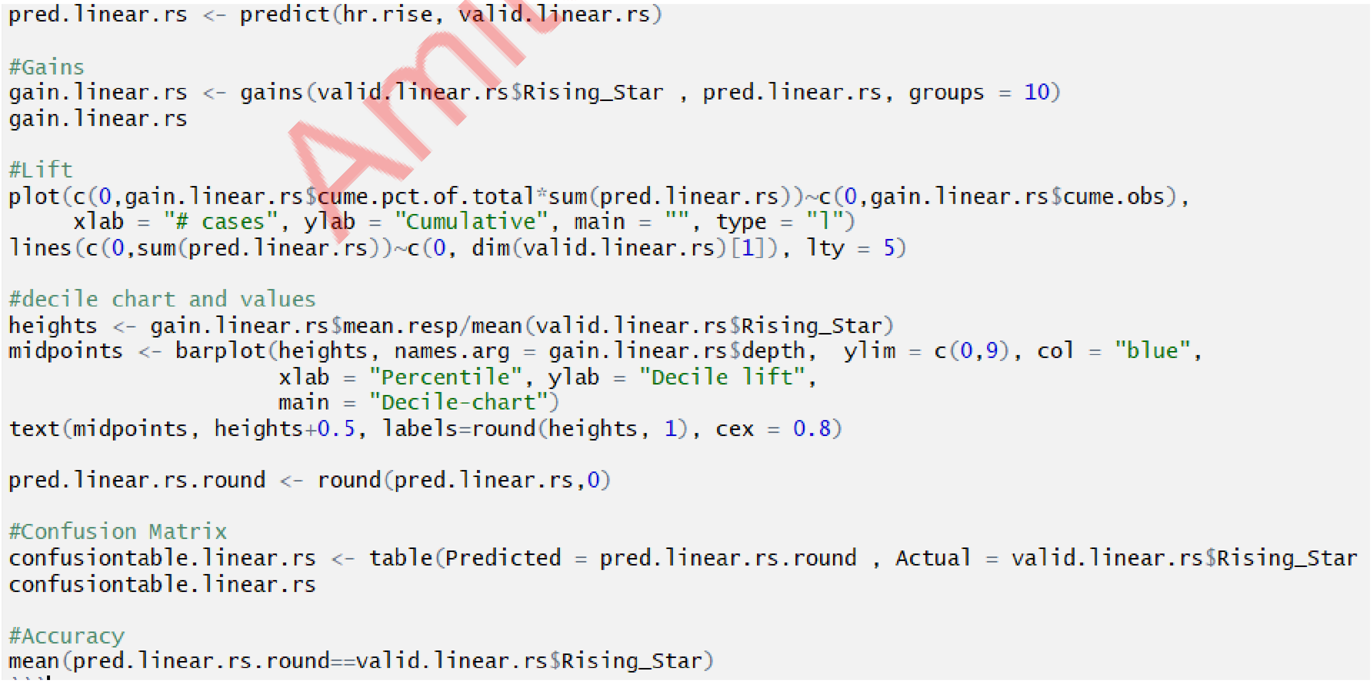
Critical: Positive coefficient signifies that if the employee is critical ("1") the likely hood of promotion ("Rising_Star) also increases in number (1 through 5). For every one-unit change in Critical value, the independent variable is effected change +0.239
Trending.perf: For every unit change in Trending.perf, there is negative 0.0082 effect on Rising_Star.
Talent Leve: For every unit change in Trending.perf, there is positive 0.3473 effect on Rising_Star.

Similarly, variables EMP_SAT_OnPRem_1, EMP_SAT_Remote1, EMP_Engagement_1, last_Evaluation, number_projects and Emp_Collaborative_1 significantly determine the output of Rising_Star.

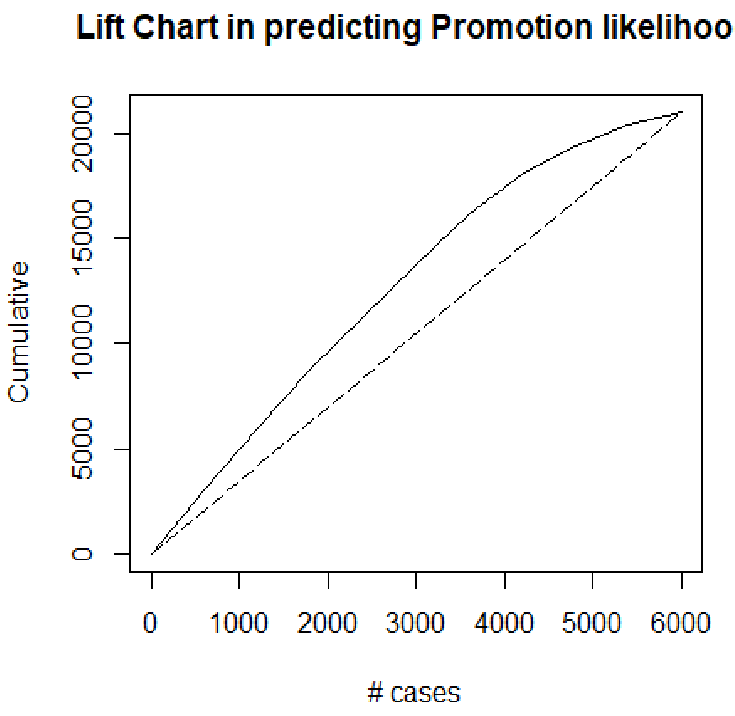
Adjusted R square value of **0.9528** can be considered as an excellent number exhibiting that approximately **95%** of the variation in Rising_Star variable is captured by the input variables.



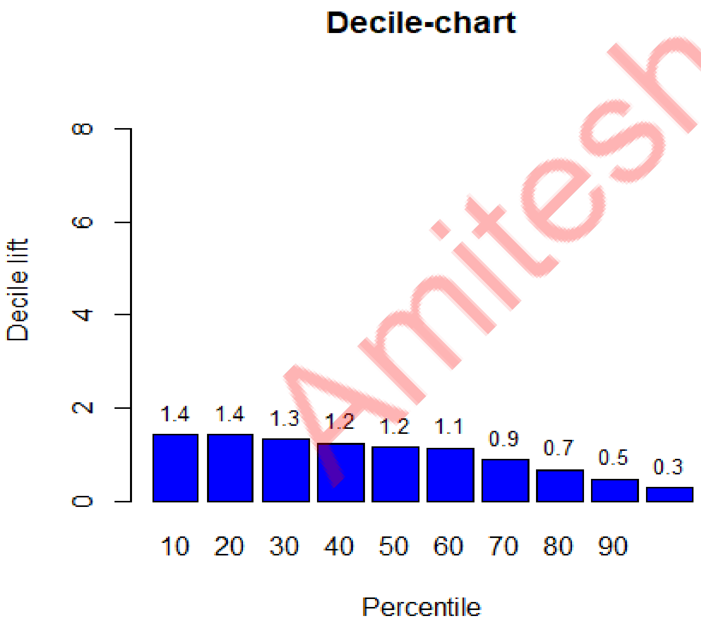
```
Coefficients:
                          Estimate   Std. Error t value            Pr(>|t|)    
(Intercept)            -0.160227851 0.061429812  -2.608            0.009114 ** 
Role                    0.012429008 0.006509563   1.909            0.056249 .  
Will_Relocate          -0.007068016 0.006330444  -1.117            0.264233    
Critical                0.239053981 0.016852727  14.185 < 0.0000000000000002 ***
Trending.Perf          -0.008287284 0.002465108  -3.362            0.000778 ***
Talent_Level            0.347302786 0.002816608 123.305 < 0.0000000000000002 ***
EMP_Sat_OnPrem_1        0.016147193 0.002132542   7.572   0.0000000000000404 ***
EMP_Sat_Remote_1        0.020028255 0.002591444   7.729   0.0000000000000120 ***
EMP_Engagement_1        0.072012753 0.004049837  17.782 < 0.0000000000000002 ***
last_evaluation         0.104104542 0.003335761  31.209 < 0.0000000000000002 ***
number_project          0.000043642 0.002831846   0.015            0.987704    
average_montly_hours    0.000008867 0.000070247   0.126            0.899558    
time_spend_company     -0.001456099 0.002124007  -0.686            0.493019    
left_Company            0.016929032 0.008713745   1.943            0.052072 .  
promotion_last_5years1  0.003809789 0.009687133   0.393            0.694119    
salary                 -0.001509074 0.005149479  -0.293            0.769488    
Gender                  0.000446806 0.006553943   0.068            0.945649    
Emp_Work_Status2       -0.004629775 0.002995547  -1.546            0.122248    
Emp_Identity            0.004083190 0.003668345   1.113            0.265701    
Emp_Role                0.000410334 0.003556289   0.115            0.908144    
Emp_Position           -0.000382038 0.003662990  -0.104            0.916936    
Emp_Title               0.007179571 0.003050805   2.353            0.018627 *  
Emp_Satisfaction        0.004933944 0.011343134   0.435            0.663593    
Emp_Competitive_1      -0.004487523 0.001956697  -2.293            0.021847 *  
Emp_Collaborative_1     0.007733397 0.002087620   3.704            0.000213 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2998 on 8975 degrees of freedom
Multiple R-squared:  0.9529,	Adjusted R-squared:  0.9528 
F-statistic:  7574 on 24 and 8975 DF,  p-value: < 0.00000000000000022
```



```
pred.linear.rs <- predict(hr.rise, valid.linear.rs)

#Gains
gain.linear.rs <- gains(valid.linear.rs$Rising_Star , pred.linear.rs, groups = 10)
gain.linear.rs

#Lift
plot(c(0,gain.linear.rs$cume.pct.of.total*sum(pred.linear.rs))~c(0,gain.linear.rs$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.rs))~c(0, dim(valid.linear.rs)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.rs$mean.resp/mean(valid.linear.rs$Rising_Star)
midpoints <- barplot(heights, names.arg = gain.linear.rs$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.rs.round <- round(pred.linear.rs,0)

#Confusion Matrix
confusiontable.linear.rs <- table(Predicted = pred.linear.rs.round , Actual = valid.linear.rs$Rising_Star
confusiontable.linear.rs

#Accuracy
mean(pred.linear.rs.round==valid.linear.rs$Rising_Star)
```

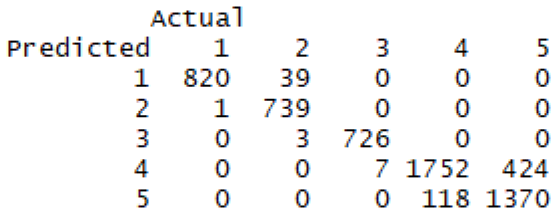

Lift Chart in predicting Promotion likelihood

- As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.


Decile-chart

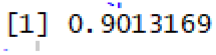
- Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 1.4 time better than the one with Naïve rule.

Confusion Matrix



```
         Actual
Predicted    1    2    3    4    5
        1  820   39    0    0    0
        2    1  739    0    0    0
        3    0    3  726    0    0
        4    0    0    7 1752  424
        5    0    0    0  118 1370
```

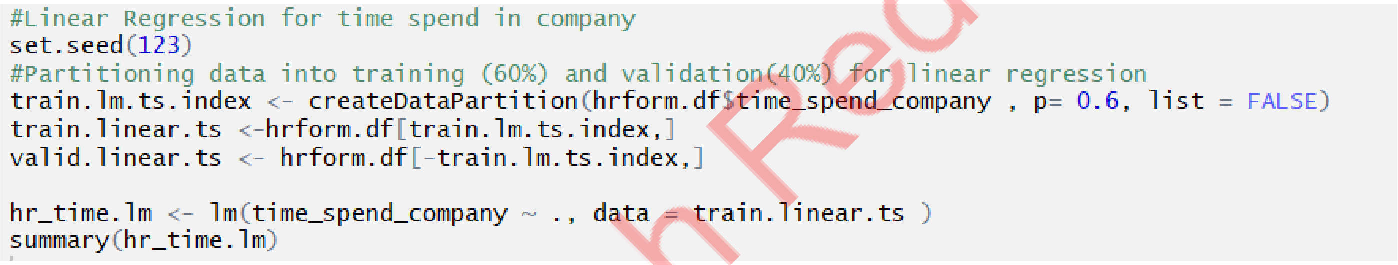
Accuracy in predicting Validation data set



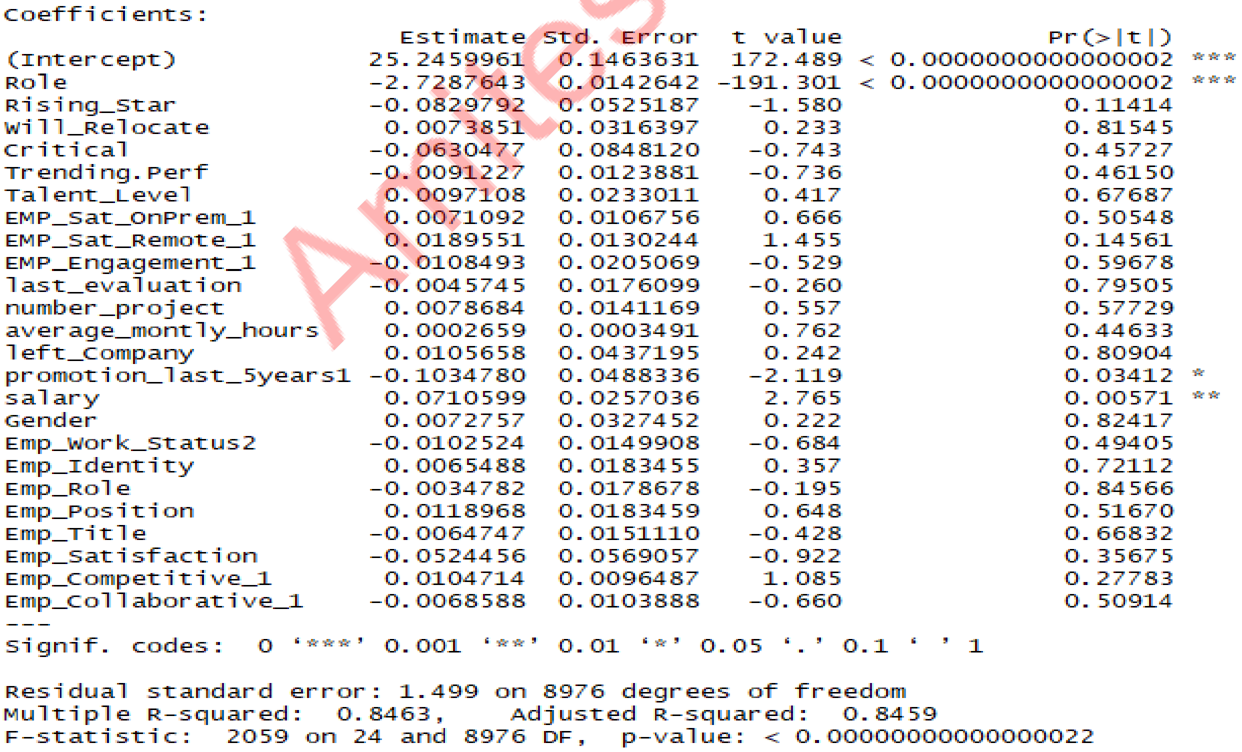
```
[1] 0.9013169
```

C) How much time will the employee spend in company?

We run the linear regression algorithm on non-categorical variables keeping “time_spend_company” as the dependent variable. The model is trained on test data that comprises 60% of the total data and validated on the rest.



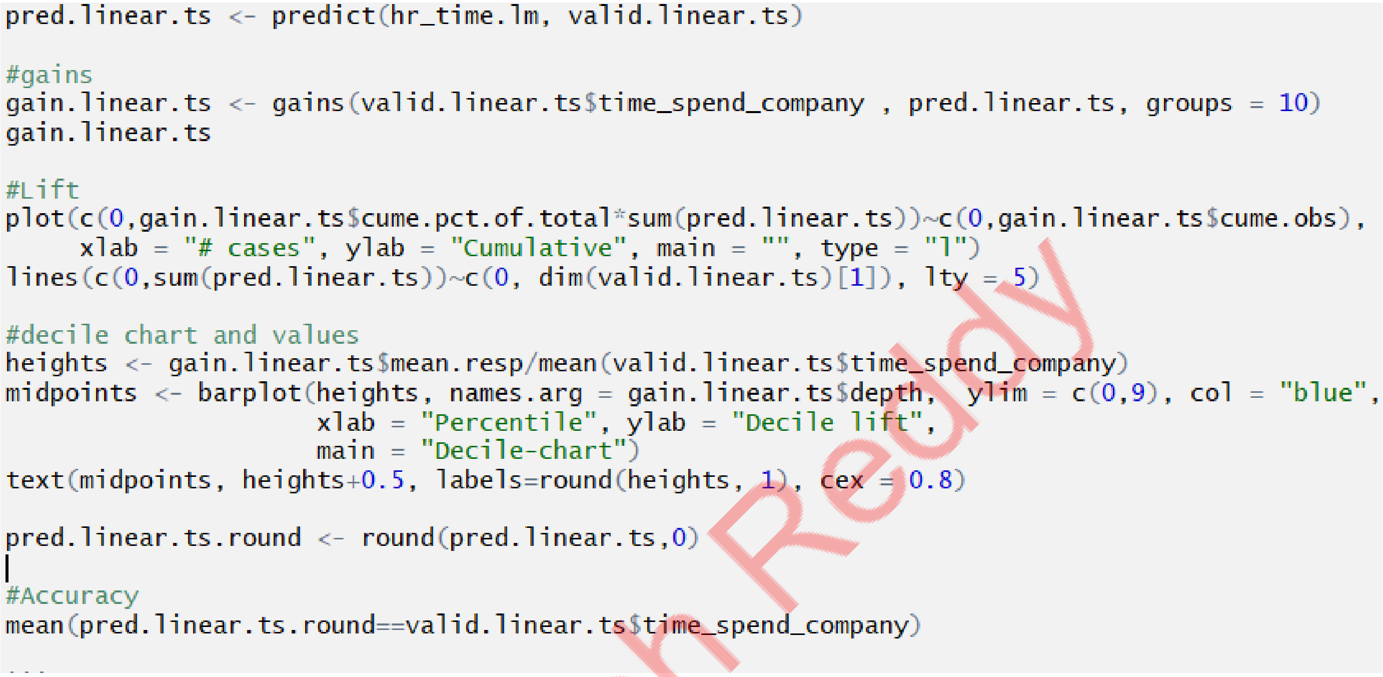
```
#Linear Regression for time spend in company
set.seed(123)
#Partitioning data into training (60%) and validation(40%) for linear regression
train.lm.ts.index <- createDataPartition(hrform.df$time_spend_company , p= 0.6, list = FALSE)
train.linear.ts <-hrform.df[train.lm.ts.index,]
valid.linear.ts <- hrform.df[-train.lm.ts.index,]

hr_time.lm <- lm(time_spend_company ~ ., data = train.linear.ts )
summary(hr_time.lm)
```



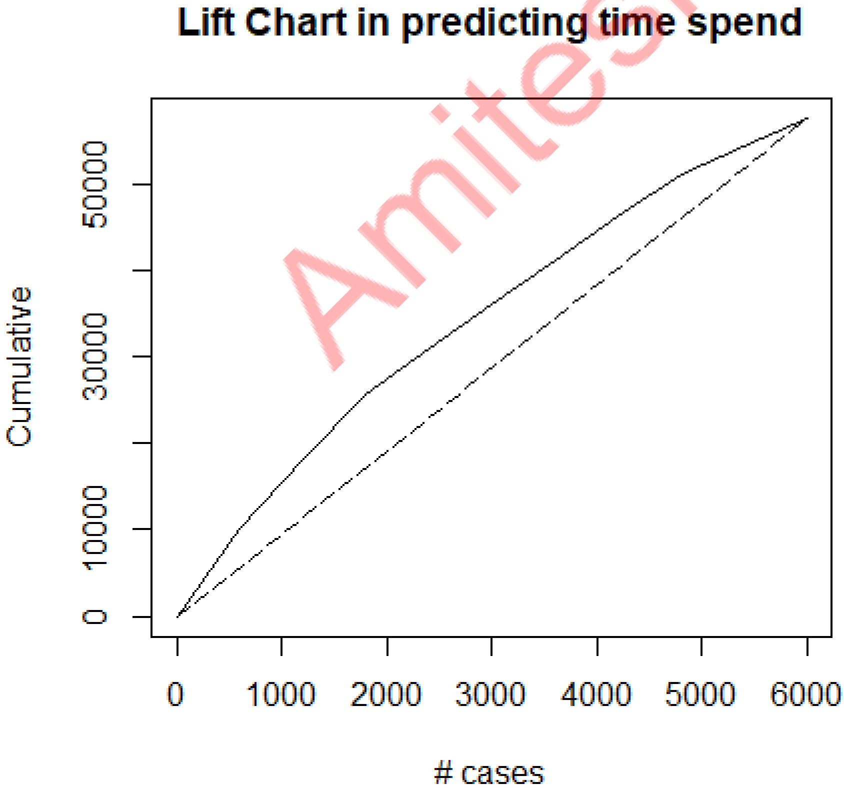
```
Coefficients:
                        Estimate Std. Error  t value              Pr(>|t|)    
(Intercept)           25.2459961  0.1463631  172.489 < 0.0000000000000002 ***
Role                  -2.7287643  0.0142642 -191.301 < 0.0000000000000002 ***
Rising_Star           -0.0829792  0.0525187   -1.580              0.11414    
Will_Relocate          0.0073851  0.0316397    0.233              0.81545    
Critical              -0.0630477  0.0848120   -0.743              0.45727    
Trending.Perf         -0.0091227  0.0123881   -0.736              0.46150    
Talent_Level           0.0097108  0.0233011    0.417              0.67687    
EMP_Sat_OnPrem_1       0.0071092  0.0106756    0.666              0.50548    
EMP_Sat_Remote_1       0.0189551  0.0130244    1.455              0.14561    
EMP_Engagement_1      -0.0108493  0.0205069   -0.529              0.59678    
last_evaluation       -0.0045745  0.0176099   -0.260              0.79505    
number_project         0.0078684  0.0141169    0.557              0.57729    
average_montly_hours   0.0002659  0.0003491    0.762              0.44633    
left_Company           0.0105658  0.0437195    0.242              0.80904    
promotion_last_5years1 -0.1034780  0.0488336   -2.119              0.03412 *  
salary                 0.0710599  0.0257036    2.765              0.00571 ** 
Gender                 0.0072757  0.0327452    0.222              0.82417    
Emp_Work_Status2      -0.0102524  0.0149908   -0.684              0.49405    
Emp_Identity           0.0065488  0.0183455    0.357              0.72112    
Emp_Role              -0.0034782  0.0178678   -0.195              0.84566    
Emp_Position           0.0118968  0.0183459    0.648              0.51670    
Emp_Title             -0.0064747  0.0151110   -0.428              0.66832    
Emp_Satisfaction      -0.0524456  0.0569057   -0.922              0.35675    
Emp_Competitive_1      0.0104714  0.0096487    1.085              0.27783    
Emp_Collaborative_1   -0.0068588  0.0103888   -0.660              0.50914    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.499 on 8976 degrees of freedom
Multiple R-squared:  0.8463,	Adjusted R-squared:  0.8459 
F-statistic:  2059 on 24 and 8976 DF,  p-value: < 0.00000000000000022
```

The significant coefficients (P Value one, two and three stars) for time_spend_company are Role, promotion_last_5years1 and salary.

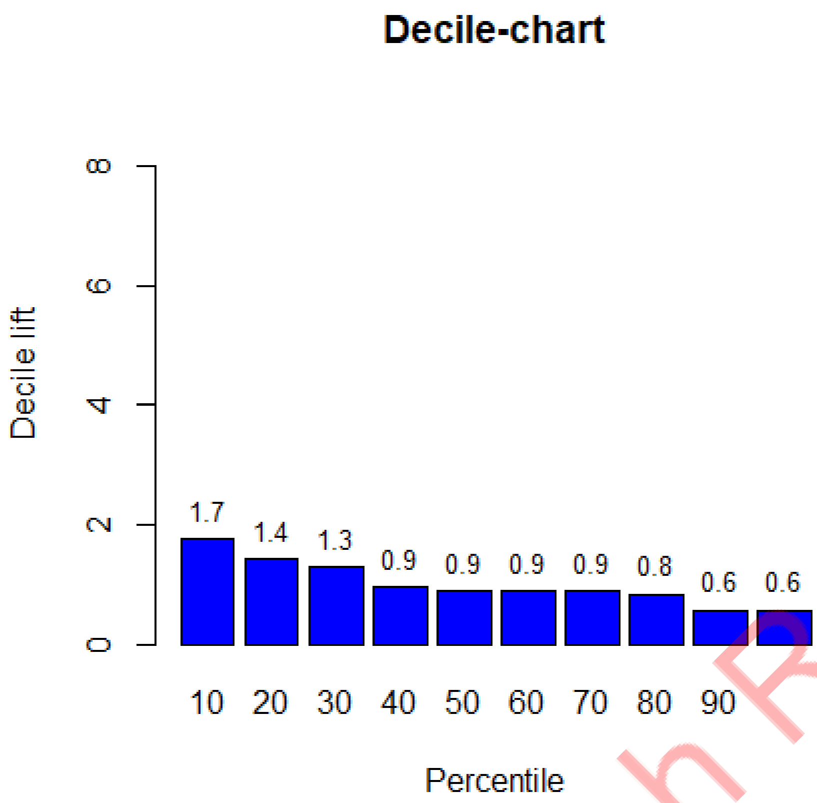
Adjusted R square value of **0.8459** can be considered as a good number exhibiting that approximately **85%** of the variation in time_spend_company variable is captured by the input variables.



```
pred.linear.ts <- predict(hr_time.lm, valid.linear.ts)

#gains
gain.linear.ts <- gains(valid.linear.ts$time_spend_company , pred.linear.ts, groups = 10)
gain.linear.ts

#Lift
plot(c(0,gain.linear.ts$cume.pct.of.total*sum(pred.linear.ts))~c(0,gain.linear.ts$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.ts))~c(0, dim(valid.linear.ts)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.ts$mean.resp/mean(valid.linear.ts$time_spend_company)
midpoints <- barplot(heights, names.arg = gain.linear.ts$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.ts.round <- round(pred.linear.ts,0)

#Accuracy
mean(pred.linear.ts.round==valid.linear.ts$time_spend_company)
```



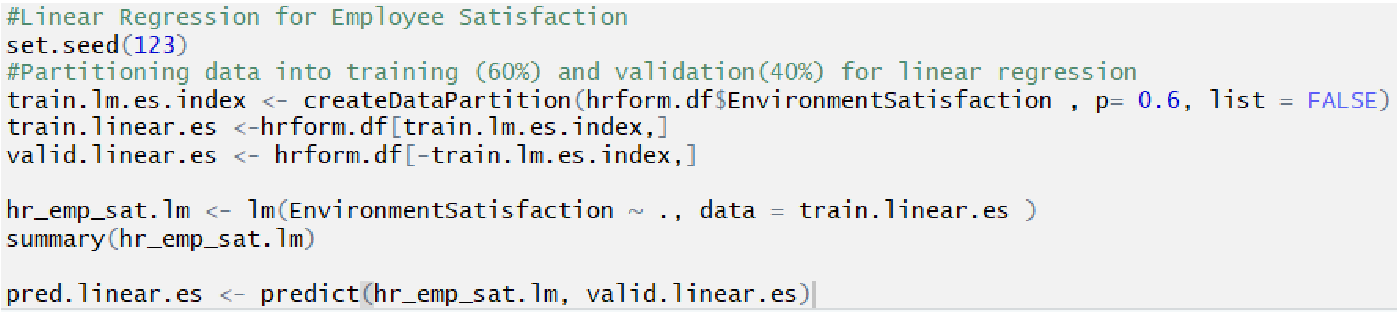
As seen from the above lift chart, it is evident that the model curve has comparatively more area(covers more variation) under it compared to the naïve rule represented by the straight line.



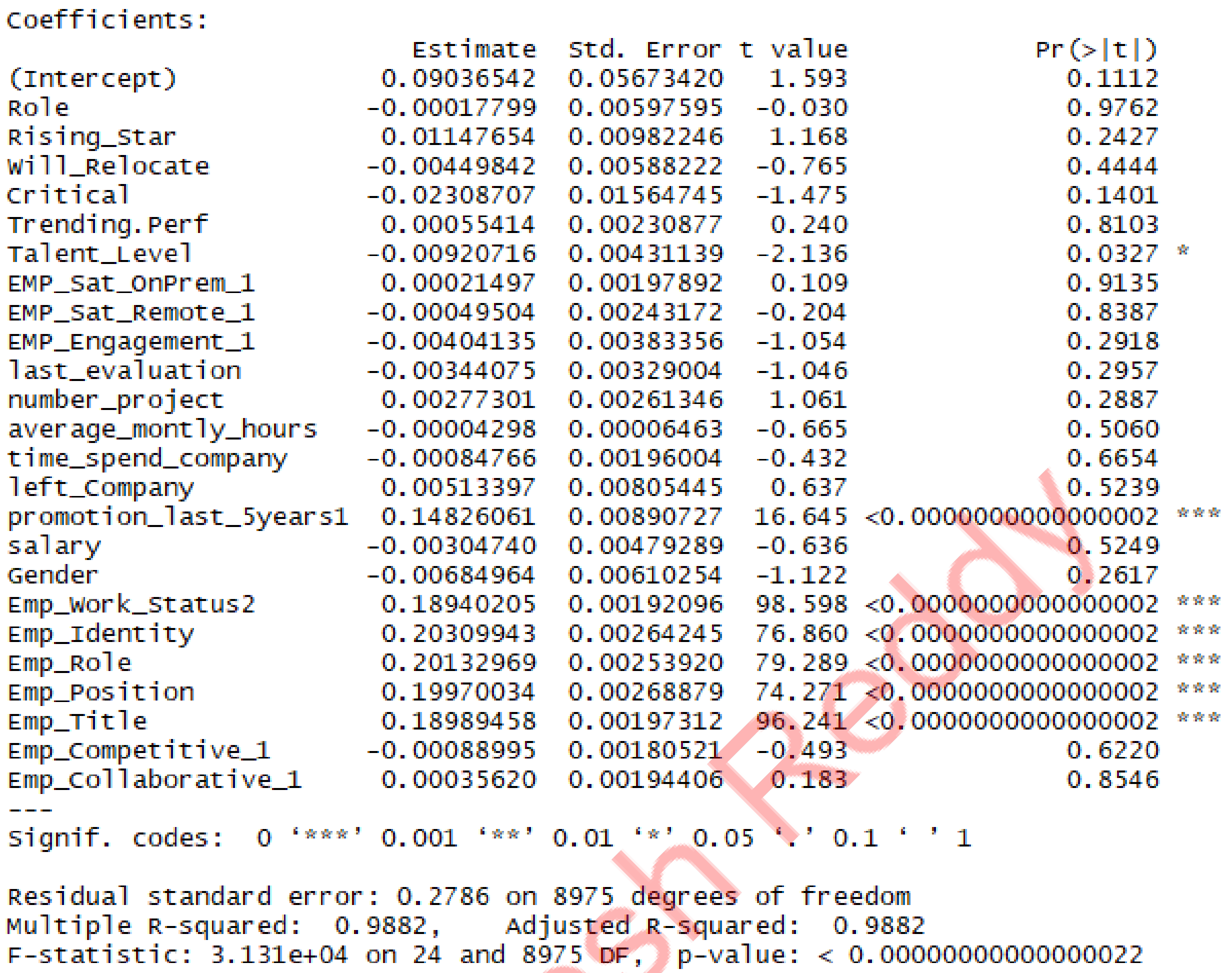
Decile chart follows an ideal structure representing maximum variation covered in initial deciles. This can be considered as good model where the deciles are decreasing in order from start to end. Looking at the first decile, we can say that this model performs 1.7 time better than the one with Naïve rule.

D) How satisfied are the employees in company?

We run the linear regression algorithm on non-categorical variables keeping "Emp_Satisfaction" as the dependent variable.
The model is trained on test data that comprises 60% of the total data and validated on the rest.



```
#Linear Regression for Employee Satisfaction
set.seed(123)
#Partitioning data into training (60%) and validation(40%) for linear regression
train.lm.es.index <- createDataPartition(hrform.df$EnvironmentSatisfaction , p= 0.6, list = FALSE)
train.linear.es <-hrform.df[train.lm.es.index,]
valid.linear.es <- hrform.df[-train.lm.es.index,]

hr_emp_sat.lm <- lm(EnvironmentSatisfaction ~ ., data = train.linear.es )
summary(hr_emp_sat.lm)

pred.linear.es <- predict(hr_emp_sat.lm, valid.linear.es)
```

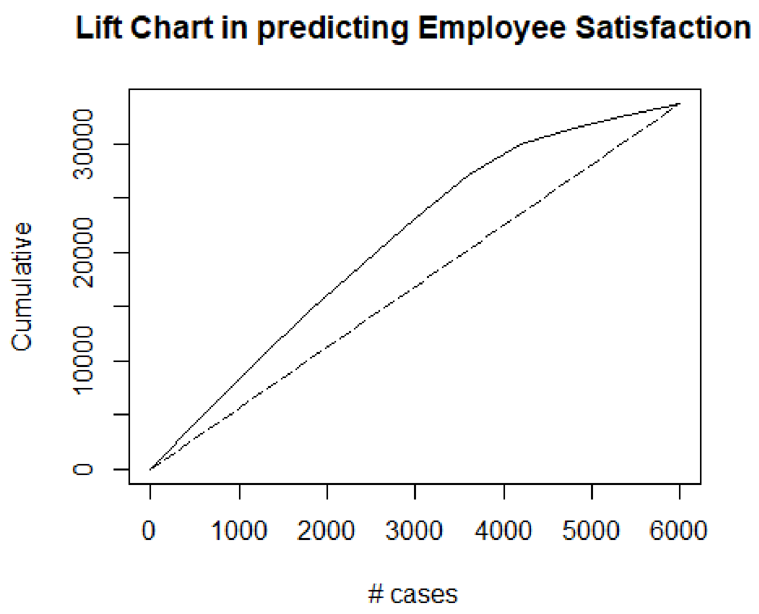


```
Coefficients:
                          Estimate  Std. Error t value           Pr(>|t|)    
(Intercept)             0.09036542  0.05673420   1.593             0.1112    
Role                   -0.00017799  0.00597595  -0.030             0.9762    
Rising_Star             0.01147654  0.00982246   1.168             0.2427    
Will_Relocate          -0.00449842  0.00588222  -0.765             0.4444    
Critical               -0.02308707  0.01564745  -1.475             0.1401    
Trending.Perf           0.00055414  0.00230877   0.240             0.8103    
Talent_Level           -0.00920716  0.00431139  -2.136             0.0327 *  
EMP_Sat_OnPrem_1        0.00021497  0.00197892   0.109             0.9135    
EMP_Sat_Remote_1       -0.00049504  0.00243172  -0.204             0.8387    
EMP_Engagement_1       -0.00404135  0.00383356  -1.054             0.2918    
last_evaluation        -0.00344075  0.00329004  -1.046             0.2957    
number_project          0.00277301  0.00261346   1.061             0.2887    
average_montly_hours   -0.00004298  0.00006463  -0.665             0.5060    
time_spend_company     -0.00084766  0.00196004  -0.432             0.6654    
left_Company            0.00513397  0.00805445   0.637             0.5239    
promotion_last_5years1  0.14826061  0.00890727  16.645 <0.0000000000000002 ***
salary                 -0.00304740  0.00479289  -0.636             0.5249    
Gender                 -0.00684964  0.00610254  -1.122             0.2617    
Emp_Work_Status2        0.18940205  0.00192096  98.598 <0.0000000000000002 ***
Emp_Identity            0.20309943  0.00264245  76.860 <0.0000000000000002 ***
Emp_Role                0.20132969  0.00253920  79.289 <0.0000000000000002 ***
Emp_Position            0.19970034  0.00268879  74.271 <0.0000000000000002 ***
Emp_Title               0.18989458  0.00197312  96.241 <0.0000000000000002 ***
Emp_Competitive_1      -0.00088995  0.00180521  -0.493             0.6220    
Emp_Collaborative_1     0.00035620  0.00194406   0.183             0.8546    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2786 on 8975 degrees of freedom
Multiple R-squared:  0.9882,	Adjusted R-squared:  0.9882 
F-statistic: 3.131e+04 on 24 and 8975 DF,  p-value: < 0.00000000000000022
```

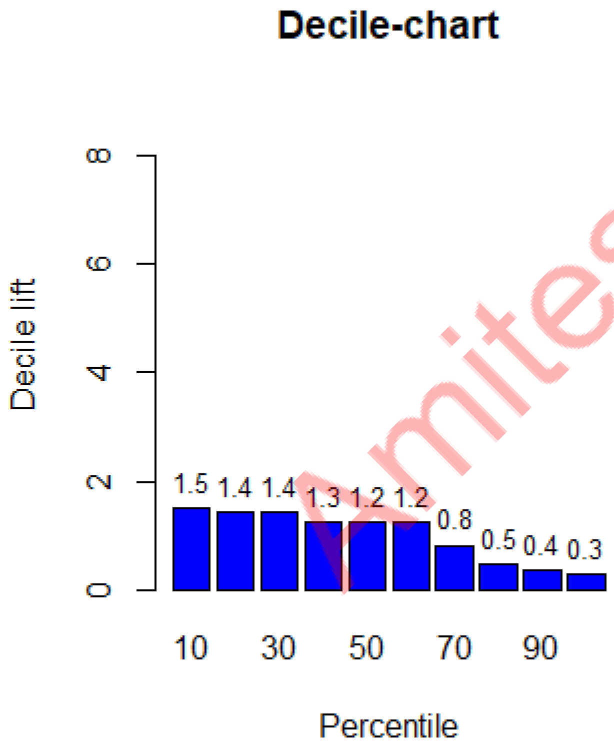
The significant coefficients (P Value one, two and three stars) for Emp_Satisfaction are Talent_Level, promotion_last_5years1, Emp_work_Status2, Emp_Identity, Emp_Role, Emp_Position and Emp_Title.

Adjusted R square value of **0.9882** can be considered as an excellent number exhibiting that approximately **99%** of the variation in Emp_Satisfaction variable is captured by the input variables.

```
#gains
gain.linear.es <- gains(valid.linear.es$EnvironmentSatisfaction , pred.linear.es, groups = 10)
gain.linear.es

#Lift
plot(c(0,gain.linear.es$cume.pct.of.total*sum(pred.linear.es))~c(0,gain.linear.es$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.es))~c(0, dim(valid.linear.es)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.es$mean.resp/mean(valid.linear.es$EnvironmentSatisfaction)
midpoints <- barplot(heights, names.arg = gain.linear.es$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.es.round <- round(pred.linear.es,0)

#Accuracy
mean(pred.linear.es.round==valid.linear.es$EnvironmentSatisfaction)
```

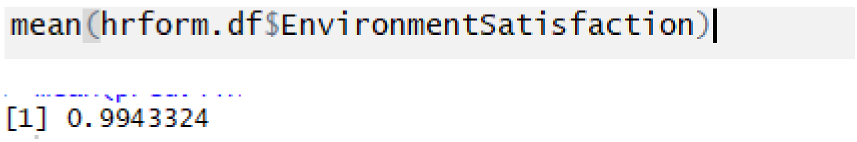


- As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.



- Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 1.5 time better than the one with Naïve rule.

Accuracy in predicting the Employee satisfaction in Validation data set



```
mean(hrform.df$EnvironmentSatisfaction)
```

```
[1] 0.9943324
```