Abstract

Human Resource Analytics using R

HR Analytics

BUAN 6356

------------------------------------------------------------------------------------------------------------------------------

Parth Rajgor, Amitesh Enugala, Jiacheng Huang, Vigneash Rajagopal, Rahul Bafna

**HR Analytics**

Submitted by

Group 9

Under the guidance of

Sourav Chatterjee



The University of Texas at Dallas

Richardson, TX 75080

**Nov 2018**

## **TABLE OF CONTENTS**

INDEX

[**TABLE OF CONTENTS** 2](#_Toc531213718)

[**ACKNOWLEDGEMENT** 3](#_Toc531213719)

[**LITERATURE** 4](#_Toc531213720)

[**BACKGROUND** 5](#_Toc531213721)

[**OBJECTIVES** 7](#_Toc531213722)

[**DATA EXPLORATION** 8](#_Toc531213723)

[**WHY DO GOOD EMPLOYEES LEAVE?** 17](#_Toc531213724)

[**MODEL ANALYSIS** 21](#_Toc531213725)

[Running various models to answer the below questions- 24](#_Toc531213726)

[A) Will the employee leave the company? 24](#_Toc531213727)

[B) What is the likelihood of employees getting a promotion? 29](#_Toc531213728)

[C) How much time will the employee spend in company? 34](#_Toc531213729)

[D) How satisfied are the employees in company? 37](#_Toc531213730)

[E) Running K-mean Clustering to find out which set of employees are more likely to exit 40](#_Toc531213731)

[**REFERENCES & CITATIONS** 42](#_Toc531213732)

## **ACKNOWLEDGEMENT**

Our project on **HR Analytics** has been a great learning experience. We were exposed to a variance of subject matter, concerns and arguments that helped us collectively assemble and shape the project.

We acknowledge Sourav Chatterjee under whose guidance we were able to complete the project and effectively present its valuable benefits.

A greater share of inputs and knowledge from **each one of us** made this project report possible to its rightful accuracy.

To all our colleagues who have helped us either directly or indirectly, we are grateful for their valuable inputs.

## **LITERATURE**

The goal of human resources analytics is to provide an organization with insights for optimum utilization and managing employees so that business goals can be reached quickly and efficiently. The challenge of human resources analytics is to identify what data should be captured and how to use the data to model and predict capabilities, so the organization gets an optimal return on investment on its human capital.

Retaining key employees is a major stake for any organization. But are there reliable ways to figure out if and why the best and most experienced employees are leaving prematurely? Most firms these days are already integrating the benefits of using analytics to introduce special efforts in regaining employees as well as hiring decisions. Lot of factors play key role in identifying significant predictors offering insights and meaning that can be interpreted using a statistical model language like R.

In our project, we have used HR Analytics dataset from Kaggle that is fictitious in nature seemingly because no company will share its personally identifiable record.

## **BACKGROUND**

**Data set**

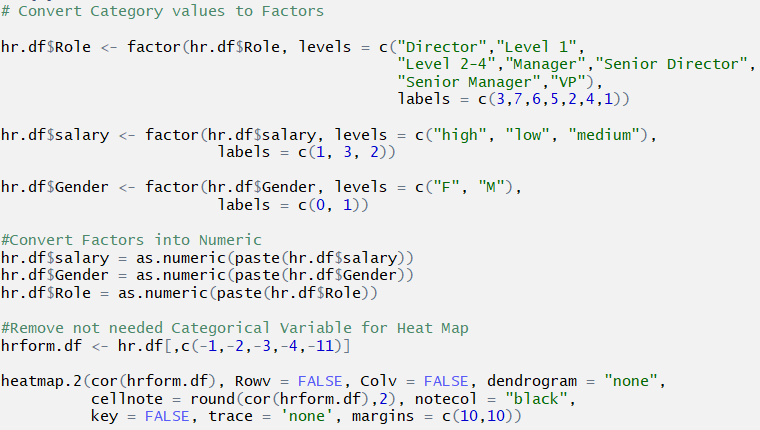
Our data set represents 14,999 employees and is composed of both currently employed and people who have already left the company with 30 variables defining the best possible way to answer the below questions and insights.

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 five other parameters from the file.

To improve the correlation significance between various predictors, we made changes against few variables. (*Rising\_Star, Left\_Company, promotion\_last\_5years, time\_spend\_company, Salary, Emp\_Satisfaction*)

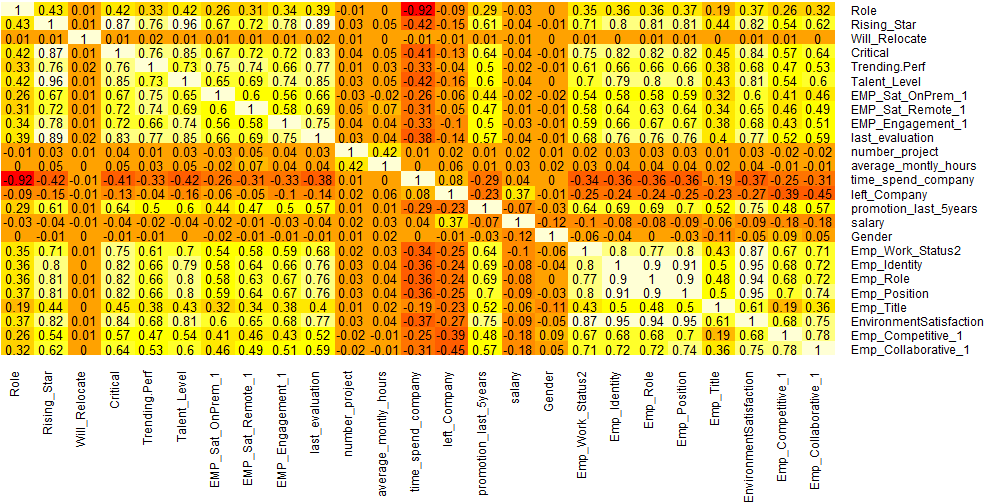
Correlation matrix before and after making changes to the dataset.

Before running the correlation, it was imperative to convert all the category variable values to factors and from factors to Numeric.



Correlation run before making changes to the dataset

Correlation run before after changes to the dataset





## **OBJECTIVES**

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

* Identify the primary reasons for employees leaving
* Why do good employees leave?
* Will the employee leave the company?
* What is the likelihood of Employee getting a promotion?
* How much time will the employee spend in company?
* How satisfied are the employees in company?

## **DATA EXPLORATION**



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

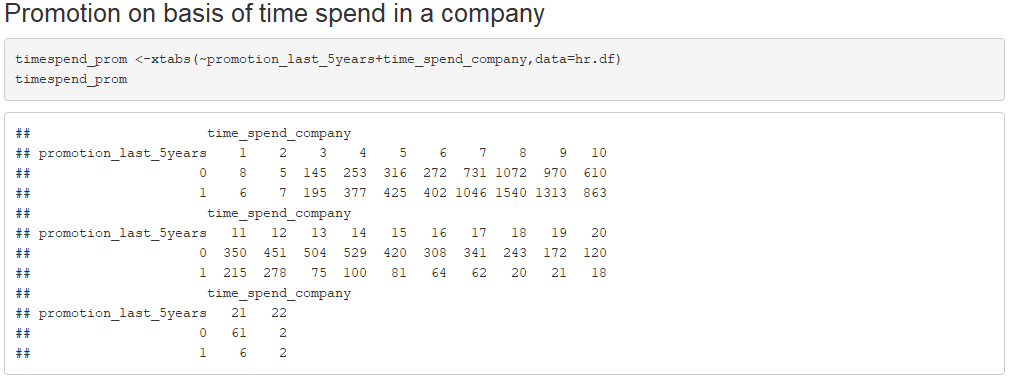
## Mean : 5.608 Mean : 4.998 Mean : 5.938

## 3rd Qu.: 8.000 3rd Qu.: 8.000 3rd Qu.: 9.000

## Max. :10.000 Max. :10.000 Max. :10.000

Meta Data

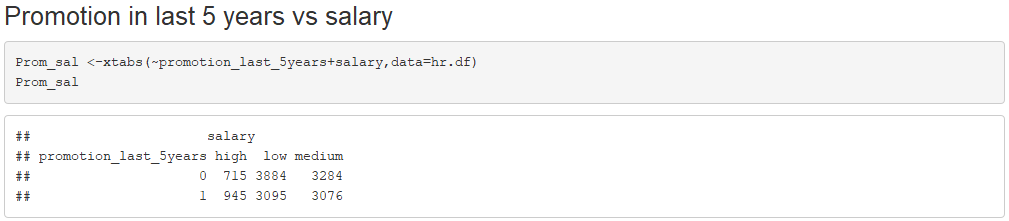
|  |  |
| --- | --- |
| **Attribute** | **Description** |
| ID | Employee ID |
| Name | Employee Name |
| Department | Department |
| GEO | Geographical location |
| Role | Current Role or title of employee |
| Rising Star | Indicates the level of promise or promote-ability the employee has. Scale(1-5) |
| Will\_Relocate | Is the employee willing to relocate? 0- No, 1- Yes |
| Critical | Is the employee critical to the organization? 0- No, 1- Yes |
| Trending Perf | How is the employee trending in performance this year? Scale (1-10) |
| Talent\_Level | This field represents a subjective level of management's view of the employee. Scale (1-10) |
| Percent\_Remote | The percentage of the employee's work that is done remotely. |
| EMP\_Sat\_OnPrem\_1 | One indicator from a survey that was sent to employees. On prem (On premise)  means that the employee maintains a high percentage of work on the corporation’s  physical work locations. Scale (1-10) |
| EMP\_Sat\_Remote\_1 | One indicator from a survey that was sent to employees. Remote (distance employee)  means that the employee does a high percentage of work away from the  corporation’s physical work locations. Scale (1-10) |
| EMP\_Engagement\_1 | One indicator from a survey that was sent to employees. Engagement represents the  employee's feeling about how they feel about being engaged in company activities. Scale(1-5) |
| last\_evaluation | The score on the last employee evaluation.Scale (1-10) |
| number\_project | The number of projects the employee works on throughout the year. |
| average\_montly\_hours | The average number of hours the employee works monthly. |
| time\_spend\_company | Years of service |
| left\_Company | Did the employee leave the company? 0- No, 1- Yes |
| promotion\_last\_5years | Did the employee get promoted in last 5 years? 0- No, 1- Yes |
| salary | Relative pay grade (low, medium, high) by role. |
| Gender | M or F |
| Emp\_Work\_Status2 | One indicator from a survey that was sent to employees. Status represents how  strongly employee feels about their status level in the organization. Scale (1-10) |
| Emp\_Identity | How the employee identifies themselves with the company. Scale (1-10) |
| Emp\_Role | How the employee identifies themselves with the importance of their role in the company. Scale (1-10) |
| Emp\_Position | How the employee identifies themselves with the importance of their position in the company. Scale (1-10) |
| Emp\_Title | How the employee feels about their title.Scale (1-10) |
| Emp\_Satisfaction | Average value of the above 5 variables. Scale out of 1-10 |
| Emp\_Competitive\_1 | One indicator from a survey that was sent to employees. How employee feels  about the competitive nature of work in the organization. Scale (1-10) |
| Emp\_Collaborative\_1 | One indicator from a survey that was sent to employees. How employee feels  about the collaborative nature of work in the organization.Scale (1-10) |



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.

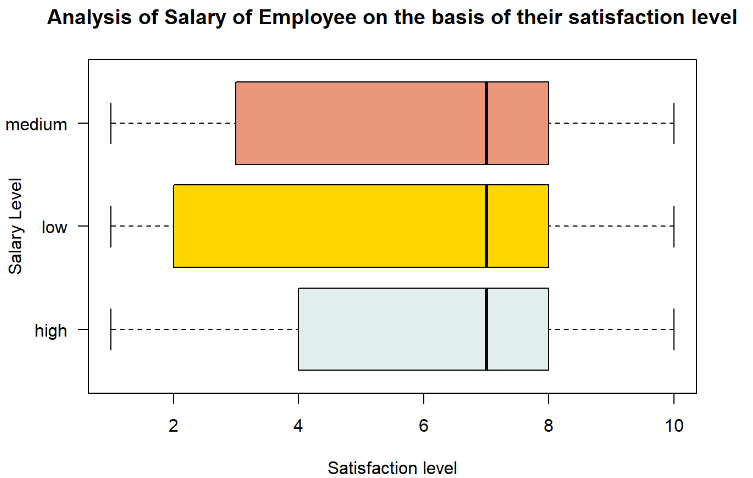
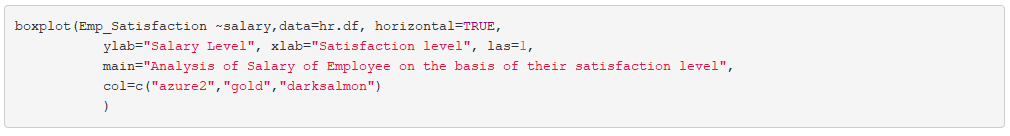


The finance department has the highest number of high-wage workers whereas the warehouse department has the highest number of low-wage workers.



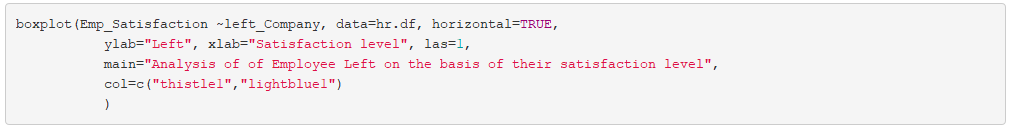
Employees getting the maximum promotions in the last 5 years have had a low to medium increase in their salary, with very few of them promoted with a high wage

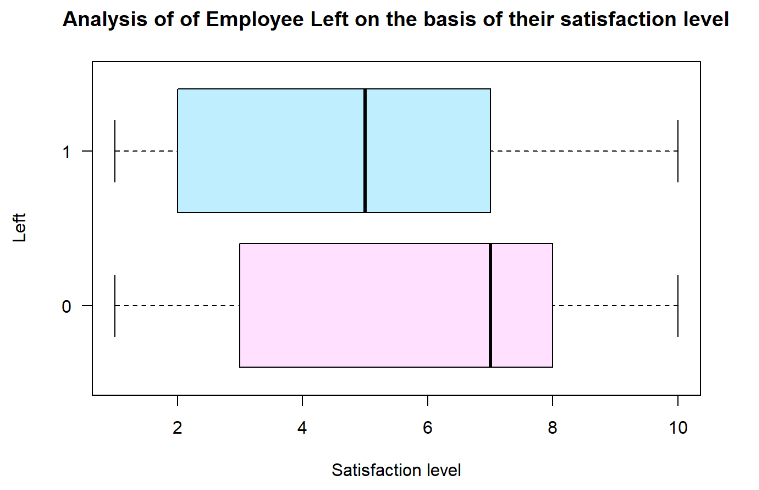
Box Plot describing relationship between Salary and Emp\_Satisfaction



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

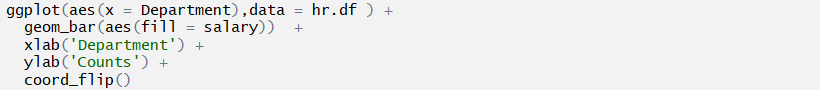
Box Plot describing relationship between Left\_company and Emp\_Satisfaction

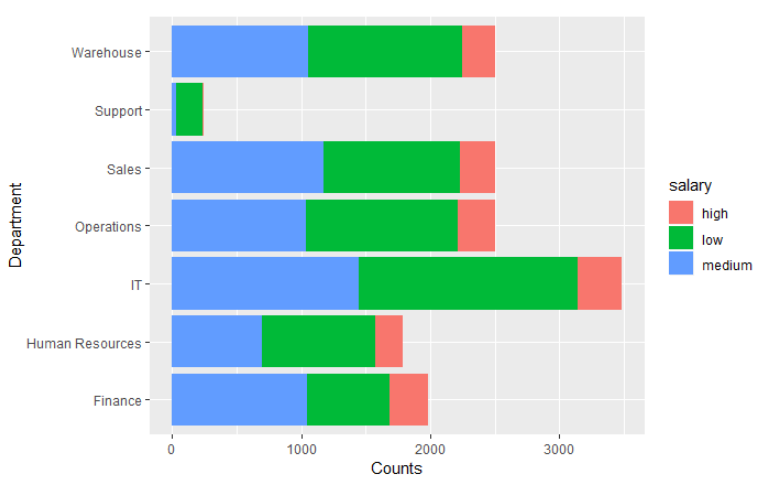




As it can be seen, employees with lower satisfaction levels tend to leave the company.

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

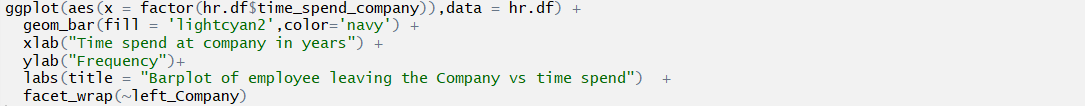


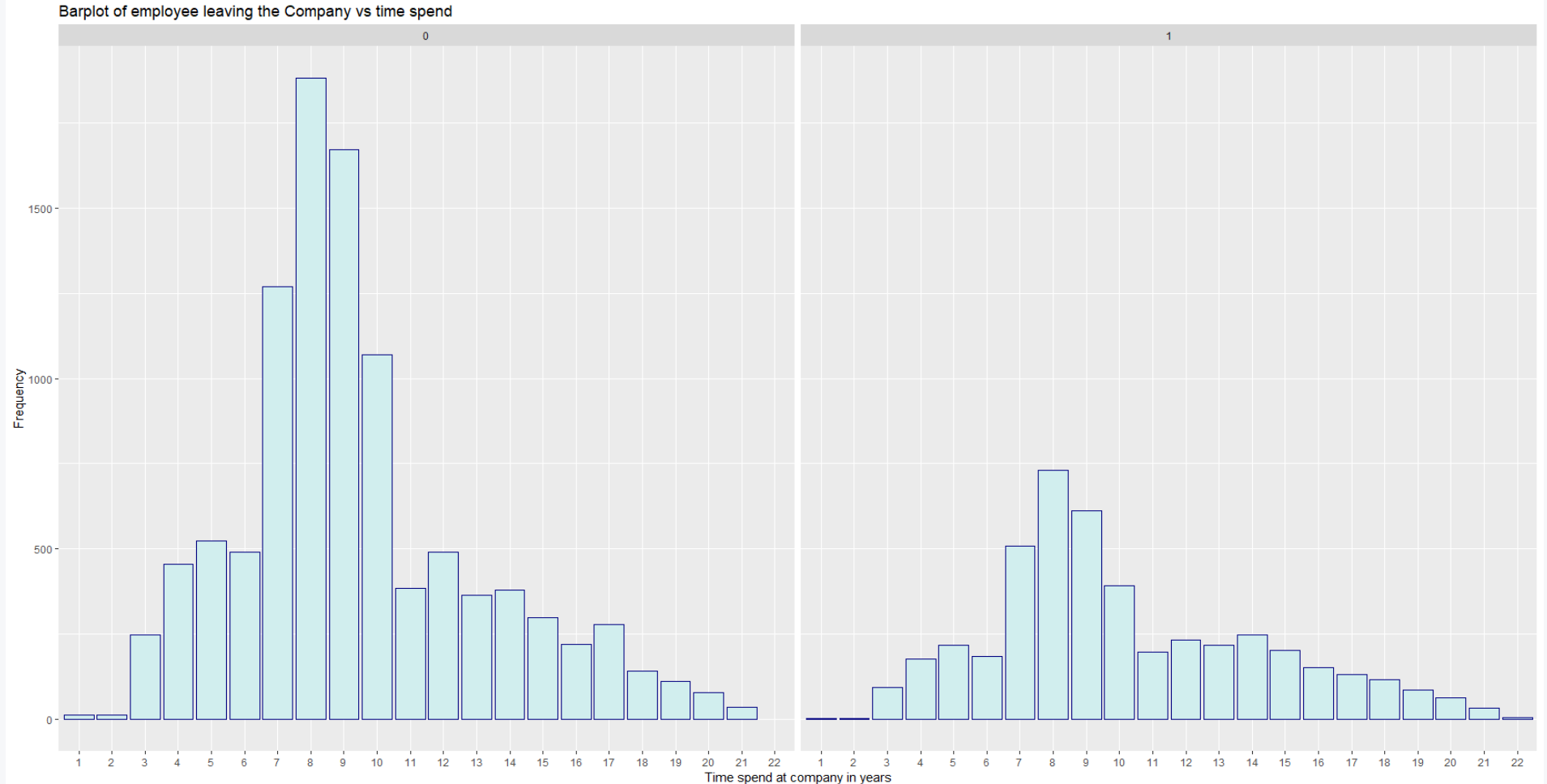


***Interpretation***

* 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

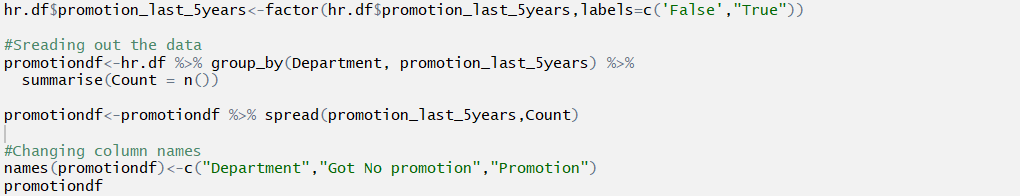


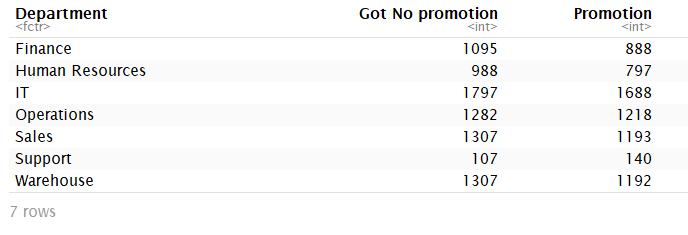


***Interpretation***

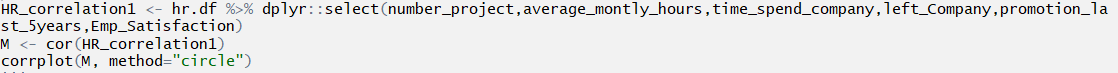
* From the second plot above that represents the employees having left the company, it is evident that employees tend to leave a company after spending 7-10 years with average being 8 years
* Very less number of employees leave the company within the first 2 years of joining
* There are employees who 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





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



***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.

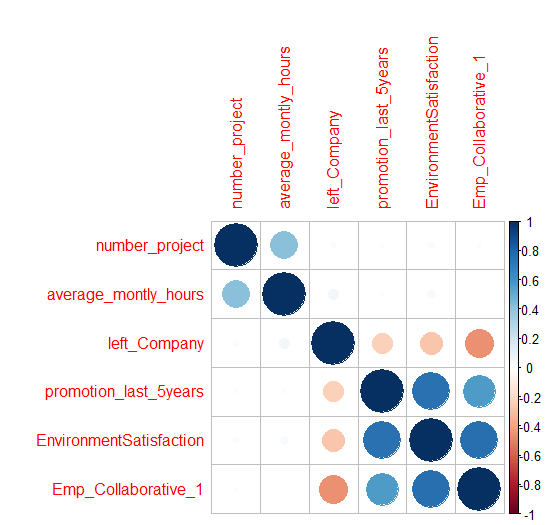
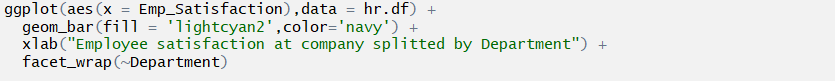
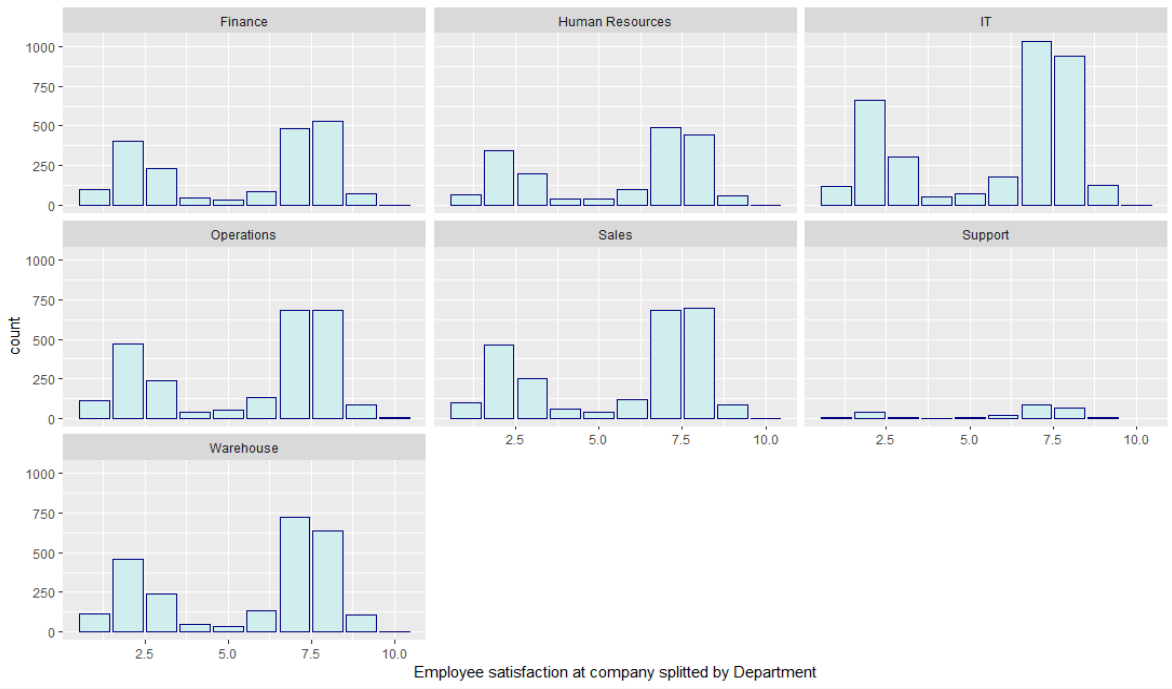
****

Table showing department wise Employee\_Satisfaction

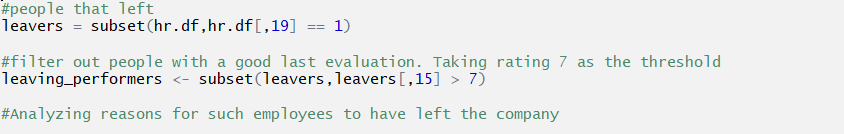




***Interpretation***

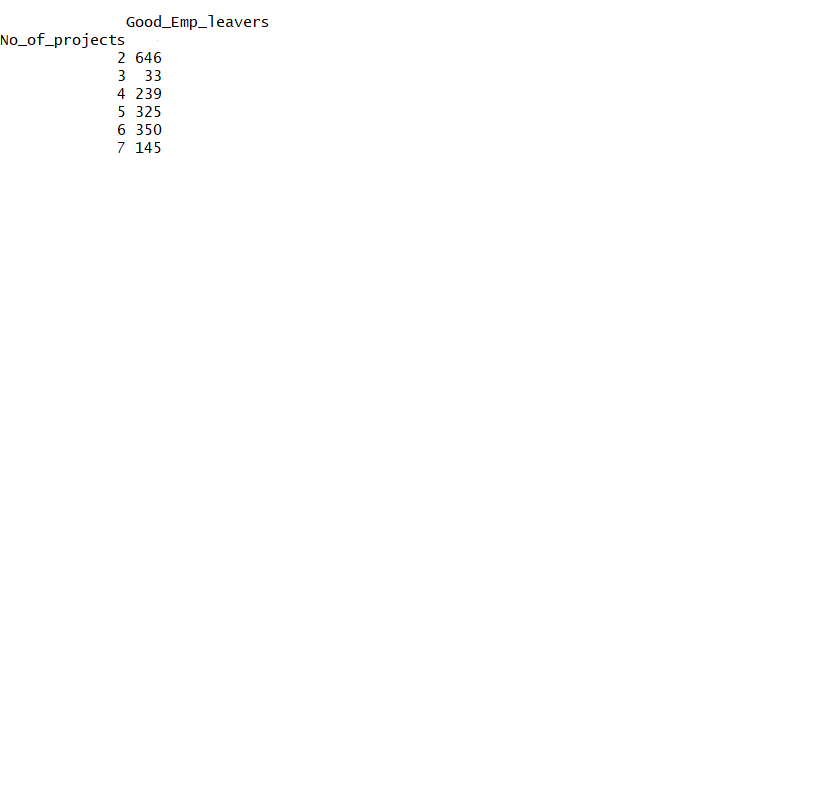
* IT department has got the most number of employees falling in both the categories(Satisfied and not satisfied) giving us takeaway that a high number of employees aren’t happy with their work.
* We see a bimodal barplot for across departments telling us that employees are either not satisfied; with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

## **WHY DO GOOD EMPLOYEES LEAVE?**



Are the number of projects employees assigned to the reason?

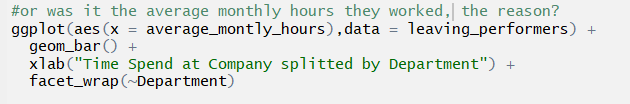


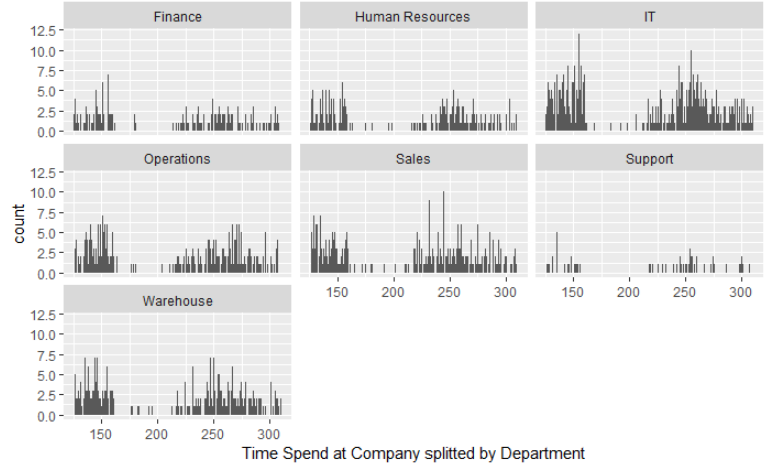


***Interpretation***

* The data shows that employees have left more when they were assigned to less number of projects.
* Probably , they felt that they were being under-utilized in the company and left the company.

Or the average monthly hours they work for across projects?

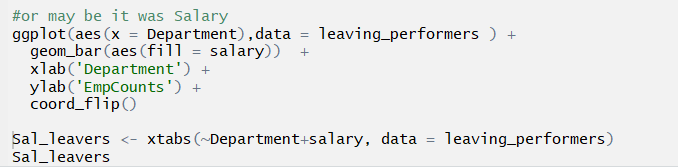


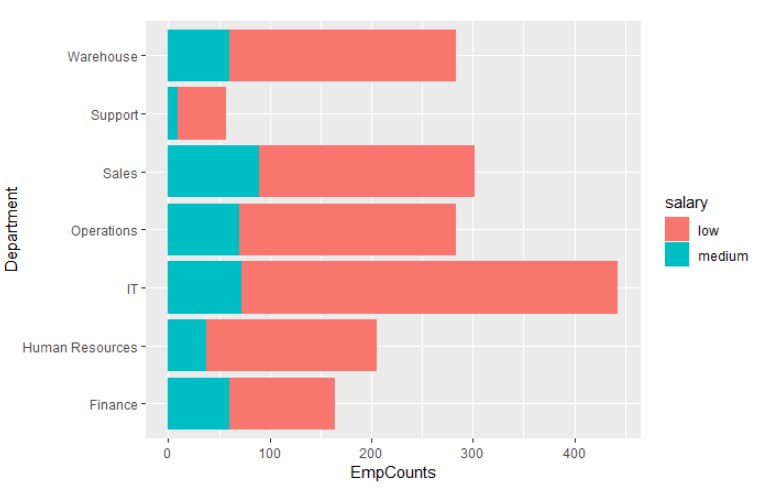


***Interpretation***

* 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 .

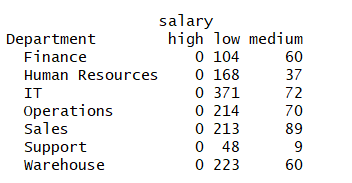
Probably salary could reveal more?





***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.



**Conclusion is that these employees are highly valuable assets that should not have been lost.**

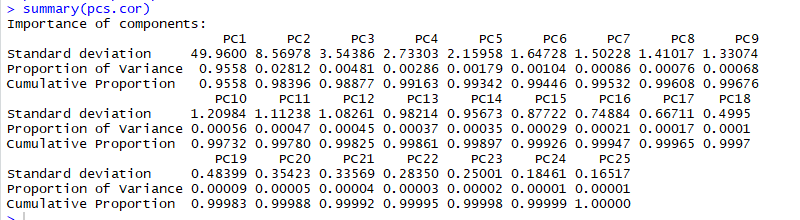
## **MODEL ANALYSIS**

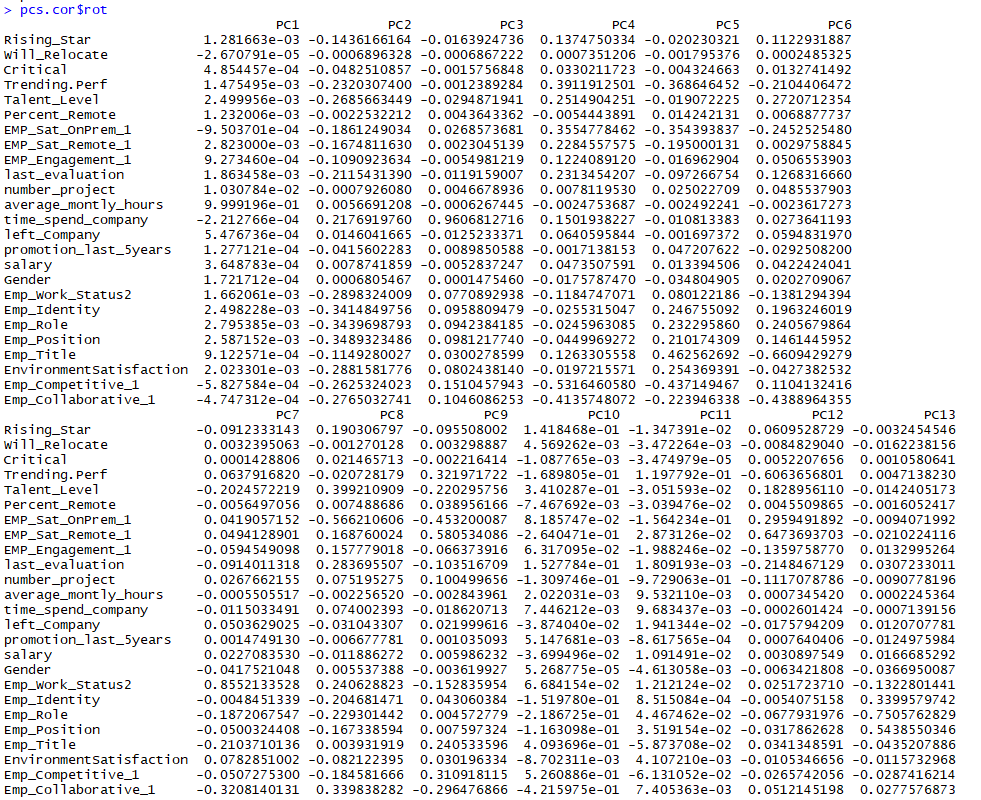
After running descriptive diagnostics on the data, we move on to predictive analytics. In this section we aim to answer the questions that will help the management to mitigate the attrition rate of employees. This analysis is important in the sense that it assists HR personnel to analyze the factors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

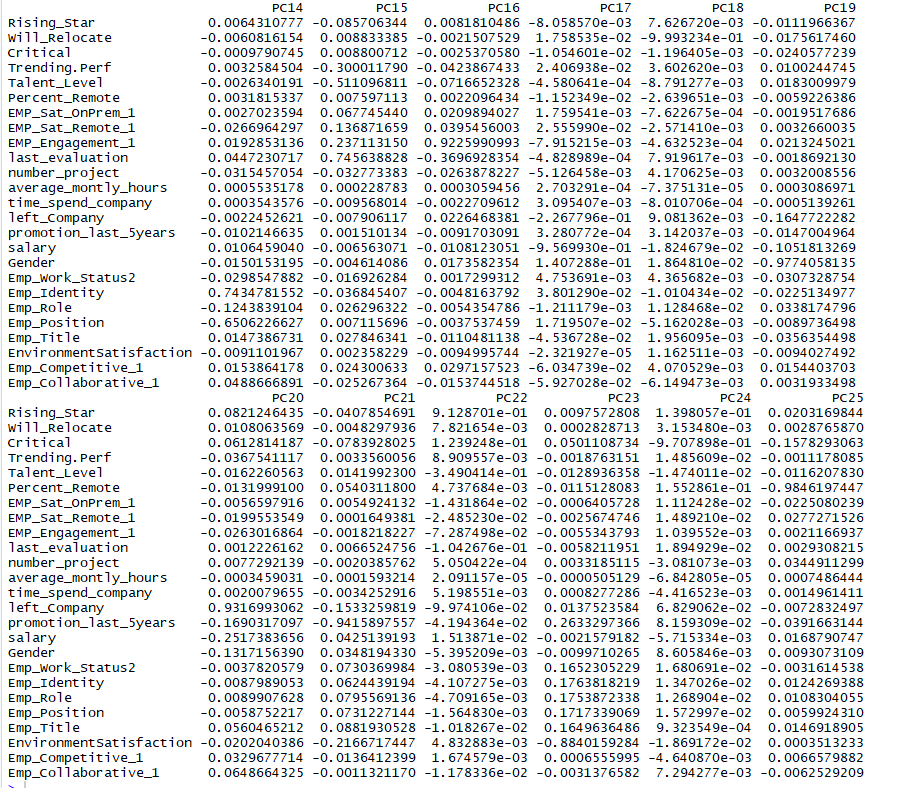
Principal Component Analysis:

The central idea of using principal component analysis (PCA) in our project is to reduce the dimensionality of the HR Analytics data set, which consists of many interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by normalizing the data and transforming to a new set of variables, the principal components (PCs), which are uncorrelated.

C:\Users\user\Desktop\Capture.PNG







*Interpretation:*

After running PCA, we find that the first PC retained almost 95.4% of the variation present in all the original variables. Also, in PC1, average\_montly\_hours is the most significant variable .

* **Strengths:** PCA is a versatile technique that works well in practice. It's fast and simple to implement, which means you can easily test algorithms with and without PCA to compare performance. In addition, PCA offers several variations and extensions (i.e. kernel PCA, sparse PCA, etc.) to tackle specific roadblocks.

* **Weaknesses:** The new principal components are not interpretable. In addition, you must still manually set or tune a threshold for cumulative explained variance. PCA simply ranks attributes by the total amount of variance that each variable contributes. So a very noisy attribute could overshadow more useful, better structured data. For building predictive models, you want to gather variables that contain more information, not more noise.

### Running various models to answer the below questions-

### Will the employee leave the company?

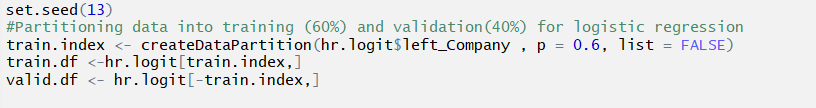
**1.Running Logistic Regression**

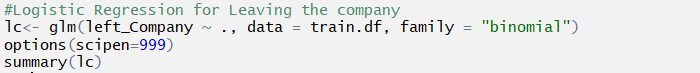
Logistic regression extends the idea of linear regression to situation where outcome variable is categorical. It is widely used , especially where a structured model is used to explain or predict.

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”.

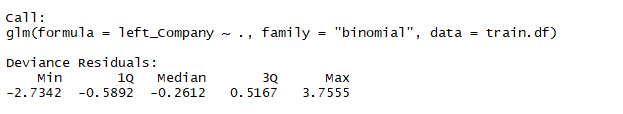


The model is trained on test data that comprises 60% of the total data and validated on the rest.

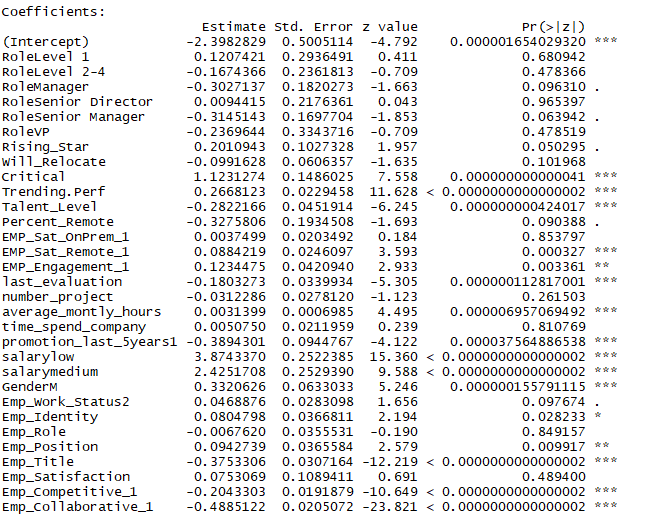


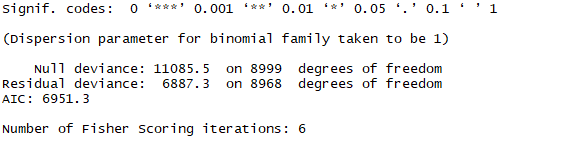


Output:



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.





*Interpretation:*

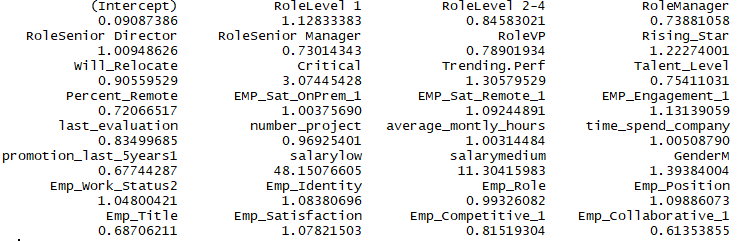
Three stars indicate an extremely low P value (approximately 0), it signifies that probability of a dependent variable occurring in a certain way in accordance with the corresponding dependent variable is very low. This suggest that there is relationship between two variables in a way that independent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

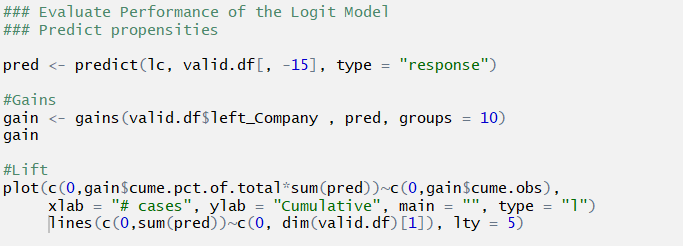
Based on the above summary and P-values of coefficient estimates it can be concluded that following predictors are important in deciding whether the employee will or will not leave the company. “Critical”, “Trending.perf”, “Talent Level”, “EMP\_Sat\_Remote\_1”, “EMP\_Engagement\_1”, “last\_evaluation”, “average\_montly\_hours”, “promotion\_last\_5years1”, “salarylow”, “salarymedium”, “GenderM”, “Emp\_Position”, “Emp\_Title”, “Emp\_Competitive\_1” and “Emp\_Collaborative\_1”



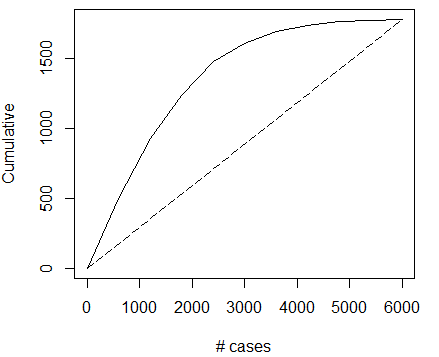


From the above values it is evident that Low salary has the highest impact on employees leaving the company

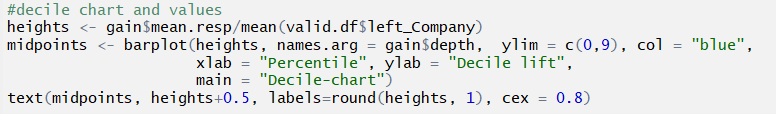
followed by medium salary and criticalness.



Lift Chart



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 straight line.



* Decile chart follows an ideal structure

representing maximum variation covered in initial deciles.

* First 5 deciles cover 90% of the

variation.

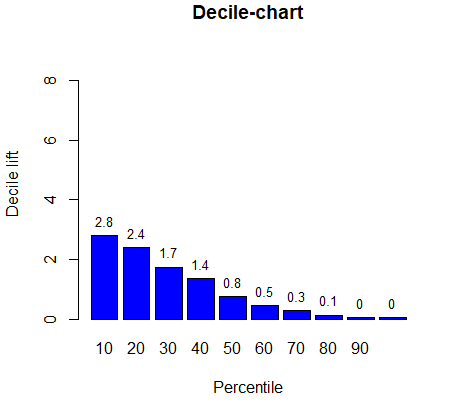
* This can be considered as good model where the deciles are decreasing in

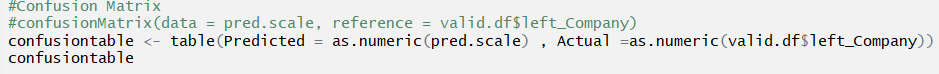
order from start to end.

* Looking at the first decile, we can say

that this model performs 2.8 time

better than the one with Naïve rule.











* **Strengths:** Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid over fitting. Logistic models can be updated easily with new data using stochastic gradient descent.

* **Weaknesses:** Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships. Logistic regression attempts to predict outcomes based on a set of independent variables but if we include the wrong independent variable, the model will have little to no predictive value. Logistic regression also works well for predicting categorical outcomes but cannot predict continuous outcomes.

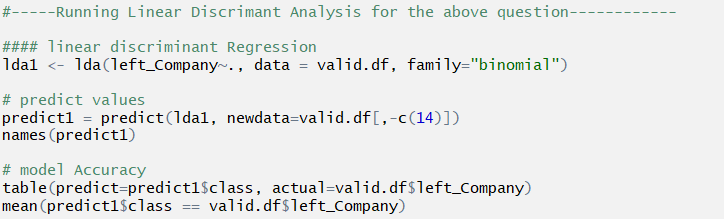
**2.Running Linear Discriminant Analysis for the same question and comparing which model is better**

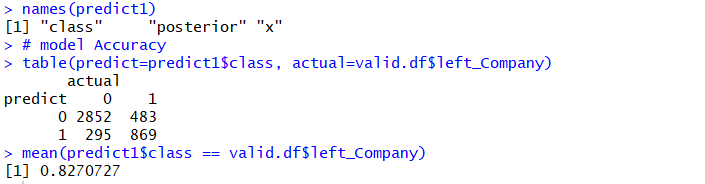
Discriminant Analysis is a classical statistic technique used for classification. It also has business data applications and can be used for profiling. Linear discriminant analysis is used to find a linear combination of the predictors that gives maximum separation between the centers of the data . It is also used to minimize the variation within each group of data.

We use the lda() function to perform linear discriminant analysis in R . It finds directions that maximize the separation between the classes , then uses these directions to predict the classifications . These linear directions are linear combinations of predictor variables.

Assumptions :

* Predictors are normally distributed
* Different classes have class-specific means and same variance/covariance structure





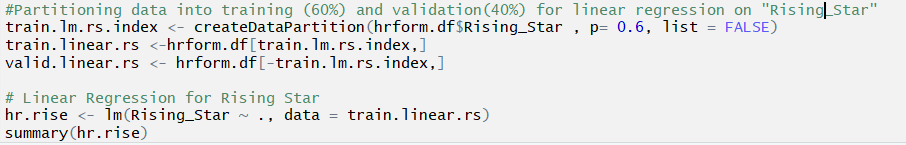
*Interpretation:*

### What is the likelihood of employees getting a promotion?

**1.Running Linear Regression**

Linear regression is the most basic and commonly used predictive analysis.  Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables.  At the center of the regression analysis is the task of fitting a single line through a scatter plot. It consists of 3 stages:  1) analyzing the correlation and directionality of the data, 2) estimating the model, i.e., fitting the line, and 3) evaluating the validity and usefulness of the model.

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.



*Interpretation:*

The significant coefficients (P Value two and three stars) for Rising\_Star are:

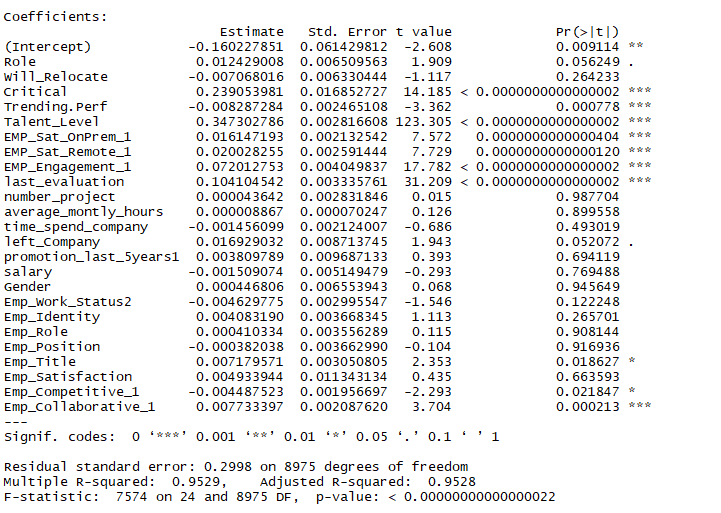
Critical: Positive coefficient signifies that if the employee is critical ( “1” ) the likely hood of promotion (“Rising\_Star) also increases in number (1 through 5). For every one-unit change in Critical value, the independent variable is affected to change +0.239

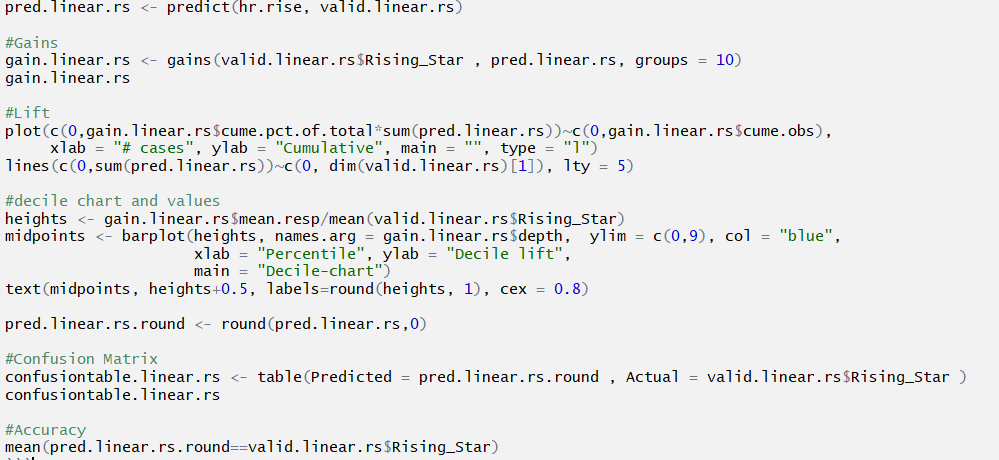
Trending.perf: For every unit change in Trending.perf, there is negative 0.0082 effect on Rising\_Star.

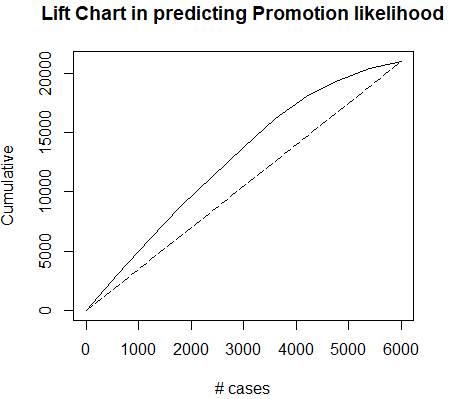
Talent Leve: For every unit change in Trending.perf, there is positive 0.3473 effect on Rising\_Star.

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.

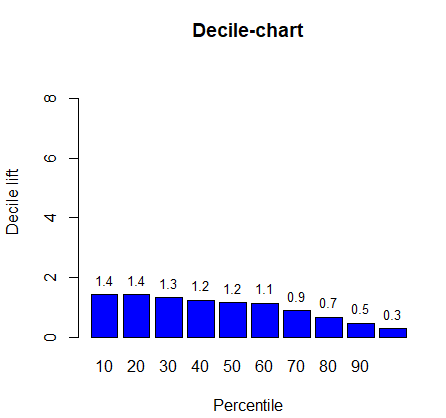
Adjusted R square value of 0.9528 can be considered as an excellent number exhibiting that approximately 95% of the variation in Rising\_Star variable is captured by the input variables.





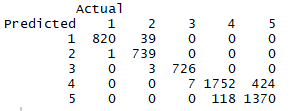


* As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.



* 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 1.4 time better than the one with Naïve rule.

Confusion Matrix



Accuracy in predicting Validation data set



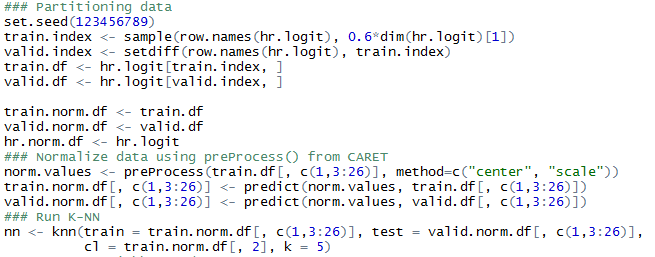
* **Strengths:** Linear regression is straightforward to understand and explain, and can be regularized to avoid over fitting. In addition, linear models can be updated easily with new data.
* **Weaknesses:**Linear regression performs poorly when there are non-linear relationships. They are not naturally flexible enough to capture more complex patterns, and adding the right interaction terms or polynomials can be tricky and time-consuming.

**2.Running Knn model for the same question and comparing which model is better**

KNN is used to classify or predict a new record based on similar records in the training data . It is a non-parametric method , it is data driven . There are no parameters to estimate as in linear regression. It is based on distance between records.

Rising star indicates the level of promise or promote-ability the employee has. Scale(1-5) 5 being the highest and 1 lowest



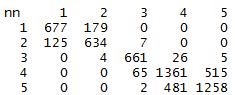


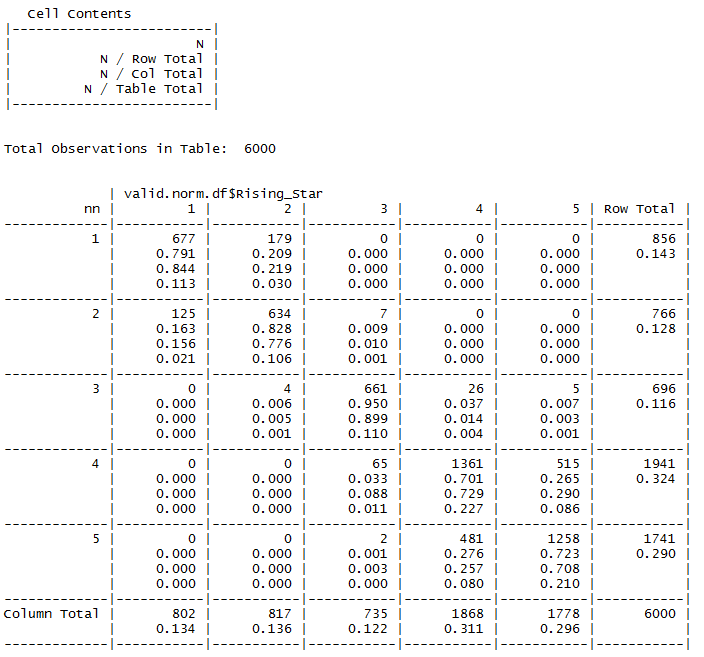


output: character(0)



output:





*Interpretation:*

accuracy = 0.78

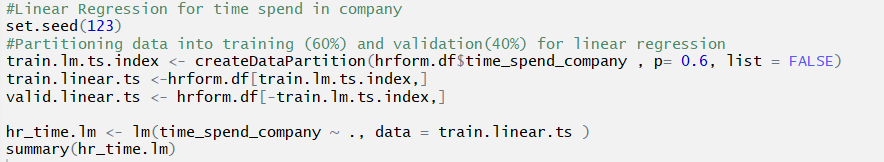
Using the 5 nearest records to predict the employee’s rising star value, the accuracy is 0.78. The prediction behaves better when employee has a low rising star score especially when their rising star is 3. It’s easy to find that 2 is more likely to be confused with 1 and 4 with 5. On the ground of that we could divide the employees into 3 parts which represent high, medium and low rising star or expectation in other words.

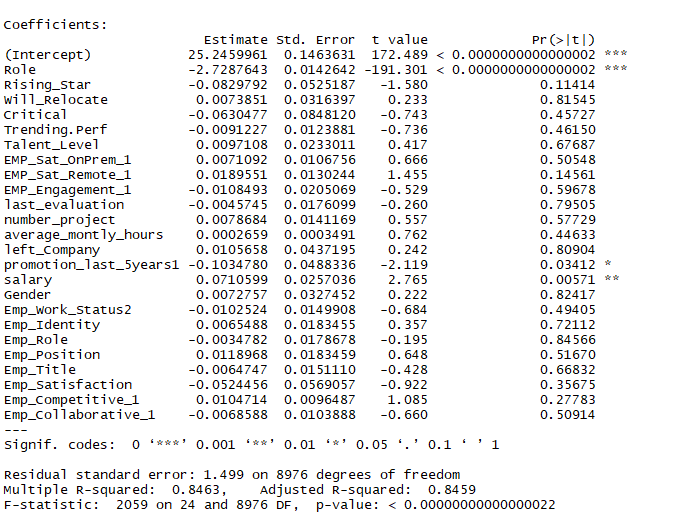
* **Strengths:**K-Means is the most popular clustering algorithm because it's fast, simple, and flexible.
* **Weaknesses:** The user must specify the number of clusters, which won't always be easy to do. In addition, if the true underlying clusters in your data are not globular, then K-Means will produce poor clusters.

### How much time will the employee spend in company?

**Running Linear Regression**

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.

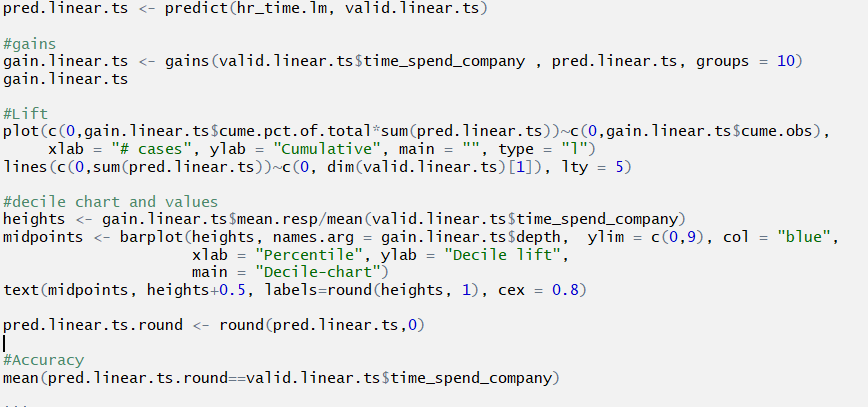


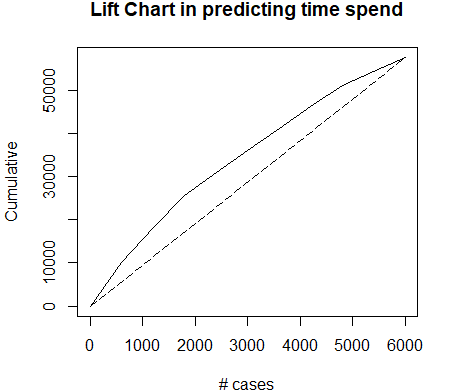


*Interpretation:*

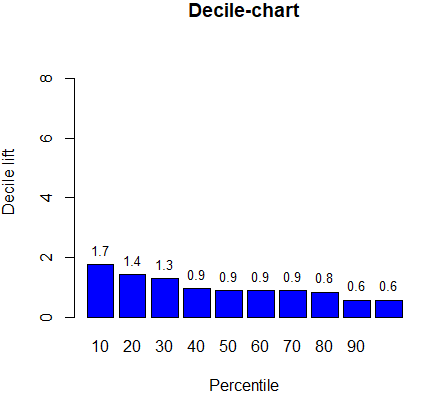
The significant coefficients (P Value one, two and three stars) for time\_spend\_company are Role, promotion\_last\_5years1 and salary.

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.





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.



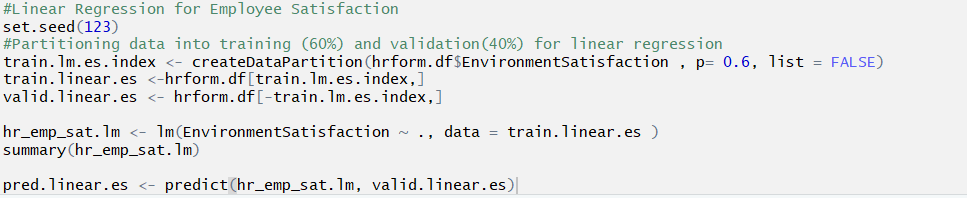
Decile chart follows an ideal structure representing maximum variation covered in initial deciles. This can be considered as good model where the deciles are decreasing in order from start to end. Looking at the first decile, we can say that this model performs 1.7 time better than the one with Naïve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

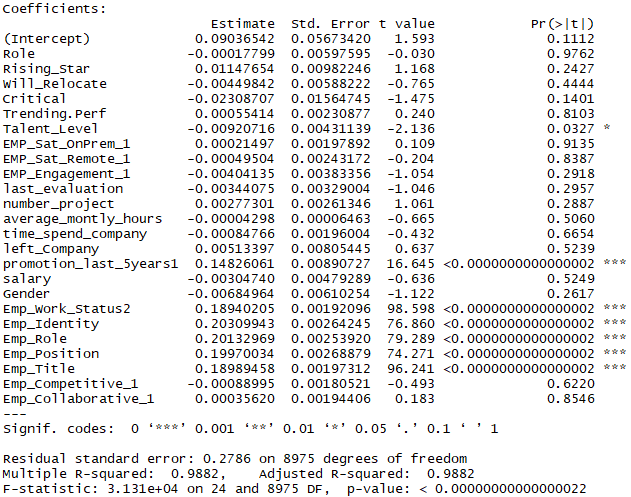
### How satisfied are the employees in company?

**Running Linear Regression**

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.

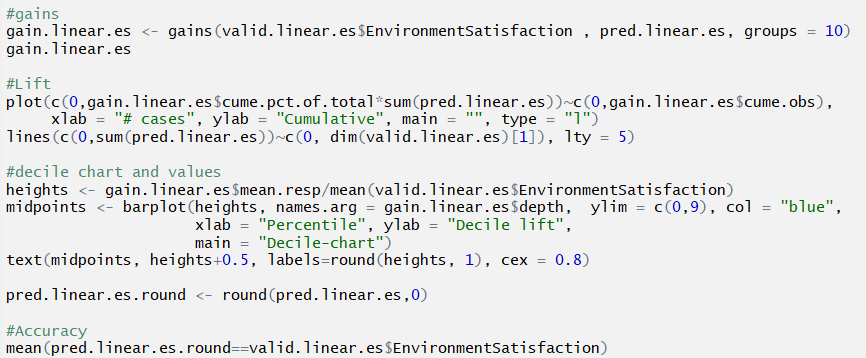


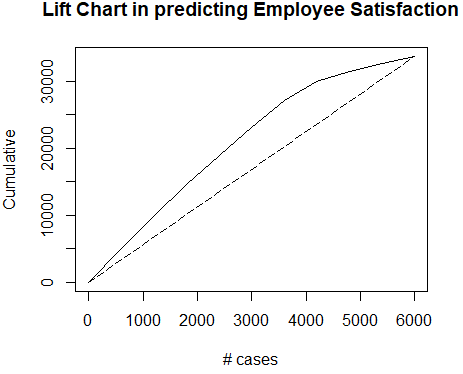


*Interpretation:*

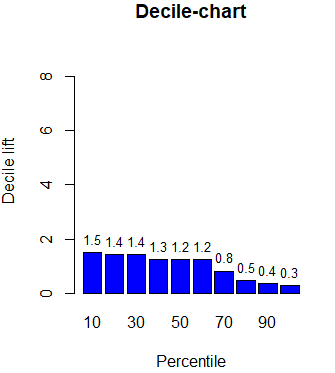
The significant coefficients (P Value one, two and three stars) for Emp\_Satisfaction are Talent\_Level, promotion\_last\_5years1, Emp\_work\_Status2, Emp\_Identity, Emp\_Role, Emp\_Position and Emp\_Title.

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.





* 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.



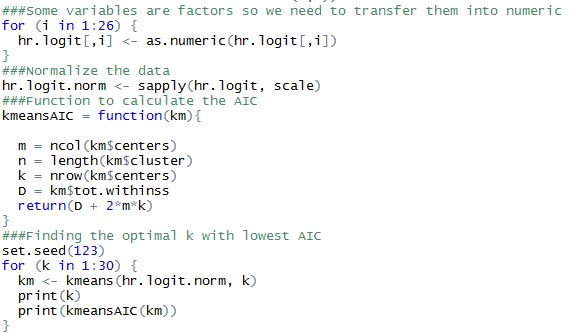
* Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
* This can be considered as good model where the deciles are decreasing in order from start to end.
* Looking at the first decile, we can say that this model performs 1.5 time better than the one with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

Accuracy in predicting the Employee satisfaction in Validation data set

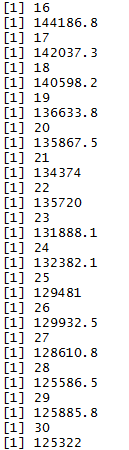
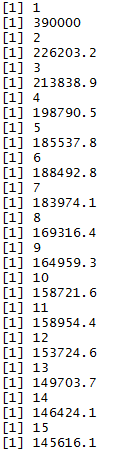




### Running K-mean Clustering to find out which set of employees are more likely to exit

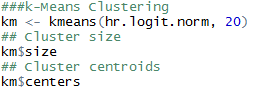


Output:



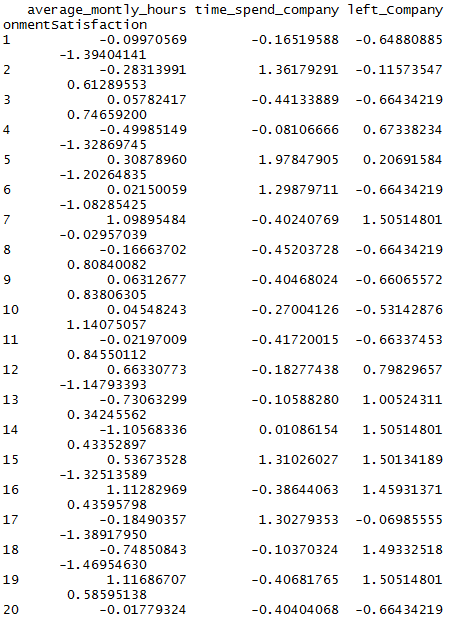
*Interpretation:*

AIC gets smaller when k increases and 20 is the optimal k since the AIC decreases much slower after it. That means the employees would be divided into 20 clusters.



output:





The employees can be divided into 20 cluster and according to the centroids of each clusters, it seems that employees in cluster 7, 13, 14, 15, 16, 18, 19 are more likely to leave the company. By using km$cluster, we could identify every employee’s group number and predict if they are more likely to leave the company.

## **REFERENCES & CITATIONS**

Website

Elite Data Science. <https://elitedatascience.com/machine-learning-algorithms>

The Classroom. https://www.theclassroom.com/disadvantages-logistic-regression-8574447.html