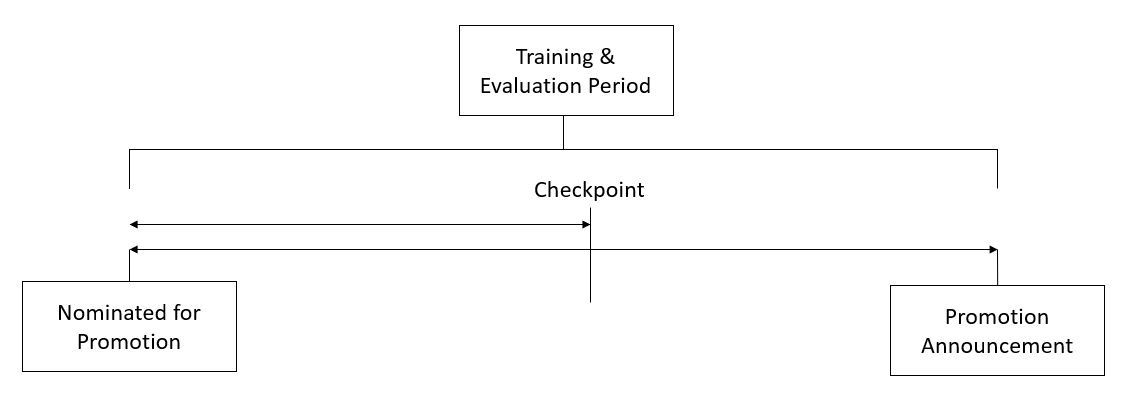
**Problem Statement**

Your client is a large MNC and they have 9 broad verticals across the organization. One of the problems your client is facing is around identifying the right people for promotion (only for manager position and below) and prepare them in time. Currently the process, they are following is:

They first identify a set of employees based on recommendations/ past performance

Selected employees go through the separate training and evaluation program for each vertical. These programs are based on the required skill of each vertical

At the end of the program, based on various factors such as training performance, KPI completion (only employees with KPIs completed greater than 60% are considered) etc., employee gets promotion

For above mentioned process, the final promotions are only announced after the evaluation and this leads to delay in transition to their new roles. Hence, company needs your help in identifying the eligible candidates at a particular checkpoint so that they can expedite the entire promotion cycle. 

They have provided multiple attributes around Employee's past and current performance along with demographics.

**Project Plan**

|  |  |
| --- | --- |
| Stelevel p 1 | Business Problem |
| Step2 | Import Data |
| Step 3 | EDA |
| Step 4 | Model Building |
| Step 5 | Model Evaluation |
| Step 6 | Deployment |
| Tools Used | R |
| Package Used |  |

**Business Problem: -** To predict whether a potential promote at checkpoint in the test set will be promoted or not after the evaluation process.

**Dataset Details: -** The following are the features and target variable available in the datasets

|  |  |  |
| --- | --- | --- |
| **Input Variables (x)** | employee\_id | Unique ID for employee |
| department | Department of employee |
| region | Region of employment (unordered) |
| education | Education Level |
| gender | Gender of Employee |
| recruitment\_channel | Channel of recruitment for employee |
| no\_of\_trainings | no of other trainings completed in previous year on soft skills, technical skills etc. |
| age | Age of Employee |
| previous\_year\_rating | Employee Rating for the previous year |
| length\_of\_service | Length of service in years |
| KPIs\_met >80% | if Percent of KPIs(Key performance Indicators) >80% then 1 else 0 |
| awards\_won? | if awards won during previous year then 1 else 0 |
| avg\_training\_score | Average score in current training evaluations |

|  |  |  |
| --- | --- | --- |
| **Dependent Variable**  **Or**  **Target Variable (y)** | is\_promoted | (Target) Recommended for promotion |

**Exploratory Data Analysis:**

Train and Test data sets is separately given,

**For train, no. of rows given is 54808**

From the above input variables, ***employee\_id*** is irrelevant as in this case there is no impact of employee id on the employee being promoted or not.

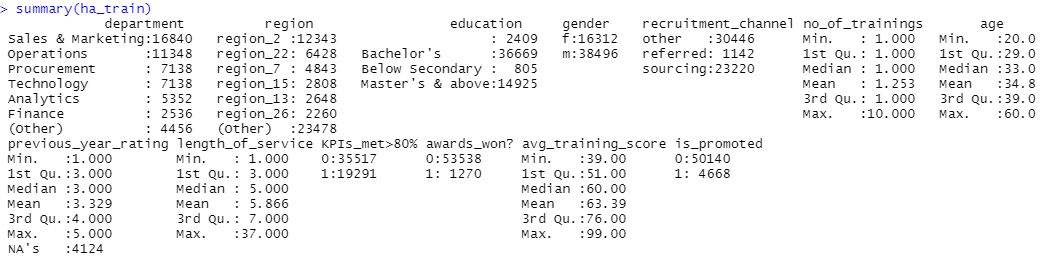
So, this feature is not considered for further analysis.

Converting the variables employee\_id, department, region, education, gender, recruitment\_channel, awards\_won., KPIs\_met..80., is\_promoted to factor variable

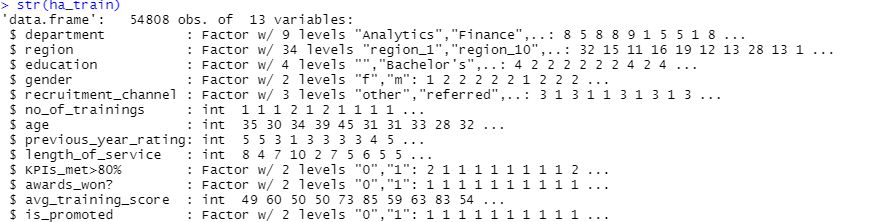
awards\_won. Variable name is changed to awards\_won?

KPIs\_met..80. Variable name is changed to KPIs\_met>80%

**Overall data Summary**



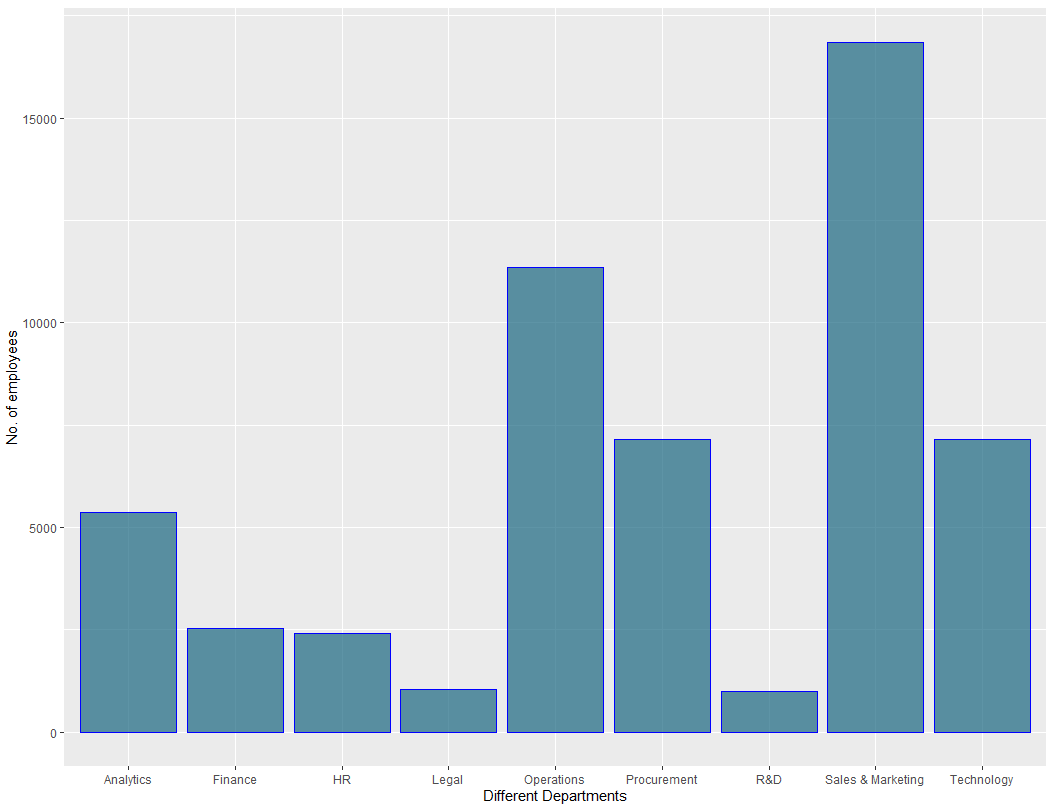
The above table shows the summary of train data.



The above snapshot from R studio represents about the datatypes of all the feature class and target variables.

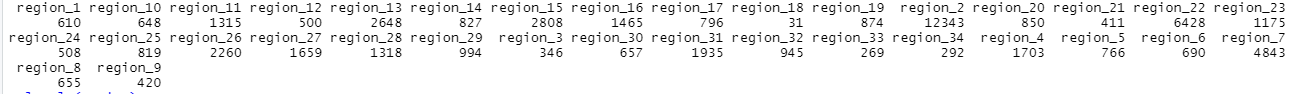
Graphical Representation

1. ) Department



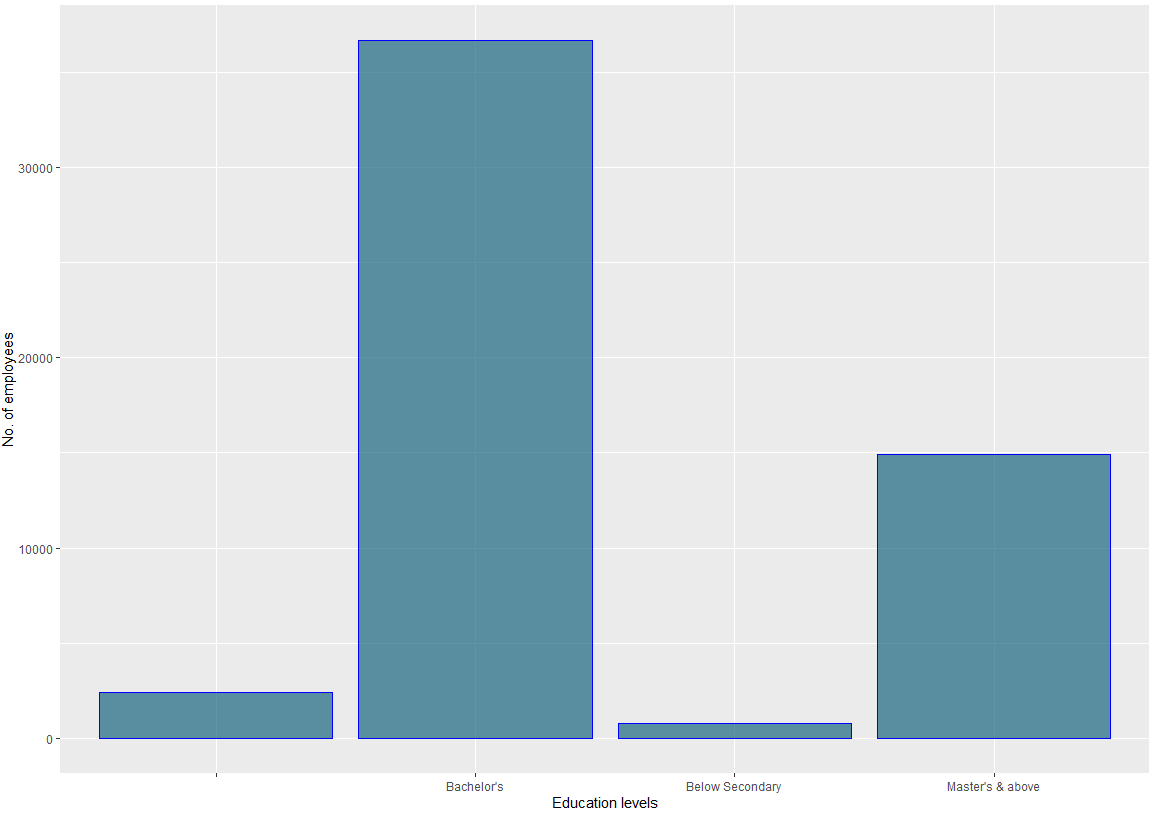
The above plot shows the Number of employees in different department. Highest number of employee is from sales and marketing department whereas lowest is from Legal department.

2.) Region



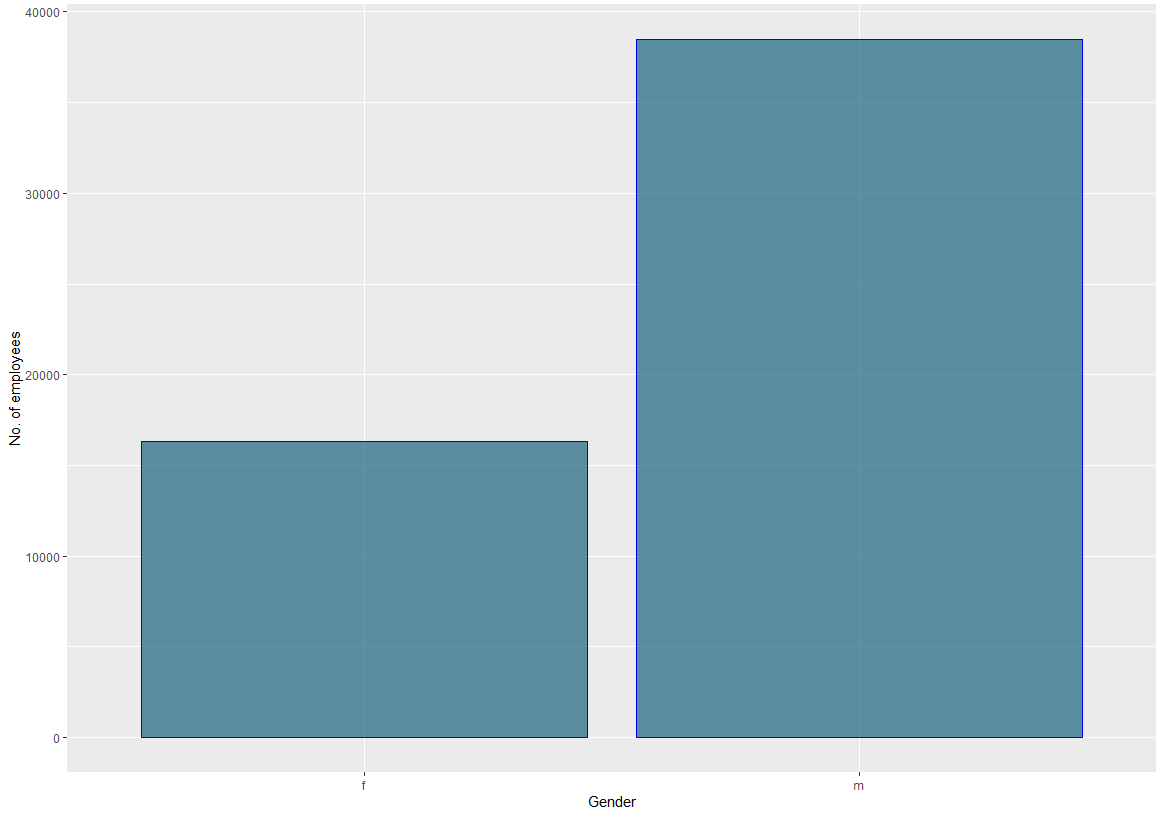
There are 34 different regions from where the employee belongs.

3) For education



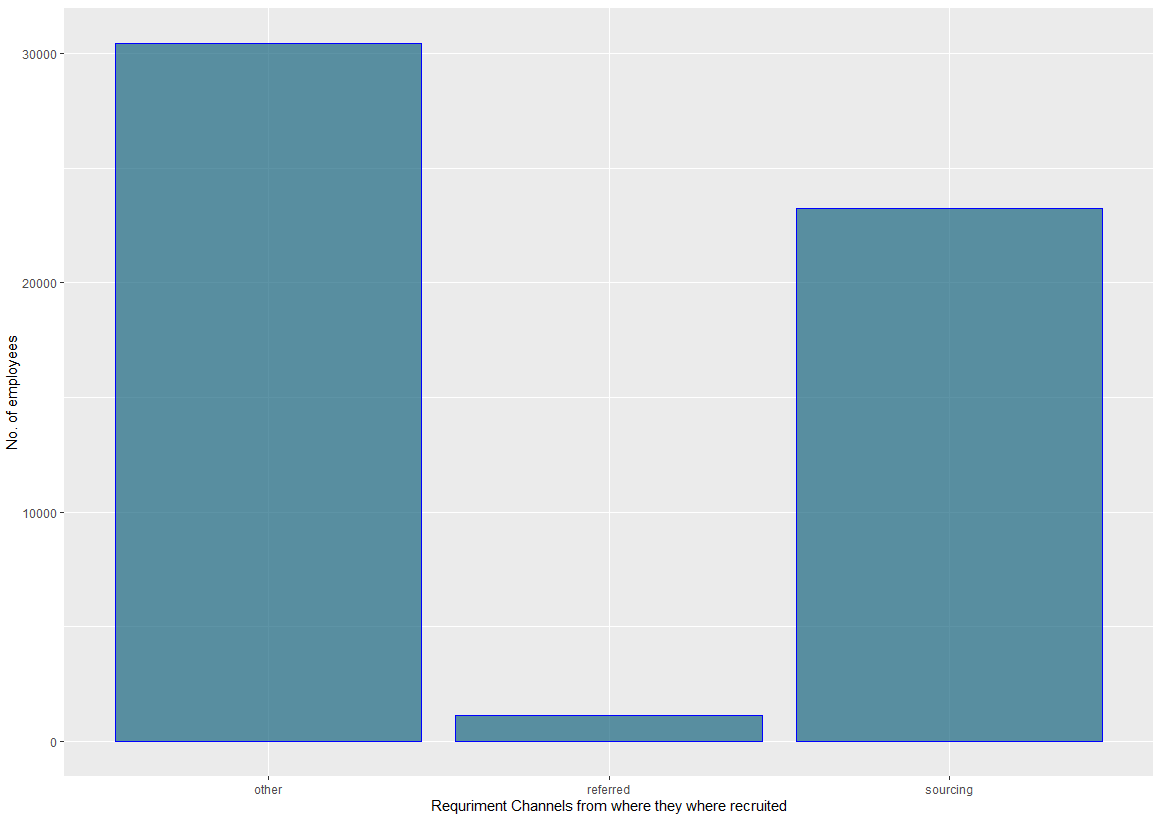
Form the above plot, Maximum people is from Bachelor’s degree and less number of people belong to below Secondary education. There are some people whose education level is not mentioned in the data. Need to impute those value which is done separately in missing value treatment.

4) For gender



From the above plot, male employees are more than female employees.

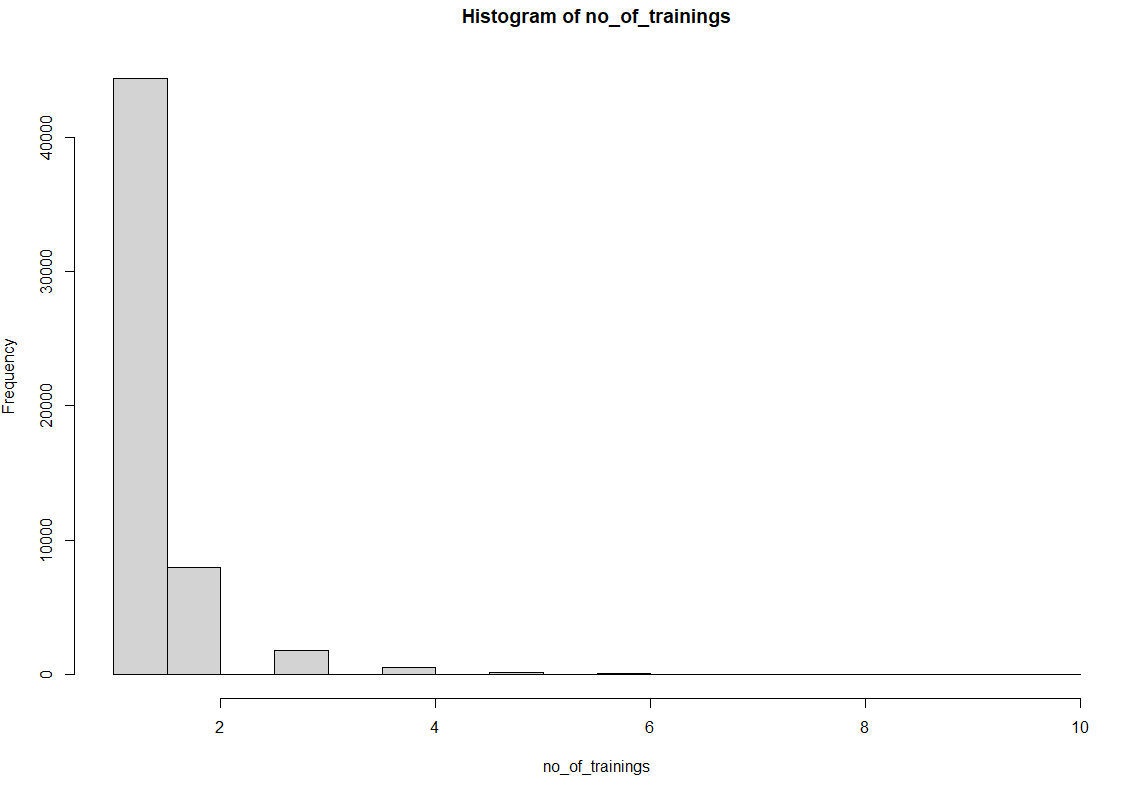
5) For recruitment\_channel



Less number of employees are recruited through reference whereas more number of employees are recruited through Other Channels.

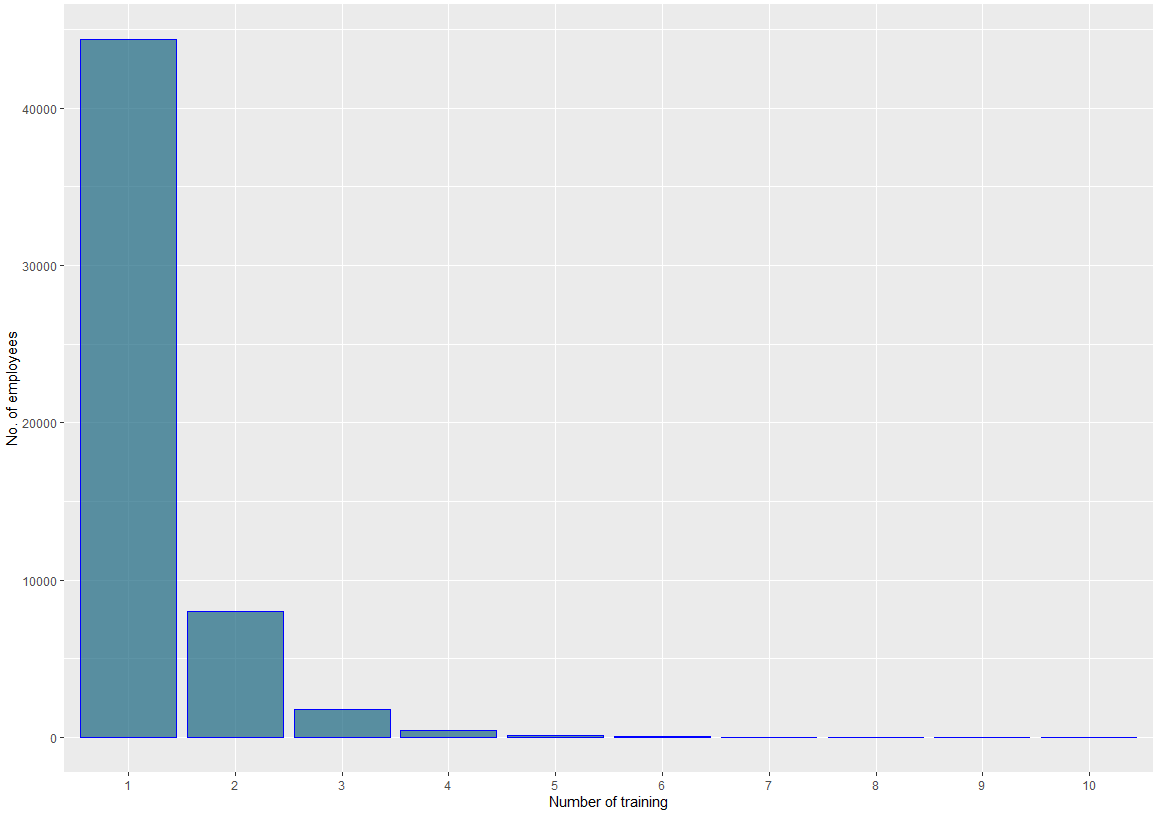
This feature may not have much influence on employee promotion so this might not be taken into consideration for model building.

6.) For No\_of\_trainings

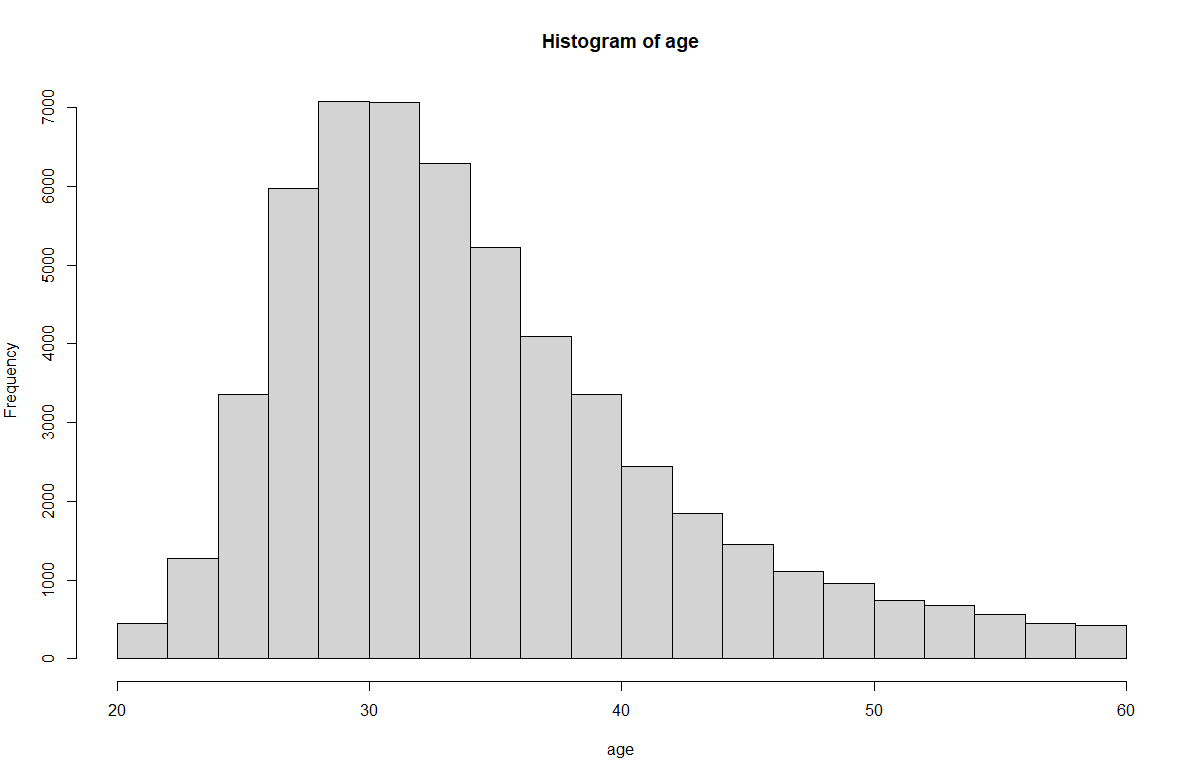


Most of the employees has been provided at least one training

max(no\_of\_trainings) =10, Maximum training provided to employee is 10 but that too very few employees. From the below graph, it can be cleared



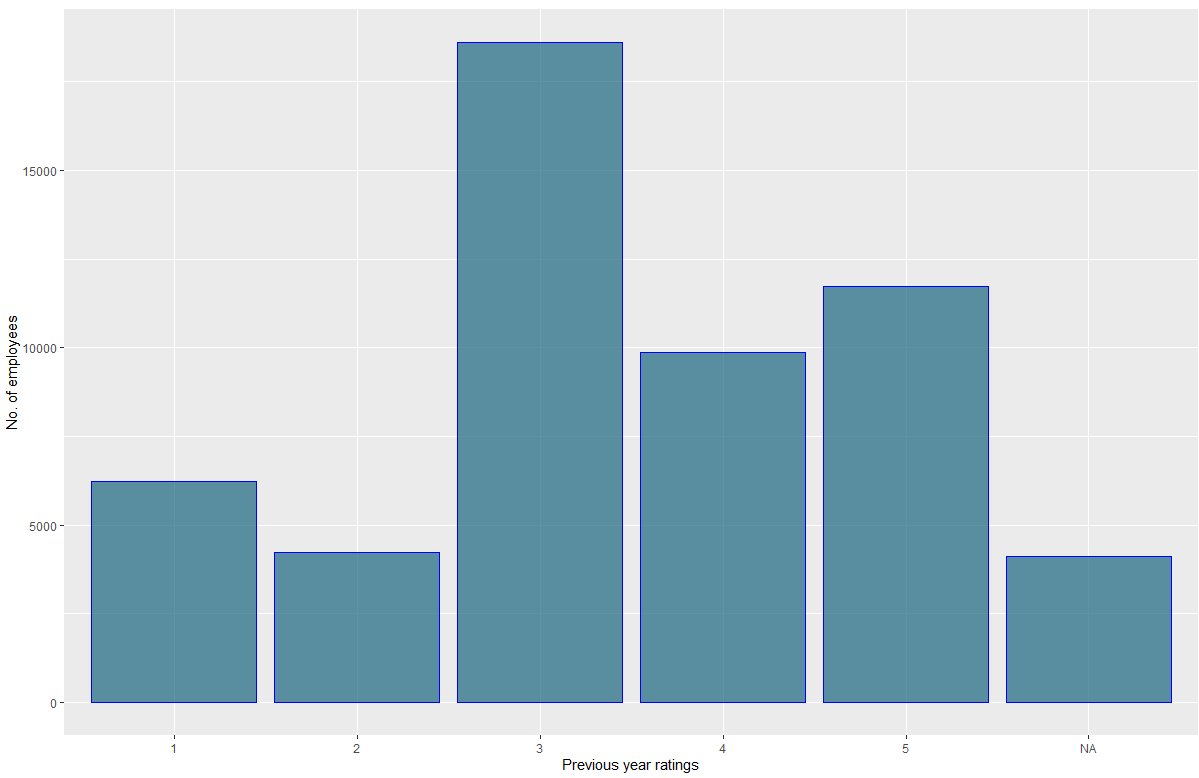
7.) For age



Maximum age of the employees is 60 years and minimum age of the employees is 20.

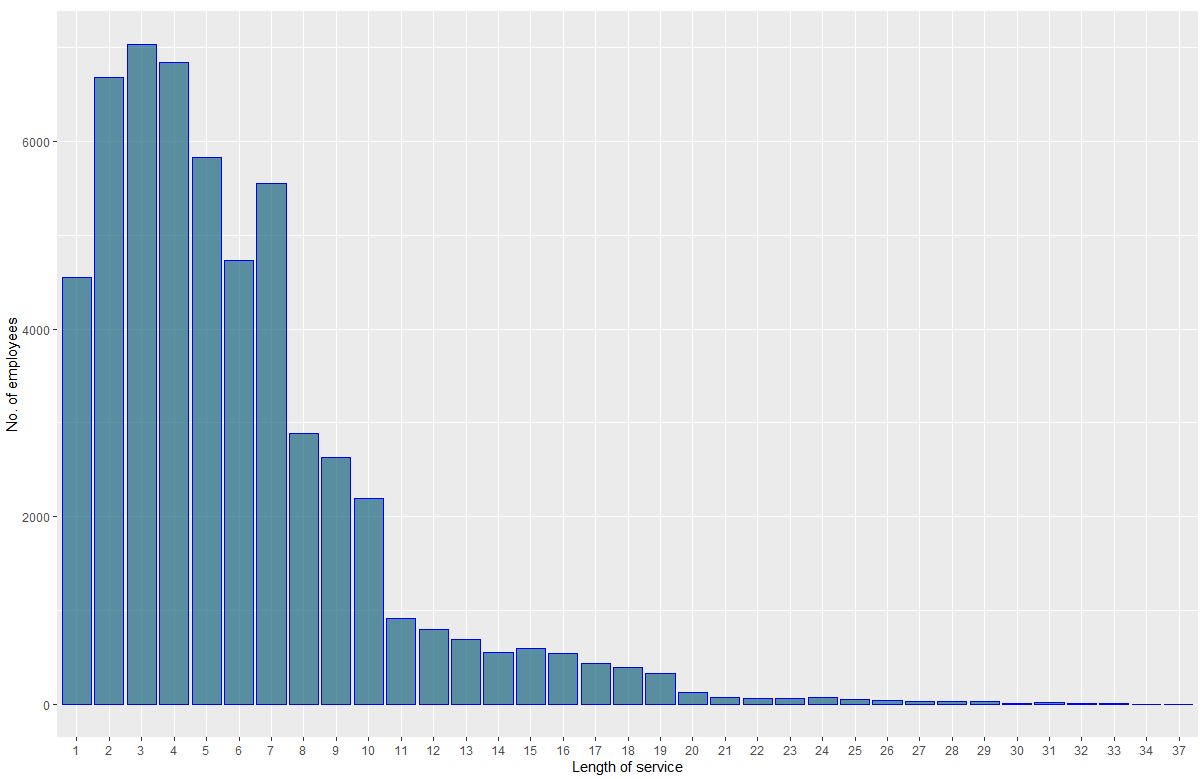
But maximum employee s belongs in the age group pf 25 to 35 years.

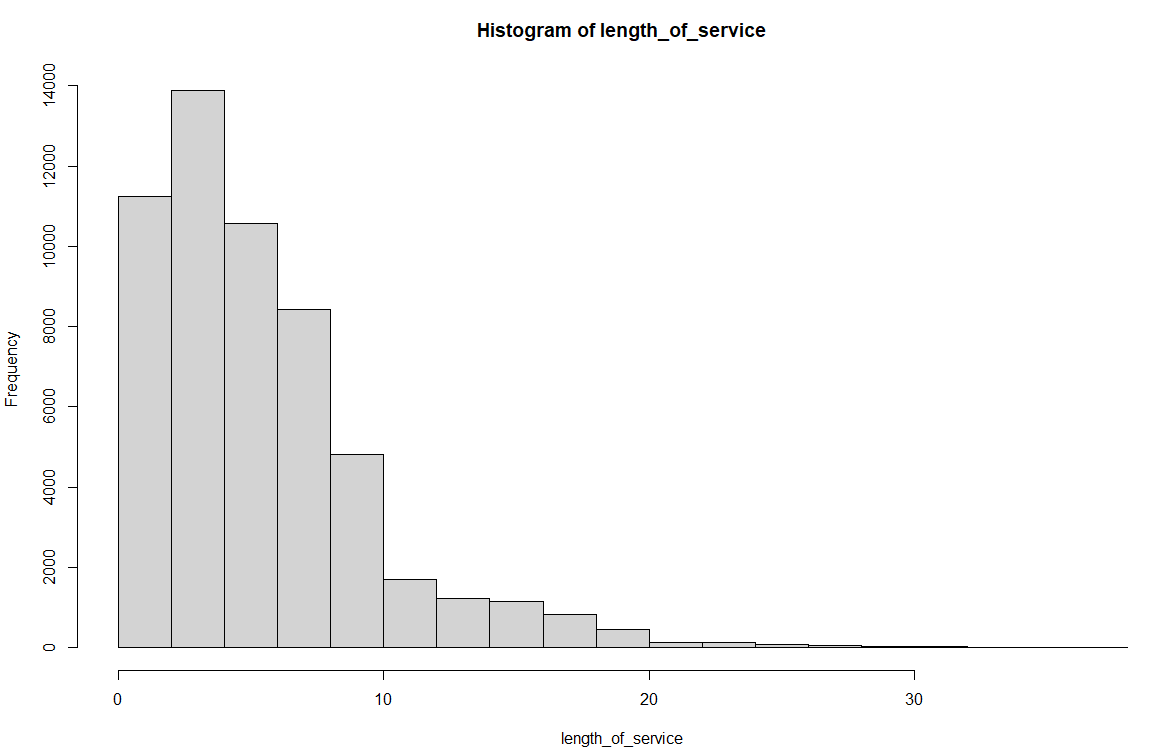
8.) Previous year ratings



From the above plot, it can be said that maximum employees they got rating in the previous year as 3 which is average rating. There are some of the NA values which need to be imputed separately

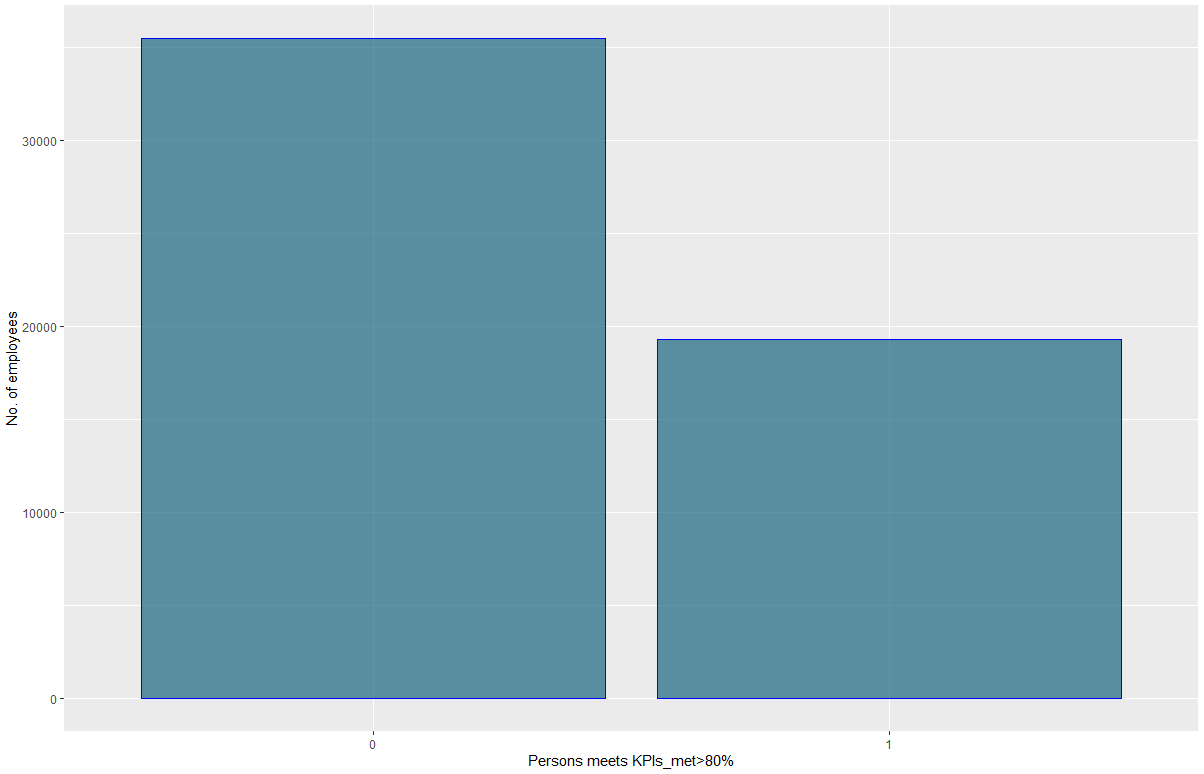
9.) Length of service





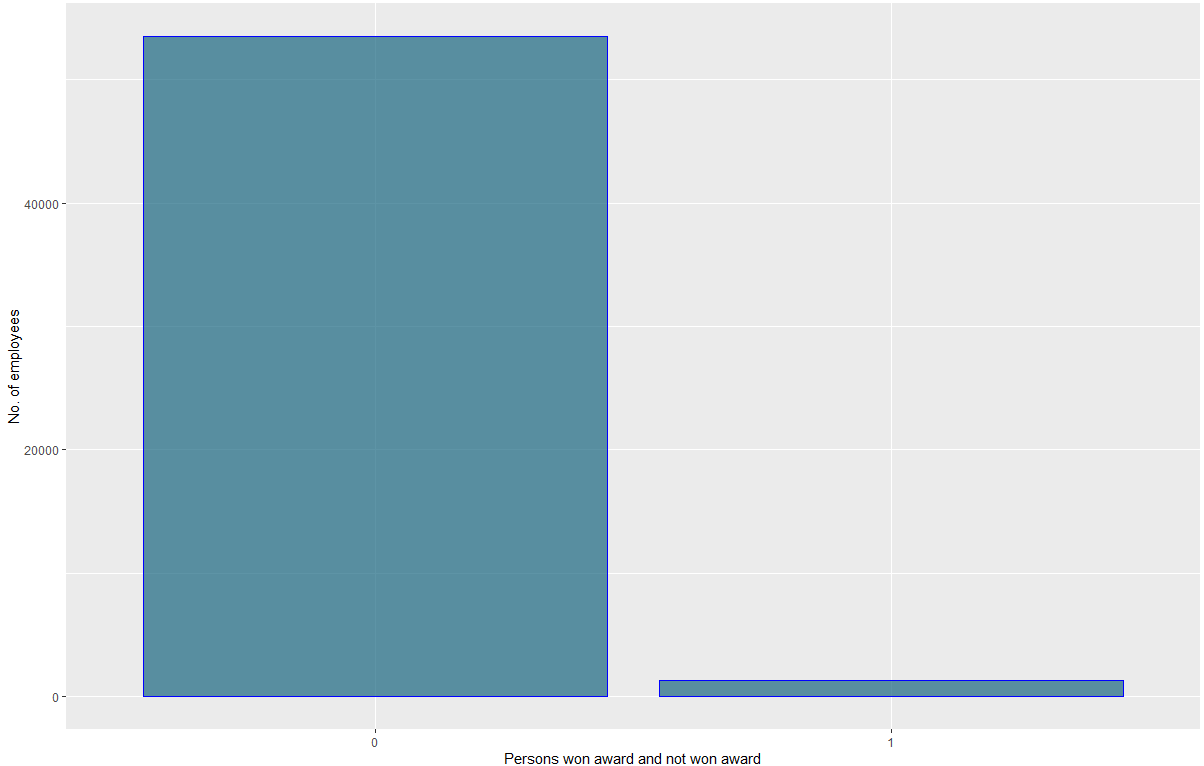
From the above plots, it can be said that, majority of the employees having work experience of 2 to 4 years. Employees having highest work experience is 37 years.

10) For KPIs\_met>80%



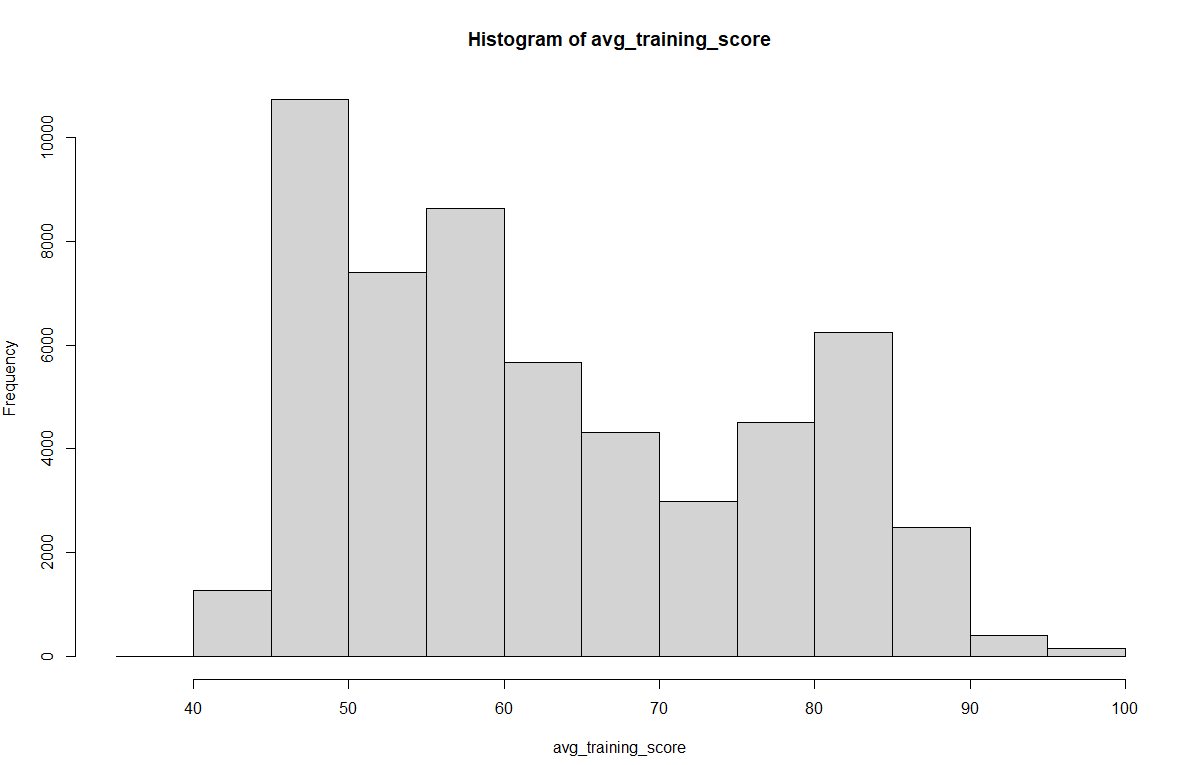
There are 35517 of employees who did not have KPI>80% and 19291 have KPI >80%, where 1 indicates Employees met KPI>80% and 0 indicates Employees did not met KPI> 80%

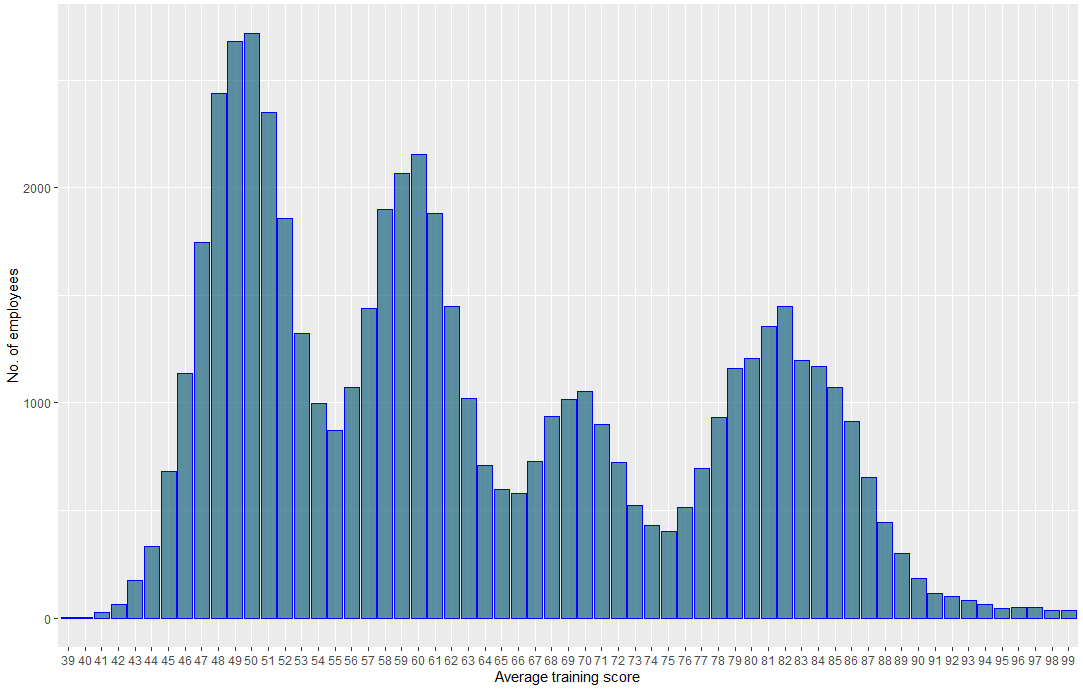
11) For Awards\_won?



There are only 1270 employees who won award and the remaining 53538 didn’t win any award where 1 indicates awarded and 0 indicates not awarded

12) Average Training Score

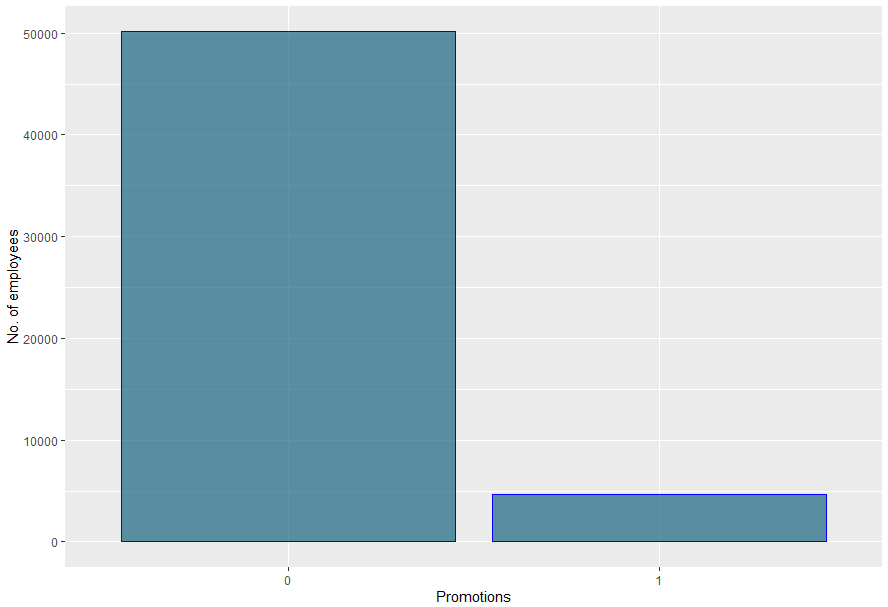




The above plots show the range of average training score that they achieve during training

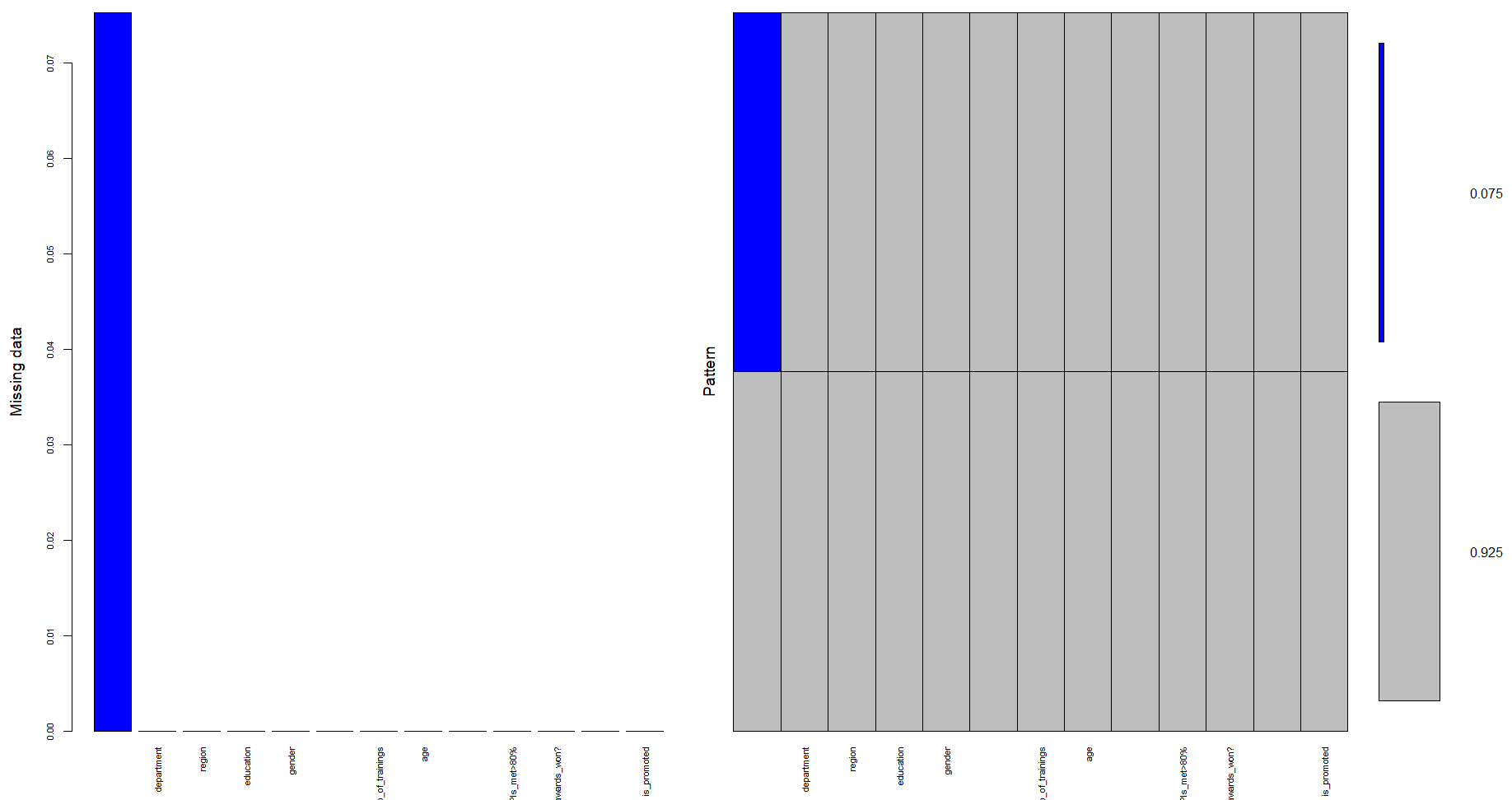
The range extends from 39 to 99. Highest number of the employees achieve score between 48 to 51.

13) For Is\_promoted



From the above plot, only 4668 was promoted and remaining 50140 is not promoted, where 1 indicates Promoted and 0 is not promoted.

**Missing value Treatment**



From the above plot, there are 92.5 % data set has no missing values. 7.5% missing values is found in previous year rating.