Employees Promotion Analysis Project

Univariant Analysis & Understanding Dataset

About Dataset:

The HR team stored data of promotion cycle last year, which consists of details of all the employees in the company working last year and also if they got promoted or not, but every time this process gets delayed due to so many details available for each employee - it gets difficult to compare and decide. this time HR team wants to utilize the stored data to make a model, that will predict if a person is eligible for promotion or not. Need to come up with a model that will help the HR team to predict if a person is eligible for promotion or not.

Objectives:

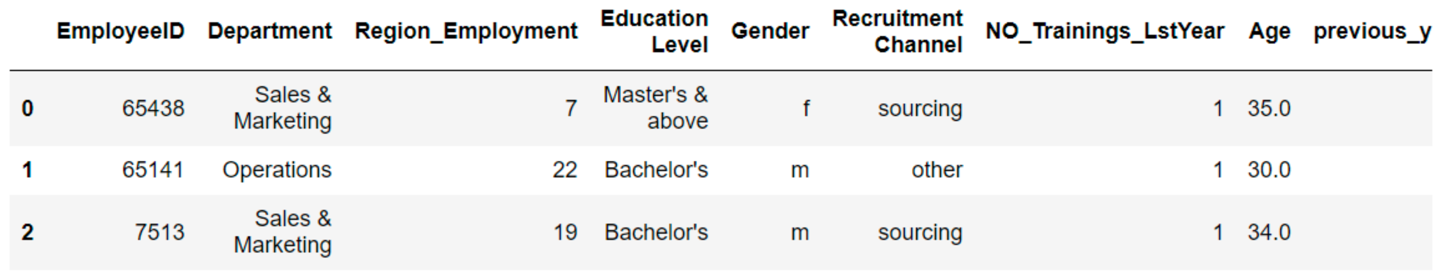
- Understanding Data
- Univariant analysis

Understanding Dataset

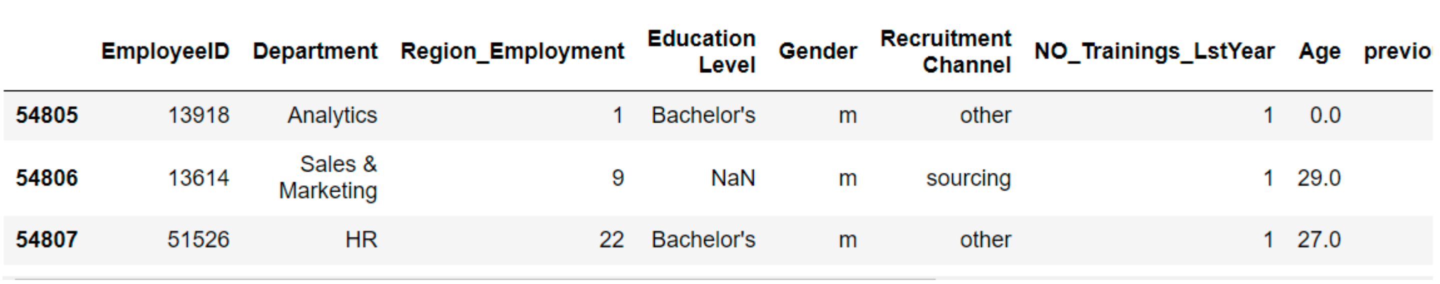
Data Feature Dictionary:

Feature Name	Description
EmployeeID	Unique ID for the employee
Department	Department of employee
Region_Employment	Region of employment (unordered)
Education Level	Education Level
Gender	Gender of Employee
Recruitment Channel	Channel of recruitment for employee
NO_Trainings_LstYear	no of other trainings completed in the previous year on soft skills, technical skills, etc.
Age	Age of Employee
previous_year_rating	Employee Rating for the previous year
Service Length	Length of service in years
Awards	if awards won during the previous year
Avg_Training_Score	Average score in current training evaluations
Is Promoted	Recommended for promotion

Sample of Data

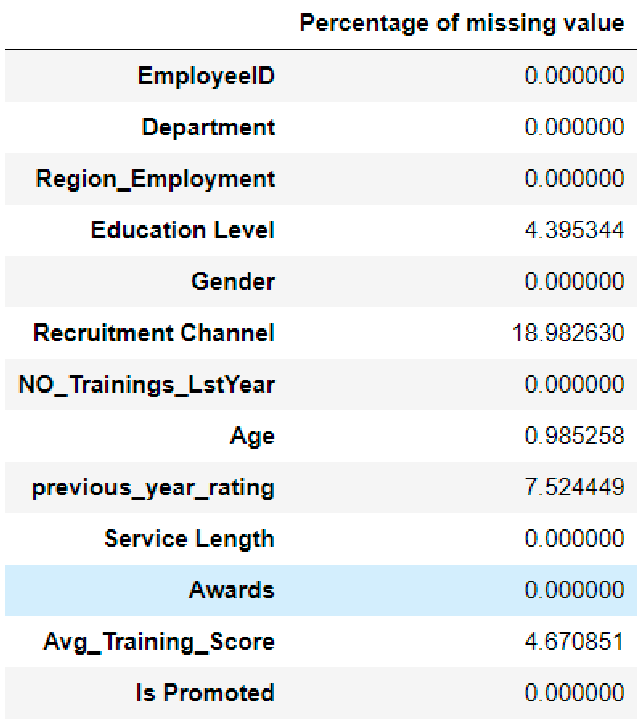


	EmployeeID	Department	Region_Employment	Education Level	Gender	Recruitment Channel	NO_Trainings_LstYear	Age	previous_y
0	65438	Sales & Marketing	7	Master's & above	f	sourcing	1	35.0	
1	65141	Operations	22	Bachelor's	m	other	1	30.0	
2	7513	Sales & Marketing	19	Bachelor's	m	sourcing	1	34.0	

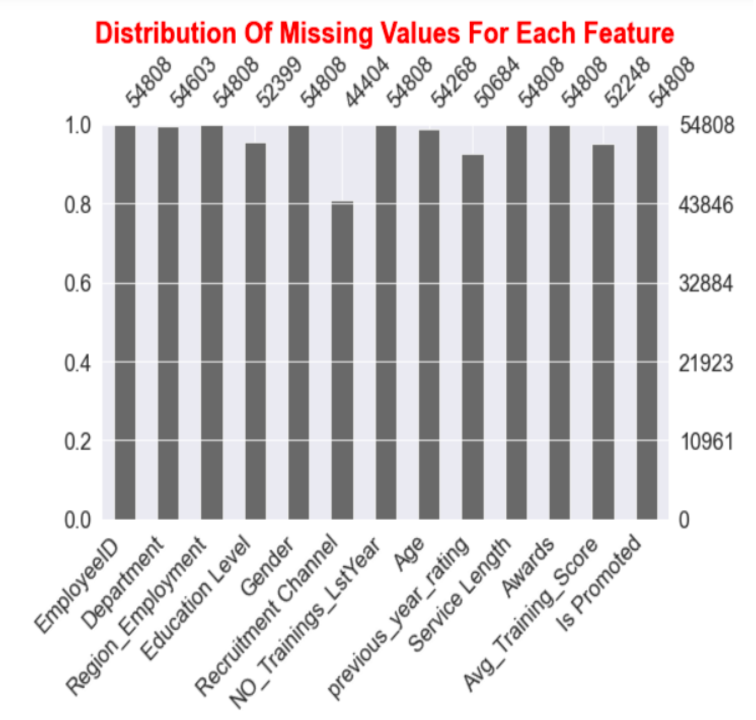


	EmployeeID	Department	Region_Employment	Education Level	Gender	Recruitment Channel	NO_Trainings_LstYear	Age	previo
54805	13918	Analytics	1	Bachelor's	m	other	1	0.0	
54806	13614	Sales & Marketing	9	NaN	m	sourcing	1	29.0	
54807	51526	HR	22	Bachelor's	m	other	1	27.0	

- Data Consist of (54808) row and (13) Columns
- Dtype: float64(3), int64(4), object (6)
- Memory usage: 5.4+ MB
- No Duplicated Records
- **Some features have missing values**

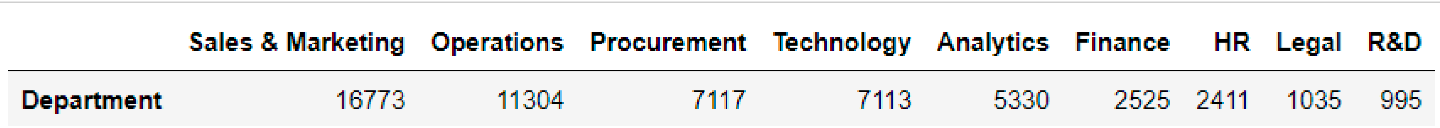


	Percentage of missing value
EmployeeID	0.000000
Department	0.000000
Region_Employment	0.000000
Education Level	4.395344
Gender	0.000000
Recruitment Channel	18.982630
NO_Trainings_LstYear	0.000000
Age	0.985258
previous_year_rating	7.524449
Service Length	0.000000
Awards	0.000000
Avg_Training_Score	4.670851
Is Promoted	0.000000

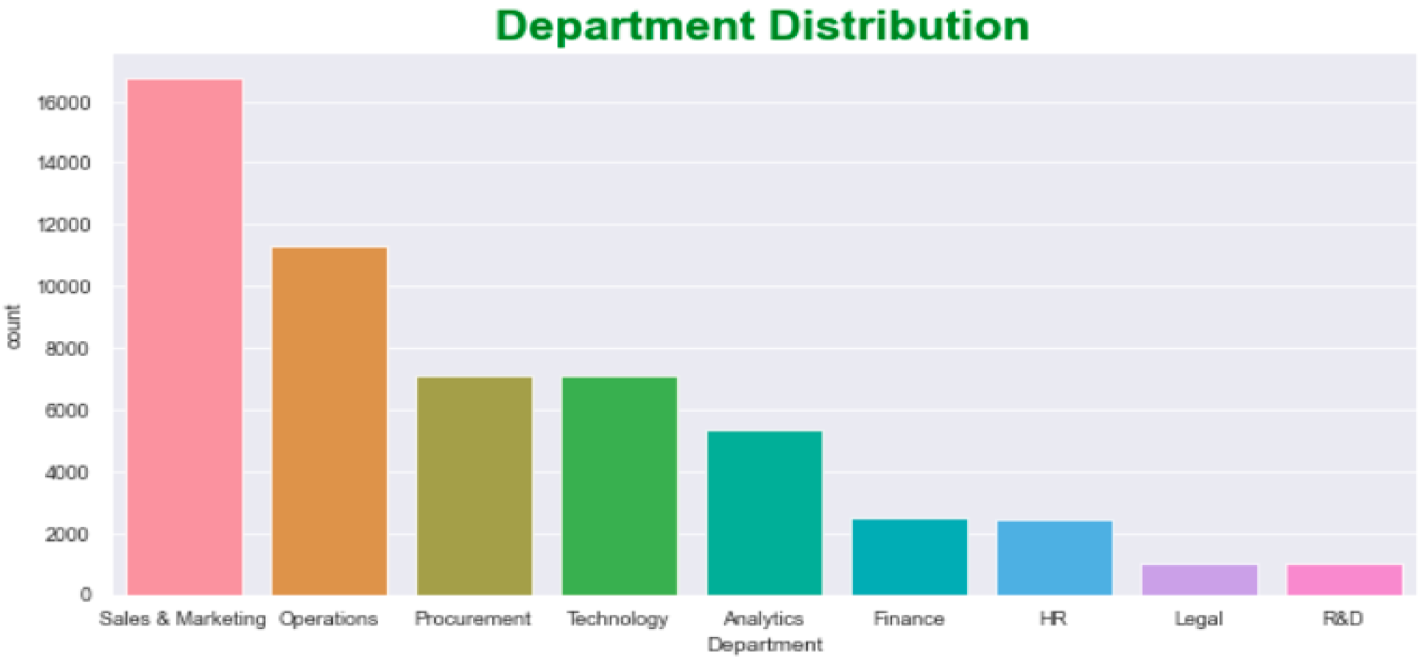


Univariant Analysis

Department Feature:

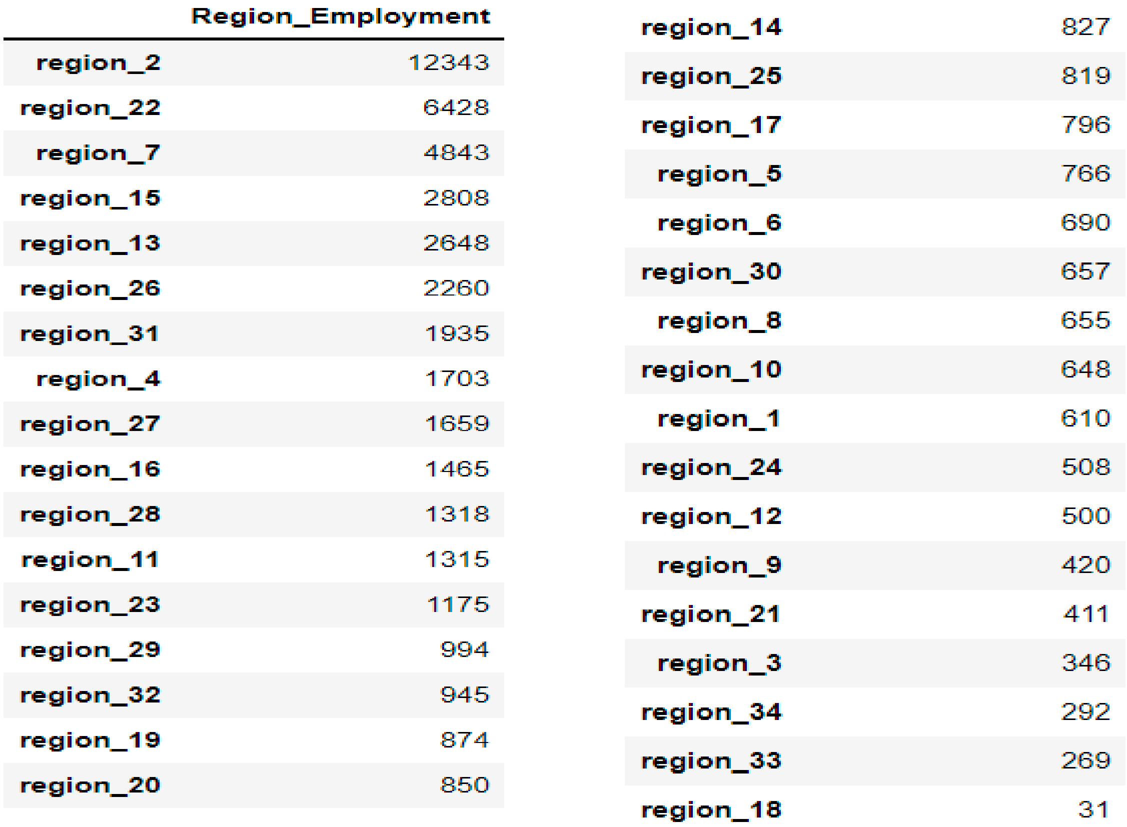


	Sales & Marketing	Operations	Procurement	Technology	Analytics	Finance	HR	Legal	R&D
Department	16773	11304	7117	7113	5330	2525	2411	1035	995

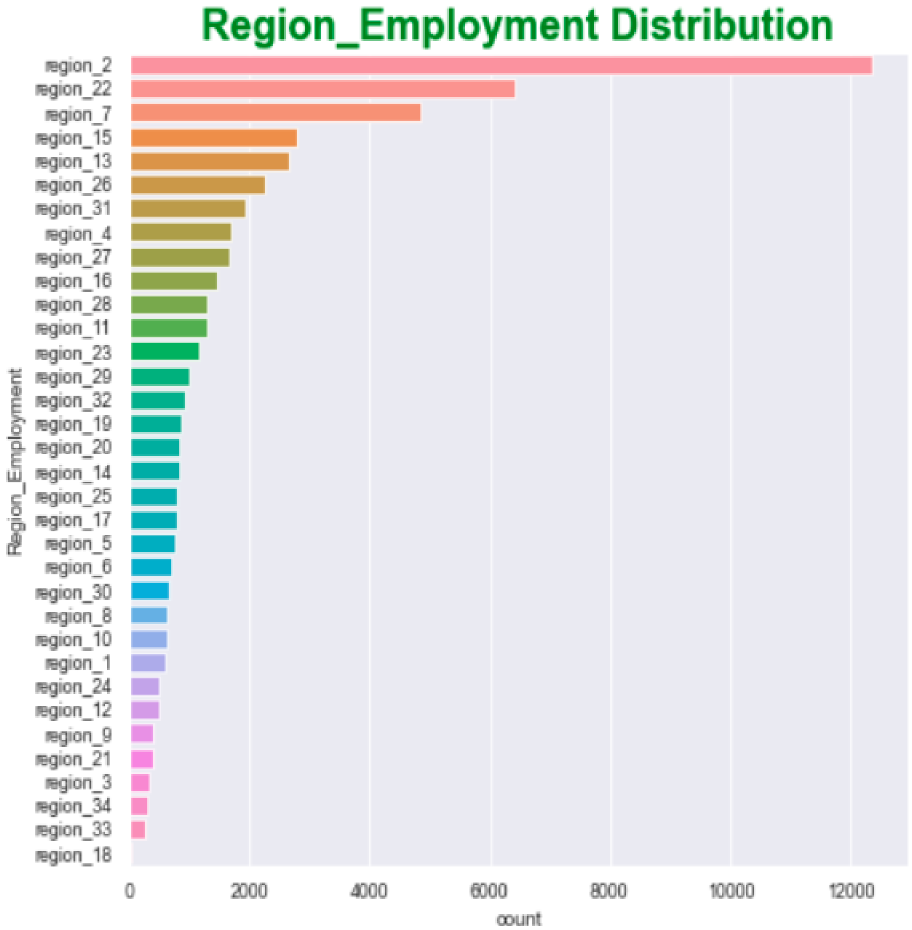


- Department Feature has 9 unique department
- has 205 (-) converted to null values
- (Sales & Marketing) is most Frequent and (R&D) is least Frequent

Region_Employment Feature Distribution

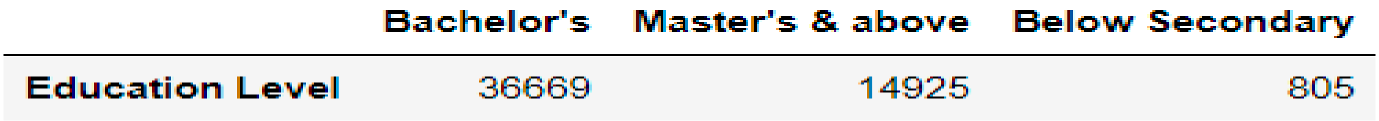


	Region_Employment
region_2	12343
region_22	6428
region_7	4843
region_15	2808
region_13	2648
region_26	2260
region_31	1935
region_4	1703
region_27	1659
region_16	1465
region_28	1318
region_11	1315
region_23	1175
region_29	994
region_32	945
region_19	874
region_20	850
region_14	827
region_25	819
region_17	796
region_5	766
region_6	690
region_30	657
region_8	655
region_10	648
region_1	610
region_24	508
region_12	500
region_9	420
region_21	411
region_3	346
region_34	292
region_33	269
region_18	31

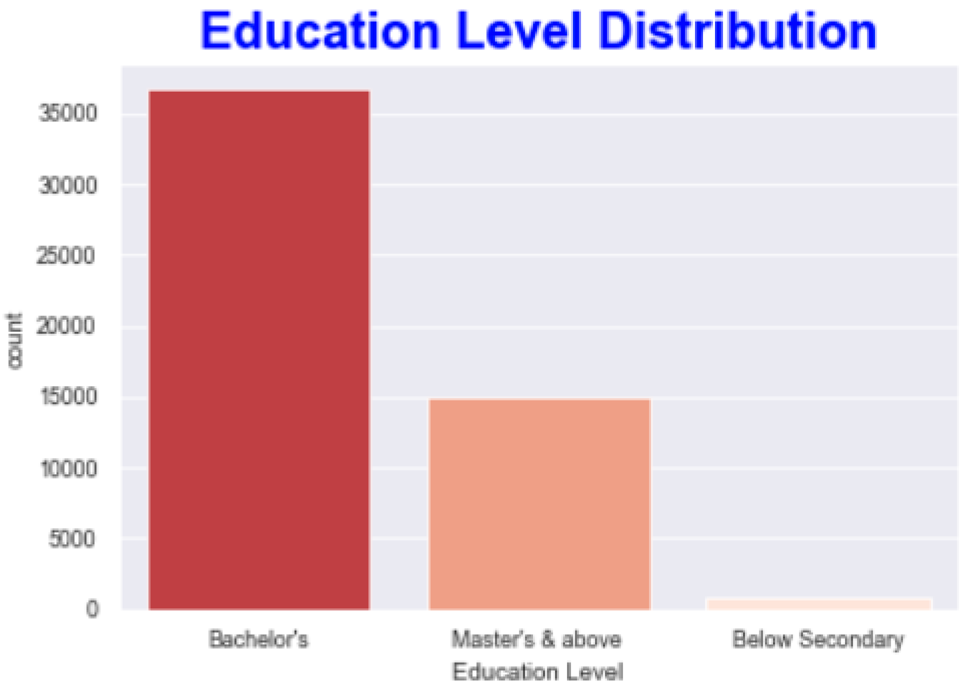


- Data type of region employment is (int64) and it is not logic this number not represent any order between regions and any priority it is just a category, so we must fix this problem
- Adding (region) word before each number
- Region Employment has no missing values
- has 34 unique regions numbered from 1 to 34
- Region_2 is most frequent and Region_18 is the least frequent

Education Level Feature Distribution

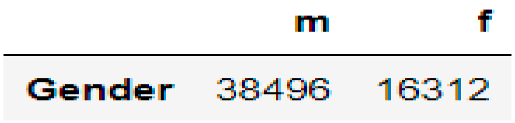


	Bachelor's	Master's & above	Below Secondary
Education Level	36669	14925	805

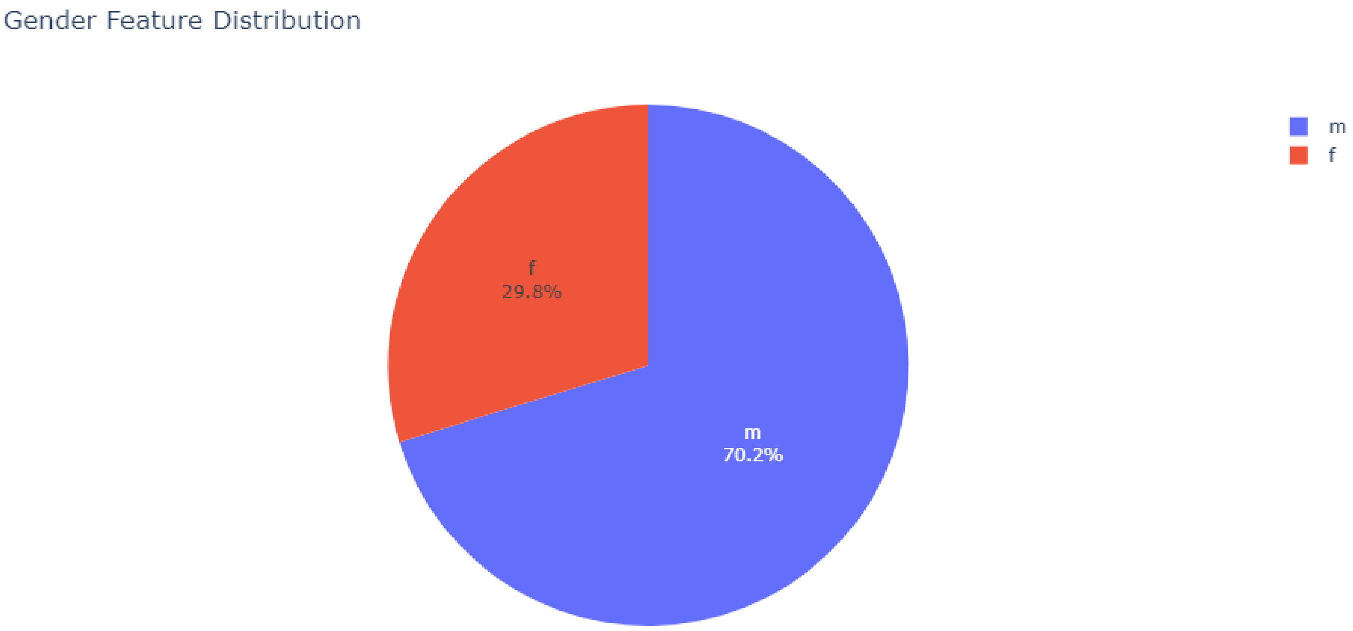


- Education Level Feature has (2409) missing value
- Has (3) unique Education level
- (Bachelor's) is the most frequent and (Below Secondary) is the least frequent

Gender Feature Distribution

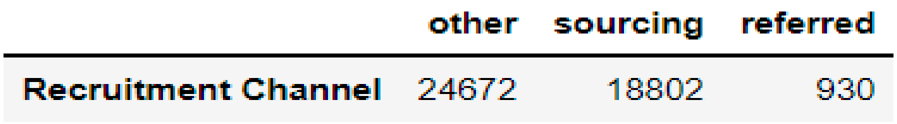


	m	f
Gender	38496	16312

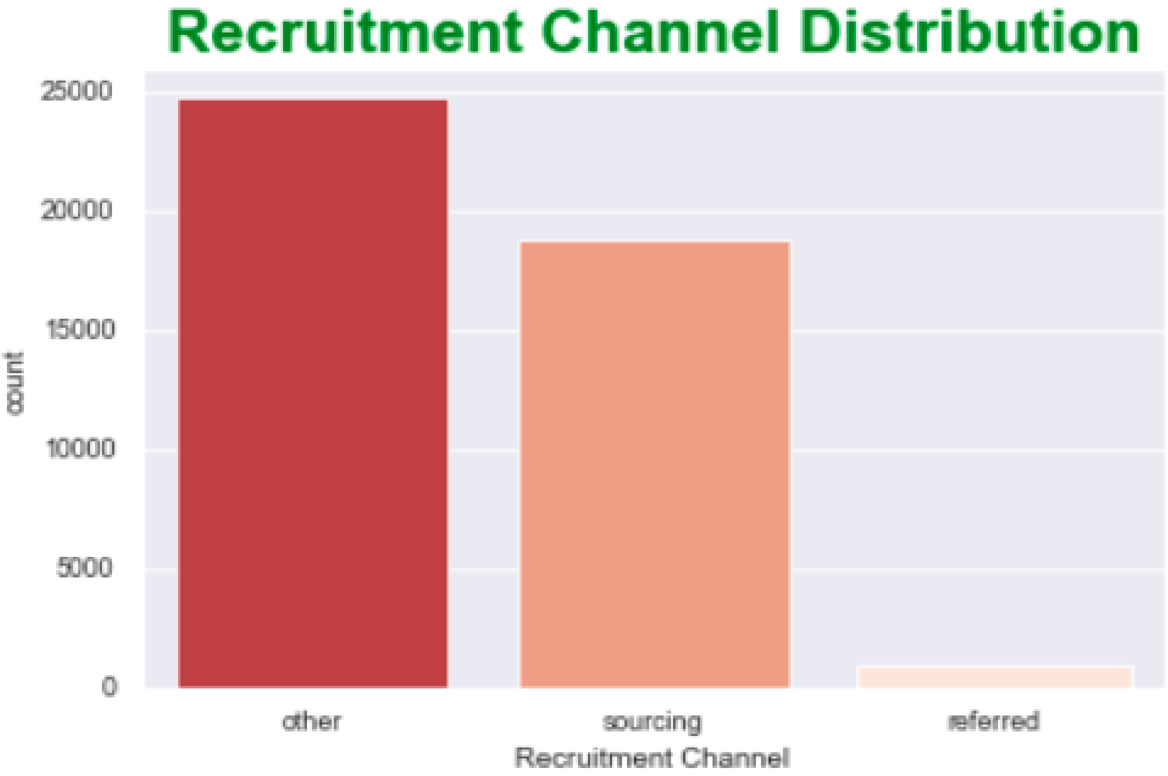


- Gender Feature has no missing values
- Gender Feature has (2) unique values (m, f)
- Male is the majority with percentage (70.2 %)

Recruitment Channel Feature Distribution

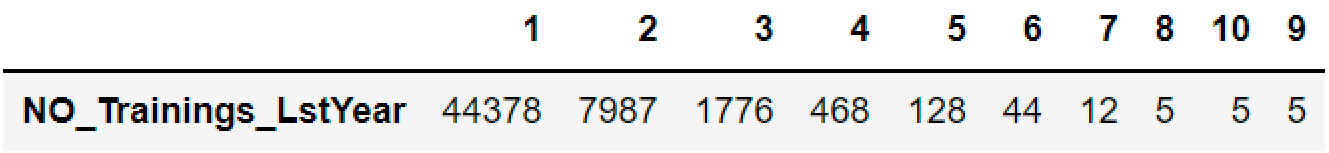


	other	sourcing	referred
Recruitment Channel	24672	18802	930

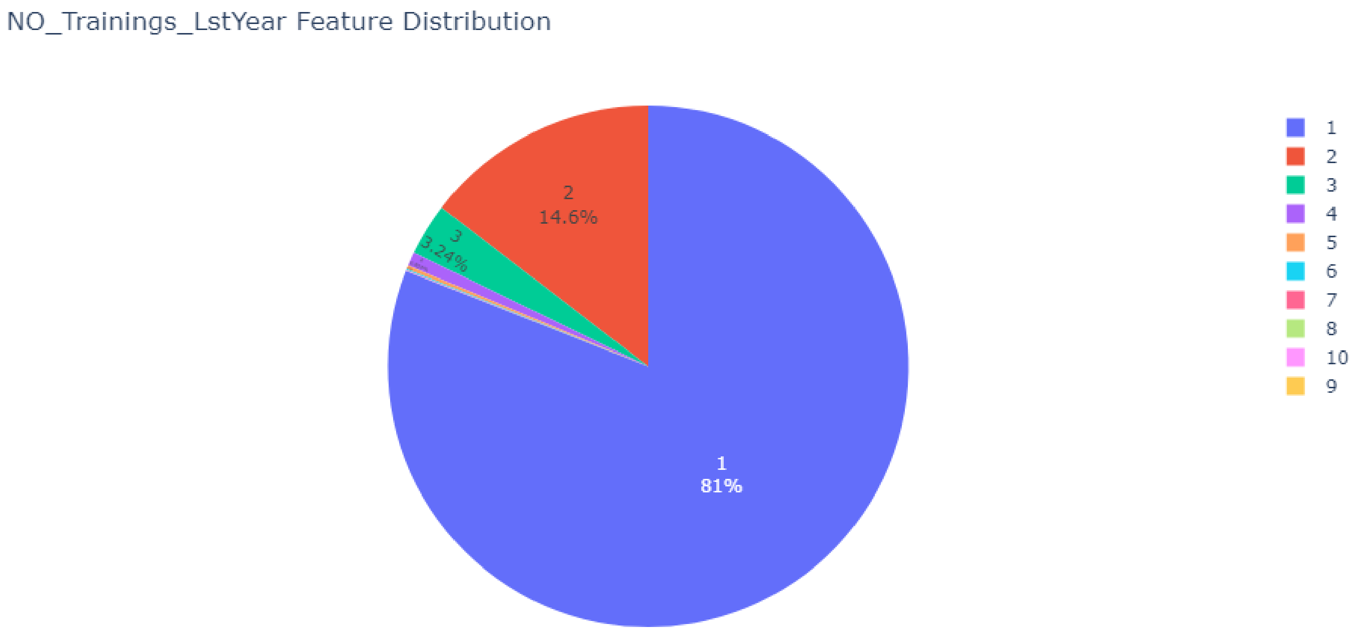


- Recruitment Channel Feature has 3 different values
- Recruitment Channel Feature has (10404) missing value with percentage (19%)
- other Recruitment is most frequent and referred is least frequent

NO_Trainings_LstYear Feature Distribution

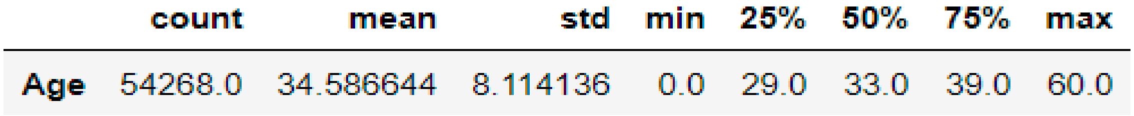


	1	2	3	4	5	6	7	8	10	9
NO_Trainings_LstYear	44378	7987	1776	468	128	44	12	5	5	5

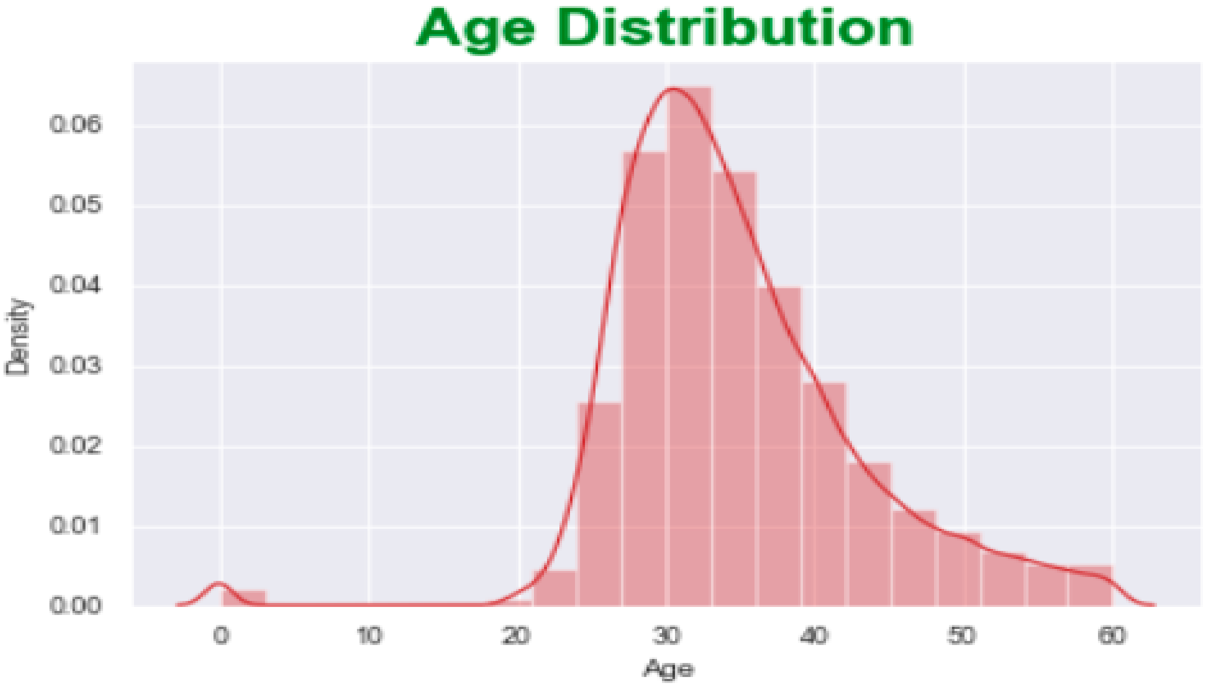


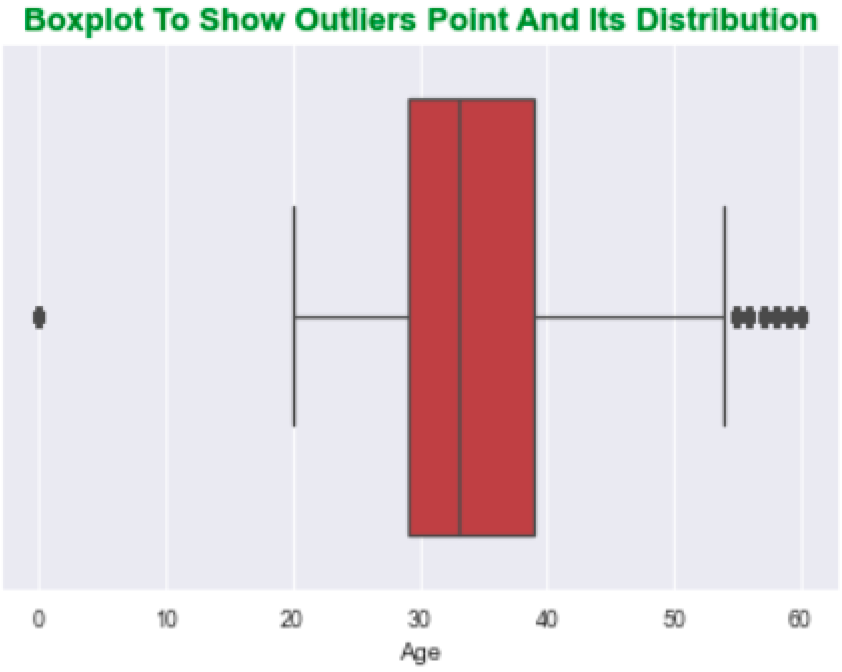
- Number of Trainings has **NO** missing values
- The majority of employees take (1) training last year
- Min value = (1) and Max value = (10)

Age Feature Distribution



	count	mean	std	min	25%	50%	75%	max
Age	54268.0	34.586644	8.114136	0.0	29.0	33.0	39.0	60.0





* Age Feature is nearly normally distributed with **skewness to the left**

* Has outlier points (1748)

* Has (540) missing values

* With (median = 33.0)

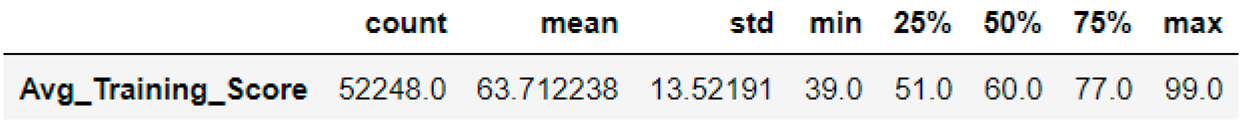
and (mean = 34.586644)

* Min value = (0)

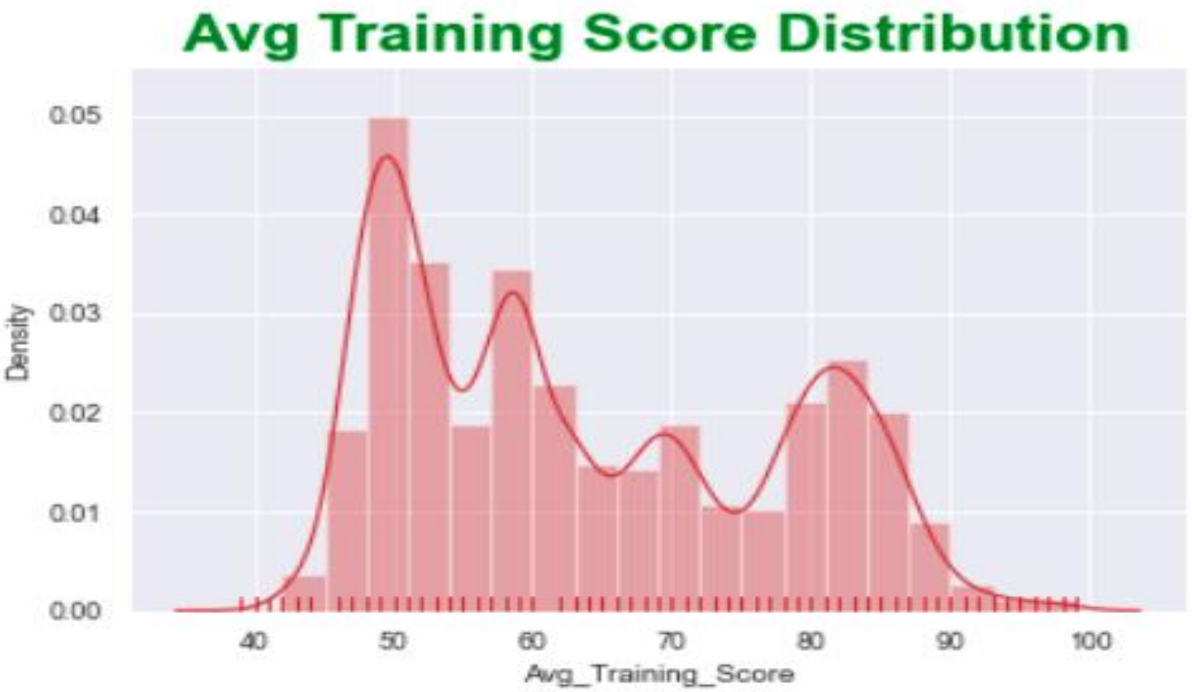
and max value = (60.0)

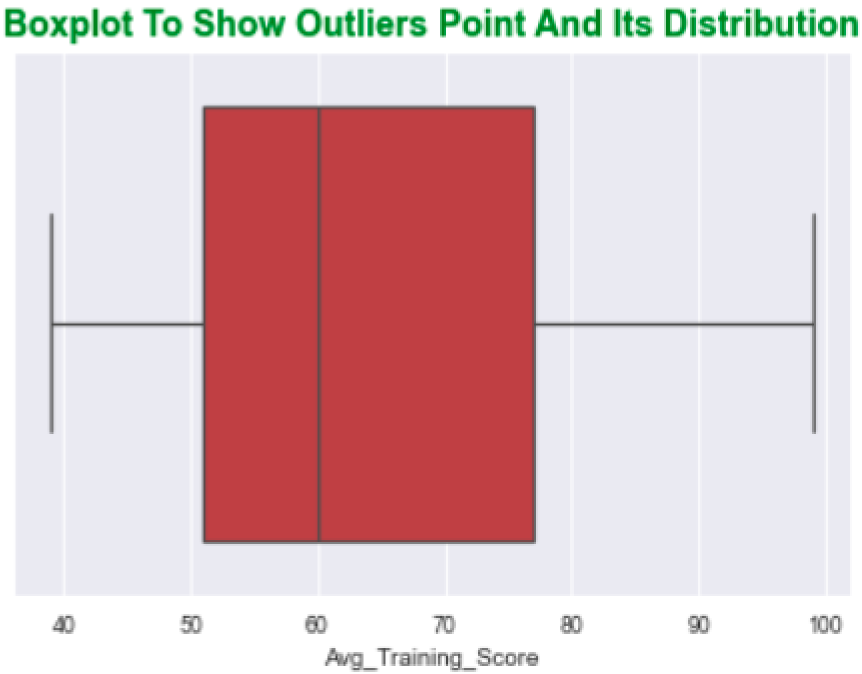
* Min value = (0) is very strange need to deal with it!

Avg_Training_Score Feature Distribution



	count	mean	std	min	25%	50%	75%	max
Avg_Training_Score	52248.0	63.712238	13.52191	39.0	51.0	60.0	77.0	99.0





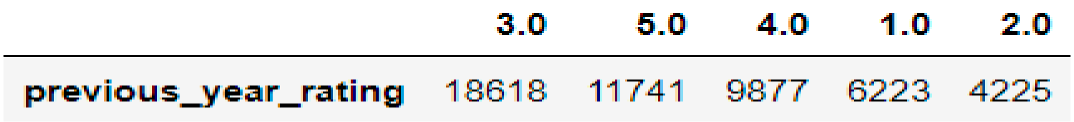
* Average Training Score Feature nearly not normally distributed

* Has (2560) points missing value

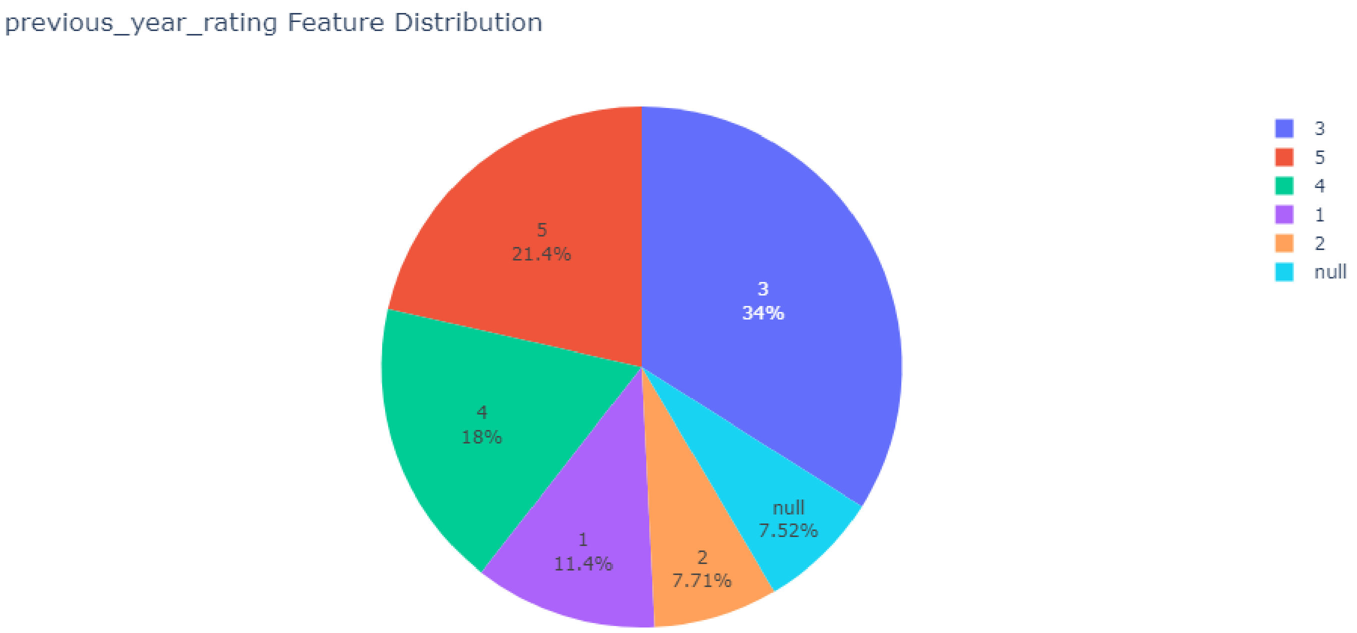
* **Has no outlier points**

* Points from (39) to (99) with mean (63.712238) and median (60.0)

previous_year_rating Feature Distribution

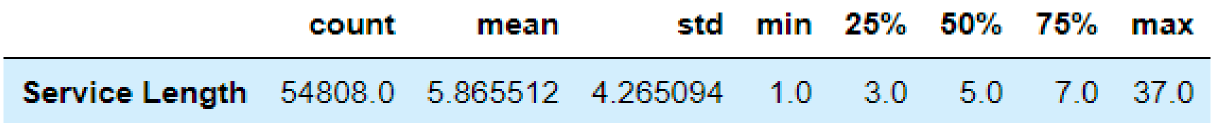


	3.0	5.0	4.0	1.0	2.0
previous_year_rating	18618	11741	9877	6223	4225

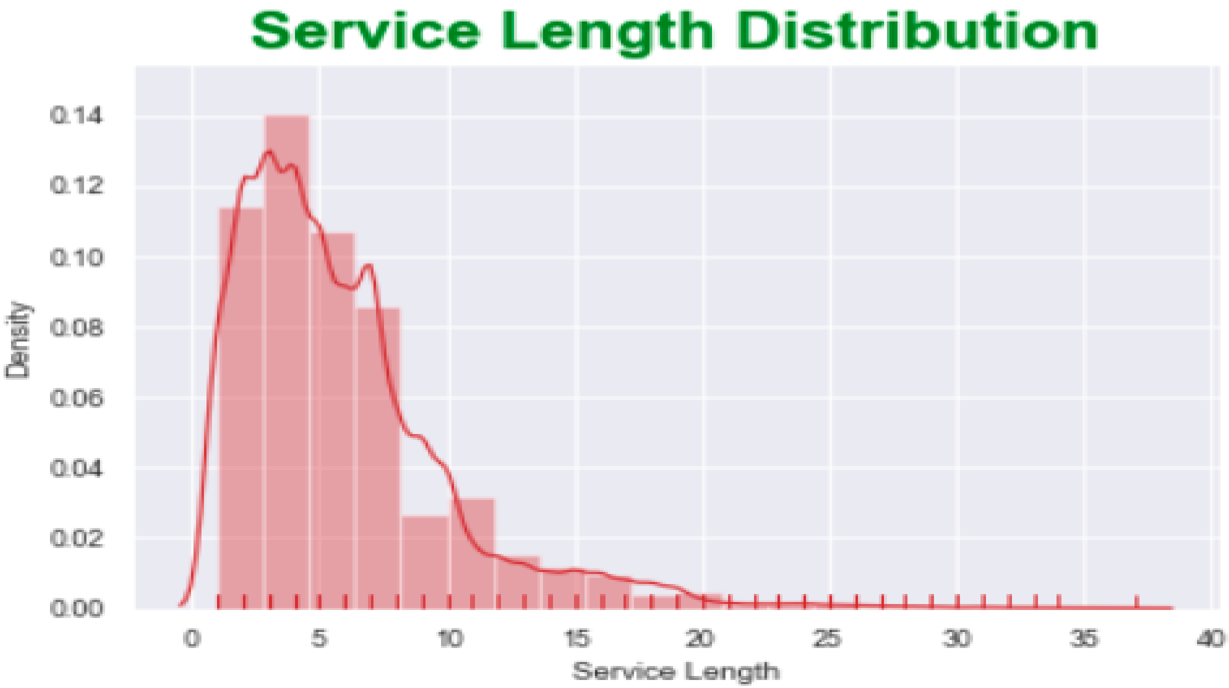


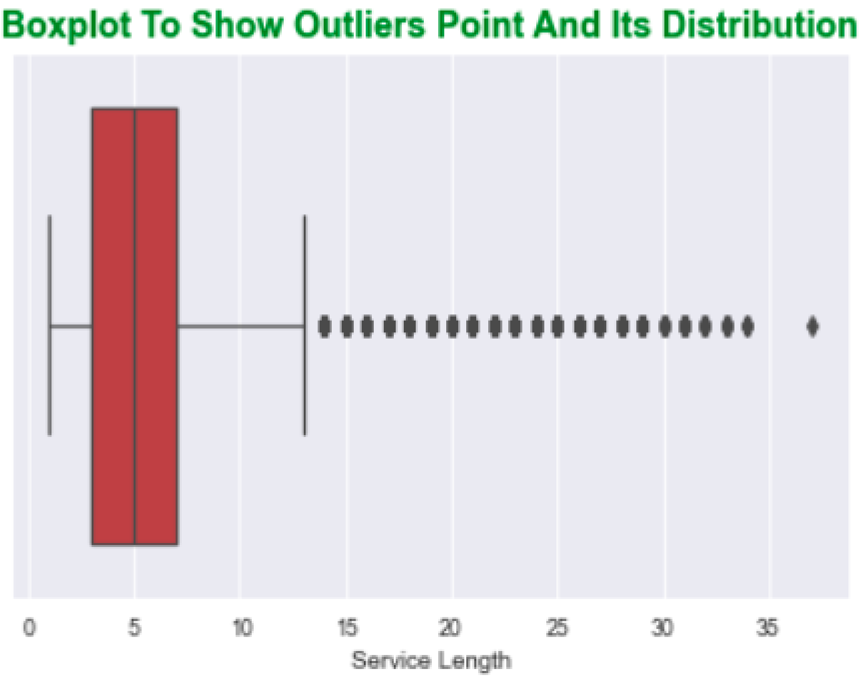
- Previous Year Rating Feature has (5) unique values from (1 - 5)
- Most of Employs has rate (3) with percentage (34%)
- Previous Year Rating has **(4124) missing value with percentage (7%)**

Service Length in (Years) Feature Distribution



	count	mean	std	min	25%	50%	75%	max
Service Length	54808.0	5.865512	4.265094	1.0	3.0	5.0	7.0	37.0





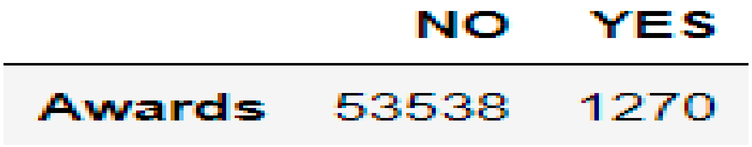
* Service Length with **skewness to the right**

* There were no missing values

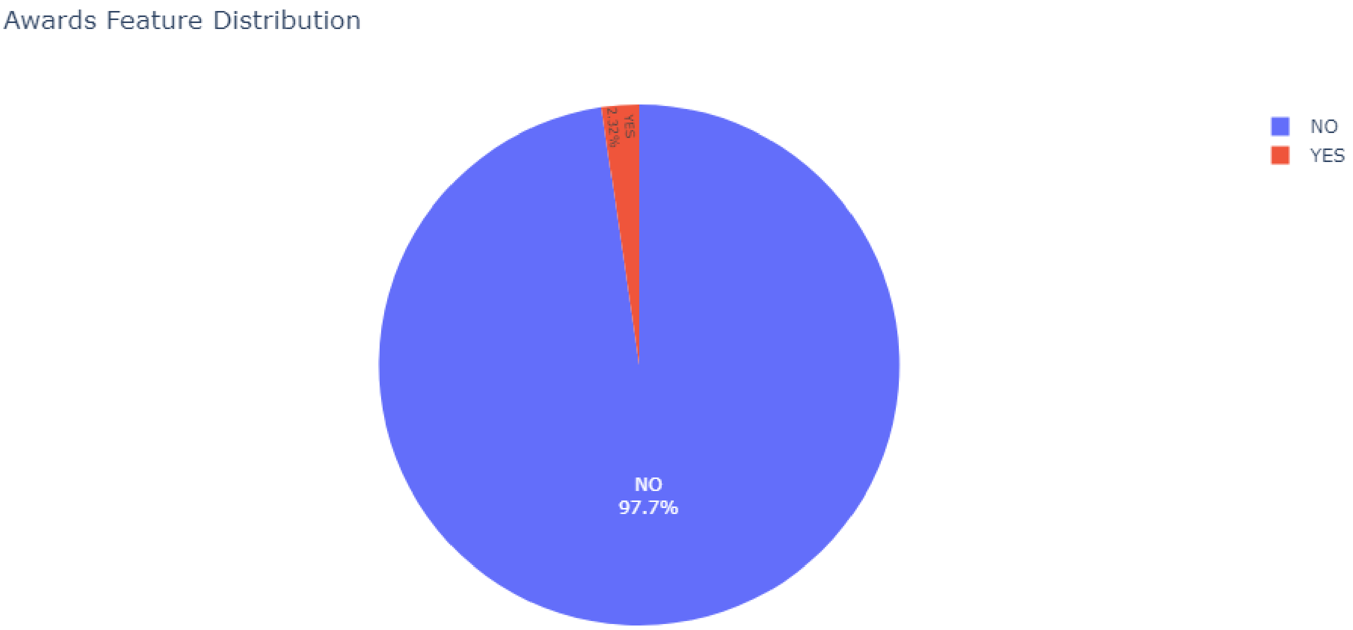
* There was **(3489) outlier points**

* Range from (1) to (37) with **mean = (5.865512)** and **median = (5.0)**

Awards Feature Distribution

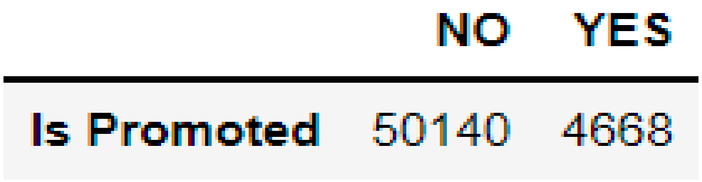


	NO	YES
Awards	53538	1270

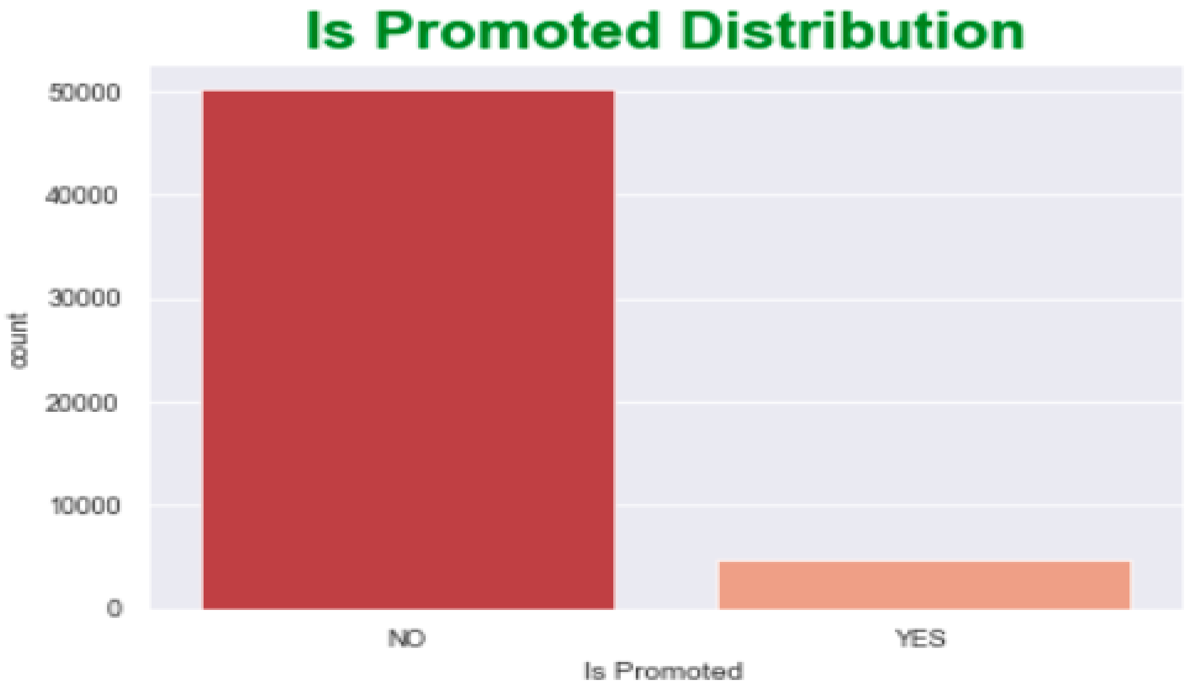


- Awards Feature has (2) unique values [yes, no]
- Most of the employees **not** awarded with **percentage (97%)**
- There was **No Missing** values

Is Promoted Feature Distribution



	NO	YES
Is Promoted	50140	4668



- Is Promoted Feature having (2) unique values [yes, no]
- Most of the Employees **Not** Promoted with **percentage (91.5%)**
- **No missing values**
- **Unbalanced Class (Label)**

Based on this report we will deal with our feature in preprocessing phase