Business Report – HR Salary Prediction

(Capstone Project)

by Anbalagan R

05-May-2024

Contents:

1.1) Introduction of the business problem:

Problem Statement:

To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles
Apart from the existing salary, there is a considerable number of factors regarding an employee's experience and other abilities to which they get evaluated in interviews. Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered..

Objective:

The objective of this exercise is to build a model, using historical data that will determine an employee's salary to be offered, such that manual judgments on selection are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles

Need of the study/project:

The study/project is needed to reduce the analysis manual judgement of Employees Offered CTC in our organization And determine the eligible employees for a company based on historical data. To proven by using treatments

Initial Data Limitations:

We only have a small amount of information available to us at first. This is not exhaustive, but it does contain historical data. If this first phase is successful, access to larger datasets for additional analysis and model improvement will become available.

Progressive Approach:

We'll start by showcasing our model's efficacy using the scant data at our disposal. The business will provide us more access to a comprehensive data lake after we demonstrate its value, enabling us to improve and grow our model for more informed decision-making.

Understanding business/social opportunity:

This model seeks to minimize human judgment with regard to salary to be offered. The business opportunity lies in building a model using historical data to determine an employee's salary to be offered, such that manual judgments on selection are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles

1.2) Data Report:

Understanding how data was collected in terms of time, frequency and methodology:

In the list of 25000 employees data collected in terms of various information based on education,role, experience,ratings, achievements,etc.

Visual inspection of data:

- The total number of Rows in the Data 25000
- The total number of Columns in the Data 29
- From below table we could identify the datatype of each column also we can see there are many missing values in the data set. Which will be treated in further proceedings.
- There are 3 float64 type , 10 int64 type, 16 object type, data columns in the data set

```
Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 29 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   IDX                                 25000 non-null  int64
 1   Applicant_ID                        25000 non-null  int64
 2   Total_Experience                    25000 non-null  int64
 3   Total_Experience_in_field_applied   25000 non-null  int64
 4   Department                          22222 non-null  object
 5   Role                                24037 non-null  object
 6   Industry                            24092 non-null  object
 7   Organization                        24092 non-null  object
 8   Designation                         21871 non-null  object
 9   Education                           25000 non-null  object
 10  Graduation_Specialization           18820 non-null  object
 11  University_Grad                     18820 non-null  object
 12  Passing_Year_Of_Graduation          18820 non-null  float64
 13  PG_Specialization                   17308 non-null  object
 14  University_PG                       17308 non-null  object
 15  Passing_Year_Of_PG                  17308 non-null  float64
 16  PHD_Specialization                  13119 non-null  object
 17  University_PHD                      13119 non-null  object
 18  Passing_Year_Of_PHD                 13119 non-null  float64
 19  Curent_Location                     25000 non-null  object
 20  Preferred_location                  25000 non-null  object
 21  Current_CTC                         25000 non-null  int64
 22  Inhand_Offer                        25000 non-null  object
 23  Last_Appraisal_Rating               24092 non-null  object
 24  No_Of_Companies_worked              25000 non-null  int64
 25  Number_of_Publications              25000 non-null  int64
 26  Certifications                      25000 non-null  int64
 27  International_degree_any            25000 non-null  int64
 28  Expected_CTC                        25000 non-null  int64
dtypes: float64(3), int64(10), object(16)
```

Table 1.1

Understanding of attributes :

- There are no duplicate lines in the data-set.

Data Summary:

	count	mean	std	min	25%	50%	75%	max
IDX	25000.0	1.250050e+04	7.217023e+03	1.0	6250.75	12500.5	18750.25	25000.0
Applicant_ID	25000.0	3.499324e+04	1.439027e+04	10000.0	22563.75	34974.5	47419.00	60000.0
Total_Experience	25000.0	1.249308e+01	7.471398e+00	0.0	6.00	12.0	19.00	25.0
Total_Experience_in_field_applied	25000.0	6.258200e+00	5.819513e+00	0.0	1.00	5.0	10.00	25.0
Passing_Year_Of_Graduation	18820.0	2.002194e+03	8.316640e+00	1986.0	1996.00	2002.0	2009.00	2020.0
Passing_Year_Of_PG	17308.0	2.005154e+03	9.022963e+00	1988.0	1997.00	2006.0	2012.00	2023.0
Passing_Year_Of_PHD	13119.0	2.007396e+03	7.493601e+00	1995.0	2001.00	2007.0	2014.00	2020.0
Current_CTC	25000.0	1.760945e+06	9.202125e+05	0.0	1027311.50	1802567.5	2443883.25	3999693.0
No_Of_Companies_worked	25000.0	3.482040e+00	1.690335e+00	0.0	2.00	3.0	5.00	6.0
Number_of_Publications	25000.0	4.089040e+00	2.606612e+00	0.0	2.00	4.0	6.00	8.0
Certifications	25000.0	7.736800e-01	1.199449e+00	0.0	0.00	0.0	1.00	5.0
International_degree_any	25000.0	8.172000e-02	2.739431e-01	0.0	0.00	0.0	0.00	1.0
Expected_CTC	25000.0	2.250155e+06	1.160480e+06	203744.0	1306277.50	2252136.5	3051353.75	5599570.0

Table 1.2

File: Data.csv
Target variable: Expected_CTC
Data dictionary:

Table 1.3

IDX	Index
Applicant_ID	Application ID
Total_Experience	Total industry experience
Total_Experience_in_field_applied	Total experience in the field applied for (past work experience that is relevant to the job)
Department	Department name of current company
Role	Role in the current company
Industry	Industry name of current field
Organization	Organization name
Designation	Designation in current company
Education	Education
Graduation_Specialization	Specialization subject in graduation
University_Grad	University or college in Graduation
Passing_Year_Of_Graduation	Year of passing Graduation
PG_Specialization	Specialization subject in Post-Graduation
University_PG	University or college in Post-Graduation
Passing_Year_Of_PG	Year of passing Post Graduation
PHD_Specialization	Specialization subject in Post-Graduation
University_PHD	University or college in Post Doctorate
Passing_Year_Of_PHD	Year of passing PHD
Curent_Location	Curent Location
Preferred_location	Preferred location to work in the company applied
Current_CTC	Current CTC
Inhand_Offer	Holding any offer in hand (Y: Yes, N:No)
Last_Appraisal_Rating	Last Appraisal Rating in current company
No_Of_Companies_worked	No. of companies worked till date
Number_of_Publications	Number of papers published
Certifications	Number of relevant certifications completed
International_degree_any	Hold any international degree (1: Yes, 0: No)
Expected_CTC	Expected CTC (Final CTC offered by Delta Ltd.)

1.3) Exploratory data analysis

✧ **Uni-variate analysis:**

Fig 1.1

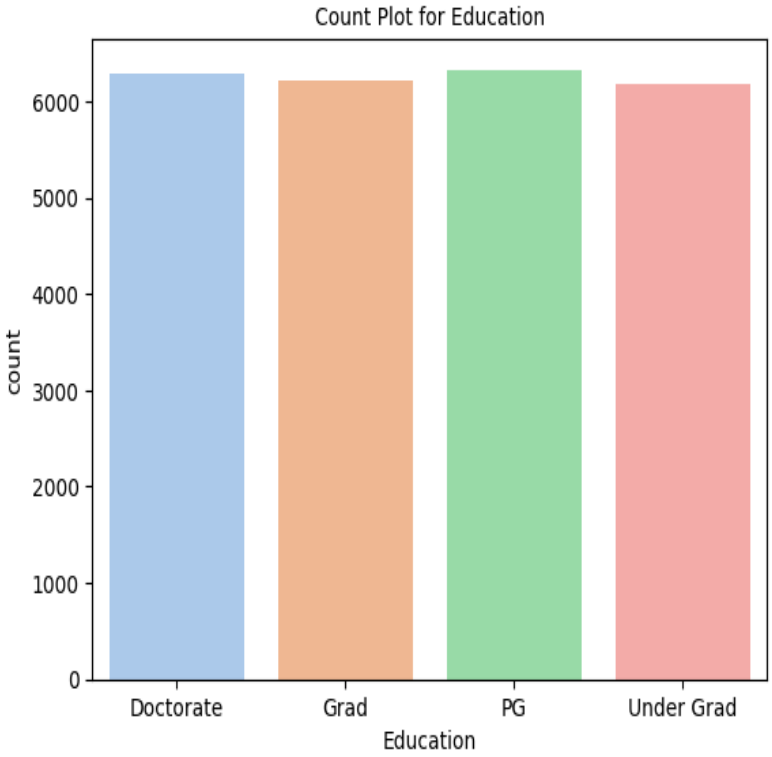


Fig 1.2

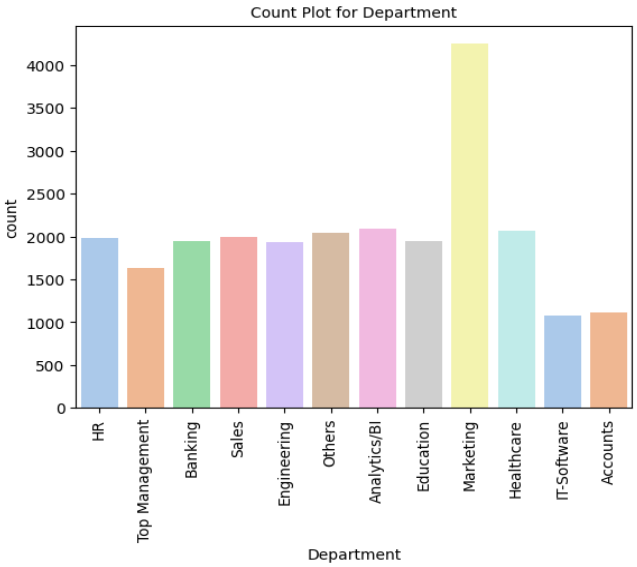
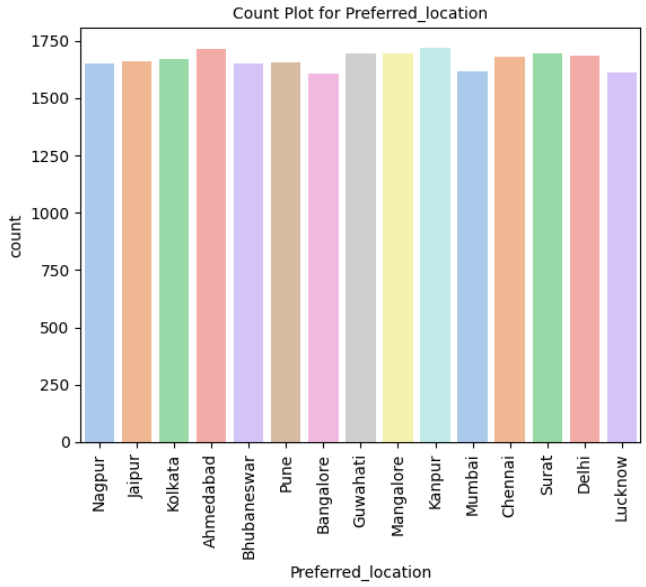


Fig 1.3 &1.4



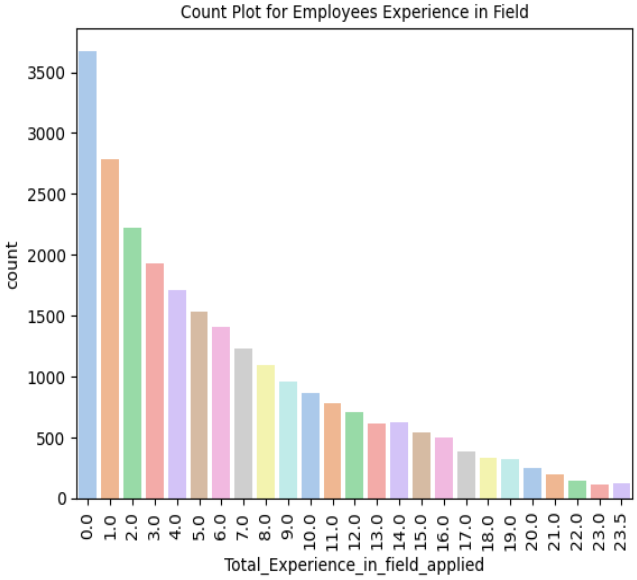


Fig 1.5 & 1.6

- High candidates applied from marketing department.
- 70 % of candidates not have a offer letter of other company.
- Based on preferred location and education not have a much count difference.
- Less than 500 candidates have more than 23 years experience

✧ **Bi-variate analysis:**

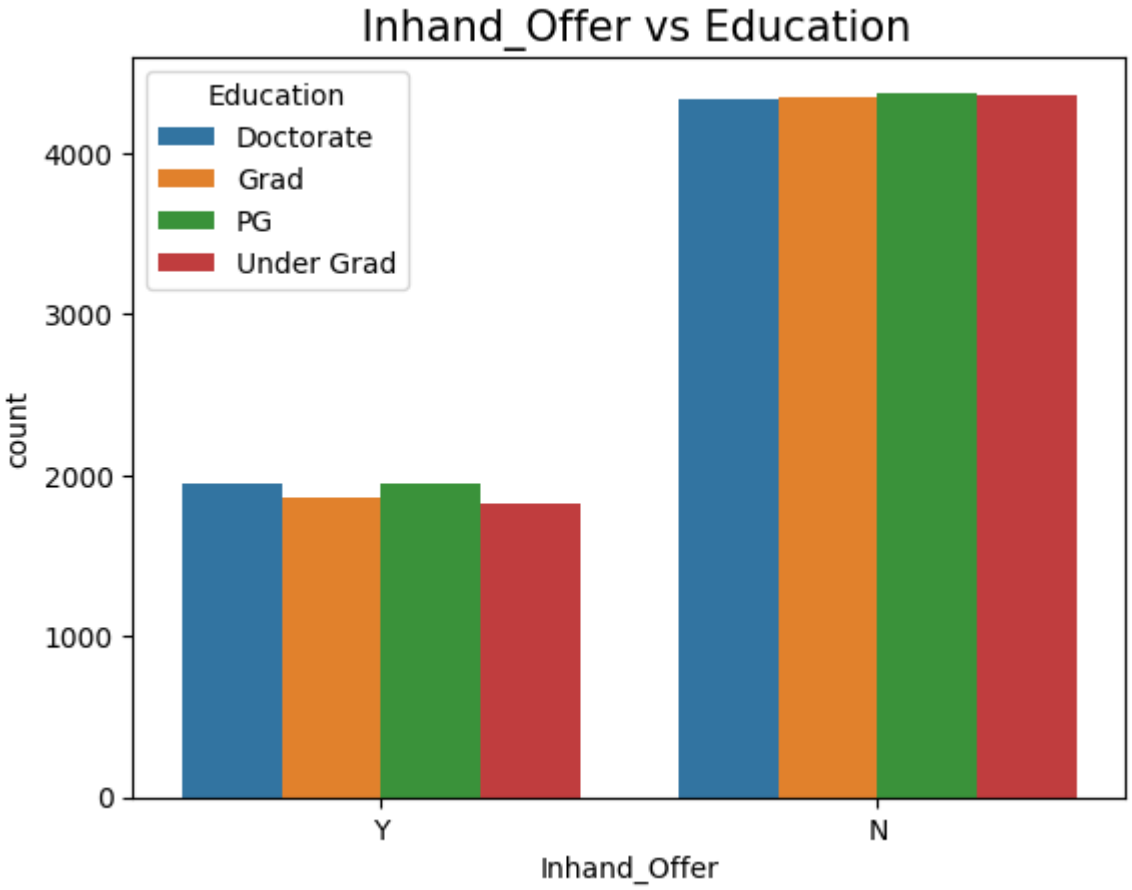


Fig 1.7

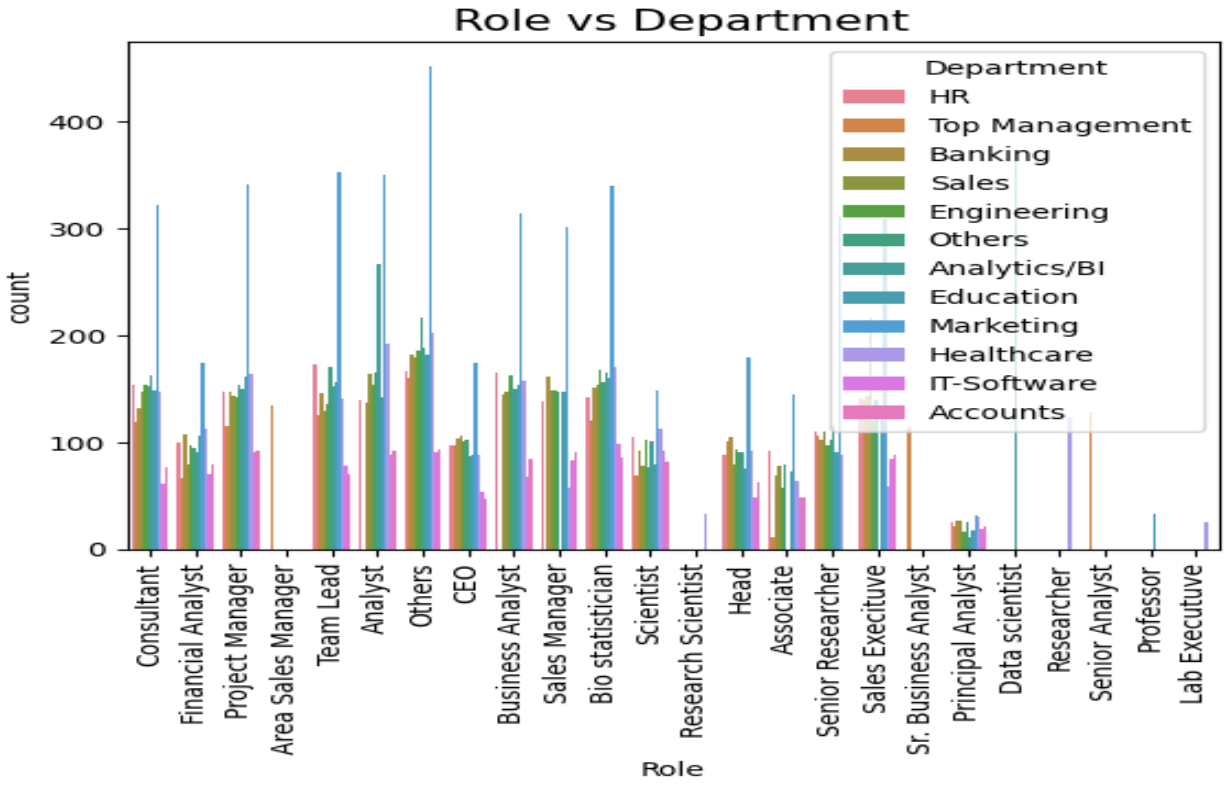


Fig 1.8

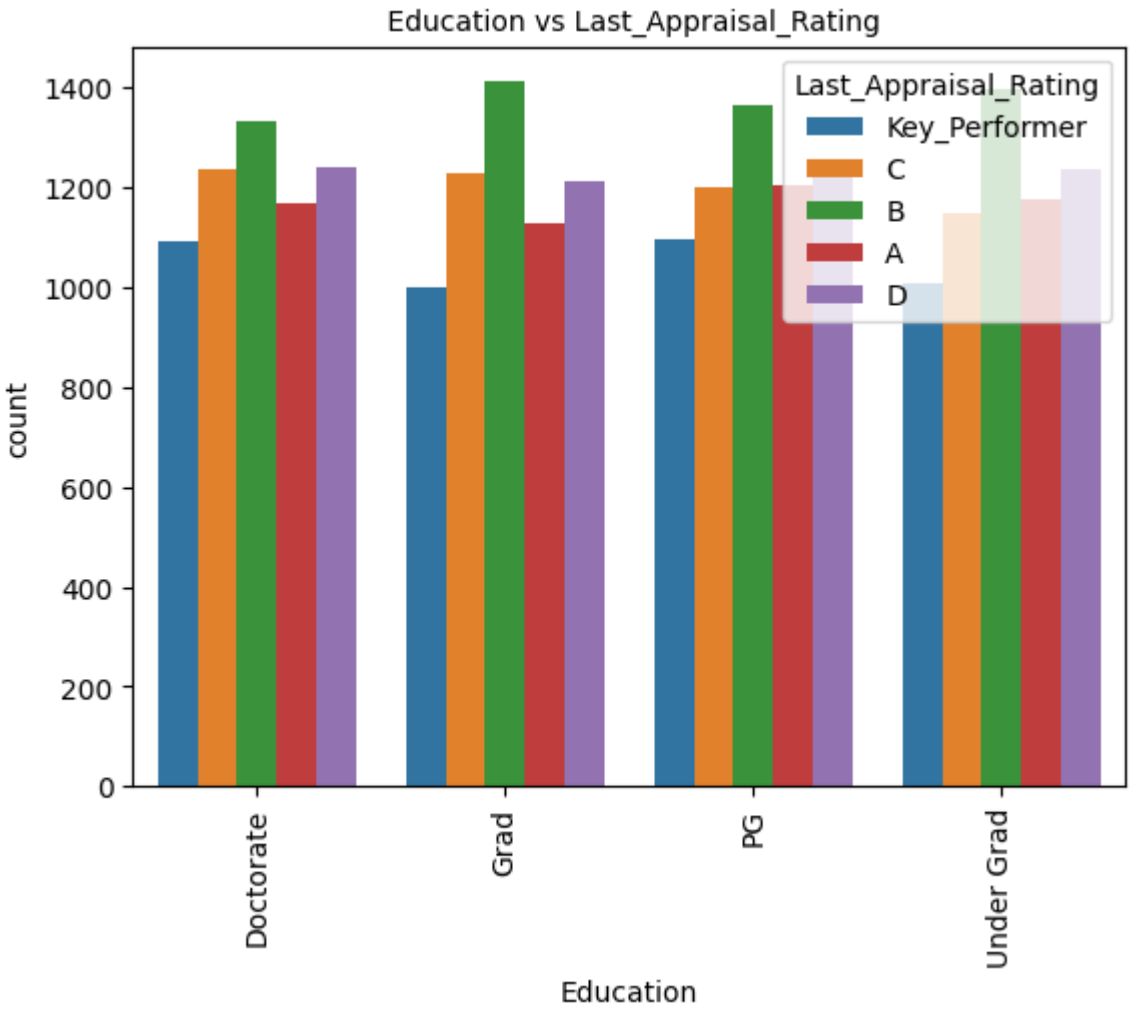


Fig 1.9

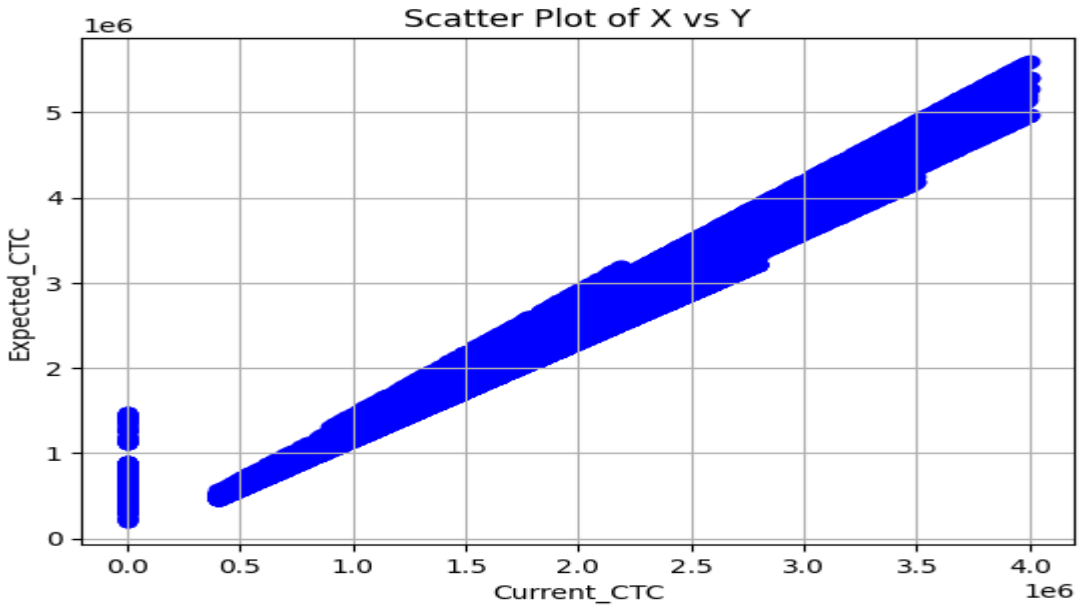


Fig 1.10

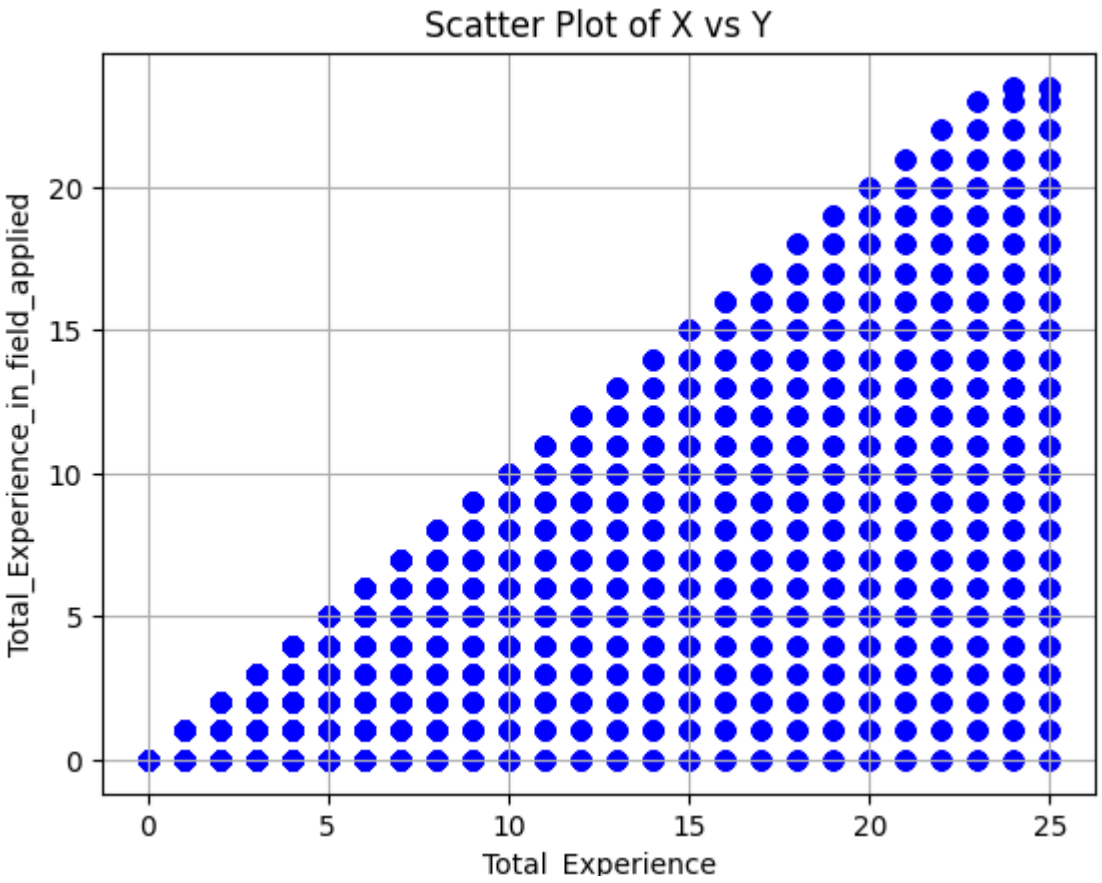


Fig 1.11

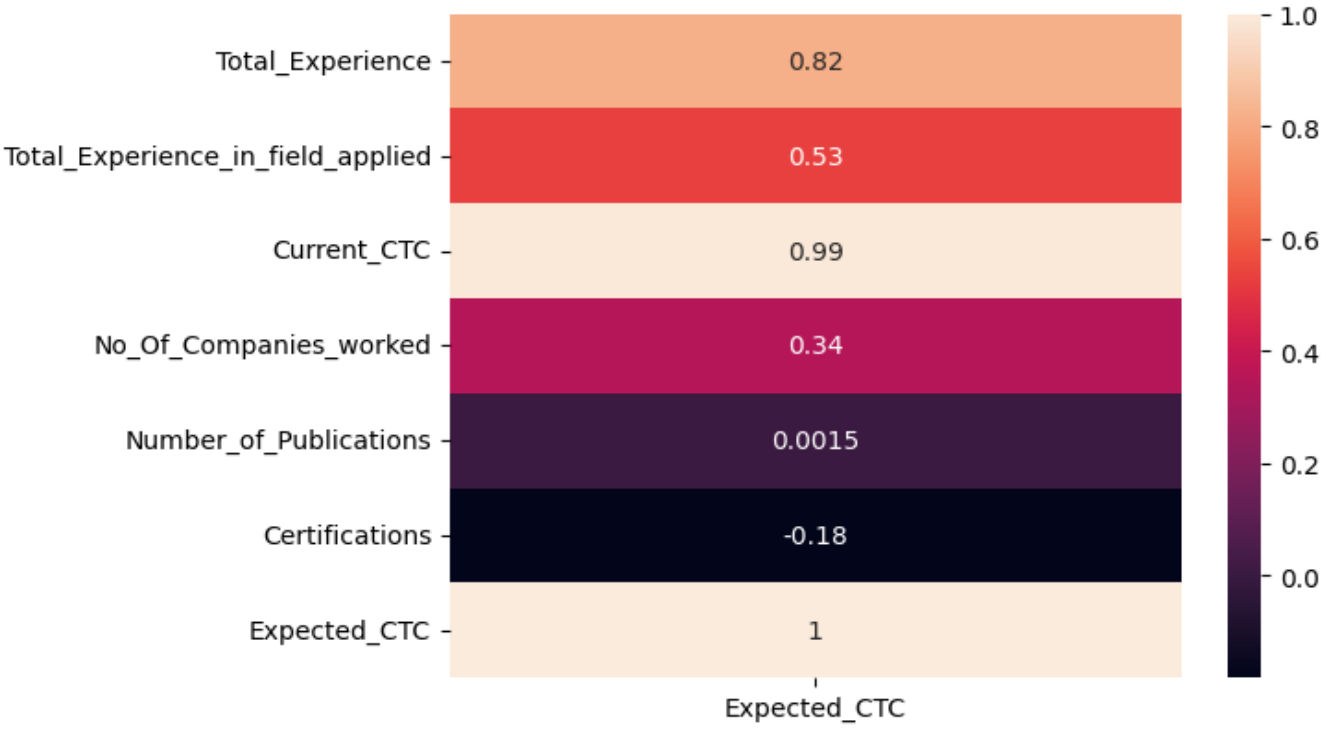


Fig 1.12

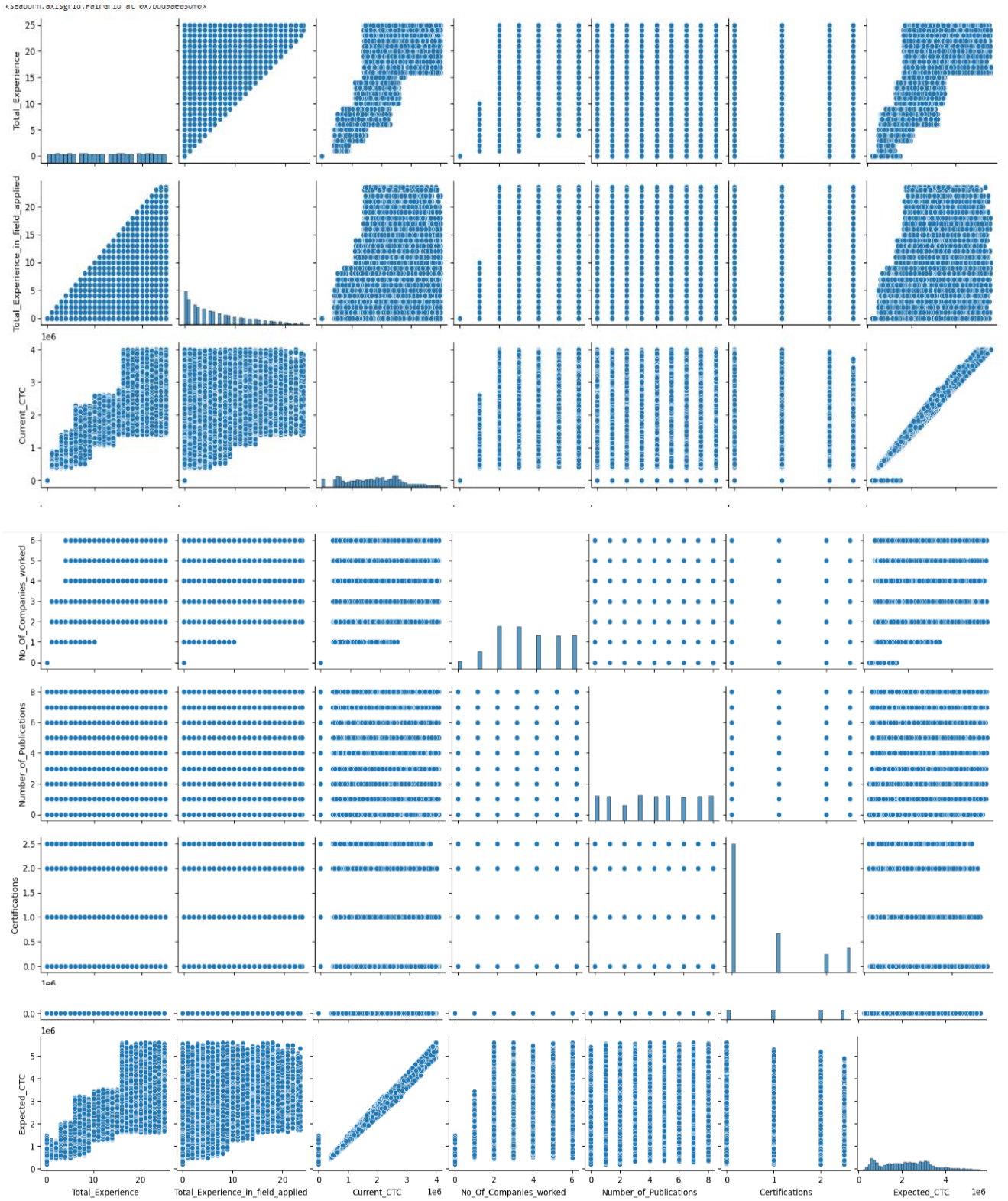

Total_Experience
Total_Experience_in_field_applied
Current_CTC
No_Of_Companies_worked
Number_of_Publications
Certifications
Expected_CTC
1e6

Fig 1.13

1.4) Data Cleaning and Pre-processing.

✧ **Removal of unwanted variables:**

Removing columns

['IDX','Applicant_ID','Organization','Graduation_Specialization','University_Grad','Passing_Year_Of_Graduation','PG_Specialization','University_PG','Passing_Year_Of_PG','PHD_Specialization','University_PHD','Passing_Year_Of_PHD'])

Above values are not have any scope for analysis in the data

✧ **Missing Value treatment:**

Below table we can see that, we have missing values in below columns and Highlighted are more than 30 per of null values

*** Percentage of null values in each column: ***

IDX	0.000
Applicant_ID	0.000
Total_Experience	0.000
Total_Experience_in_field_applied	0.000
Department	11.112
Role	3.852
Industry	3.632
Organization	3.632
Designation	12.516
Education	0.000
Graduation_Specialization	24.720
University_Grad	24.720
Passing_Year_Of_Graduation	24.720
PG_Specialization	30.768
University_PG	30.768
Passing_Year_Of_PG	30.768
PHD_Specialization	47.524
University_PHD	47.524
Passing_Year_Of_PHD	47.524
Curent_Location	0.000
Preferred_location	0.000
Current_CTC	0.000
Inhand_Offer	0.000
Last_Appraisal_Rating	3.632
No_Of_Companies_worked	0.000
Number_of_Publications	0.000
Certifications	0.000
International_degree_any	0.000
Expected_CTC	0.000

dtype: float64

Table 1.4

- We removed the columns more than 30 percentage of null values and for balance replace the null value by using mode

✧ **Outlier treatment:**

Before Outlier treatment

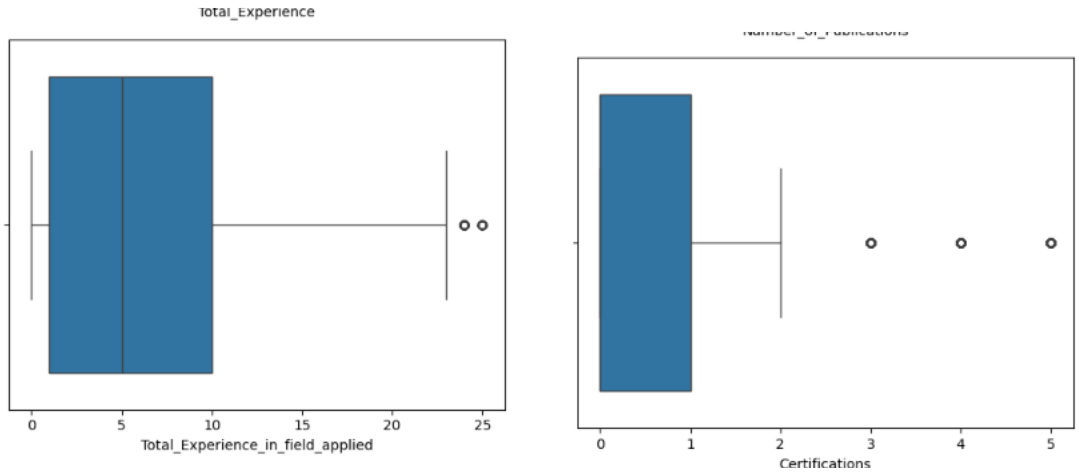


Table 1.5

After outlier treatment:

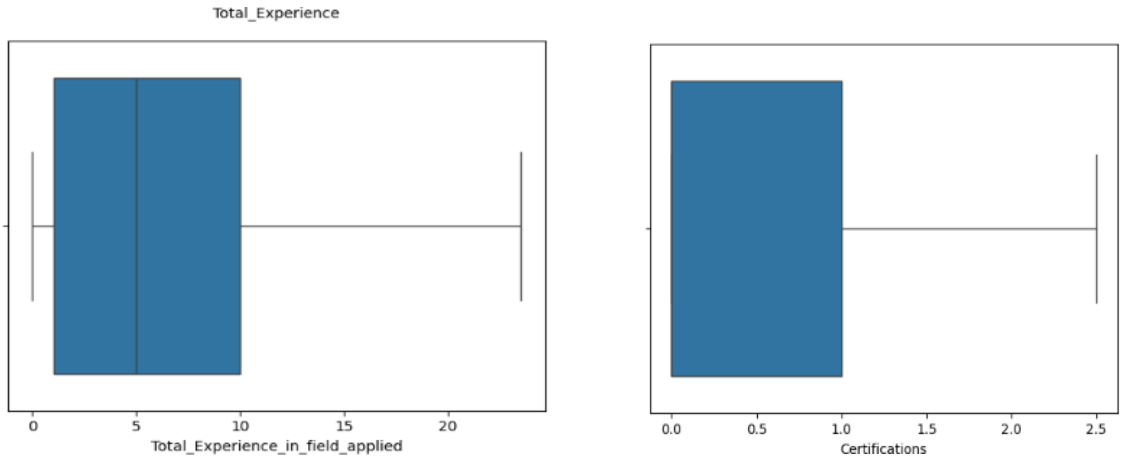


Table 1.6

✧ **Duplicate:**

After removing the unique values, it has been determined that there are **no duplicate** lines present in the data-set.

.

1.5) Model building.

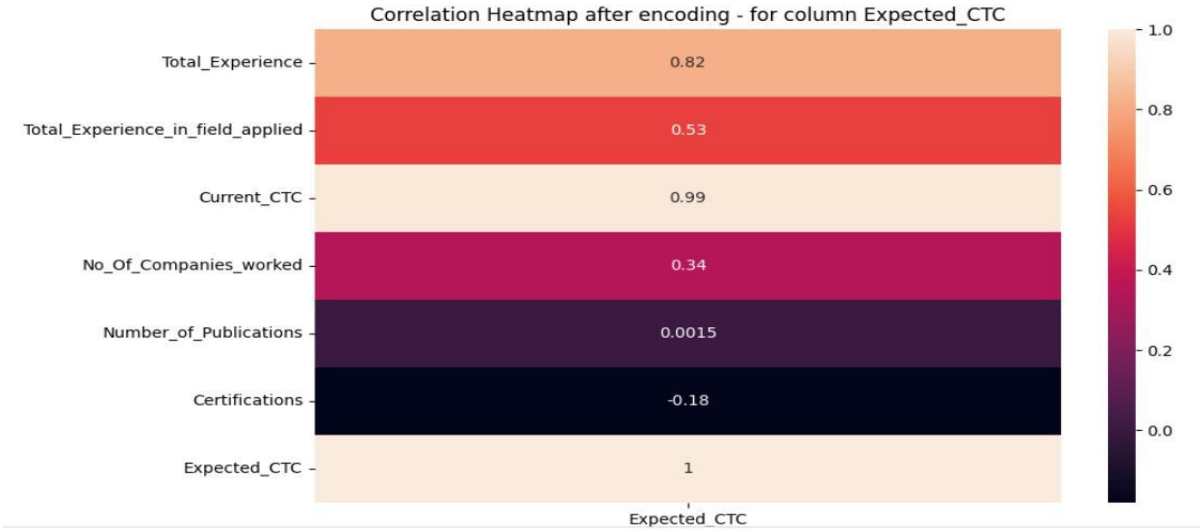
✧ Encoding :

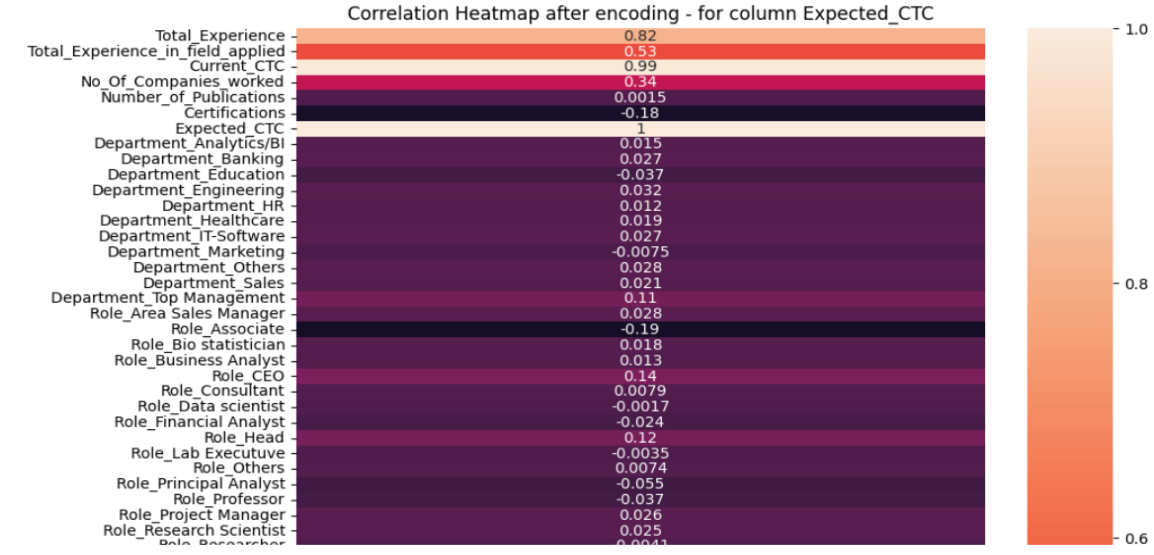
Encoded categorical variables using one-hot encoding. Now, we can proceed to build various models based on the encoded data

```
encoded_data.head()
```

	Total_Experience	Total_Experience_in_field_applied	Current_CTC	No_Of_Companies_worked	Number_of_Publications	Certifications	Expected_CTC	Department_Analytics/BI	Department_Banking
1	23.0	14.0	2702664.0	2.0	4.0	0.0	3783729.0	False	False
2	21.0	12.0	2236661.0	5.0	3.0	0.0	3131325.0	False	False
3	15.0	8.0	2100510.0	5.0	3.0	0.0	2608833.0	False	True
4	10.0	5.0	1931644.0	2.0	3.0	0.0	2221390.0	False	False
5	16.0	3.0	3511167.0	5.0	4.0	0.0	4522383.0	False	False

5 rows × 104 columns





- **Need Split the data into features (X) and target variable (y)**
- **And Split the data into training and testing sets**
- **After standardize the data. Then execution we got the train and test data**

```
y_train
X_test
X_train
```

```
array([[ 0.0656527 ,  0.3015644 , -0.03972808, ...,  2.02681959,
        -0.49239873, -0.44899106],
       [-0.3367719 , -0.21554667,  0.02721713, ..., -0.49338382,
        -0.49239873, -0.44899106],
       [-1.67818724, -1.07739844, -1.91453108, ..., -0.49338382,
        -0.49239873, -0.44899106],
       ...,
       [ 1.13878497,  1.68052724,  1.01688891, ..., -0.49338382,
        -0.49239873, -0.44899106],
       [ 1.40706804,  0.99104582,  1.97471664, ..., -0.49338382,
        -0.49239873,  2.22721585],
       [ 1.67535111,  0.64630511,  0.25436725, ..., -0.49338382,
        -0.49239873, -0.44899106]])
```

Here are a few model types we can consider building:

• Linear Regression
• Decision Tree Regression
• Random Forest Regression
• Gradient Boosting Regression

Regression Models:

We will create and evaluate models for Linear Regression, Decision Tree, Random Forest, Naive Bayes, KNN, AdaBoost, Gradient Boosting, SVM (Support Vector Machine), and XGBoost, with "number" as the target variable. Adjust the models as needed based on your requirements and data

The below are performance output of various models

```
Linear Regression:
Mean Squared Error: 8470038860.953552
R-squared: 0.9937109654893629

Decision Tree:
Mean Squared Error: 2286179737.6734
R-squared: 0.9983025032702

Random Forest:
Mean Squared Error: 1342387003.9047267
R-squared: 0.9990032727909778

Naive Bayes:
Mean Squared Error: 2159013511760.7488
R-squared: -0.6030753467955903

KNN:
Mean Squared Error: 650660093008.285
R-squared: 0.5168825259483165

AdaBoost:
Mean Squared Error: 31701214426.1499
R-squared: 0.9764617335495058

Gradient Boosting:
Mean Squared Error: 2742060224.485547
R-squared: 0.9979640103587325

SVM:
Mean Squared Error: 1346570818386.8357
R-squared: 0.00016629358199682365

XGBoost:
Mean Squared Error: 1181434382.750368
R-squared: 0.9991227806947354
```

Model	R-squared	Mean Squared Error
Linear Regression:	0.9937	8470038860.95
Decision Tree:	0.9983	2286179737.67
Random Forest:	0.9990	1342387003.90
KNN:	0.5169	650660093008.29
AdaBoost:	0.9765	31701214426.15
Gradient Boosting:	0.9980	2742060224.49
SVM	0.0002	1346570818386.83
XGBoost:	**0.9991**	**1181434382.75**

XGBoost Regressor gives the best result as compared to others

Tuning : (Grid Search)

In Model tuning, XGBoost is the best performance

Define a parameter grid containing various hyperparameters for XGBoost
initialize an XGBoost regressor
set up a GridSearchCV object with the XGBoost regressor and the parameter grid, specifying the number of cross-validation folds and the scoring metric
fit the best model to the training data.
make predictions on the test data.
Below are result of tuned of XGBoost Model

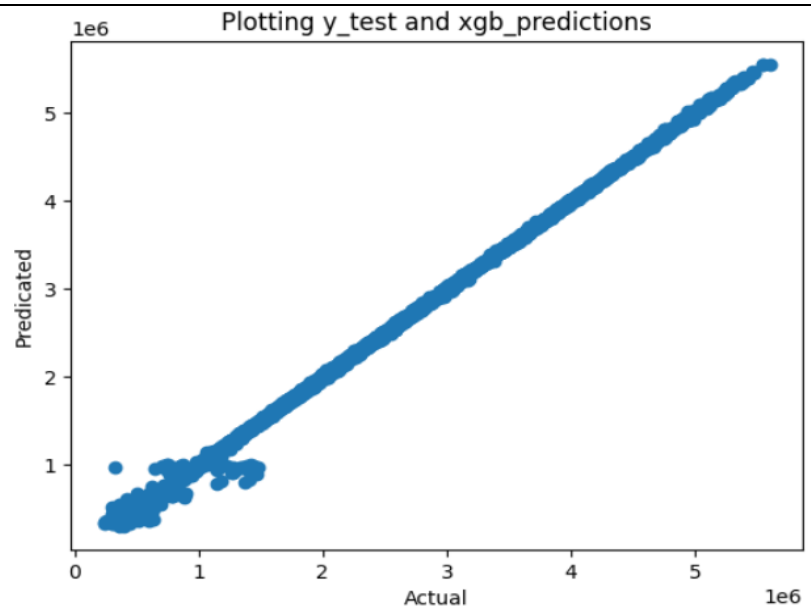
```
MAE: 13282.212008333334
MSE: 1112296503.3477023
RMSE: 33351.10947701294
R-squared: 0.9991741158204288
```

Model validation.

- Among the tree-based models, XG Boost Regressor stands out as the top performer with the highest R2 score and lowest RMSE. Therefore, it's recommended to prioritize the implementation of XG Boosting Regressor due to its superior predictive capabilities.

- MAE: 13282.212008333334
- MSE: 1112296503.3477023
- RMSE: 33351.10947701294
- R-squared: 0.9991741158204288

- While XG Boost seems best for now, it's essential to keep testing and refining models as the Employees data changes or as we learn more about what works best for our specific situation.



1.6) Final interpretation / recommendation.

- **Mostly , Marketing field applicants are highly received**
  - **Based on Inhand offer, most of the 70 % of the employees have another job in hand . so we need to focus the other peoples**
  - **Based on skill key performer applicants need a priority for the job**
  - **Give a importance to Total_Experience_in_field_applied and did not have Inhand_Offer of candidates**
  - **Because have a risk factor of Already placed candidates to accept the job**
  - **across different location, with the Ahmedabad and Kanpur are mostly preferred by candidates**
  - **Mostly focusing on key performer who have did not placed in any companies with high experience**
  - **As per education , mostly doctarate candidates have a high package in companies**
  - **Candidate who applied for Marketing dept is high in count and IT dept have low in count. So for marketing we need to filter the peoples based on rating and performance**