Interim Report of

HR Data Capstone Project

Submitted By

Satya Raga Sudha

Akashatra Sharma

Research Supervisor

Surya Prakash

Contents

Table OF Figures

1. Introduction

HR team plays a crucial role in determining the salary of an employee in an organization. If any of the judgement or consideration goes wrong, it will affect the performance due to employee dissatisfaction which might lead to disengagement of employee. HR team need to manage well to retain the talent in any organization.

In the current situation, people are moving out of organizations frequently and the organization need replacements as well as for new project requirements.

To overcome such problems, if we have some prediction tools which can probably predict the salary details of each employee recruited by the company, such that it will reduce the stress or work carried out by the HR team for negotiating the salary and avoid discrimination in the company.

2. Problem Statement, Objective, Scope & Significance of the project

We have a problem statement related to an organization Delta ltd. The HR team of Delta want to have a system, which can predict the salary of employees, which will lead to no discrimination & employee satisfaction based on their past data, easy to use, avoid manual judgement & effective tool with minimal involvement.

We have a scope of developing a tool, which help them out in solving their issue & reduce their effort in salary calculation. It will be easy to use and avoid manual work out.

The objective, we have here is, to collect past data of all employees of Delta ltd, which are presently used for estimation of Annual salary of an employee by HR. then we understand the data &analyze the data & prepare a model to predict the salary of new employee with similar kind of profile & avoid manual judgement. We test the model by comparing it with existing data as confirmation

3. Data Source and Description

We have collected data (25000 Applicants) from the HR team of Delta ltd. It contains 29 different parameter on which the salary judgement(Expected CTC) is processed. We have observed it contains both numerical & categorical data.

Numerical data – There are 12 Parameters such as Index, Application ID, Total experience, Experience in field, passing years of graduation, PG & PHD, Current CTC, No. of companied worked, No. of publication, certification & expected CTC.

Categorical data - Remaining 17 out of 29 are categorical data. Ordinal categorical data are – Education, Appraisal Rating and Designation. We do have Missing values in Department, Roles, Designation, education, education related columns. Most of the missing values have arisen due to freshers & under graduates. The fresher are outliers.

4. Data Pre-processing

I. Data Info

The data info gave us multiple information. They were as follows :

```
<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)
memory usage: 5.5+ MB
```

Figure 1 : Data Info

- There 3 float data type, 10 integer data type & 16 object data type.
- Many variables were representing null values. Hence, they must be checked upon and a solution to it shall be found out.

```
a. Department
b. Role
c. Industry
d. Organization
e. Designation
f. Graduation_Specialization
g. University_Grad
h. Passing_Year_Of_Graduation
i.PG_Specialization
j. University_PG
K. Passing_Year_Of_PG
L. PHD_Specialization
M. University_PHD
N. Passing_Year_Of_PHD
O. Last_Appraisal_Rating
```

Figure 2 : Null Values Variable Names

- There were 25000 entries with total columns equal to 29. Thus, the data shape is (25000,29).

```
no. of rows:  25000
no. of columns:  29
```

II. Descriptive Statistics

	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

Figure 3 : Data - Descriptive Statistics

Interpretations :

- By comparing mean and median(50%) from the describe function, we got to know that most of the columns are very close to normal distribution . Very slight difference is there in the values of mean and median which put each of the variable into left skewed and right skewed distribution.
- **International_degree_any** had maximum skewness followed by **Certifications** which had 2nd highest skewness in the dataset.

III. Skewness

The skewness of the following variables were as follows:

```
IDX                                  0.000000
Applicant_ID                         0.003409
Total_Experience                     0.004109
Total_Experience_in_field_applied    0.961951
Passing_Year_Of_Graduation           0.061408
Passing_Year_Of_PG                  -0.066166
Passing_Year_Of_PHD                  0.014436
Current_CTC                          0.097643
No_Of_Companies_worked              -0.068026
Number_of_Publications              -0.075217
Certifications                       1.610987
International_degree_any             3.054017
Expected_CTC                         0.331972
dtype: float64
```

Figure 4 : Skewness

IV. Checking For Anomalies / Bad Data

- We first separated the categorical and numerical variables and then check for anomalies/bad data.
- For categorical , the output was as follows :

```
Department : 0
Role : 0
Industry : 0
Organization : 0
Designation : 0
Education : 0
Graduation_Specialization : 0
University_Grad : 0
PG_Specialization : 0
University_PG : 0
PHD_Specialization : 0
University_PHD : 0
Curent_Location : 0
Preferred_location : 0
Inhand_Offer : 0
Last_Appraisal_Rating : 0
```

Figure 5 : Categorical Variable - Check for Anomalies

- For numerical variables, output was as follows :

```
IDX : 0
Applicant_ID : 0
Total_Experience : 0
Total_Experience_in_field_applied : 0
Passing_Year_Of_Graduation : 0
Passing_Year_Of_PG : 0
Passing_Year_Of_PHD : 0
Current_CTC : 0
No_Of_Companies_worked : 0
Number_of_Publications : 0
Certifications : 0
International_degree_any : 0
Expected_CTC : 0
```

Figure 6 : Numerical Variables - Check For Anomalies

Interpretation :

- No external variables was present in the categorical & numerical dataset.
- Hence as per the above code, there were no anomalies/ '?' present in the data set.

V. Descriptive Stats For categorical and numerical variables

- For categorical , it was as follows :

	count	unique	top	freq
Department	22222	12	Marketing	2379
Role	24037	24	Others	2248
Industry	24092	11	Training	2237
Organization	24092	16	M	1574
Designation	21871	18	HR	1648
Education	25000	4	PG	6326
Graduation_Specialization	18820	11	Chemistry	1785
University_Grad	18820	13	Bhubaneswar	1510
PG_Specialization	17308	11	Mathematics	1800
University_PG	17308	13	Bhubaneswar	1377
PHD_Specialization	13119	11	Others	1545
University_PHD	13119	13	Kolkata	1069
Curent_Location	25000	15	Bangalore	1742
Preferred_location	25000	15	Kanpur	1720
Inhand_Offer	25000	2	N	17418
Last_Appraisal_Rating	24092	5	B	5501

Figure 7 : Descriptive Stats : Categorical

- For numerical, it was as follows :

	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

Figure 8 : Descriptive Stats : Numerical

VI. Check For Duplicate Data

- We checked for the duplicate data in the dataset, but there were no duplicates present.

```
Number of duplicate rows = 0
```

Figure 9 : Check For Duplicates

VII. Creating Duplicate Dataset

- We created a duplicate dataset named as "data" and we worked everything on it.
- This was a predetermined precaution in order to avoid any harms/unchangeable actions to our main dataset.

VIII. ANOVA Test

- **Analysis of variance** (**ANOVA**)Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.
- We converted the object data type into categorical data type before performing ANOVA test.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 28 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Applicant_ID                     25000 non-null  int64
 1   Total_Experience                 25000 non-null  int64
 2   Total_Experience_in_field_applied 25000 non-null  int64
 3   Department                       22222 non-null  category
 4   Role                             24037 non-null  category
 5   Industry                         24092 non-null  category
 6   Organization                     24092 non-null  category
 7   Designation                      21871 non-null  category
 8   Education                        25000 non-null  object
 9   Graduation_Specialization        18820 non-null  object
 10  University_Grad                  18820 non-null  object
 11  Passing_Year_Of_Graduation       18820 non-null  float64
 12  PG_Specialization                17308 non-null  object
 13  University_PG                    17308 non-null  object
 14  Passing_Year_Of_PG               17308 non-null  float64
 15  PHD_Specialization               13119 non-null  object
 16  University_PHD                   13119 non-null  object
 17  Passing_Year_Of_PHD              13119 non-null  float64
 18  Curent_Location                  25000 non-null  object
 19  Preferred_location               25000 non-null  object
 20  Current_CTC                      25000 non-null  int64
 21  Inhand_Offer                     25000 non-null  object
 22  Last_Appraisal_Rating            24092 non-null  object
 23  No_Of_Companies_worked           25000 non-null  int64
 24  Number_of_Publications           25000 non-null  int64
 25  Certifications                   25000 non-null  int64
 26  International_degree_any         25000 non-null  int64
 27  Expected_CTC                     25000 non-null  int64
dtypes: category(5), float64(3), int64(9), object(11)
memory usage: 4.5+ MB
```

Figure 10 : Object to Categorical Conversion

- We then performed ANOVA Test on all the relevant variables. They were as follows :

a) Variable : 'Department' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Department' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Department' for the various categories of education are unequal.

b) Variable :'Role' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Role' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Role' are unequal.

c) Variable :'Industry' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Industry' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Industry' are unequal.

Figure 11 :ANOVA TEST 1

d) Variable :'Organization' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Organization' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Organization' are unequal.

e) Variable :'Designation' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Designation' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Designation' are unequal.

Figure 12 : ANOVA TEST 2

- We assumed Level of Significance = 0.05 by default.

Level of significance:

$\alpha = 0.05$

Figure 13 : Level OF Significance

- **Conclusion**:

- As per the ANOVA results, the variables which are found to be in significant are - **'Department'** , **'Role'** , **'Designation'**.
- We also inferred that , the categorical variables which are found to be significant are - **'Industry'** & **'Organization'** .
- As a result, the insignificant variables are not needed in the model building and thus they needed to be dropped from the table.

- We then did the same procedure for other variables :

a) Variable :'Graduation_Specialization' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Graduation_Specialization' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Graduation_Specialization' are unequal.

)b Variable :'University_Grad' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'University_Grad' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'University_Grad' are unequal.

c) Variable :'PG_Specialization' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'PG_Specialization' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'PG_Specialization' are unequal.

Figure 14 : ANOVA TEST 3

d) Variable :'University_PG' -

H_0: μ_{M1} = μ_{M2} = μ_{M3}

H_0: Mean 'University_PG' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'University_PG' are unequal.

e) Variable :'PHD_Specialization' -

H_0: μ_{M1} = μ_{M2} = μ_{M3}

H_0: Mean 'PHD_Specialization' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'PHD_Specialization' are unequal.

f) Variable :'University_PHD' -

H_0: μ_{M1} = μ_{M2} = μ_{M3}

H_0: Mean 'University_PHD' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'University_PHD' are unequal.

Figure 15 : ANOVA TEST 4

- All three variables ‘Passing_Year_Of_Graduation’, ‘Passing_Year_Of_PG’ , ‘Passing_Year_Of_PHD’ are time stamps and are discrete values. Thus, they can be considered as categorical variables also.
- As a result , we would do ANOVA Test to them as well to consider every possibility.

a) Variable :'Passing_Year_Of_Graduation' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Passing_Year_Of_Graduation' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Passing_Year_Of_Graduation' are unequal.

b) Variable :'Passing_Year_Of_PG' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Passing_Year_Of_PG' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Passing_Year_Of_PG' are unequal.

c) Variable :'Passing_Year_Of_PHD' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Passing_Year_Of_PHD' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Passing_Year_Of_PHD' are unequal.

Figure 16 : ANOVA TEST 5

- **Conclusion :**

- As per the ANOVA test, the variables 'Graduation_Specialization', 'University_Grad', 'PG_Specialization' , 'University_PG' , 'PHD_Specialization' & 'University_PHD' were found to be significant and thus they shall not be dropped.

- But as per the ANOVA results for other variables that is the variables which are found to be insignificant are - 'Passing_Year_Of_Graduation' , 'Passing_Year_Of_PG' , 'Passing_Year_Of_PHD'.
- We know that 'Graduation_Specialization' & 'University_Grad' are correlated to 'Passing_Year_Of Graduation' and same for 'PG_Specialization' & 'PHD_Specialization'. So, as 'Passing_Year_Of_Grad' , 'Passing_Year_Of_PG' & 'Passing_Year_Of_PHD' were found to be insignificant that means there correlated variables must also become irrelevant for model building.
- As a result, the insignificant variables are not needed in the model building and thus they needed to be dropped from the table.
- At last we did for these two variables :

a) Variable :'Curent_Location' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Curent_Location' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Curent_Location' are unequal.

b) Variable :'Preferred_location' -

H_0: $\mu_{M1} = \mu_{M2} = \mu_{M3}$

H_0: Mean 'Preferred_location' for an for all the Employees/Individuals are equal.

H_1: $\mu_{M1} \neq \mu_{M2} = \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M2} \neq \mu_{M3}$ or H_1: $\mu_{M1} = \mu_{M3} \neq \mu_{M2}$ or H_1: $\mu_{M1} \neq \mu_{M2} \neq \mu_{M3}$

H_1: At least one of the mean 'Preferred_location' are unequal.

Figure 17 : ANOVA TEST 6

Conclusion :

- As per the ANOVA test, the variables 'Curent_Location' & 'Preferred_location' were found to be significant and thus they shall not be dropped.
- We will group them so as to make work easy and efficient.

IX. Label Encoding

- We did label encoding to ‘Education’ & ‘Inhand_Offer’ variables and transformed it into numerical(integer) format.

X. Dropping All Insignificant Variables

- We dropped all the insignificant variables and then we finish encoding the categorical variables into numerical format so as to build the model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 16 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Applicant_ID                       25000 non-null  int64
 1   Total_Experience                   25000 non-null  int64
 2   Total_Experience_in_field_applied  25000 non-null  int64
 3   Industry                           24092 non-null  category
 4   Organization                       24092 non-null  category
 5   Education                          25000 non-null  int64
 6   Curent_Location                    25000 non-null  object
 7   Preferred_location                 25000 non-null  object
 8   Current_CTC                        25000 non-null  int64
 9   Inhand_Offer                       25000 non-null  object
 10  Last_Appraisal_Rating              24092 non-null  object
 11  No_Of_Companies_worked             25000 non-null  int64
 12  Number_of_Publications             25000 non-null  int64
 13  Certifications                     25000 non-null  int64
 14  International_degree_any           25000 non-null  int64
 15  Expected_CTC                       25000 non-null  int64
dtypes: category(2), int64(10), object(4)
memory usage: 2.7+ MB
```

Conclusion :

- All the insignificant variables had been dropped out successfully.
- Now, we will proceed to further make changes whereever necessary for model building.

Figure 18 :Dropping the Insignificant Variables

- We imputed values in ‘Last_Appraisal_Rating’, ‘Industry’, ‘ Organization’ variables using the mode as there were less than 1% missing values.
- There was no need to do KNN imputation as the they were categorical variables at that time and KNN imputation is valid for only numerical. Moreover atleast 25% of missing values or less should be present in order to use KNN.
- Afterwards, we performed ordinal encoding in it and convert it into numerical format.

XI. Grouping

- We grouped the cities of these variables : ‘Curent_Location’ & ‘Preferred_location’ into Tier_1, Tier2, Tier_3.

	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Industry	Organization	Education	Curent_Location	Preferred_location	Current_CTC	Inhar
0	22753	0	0	NaN	NaN	2	Tier_3	Tier_1	0	
1	51087	23	14	Analytics	H	3	Tier_1	Tier_3	2702664	
2	38413	21	12	Training	J	3	Tier_1	Tier_2	2236661	
3	11501	15	8	Aviation	F	3	Tier_2	Tier_1	2100510	
4	58941	10	5	Insurance	E	1	Tier_1	Tier_1	1931644	

Figure 19 : Grouping

- This will help us easy to work and much reliable than creating dummies as it would become messier if we do with that approach.
- After grouping, we transformed both of them into numerical variables using ordinal encoding.
- We dropped ‘IDX’, ‘Applicant_ID’ also as they were of no use to us.

- Thus, we can now proceed to EDA and after that model building.

5. Exploratory Data Analysis

Graph shown below department & organization as independent variable with reference to expected CTC.

We have considered “Median of expected CTC” for identification of correlation with independent variable**.**

OUTLIERS

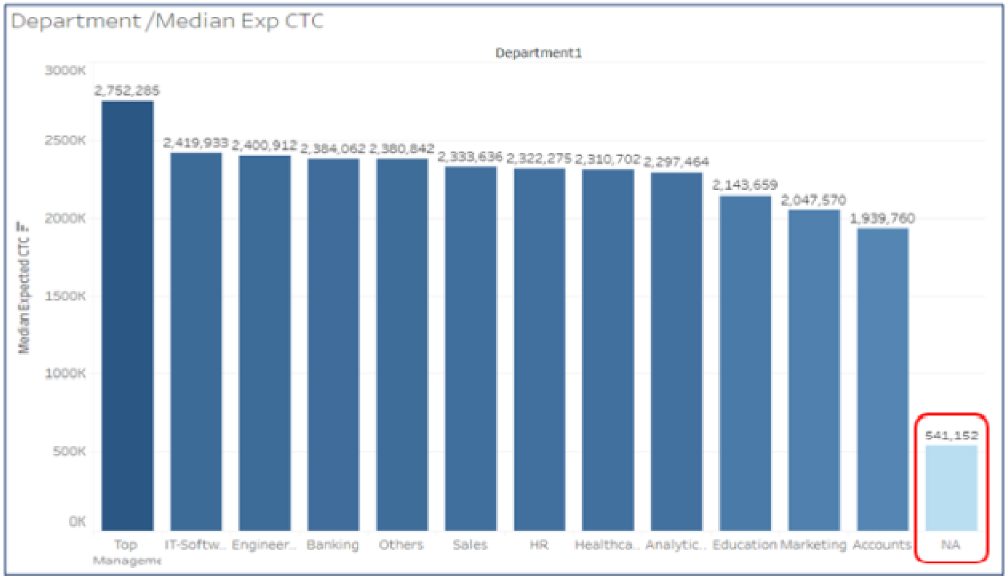


Figure 20 : Department vs Expected_CTC – Tableau

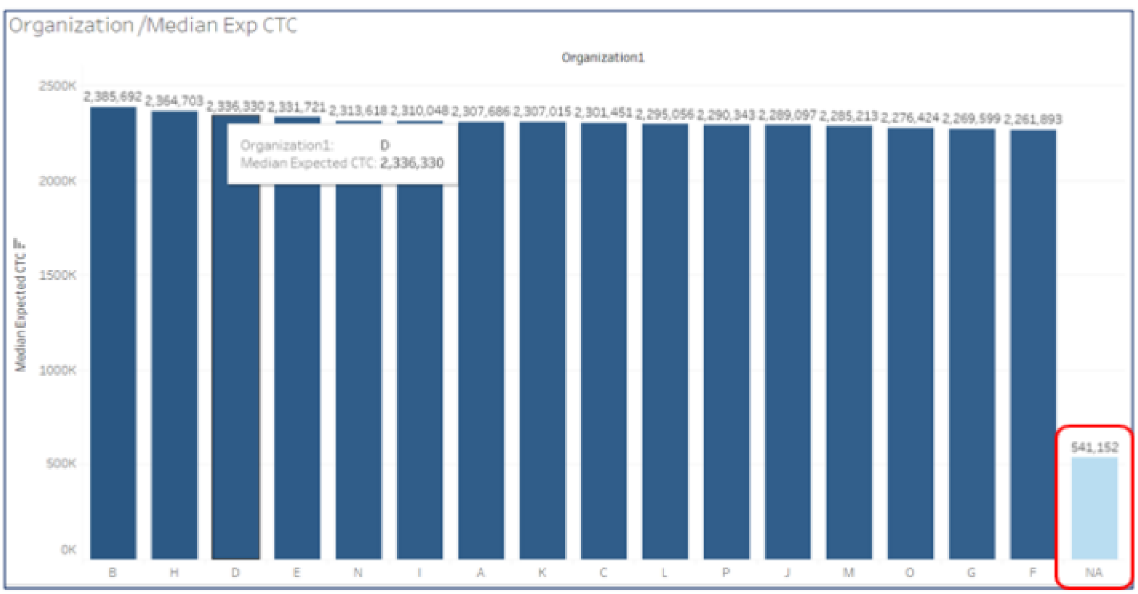


Figure 21 : Organization vs Expected_CTC - Tableau

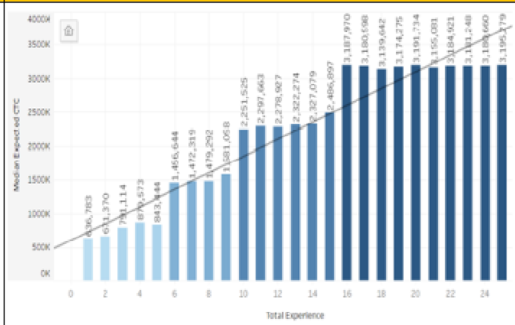
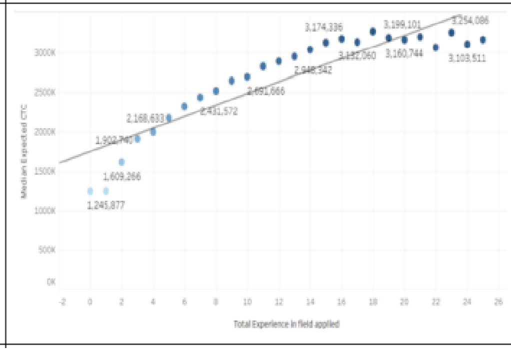
Sl.no.	Column details	Type	Corelation	Insights
1	Total_Experience	Numerical		We have observed quite linear relationship with total experience. Lowest exp median CTC is - 636783. Highest exp median CTC is - 3195279. 80% increase in expected median CTC.
2	Total_Experience_in_field_applied	Numerical		We have observed quite linear relationship with total experience. Lowest exp median CTC is - 1245877. Highest exp median CTC is - 3254086. 61.71% increase in expected median CTC.

Figure 22 : Tableau Insights 1

Correlation Using Python

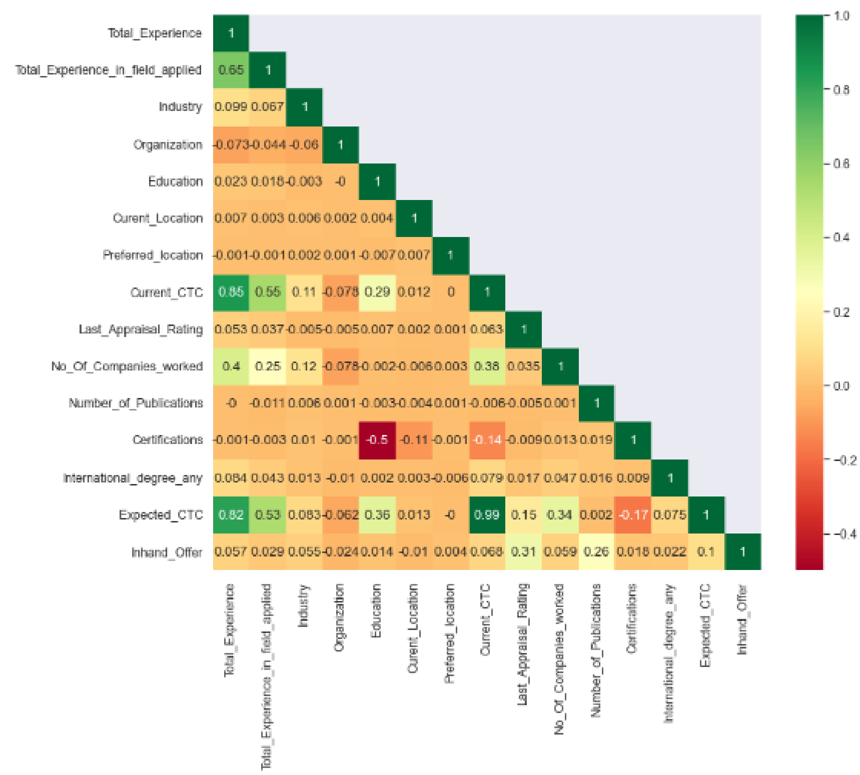


Figure 23 : Correlation Heatmap

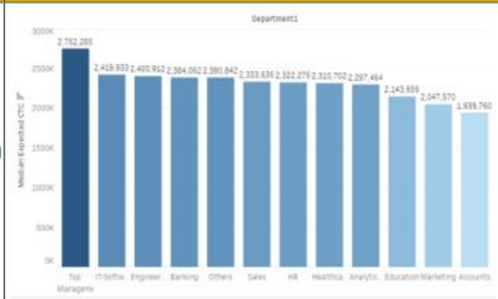
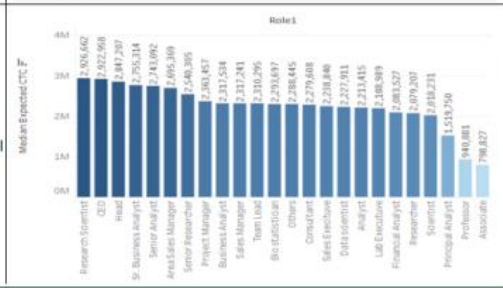
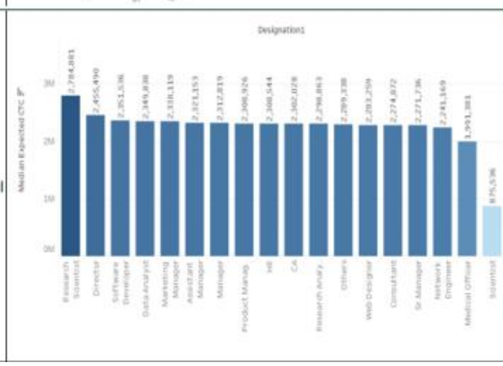
Sl.no.	Column details	Type	Corelation	Insights
3	Department	Categorical		We have observed Top Management has highest exp median CTC- 2752285/- Accounts has lowest exp median CTC - 1939760/- 29% increase in expected median CTC.
4	Role	Categorical		We have observed Reaserch scientist highest exp median CTC- 2926662/- Associate lowest exp median CTC - 798827/- 72.7% Increase in expected median CTC observed in role.
7	Designation	Categorical		We have observed Reaserch Scientist has highest exp median CTC- 2784881/- Scientist has lowest exp median CTC - 875536/- Next to scientist, medical officer has highest salary of 1991381/- further we can consider scientist as outlier & check model accuracy 68.5 % Increase in CTC is observed across designation (considering scientist) 28.5 % Increase in CTC is observed across designation (considering scientist as outlier)

Figure 24 : Correlation - Tableau 1

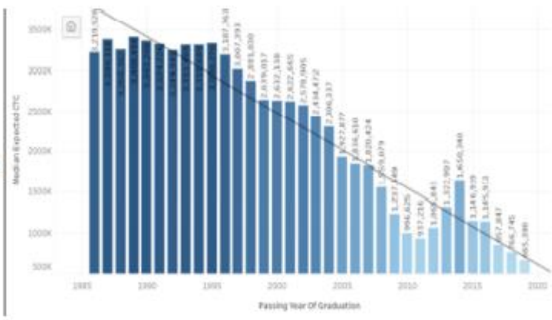
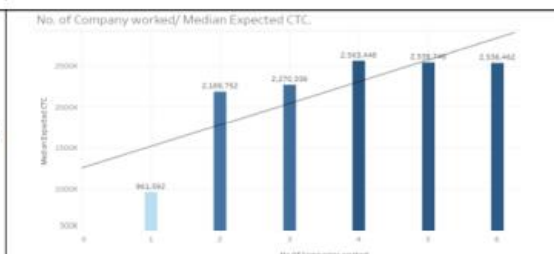
11	Passing_Year_Of_Graduation	Numerical		we have observed 1985 Graduation has higher Median expected CTC of 3219926/- 2019 Graduation has lower Median expected CTC of 665380/- 79.3% Median Expected CTC increase is observed.
23	No_Of_Companies_worked	Numerical		we have observed we have linear relationship with increase with no. of company from 1 year to 961592/- to 6 years 2536462/- 62% increase is observed. But we see a gradual increase in CTC till 4 years , after 4 years we don't see significant increase in CTC.

Figure 25 : Correlation - Tableau 2

Univariate Analysis

- We created histograms for all the variables and then go for the modelling part.

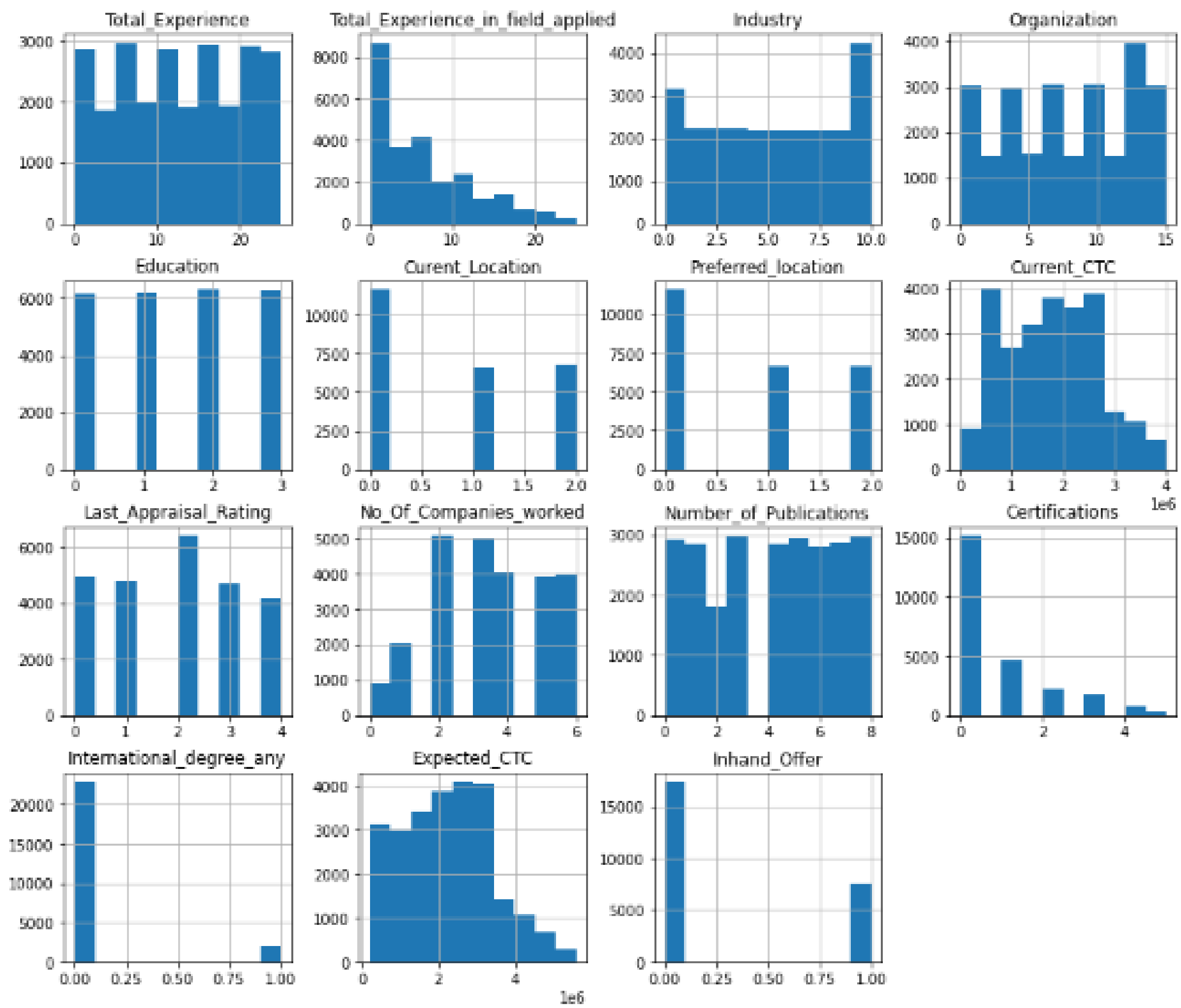


Figure 26 : Univariate Analysis – Histogram

Pairplot

- We plotted the pairplot as it is very important for providing useful insights.
- It also gives us the approach and relation between the variables . Pairplot is as follows :

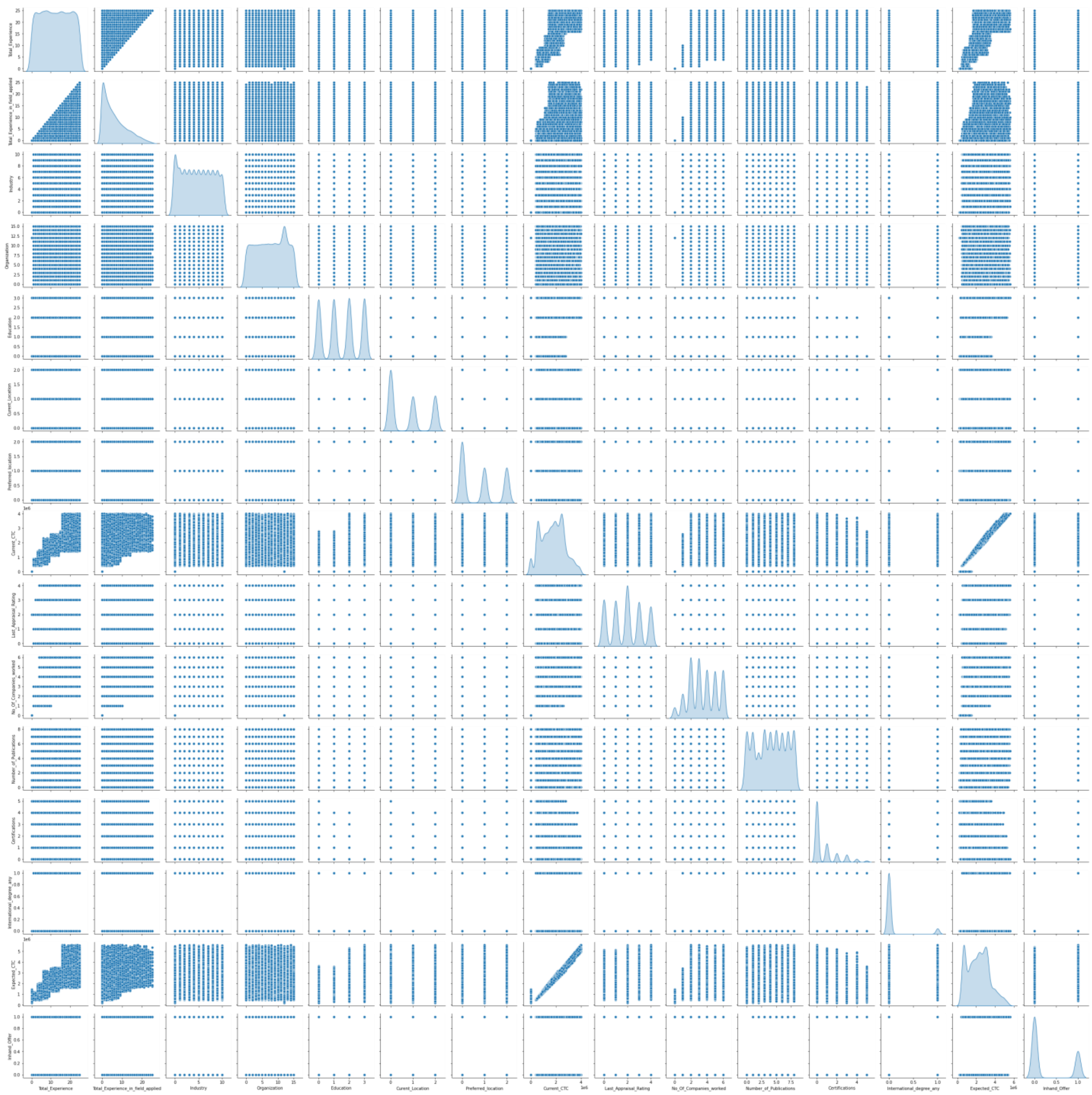


Figure 27 : Pairplot

6. Modelling Approach

- After all data pre-processing and EDA, we finally entered into the model building. But we must know what approach we should take during model building.
- For our problem statement, Linear Regression is the best approach as it will help us predict the correct candidate for the job and get our target variable “Expected_CTC” accurately.
- This problem is a regression based problem so we will go with linear regression . Logistic regression would not be applied as it is not a classification problem.

-	PCA is not required in it as , we are not determined to do dimension reduction. Here we need to build a model that can predict the Expected CTC for a candidate who will apply for the respective company.

-	We did check for outliers after all data pre-processing. We did find in some of them but they were treated .

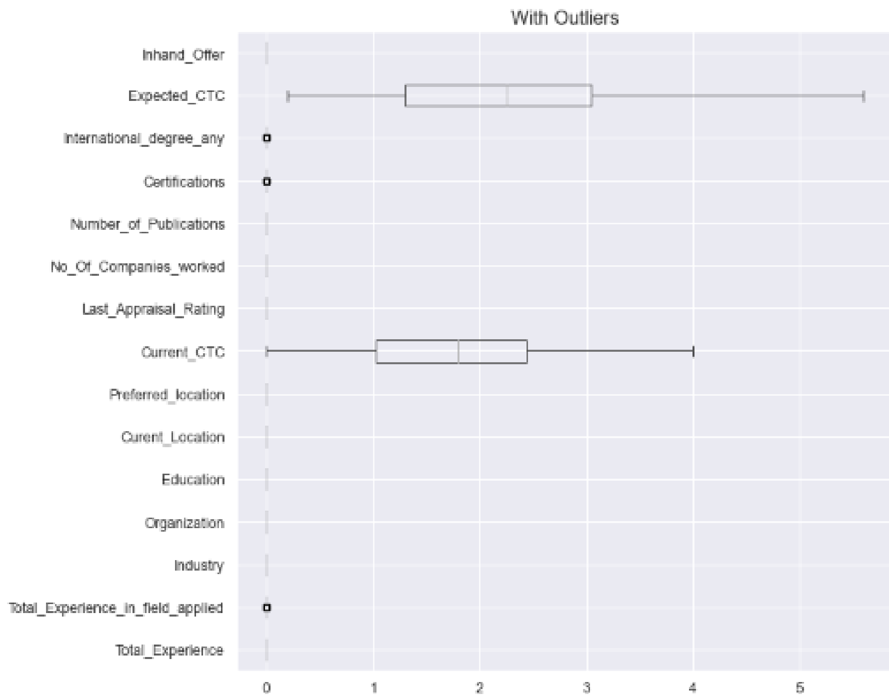


Figure 28 : Outliers Present

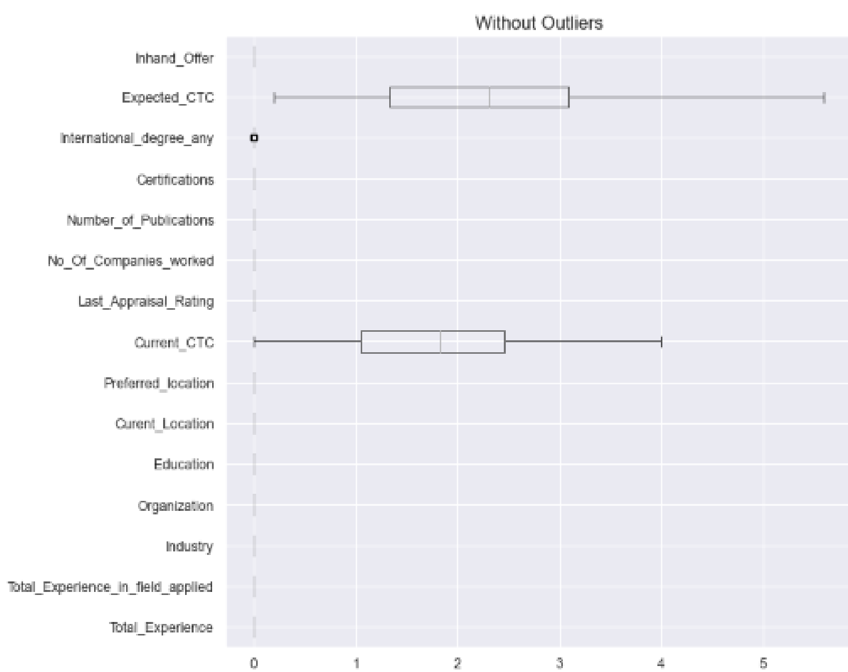


Figure 29 : Outliers not present

Interpretations:

- After inferring insights, it was confirmed that in **"International_degree_any"** no outliers were present. The dots representing them are actual values and that are 0 & 1 only. They cannot be treated as outliers.
- Apart from that, outliers from other variables had been treated well.
- The boxplot with outliers and without outliers can easily be seen after comparing both boxplots.
- After outlier treatment, we builded the base linear regression model using statsmodel library. The output was as follows :

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	1.216e+05
Date:	Sun, 20 Nov 2022	Prob (F-statistic):	0.00
Time:	21:50:05	Log-Likelihood:	-2.8998e+05
No. Observations:	21944	AIC:	5.800e+05
Df Residuals:	21929	BIC:	5.801e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.242e+05	4166.603	-29.818	0.000	-1.32e+05	-1.16e+05
Total_Experience	79.6060	272.488	0.292	0.770	-454.491	613.703
Total_Experience_in_field_applied	-32.7446	204.342	-0.160	0.873	-433.269	367.780
Industry	-7237.4487	281.327	-25.726	0.000	-7788.869	-6686.028
Organization	3087.0167	195.558	15.786	0.000	2703.710	3470.324
Education	7.915e+04	1021.596	77.479	0.000	7.72e+04	8.12e+04
Curent_Location	1919.8534	1073.752	1.788	0.074	-184.778	4024.484
Preferred_location	352.6345	1075.825	0.328	0.743	-1756.061	2461.330
Current_CTC	1.2286	0.002	576.520	0.000	1.224	1.233
Last_Appraisal_Rating	7.479e+04	701.899	106.560	0.000	7.34e+04	7.62e+04
No_Of_Companies_worked	-1.906e+04	584.104	-32.624	0.000	-2.02e+04	-1.79e+04
Number_of_Publications	3040.5282	357.278	8.510	0.000	2340.238	3740.818
Certifications	2767.3163	1473.604	1.878	0.060	-121.054	5655.687
International_degree_any	-1.16e+04	3292.626	-3.524	0.000	-1.81e+04	-5151.016
Inhand_Offer	2.05e+04	2159.064	9.496	0.000	1.63e+04	2.47e+04

Omnibus:	11935.057	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	153747.662
Skew:	2.343	Prob(JB):	0.00
Kurtosis:	15.091	Cond. No.	9.57e+06

Figure 30 : Base Model - Linear Regression

- With this we need to check for multi-colinearity, so we used VIF to check it.

```
Total_Experience  VIF =  5.14
Total_Experience_in_field_applied  VIF =  1.7
Industry  VIF =  1.03
Organization  VIF =  1.01
Education  VIF =  1.52
Curent_Location  VIF =  1.0
Preferred_location  VIF =  1.0
Current_CTC  VIF =  4.92
Last_Appraisal_Rating  VIF =  1.13
No_Of_Companies_worked  VIF =  1.22
Number_of_Publications  VIF =  1.09
Certifications  VIF =  1.19
International_degree_any  VIF =  1.01
Inhand_Offer  VIF =  1.22
```

Figure 31 : VIF - Base Model

- Then we inferred that p_value for many variables were higher than level of significance , so the most insignificant variable must be removed one at a time and better model was created.
- WE did the same procedure and removed insignificant variables until all the p values were gone below level of significance (0.05).
- Then , that model was our **best model.** It was as follows :

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.987
Method:	Least Squares	F-statistic:	1.892e+05
Date:	Sun, 20 Nov 2022	Prob (F-statistic):	0.00
Time:	21:50:09	Log-Likelihood:	-2.8998e+05
No. Observations:	21944	AIC:	5.800e+05
Df Residuals:	21934	BIC:	5.801e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.201e+05	3777.338	-31.792	0.000	-1.27e+05	-1.13e+05
Industry	-7227.1469	281.269	-25.695	0.000	-7778.454	-6675.840
Organization	3092.5457	195.549	15.815	0.000	2709.256	3475.836
Education	7.832e+04	870.851	89.941	0.000	7.66e+04	8e+04
Current_CTC	1.2291	0.001	1113.463	0.000	1.227	1.231
Last_Appraisal_Rating	7.48e+04	701.874	106.567	0.000	7.34e+04	7.62e+04
No_Of_Companies_worked	-1.903e+04	580.300	-32.800	0.000	-2.02e+04	-1.79e+04
Number_of_Publications	3039.9551	357.238	8.510	0.000	2339.743	3740.167
International_degree_any	-1.148e+04	3290.858	-3.487	0.000	-1.79e+04	-5025.612
Inhand_Offer	2.048e+04	2158.925	9.487	0.000	1.62e+04	2.47e+04

Omnibus:	11923.901	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	153498.521
Skew:	2.341	Prob(JB):	0.00
Kurtosis:	15.082	Cond. No.	8.61e+06

Figure 32 : Best Model - Linear Regression

-	Now, all statistical parameters being checked , so this model had all the appropriate things which must be required in our linear regression model.
-	Then we checked for the scatter plot for the predicted values and actual values.

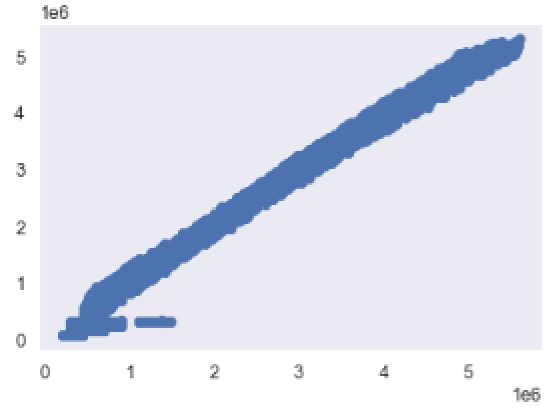


Figure 33 : Scatter Plot

- It was a linear strong relationship scatter plot.

- As per the scatter plot, we inferred that the predicted and actuals are very close to eachother. Hence the R2 is high.

- Density plot was as follows :

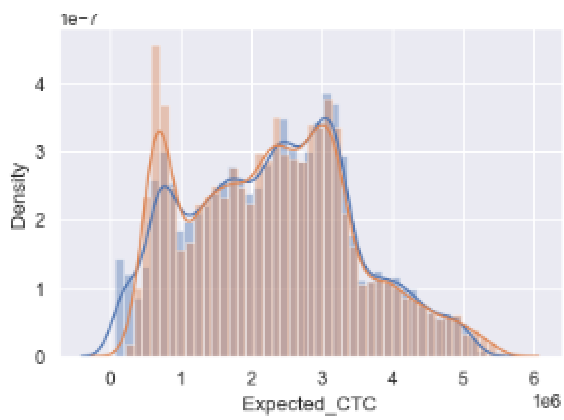


Figure 34 : Density Plot - Expected CTC

Conclusion

- From the distplot, we inferred that the predicted values and actual values were predicting very much likely or in other words we can say they are very much similar to each other.
- It means that our linear regression model was very good.

Note : We also build the linear regression model using Sklearn Library, you can refer to it on Jupyter notebook attached with this report.

7. Actionable insights and recommendations to the stakeholder

We need to identify few insights from EDA & Reason being for such pattern observation.
We need to reduce further the MAE & RMSE values & reduce the difference within them which can be done by identifying further outliers, by elimination of parameter which has minimal relationship with dependent variables & by model tuning.

We convert the data into 70:30 ratio to train, test & verify the model as user experience to validate the model accuracy.

8. References and Bibliography

1. Great Learning class videos
2. Tableau

9. Appendix

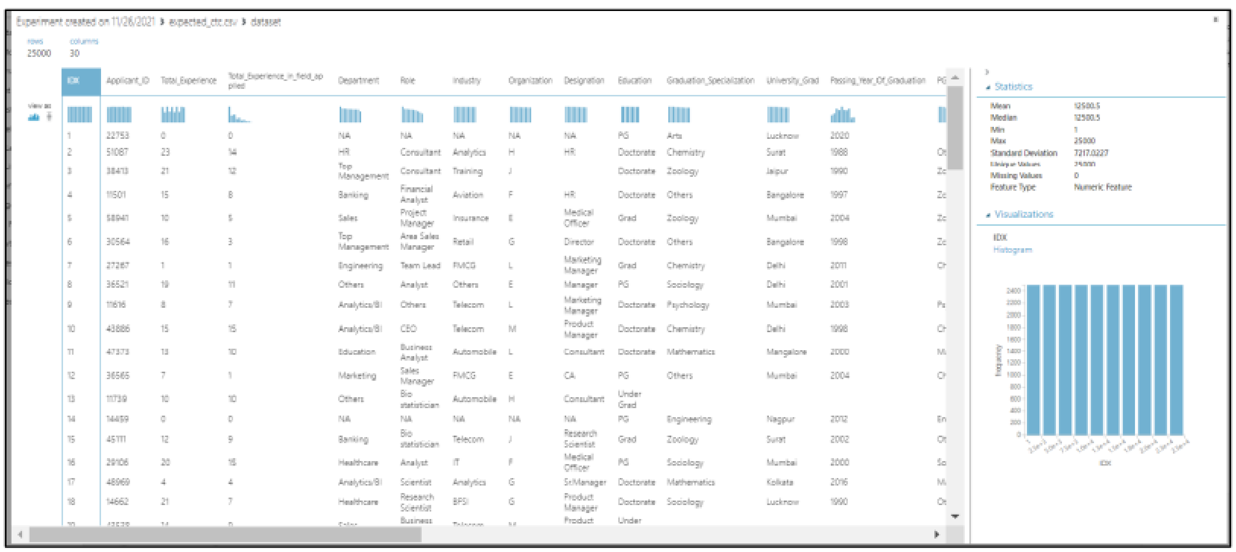


Figure 35 :EDA 1 - Tableau

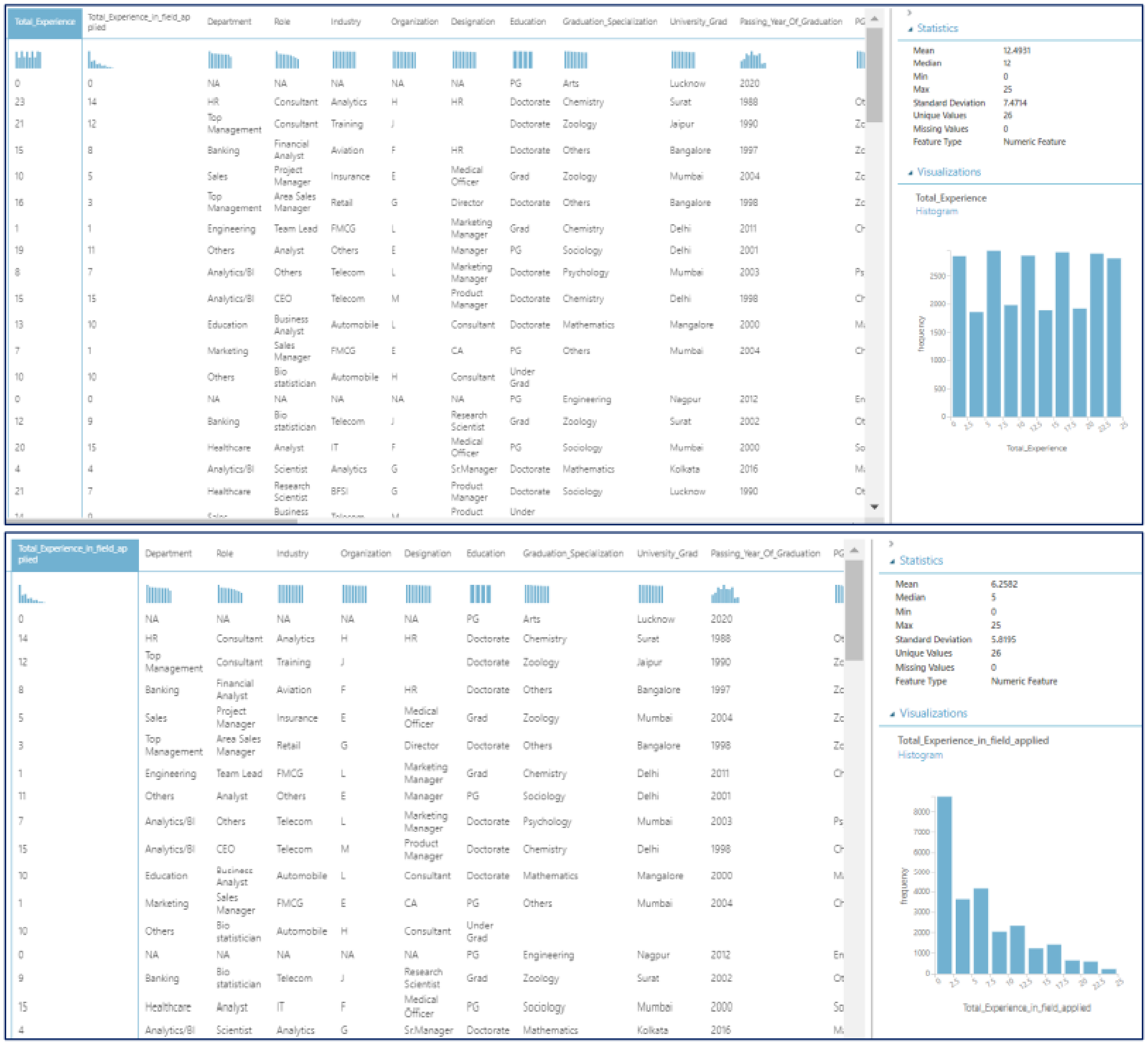


Figure 36 : EDA 2 - Tableau

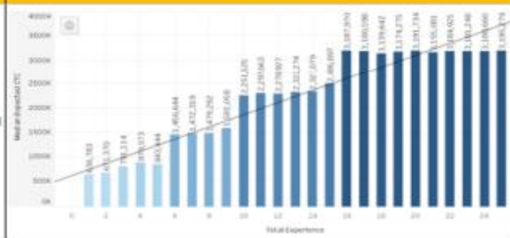
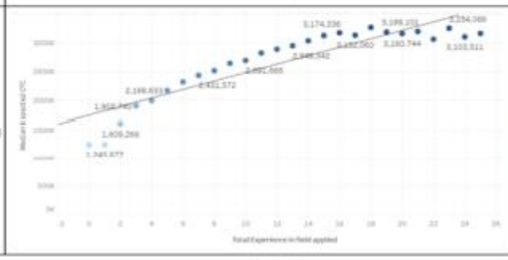
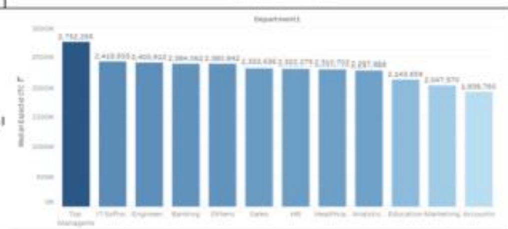
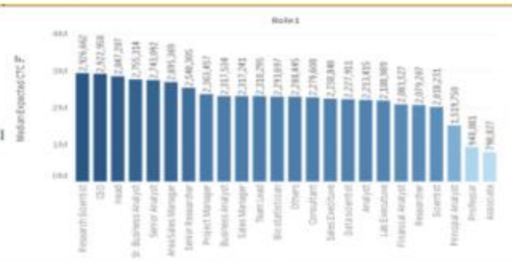
Sl.no.	Column details	Type	Corelation	Insights
1	Total_Experience	Numerical		We have observed quite linear relationship with total experience. Lowest exp median CTC is - 636783. Highest exp median CTC is - 3195279. 80% increase in expected median CTC.
2	Total_Experience_in_field_applied	Numerical		We have observed quite linear relationship with total experience. Lowest exp median CTC is - 1245877. Highest exp median CTC is - 3254086. 61.71% increase in expected median CTC.
3	Department	Categorical		We have observed Top Management has highest exp median CTC- 2752285/- Accounts has lowest exp median CTC - 1939760/- 29% increase in expected median CTC.
4	Role	Categorical		We have observed Reaserch scientist highest exp median CTC- 2926662/- Associate lowest exp median CTC - 798827/- 72.7% Increase in expected median CTC observed in role.

Figure 37 : EDA 3 - Tableau

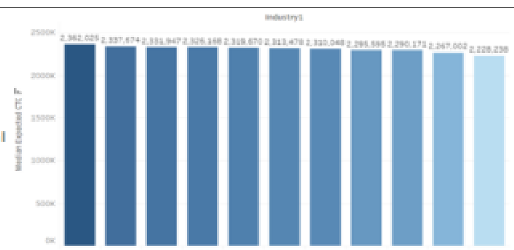
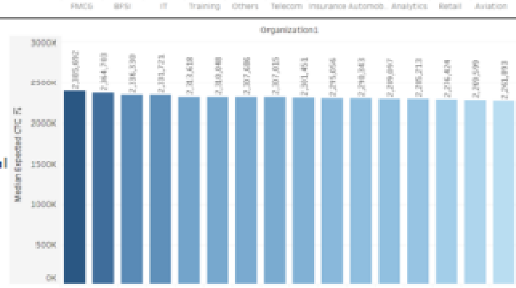
5	Industry	Categorical		We have observed FMCG Industry has highest CTC- 2362025/- Aviation industry has lowest CTC - 2228238/- 6 % Increase in expected median CTC is observed in industry.
6	Organization	Categorical		We have observed B organization has highest exp median CTC- 2385692/- F organization has lowest exp median CTC - 2261893/- 5 % Increase in expected median CTC is observed across Organization.

Figure 38 : EDA 4 - Tableau