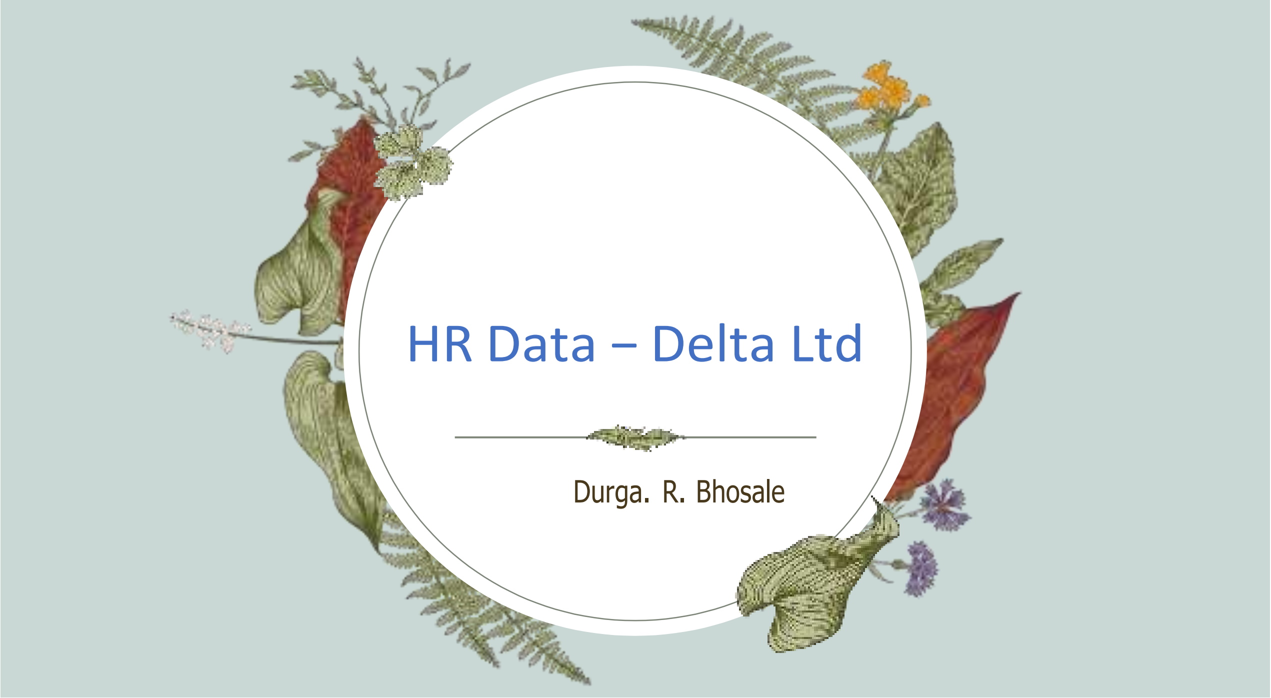


HR Data – Delta Ltd

Durga. R. Bhosale

Introduction of the business problem

a) Defining problem statement
b) Need of the study/project
c) Understanding business/social opportunity

1) **Data Report**

a) Understanding how data was collected in terms of time, frequency and methodology
b) Visual inspection of data (rows, columns, descriptive details)
c) Understanding of attributes (variable info, renaming if required)

2) **Exploratory data analysis**

a) Removal of unwanted variables (if applicable)

b) Missing Value treatment (if applicable)

c) Outlier treatment (if required)

d) Variable transformation (if applicable)

e) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

f) Bivariate analysis (relationship between different variables, correlations)

g) Addition of new variables (if required)

4) Business insights from EDA

a) Business insights using clustering (if applicable)

1) Introduction of the business problem

a) Defining problem statement:

We have data set from the company Delta Ltd of employees related to individuals who applied for job, in order to maintain a salary range for each employee with similar profiles apart from the existing salary, there are various factors which is related to employee’s experience and other metrics like performance evaluation on interviews. By building models we can determine salary that can be offered to the candidate who is selected in the company.

b) Need of the study/project:

Model will help us to minimize human judgement in regards to salary that has been offered.

c) Understanding business/social opportunity:

- Building such models will help us to reduce the discrimination between employees and also it we will be free from human judgements.
- In future suggestions from the model can be considered and then will offer candidates accordingly.
- Every now and then such models as to be to get updated suggestion.11:30 AM

2) Data Report

a) Understanding how data was collected in terms of time, frequency and methodology. Data is collected form the human resource department of Delta Ltd.

b) Visual inspection of data (rows, columns, descriptive details)

Reading and Exploring data:

There are 25000 rows 29 columns in the data set.

c) Understanding of attributes

Data information:

```
<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-    int64
                                          null
```

```
1   Applicant_ID                        25000 non-   int64
                                           null
2   Total_Experience                    25000 non-   int64
                                           null
3                                       25000 non-   int64
    Total_Experience_in_field_appl         null
ied
4   Department                          22222 non-   objec
                                           null      t
5   Role                                24037 non-   objec
                                           null      t
```

```
 6   Industry                        24092 non-     object
                                           null
 7   Organization                    24092 non-     object
                                           null
 8   Designation                     21871 non-     object
                                           null
 9   Education                       25000 non-     object
                                           null
 10                                  18820 non-     object
 Graduation Specialization                 null
 11  University_Grad                 18820 non-     object
                                           null
 12                                  18820 non-     float6
 Passing Year Of Graduation                null     4
 13  PG_Specialization               17308 non-     object
                                           null
 14  University_PG                   17308 non-     object
                                           null
 15  Passing_Year_Of_PG              17308 non-     float6
                                           null     4
 16  PHD_Specialization              13119 non-     object
                                           null
 17  University_PHD                  13119 non-     object
                                           null
 18  Passing_Year_Of_PHD             13119 non-     float6
                                           null     4
 19  Curent_Location                 25000 non-     object
                                           null
 20  Preferred_location              25000 non-     object
                                           null
 21  Current_CTC                     25000 non-     int64
                                           null
 22  Inhand_Offer                    25000 non-     object
                                           null
 23  Last_Appraisal_Rating           24092 non-     object
                                           null
 24  No_Of_Companies_worked          25000 non-     int64
                                           null
 25  Number_of_Publications          25000 non-     int64
                                           null
 26  Certifications                  25000 non-     int64
                                           null
 27                                  25000 non-     int64
 International degree any                  null
 28  Expected_CTC                    25000 non-     int64
                                           null
dtypes: float64(3),        object(16)
int64(10),
       memory usage: 5.5+
       MB
```

Insights:

From the above table we can observe that 16 variables have object data type, 12 variables have numeric data type, also we can observe data set has null values.

Describe the columns :

ID **X**	**Applicant_ID**

	count	mean	std	min	25%	50%	75%	max
	25000.0	1.250050e+04	7.217023e+03	1.0	6250.75	12500.5	18750.25	25000.0
	25000.0	3.499324e+04	1.439027e+04	10000.0	22563.75	34974.5	47419.00	60000.0
Ťotal_Expeíience	25000.0	1.249308e+01	7.471398e+00	0.0	6.00	12.0	19.00	25.0
Ťotal_Expeíience_in_field_applied	25000.0	6.258200e+00	5.819513e+00	0.0	1.00	5.0	10.00	25.0
Passing_Yeaí_Of_Gíaduation	18820.0	2.002194e+03	8.316640e+00	1986.0	1996.0	2002.0	2009.0	2020.0
Passing_Yeaí_Of_PG	17308.0	2.005154e+03	9.022963e+00	1988.0	1997.0	2006.0	2012.0	2023.0

	count	mean	std	min	25%	50%	75%	max
Passing_Yeaí_Of_PHD	13119.0	2.007396e+03	7.493601e+00	1995.0	2001.00	2007.0	2014.00	2020.0
Cuííent_CI̋C	25000.0	1.760945e+06	9.202125e+05	0.0	1027311.50	1802567.5	2443883.25	3999693.0
No_Of_Companies_woíked	25000.0	3.482040e+00	1.690335e+00	0.0	2.00	3.0	5.00	6.0
Numbeí_of_Publications	25000.0	4.089040e+00	2.606612e+00	0.0	2.00	4.0	6.00	8.0
Ceítifications	25000.0	7.736800e-01	1.199449e+00	0.0	0.00	0.0	1.00	5.0
Inteínational_degíee_any	25000.0	8.172000e-02	2.739431e-01	0.0	0.00	0.0	0.00	1.0
Expected_CI̋C	25000.0	2.250155e+06	1.160480e+06	203744.0	1306277.50	2252136.5	3051353.75	5599570.0

From the above table we can find 5-point summary of the data and we can also detect outliersforfew variables. However, in our case we are not going to treat null values because we need actual data for further analysis to predict accurately.

3) Exploratory data analysis

a) Removal of unwanted variables.

Null Values Check:

```
IDX                                    0
Applicant_ID                           0
Total_Experience                       0
Total_Experience_in_field_appl         0
ied
Department                          2778
Role                                 963
Industry                             908
Organization                         908
Designation                         3129
Education                              0
```

```
 Graduation_Specialization        6180
 University_Grad                  6180
 Passing_Year_Of_Graduation       6180
 PG_Specialization                7692
 University_PG                    7692
 Passing_Year_Of_PG               7692
 PHD_Specialization              11881
 University_PHD                  11881
 Passing_Year_Of_PHD             11881
 Curent_Location                     0
 Preferred_location                  0
 Current_CTC                         0


Inhand_Offer                        0
Last_Appraisal_Rating             908
No_Of_Companies_worked              0
Number_of_Publications              0
Certifications                      0
International_degree_any            0
Expected_CTC                        0
dtype: int64
```

Inference:

From the above table we can observe that variables such as Department, Role, Industry, Organization, Designation, Graduation specialization, University grad, passing year ofgraduation, PG Specialization, University PG, passing year of PG, PHD Specialization, University PHD, Passing year of PHD, Last Appraisal rating have null values, for our analysis purpose we do not required variables as follows IDX, Applicant ID, Organization ,Graduationspecialization, University grad, passing year of graduation, PG Specialization, University PG, passing year of PG, PHD Specialization, University PHD, Passing year of PHD. Hence, we are dropping those variables.

Rest we have Department, Role, Industry, Designation, Last Appraisal rating, these columnsnull values have to be treated.

No Duplicated data detected from the data set.

Treating Null Values.

Variable ‘Department’ was treated by imputing value ‘Others ‘in place of Null. So, it gets included in ‘Others’ category.

Variable ‘Role’ was treated by imputing value ‘Others ‘in place of Null. So, it gets included in ‘Others’ category.

Variable ‘Industry’ was treated by imputing value ‘Fresher’ in place of Null.

Variable ‘Designation’ was treated by imputing value ‘Others ‘in place of Null. So, it gets included in ‘Others’ category.

Variable ‘Last Appraisal rating’ was treated by imputing value ‘Fresher’ in place of Null.

After treating Null Values

```
Total_Experience                 0
Total_Experience_in_field_applied 0
Department                       0
Role                             0
Industry                         0
Designation                      0
Education                        0
Curent_Location                  0
Preferred_location               0
Current_CTC                      0
Inhand_Offer                     0
Last_Appraisal_Rating            0
```

```
No_Of_Companies_worked          0
Number_of_Publications          0
Certifications                  0
International_degree_any        0
Expected_CTC                    0
dtype: int64
```

Inference:

From the above table we can observe there is no values in the data set.
In this data set we can observe there is no value in the data set.

- Outlier treatment

Outlier treatment not require in this data set, because we need accurate value to build better model.

- Variable transformation.

Data Information:

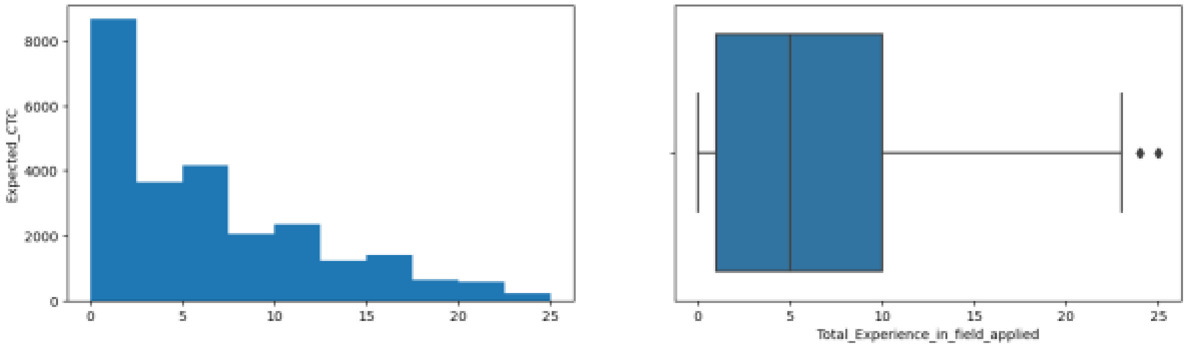
```
<class
'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to
24999 Data  columns  (total  17
columns):
 #   Column                        Non-Null Count  Dtype
-----  ---------                   ----------------  -------
 0   Total_Experience              25000 non-null  int64
 1   Total_Experience_in_field_applied  25000 non-null  int64
 2   Department                    25000 non-null  object
 3   Role                          25000 non-null  object
 4   Industry                      25000 non-null  object
 5   Designation                   25000 non-null  object
 6   Education                     25000 non-null  object
 7   Curent_Location               25000 non-null  object
 8   Preferred_location            25000 non-null  object
 9   Current_CTC                   25000 non-null  int64
 10  Inhand_Offer                  25000 non-null  int8
 11  Last_Appraisal_Rating         25000 non-null  object
 12  No_Of_Companies_worked        25000 non-null  category
 13  Number_of_Publications        25000 non-null  int64
 14  Certifications                25000 non-null  int64
 15  International_degree_any      25000 non-null  category
 16  Expected_CTC                  25000 non-null
int64dtypes: category(2), int64(6), int8(1),
object(8)
memory usage: 2.7+ MB
```

Observation:

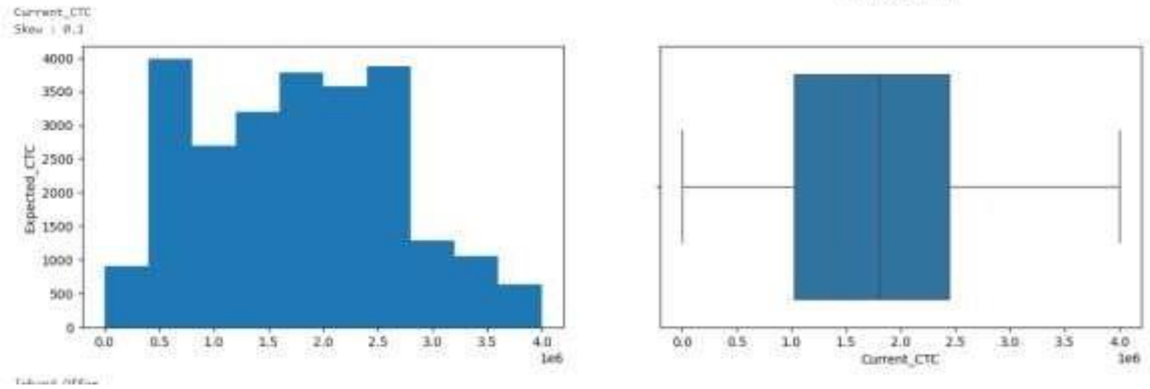
From the above table we can observe that No of companies worked and international degree have been changed from integer to categorical variable for further Analysis.

- Univariate analysis (distribution and spread for every continuous attribute, distribution of data incategories for categorical ones)

Bivariate analysis for total Experience in applied field:

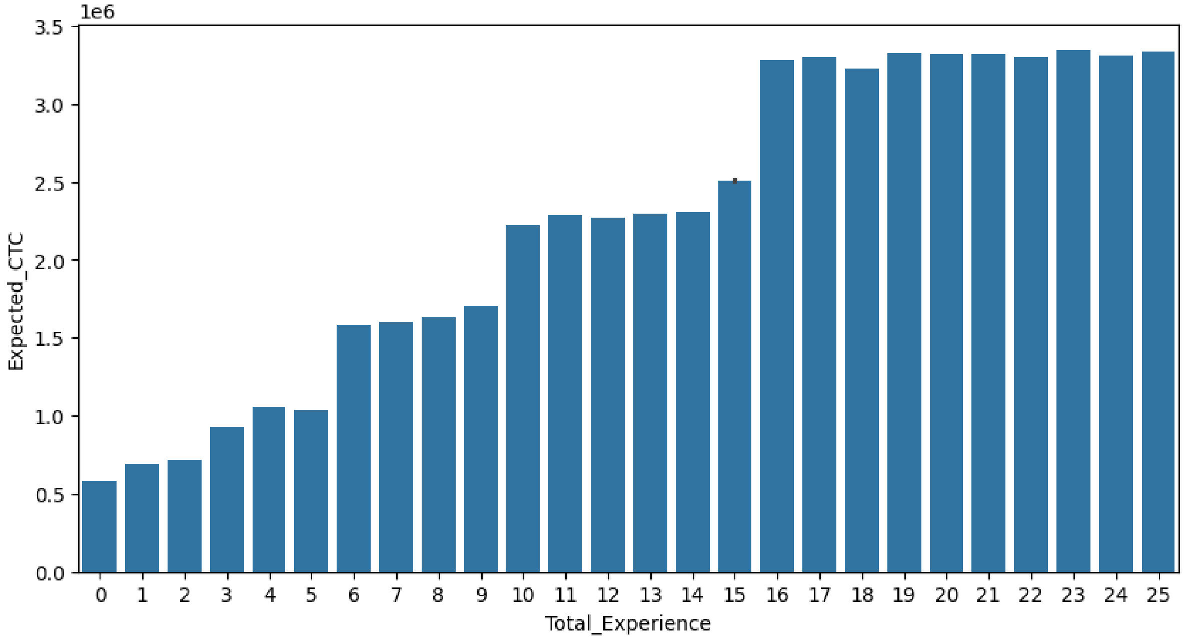


Bivariate analysis for current CTC:



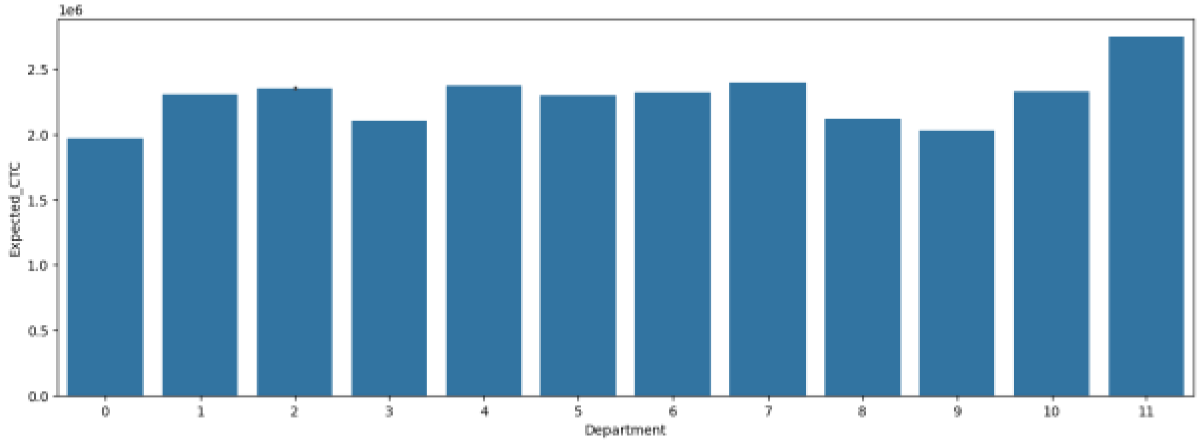
Candidates who have less CTC also expects high expectations.

Bivariate for total experience:



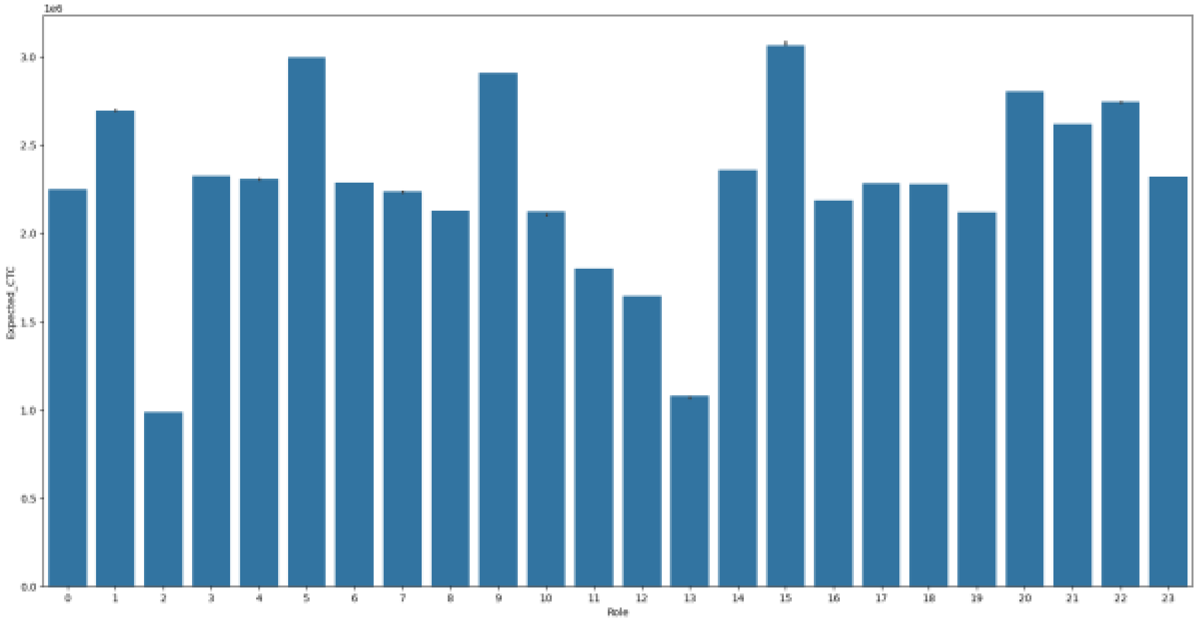
Candidates who have high experience have high CTC expectations.

Department:



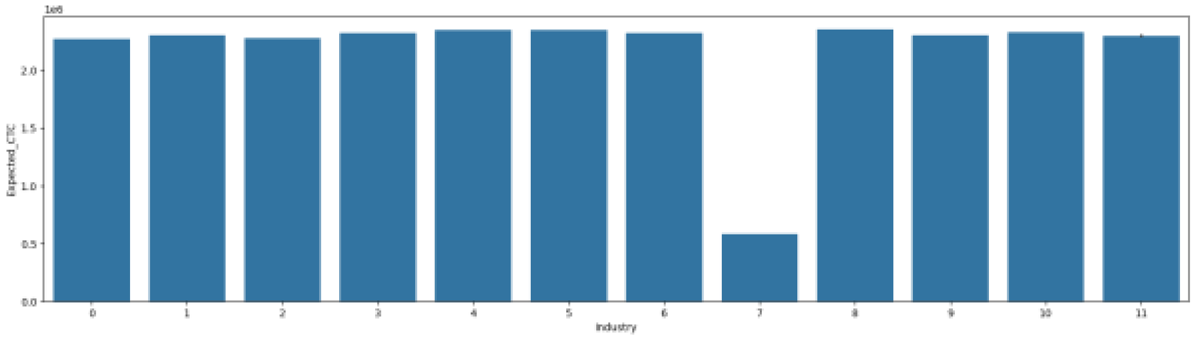
Candidates from top management have high CTC Expectations.

Role:



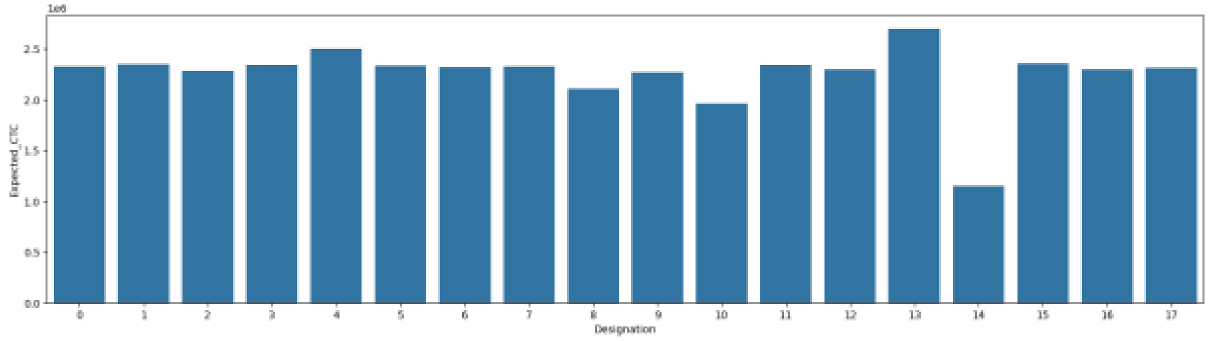
Research scientist have high CTC expectations, whereas associate professors have lessCTCexpectations.

Industry:



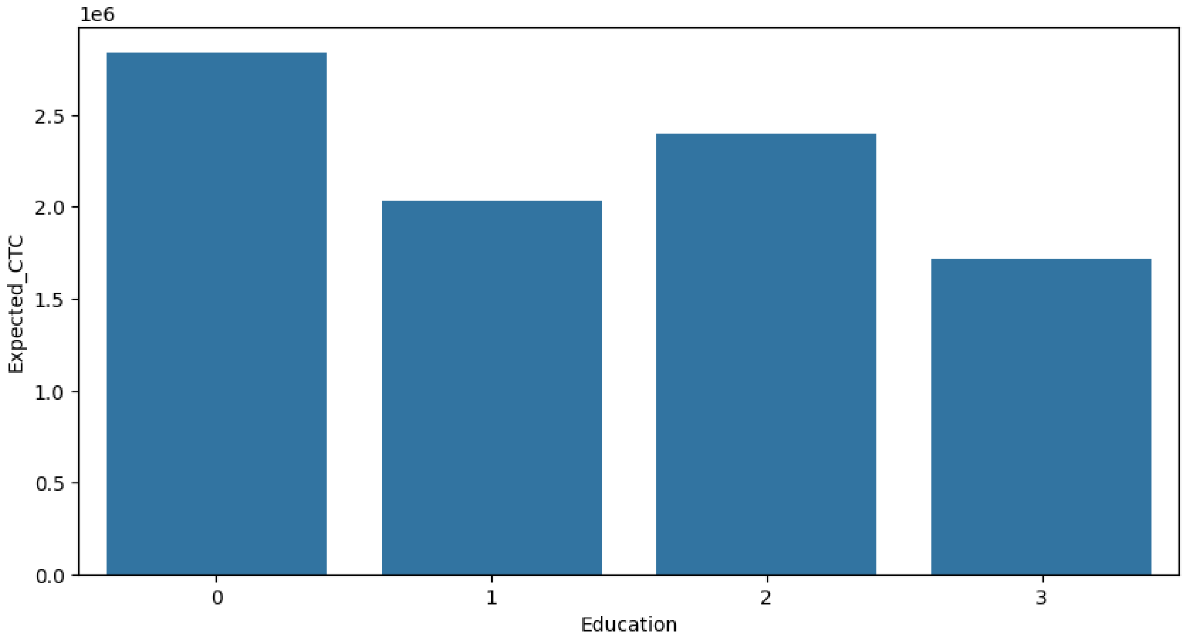
Candidates from FMCG and Others have high CTC expectations.

Designation:



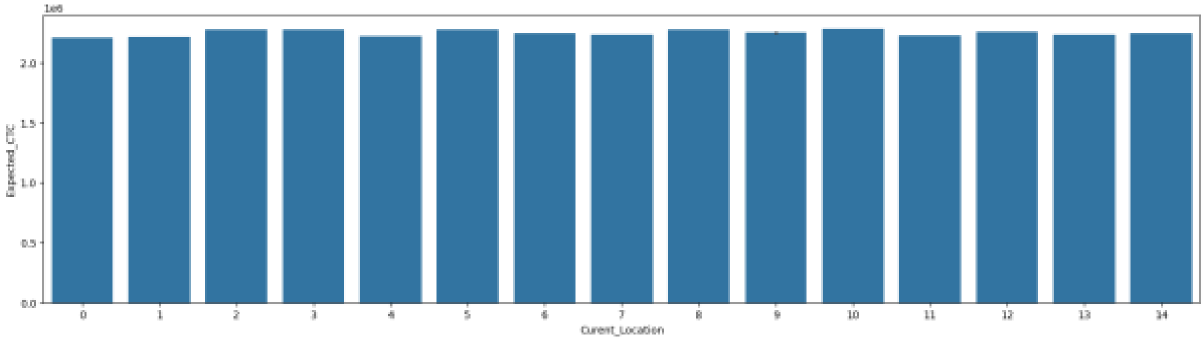
Research Scientists have high CTC expectations.

Education:



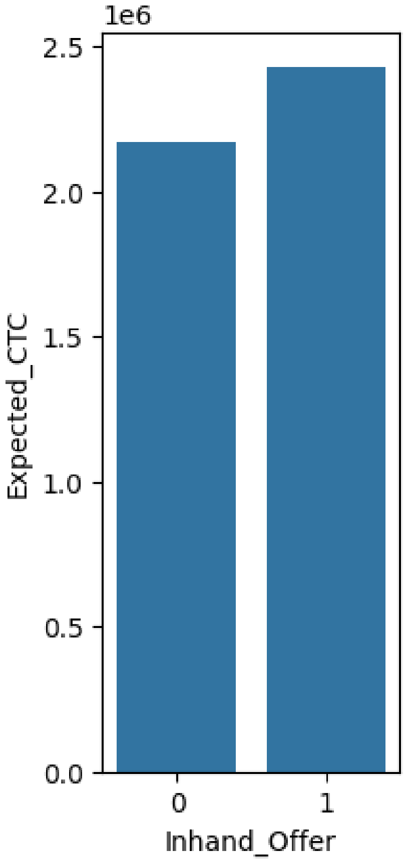
Candidates with doctorate have high CTC expectations.

Current Location:



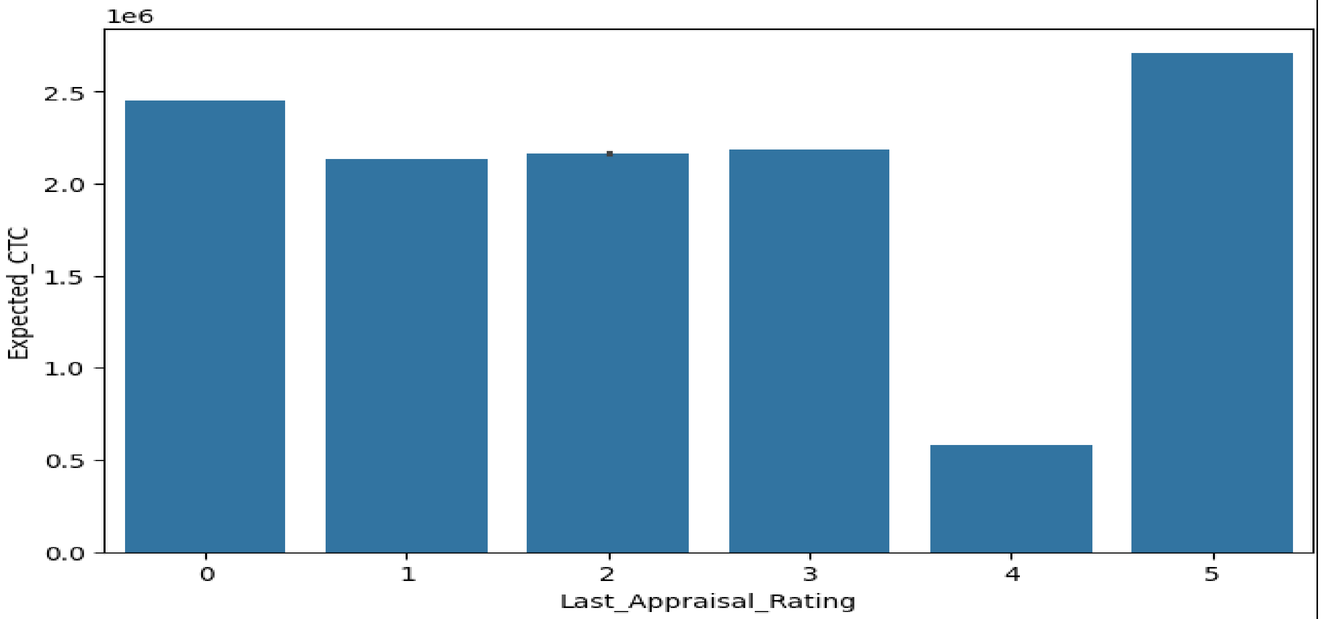
Candidates from Kolkata, Guwahati, Mangalore, Bhubaneshwar have high CTC expectations thanother locations.

In hand offer:



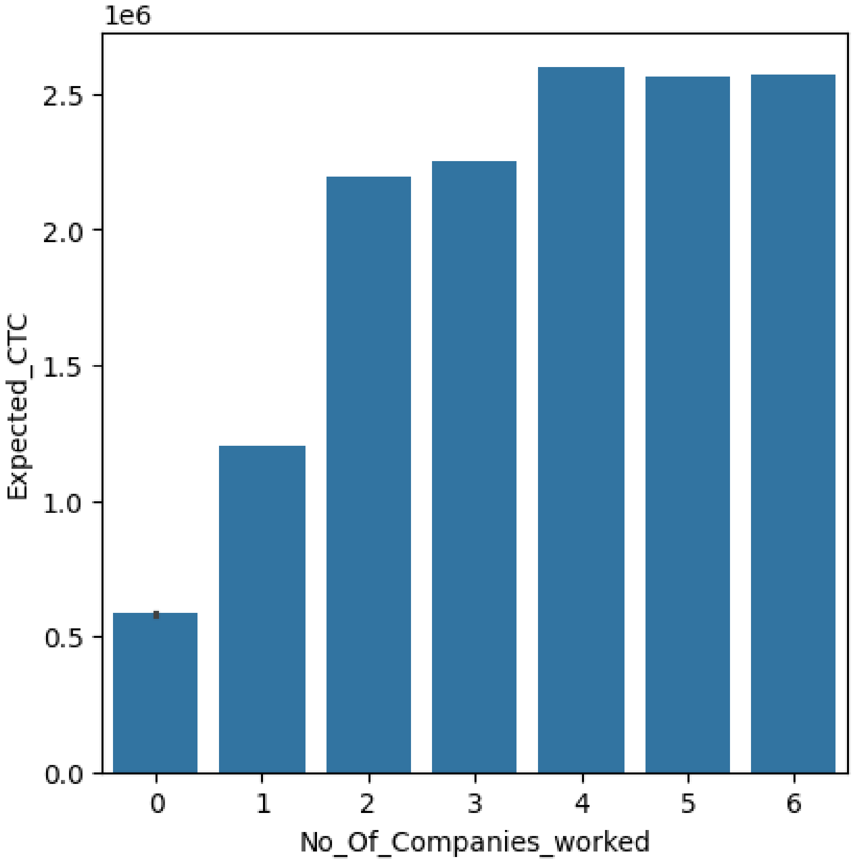
Candidates who holds offer have high CTC expectations.

Last appraisal rating:



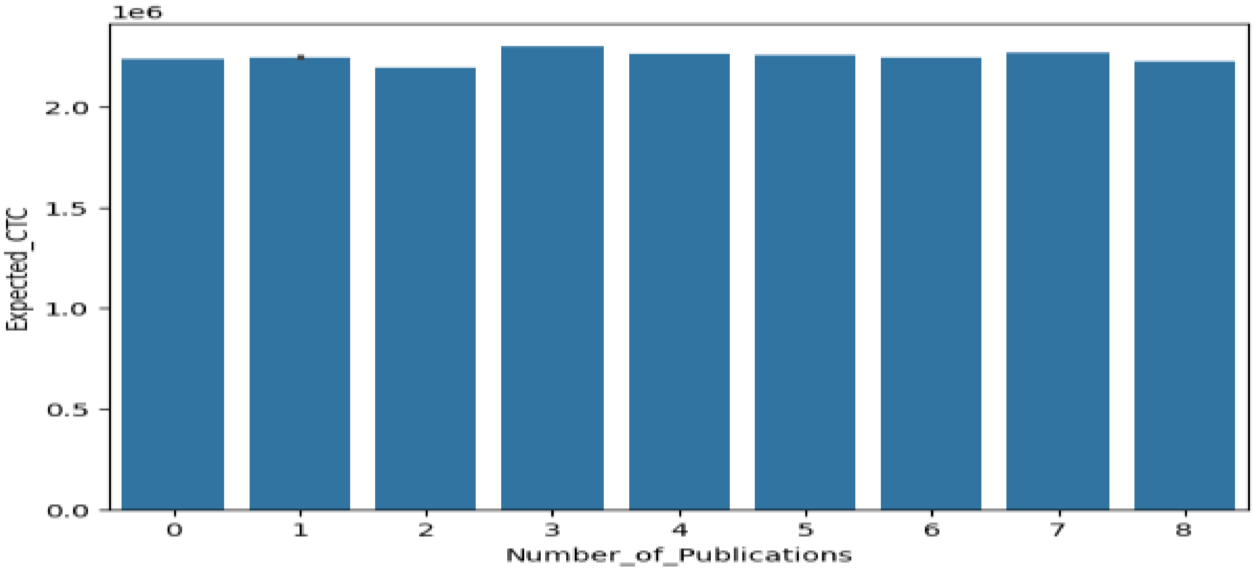
Candidates with key performer appraisal rating have high CTC expectations.

Company's worked



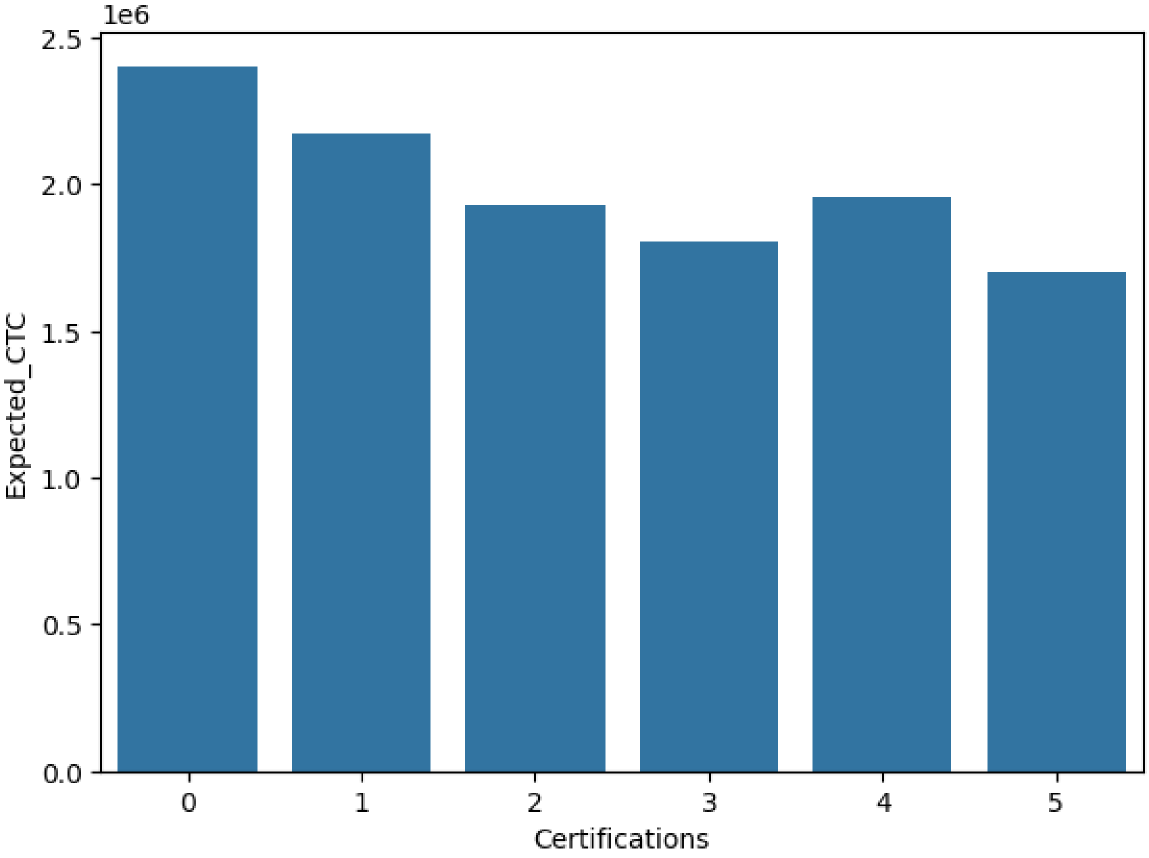
Candidates who worked in 4 company have high CTC expectations even 5&6 companies also have high expectation.

Number of published:



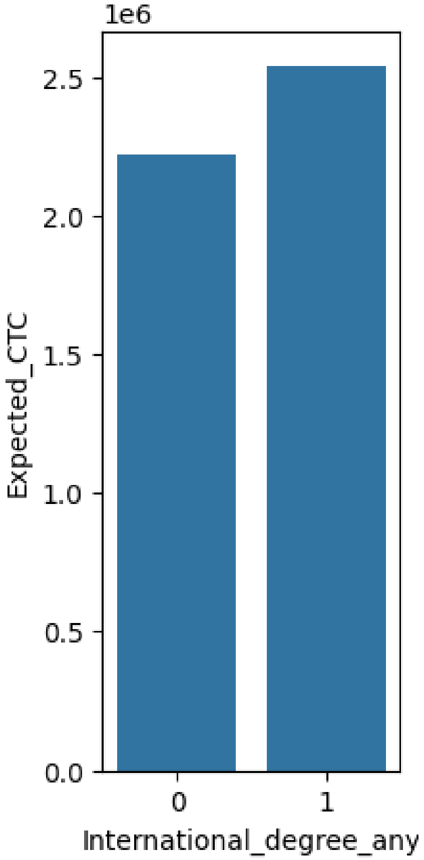
Candidates who have done 3 publications have high CTC expectations.

Certification:



Whereas candidates with 3certifications have less expectations. Candidates with 0 certifications have high CTC expectations.

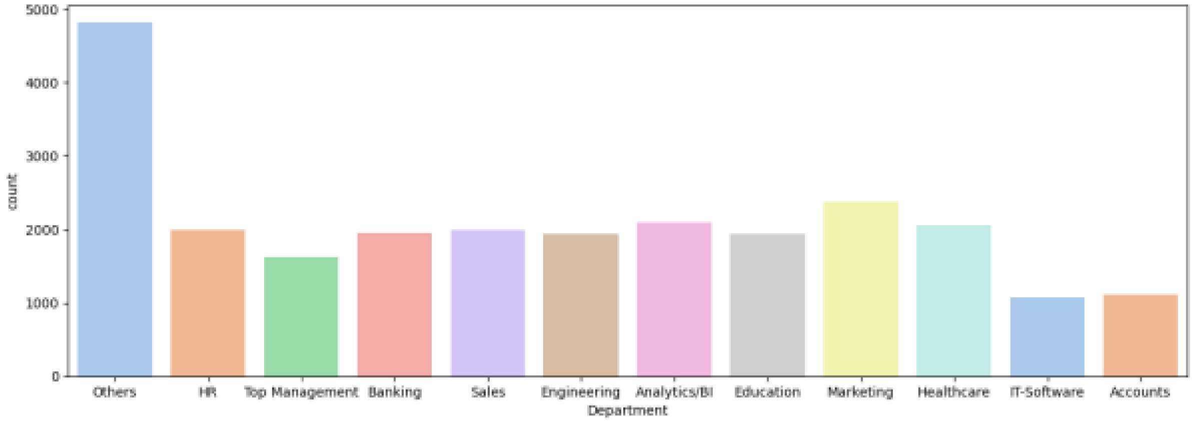
International Degree:



Candidates with international degree have high CTC expectations.

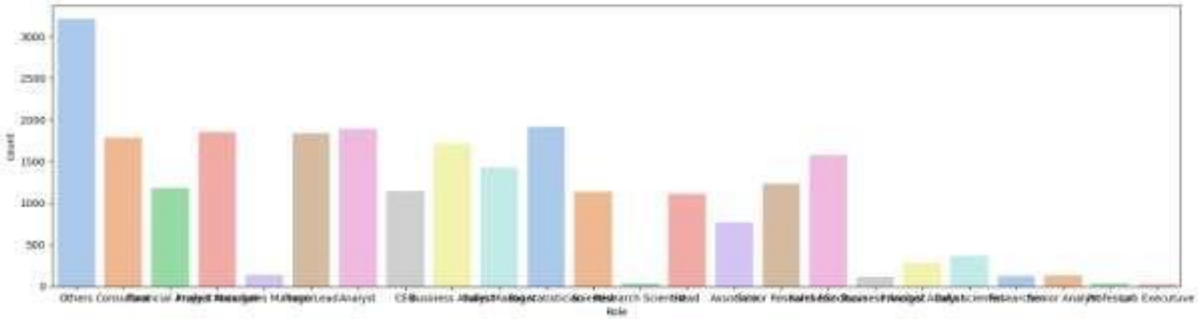
b) Univariate analysis (distribution and spread for every continuous attribute, distribution of data incategories for categorical ones)

Let’s check Univariate for Department:



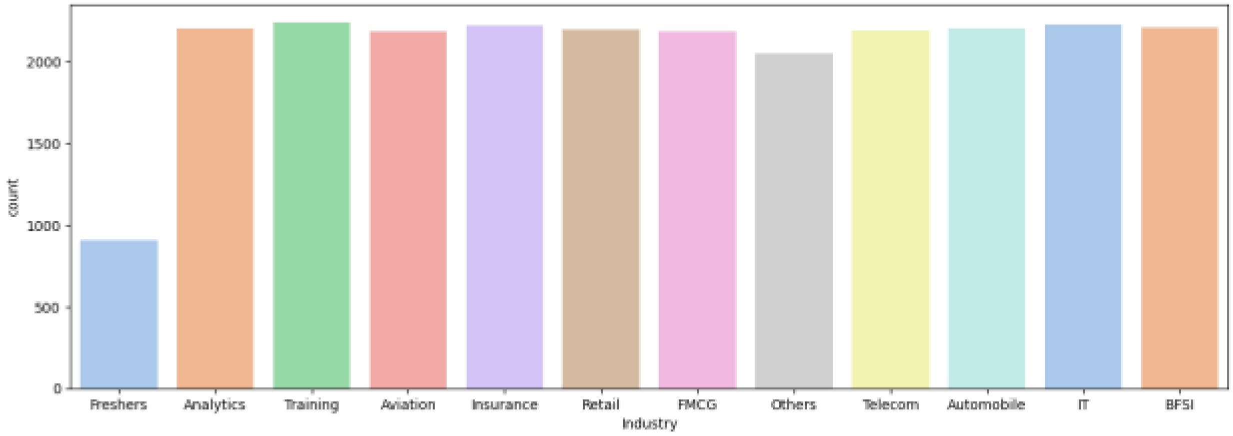
In our data set “others” category has huge number of data which is somewhere around 4800, nextto that “marketing” category have count of 2200 approximately.

Role:



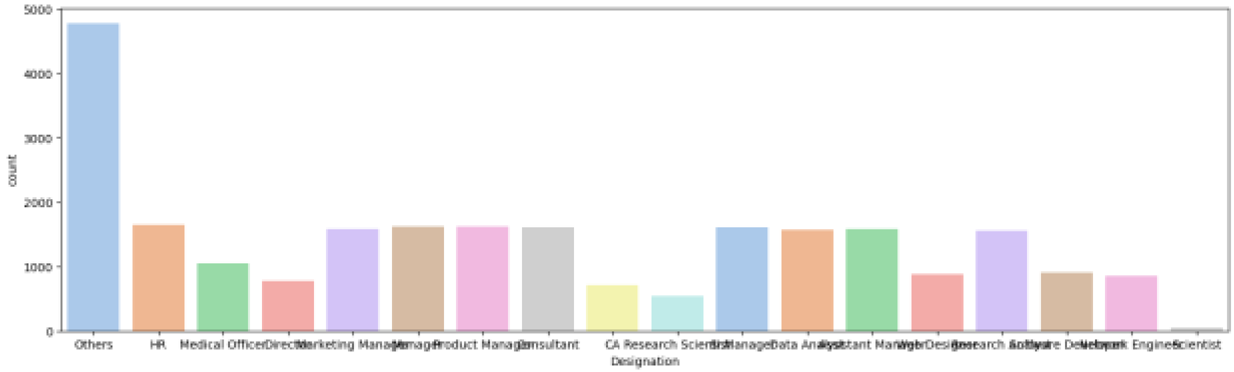
In our data set “others” category has huge number of data which is somewhere around 3500,next to that “Analyst” category have count of 1800 approximately. Research scientist, Lab executive andProfessors have very less count which is less than 50.

Industry :



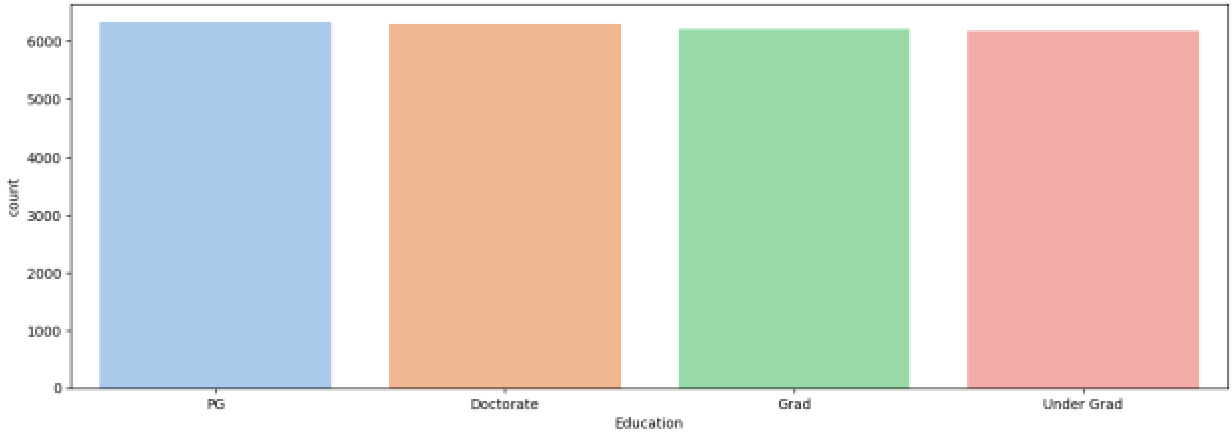
Candidates from training, insurance, IT and BFSI have high count of around 2500 approximately fromeach industry. Fresher candidates have count of 1800 approximately.

Designation:



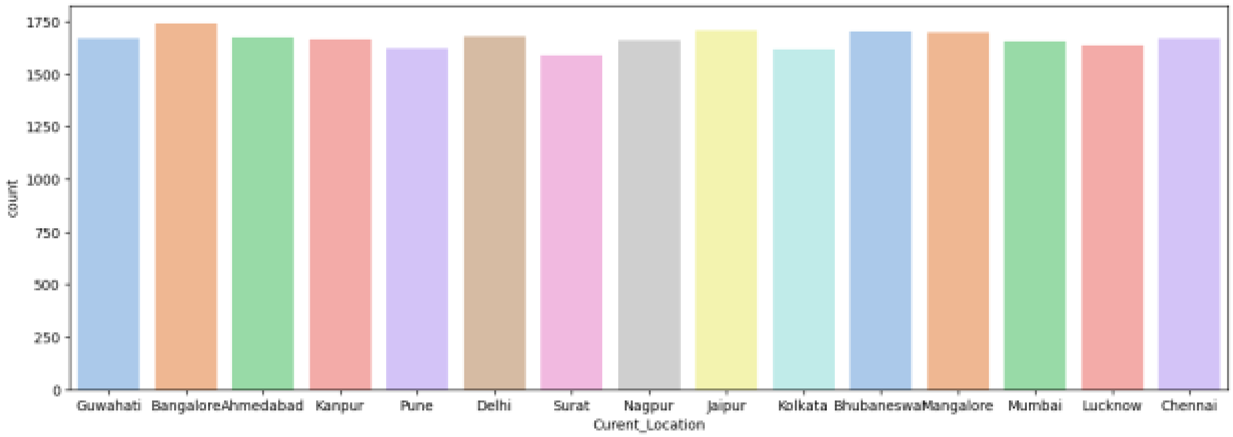
In “others” category around 4800 count for designation, next to that marketing manager, manager,product manager and consultant were around 1800.

Education:



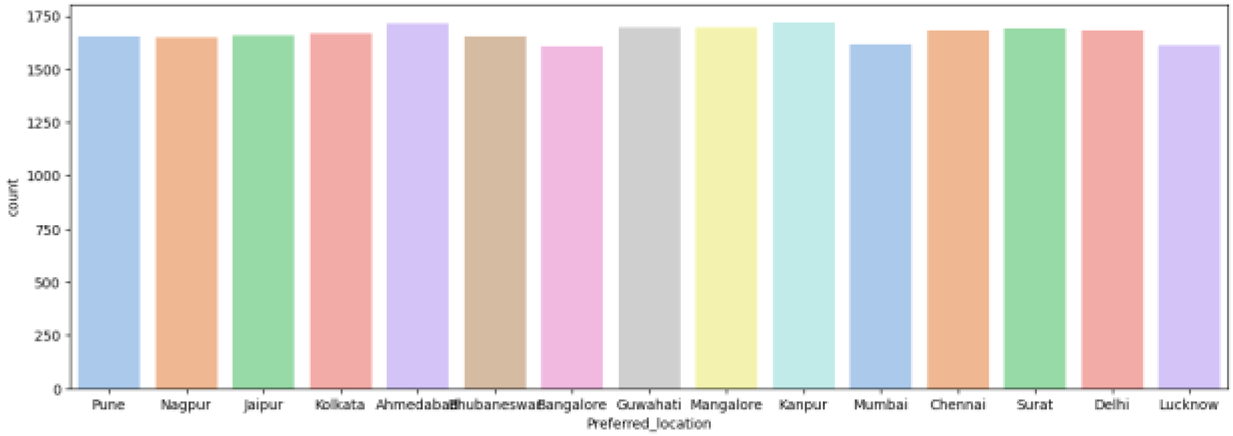
In our data set Post graduate and doctorate candidates have high count of more than 6500,difference between other grads if very minimal.

Current Location:



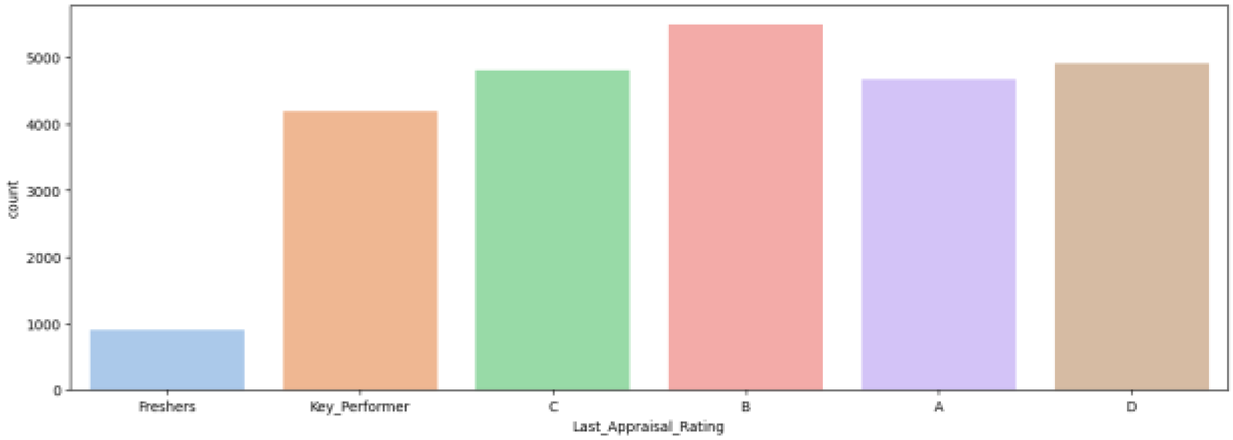
Candidates from Bangalore, Jaipur, Bhubaneshwar and Mangalore are high, from Pune its little lowerthan other locations.

Preferred Location:



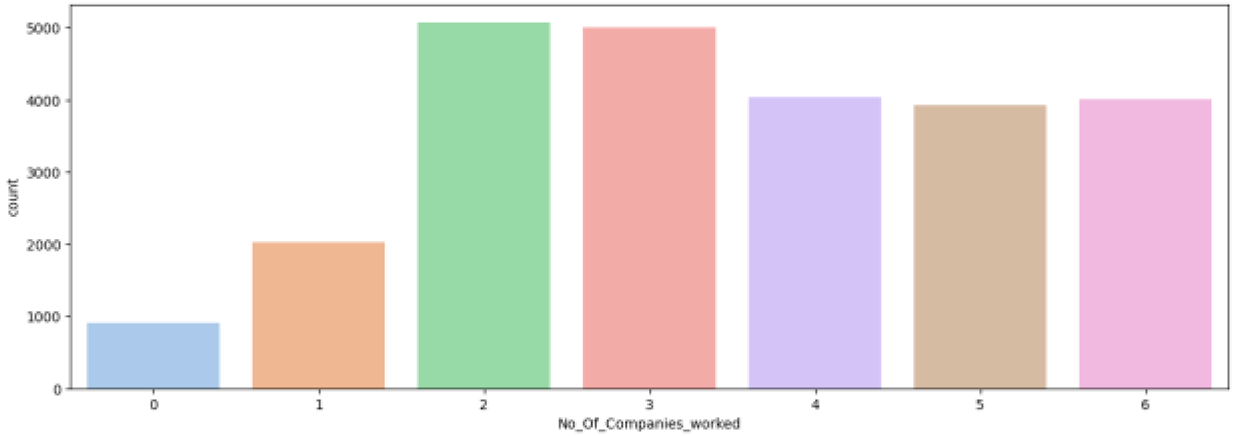
Candidates highly prefers Ahmedabad, Guwahati, Mangalore as their job preferred location.

Last appraisal rating:



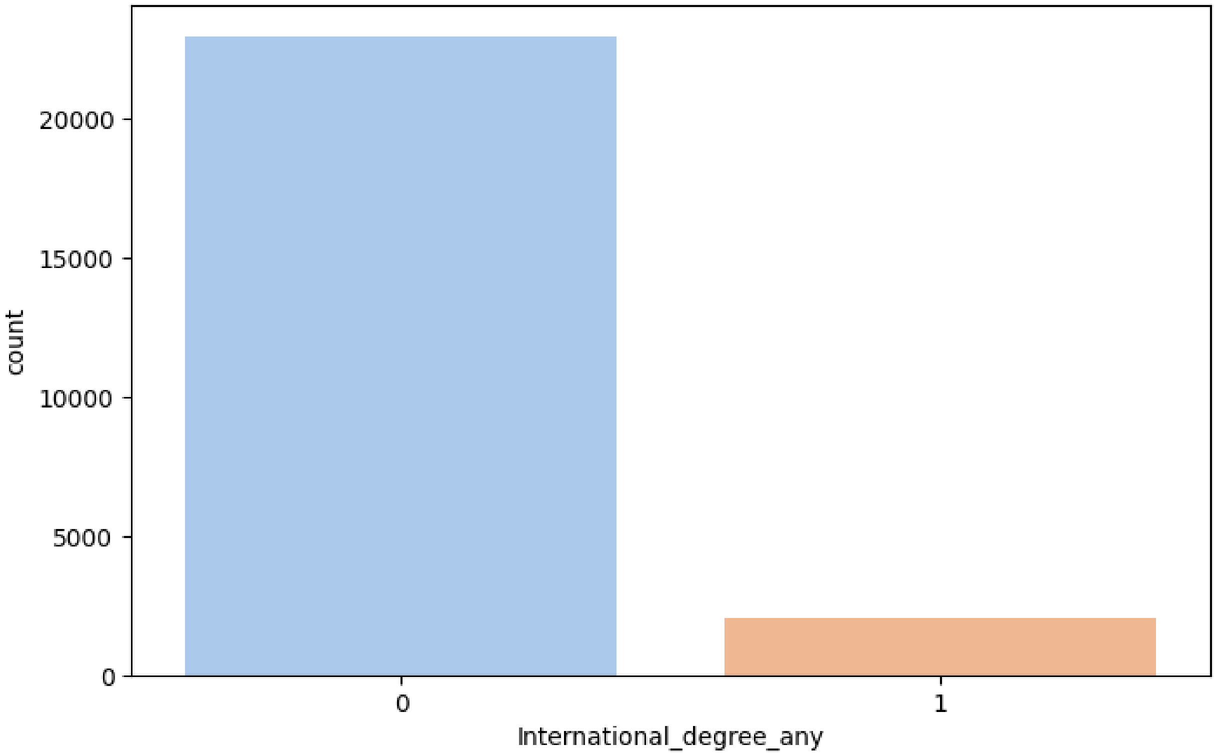
In out data set people who got ‘B’ appraisal are high, Key performers were around 4200.

Number of companies worked:



Two and three companies working persons are more. As comparer to number Five company number four and six companies working experience persons are more .

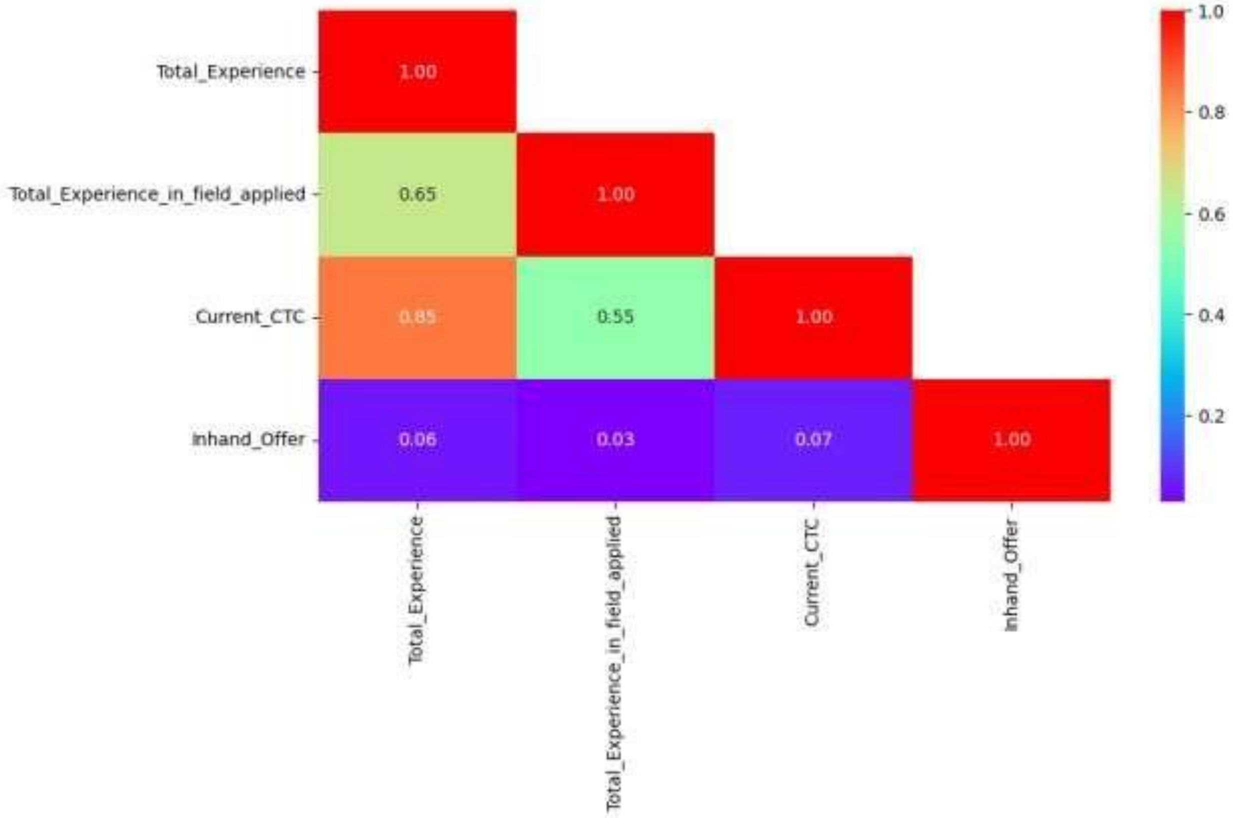
International degree:



Candidates with international degree are very in our data set, more than 90% of population arewithout international degree.

c)Bivariate analysis (relationship between different variables, correlations)

Correlation Map:



From the above heat map, we can observe that Total experience of field applied and Total experience, Current CTC and Total experience of field applied have high correlation. In hand offerand Number of publications have medium correlation.

4) Business insights

fromEDA

Feature Selection:

Before selecting features need to convert data types as integer.

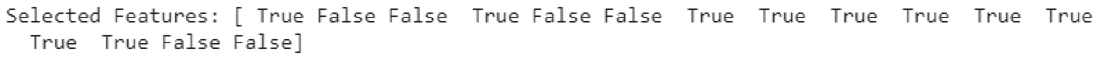
```
<class
'pandas.core.frame.DataFram
e'>
RangeIndex  25000        0  to
:          entries,
24999       columns        17
Data                 (tota
            l
columns):
 #   Column                            Non-Null  Dtyp
                                          Count  e
------                            ---------------------------------------
 0   Total_Experience                 25000 non-  int6
                                          null  4
 1   Total_Experience_in_field_app    25000 non-  int6
     lied                                 null  4
 2   Department                       25000 non-  int8
                                          null
 3   Role                           25000 non-     int8
                                          null
 4   Industry                       25000 non-     int8
                                          null
 5   Designation                    25000 non-     int8
                                          null
 6   Education                      25000 non-     int8
                                          null
 7   Curent_Location                25000 non-     int8
                                          null
 8   Preferred_location             25000 non-     int8
                                          null
 9   Current_CTC                    25000 non-     int6
                                          null     4
 10   Inhand_Offer                  25000 non-     int8
                                          null
 11  Last_Appraisal_Rating          25000 non-     int8
                                          null
 12  No_Of_Companies_worked         25000 non-     int8
                                          null
 13  Number_of_Publications         25000 non-     int6
                                          null     4
 14  Certifications                 25000 non-     int6
                                          null     4
 15  International_degree_any       25000 non-     int8
                                          null
 16  Expected_CTC                   25000 non-
                                          null
int64dtypes: int64(6),
int8(11)memory usage: 1.4
MB
```

From the above table we can observe that every variable has integer data type. Its good to performRFE analysis for feature selection.

RFE Analysis :

- Before performing RFE analysis need to segregate target variable separately from dataframe.
- Target variable should be assigned separately.
- Rest of the variables should be fit in to RFE analysis.
- For estimator considered Random Forest Regressor since our target variable is continuous.

RFE Analysis:



```
Selected Features: [ True False False  True False False  True  True  True  True  True  True
  True  True False False]
```

Above results are in Boolean values, true which denotes to select that features. However Falserepresents that not to select those features.

List of features considering for further analysis:

Total_Experi

enceRole

Education

Curent_Location

Preferred_locatio

nCurrent_CTC

Inhand_Offer

Last_Appraisal_R

ating

No_Of_Companies_workedNumber_of_Publications

Inference:

- Above mentioned features are suggested by RFE analysis as best features.
- From this set of features going to perform further Analysis.

a) Business insights using clustering

Scaling Data:

- Using Standard scalar scaling data before performing clustering.
- Since our data set is larger using K- means clustering.

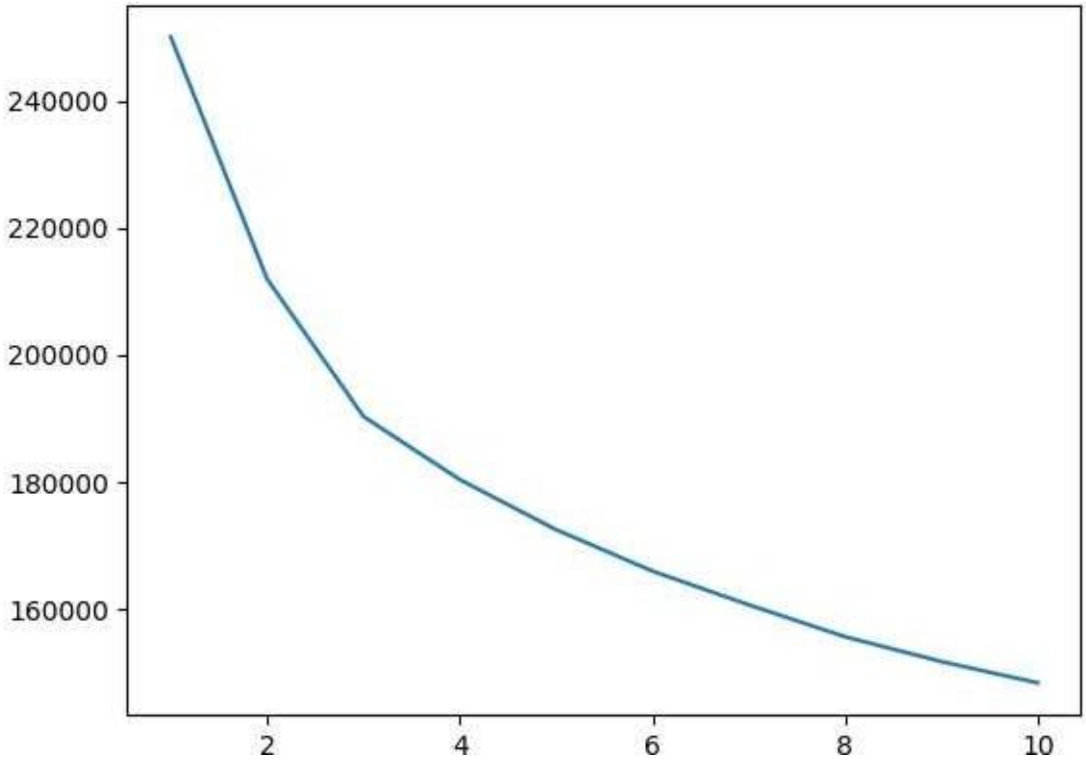
K Means Clustering:

1. 249999.9999999999,
2. 211893.92438503072,
3. 190242.21739650163,
4. 180302.15930825716,
5. 172413.81300687417,
6. 165894.29133165703,
7. 160576.8311867193,
8. 155551.54410477262,

9. 151642.85369396262,
10. 148290.84740284178

From the above table no of clusters and inertia values respectively.

Elbow plot:



From the above elbow plot we can observe that no stable clusters till cluster 10. However, wechoose number of clusters based on silhouette scores.

Number of clusters 3 and 2 has good silhouette scores which is 0.142670587829086.

Note-II

Parametric Model:

Non - Parametric Models

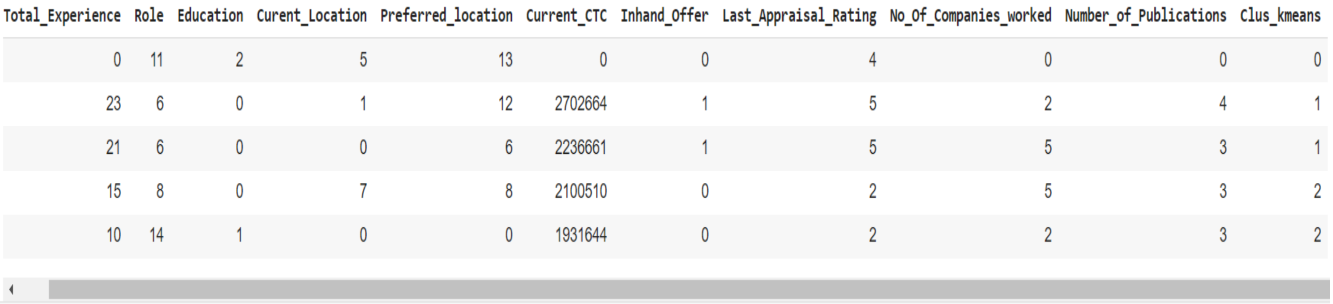
Non Parametric Models

Goal & 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

Model building:

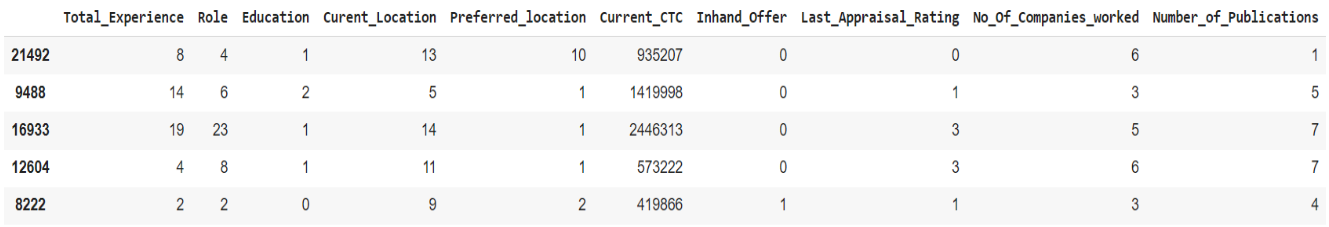
For building the model we have to split the data set in to 30% for test and 70% for train.

Before the splitting data scale the data.



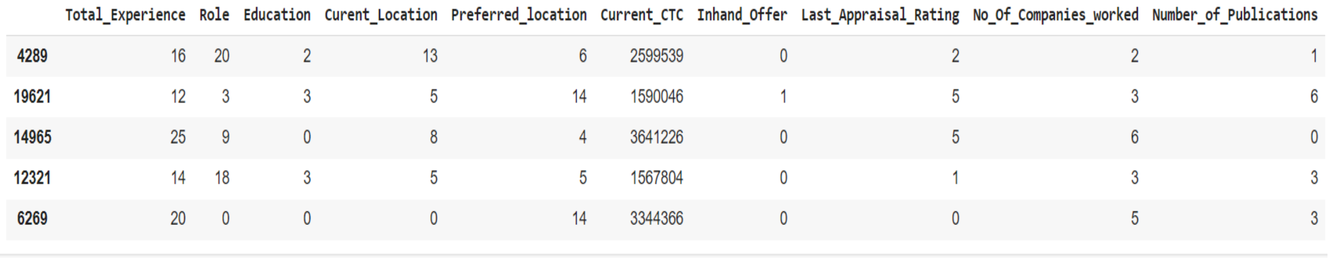
Total_Experience	Role	Education	Curent_Location	Preferred_location	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked	Number_of_Publications	Clus_kmeans
0	11	2	5	13	0	0	4	0	0	0
23	6	0	1	12	2702664	1	5	2	4	1
21	6	0	0	6	2236661	1	5	5	3	1
15	8	0	7	8	2100510	0	2	5	3	2
10	14	1	0	0	1931644	0	2	2	3	2

Result of the 30% data set



	Total_Experience	Role	Education	Curent_Location	Preferred_location	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked	Number_of_Publications
21492	8	4	1	13	10	935207	0	0	6	1
9488	14	6	2	5	1	1419998	0	1	3	5
16933	19	23	1	14	1	2446313	0	3	5	7
12604	4	8	1	11	1	573222	0	3	6	7
8222	2	2	0	9	2	419866	1	1	3	4

Result of the 70% dataset



	Total_Experience	Role	Education	Curent_Location	Preferred_location	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked	Number_of_Publications
4289	16	20	2	13	6	2599539	0	2	2	1
19621	12	3	3	5	14	1590046	1	5	3	6
14965	25	9	0	8	4	3641226	0	5	6	0
12321	14	18	3	5	5	1567804	0	1	3	3
6269	20	0	0	0	14	3344366	0	0	5	3

Use parametric Model:

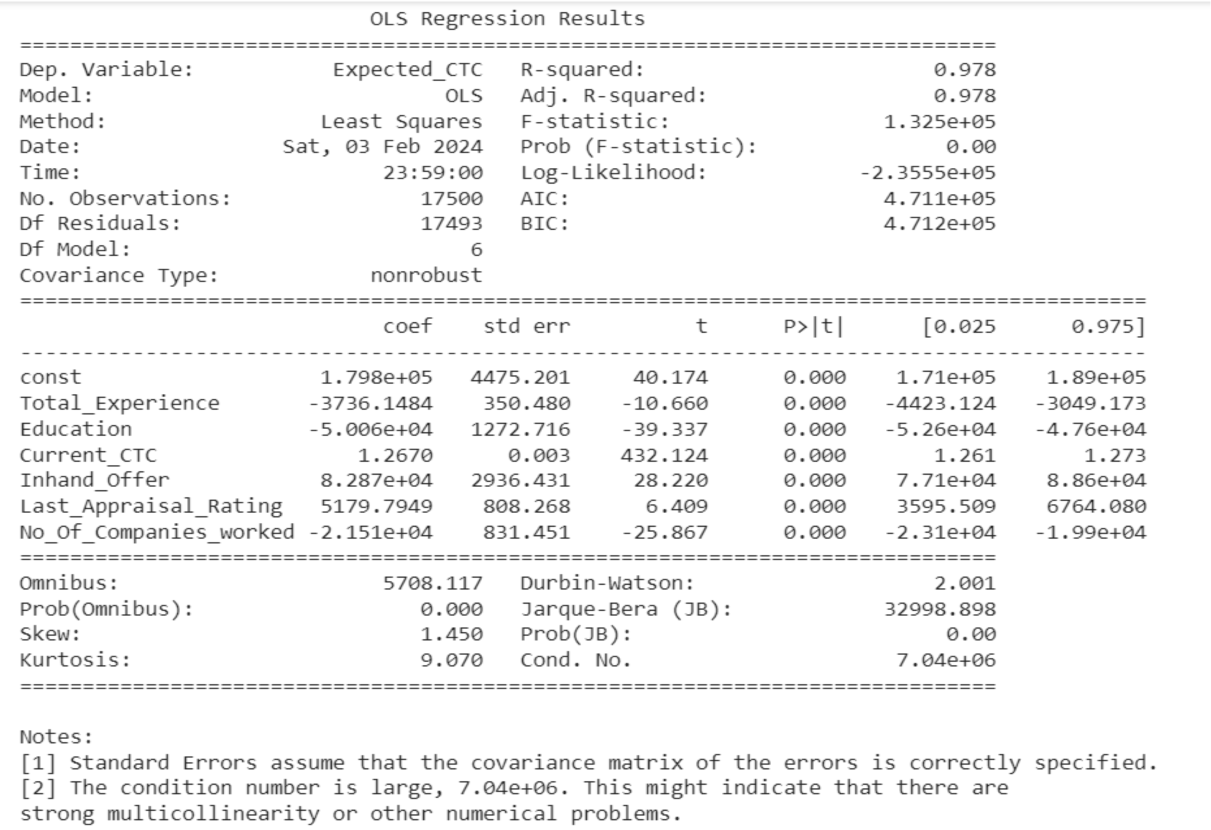
```
                            OLS Regression Results                            
=======================================================================
=======
Dep. Variable:            Expected_CTC   R-squared:                       
0.978
Model:                             OLS   Adj. R-squared:                  
0.978
Method:                  Least Squares   F-statistic:                     
7.951e+04
Date:                 Sat, 03 Feb 2024   Prob (F-statistic):              
0.00
Time:                         23:59:00   Log-Likelihood:                -
2.3555e+05
No. Observations:                17500   AIC:                             
4.711e+05
Df Residuals:                    17489   BIC:                             
4.712e+05
Df Model:                           10                                         
Covariance Type:             nonrobust                                         
=======================================================================
===================
                             coef    std err          t      P>|t|      
[0.025      0.975]
-----------------------------------------------------------------------
-------------------
const                   1.791e+05   6040.037     29.660      0.000    
1.67e+05    1.91e+05
Total_Experience       -3753.5562    350.681    -10.704      0.000   -
4440.925   -3066.187
Role                     173.1043    181.889      0.952      0.341    -
183.416     529.624
Education              -5.002e+04   1272.875    -39.295      0.000   -
5.25e+04   -4.75e+04
Curent_Location          384.1913    298.323      1.288      0.198    -
200.551     968.933
Preferred_location      -543.9333    296.746     -1.833      0.067   -
1125.585      37.719
Current_CTC                1.2670      0.003    432.001      0.000     
1.261       1.273
Inhand_Offer            8.298e+04   3053.886     27.172      0.000     
7.7e+04     8.9e+04
```

```
Last_Appraisal_Rating   5174.0534    810.915      6.381      0.000    
3584.579    6763.528
No_Of_Companies_worked -2.152e+04    831.624    -25.877      0.000   -
2.31e+04   -1.99e+04
Number_of_Publications     3.0250    509.956      0.006      0.995    -
996.539    1002.589
=======================================================================
=======
Omnibus:                     5704.051   Durbin-Watson:                 
2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          
32936.367
Skew:                           1.449   Prob(JB):                      
0.00
Kurtosis:                       9.064   Cond. No.                      
9.41e+06
=======================================================================
=======

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 9.41e+06. This might indicate that
there are
strong multicollinearity or other numerical problems.

CTC   R-squared:                       0.978
```

Parametric model shows the R-squared and adj.R-squared value 97% .which is the good for dataset.

Observation from the predictor:
'Number_of_Publications','Curent_Location','Preferred_location','Role'has p-value>0.05 we remove those columns and build the model.

After dropping the columns parametric model:

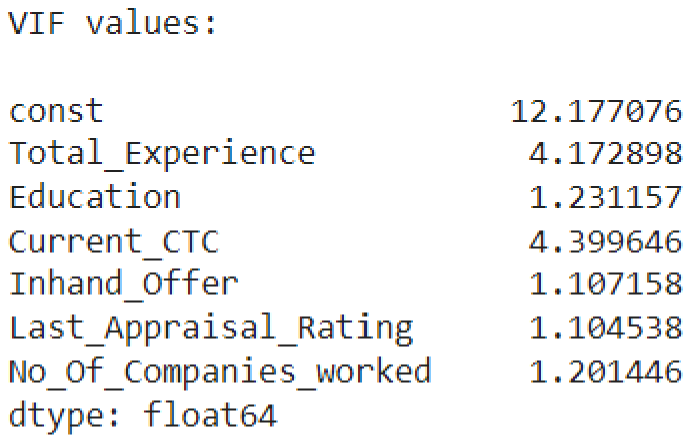


```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Expected_CTC   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.978
Method:                 Least Squares   F-statistic:                 1.325e+05
Date:                Sat, 03 Feb 2024   Prob (F-statistic):               0.00
Time:                        23:59:00   Log-Likelihood:            -2.3555e+05
No. Observations:               17500   AIC:                         4.711e+05
Df Residuals:                   17493   BIC:                         4.712e+05
Df Model:                           6
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 1.798e+05   4475.201     40.174      0.000    1.71e+05    1.89e+05
Total_Experience     -3736.1484    350.480    -10.660      0.000   -4423.124   -3049.173
Education            -5.006e+04   1272.716    -39.337      0.000   -5.26e+04   -4.76e+04
Current_CTC              1.2670      0.003    432.124      0.000       1.261       1.273
Inhand_Offer          8.287e+04   2936.431     28.220      0.000    7.71e+04    8.86e+04
Last_Appraisal_Rating 5179.7949    808.268      6.409      0.000    3595.509    6764.080
No_Of_Companies_worked -2.151e+04   831.451    -25.867      0.000   -2.31e+04   -1.99e+04
==============================================================================
Omnibus:                     5708.117   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            32998.898
Skew:                           1.450   Prob(JB):                         0.00
Kurtosis:                       9.070   Cond. No.                     7.04e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.04e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

After removing high VIF values feature “Total experience” R-Squared value remains same. However, if we remove “Current CTC” there is drastic dip in R-Squared value.

- Now we will check VIF predictor:

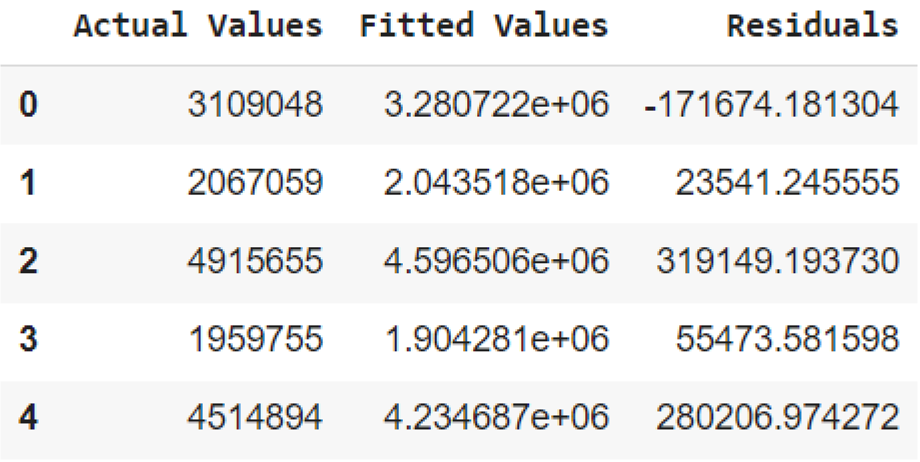
VIF values are:



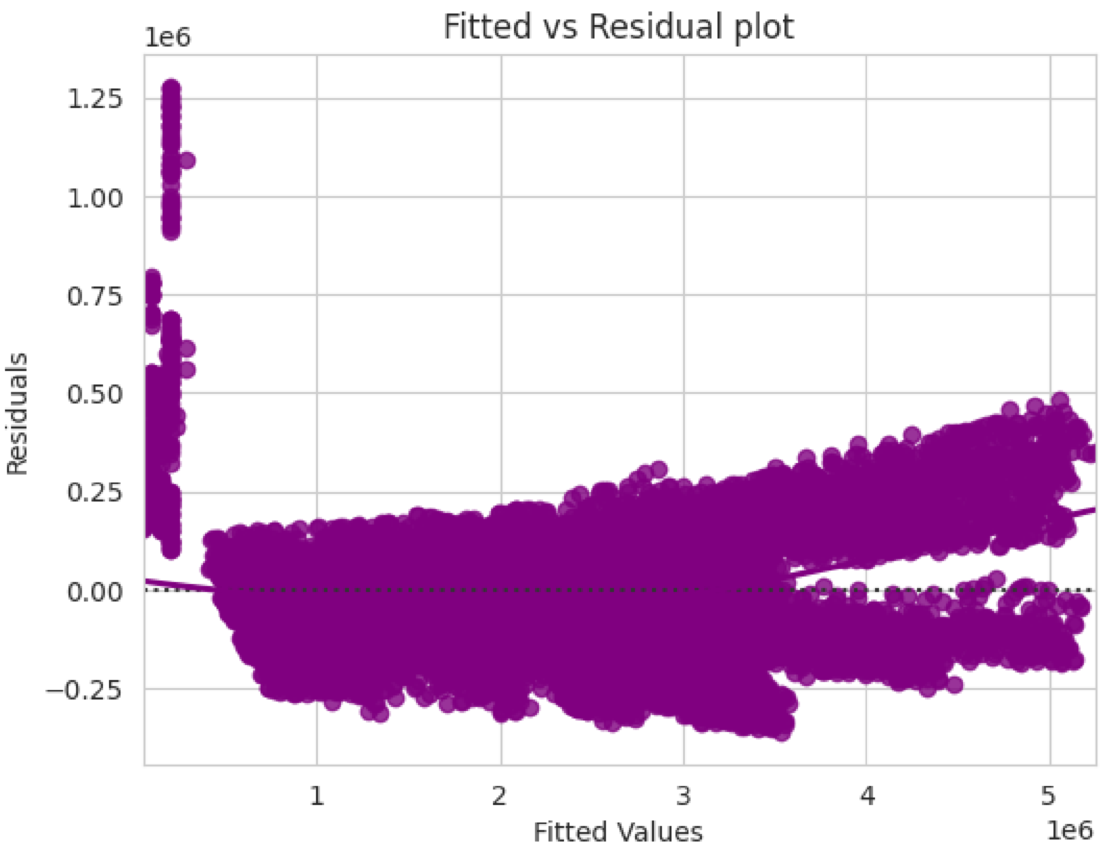
```
VIF values:

const                     12.177076
Total_Experience           4.172898
Education                  1.231157
Current_CTC                4.399646
Inhand_Offer               1.107158
Last_Appraisal_Rating      1.104538
No_Of_Companies_worked     1.201446
dtype: float64
```

From OLS stats model we can fair R-Squared and Adjusted R-Squared value. However, a few variables have VIF values > 2 therefore some multicolinearity in the data. Hence those features are important for the analysis we cannot drop those variables.

- Linearity and Independence predictor:



	Actual Values	Fitted Values	Residuals
0	3109048	3.280722e+06	-171674.181304
1	2067059	2.043518e+06	23541.245555
2	4915655	4.596506e+06	319149.193730
3	1959755	1.904281e+06	55473.581598
4	4514894	4.234687e+06	280206.974272



- **Test normality**

Since p-value < 0.05, the residuals are not normal as per shapiro test.

ShapiroResult(statistic=0.9244317412376404, pvalue=0.0)

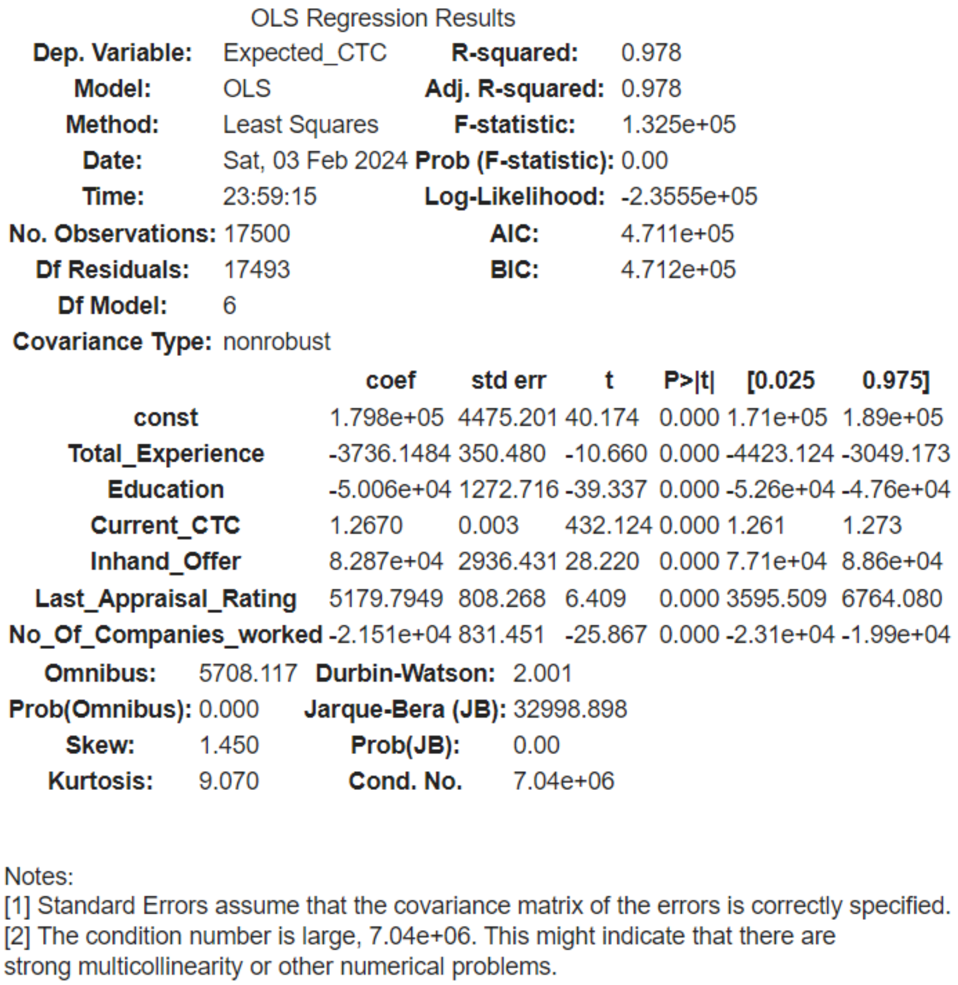
- Test Homoscedasticity:

0.9790227865600101

Since p-value > 0.05 we can say that the residuals are homoscedastic.

The model built Linear_OLS_model2 satisfies all assumptions of Linear Regression

- Build linear OLS model:



```
                          OLS Regression Results
     Dep. Variable:  Expected_CTC          R-squared:  0.978
            Model:   OLS               Adj. R-squared:  0.978
           Method:   Least Squares         F-statistic:  1.325e+05
             Date:   Sat, 03 Feb 2024 Prob (F-statistic): 0.00
             Time:   23:59:15          Log-Likelihood:  -2.3555e+05
No. Observations: 17500                        AIC:  4.711e+05
     Df Residuals:   17493                        BIC:  4.712e+05
         Df Model:   6
Covariance Type: nonrobust
                                  coef     std err     t     P>|t|   [0.025    0.975]
             const             1.798e+05  4475.201 40.174   0.000 1.71e+05  1.89e+05
      Total_Experience         -3736.1484 350.480  -10.660  0.000 -4423.124 -3049.173
          Education            -5.006e+04 1272.716 -39.337  0.000 -5.26e+04 -4.76e+04
         Current_CTC           1.2670     0.003    432.124  0.000 1.261     1.273
         Inhand_Offer          8.287e+04  2936.431 28.220   0.000 7.71e+04  8.86e+04
    Last_Appraisal_Rating      5179.7949  808.268  6.409    0.000 3595.509  6764.080
No_Of_Companies_worked         -2.151e+04 831.451  -25.867  0.000 -2.31e+04 -1.99e+04
      Omnibus:      5708.117  Durbin-Watson:    2.001
Prob(Omnibus): 0.000      Jarque-Bera (JB): 32998.898
        Skew:       1.450         Prob(JB):      0.00
     Kurtosis:      9.070         Cond. No.      7.04e+06


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.04e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Equation of linear regression:

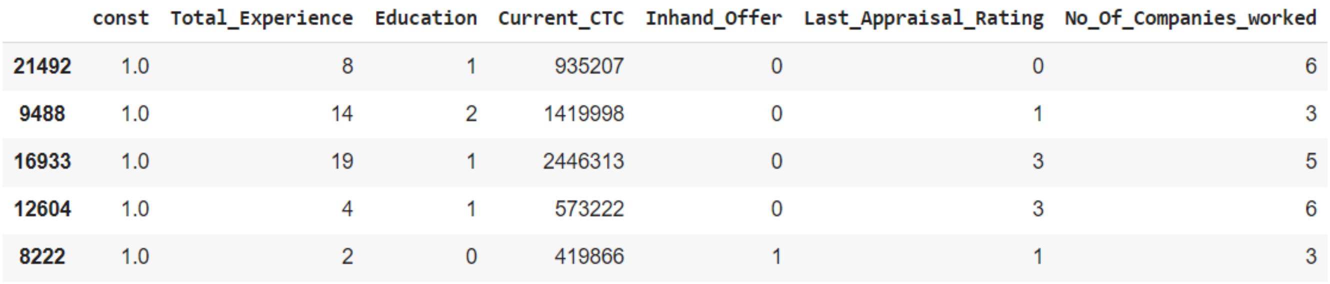
```
log_price = 179788.9299403612 + -3736.1484458393634 * ( Total_Experience ) +  -
50064.87500993241 * ( Education ) +  1.266954141958229 * ( Current_CTC ) +
82865.26557903447 * ( Inhand_Offer ) +  5179.794886646088 * (
Last_Appraisal_Rating ) +  -21507.45824393902 * ( No_Of_Companies_worked )
```

let's make predictions on the test set

After predicting the RMSE On the train data:

169618.48115250297

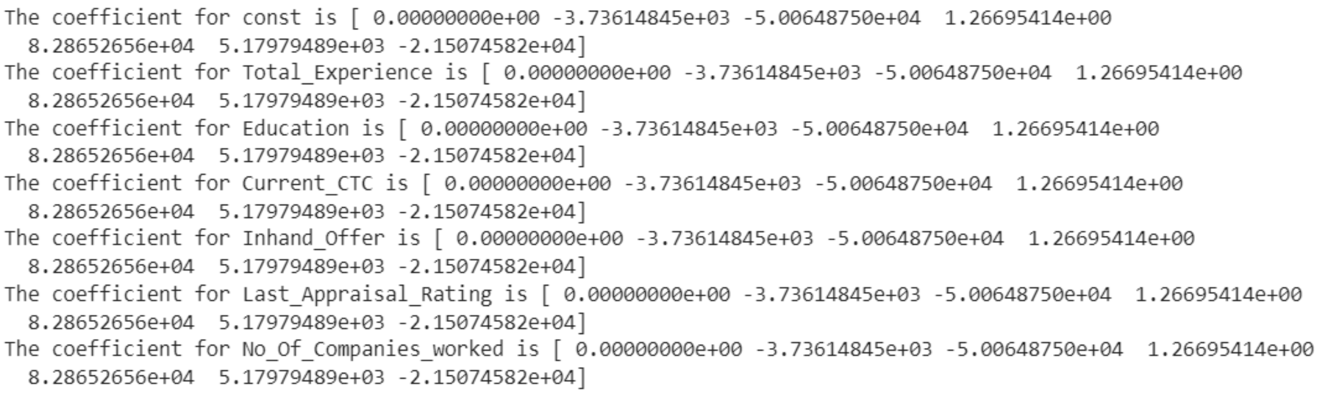
it means that, on average, the predictions of your regression model have an error of approximately 170168.53 units in the same scale as your target variable. A lower RMSE indicates better model performance, as it reflects smaller prediction errors.



	const	Total_Experience	Education	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked
21492	1.0	8	1	935207	0	0	6
9488	1.0	14	2	1419998	0	1	3
16933	1.0	19	1	2446313	0	3	5
12604	1.0	4	1	573222	0	3	6
8222	1.0	2	0	419866	1	1	3

Above table is shows the x_test values

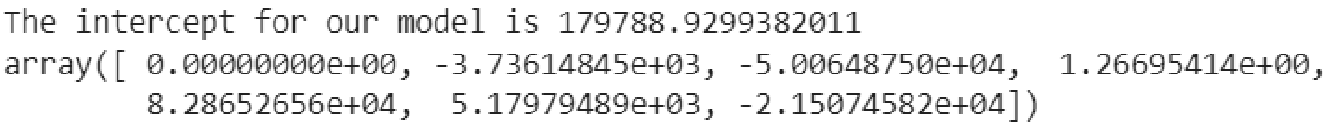
Now showing some coefficient values are:



```
The coefficient for const is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for Total_Experience is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for Education is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for Current_CTC is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for Inhand_Offer is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for Last_Appraisal_Rating is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
The coefficient for No_Of_Companies_worked is [ 0.00000000e+00 -3.73614845e+03 -5.00648750e+04  1.26695414e+00
  8.28652656e+04  5.17979489e+03 -2.15074582e+04]
```

From the above table we can observe that Current CTC, Inhand offer, Last appraisal rating seems to have good coefficient values towards target variable.

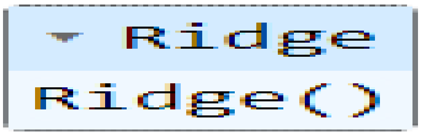
The model is performing intercept



```
The intercept for our model is 179788.9299382011
array([ 0.00000000e+00, -3.73614845e+03, -5.00648750e+04,  1.26695414e+00,
        8.28652656e+04,  5.17979489e+03, -2.15074582e+04])
```

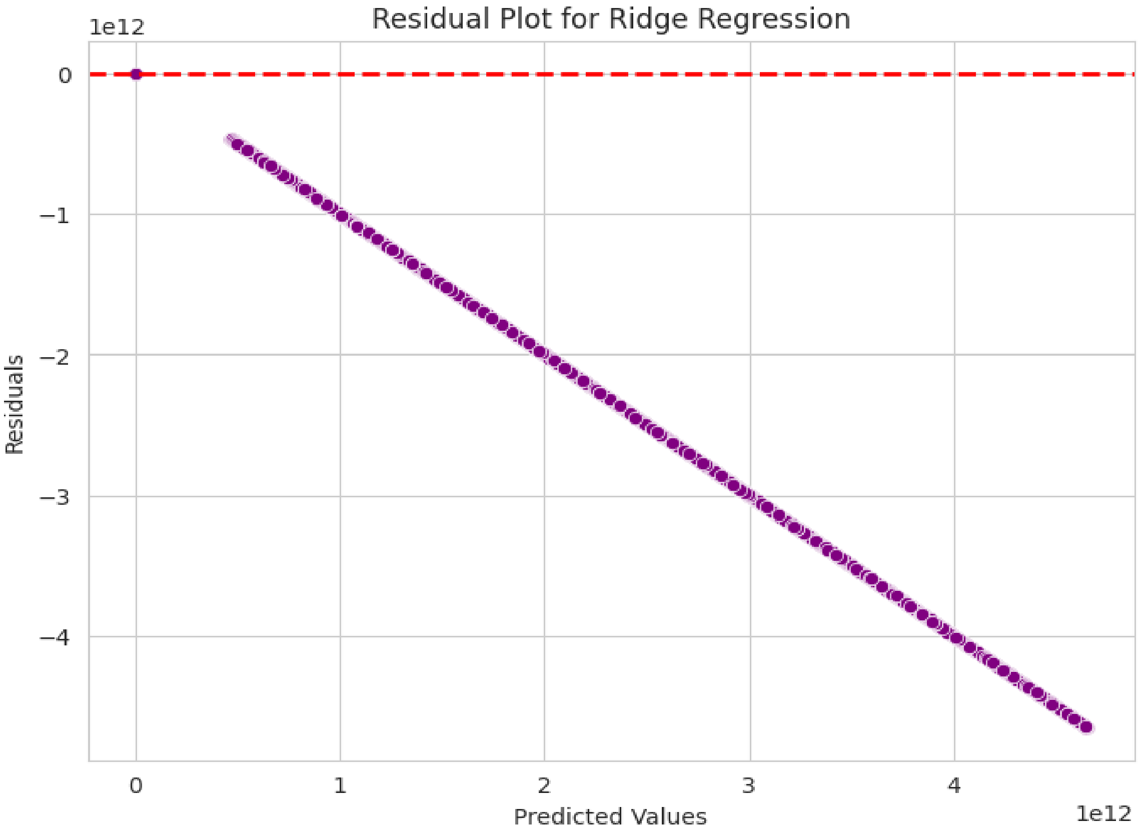
Building Ridge model:

modeling technique that can significantly improve the performance of linear regression **model.** Fit the Ridge model to the training data:



```
▾ Ridge
Ridge()
```

Making prediction on test data.



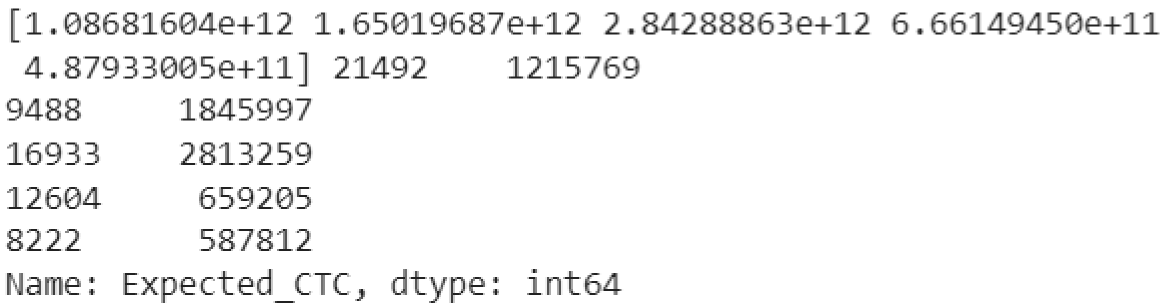
check the RMSE on the train data:

2305304856932.5674

Check the RMSE on the test data:

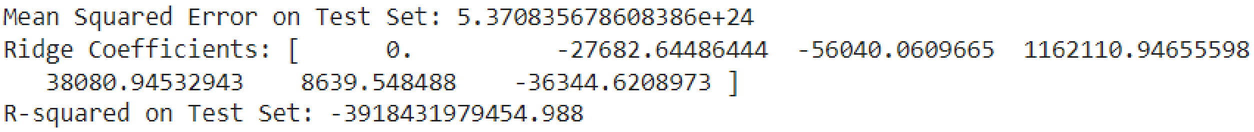
2317506349205.6255

Checking the five columns prediction ridge on test data:



```
[1.08681604e+12 1.65019687e+12 2.84288863e+12 6.66149450e+11
 4.87933005e+11] 21492    1215769
9488     1845997
16933    2813259
12604     659205
8222      587812
Name: Expected_CTC, dtype: int64
```

From above mentioned Applicant id number , expected CTC According to there application id we need to increase Or decrease the salary.

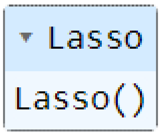


```
Mean Squared Error on Test Set: 5.370835678608386e+24
Ridge Coefficients: [      0.          -27682.64486444  -56040.0609665   1162110.94655598
   38080.94532943     8639.548488     -36344.6208973 ]
R-squared on Test Set: -3918431979454.988
```

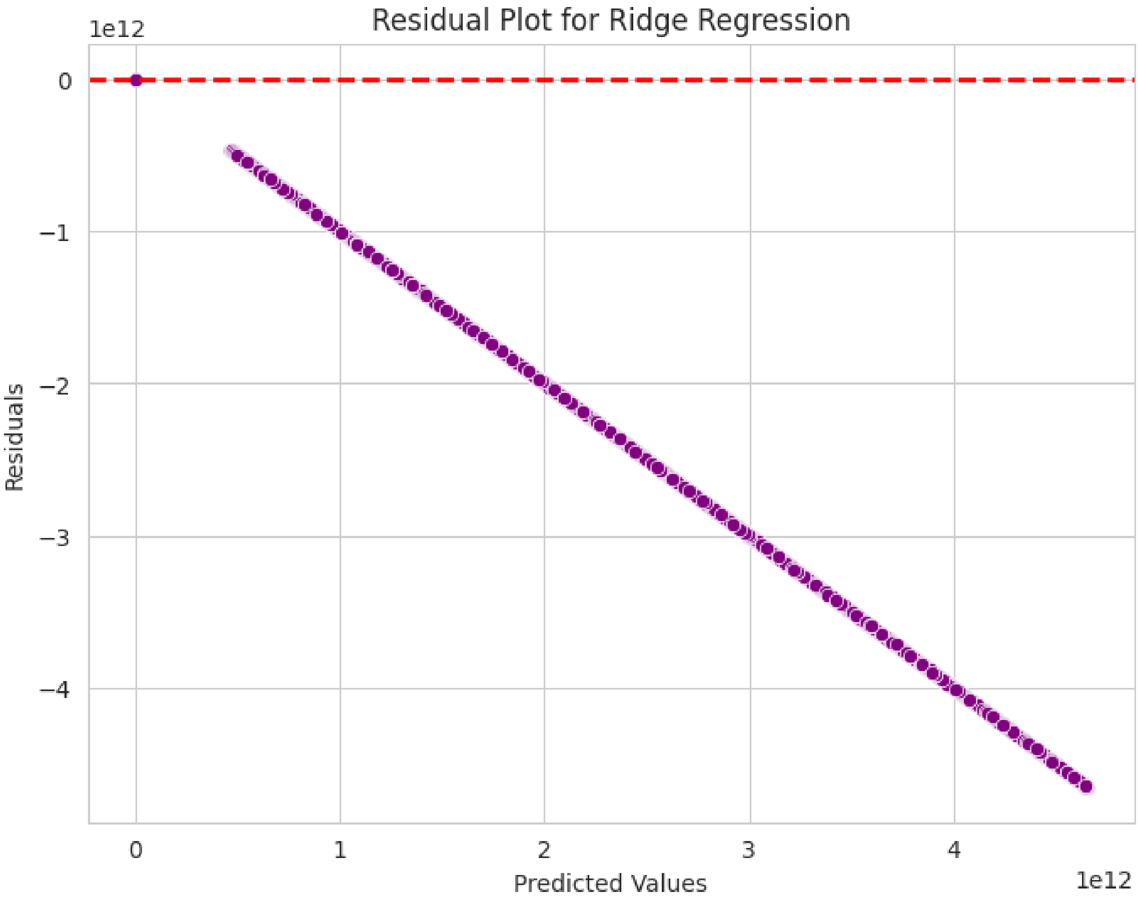
From the above table we can observe that Current CTC, Inhand offer, Last appraisal rating seems to have good coefficient values towards target variable.

Building LASSO model:

fit the dataset set in LASSO model:
LASSO regression is a regularization technique. It is used over regression methods for a more accurate prediction.



```
▾ Lasso
Lasso()
```



Ridge model distributed residual only on negative side not on positive data.

Coefficient of regression test:

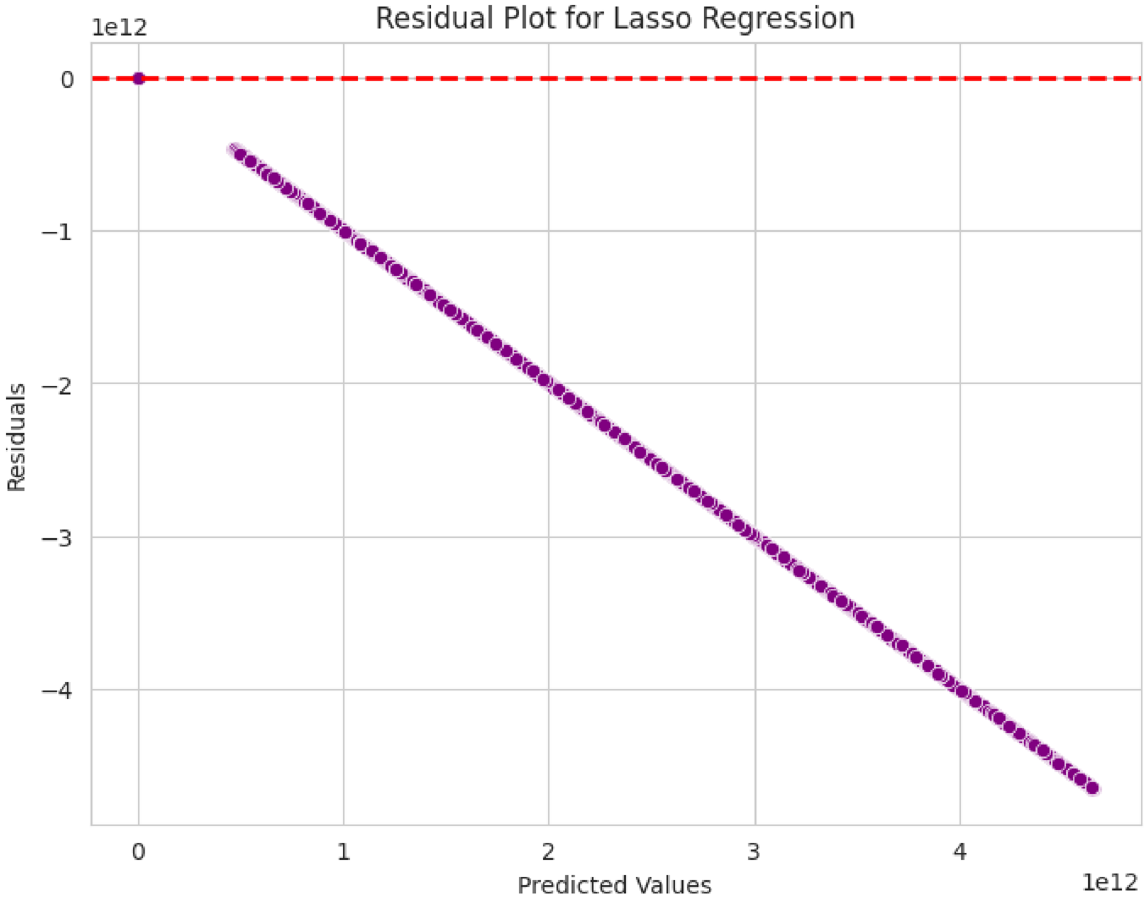
```
Mean Squared Error on Test Set: 5.373493875810108e+24
Lasso Coefficients: [      0.         -27919.7775436   -55976.40719441 1162398.49531821
   38079.36732673    8636.86139511  -36360.73927788]
True
```

From the above table we can observe that Current CTC, Inhand offer, Last appraisal rating seems to have good coefficient values towards target variable.

RMSE for lasso model for Train:

```
2305875270655.3564
```

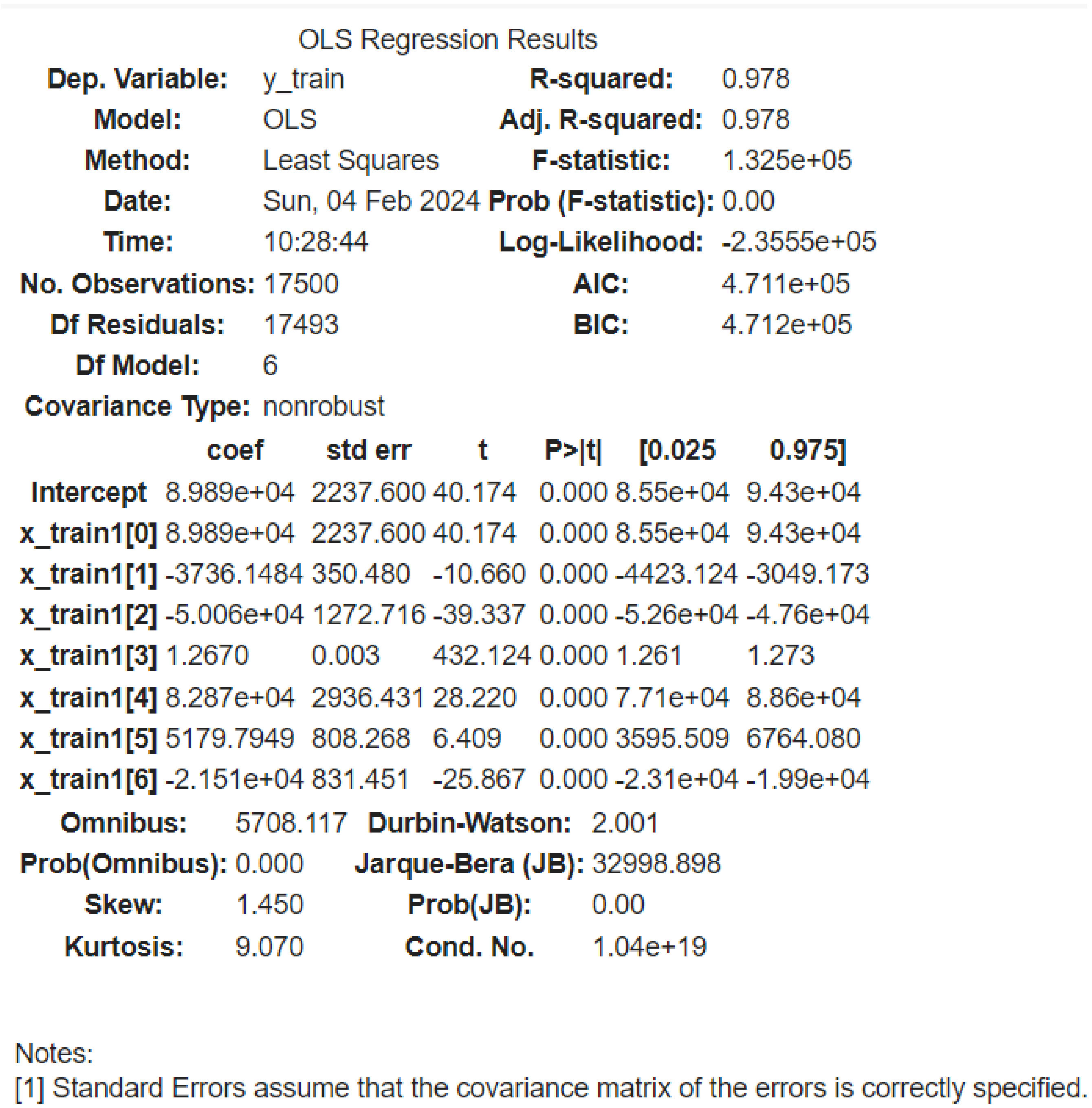
A lower RMSE indicates better model performance, as it reflects smaller prediction errors.



Lasso model distributed residuals only on negative side not on positive side.

OLS test model:

OLS test summery:



```
                         OLS Regression Results
     Dep. Variable:   y_train              R-squared:          0.978
             Model:   OLS              Adj. R-squared:         0.978
            Method:   Least Squares        F-statistic:        1.325e+05
              Date:   Sun, 04 Feb 2024 Prob (F-statistic):     0.00
              Time:   10:28:44         Log-Likelihood:         -2.3555e+05
  No. Observations:   17500                    AIC:            4.711e+05
      Df Residuals:   17493                    BIC:            4.712e+05
          Df Model:   6
   Covariance Type:   nonrobust
                    coef       std err     t       P>|t|   [0.025     0.975]
     Intercept   8.989e+04   2237.600  40.174   0.000  8.55e+04   9.43e+04
   x_train1[0]   8.989e+04   2237.600  40.174   0.000  8.55e+04   9.43e+04
   x_train1[1]  -3736.1484   350.480  -10.660   0.000  -4423.124  -3049.173
   x_train1[2]  -5.006e+04  1272.716  -39.337   0.000  -5.26e+04  -4.76e+04
   x_train1[3]   1.2670        0.003   432.124  0.000  1.261      1.273
   x_train1[4]   8.287e+04  2936.431  28.220   0.000  7.71e+04   8.86e+04
   x_train1[5]   5179.7949   808.268   6.409    0.000  3595.509   6764.080
   x_train1[6]  -2.151e+04   831.451  -25.867   0.000  -2.31e+04  -1.99e+04
       Omnibus:      5708.117   Durbin-Watson:     2.001
  Prob(Omnibus):     0.000     Jarque-Bera (JB):   32998.898
          Skew:      1.450          Prob(JB):      0.00
      Kurtosis:      9.070         Cond. No.       1.04e+19


  Notes:
  [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

[2] The smallest eigenvalue is 1.21e-20. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

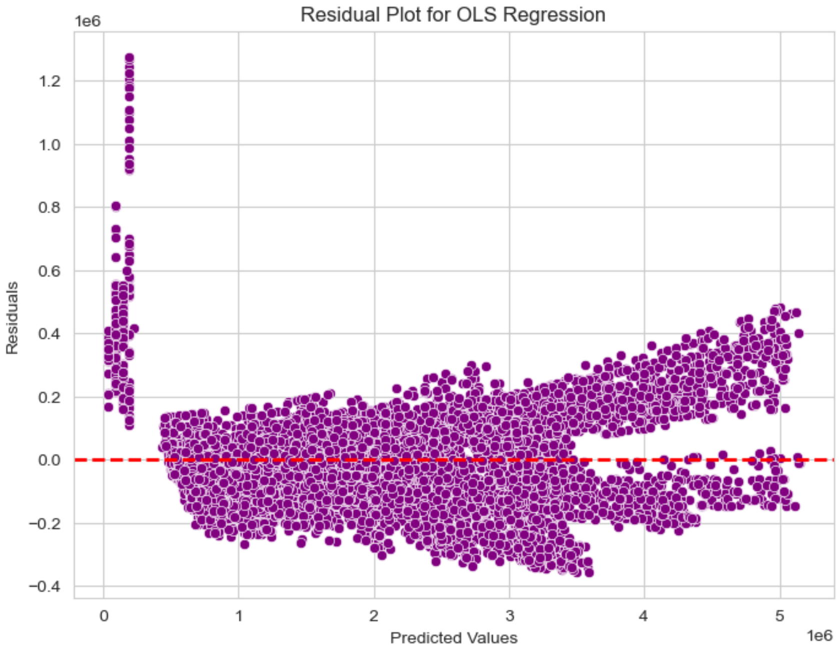
RSME OLS test:

167388.84393706292

It means that, on average, the predictions of your regression model have an error of approximately 170168.

53 units in the same scale as your target variable. A lower RMSE indicates better model performance, as it reflects smaller prediction errors.

Residual for OLS test:



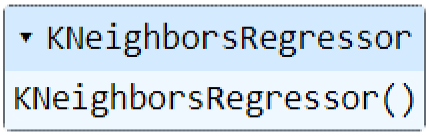
In OLS model by observing the plot we can get see that residuals have been distributed almost equally on positive and negative.

Non-Parametric Models

Performing K Nearest Neighbors Regression model:

we take the k nearest values of the target variable and compute the mean of those values. Those k nearest values act like regressors of linear regression.

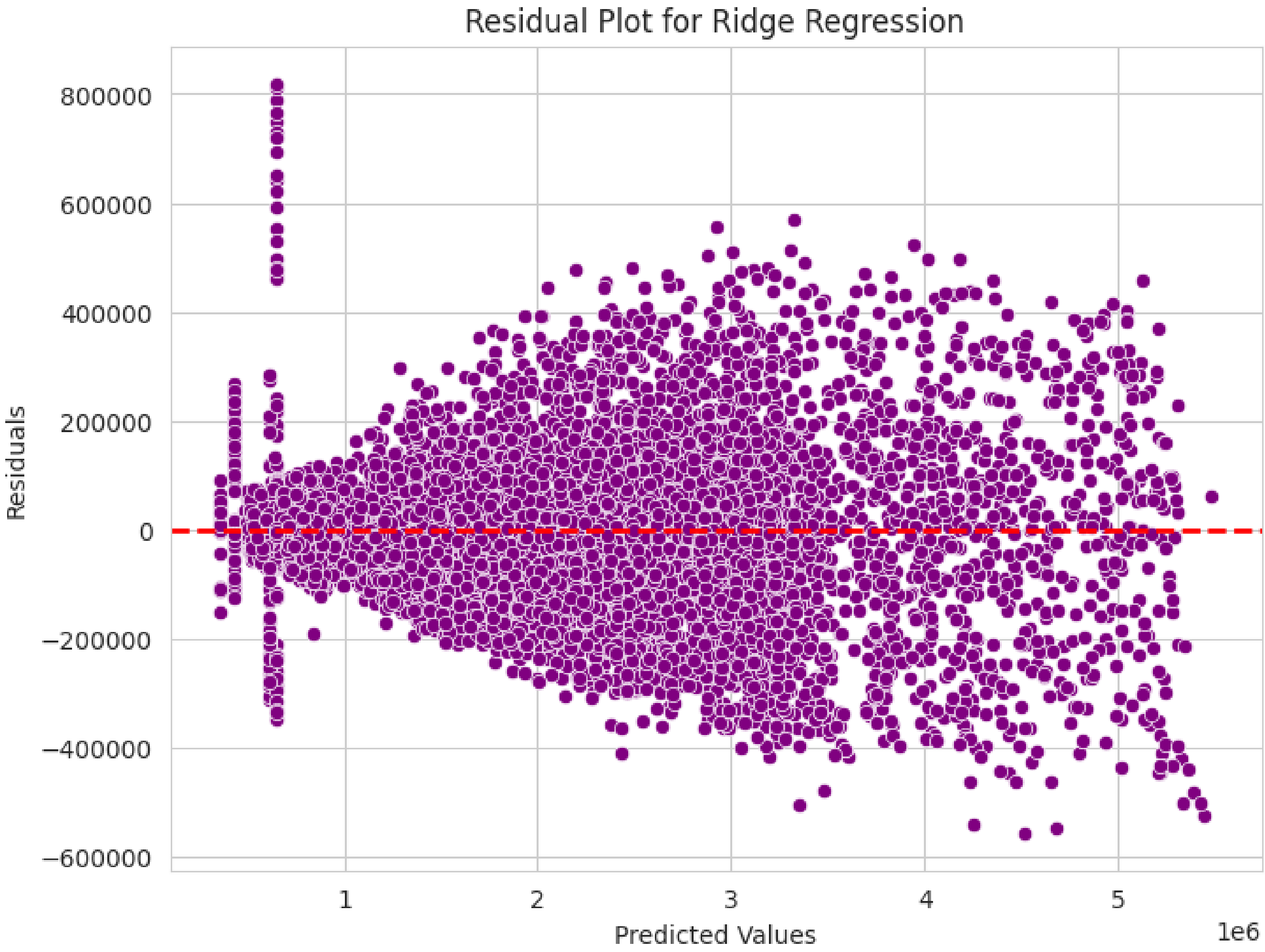
Fit dataset in the KNN model:



```
▾ KNeighborsRegressor
KNeighborsRegressor()
```

Now we calculate the mean square error on test set:

```
Mean Squared Error on Test Set: 24595411848.201466
```



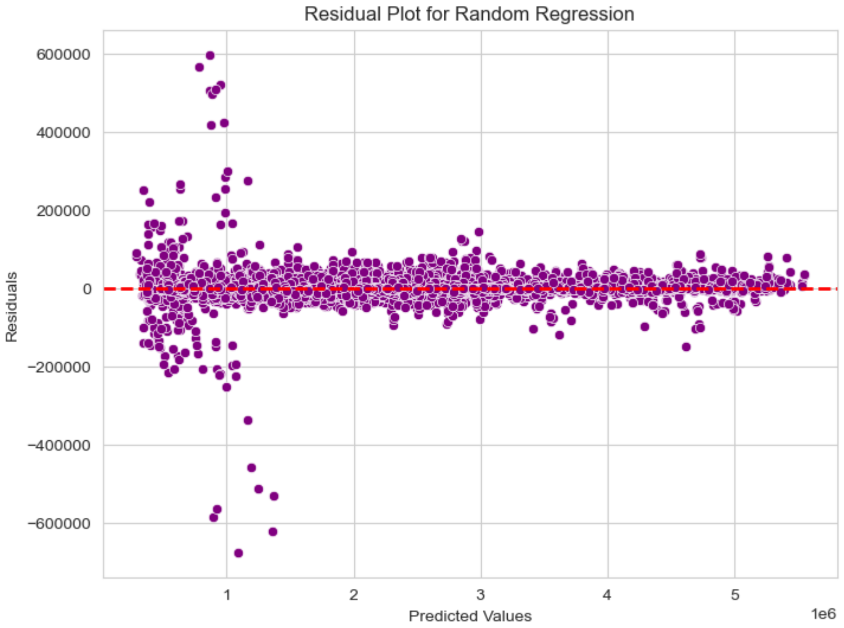
In KNN regression model by observing the plot we can get clear picture that residuals have been distributed equally on positive and negative sides.

RMSE KNN test:

127214.58938566186

RMSE of 127214.58938566186 means that, on average, the predictions of your regression model have an error of approximately 170168.53 units in the same scale as your target variable. A lower RMSE indicates better model performance, as it reflects smaller prediction errors.

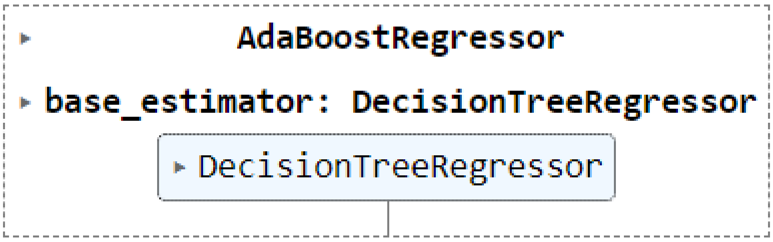
Random Forest test:



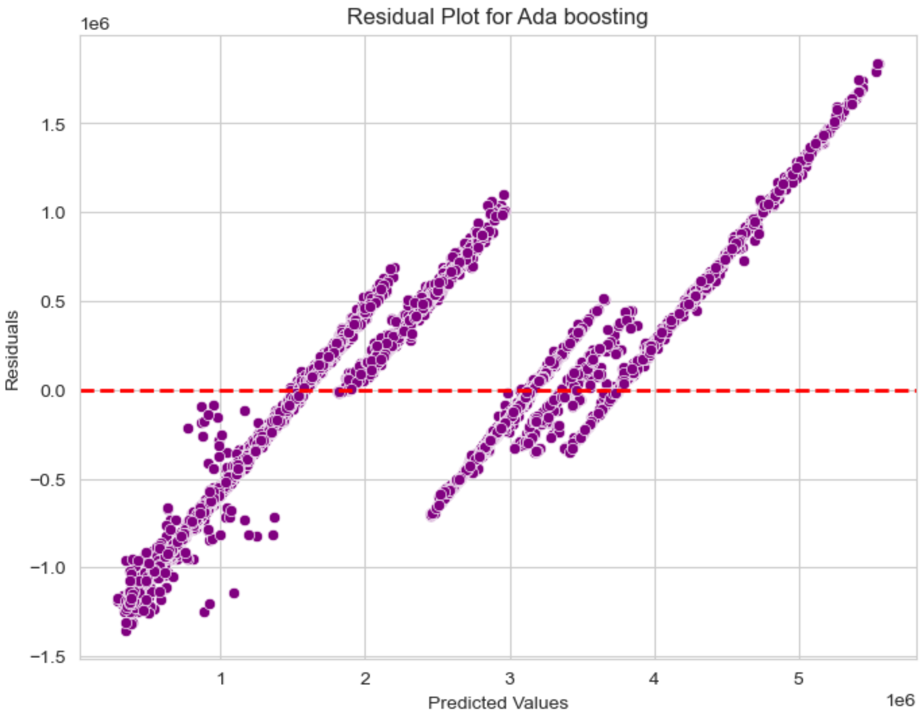
In Random Forest regression model by observing the plot we can get clear picture that residuals has been distributed equally on positive and negative sides.

Ada Boosting Model test:

Fit the data set :



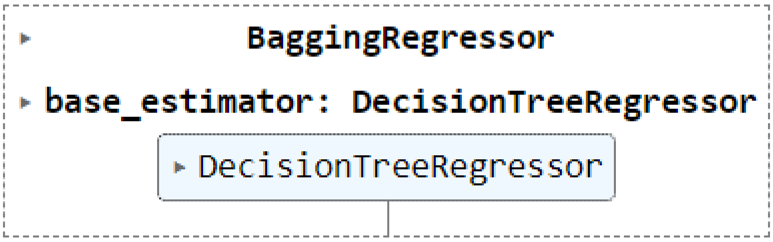
```
▸ AdaBoostRegressor
▸ base_estimator: DecisionTreeRegressor
▸ DecisionTreeRegressor
```



From the above plot we can observe so splitting pattern, so in this case residuals are not normal. Hence, model is not performing that great.

Bagging Model test:

Fit the dataset:



```
▸ BaggingRegressor
▸ base_estimator: DecisionTreeRegressor
▸ DecisionTreeRegressor
```

RMSE for Bagging train: 557786.298959022

RMSE of **557786.298959022**, it means that, on average, the predictions of your regression model have an error of approximately **557786.298959022**units in the same scale as your target variable. A lower RMSE indicates better model performance, as it reflects smaller prediction errors.

9.2 Mean Squared error Bagging test:

```
Mean Squared Error on Test Set: 1218804212.4628923
```

Interpretation of the models:

Parametric models	**RMSE Train**	**RMSE Test**	**R-Squared Train**	**R-Squared Test**
Linear Regression Stats	170168.53	168280.97	0.978	0.979
Linear Regression Scikit	170168.53	168280.97	0.978331241	0.979339517
Ridge Regression	2.26E+12	2.27E+12	-3.81E+12	-3.76E+12
Lasso Regression	2.26E+12	2.27E+12	-3.81E+12	-3.77E+12
OLS Regression	170168.53	168280.97	0.978331241	0.979339517
Non Parametric models	**RMSE Train**	**RMSE Test**	**R-Squared Train**	**R-Squared Test**
K Nearest Neighbors Regression	127799.88	127214.59	0.987778164	0.988192872
Random Forest Model	14318.22	34553.27	0.99984659	0.999128939
Ada Boosting Model	547722.15	557786.3	0.77551042	0.77301031
Bagging Model	547722.15	557786.3	0.77551042	0.77301031

Insights and Recommendations for Clustering:

- Had good inertia value by increasing number of clusters from 1 to 10. However, silhouettescore is good for cluster 2 and 3. Hence considering number of clusters as 3 for further analysis.
- Number of clusters more than 3, the projection of silhouette score is not good it’s goingbelow 0.10.
- Hence, number of clusters 3 is best to perform analysis.

Interpretations based on RMSE Score:

- The root mean squared value (RMSE) is a commonly used metric to evaluate the performance of a regression model. In the context of a regression model, RMSE measures the average magnitude of the errors between predicted values and actual values. Specifically, it calculates the square root of the average of the squared differences between predicted and actual values.
- Above all model’s Random forest has very minimal RMSE score which means it has less errors between predicted values and actual values. However, there is huge difference between train and test data set, train set reflects RMSE as **14318.22** whereas test set reflects **34553.27.** Hence its not performing that great.
- In next place KNN regression model it has train RMSE as **127799.88** and test RMSE as **127214.59.** Hence this model is performing good. We consider this model for business insights and recommendations.
- Residual plot also in KNN regression model by observing the plot we can get clear picture that residuals has been distributed equally on positive and negative sides. Hence data has been normally distributed.

Business implications:

Business Insights and Recommendations:

- During the hiring process, pay close attention to the candidates' Current CTC, Inhand offer, and Last Appraisal Ratings. These factors can be used to assess the candidate's expectations and potential fit within the organization's salary structure.

- Utilize the insights from the model to inform compensation policies. Consider adjusting salary structures based on the importance of Current CTC and Inhand offer. Additionally, use performance metrics associated with Last Appraisal Ratings to guide compensation decisions.

- From all the models we can observe strong coefficient for Current CTC, Inhand offer and Last Appraisal Ratings towards Expected CTC.

- Clearly communicate to employees how their Current CTC, Inhand offer, and Last Appraisal Ratings contribute to the determination of their Expected CTC. Transparency in salary calculations can foster trust and understanding among employees.

- Encourage employees to focus on improving their performance to receive higher appraisal ratings. A strong coefficient for Last Appraisal Ratings indicates that it strongly impacts the Expected CTC. Training programs, mentorship, and performance feedback can help employees enhance their skills and performance.

- Regularly monitor the model's performance and update it as necessary. Business conditions, industry standards, and employee expectations may change over time, so the model should be adapted to reflect these changes.

- For employees negotiating their salary or job offer, emphasize the importance of having a competitive Current CTC and Inhand offer. These variables seem to play a crucial role in determining the Expected CTC**.**