**Interim Report of**

**HR Data**

**Submitted By**

**Group No. 12 [Batch:** **2022 Location :Vadodara]**

**Group Members**

1. **Brijal Panchal**
2. **Dhev Sarath L.K**

**Research Supervisor**

**Uday kumar**

Contents

[1. Introduction to the Business Problem 3](#_Toc103791383)

[1.1. Problem Statement, Scope and Objective 3](#_Toc103791384)

[2. Data Report 3](#_Toc103791385)

[2.1. Data Preprocessing 5](#_Toc103791386)

[3. Exploratory Data Analysis (including insights from EDA) 11](#_Toc103791387)

[4. Modelling Approach: Initial model building 26](#_Toc103791388)

[4.1. Actionable insights and recommendations to the stakeholder 43](#_Toc103791389)

[5. References and Bibliography 43](#_Toc103791390)

[6. Appendix 43](#_Toc103791391)

1. Introduction to the Business Problem

Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles.

* 1. Problem Statement, Scope and Objective

To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles. Apart from the existing salary, there is a considerable number of factors regarding an employee’s experience and other abilities to which they get evaluated in interviews.

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.

1. Data Report

Data Dictionary

1. **IDX**: Row index
2. **Applicant\_ID**: Application ID
3. **Total\_Experience**: Total industry experience
4. **Total\_Experience\_in\_field\_applied**: Total experience in the field applied for (past work experience that is relevant to the job)
5. **Department**: Department name of current company
6. **Role**: Role in the current company
7. **Industry**: Industry name of current field
8. **Organization**: Organization name
9. **Designation**: Designation in current company
10. **Education**: Education
11. **Graduation\_Specialization**: Specialization subject in graduation
12. **University\_Grad**: University or college in Graduation
13. **Passing\_Year\_Of\_Graduation**: Year of passing Graduation
14. **PG\_Specialization**: Specialization subject in Post-Graduation
15. **University\_PG**: University or college in Post-Graduation
16. **Passing\_Year\_Of\_PG**: Year of passing Post Graduation
17. **PHD\_Specialization**: Specialization subject in Post-Graduation
18. **University\_PHD**: University or college in Post Doctorate
19. **Passing\_Year\_Of\_PHD**: Year of passing PHD
20. **Curent\_Location**: Curent Location
21. **Preferred\_location**: Preferred location to work in the company applied
22. **Current\_CTC**: Current CTC
23. **Inhand\_Offer**: Holding any offer in hand (Y: Yes, N:No)
24. **Last\_Appraisal\_Rating**: Last Appraisal Rating in current company
25. **No\_Of\_Companies\_worked**: No. of companies worked till date
26. **Number\_of\_Publications**: Number of papers published
27. **Certifications**: Number of relevant certifications completed
28. **International\_degree\_any**: Hold any international degree (1: Yes, 0: No)
29. **Expected\_CTC**: Expected CTC (Final CTC offered by Delta Ltd.

Load the required packages, set the working directory and load the data file.

Dataset has 25000 rows and 29 columns

Let us start the data exploration step with the head method to look at the first five rows.



*Table1.1: DataFrame head with top five rows and first ten columns*



*Table1.2: DataFrame head with top five rows and middle ten columns*



*Table1.3: DataFrame head with top five rows and last ten columns*

A quick look at the dataset and data dictionary tells us that the independent variables consist of sixteen (16) categorical variables, and thirteen (13) Continuous (integer, float) variables. Also all categorical variables are not encoded (i.e. have numerical values). The target variable is a Continuous categorical variable (Expected\_CTC)

We have NULL values present in the data in columns PHD\_Specialization (11881), University\_PHD (11881), Passing\_Year\_Of\_PHD(11881), PG\_Specialization(7692), University\_PG (7692), Passing\_Year\_Of\_PG (7692), Graduation\_Specialization (6180), University\_Grad (6180), Passing\_Year\_Of\_Graduation (6180), Designation (3129), Department (2778), Role (963), Industry (908), Organization (908), Last\_Appraisal\_Rating (908). Most of the rows are associated with freshers or employees who are graduates.

We do not have duplicated rows.

* 1. Data Preprocessing

There is missing value present in PHD\_Specialization, University\_PHD, PG\_Specialization, University\_PG, Graduation\_Specialization, University\_Grad, most of this columns are from Under graduation education or have done graduation or post-graduation category. So imputed them with ‘None’ as a value. As well as same rows are having Passing\_Year\_Of\_PHD, Passing\_Year\_Of\_PG, Passing\_Year\_Of\_Graduation empty, imputed those rows with 0 value.

Few of the rows(4362) have done education as Grad but in the details we found details of PHD\_Specialization, corrected those rows with Doctorate in Education column

Some of the rows(1363) have done education as Grad but in the details we found details of PG \_Specialization, corrected those rows with PG in Education column

Some of the rows(4390) have done education as PG but in the details we found details of PHD \_Specialization, corrected those rows with Doctorate in Education column

Some of the rows(1412) have done education as Doctorate but in the details we do not found details of PHD \_Specialization but PG \_Specialization is not empty, corrected those rows with PG in Education column

Some of the rows(522) have done education as PG but in the details we do not found details of PG \_Specialization but Graduation \_Specialization is not empty, corrected those rows with Grad in Education column

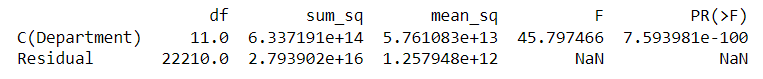
Imputed missing value of Designation, Department, Last\_Appraisal\_Rating, Organization, Industry, Role As ‘None’, will update the missing value imputation using KNN model to find the prediction of those missing values

Dropped IDX, Applicant\_ID column as it is not play vital role in prediction model

Checking correlation of categorical independent variable with target variable using Anova test and label encoding according to its level with respect to target variable or dropping the column if it do not have correlation with target variable

Anova test:

1. **Department**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means at least one particular category in the 'Department' variable has different mean of rate as compared to the other categories.

Label encoding: {' Group 4':0, 'Group 3': 1, 'Group 2': 2, 'Group 1':3}

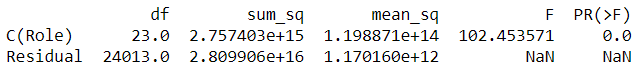
Group 1: Top Management

Group 2: Analytics/BI, Banking, Engineering, HR, Healthcare, IT-Software, Others, Sales

Group 3: Accounts, Education, Marketing

Group 4: None

1. **Role**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means at least one particular category in the 'Role' variable has different mean of rate as compared to the other categories.

Label encoding: {' Group 4':0, 'Group 3': 1, 'Group 2': 2, 'Group 1':3}

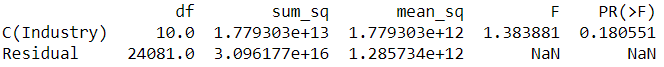
Group 1: Research Scientist, CEO, Head, Sr. Business Analyst, Senior Analyst, Area Sales Manager, Senior Researcher

Group 2: Project Manager, Business Analyst, Others, Sales Manager, Team Lead, Bio statistician, Consultant, Sales Execituve, Data scientist, Analyst, Lab Executuve, Financial Analyst, Researcher, Scientist, Principal Analyst,

Group 3: Professor, Associate

Group 4: None

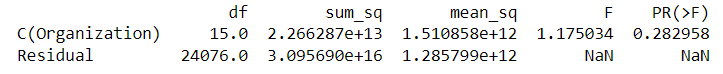
1. **Industry**:



At 95% confidence level, We see that the p-value is greater than alpha (0.05). Thus, we Accept the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means all category in the 'Industry' variable has same mean of rate.

**Dropping Industry column**

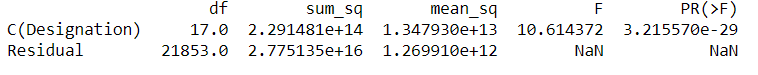
1. **Organization**:



At 95% confidence level, We see that the p-value is greater than alpha (0.05). Thus, we Accept the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means all category in the 'Organization' variable has same mean of rate.

**Dropping Organization column**

1. **Designation**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means at least one particular category in the 'Designation' variable has different mean of rate as compared to the other categories.

Label encoding: {' Group 5':0, 'Group 4': 1, 'Group 3': 2, 'Group 2': 3, 'Group 1':4}

Group 1: Research Scientist

Group 2: Network Engineer,Sr.Manager,Consultant,Web Designer,Research Analyst,CA,HR,Product Manager,Manager,Assistant Manager,Marketing Manager,Data Analyst,Others,Software Developer,Director

Group 3: Medical Officer

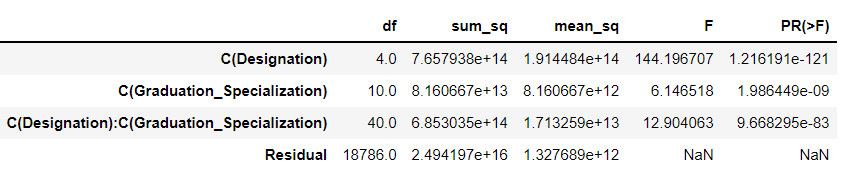
Group 4: Scientist

Group 5: None

1. **Education**:

Label encoding: {'Under Grad':0, 'Grad': 1, 'PG': 2, 'Doctorate': 3}

1. **Graduation\_Specialization**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we reject the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means 𝑇ℎ𝑒𝑟𝑒 𝑖𝑠 𝑖𝑛𝑡𝑒𝑟𝑎𝑐𝑡𝑖𝑜𝑛 𝑏𝑒𝑡𝑤𝑒𝑒𝑛 Designation 𝑎𝑛𝑑 Graduation\_Specialization

Label encoding: { Group 4':0, 'Group 3': 1, 'Group 2': 2, 'Group 1': 3})

Group 1: Sociology, Botony

Group 2: Engineering, Psychology, Zoology, Economics, Chemistry, Arts, Mathematics

Group 3: Others, Statistics

1. **University\_Grad,University\_PHD,University\_PG,Curent\_Location,Preferred\_location**:

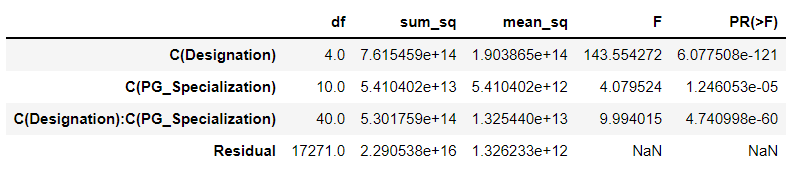
According to https://en.wikipedia.org/wiki/Classification\_of\_Indian\_cities converting city into tier -1, tier -2, tier -3

Label encoding: {'None':0, 'tier-3': 1, 'tier-2': 2, 'tier-1': 3}

tier-1: Ahmedabad, Bengaluru, Chennai, Delhi, Kolkata, Mumbai, Pune

tier-2: Surat, Kanpur, Lucknow, Nagpur, Jaipur, Guwahati, Bhubaneswar

tier-3: Mangalore

1. **PG\_Specialization**: 

At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we reject the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means 𝑇ℎ𝑒𝑟𝑒 𝑖𝑠 𝑖𝑛𝑡𝑒𝑟𝑎𝑐𝑡𝑖𝑜𝑛 𝑏𝑒𝑡𝑤𝑒𝑒𝑛 Designation 𝑎𝑛𝑑 PG\_Specialization

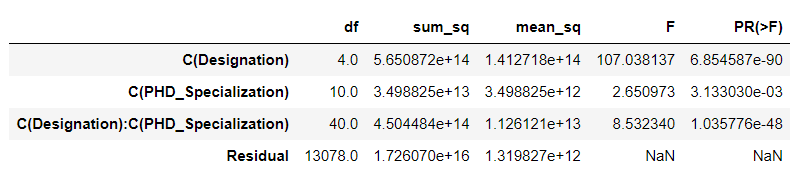
Label encoding: { Group 4':0, 'Group 3': 1, 'Group 2': 2, 'Group 1': 3})

Group 1: Sociology

Group 2: Psychology, Botony, Engineering, Others, Chemistry, Economics, Zoology, Arts, Mathematics

Group 3: Statistics

1. **PHD\_Specialization**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we reject the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means 𝑇ℎ𝑒𝑟𝑒 𝑖𝑠 𝑖𝑛𝑡𝑒𝑟𝑎𝑐𝑡𝑖𝑜𝑛 𝑏𝑒𝑡𝑤𝑒𝑒𝑛 Designation 𝑎𝑛𝑑 PHD\_Specialization

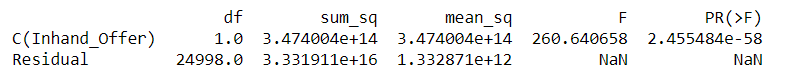
Label encoding: { Group 4':0, 'Group 3': 1, 'Group 2': 2, 'Group 1': 3})

Group 1: Sociology

Group 2: Engineering, Others, Economics, Zoology, Arts, Botony, Psychology, Statistics, Chemistry

Group 3: Mathematics

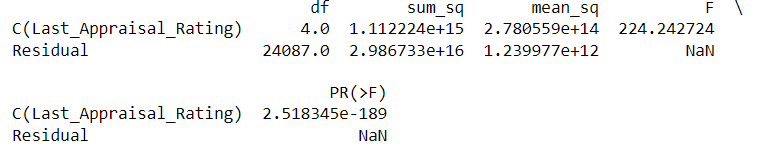
1. **Inhand\_Offer**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means at least one particular category in the 'Inhand\_Offer' variable has different mean of rate as compared to the other categories.

Label encoding: {'N':0, 'Y': 1}

1. **Last\_Appraisal\_Rating**:



At 95% confidence level, We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ). This means at least one particular category in the 'Last\_Appraisal\_Rating' variable has different mean of rate as compared to the other categories.

Label encoding: {'None':0, 'D': 1, 'C': 2, 'B': 3, 'A': 4, 'Key\_Performer': 5}

1. Exploratory Data Analysis (including insights from EDA)

Let us now take a look at the Summary Statistics using the describe method:



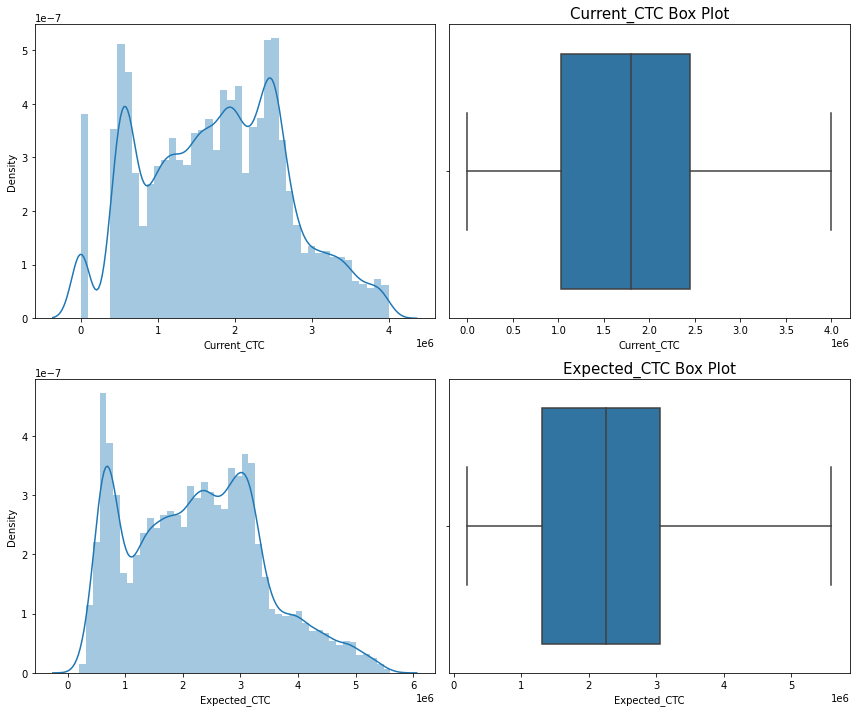
*Table2: Describe method with ‘include’ = ‘all’*

The mean/median Total\_Experience is close to 12 indicating a more experience candidates are applied. In addition, Total\_Experience\_in\_field\_applied median is 5 years, indicating the possibility of a healthy retention rate. Current\_CTC mean and median are 1760945.38 and 1802567.5 as well as standard deviation is very high indicating skewness in the data. Expected\_CTC mean and median are 2250154.51 and 2252136.5 as well as standard deviation is very high indicating skewness in the data. Minimum Expected\_CTC is 203744.0 and Maximum is 5599570.0.

Univariate analysis

Continuous Variables-

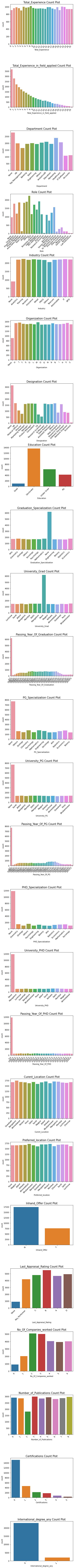
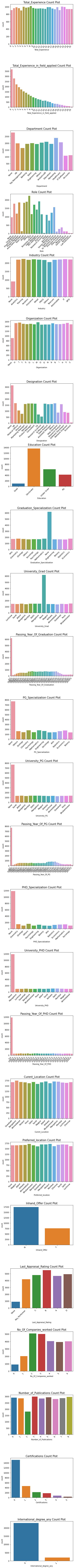
For performing Univariate analysis we will take a look at the Box plot and kde plot to get a fair understanding of the distributions -

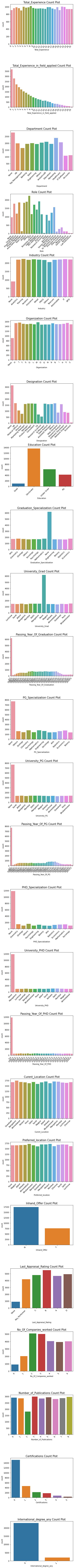
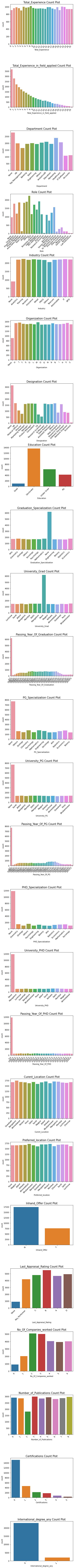


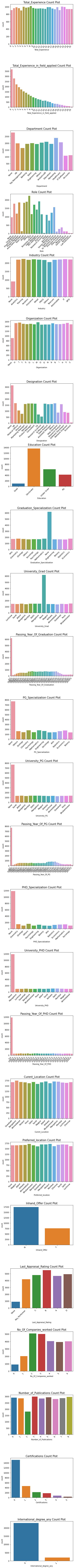
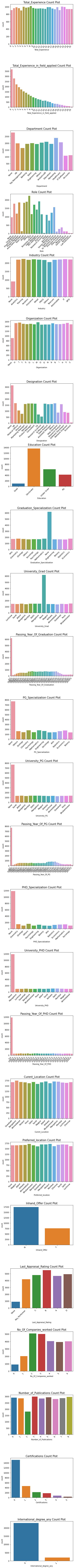
*Figure1: Kde plot and Box Plots of Continuous variables*

Categorical Variables –

For categorical variables let’s take a look at the count plots (Bar type) –

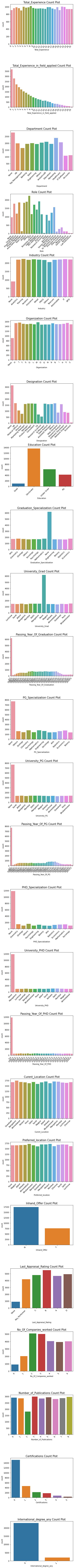


Figure 2: Count plot of all the categorical variables

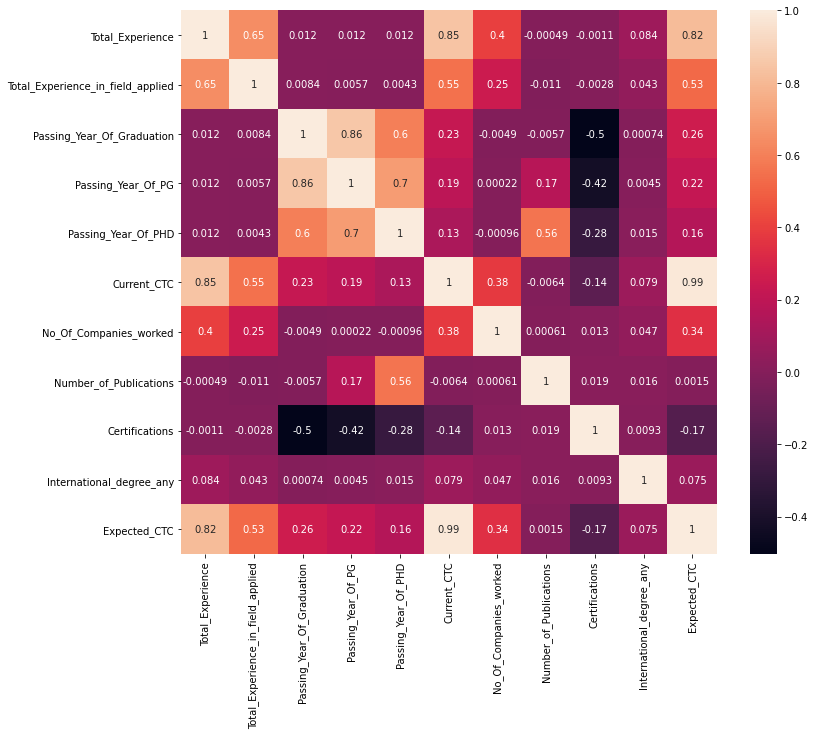
Inferences from Univariate analysis-

1. No Outliers are present in all Two continuous variables.
2. Expected\_CTC is positively skewed, whereas Current\_CTC is near normal distribution
3. All the continuous variables are in different scales so scaling will be performed (Not required for linear models though)
4. All the categorical variables have one clear majority class :
5. We do not have a high imbalance in the target variable

Bi-variate analysis –

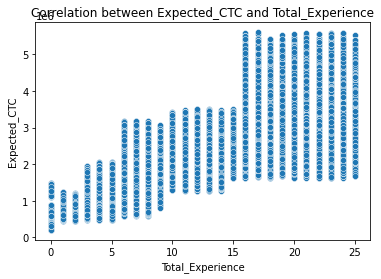
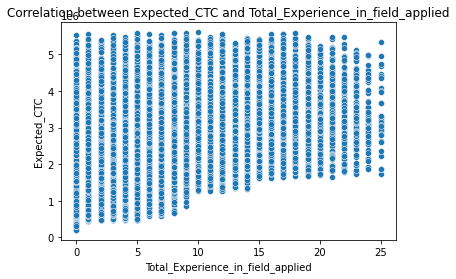
Continuous variables -

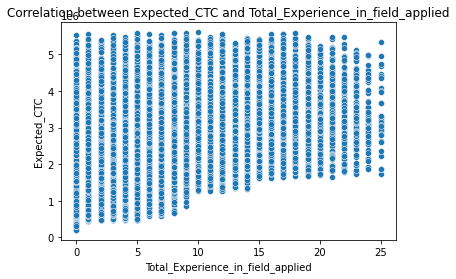
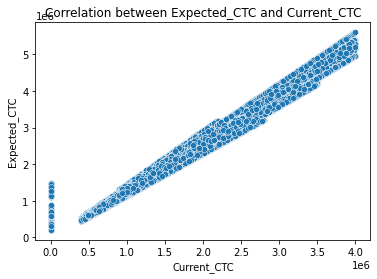
For Bi-variate analysis, lets take a look at the correlation, followed by scatter plot significant correlations –

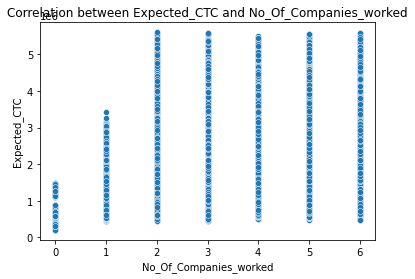
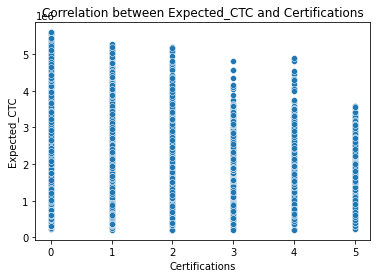


*Figure 3: Heatmap of correlation coefficients*

Scatter plots –

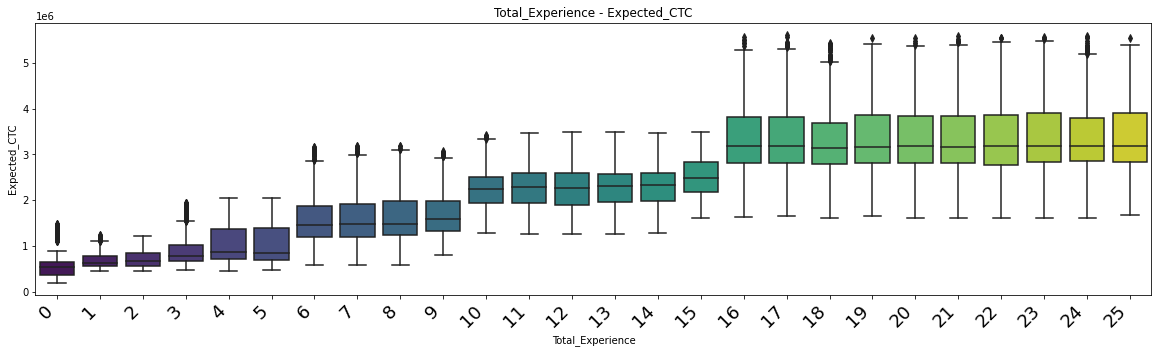
 

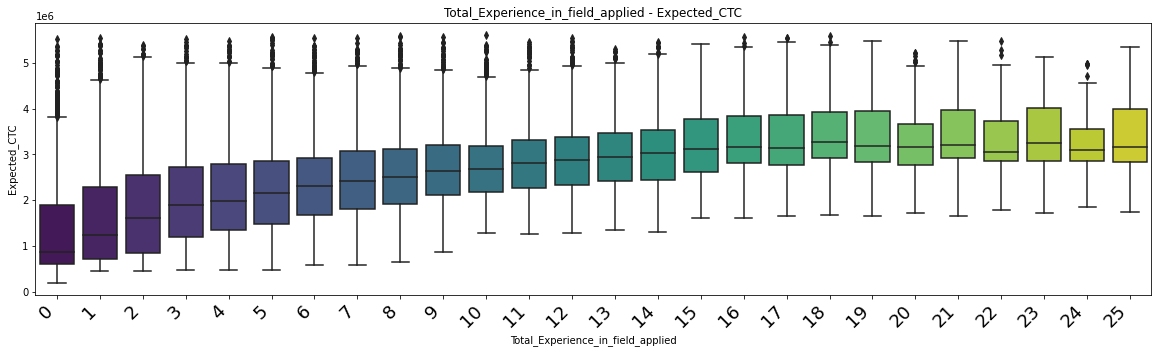
 

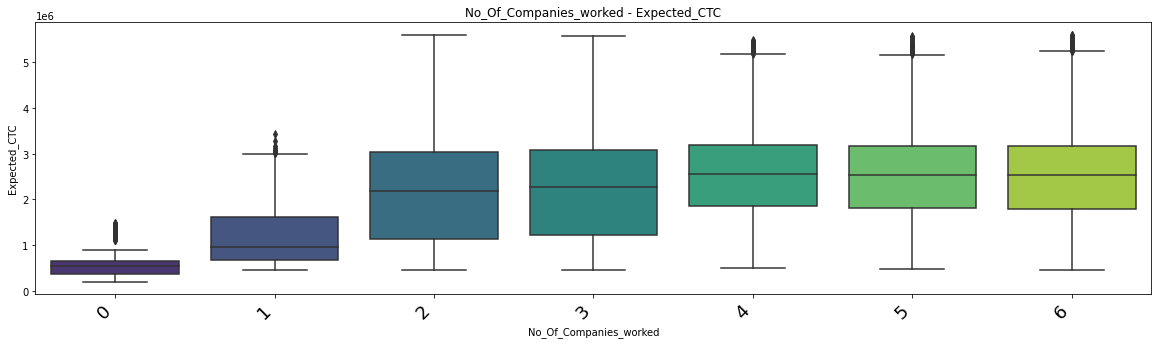
 

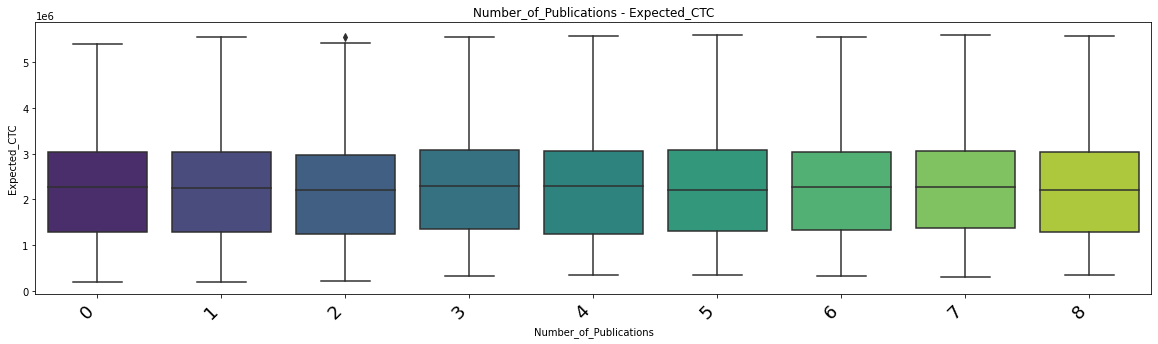
*Figure 4: Scatter plot of correlated features*

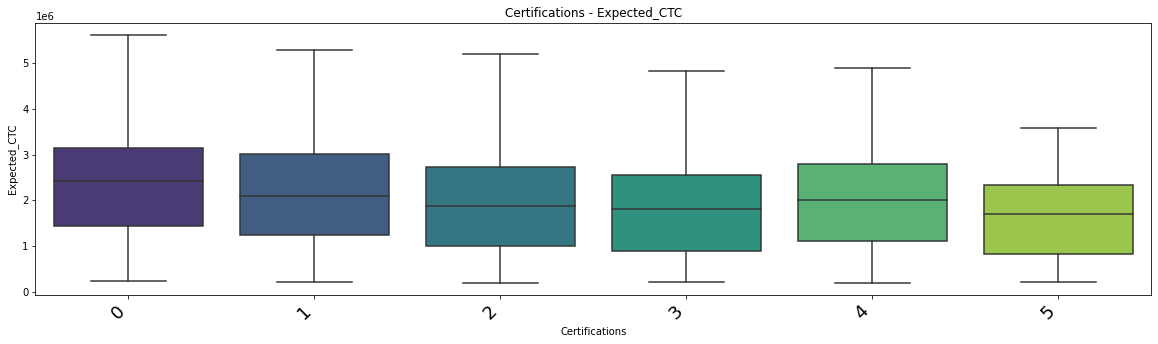
Continuous variables vs. Target variables Lets take a look at the box plots –

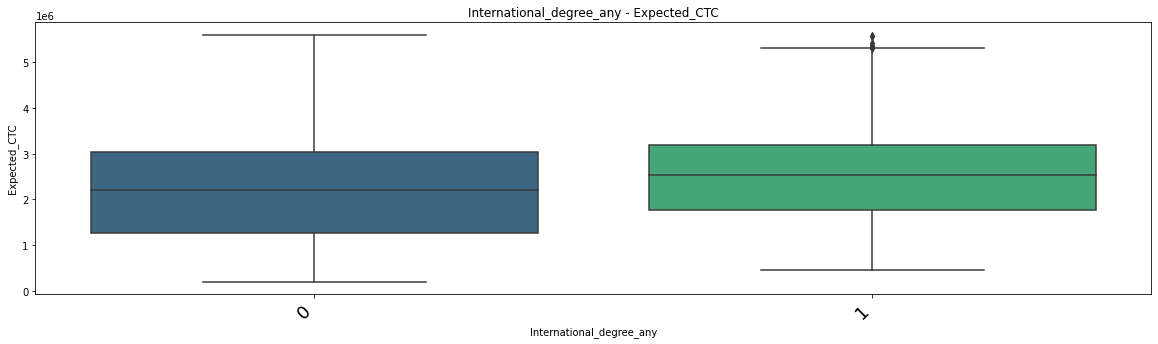






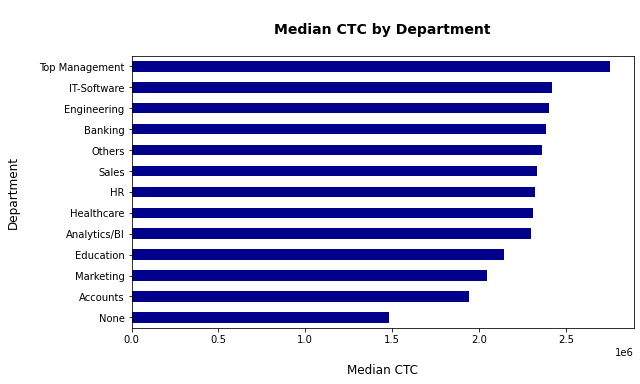


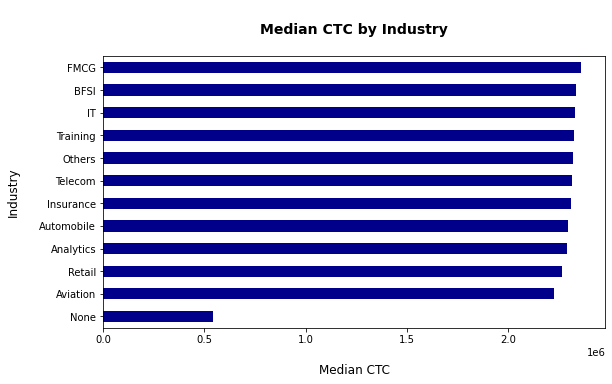
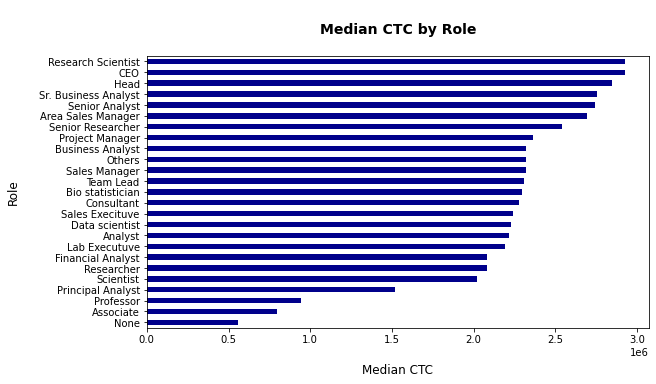


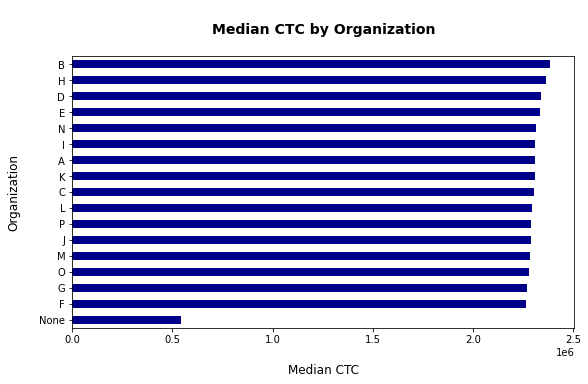


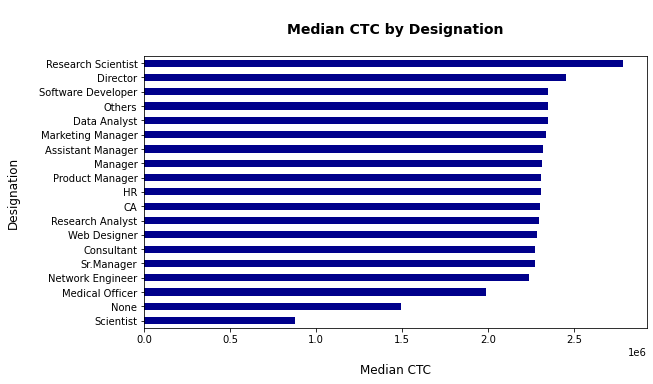
*Figure5: Box plots of Target variable vs. Continuous variables*

Target variable vs Categorical variables -

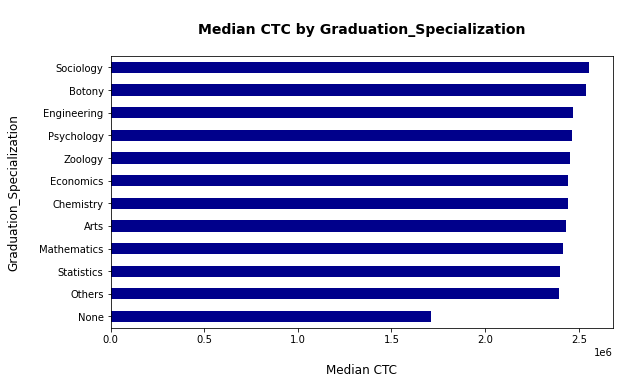


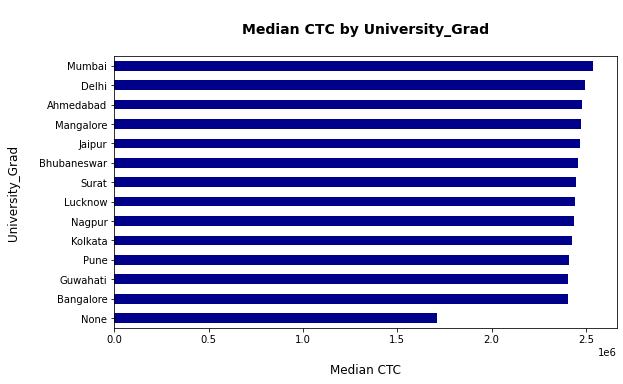


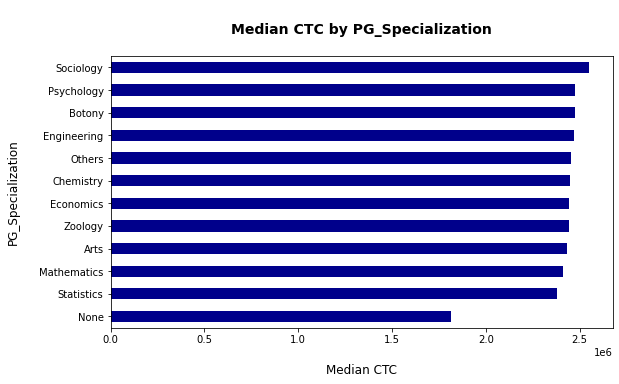


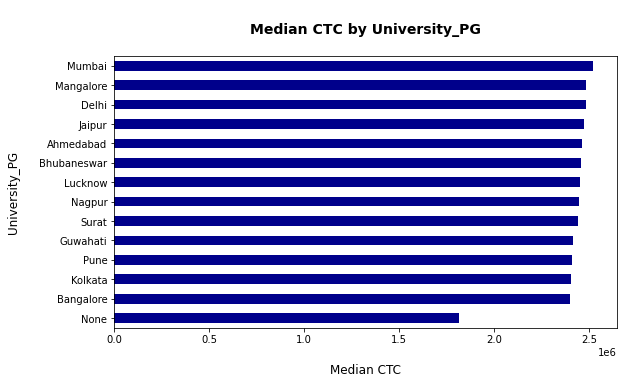


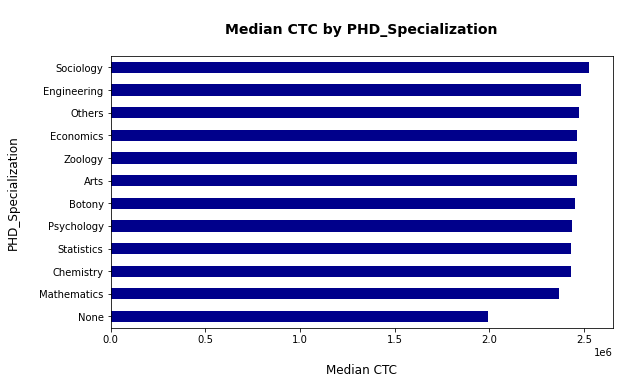


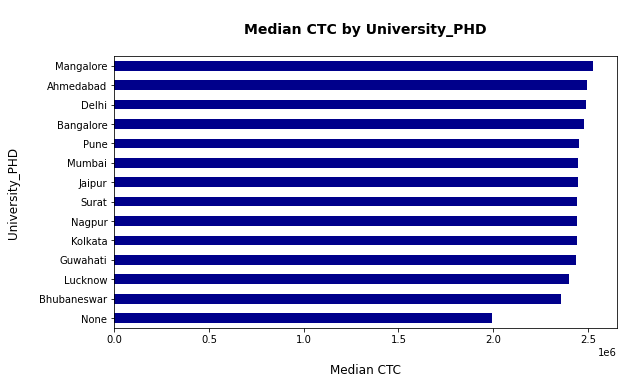


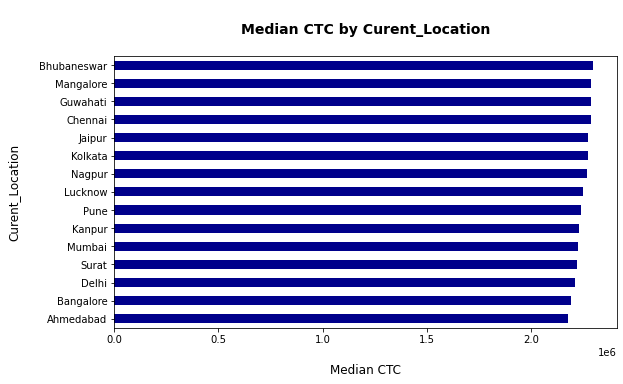


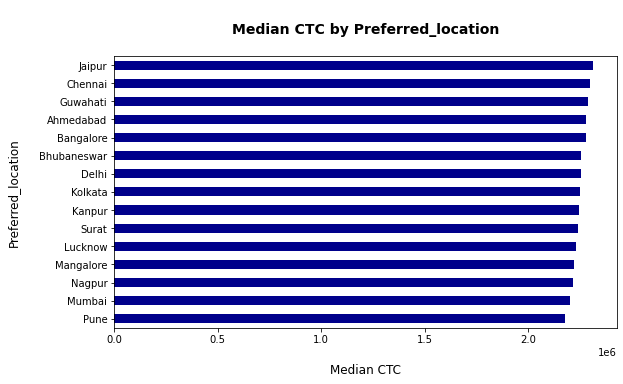




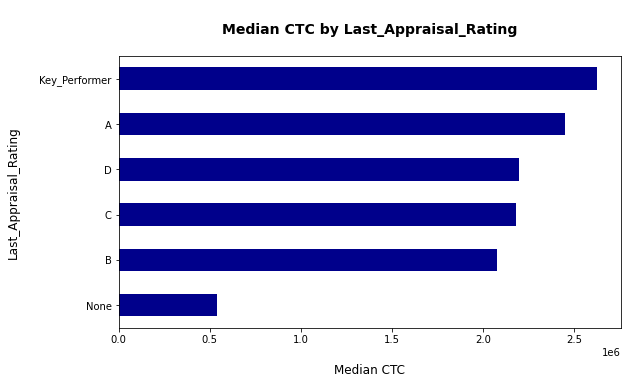












*Figure 6: median plot of Categorical variables using Target variable for color encoding*

Inferences from Bivariate Analysis-

1. As expected Total\_Experience, Total\_Experience\_in\_field\_applied, No\_Of\_Companies\_worked, Certifications, and Current\_CTC are all highly correlated to one another
2. It is evident from the bivariate plots that outliers detected in box plots are not so distant from the normal observations. (Also, the number of outliers seem to be reduced in Bi-variate box plots)
3. A strong linear relationship is observed in the scatter plots as well
4. Top Management, Research Scientist, is highest mean of expected salary
5. Modelling Approach: Initial model building

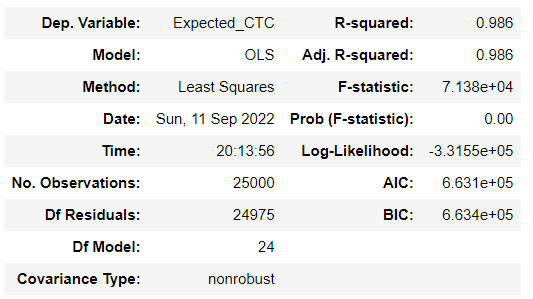
We are using Linear regression model to find the continuous prediction variable name Expected\_CTC.

As per the Heat-map we are seeing high correlation between independent variable

check for Multicolinearity-

OLS Regression Results for Formula

'Expected\_CTC~Total\_Experience+Total\_Experience\_in\_field\_applied+Department+Role+Designation+Education+Graduation\_Specialization+University\_Grad+Passing\_Year\_Of\_Graduation+PG\_Specialization+University\_PG+Passing\_Year\_Of\_PG+PHD\_Specialization+University\_PHD+Passing\_Year\_Of\_PHD+Curent\_Location+Preferred\_location+Current\_CTC+Inhand\_Offer+Last\_Appraisal\_Rating+No\_Of\_Companies\_worked+Number\_of\_Publications+Certifications+International\_degree\_any'



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 2.707e+05 | 6679.680 | 40.528 | 0.000 | 2.58e+05 | 2.84e+05 |
| **Total\_Experience** | -4718.8132 | 259.626 | -18.175 | 0.000 | -5227.695 | -4209.931 |
| **Total\_Experience\_in\_field\_applied** | -88.2988 | 198.069 | -0.446 | 0.656 | -476.526 | 299.929 |
| **Department** | -2.974e+04 | 1256.049 | -23.680 | 0.000 | -3.22e+04 | -2.73e+04 |
| **Role** | -1.331e+05 | 1945.973 | -68.375 | 0.000 | -1.37e+05 | -1.29e+05 |
| **Designation** | -3.598e+04 | 954.710 | -37.689 | 0.000 | -3.79e+04 | -3.41e+04 |
| **Education** | 5.615e+04 | 3813.480 | 14.724 | 0.000 | 4.87e+04 | 6.36e+04 |
| **Graduation\_Specialization** | -9506.6463 | 2270.916 | -4.186 | 0.000 | -1.4e+04 | -5055.517 |
| **University\_Grad** | 2201.6722 | 5671.236 | 0.388 | 0.698 | -8914.284 | 1.33e+04 |
| **Passing\_Year\_Of\_Graduation** | -28.4658 | 8.196 | -3.473 | 0.001 | -44.531 | -12.400 |
| **PG\_Specialization** | 5186.4199 | 3310.421 | 1.567 | 0.117 | -1302.201 | 1.17e+04 |
| **University\_PG** | -2508.9437 | 5916.314 | -0.424 | 0.672 | -1.41e+04 | 9087.380 |
| **Passing\_Year\_Of\_PG** | -6.8110 | 8.177 | -0.833 | 0.405 | -22.839 | 9.216 |
| **PHD\_Specialization** | -2.982e+04 | 2983.840 | -9.993 | 0.000 | -3.57e+04 | -2.4e+04 |
| **University\_PHD** | 347.8587 | 1948.481 | 0.179 | 0.858 | -3471.278 | 4166.996 |
| **Passing\_Year\_Of\_PHD** | 8.2466 | 4.533 | 1.819 | 0.069 | -0.638 | 17.132 |
| **Curent\_Location** | 480.9537 | 1448.795 | 0.332 | 0.740 | -2358.770 | 3320.678 |
| **Preferred\_location** | 541.9421 | 1437.999 | 0.377 | 0.706 | -2276.621 | 3360.506 |
| **Current\_CTC** | 1.3107 | 0.002 | 644.299 | 0.000 | 1.307 | 1.315 |
| **Inhand\_Offer** | 4.705e+04 | 2182.461 | 21.560 | 0.000 | 4.28e+04 | 5.13e+04 |
| **Last\_Appraisal\_Rating** | 5.225e+04 | 676.720 | 77.205 | 0.000 | 5.09e+04 | 5.36e+04 |
| **No\_Of\_Companies\_worked** | -1.318e+04 | 587.642 | -22.433 | 0.000 | -1.43e+04 | -1.2e+04 |
| **Number\_of\_Publications** | 440.5329 | 482.691 | 0.913 | 0.361 | -505.571 | 1386.637 |
| **Certifications** | -1.453e+04 | 862.714 | -16.842 | 0.000 | -1.62e+04 | -1.28e+04 |
| **International\_degree\_any** | 5.57e+04 | 3348.553 | 16.635 | 0.000 | 4.91e+04 | 6.23e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 8515.323 | **Durbin-Watson:** | 2.006 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 65585.888 |
| **Skew:** | 1.430 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 10.402 | **Cond. No.** | 1.83e+07 |

The condition number is large, 1.83e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Calculating VIF:

| **variables** | **VIF** |
| --- | --- |
| **23** | International\_degree\_any | 1.175076 |
| **18** | Inhand\_Offer | 1.861472 |
| **22** | Certifications | 1.914982 |
| **1** | Total\_Experience\_in\_field\_applied | 3.697701 |
| **19** | Last\_Appraisal\_Rating | 5.907272 |
| **21** | Number\_of\_Publications | 6.468275 |
| **2** | Department | 6.524183 |
| **20** | No\_Of\_Companies\_worked | 6.631554 |
| **4** | Designation | 9.004940 |
| **16** | Preferred\_location | 12.023251 |
| **15** | Curent\_Location | 12.627157 |
| **13** | University\_PHD | 15.723590 |
| **0** | Total\_Experience | 18.420925 |
| **3** | Role | 20.316974 |
| **17** | Current\_CTC | 21.026044 |
| **6** | Graduation\_Specialization | 21.891485 |
| **12** | PHD\_Specialization | 24.484539 |
| **9** | PG\_Specialization | 40.301653 |
| **14** | Passing\_Year\_Of\_PHD | 55.919184 |
| **5** | Education | 105.397896 |
| **7** | University\_Grad | 188.855947 |
| **10** | University\_PG | 188.983705 |
| **11** | Passing\_Year\_Of\_PG | 240.177851 |
| **8** | Passing\_Year\_Of\_Graduation | 259.713722 |

Dropping Education column as it has High VIF and p-value:

OLS Regression Results for Formula

'Expected\_CTC~Total\_Experience+Total\_Experience\_in\_field\_applied+Department+Role+Designation+Graduation\_Specialization+University\_Grad+Passing\_Year\_Of\_Graduation+PG\_Specialization+University\_PG+Passing\_Year\_Of\_PG+PHD\_Specialization+University\_PHD+Passing\_Year\_Of\_PHD+Curent\_Location+Preferred\_location+Current\_CTC+Inhand\_Offer+Last\_Appraisal\_Rating+No\_Of\_Companies\_worked+Number\_of\_Publications+Certifications+International\_degree\_any'

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Expected\_CTC | **R-squared:** | 0.986 |
| **Model:** | OLS | **Adj. R-squared:** | 0.985 |
| **Method:** | Least Squares | **F-statistic:** | 7.383e+04 |
| **Date:** | Sun, 11 Sep 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:13:58 | **Log-Likelihood:** | -3.3166e+05 |
| **No. Observations:** | 25000 | **AIC:** | 6.634e+05 |
| **Df Residuals:** | 24976 | **BIC:** | 6.636e+05 |
| **Df Model:** | 23 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 2.709e+05 | 6708.459 | 40.386 | 0.000 | 2.58e+05 | 2.84e+05 |
| **Total\_Experience** | -4948.2378 | 260.275 | -19.012 | 0.000 | -5458.392 | -4438.083 |
| **Total\_Experience\_in\_field\_applied** | -63.9988 | 198.916 | -0.322 | 0.748 | -453.886 | 325.889 |
| **Department** | -2.98e+04 | 1261.458 | -23.621 | 0.000 | -3.23e+04 | -2.73e+04 |
| **Role** | -1.328e+05 | 1954.309 | -67.976 | 0.000 | -1.37e+05 | -1.29e+05 |
| **Designation** | -3.617e+04 | 958.737 | -37.729 | 0.000 | -3.81e+04 | -3.43e+04 |
| **Graduation\_Specialization** | -9410.1172 | 2280.696 | -4.126 | 0.000 | -1.39e+04 | -4939.818 |
| **University\_Grad** | 3229.1970 | 5695.252 | 0.567 | 0.571 | -7933.834 | 1.44e+04 |
| **Passing\_Year\_Of\_Graduation** | 15.8144 | 7.658 | 2.065 | 0.039 | 0.805 | 30.824 |
| **PG\_Specialization** | 5037.0211 | 3324.676 | 1.515 | 0.130 | -1479.541 | 1.16e+04 |
| **University\_PG** | -3530.4604 | 5941.410 | -0.594 | 0.552 | -1.52e+04 | 8115.053 |
| **Passing\_Year\_Of\_PG** | 3.8300 | 8.180 | 0.468 | 0.640 | -12.204 | 19.864 |
| **PHD\_Specialization** | -2.995e+04 | 2996.690 | -9.994 | 0.000 | -3.58e+04 | -2.41e+04 |
| **University\_PHD** | 359.3472 | 1956.880 | 0.184 | 0.854 | -3476.253 | 4194.948 |
| **Passing\_Year\_Of\_PHD** | 36.5829 | 4.122 | 8.875 | 0.000 | 28.504 | 44.662 |
| **Curent\_Location** | 593.4731 | 1455.020 | 0.408 | 0.683 | -2258.453 | 3445.399 |
| **Preferred\_location** | 609.1449 | 1444.191 | 0.422 | 0.673 | -2221.555 | 3439.845 |
| **Current\_CTC** | 1.3133 | 0.002 | 645.200 | 0.000 | 1.309 | 1.317 |
| **Inhand\_Offer** | 4.73e+04 | 2191.802 | 21.583 | 0.000 | 4.3e+04 | 5.16e+04 |
| **Last\_Appraisal\_Rating** | 5.215e+04 | 679.603 | 76.730 | 0.000 | 5.08e+04 | 5.35e+04 |
| **No\_Of\_Companies\_worked** | -1.323e+04 | 590.166 | -22.416 | 0.000 | -1.44e+04 | -1.21e+04 |
| **Number\_of\_Publications** | 443.5094 | 484.772 | 0.915 | 0.360 | -506.673 | 1393.692 |
| **Certifications** | -1.513e+04 | 865.461 | -17.483 | 0.000 | -1.68e+04 | -1.34e+04 |
| **International\_degree\_any** | 5.614e+04 | 3362.856 | 16.694 | 0.000 | 4.95e+04 | 6.27e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 8426.541 | **Durbin-Watson:** | 2.005 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 63574.095 |
| **Skew:** | 1.419 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 10.279 | **Cond. No.** | 1.83e+07 |

The condition number is large, 1.83e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Calculating VIF:

|  | **variables** | **VIF** |
| --- | --- | --- |
| **22** | International\_degree\_any | 1.174986 |
| **17** | Inhand\_Offer | 1.861360 |
| **21** | Certifications | 1.910651 |
| **1** | Total\_Experience\_in\_field\_applied | 3.697443 |
| **18** | Last\_Appraisal\_Rating | 5.906697 |
| **20** | Number\_of\_Publications | 6.468267 |
| **2** | Department | 6.524135 |
| **19** | No\_Of\_Companies\_worked | 6.631373 |
| **4** | Designation | 9.003333 |
| **15** | Preferred\_location | 12.022950 |
| **14** | Curent\_Location | 12.626489 |
| **12** | University\_PHD | 15.723588 |
| **0** | Total\_Experience | 18.354658 |
| **3** | Role | 20.315576 |
| **16** | Current\_CTC | 20.869683 |
| **5** | Graduation\_Specialization | 21.891302 |
| **11** | PHD\_Specialization | 24.484313 |
| **8** | PG\_Specialization | 40.301279 |
| **13** | Passing\_Year\_Of\_PHD | 45.816870 |
| **6** | University\_Grad | 188.827509 |
| **9** | University\_PG | 188.957858 |
| **7** | Passing\_Year\_Of\_Graduation | 224.451622 |
| **10** | Passing\_Year\_Of\_PG | 238.302942 |

Dropping Passing\_Year\_Of\_Graduation,Passing\_Year\_Of\_PG,Passing\_Year\_Of\_PHD

OLS Regression Results for Formula

‘Expected\_CTC~Total\_Experience+Total\_Experience\_in\_field\_applied+Department+Role+Designation+Graduation\_Specialization+University\_Grad+PG\_Specialization+University\_PG+PHD\_Specialization+University\_PHD+Curent\_Location+Preferred\_location+Current\_CTC+Inhand\_Offer+Last\_Appraisal\_Rating+No\_Of\_Companies\_worked+Number\_of\_Publications+Certifications+International\_degree\_any'

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Expected\_CTC | **R-squared:** | 0.985 |
| **Model:** | OLS | **Adj. R-squared:** | 0.985 |
| **Method:** | Least Squares | **F-statistic:** | 8.442e+04 |
| **Date:** | Sun, 11 Sep 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:13:59 | **Log-Likelihood:** | -3.3173e+05 |
| **No. Observations:** | 25000 | **AIC:** | 6.635e+05 |
| **Df Residuals:** | 24979 | **BIC:** | 6.637e+05 |
| **Df Model:** | 20 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 2.753e+05 | 6637.037 | 41.481 | 0.000 | 2.62e+05 | 2.88e+05 |
| **Total\_Experience** | -5116.0797 | 259.889 | -19.686 | 0.000 | -5625.477 | -4606.682 |
| **Total\_Experience\_in\_field\_applied** | -59.7479 | 199.468 | -0.300 | 0.765 | -450.716 | 331.220 |
| **Department** | -2.943e+04 | 1264.590 | -23.271 | 0.000 | -3.19e+04 | -2.69e+04 |
| **Role** | -1.345e+05 | 1954.373 | -68.810 | 0.000 | -1.38e+05 | -1.31e+05 |
| **Designation** | -3.544e+04 | 958.730 | -36.965 | 0.000 | -3.73e+04 | -3.36e+04 |
| **Graduation\_Specialization** | -6685.5956 | 2127.109 | -3.143 | 0.002 | -1.09e+04 | -2516.336 |
| **University\_Grad** | 1.185e+04 | 2605.691 | 4.546 | 0.000 | 6739.088 | 1.7e+04 |
| **PG\_Specialization** | 1.182e+04 | 2708.545 | 4.363 | 0.000 | 6507.489 | 1.71e+04 |
| **University\_PG** | -4868.4723 | 2819.506 | -1.727 | 0.084 | -1.04e+04 | 657.926 |
| **PHD\_Specialization** | -1.026e+04 | 2082.399 | -4.928 | 0.000 | -1.43e+04 | -6181.365 |
| **University\_PHD** | 1.263e+04 | 1587.768 | 7.956 | 0.000 | 9519.625 | 1.57e+04 |
| **Curent\_Location** | 856.6748 | 1458.510 | 0.587 | 0.557 | -2002.090 | 3715.439 |
| **Preferred\_location** | 522.4015 | 1448.256 | 0.361 | 0.718 | -2316.266 | 3361.069 |
| **Current\_CTC** | 1.3148 | 0.002 | 647.726 | 0.000 | 1.311 | 1.319 |
| **Inhand\_Offer** | 4.963e+04 | 2189.372 | 22.670 | 0.000 | 4.53e+04 | 5.39e+04 |
| **Last\_Appraisal\_Rating** | 5.177e+04 | 680.798 | 76.050 | 0.000 | 5.04e+04 | 5.31e+04 |
| **No\_Of\_Companies\_worked** | -1.328e+04 | 591.785 | -22.448 | 0.000 | -1.44e+04 | -1.21e+04 |
| **Number\_of\_Publications** | 909.1718 | 464.034 | 1.959 | 0.050 | -0.362 | 1818.706 |
| **Certifications** | -1.609e+04 | 856.566 | -18.785 | 0.000 | -1.78e+04 | -1.44e+04 |
| **International\_degree\_any** | 5.743e+04 | 3369.944 | 17.041 | 0.000 | 5.08e+04 | 6.4e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 8620.179 | **Durbin-Watson:** | 2.005 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 67560.326 |
| **Skew:** | 1.445 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 10.517 | **Cond. No.** | 1.52e+07 |

The condition number is large, 1.52e+07. This might indicate that there are strong multicollinearity or other numerical problems.

|  | **variables** | **VIF** |
| --- | --- | --- |
| **19** | International\_degree\_any | 1.173998 |
| **14** | Inhand\_Offer | 1.847258 |
| **18** | Certifications | 1.872342 |
| **1** | Total\_Experience\_in\_field\_applied | 3.696954 |
| **15** | Last\_Appraisal\_Rating | 5.893259 |
| **17** | Number\_of\_Publications | 5.979910 |
| **2** | Department | 6.516288 |
| **16** | No\_Of\_Companies\_worked | 6.628451 |
| **4** | Designation | 8.923596 |
| **10** | University\_PHD | 10.287157 |
| **9** | PHD\_Specialization | 11.681645 |
| **12** | Preferred\_location | 11.906823 |
| **11** | Curent\_Location | 12.440706 |
| **0** | Total\_Experience | 18.207099 |
| **5** | Graduation\_Specialization | 18.898354 |
| **3** | Role | 20.244394 |
| **13** | Current\_CTC | 20.667091 |
| **7** | PG\_Specialization | 26.452173 |
| **6** | University\_Grad | 38.727063 |
| **8** | University\_PG | 42.098268 |

Dropping column Current\_CTC

OLS Regression Results for formula

'Expected\_CTC~Total\_Experience+Total\_Experience\_in\_field\_applied+Department+Role+Designation+Graduation\_Specialization+University\_Grad+PG\_Specialization+University\_PG+PHD\_Specialization+University\_PHD+Curent\_Location+Preferred\_location+Inhand\_Offer+Last\_Appraisal\_Rating+No\_Of\_Companies\_worked+Number\_of\_Publications+Certifications+International\_degree\_any'

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Expected\_CTC | **R-squared:** | 0.741 |
| **Model:** | OLS | **Adj. R-squared:** | 0.740 |
| **Method:** | Least Squares | **F-statistic:** | 3753. |
| **Date:** | Sun, 11 Sep 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:14:01 | **Log-Likelihood:** | -3.6772e+05 |
| **No. Observations:** | 25000 | **AIC:** | 7.355e+05 |
| **Df Residuals:** | 24980 | **BIC:** | 7.356e+05 |
| **Df Model:** | 19 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 1.117e+05 | 2.8e+04 | 3.993 | 0.000 | 5.69e+04 | 1.67e+05 |
| **Total\_Experience** | 1.217e+05 | 721.034 | 168.774 | 0.000 | 1.2e+05 | 1.23e+05 |
| **Total\_Experience\_in\_field\_applied** | 805.5948 | 841.426 | 0.957 | 0.338 | -843.651 | 2454.840 |
| **Department** | 4.711e+04 | 5311.277 | 8.870 | 0.000 | 3.67e+04 | 5.75e+04 |
| **Role** | 3833.5619 | 8195.076 | 0.468 | 0.640 | -1.22e+04 | 1.99e+04 |
| **Designation** | -1.622e+04 | 4042.423 | -4.013 | 0.000 | -2.41e+04 | -8297.161 |
| **Graduation\_Specialization** | 5.296e+04 | 8964.702 | 5.908 | 0.000 | 3.54e+04 | 7.05e+04 |
| **University\_Grad** | 1.593e+05 | 1.09e+04 | 14.551 | 0.000 | 1.38e+05 | 1.81e+05 |
| **PG\_Specialization** | 9.84e+04 | 1.14e+04 | 8.623 | 0.000 | 7.6e+04 | 1.21e+05 |
| **University\_PG** | -8.895e+04 | 1.19e+04 | -7.486 | 0.000 | -1.12e+05 | -6.57e+04 |
| **PHD\_Specialization** | 1.022e+04 | 8783.494 | 1.163 | 0.245 | -6997.061 | 2.74e+04 |
| **University\_PHD** | 2.008e+04 | 6697.753 | 2.998 | 0.003 | 6950.911 | 3.32e+04 |
| **Curent\_Location** | 4634.3974 | 6152.608 | 0.753 | 0.451 | -7425.077 | 1.67e+04 |
| **Preferred\_location** | -874.4488 | 6109.397 | -0.143 | 0.886 | -1.28e+04 | 1.11e+04 |
| **Inhand\_Offer** | 7.109e+04 | 9234.709 | 7.698 | 0.000 | 5.3e+04 | 8.92e+04 |
| **Last\_Appraisal\_Rating** | 7.434e+04 | 2868.156 | 25.918 | 0.000 | 6.87e+04 | 8e+04 |
| **No\_Of\_Companies\_worked** | 6233.7992 | 2493.179 | 2.500 | 0.012 | 1347.021 | 1.11e+04 |
| **Number\_of\_Publications** | -6262.8071 | 1956.950 | -3.200 | 0.001 | -1.01e+04 | -2427.070 |
| **Certifications** | -7.402e+04 | 3593.640 | -20.596 | 0.000 | -8.11e+04 | -6.7e+04 |
| **International\_degree\_any** | 1.917e+04 | 1.42e+04 | 1.349 | 0.177 | -8687.691 | 4.7e+04 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Omnibus:** | 1405.509 | **Durbin-Watson:** | | 2.015 | |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | | 2188.650 | |
| **Skew:** | 0.478 | **Prob(JB):** | | 0.00 | |
| **Kurtosis:** | 4.089 | **Cond. No.** | | 139. | |
|  | **variables** | | **VIF** | | |
| **18** | International\_degree\_any | | 1.173745 | |
| **13** | Inhand\_Offer | | 1.846753 | |
| **17** | Certifications | | 1.849656 | |
| **1** | Total\_Experience\_in\_field\_applied | | 3.696798 | |
| **14** | Last\_Appraisal\_Rating | | 5.880166 | |
| **16** | Number\_of\_Publications | | 5.972427 | |
| **2** | Department | | 6.462969 | |
| **15** | No\_Of\_Companies\_worked | | 6.613229 | |
| **0** | Total\_Experience | | 7.871811 | |
| **4** | Designation | | 8.917718 | |
| **10** | University\_PHD | | 10.286466 | |
| **9** | PHD\_Specialization | | 11.677702 | |
| **12** | Preferred\_location | | 11.899397 | |
| **11** | Curent\_Location | | 12.436950 | |
| **5** | Graduation\_Specialization | | 18.865496 | |
| **3** | Role | | 20.030629 | |
| **7** | PG\_Specialization | | 26.394512 | |
| **6** | University\_Grad | | 38.457538 | |
| **8** | University\_PG | | 42.018922 | |

Dropping Designation as it has high missing value present and have multicolinearity

OLS Regression Results for formula:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Expected\_CTC | **R-squared:** | 0.740 |
| **Model:** | OLS | **Adj. R-squared:** | 0.740 |
| **Method:** | Least Squares | **F-statistic:** | 3958. |
| **Date:** | Sun, 11 Sep 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:14:02 | **Log-Likelihood:** | -3.6772e+05 |
| **No. Observations:** | 25000 | **AIC:** | 7.355e+05 |
| **Df Residuals:** | 24981 | **BIC:** | 7.356e+05 |
| **Df Model:** | 18 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 9.537e+04 | 2.77e+04 | 3.444 | 0.001 | 4.11e+04 | 1.5e+05 |
| **Total\_Experience** | 1.216e+05 | 721.154 | 168.680 | 0.000 | 1.2e+05 | 1.23e+05 |
| **Total\_Experience\_in\_field\_applied** | 808.4574 | 841.680 | 0.961 | 0.337 | -841.286 | 2458.201 |
| **Department** | 4.461e+04 | 5276.124 | 8.455 | 0.000 | 3.43e+04 | 5.5e+04 |
| **Role** | -3076.6700 | 8014.506 | -0.384 | 0.701 | -1.88e+04 | 1.26e+04 |
| **Graduation\_Specialization** | 5.611e+04 | 8933.059 | 6.281 | 0.000 | 3.86e+04 | 7.36e+04 |
| **University\_Grad** | 1.572e+05 | 1.09e+04 | 14.372 | 0.000 | 1.36e+05 | 1.79e+05 |
| **PG\_Specialization** | 9.674e+04 | 1.14e+04 | 8.480 | 0.000 | 7.44e+04 | 1.19e+05 |
| **University\_PG** | -8.826e+04 | 1.19e+04 | -7.427 | 0.000 | -1.12e+05 | -6.5e+04 |
| **PHD\_Specialization** | 1.094e+04 | 8784.285 | 1.246 | 0.213 | -6272.733 | 2.82e+04 |
| **University\_PHD** | 1.924e+04 | 6696.543 | 2.874 | 0.004 | 6118.396 | 3.24e+04 |
| **Curent\_Location** | 4695.1448 | 6154.449 | 0.763 | 0.446 | -7367.938 | 1.68e+04 |
| **Preferred\_location** | -920.1205 | 6111.232 | -0.151 | 0.880 | -1.29e+04 | 1.11e+04 |
| **Inhand\_Offer** | 6.771e+04 | 9198.931 | 7.361 | 0.000 | 4.97e+04 | 8.57e+04 |
| **Last\_Appraisal\_Rating** | 7.373e+04 | 2864.982 | 25.734 | 0.000 | 6.81e+04 | 7.93e+04 |
| **No\_Of\_Companies\_worked** | 5379.7394 | 2484.828 | 2.165 | 0.030 | 509.330 | 1.03e+04 |
| **Number\_of\_Publications** | -6414.6201 | 1957.175 | -3.277 | 0.001 | -1.03e+04 | -2578.441 |
| **Certifications** | -7.455e+04 | 3592.230 | -20.754 | 0.000 | -8.16e+04 | -6.75e+04 |
| **International\_degree\_any** | 2.104e+04 | 1.42e+04 | 1.481 | 0.139 | -6810.670 | 4.89e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1406.667 | **Durbin-Watson:** | 2.016 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2190.476 |
| **Skew:** | 0.478 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 4.090 | **Cond. No.** | 137. |

**KNN model for imputing missing value of Department and Role**

**Department:**

**Usage of cross\_val\_score for computing cross-validation scores**

CV accuracy scores: [0.68556701 0.68556701 0.68492268 0.68556701 0.68556701 0.68621134

0.68536428 0.68536428 0.68536428 0.68471954]

CV accuracy: 0.685 +/- 0.000

Train Accuracy is : 0.68574374838876

Test Accuracy is : 0.6785445797624418

Imputed 0 by predicting value using model prediction

**Role:**

**Usage of cross\_val\_score for computing cross-validation scores**

CV accuracy scores: [0.83006536 0.83125371 0.83897802 0.82947118 0.82471777 0.83115339

0.83174792 0.82461356 0.83115339 0.83055886]

CV accuracy: 0.830 +/- 0.004

Train Accuracy is : 1.0

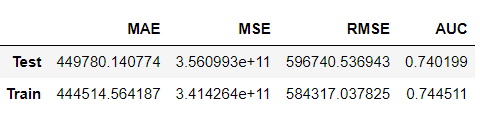
Test Accuracy is : 0.831946755407654

Imputed 0 by predicting value using model prediction

**Linear Regression model For Target variable Expected CTC:**

We now proceed with splitting the data set into train and test. We use the train\_test\_split function with arguments test\_size = 0.3 to get 70:30 split , random\_state = 1 for reproduceable results and stratify = y(target variable) so that target variable has equal proportion of classes in train and test set. The train set now has 17500 rows, 18 features whereas the test set has 7500 rows, 18 features. All the values are now numerical; hence, our data is ready for modelling and testing.

The results after applying Linear Regression on the train and test sets are as follows –



Hence we found model with accuracy of 74% and RMSE score of 596740.536943.

* 1. Actionable insights and recommendations to the stakeholder

1.

1. References and Bibliography

1. https://en.wikipedia.org/wiki/Classification\_of\_Indian\_cities

1. Appendix

KNN For Department

params = {'n\_neighbors': [45,49,51],

'leaf\_size': [10,20],

'p': [1,2],

'weights': ['uniform', 'distance'],

'metric': ['minkowski', 'chebyshev']}

{'leaf\_size': 10, 'metric': 'minkowski', 'n\_neighbors': 49, 'p': 2, 'weights': 'uniform'}

KNN For Role

params = {'n\_neighbors': [41,45,51],

'leaf\_size': [10,20],

'p': [1,2],

'weights': ['uniform', 'distance'],

'metric': ['minkowski', 'chebyshev']}

{'leaf\_size': 10, 'metric': 'minkowski', 'n\_neighbors': 45, 'p': 2, 'weights': 'distance'}