CAPSTONE PROJECT

FINAL PROJECT REPORT

SALARY PREDICTION FOR HR DEPARTMENT OF DELTA LTD.

Submitted by: SABITA NAIR PANCHAL

Submitted on: 5th June, 2022

List of Tables:

List of Figures:

Salary Prediction - Delta Ltd.

1. Introduction of the business problem:

Problem statement:

To build a machine learning model that will accurately predict the salary to be offered to new recruits at Delta Ltd. based on various parameters like their education, job roles, current salary, work experience, etc.

Need of the study:

It is a well-known fact that salary prediction, especially in large organisations, is a tedious process. HR & Recruitment personnel often have to base salary predictions on careful and exhaustive scrutiny of past recruitment data, or on carefully studies salary benchmarks benchmarks prevalent in the job market. Their predictions are especially crucial to the organisation, since it may prove to be a deciding factor between the continuity and discontinuity of service of a new recruit.

A machine learning algorithm to predict salary could address the following organisational needs effectively:

- Make salary offerings more efficient and transparent by having a standard formula for predictions, rather than speculating and anticipating.
- Eradicate discrimination in the salary prediction procedure - candidates with similar credentials and background get similar pay.
- To establish empirically robust and feasible salary benchmarks within the organisation, making the recruitment process itself more reliable.

Business opportunity involved:

The benefits of a robust salary prediction algorithm are numerous, and the resulting business impact on Delta Ltd. will be immense. Some of the benefits and their business implications are enlisted below:

- Substantial time and effort saved of the HR/Recruitment personnel, resulting in valuable man-hours and working hours for the company. This leads to considerable improvement in operational efficiencies for the organisation.
- Creation of a uniform system of salary prediction for new employees that draws from database of previous recruitments, in order to minimise human error/judgement. This will translate to system improvements, making the recruitment process itself more transparent and objective.
- Eliminates discrimination and probability of bias from the process of salary prediction, resulting in creation of goodwill for the company among new and existing employees, which is a very valuable non-financial capital for the company.

2. Exploratory data analysis (EDA):

Understanding the data:

This dataset (*'expected_ctc.xlxs'*) is the collection of recruitment data from various branches of Delta Ltd.. Consult 'APPENDIX - A' for a snapshot of the dataset. It contains names and details of candidates who had previously applied for jobs at Delta Ltd.

Data structuring:

It contains approximately 7,25,000 entries, organised in the following way:

No. of rows	25,000
No. of columns	29

Attribute / variable distribution:

The variables that make the dataset are a mix of numerical and categorical, originally bifurcated as given below. Consult 'APPENDIX - B' for the bifurcation table of the actual variables.

No. of categorical variables	16
No. of numerical variables	13

Target variable identification:

The target variable is *'Expected_CTC'* - which is the variable to be predicted using the supervised learning algorithm that will be built later in the project.

Data hygiene checks:

- The *'isna().sum()*' function was used to identify variables with null/missing values in them. There were numerous missing values, mostly corresponding to candidates who had zero years of experience.

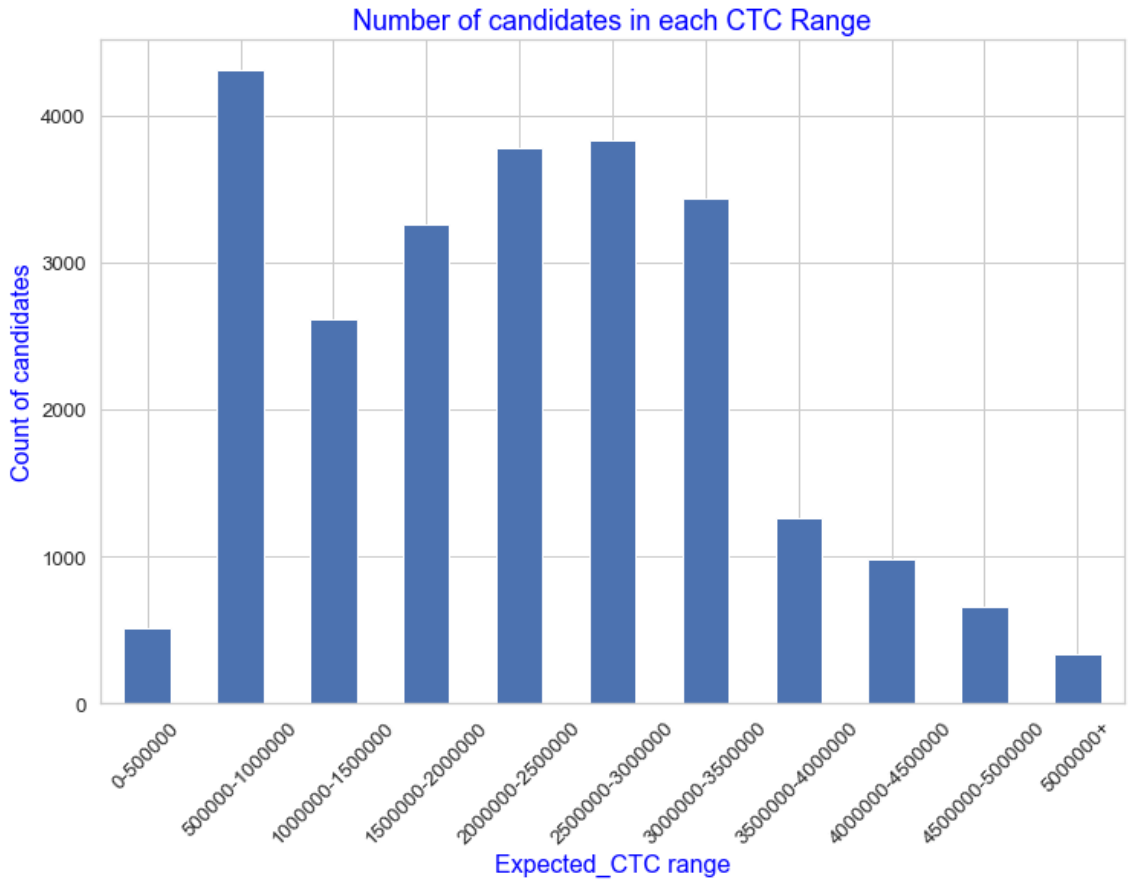
- On using the function *'.duplicate()*', one could see that there are no duplicate rows in the dataset.

Target variable analysis:

- The distribution of target variable 'Expected_CTC' was checked using a histogram, which derived a right-skewed distribution. Log transformation of 'Expected_CTC' was done to try and normalise the distribution. Check 'APPENDIX - C' for the normal and log-transformed histograms.

- Using *'pd.cut()*' function, the number of candidates in 11 Expected_CTC ranges was derived, as is shown in the following figure:

Figure: 2.1



The statistical summary of target variable, and its percentile distribution are given in the following tables:

Table: 2.1

5% candidates have an expected_ctc lower than 577473.50
10% candidates have an expected_ctc lower than 681085.10
25% candidates have an expected_ctc lower than 1306277.50
50% candidates have an expected_ctc lower than 2252136.50
75% candidates have an expected_ctc lower than 3051353.75
90% candidates have an expected_ctc lower than 3796165.40
95% candidates have an expected_ctc lower than 4360145.10

Table: 2.2

count	25,000
mean	22,50,155
std	11,60,480
min	2,03,744
25%	13,06,278
50%	22,52,136
75%	30,51,354
max	55,99,570

Business implications:

- Average salary expectation is approx. Rs. 22,52,136
- While the maximum expected salary is approx. Rs. 56,00,000, there are only about 5% candidates (1250) expecting more than Rs. 43.60.000.
- The minimum / starting salary offered by Delta Ltd. is just over Rs. 2,00,000.
- Only about 5% candidates get salaries higher than Rs. 43,00,000.
- About 3/4 of the candidates have salaries less than Rs. 30,50,000.

For detailed analysis, the dataset was divided into two sets - Numerical (continuous variables) and Categorical (discrete variables), as follows:

Variable type	Column names
Numerical	Total_Experience, Total_Experience_in_field_applied, Current_CTC, No_Of_Companies_worked, Number_of_Publications, Certifications, International_degree_any, Expected_CTC
Categorical	Department, Role, Industry, Organization, Designation, Education, Graduation_Specialization, University_Grad, PG_Specialization, University_Grad, PG_Specialization, University_PG, PHD_Specialization, University_PHD, Current_Location, Preferred_Location, Inhand_Offer, Last_Appraisal_Rating, Passing_Year_Of_Graduation, Passing_Year_Of_PG, Passing_Year_Of_PHD

Univariate analysis - Numerical variables:

- Data hygiene was ascertained by checking unique values of continuous variables, to rule out presence of any unwanted characters/strings in the data.
- Missing value check returned that there were no missing values in the continuous variables.
- The describe() function returned the following statistical summary:

Table: 2.3

	count	mean	std	min	25%	50%	75%	max
Total_Experience	25,000	1.25E+01	7.47E+00	0	6	12	19	25
Total_Experience_in_field_applied	25,000	6.26E+00	5.82E+00	0	1	5	10	25
No_Of_Companies_worked	25,000	3.48E+00	1.69E+00	0	2	3	5	6
Number_of_Publications	25,000	4.09E+00	2.61E+00	0	2	4	6	8
Certifications	25,000	7.74E-01	1.20E+00	0	0	0	1	5
International_degree_any	25,000	8.17E-02	2.74E-01	0	0	0	0	1
Current_CTC	25,000	1.76E+06	9.20E+05	0	1027311.5	1802567.5	2443883.25	3999693
Expected_CTC	25,000	2.25E+06	1.16E+06	203744	1306277.5	2252136.5	3051353.75	5599570

- Furthermore, histograms of the continuous variables were plotted to check their distribution and spread.
- From the histograms, it was evident that variables 'No_Of_Companies_worked', 'Number_of_Publications', 'Certifications' and 'International_degree_any' are discrete in nature, so they were converted to 'object' datatype.
- Outlier check was done using box plots - few outliers were seen in 'Total Experience_in_field_applied'.
- Skewness of the continuous variables was checked using *'skew()'* function, with results as shown below:

Table: 2.4

Variable	Skewness
Total_Experience	0.004109
Total_Experience_in_field_applied	0.951124
Current_CTC	0.097643
Expected_CTC	0.331972

- High skewness was found in 'Total_Experience_in_field_applied'. Rectification of skewness was tried using log and square root transformation, none of which succeeded in reducing the skewness substantially. Check 'APPENDIX - D' for the related charts.

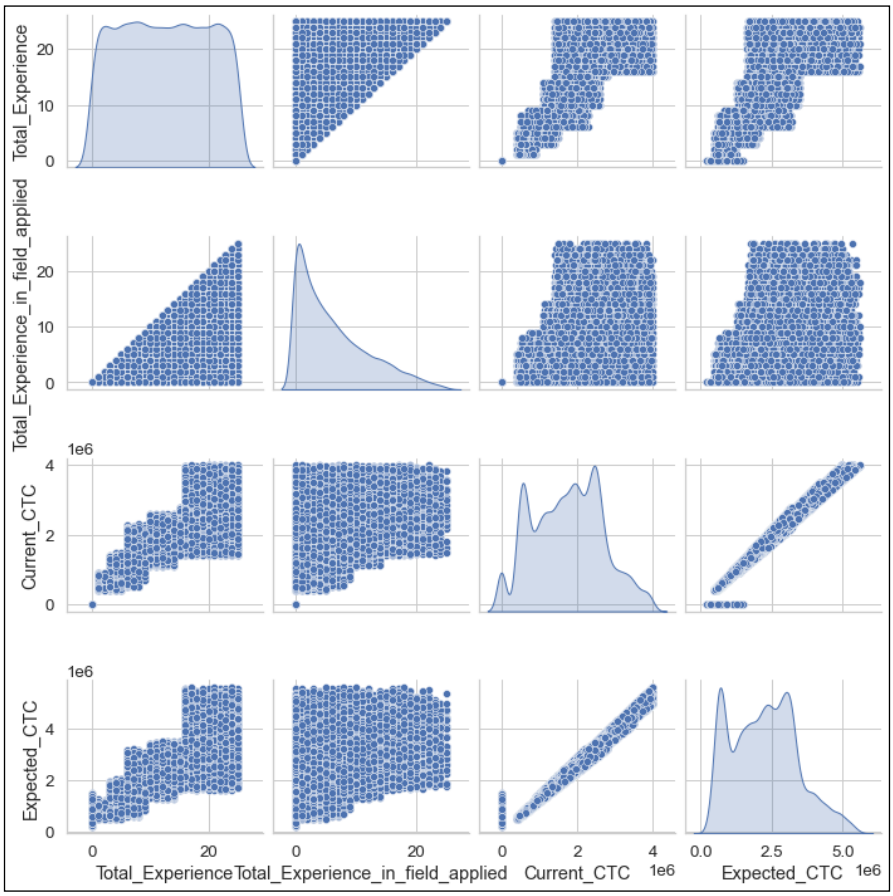
Business implications:

- Candidates' experience ranges from 0 to 25 years, both in terms of total work experience, as well as in terms of role specific experience.
- While fresh candidates (without a current salary) are getting a salary of over Rs. 2,00,000, the leap in expectation rises multifold with increase in current salary. For a candidate who is already earning around Rs. 40,00,000, the expected salary is approx. Rs. 56,00,000 - which is almost a 40% hike in salary!

Bivariate analysis - Numerical variables:

- Histograms were used to check the distribution of the continuous variables - check 'APPENDIX - E' for the plots.
- A pairplot was used to visualise the relation between all continuous variables, as given below:

Figure: 2.2



- Using the corr() function and a heat-map, the correlation of these variables was checked, which derived the following results:

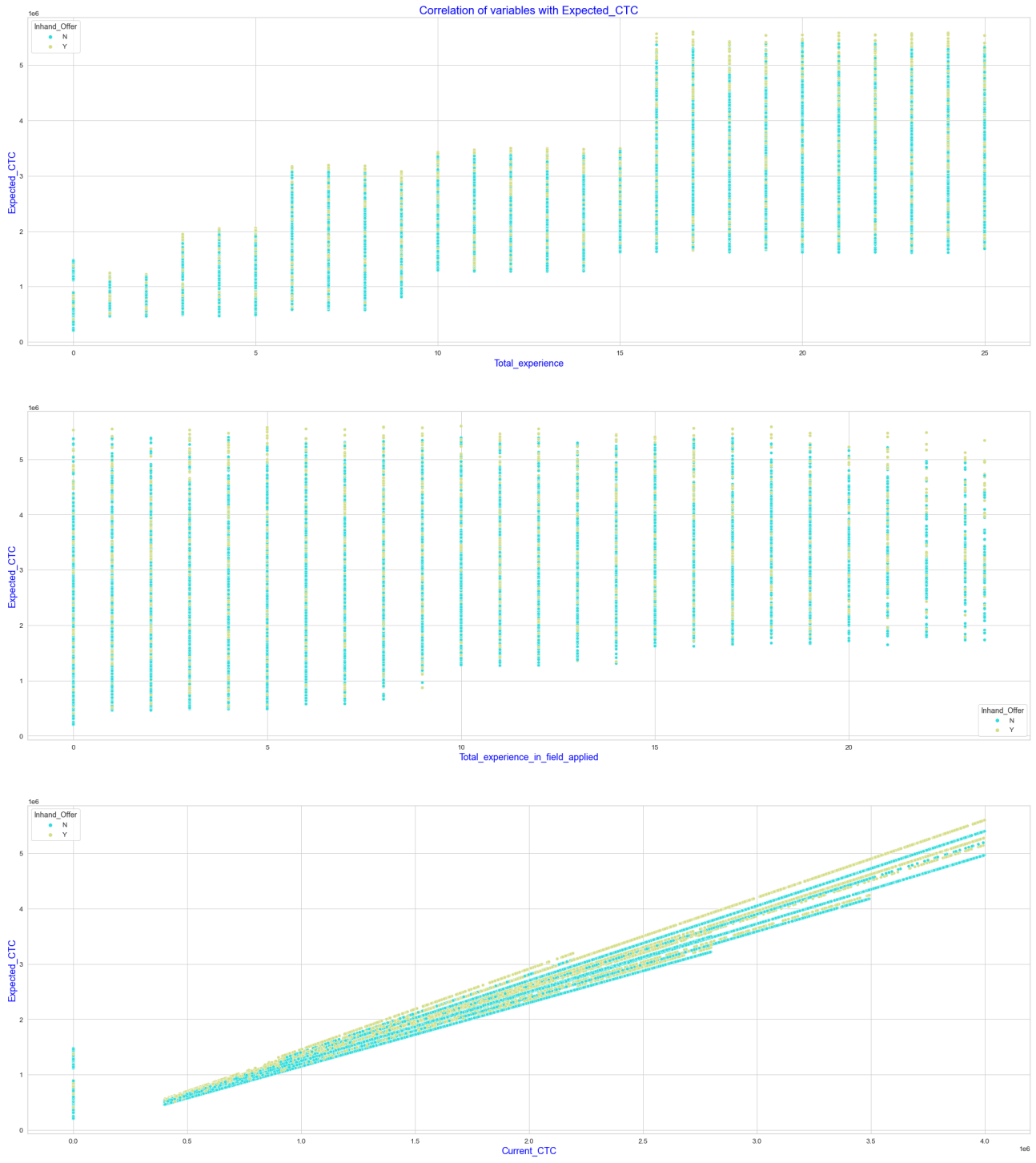
Table: 2.5

	Total_Experience	Total_Experience_in_field_applied	Current_CTC	Expected_CTC
Total_Experience	1	0.645135	0.846476	0.816593
Total_Experience_in_field_applied	0.645135	1	0.548017	0.529115
Current_CTC	0.846476	0.548017	1	0.986718
Expected_CTC	0.816593	0.529115	0.986718	1

- Very high correlation is evident between 'Expected_CTC' and 'Current_CTC', meaning higher the current salary, greater will be the predicted salary. Refer 'APPENDIX - F' for the heat-map / correlation matrix.

- Scatterplots were plotted to check the variance of the continuous variables against the target variable 'Expected_CTC', as seen in the following figure:

Figure: 2.3



- Apart from ‘Current_CTC’, ‘Total_Experience’ also impacts the salary expectation. There seems to be a clear increase in expected salary as the number of years of total work experience increases.

EDA - Categorical variables:

Univariate analysis - Categorical variables:

- Countplots were created to check the distribution of the various categorical variables (examples are given below). Check 'APPENDIX - G' for the countplots of other categorical variables.

Figure: 2.4

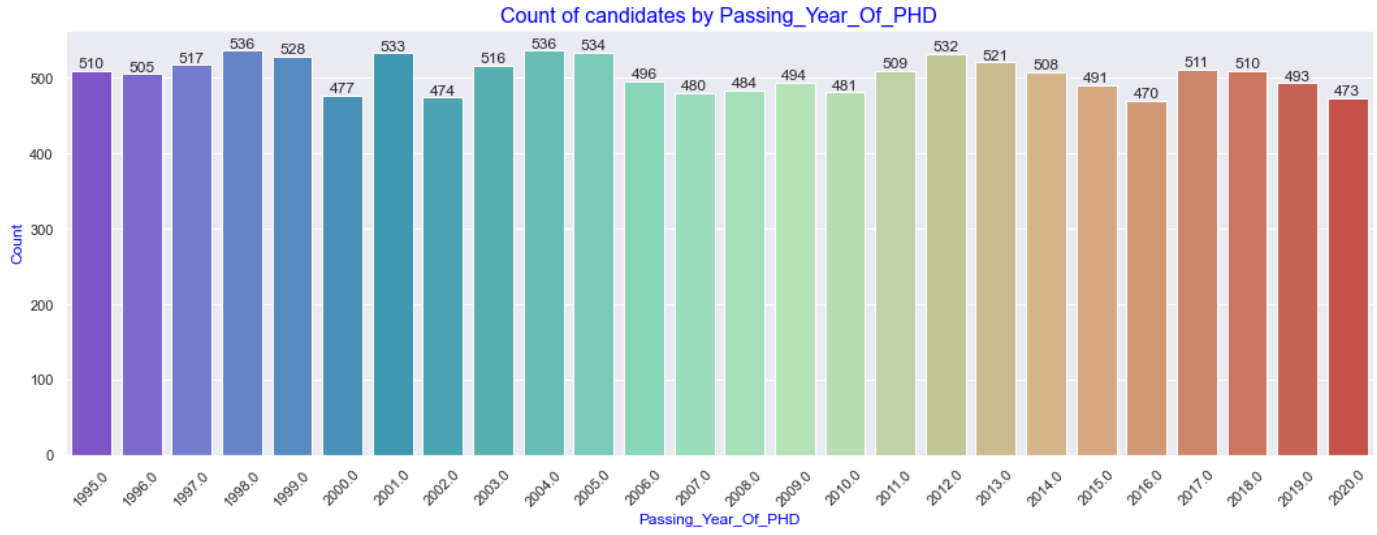
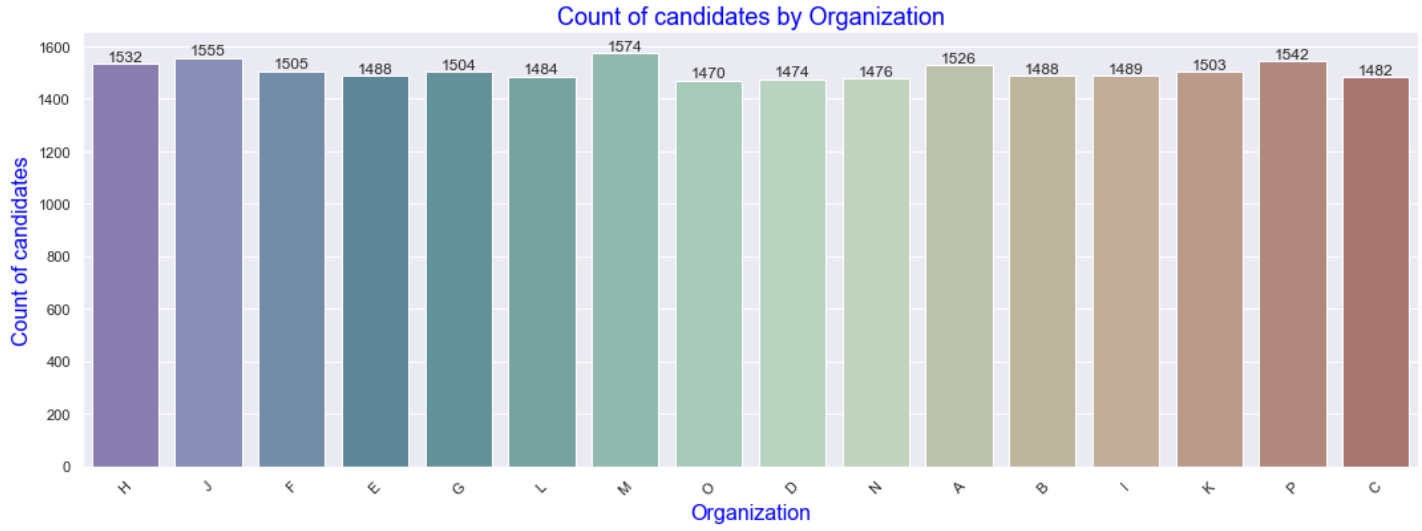


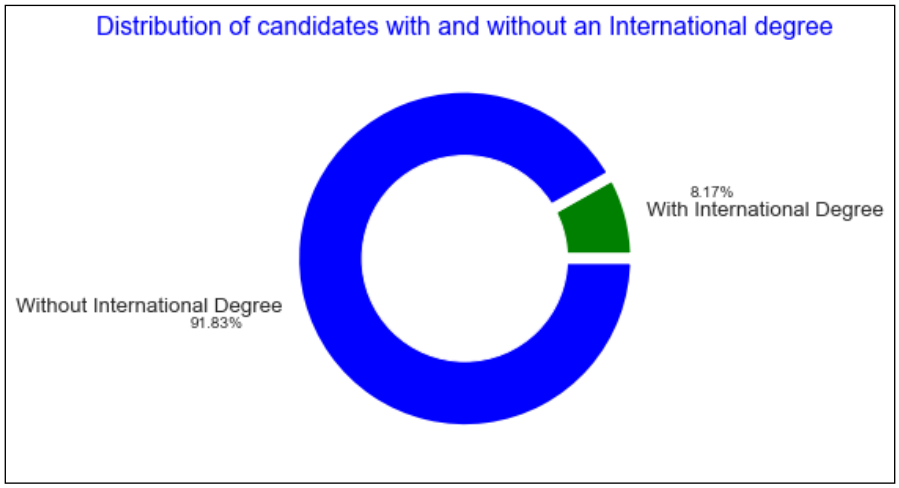
Figure: 2.5



- Features such as "Department", "Organisation", "Education" and "Passing Year Of PHD" showed more or less equal distribution, while other variables showed unequal distribution.

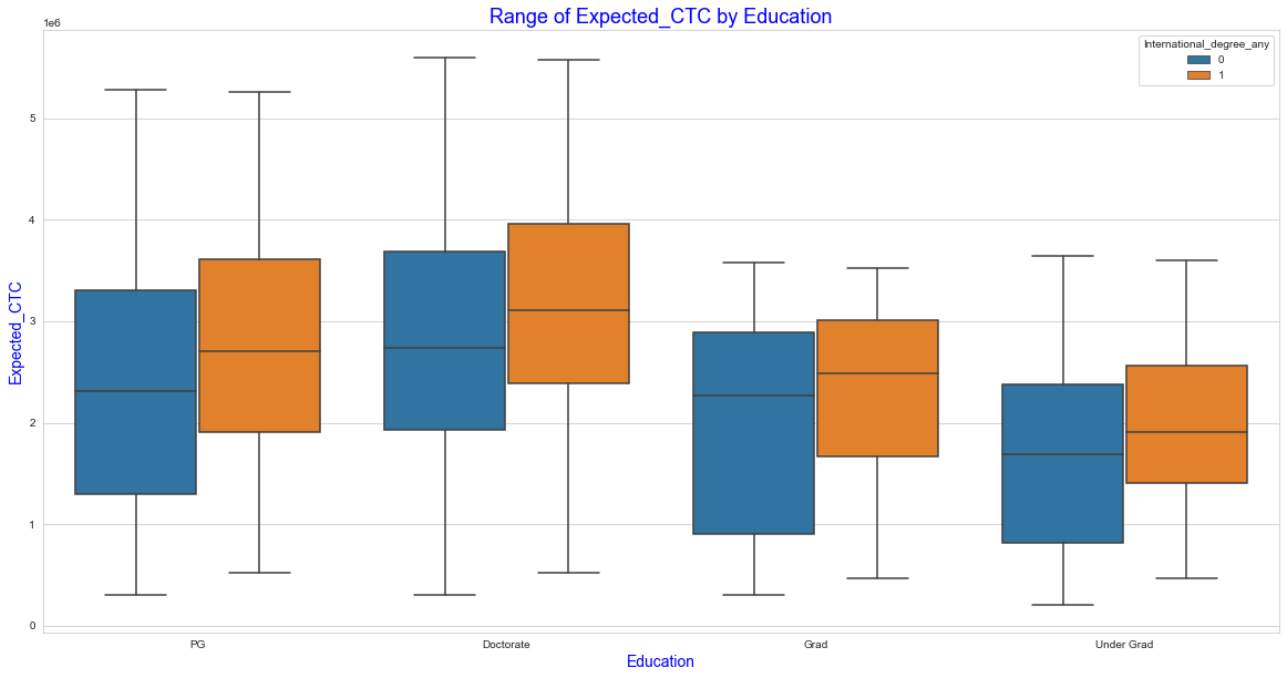
- Another feature is the presence of an international degree - a very minute section of candidates (8.17%) have an international degree, as shown in the pie chart below:

Figure: 2.6



Bivariate analysis - Categorical variables:

Figure: 2.7

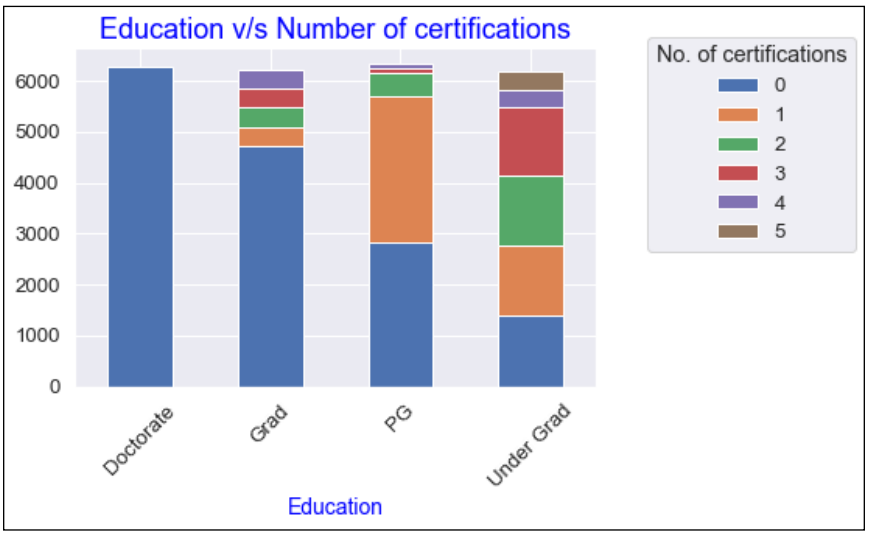


- Boxplots of education level against expected salary shows that Doctorate level candidates enjoy higher salaries, followed by the post-graduates, while under-graduates and graduates have mid-range salary expectations.

- In the above figure, impact of having an international degree is seemingly more, i.e. - candidates with international degrees definitely earn more.

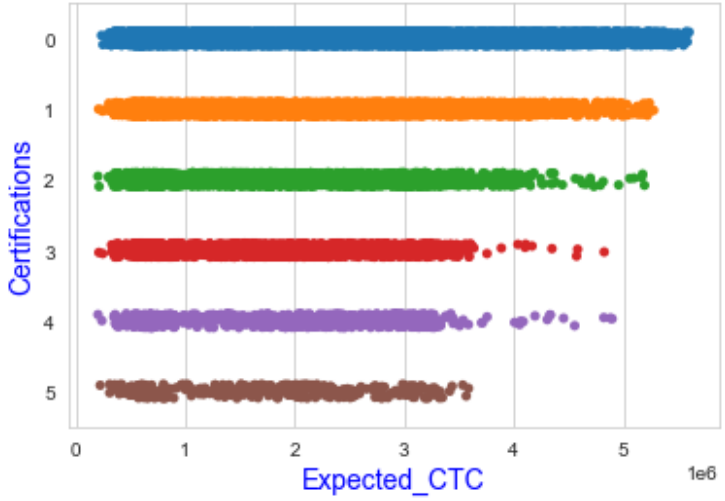
- Stacked bar chart of educational qualifications with number of certifications (given below), clearly shows that candidates holding highest qualification (PHD/Doctorate) have no certifications. While candidates with the lowest credentials (under-grads) have the most number of certifications. Thus, education seems to be inversely proportional to the number of certifications.

Figure: 2.8



- Strip plots of independent categorical variables against target variable showed interesting correlations. For example, as evident in below plot, candidates with more certifications have lesser salaries, which those without certifications have recorded the highest salaries.

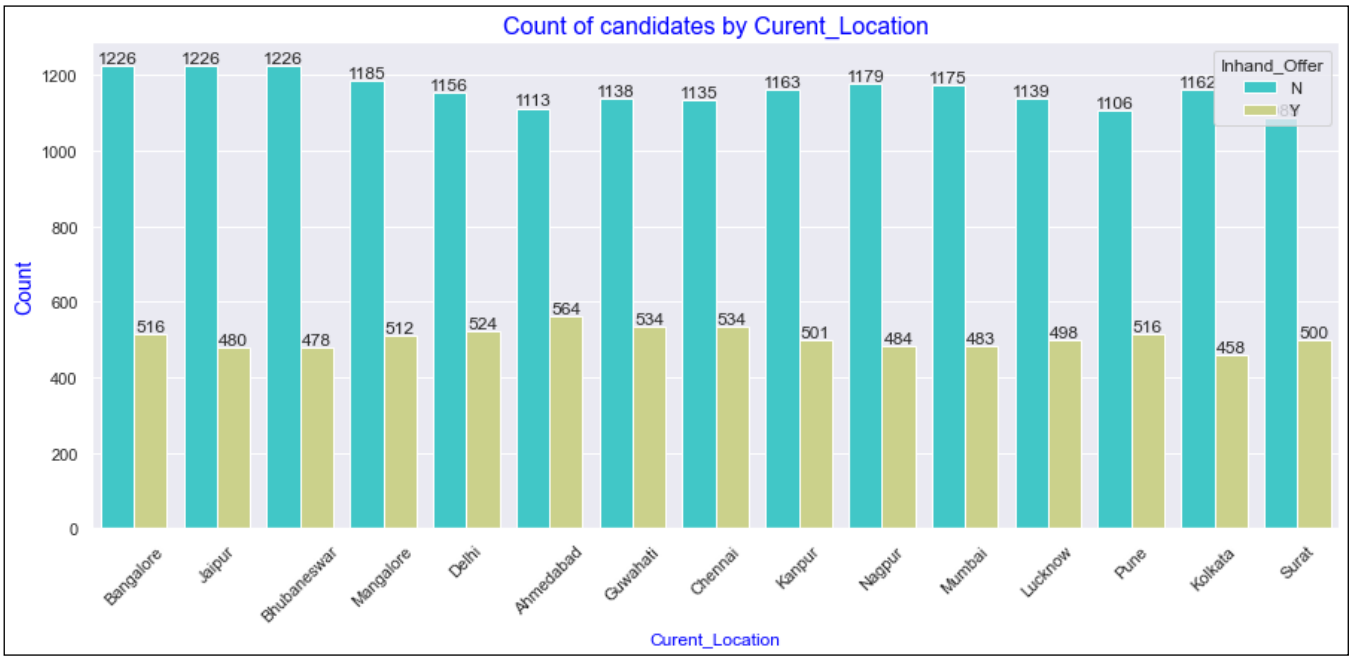
Figure: 2.9

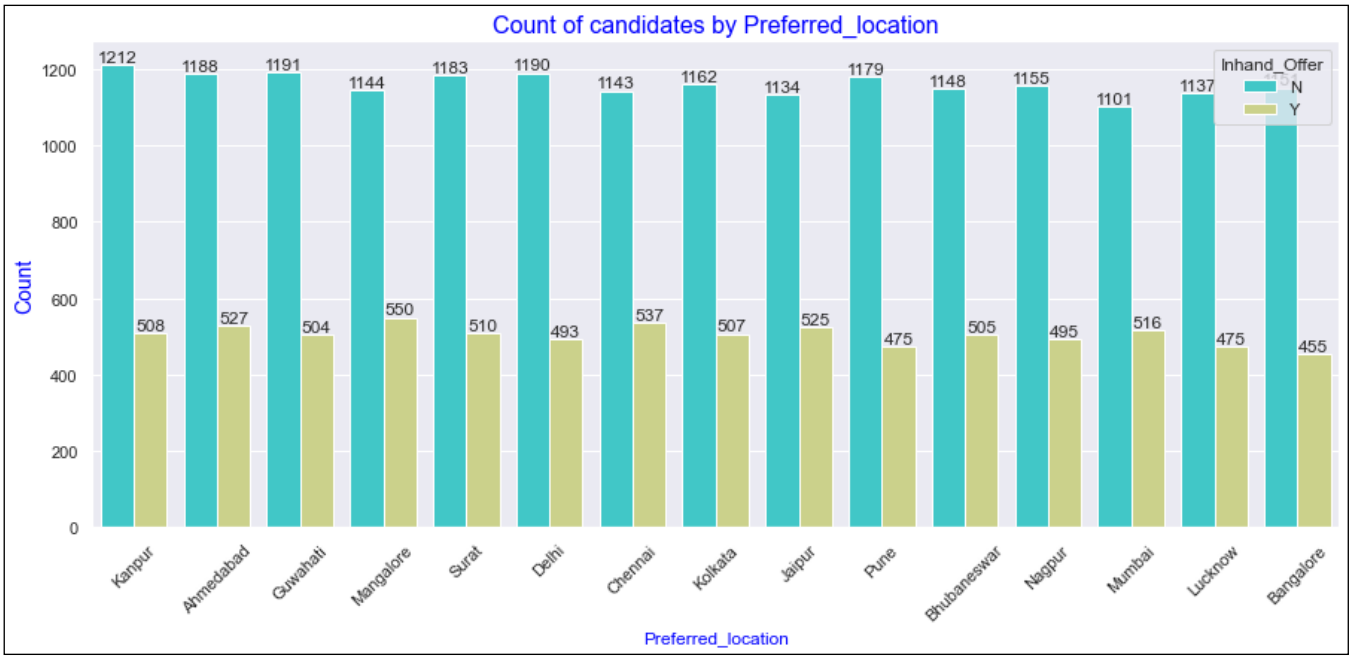


- Similarly, although candidates having recently completed Graduation and PHD earn less salaries, but in case of Post-graduates, there is an increase in salary, I.e., pass-outs from 2015 onwards earn more salaries than experienced people.

Multi-variate analysis of variables:

Figure: 2.10





- Above plot shows that only less than half the candidates seem to already have an offer of employment at hand.
- While current work location of majority of candidates is Bangalore, the most preferred work location is Kanpur.

- Figure 3.2.8 (next page) shows that majority of candidates receive appraisal rating D in the previous appraisal. The highest salaries go to the Key Performers, followed by appraisal rating A, B and C.

- Figure 3.2.9 demonstrates that the candidates with doctorates earn highest salaries. But, one must note that Under-graduates with 15-25 years of experience earn more than Graduates, Post-Graduates and PHD holders who have between 10-15 years of experience.

Figure: 2.11

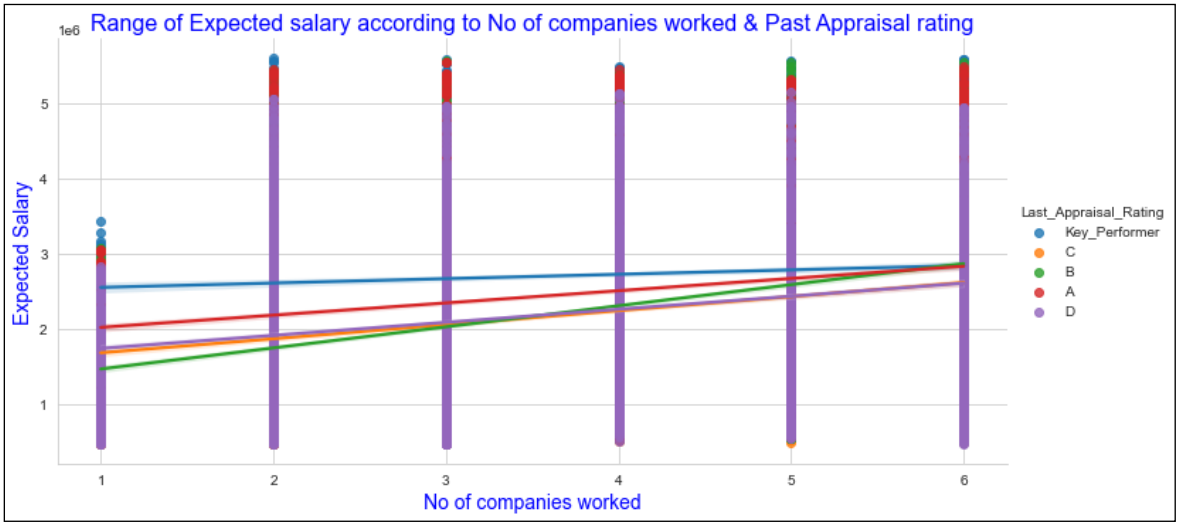


Figure: 2.12

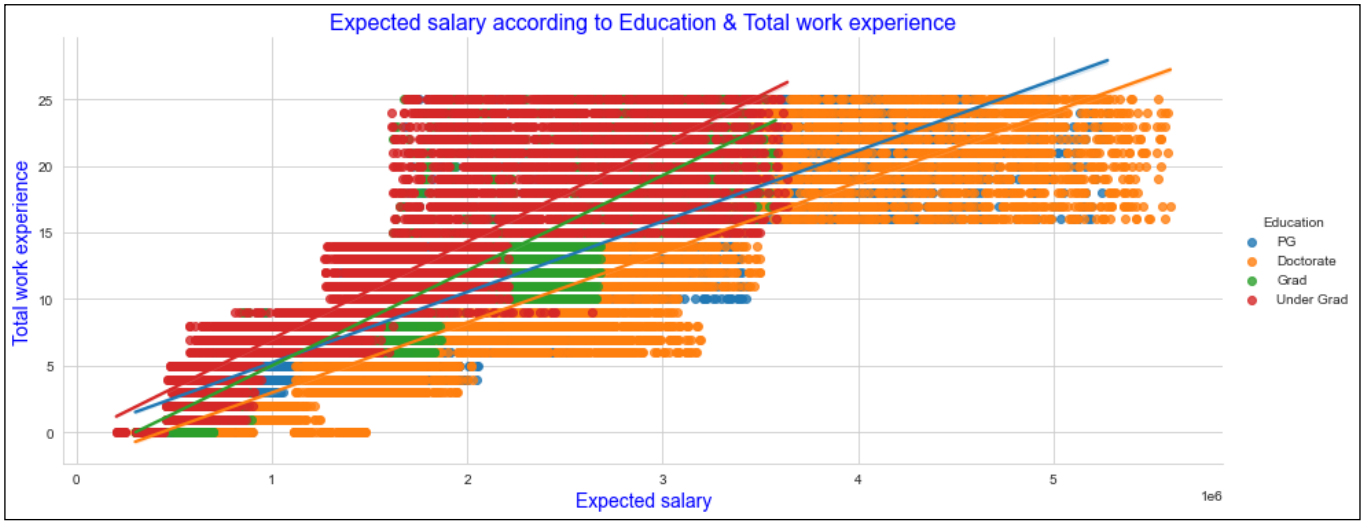
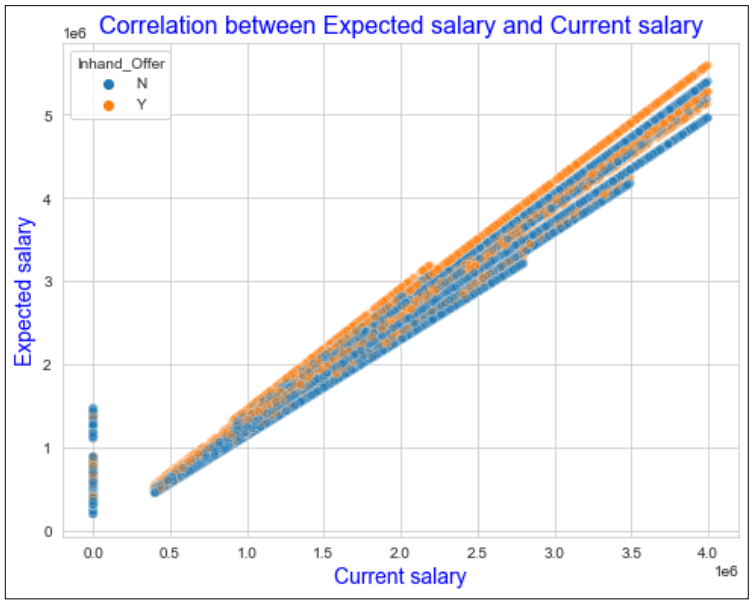


Figure: 2.13



- From Figure 2.13, one can infer that current salary is directly proportional to expected salary, and that a person who has another offer of employment has more earning potential.

- Profile of the candidates with the highest and lowest expected salary:

Table: 2.6

	Role	Designation	Education	Last_Appraisal_Rating	Total_Experience	Inhand_Offer	No_Of_Companies_worked
Candidate with highest expected salary	Financial Analyst	Marketing Manager	Doctorate	Key_Performer	17	Y	2
Candidate with lowest expected salary	NaN	NaN	Under Grad	NaN	0	N	0

Overall Business implications derived from EDA:

- While fresh candidates (without a current salary) are getting a salary of over Rs. 2,00,000, the leap in expectation rises multifold with increase in current salary. For a candidate who is already earning around Rs. 40,00,000, the expected salary is approx. Rs. 56,00,000 - which is almost a 40% hike in salary!
- Higher a candidate's current salary, higher salary expectation he/she will have.
- Education plays a major role in deciding salary - higher the education credential, higher will be the salary expectation.
- With work experience also, one's salary expectation will increase.
- Candidates with an existing employment offer in-hand will negotiate for higher salaries.
- Last appraisal ratings will also be decisive in salary expectation - 'key performers' are very few in the sample, but they take home the highest salaries.
- For Graduates and PHD holders, there is a linear relationship between year of passing out and expected salary - earlier the passing-out year, higher the salary. (Check 'APPENDIX - H' for point-plots of Year-of-passing versus 'Expected_CTC').
- However, for Post-graduates, there is a pattern - older pass-outs do have the highest salaries, but there is a surge in salary for those who completed PG post 2015.
- Certain variables like 'Department', 'Role', 'Organization', 'Designation', etc. will have a very minor impact on the expected salary.
- Candidates who have an international degree will be bound to ask for more salary. (Check 'APPENDIX - I').

3. Data cleaning & Pre-processing:

The following feature transformations were done in preparation for the model building exercises to follow:

Removal of unwanted variables: Columns 'IDX' and 'Applicant_ID' were removed as they are just unique identifiers that will not add any value to the model.

Missing value treatment:

- Most missing values in columns "Industry", "Last Appraisal Rating", "Organization", "Department", "Role" & "Designation" were correspondent to candidates with total work experience of 0 years. Hence, those values were replaced with 'None', meaning - no experience.

- In columns 'Graduation Specialization', 'University Grad', 'Passing Year Of Graduation', 'PG Specialization', 'University PG', 'Passing Year Of PG', 'PHD Specialization', 'University PHD' & 'Passing_Year_Of_PHD', the NaN values were correspondent to those candidates who did not have the relevant qualification - such NaN values were replaced with 'Not Applicable'.

Outlier treatment: Outliers in the variable 'Total_Experience_in_field_applied' were imputed using the IQR method. Refer 'APPENDIX - J' to see box-plots before and after outlier treatment.

Variable tranformation:

- *'Inhand_Offer'* values (Y,N) were converted to boolean values (1,0).

- Nominal encoding was performed on the variable *'Last_Appraisal_Rating'*, and the alphabetic codes were replaced with random nominal codes.

Addition of new variable: There seem to be errors/discrepancies in the '*Education*' column, since it was not capturing many of the qualifications of candidates, even though they had passed the said qualifications. Assuming that '*Education*' had errors in it, a new column *'Edu_Qualification*' was created using the *'select (conditions, values)'* function in *numpy* to capture the real scenario. *'Education'* was later deleted.

Refer to 'APPENDIX - K' to understand the difference in deleted variable *'Education'* and the newly created variable *'Edu_Qualification'.*

Data preprocessing in readiness for model-building:

The following steps were taken in order to get the data ready for model building:

- Ordinal encoding was done for the variable 'Edu_Qualification' since EDA had shown a distinct linear, progressing relationship between educational qualification and expected salary - higher the educational credentials, higher the expected CTC, hence there is an inherent order/rank to values. The coding is shown below:

Table: 3.1

Edu_Qualification	PHD	Post_Grad	Grad	Under_Grad
Code	3	2	1	0

- Before proceeding to dummy encoding of categorical variables, there was the necessity to reduce dimensionality for error-free performance of the model.

- For example, city names across the columns "Current Location", "Preferred location", "University PHD", "University PG", "University Grad", were categorised to Tier-1 and Tier-2, depending on their tier status. The following segregation was done:

Table: 3.2

Tier	Cities
Tier-1	Bangalore, Chennai, Hyderabad, Mumbai, Kolkata, Delhi
Tier-2	Mangalore, Jaipur, Bhubaneswar, Ahmedabad, Guwahati, Kanpur, Nagpur, Lucknow, Pune, Surat

- In terms of 'Passing Year Of Graduation', 'Passing Year Of PG & 'Passing Year Of PHD', these 3 variables contained altogether 97 unique values, which on encoding would have added 95+ columns to the dataset, increasing the dimensionality. Thus, the year values were cut to bins of 5 years each, like 1980-1985, 1985-1990, 1990-1995, 1995-2000, 2000-2005, 2005-2010, 2010-2015, 2015-2020, 2020-2025, in order to cut down on the dimensions.

- Similarly, for variables 'Graduation Specialization', 'PG Specialization' & 'PHD Specialization', there were about 13-15 subjects in each column. These would have increased multifold the number of columns post dummy encoding. Hence, subjects like 'Chemistry', 'Zoology' & 'Botany' were clubbed into 'Pure_sciences; 'Arts', 'Psychology' & 'Sociology' were clubbed into 'Arts_Humanities'; and 'Mathematics'& 'Statistics' were clubbed into 'Maths_Stats'. Hence each of the above mentioned 3 columns ended up having just 7 categories, instead of 13/15.

- Dummy encoding was applied on the following categorical variables with drop_first() function to ensure that dimensionality is limited - 'Department', 'Role', 'Industry', 'Organization', 'Designation', 'Last_Appraisal_Rating', 'Graduation Specialization', 'University Grad', 'PG Specialization', 'University PG', 'PHD Specialization', 'University PHD', 'Curent Location', 'Preferred location', 'Year Graduation bin', 'Year PG bin', 'Year PHD bin'.

- Separate vectors 'x' (containing all independent variables) & 'y' (containing only the dependent variable 'Expected_CTC') were created to enable the splitting of data into training and test sets.

- Thereafter, using the train_test_split() function of sklearn.model_selection, the data was split into training set and test set in the ratio 70:30 respectively. Post splitting, the dimensions of the sets were as follows:

Table: 3.3

Set	Dimensions
x_train	(17500, 142)
x_test	(7500, 142)
y_train	(17500,)
y_test	(7500,)

4. Model Building

In this business problem, we are predicting the values for a continuous variable (Expected CTC) that is already part of the dataset, thus we used models for supervisory predictive analysis. Theoretically, the ideal model for such a business problem is Linear Regression.

However, to empirically test which model is best, a host of other models were also built. The models essentially predicted the Expected CTC by establishing relations between the input and output variables (they adjusted the weights allotted to the input variables until the model was fitted appropriately i.e. a near perfect prediction of output variable is derived). The various models that were built are given below:

1. Linear regression (LR):

It is considered the most appropriate model for supervised predictive analysis for continuous output variables. The model was built using the 'LinearRegression()' function from the 'sklearn,model_selection' package.

Model interpretation: On running the model, coefficients (weights assigned to each input variable) were derived. The coefficient of a variable determines its impact on the target variable. Implications of some of the highest (positive & negative) coefficients is given in the next page, as a sample:

Table: 4.1

Variable / feature	Change in 'Expected_CTC' due to a unit increase in feature
Current_CTC	Increases by 1.35
Edu_qualification	Increases by 11462.08
No_Of_Companies_worked	Decreases by 109.35
Certifications	Decreases by 14655.69

The intercept for this LR model = 83923.89

Based on these results, we can say that the linear equation for the dependent-independent variable relationship is as follows:

Expected_CTC = Intercept (83923.89) + (-8878.50) * Total_Experience + (109.86) * Total Experience in field applied + (1.34) * Current_CTC + (58203.14) * Inhand Offer + ……+ (coeff_142) * last variable

Model evaluation: Evaluation metrics such as R^2 (mean residual square) and RSME (root mean square error) were computed to check the effectiveness of the LR model.

Table: 4.2

LR Metric	Training data	Test data
R^2	0.9929	0.993
RMSE	97582.27	97934.34

The RMSE of both train and test data have a very small difference between them, hinting that train data predictions and the test data predictions are at almost similar distance from the best fit line derived by the LR model.

The R^2 (which represents the model score, or accuracy) is more or less same for training and test data. To try and enhance model performance, several other models were built, which are explained hereafter.

2. Linear regression using Ordinary Least Squares (OLS):

This model was built using the Ordinary Least Squares function from the 'statsmodel' library. This model is needed because sometimes R^2 can often be misleading - increasing with the increase in the number of variables (as is the case here), irrespective of their contribution to the prediction of the target variable. A more dependable metric is the 'Adjusted R^2', computed using the OLS model for Linear regression. Unlike R^2, Adjusted R^2 takes into account only the impact of those independent variables that actually have an effect on the dependent variable, and gives a more realistic idea of the model effectiveness.

Model interpretation:

The OLS model summary snapshot is given below.

Figure: 4.1

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.993
Method:	Least Squares	F-statistic:	1.819e+04
Date:	Sat, 14 May 2022	Prob (F-statistic):	0.00
Time:	11:11:11	Log-Likelihood:	-2.2588e+05
No. Observations:	17500	AIC:	4.520e+05
Df Residuals:	17366	BIC:	4.531e+05
Df Model:	133		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.742e+04	1742.260	9.997	0.000	1.4e+04	2.08e+04
Total_Experience	-8878.5012	270.960	-32.767	0.000	-9409.610	-8347.392
Total_Experience_in_field_applied	109.8575	167.046	0.658	0.511	-217.570	437.285
Current_CTC	1.3451	0.002	715.721	0.000	1.341	1.349
Inhand_Offer	5.82e+04	2051.436	28.372	0.000	5.42e+04	6.22e+04

The metrics of importance here are the following:

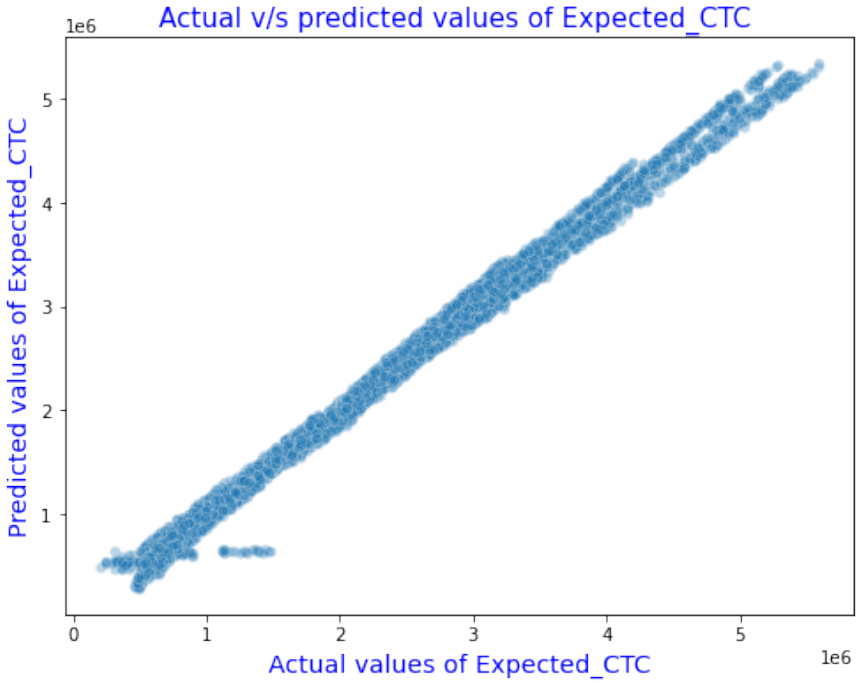
- **R^2** - also known as the coefficient of determination, R^2 tells us how much of the variation in target variable is caused by changes in the independent variables. In this case, our model explains 99.3% of the variation in 'Expected_CTC', which is a very robust figure.
- **Adjusted R^2** - At 99.3%, this measure is equivalent to the R^2, hence it is a reassurance that all our input variables do, in reality, contribute to the variance in out target/dependent variable. An R^2 equivalent to Adjusted R^2 is a sign of model robustness.
- **Probability (F-statistic)** = 0 (< 0.05), meaning that we can reject the null hypothesis that independent variables have no impact on the dependent variable.

In addition, the **Mean Squares Error (MSE)** metric was also calculated, which gives the average variance between the actual target values and the predicted target values.

Here, **MSE = 97934.336** means that there is an average difference of 97,934 Rs. between the real 'Expected_CTC' values and the values predicted by this model, which is not very high considering the scale of our target variable.

A **scatterplot of actual versus predicted values of target variable** gave us the following:

Figure: 4.2



From this scatterplot it is evident that apart from a few instances at the lower spectrum, the predicted values are very close by to the best fit line of linear regression model.

Multicollinearity check: to check multicollinearity, the test VIF (Variation Inflation Factor) was done (a snapshot of the first few lines of the results is given below):

Figure: 4.3

```
Total_Experience  VIF =  7.48
Total_Experience_in_field_applied  VIF =  1.74
Current_CTC  VIF =  5.42
Inhand_Offer  VIF =  1.62
No_Of_Companies_worked  VIF =  1.37
Number_of_Publications  VIF =  2.07
Certifications  VIF =  1.48
International_degree_any  VIF =  1.88

<ipython-input-68-006ae825bd87>:8: RuntimeWarning: di
  vif=round(1/(1-rsq),2)

Edu_qualification  VIF =  inf
Department_Analytics_BI  VIF =  3.03
Department_Banking  VIF =  2.68
Department_Education  VIF =  2.68
Department_Engineering  VIF =  2.62
Department_HR  VIF =  2.68
Department_Healthcare  VIF =  2.82
Department_IT_Software  VIF =  1.94
```

Here, most values are contained within 5, which does not indicate very high levels of collinearity - meaning that the independent variables are not affected by a high level of correlation with other independent variables. Only concerning variable is '**Total Experience' with VIF of 7.48.**

In order to test whether the multicollinearity of 'Total_Experience' is impacting the robustness of the model, this feature was removed and the OLS regression summary was performed again to check if evaluation metrics show any significant change. The new OLS summary snapshot is given below:

Figure: 4.4

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.992
Model:	OLS	Adj. R-squared:	0.992
Method:	Least Squares	F-statistic:	1.726e+04
Date:	Sat, 14 May 2022	Prob (F-statistic):	0.00
Time:	13:45:08	Log-Likelihood:	-2.2640e+05
No. Observations:	17500	AIC:	4.531e+05
Df Residuals:	17367	BIC:	4.541e+05
Df Model:	132		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4277.3370	1747.050	2.448	0.014	852.942	7701.732
Total_Experience_in_field_applied	-1585.4549	163.663	-9.687	0.000	-1906.252	-1264.658
Current_CTC	1.3091	0.002	832.974	0.000	1.306	1.312
Inhand_Offer	5.808e+04	2113.836	27.476	0.000	5.39e+04	6.22e+04
No_Of_Companies_worked	-2308.9448	524.106	-4.405	0.000	-3336.246	-1281.644

We can see from the above table that R^2 and Adjusted R^2 have slightly diminished. We can summarise the findings in the table below:

Table: 4.3

OLS Summary Metrics	With 'Total_Exeperience'	Without 'Total_Experience'
R^2	0.993	0.992
Adjusted R^2	0.993	0.992
Prob (F-statistic)	0	0

We can observe that the change in the metrics is highly insignificant (0.001), thus we cannot say decisively that there will be any impact on the model if the variable 'Total_Experience' is removed. So, we will keeps all variables intact for all future models.

2. Decision Tree (Classification & Regression Tree - CART)

A continuous variable decision tree was built using criterion 'Gini'. The codes for the CART model was saved in a .dot file ('tree_regularized.dot'). These codes were used to generate the actual decision tree on the website 'www.webgraphviz.com'.

Further, feature importance of the various features was also calculated to understand the impact of each variable on the target variable. It shows that the variable 'Current_CTC' is the most important independent variable in terms of predicting the Expected salary of an individual. It is followed by the variables 'Last_Appraisal_Rating' (D & C), 'Total Experience', 'Edu_qualification' and so on. A list of the ten most important features, as given by their Gini scores, is given below:

Figure: 4.5

	Imp
Current_CTC	9.878782e-01
Last_Appraisal_Rating_D	4.387552e-03
Last_Appraisal_Rating_C	4.159063e-03
Total_Experience	8.723766e-04
Edu_qualification	6.599851e-04
Inhand_Offer	6.560692e-04
Certifications	3.051328e-04
Year_Graduation_bin_Not_Applicable	1.897096e-04
Graduation_Specialization_Not_Applicable	1.595508e-04
Number_of_Publications	8.353873e-05

The performance metrics of the CART model are as follows:

Table: 4.4

CART Metric	Training data	Test data
R^2	0.999	0.994
RMSE	10608.936	93445.970

3. Artificial Neural Network (ANN):

An Artificial Neural Network model was built using the MLPRegressor, with 300 hidden layers and 5000 iterations. Before building this model, the training data was scaled using StandardScalar(), since the ANN model is sensitive to variations in scale. The performance metrics of this model are given below:

Table: 4.5

ANN Metric	Training data	Test data
R^2	0.996	0.995
RMSE	68880.980	82933.139

4. Random Forest Regressor (RF):

A random forest regression model was built using the RandomForestRegressor(). The following performance metrics were derived:

Table: 4.6

RF Metric	Training data	Test data
R^2	0.999	0.996
RMSE	28114.074	69852.485

Model comparison:

The comparative results of the four models are given below:

Table: 4.7

Model	Training RMSE	Test RMSE	Training accuracy	Test accuracy
Linear Regression	97582.273	97934.336	0.9929	0.9930
Decision Tree Regressor	10608.936	93445.970	0.9990	0.9936
Random Forest Regressor	28114.074	69852.485	0.9994	0.9964
ANN Regressor	68880.980	82933.140	0.9965	0.9950

Inferences:

- The RMSE variation is the least in the Linear Regression model - which indicates that the model predictions are similar to the actual model predictions (the training set predictions and test set predictions are more or less equally closer to the best fit line derived by the model).

- For the other models, the RMSE values are highly varied, with test data RMSE being way higher than training data RMSE - this means that the test predictions are located far away from the best fit line than the training set predictions.

- The R^2 values are more or less similar in all the four models, showing very high (near perfect) accuracy of the models. Since both training and test R^2 values are close, it rules out the possibility of the models overfitting. But it is the LR model where R^2 values are exactly the same, meaning similar accuracy for the training as well as the test data sets.

- Based on these results, one can easily say that the Linear Regression seems to be the best model.

5. Model Tuning & Validation:

Although we have seen very high performance in all the models so far, some model tuning measures were carried out so as to see if model performance can be enhanced in any way.

1. Regularised Linear Regression using Ridge & Lasso regression:

In the previous section, we derived the coefficients for the various variables through the LR model. However, since our dataset has a vey high number of variables and some of our coefficients were very large and varied, it could lead to over-fitting on the training data set. Thus we will use certain techniques for regularisation, which means penalising the features with high coefficients.

Ridge regressor - the Ridge regressor is a shrinkage method that works by bringing down / suppressing the magnitude of the coefficients where they are very high. Many of the coefficients will be reduced to a value close to zero.

Lasso regressor - the Lasso regressor not only reduces the coefficients, but also drops several of the variables by making their coefficients zero. This is done in order to avoid the curse of multidimensionality.

Further, polynomial feature expansion was also tried with these two models to check the performance metrics. The accuracy value for these models, are given below:

Table: 5.1

Model	Training accuracy	Test accuracy
Ridge regressor (alpha=50)	0.9928	0.9930
Lasso regressor (alpha=50)	0.9929	0.9930
Ridge - Poly expansion (alpha=50)	0.9975	0.9905
Lasso - Poly expansion (alpha=50)	0.9974	0.9945
Lasso - Poly expansion (alpha=200)	0.9970	0.9953

2. Grid Search & Cross Validation:

Grid search was employed on the CART model and the Random Forest model, wherein a variety of hyper-parameters were suggested for building the model, and thereafter, a cross-validation loop of 3 was set in order for the model to select the best parameters for building the most appropriate model.

The best hyper-parameters derived for the two models are given below:

Figure: 5.1

CART model best params:
max_depth - 20
min_samples_leaf - 10
min_samples_split - 10

Random Forest best params:
max_depth - 15
max_features - 20
min_samples_leaf - 5
min_samples_split - 20
n_estimators - 300

3. Ensemble models:

In addition to the Random Forest (RF) regressor, various other ensemble models were built to check the accuracies, since ensemble models combine the results of different algorithms to give the best scores.

Gradient Boosting - the Gradient Boosting ensemble technique was used for building a model using the 'GradientBoostingRegressor', with n_estimators=50.

Xtreme Gradient Boosting - Another important ensemble technique 'Xtreme Gradient Boosting' was used to build a model using n_estimators=500 and max_depth=7.

The results of these two models are given below:

Table: 5.2

Model accuracy	Training data	Test data
Gradient Boosting	0.99425	0.99424
Xtreme gradient Boosting	0.99947	0.99636

Model Evaluation:

Given below is a comparative table of model accuracy scores of all the models built.

Table: 5.3

Model	Training accuracy	Test accuracy
Linear Regression	0.9929	0.9930
Decision Tree Regressor	0.9990	0.9936
Random Forest Regressor	0.9994	0.9964
ANN Regressor	0.9965	0.9950
Ridge regressor (alpha=50)	0.9928	0.9930
Lasso regressor (alpha=50)	0.9929	0.9930
Ridge - Poly expansion (alpha=50)	0.9975	0.9905
Lasso - Poly expansion (alpha=50)	0.9974	0.9945
Lasso - Poly expansion (alpha=200)	0.9970	0.9953
Gradient Boosting	0.9943	0.9942
Xtreme gradient Boosting	0.9995	0.9964

We can see that the accuracy for all above models are centered around 99% with only slight variations. However, we can eliminate all those models wherein training accuracy is higher than the test accuracy, since it may be indicative of slight overfitting.

Thus we can shortlist the following models wherein test accuracy is higher than training accuracy:

- Linear regression
- Ridge regressor
- Lasso regressor

These three models have almost similar model scores for both training and test sets. However, for the sake of choosing a final model, we can select Linear Regression, since RMSE scores for Linear regression were the most consistent among other models (refer Table-10). This implies that the LR model predictions for both training as well as test sets are the closest to the best fit line.

6. Business Implications:

- Delta Ltd.'s recruitment data is very efficient as it has given us some very robust models.

- The Linear Regression model can be deployed on the recruitment dataset for predicting the expected salary of future employees.

- Current salary of a candidate has the highest bearing on his/her expected salary prediction, with expected salary expectations increasing with increase in current salary.

- Educational qualification of a candidate is also very crucial, with increase in education credentials being linked to a very high order of increase in the expected salary.

- PHD candidates have highest salary expectations, followed by Post graduates, then Graduates and finally Under graduates.

- Total work experience of a candidate is a more important deciding factor of expected salary than the total experience in the field/job applied. This means that the company values the overall work experience of a candidate, not just his/her experience in a certain field.

- A crucial factor deciding the salary expectations of a candidate is whether he/she already has an employment offer in hand. Candidates with a valid in-hand offer get higher salaries than those who have no offer in hand.

- The last appraisal rating of a candidate is important to deciding the expected salary, however, it is the candidates with average appraisal ratings (D & C) that manage to earn higher salaries.

- Some of the education-subject specialisation combinations that have high expected salary are:
    - PHD in Engineering
    - PHD in Maths/Stats
    - PG in Pure Sciences (Physics, Chemistry, Botany)
    - PG in Maths/Stats

- Factors such as a candidate's job role, his current organisation, the industry he/she currently works in and his/her current designation, have no impact on the expected salary. The table below shows the various variables as salary predictors:

Table: 6

Weak predictors	Moderate predictors	Strong predictors
Department Industry Organization University Location Number of publications International degree	Role Designation In-hand offer Specialisation subject Number of companies worked Total experience in field applied	Current CTC Education Last appraisal rating Total work experience Year of passing

APPENDIX

CONTENT LIST

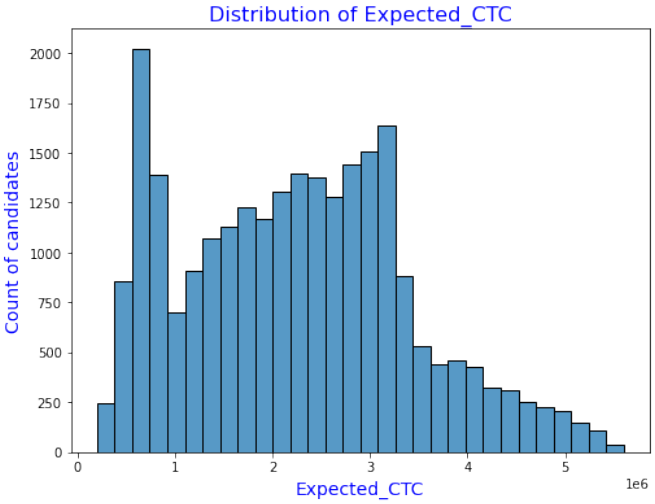
APPENDIX - A: Snapshot of the dataset

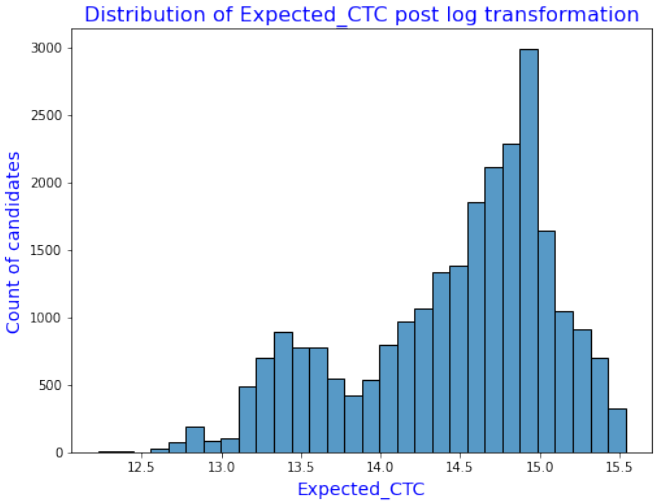
IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	University_Grad	Passing_Year_Of_Graduation	PG_Specialization	University_PG	Passing_Year_Of_PG	PHD_Specialization	University_PHD	Passing_Year_Of_PHD	Current_Location	Preferred_location	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked	Number_of_Publications	Certifications	International_degree_any	Expected_CTC
1	22753	0	0	NA	NA	NA	NA	NA	PG	Arts	Lucknow	2020							Guwahati	Pune	0	N	NA	0	0	0	0	384551
2	51087	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Chemistry	Surat	1988	Others	Surat	1990	Chemistry	Mangalore	1997	Bangalore	Nagpur	2702664	Y	Key_Performer	2	4	0	0	3783729
3	38413	21	12	Top Management	Consultant	Training	J		Doctorate	Zoology	Jaipur	1990	Zoology	Jaipur	1992	Zoology	Lucknow	1999	Ahmedabad	Jaipur	2236661	Y	Key_Performer	5	3	0	0	3131325
4	11501	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Others	Bangalore	1997	Zoology	Bangalore	1999	Chemistry	Guwahati	2005	Kanpur	Kolkata	2100510	N	C	5	3	0	0	2608833
5	58941	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Zoology	Mumbai	2004	Zoology	Mumbai	2006	Zoology	Bangalore	2010	Ahmedabad	Ahmedabad	1931644	N	C	2	3	0	0	2221390
6	30564	16	3	Top Management	Area Sales Manager	Retail	G	Director	Doctorate	Others	Bangalore	1998	Zoology	Bangalore	2000	Others	Bhubaneswar	2004	Pune	Bhubaneswar	3511167	Y	C	5	4	0	0	4522383
7	27267	1	1	Engineering	Team Lead	FMCG	L	Marketing Manager	Grad	Chemistry	Delhi	2011	Chemistry	Delhi	2013	Chemistry	Mangalore	2019	Delhi	Pune	492917	Y	B	3	3	0	0	630933
8	36521	19	11	Others	Analyst	Others	E	Manager	PG	Sociology	Delhi	2001							Surat	Bangalore	2693459	N	C	6	1	1	0	3221376
9	11616	8	7	Analytics/BI	Others	Telecom	L	Marketing Manager	Doctorate	Psychology	Mumbai	2003	Psychology	Mumbai	2005	Psychology	Pune	2012	Nagpur	Guwahati	1630225	N	A	6	6	0	0	2288835
10	43886	15	15	Analytics/BI	CEO	Telecom	M	Product Manager	Doctorate	Chemistry	Delhi	1998	Chemistry	Delhi	2000	Chemistry	Jaipur	2005	Bangalore	Bhubaneswar	1761797	Y	A	5	6	0	1	2466515
11	47373	13	10	Education	Business Analyst	Automobile	L	Consultant	Doctorate	Mathematics	Mangalore	2000	Mathematics	Mangalore	2002	Others	Guwahati	2007	Bangalore	Nagpur	2232801	Y	Key_Performer	3	3	0	0	3125921
12	36565	7	1	Marketing	Sales Manager	FMCG	E	CA	PG	Others	Mumbai	2004	Chemistry	Mumbai	2006	Others	Surat	2013	Jaipur	Mangalore	1351191	N	B	2	8	1	0	1756548
13	11739	10	10	Others	Bio statistician	Automobile	H	Consultant	Under Grad										Surat	Kanpur	1446451	N	C	3	2	1	0	1663418
14	14459	0	0	NA	NA	NA	NA	NA	PG	Engineering	Nagpur	2012	Engineering	Nagpur	2014	Engineering	Kolkata	2020	Nagpur	Bangalore	0	N	NA	0	6	0	0	639655
15	45111	12	9	Banking	Bio	Telecom	J	Research	Grad	Zoology	Surat	2002	Others	Surat	2004	Others	Delhi	2008	Jaipur	Mumbai	1990865	N	C	6	3	0	0	2289494

APPENDIX - B: Variable description and type

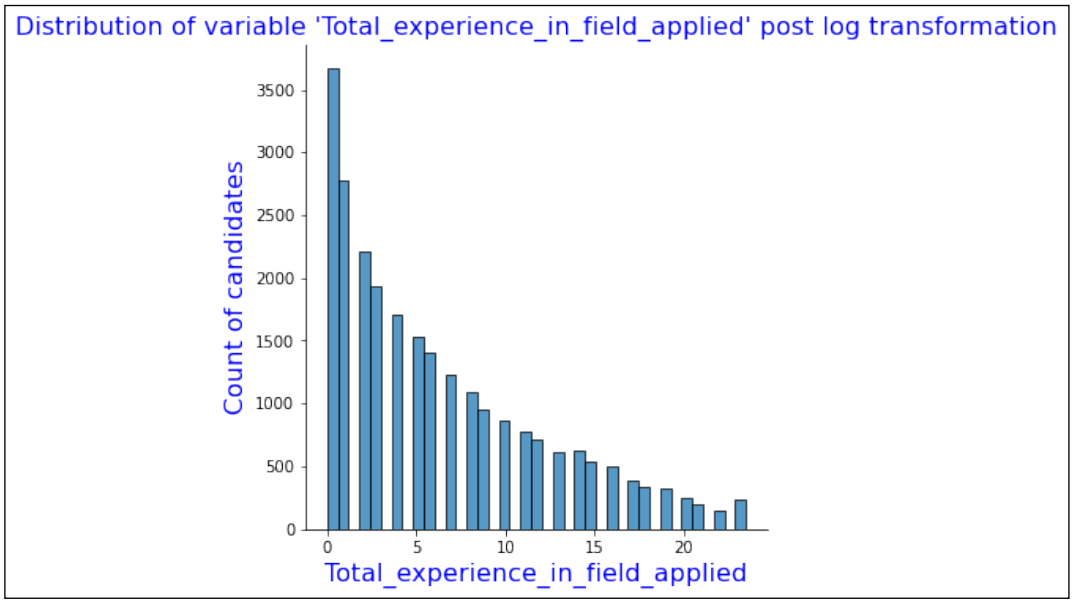
Sr no.	Variable name	Description	Variable type
1	IDX	Index no.	Numerical
2	Applicant_ID	Unique identifier for applicants	Numerical
3	Total_Experience	Total industry experience of the candidate	Numerical
4	Total_Experience_in_ field_ applied	Experience in the field applied / past experience relevant to the job	Numerical
5	Department	Department name of current company	Categorical
6	Role	Job role in the current company	Categorical
7	Industry	Industry name of the current work field	Categorical
8	Organization	Name of current organisation	Categorical
9	Designation	Designation in the current company	Categorical
10	Education	Highest educational credential	Categorical
11	Graduation_ specialisation	Specialisation subject in graduation	Categorical
12	University_Grad	City of Graduation university / college	Categorical
13	Passing_year_of_ graduation	Year of passing graduation	Numerical
14	PG_ specialisation	Specialisation subject in Post-graduation	Categorical
15	University_PG	City of Post-Graduation university / college	Categorical
16	Passing_year_of_ PG	Year of passing Post-graduation	Numerical
17	PHD_specialisation	Specialisation subject in PHD	Categorical
18	University_PHD	City of PHD university / college	Categorical
19	Passing_year_of_PHD	Year of passing PHD	Numerical
20	Current_location	City of current job	Categorical
21	Preferred_location	Preferred job location	Categorical
22	Current_CTC	Current salary (in INR)	Numerical
23	Inhand_offer	Whether candidate currently has another job offer	Categorical
24	Last_appraisal_rating	Last appraisal rating at the current job	Categorical
25	No_of_companies_ worked	Total number of companies candidate has previously worked in	Numerical
26	Number_of_ publications	Number of papers candidate has published	Numerical
27	Certifications	Number of certifications candidate has completed	Numerical
28	International_degree_ any	International degrees held by the candidate	Numerical
29	Expected_CTC	Final salary offered by Delta Ltd. to candidate	Numerical

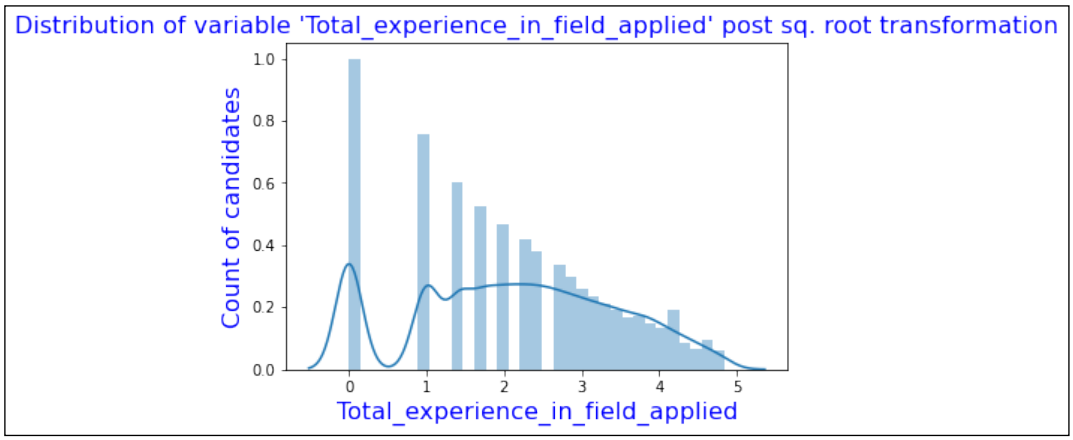
APPENDIX - C: Target variable histograms



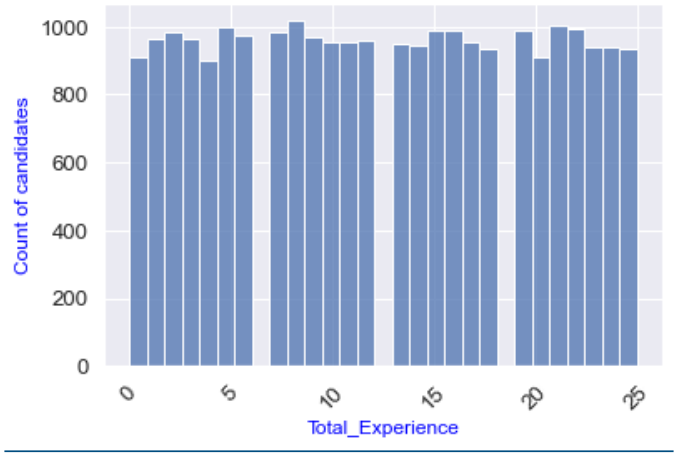


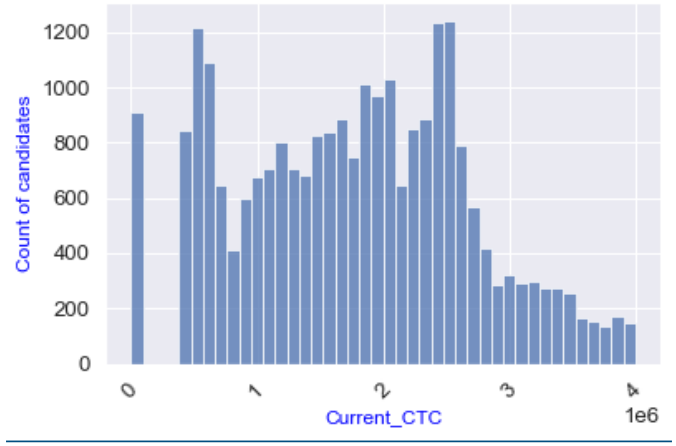
APPENDIX - D: Skew rectification using log and sq. root transformation

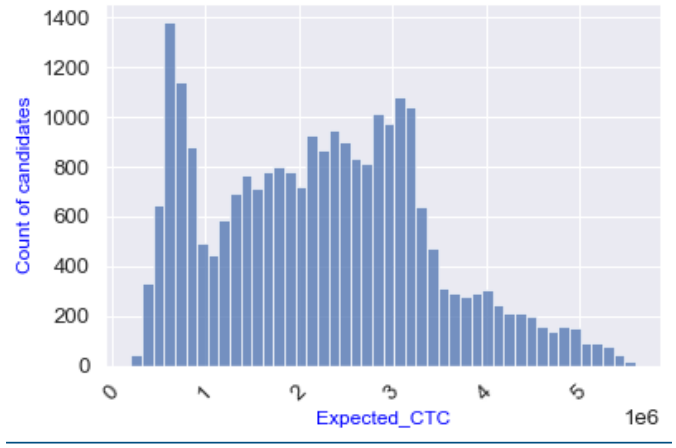




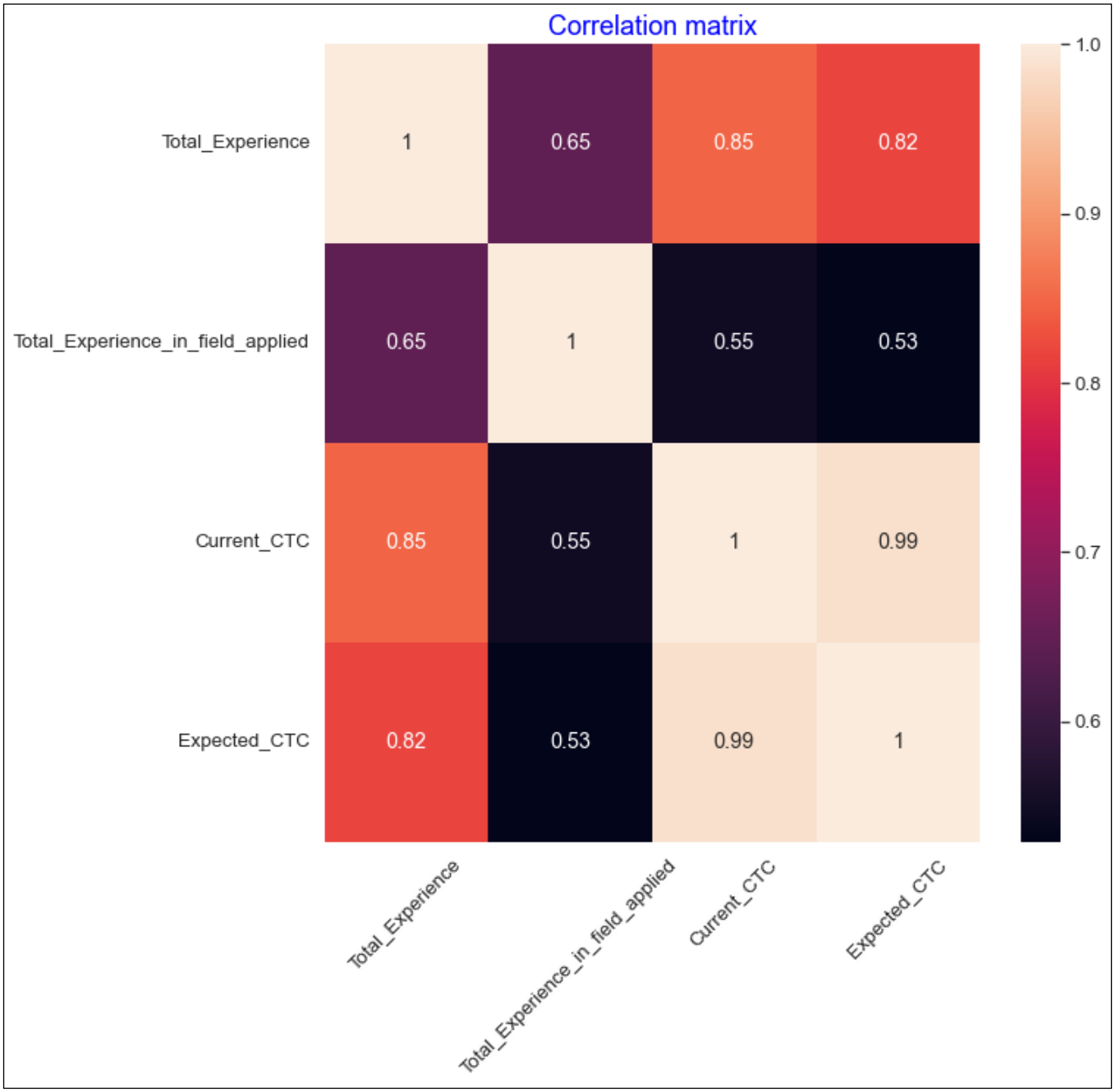
APPENDIX - E: Distribution of continuous variables



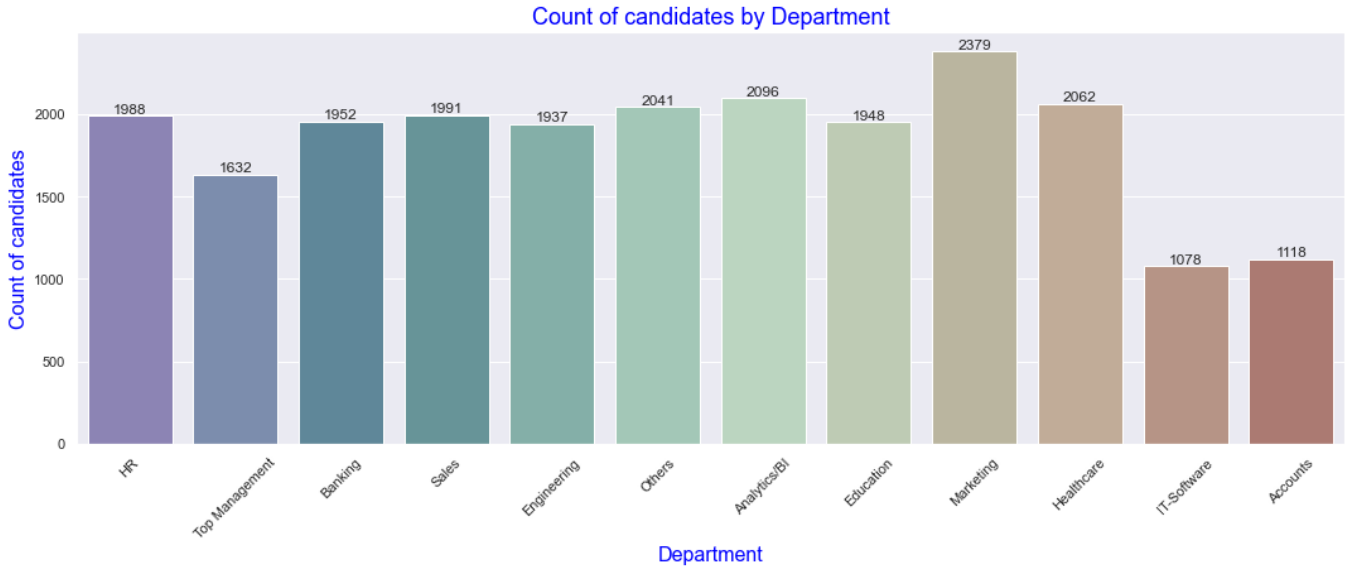


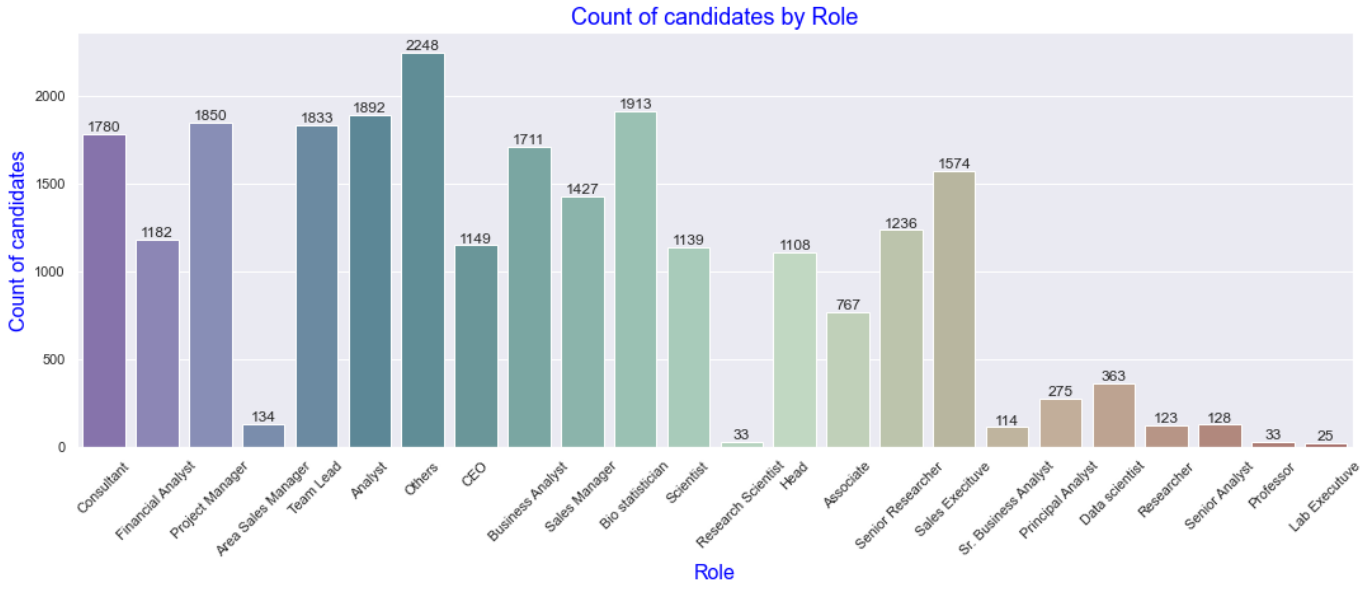


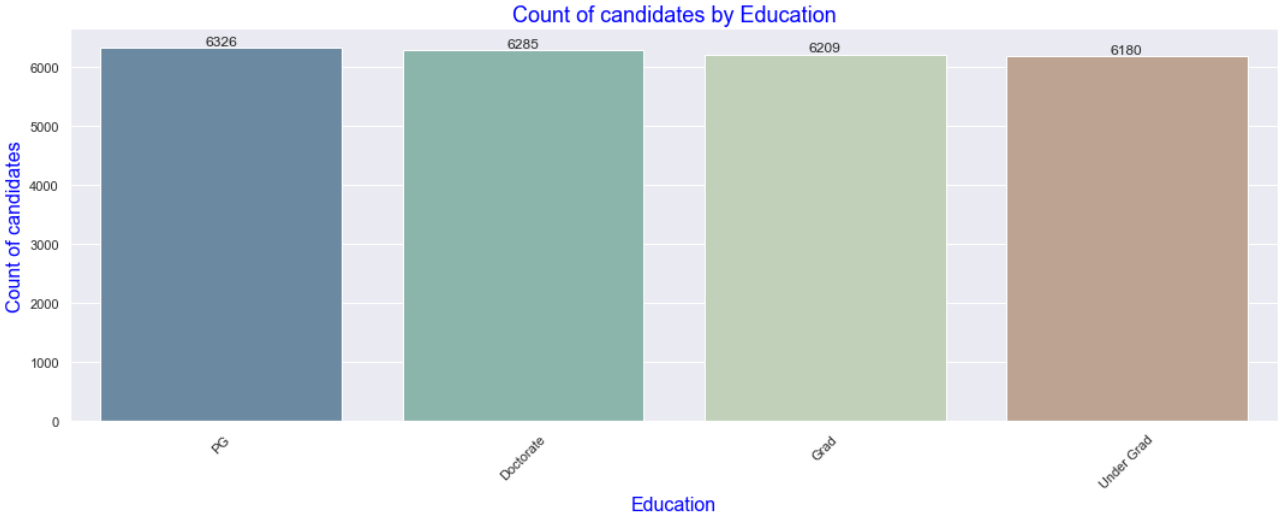
APPENDIX - F: Heat-map of continuous variables

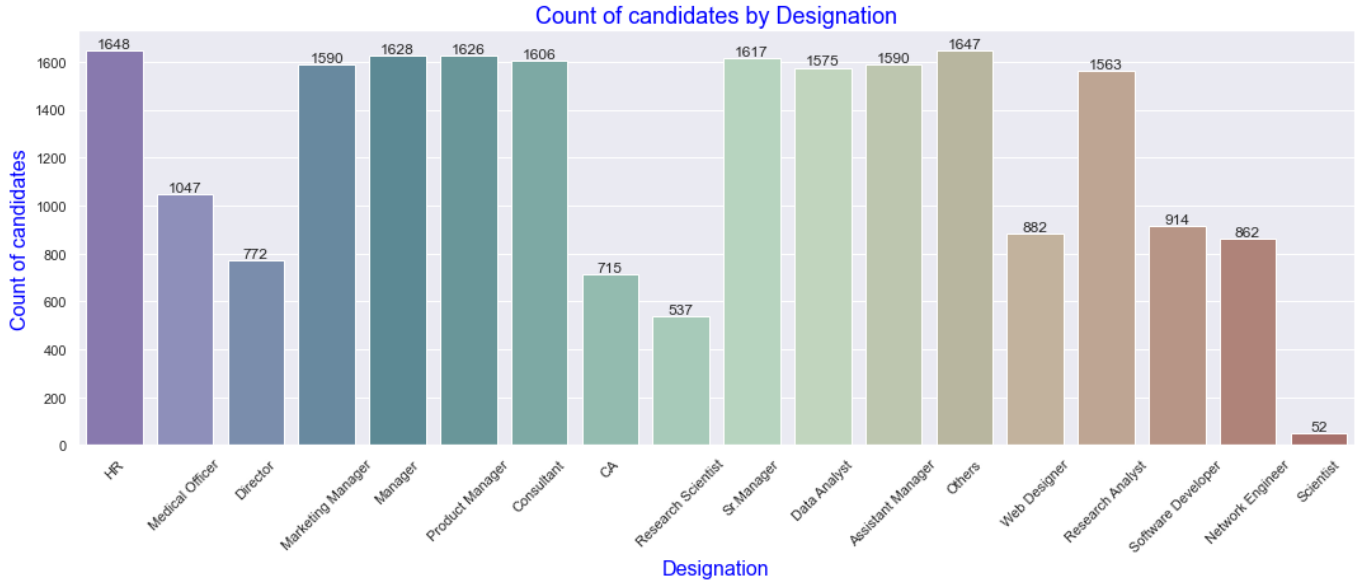


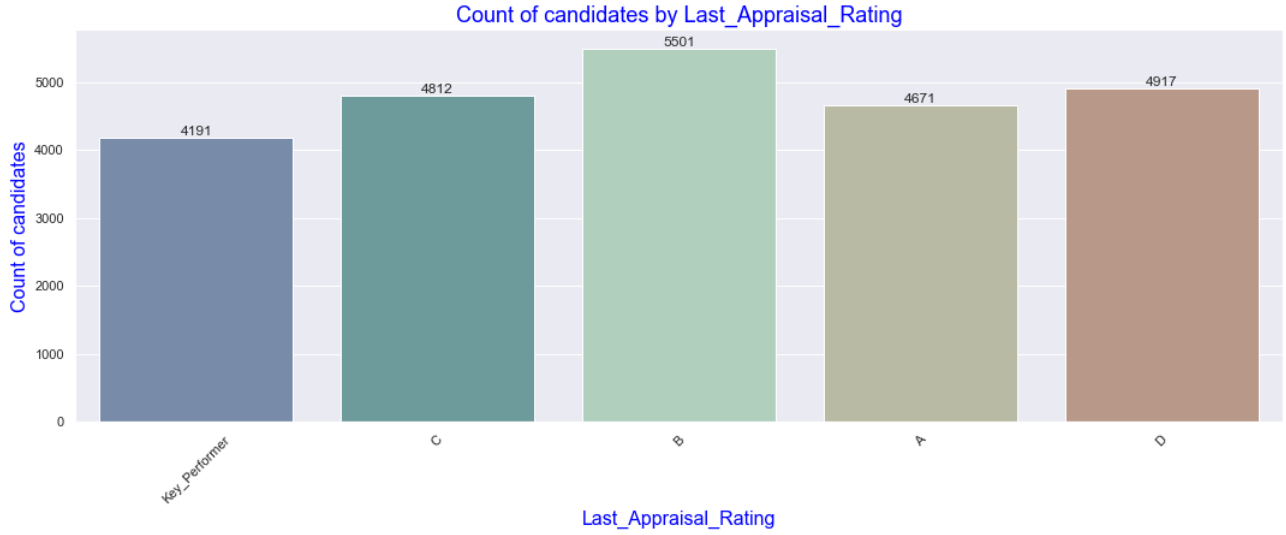
APPENDIX - G: Countplots of categorical variables

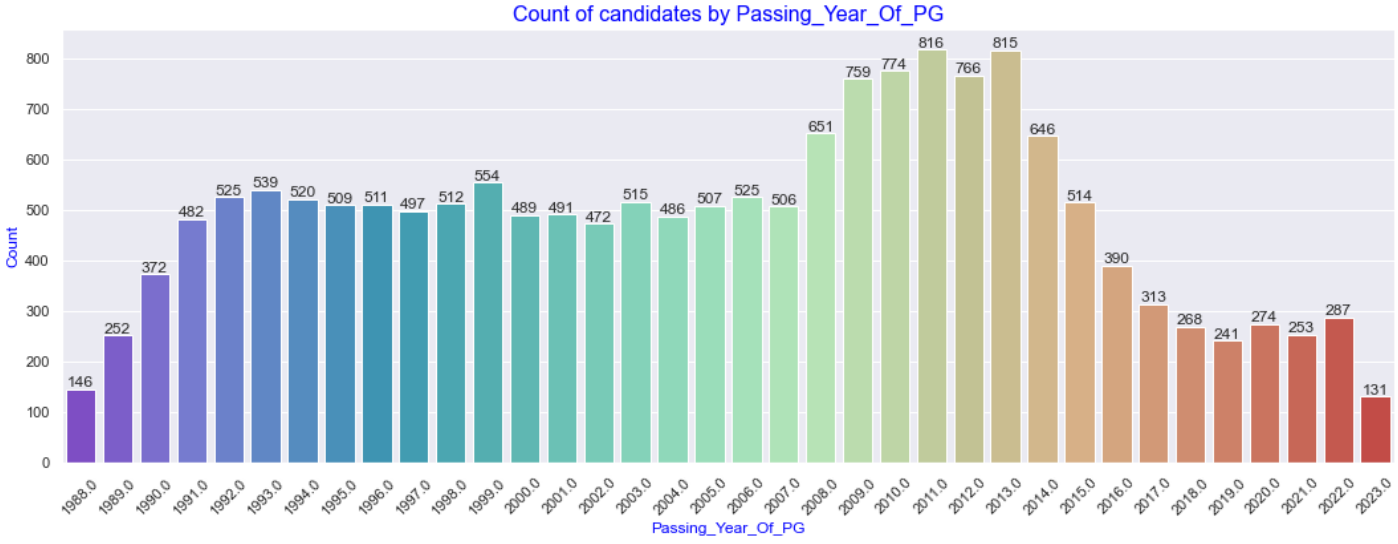




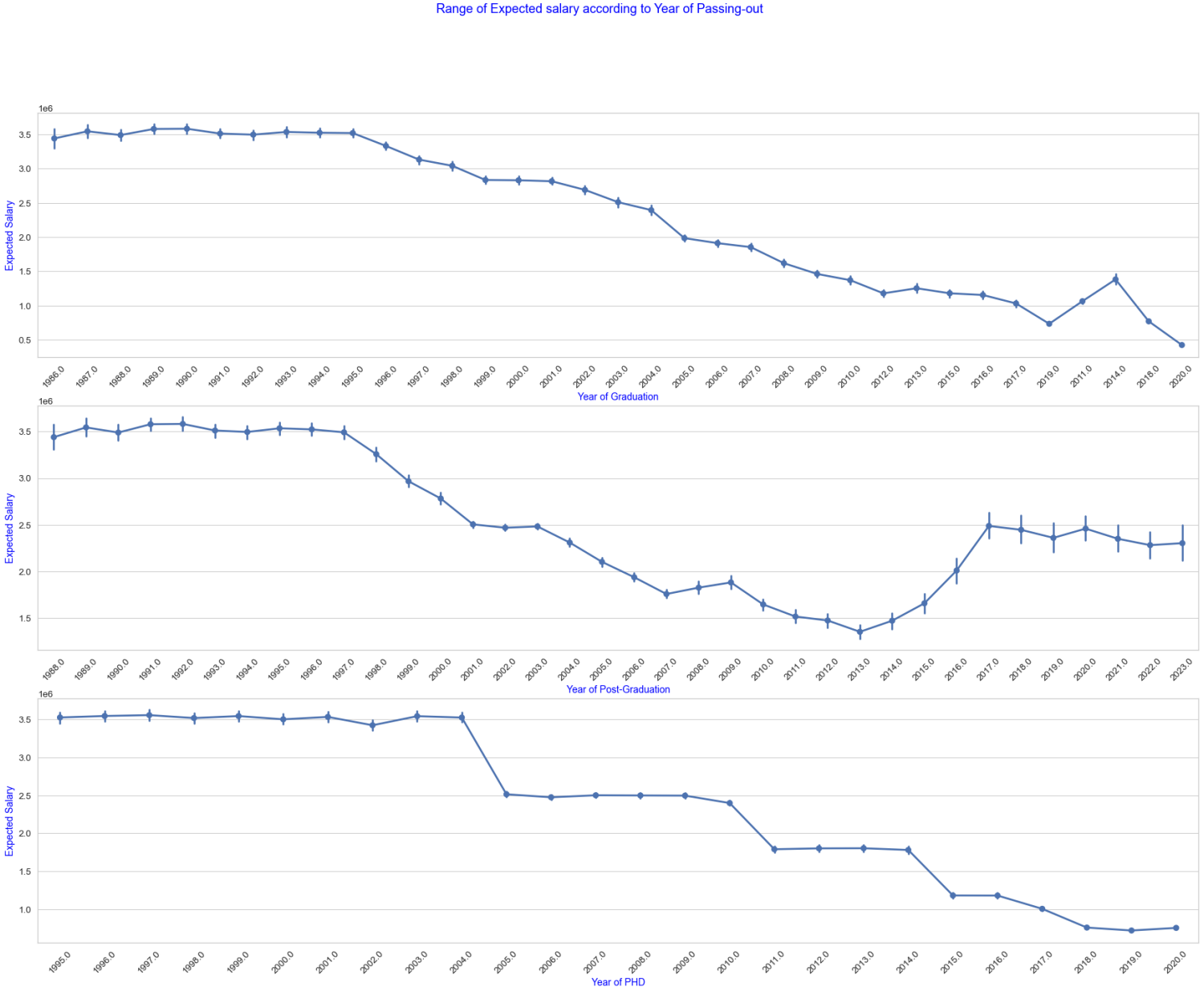



Count of candidates by Designation
Count of candidates
Designation
HR
Medical Officer
Director
Marketing Manager
Manager
Product Manager
Consultant
CA
Research Scientist
Sr.Manager
Data Analyst
Assistant Manager
Others
Web Designer
Research Analyst
Software Developer
Network Engineer
Scientist
1648
1047
772
1590
1628
1626
1606
715
537
1617
1575
1590
1647
882
1563
914
862
52

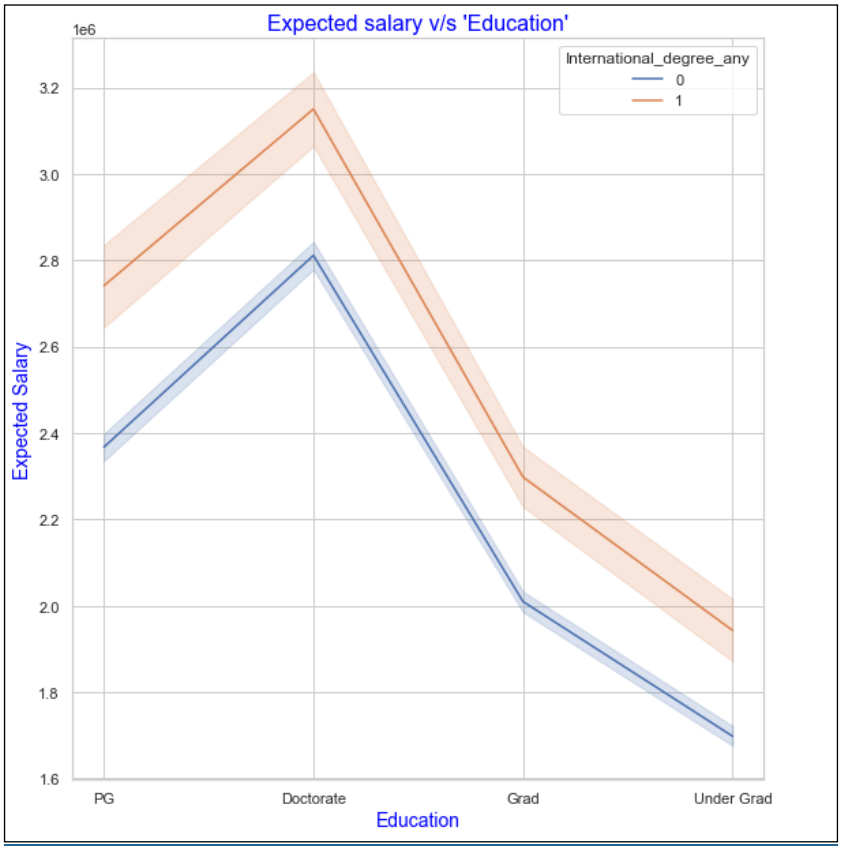

Count of candidates by Last_Appraisal_Rating
Count of candidates
Last_Appraisal_Rating
Key_Performer
C
B
A
D
4191
4812
5501
4671
4917


Count of candidates by Passing_Year_Of_PG
Count
Passing_Year_Of_PG
146
252
372
482
525
539
520
509
511
497
512
554
489
491
472
515
486
507
525
506
651
759
774
816
766
815
646
514
390
313
268
241
274
253
287
131

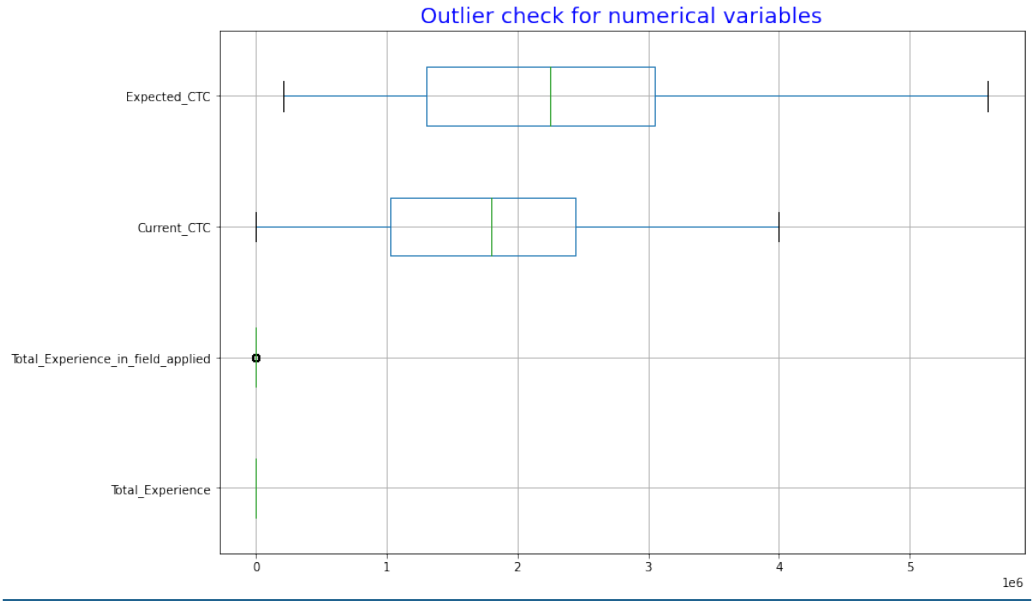
APPENDIX - H: Year-of-passing-out versus 'Expected_CTC'

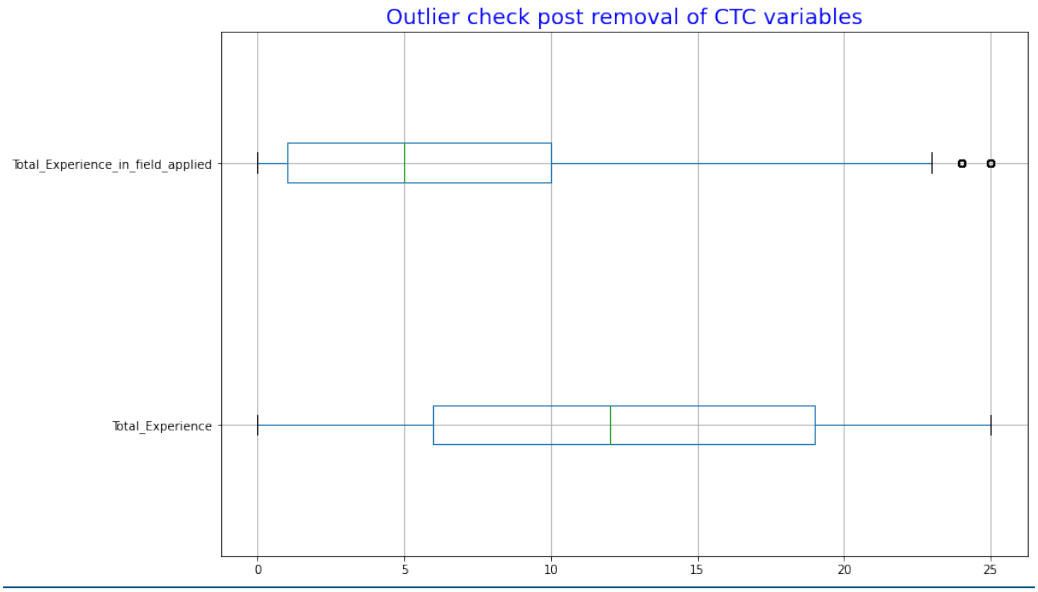


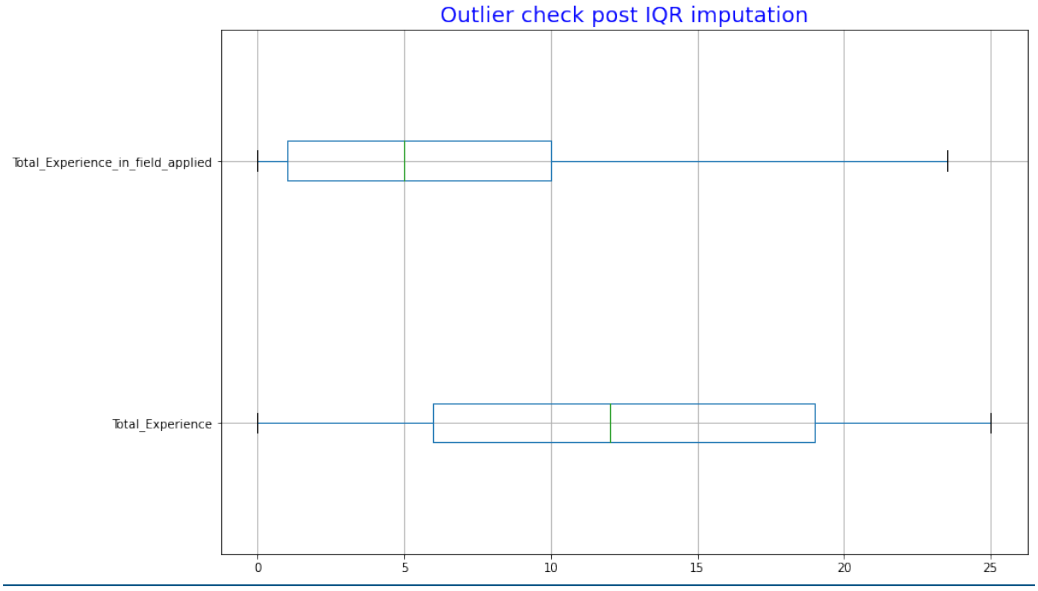
APPENDIX - I: Expected_salary versus Education



APPENDIX - J: Box-plots before and after outlier treatment







APPENDIX - K: Difference between new variable 'Edu_qualification' and old variable 'Education'