Final Report for Capstone

Determining the CTC to be offered for Employees joining Delta Ltd.

Submitted By

Yash Saxena [Batch: PGP DSBA June’20]

Project Mentor

Mr. Amitabh Sharma

Contents

List of tables and figures

1. Introduction

- Problem Statement: The HR department of Delta Ltd. is looking to be more data oriented in terms of their CTC offering to the new employees in order to meet the industry standards, align with the expectations of the candidate, and to reduce biasness.

- Need of the study/project: One of the most important aspects of onboarding new talent(s) to the company is offering a fair and competitive compensation which motivates the talent(s) to join the organization and work towards the success of it.

In this highly competitive talent-hunt market, many HR personnel have seen losing the great talent to the other companies by the virtue of offering a lower compensation than their competitor company. Therefore, in this project I have taken up solving this problem of utilising the Data Science tools to help the hiring team make an informed decision about ***"what would be the ideal CTC for a particular candidate"***. Some of the use cases of this project are:

1. Estimating a fair and competitive CTC for a new hire
2. Assessing the monetary loss if a new candidate is to be hired as a replacement and thus, take necessary interventional measures
3. Evaluating the parameters which affect the CTC offered which could be used extensively in salary negotiation, making the workplace more employee friendly, and in improving the hiring process altogether
4. Calculating the cost of adding new members with certain criteria
5. Aiding the process of hunting new candidates by giving another dimension to look at

In terms of social opportunity, nowadays, candidates expect the hiring team to show the CTC in the job description to help the decision making of the candidate. However, basis a recent data (as scraped from Indeed.com) of 10,000+ jobs in the US for analytics, only 8% had shown the CTC offered. This demand (of showing CTC in JD) of candidates has taken the LinkedIn and other social media sites by storm and it was concluded that almost all good candidates expect the hiring teams to do so.

Therefore, this will bring a **calculated transparency** not just some random range.

2. EDA and Business Implication

❖ Visual Inspection of the data:

- The total number of rows and columns this data has are given below:

No. of candidates	25000
No. of features	29

Table 1:Shape of data

- This dataset has:

No. of categorical columns	16
No. of numerical columns	29

Table 2:No. of categorical and continuous columns

- Null values in the dataset:

  1. There are **15 features** which have null values in them. (Refer to Appendix figure 1 for the visual)
  2. Out of these 15, **12 features** are categorical and **3 features** are continuous data types
  3. The percentage of null values in these features are as given in figure 2. (Refer to Appendix figure 1 for the visual)

- Basic descriptive statistics of numerical variables is given in Table 3. (Refer to appendix table 3)
- Basic descriptive statistics of categorical variables is given in Table 4. (Refer to appendix table 4):

Some inferences made from the descriptive statistics of the features:

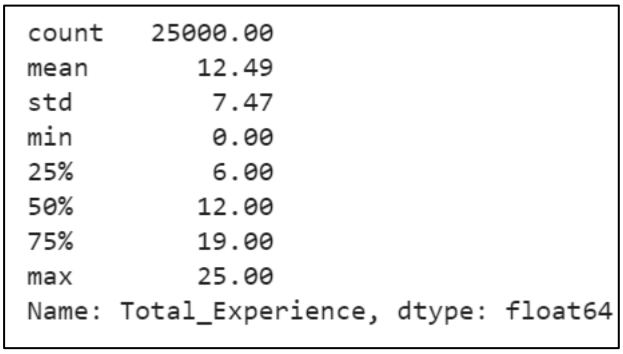
1. Average experience of the candidates interviewed is **12 years**
2. 25% of the candidates have 1 or less than 1 year of experience in the field of their application
3. Company hires freshers too
4. By just looking at the descriptive statistics of total experience in the field applied feature, there is a relatively large gap between 75th percentile and the maximum which indicates the presence of outliers
5. Not many candidates have any certifications. About 50% candidates do not have any certifications
6. Highest number (**10%**) of the candidates applied are from marketing department in their current company

Note: "Field_experience_range" is renamed to "Relevant_experience"

❖ Univariate Analysis:

1. **Percentage of candidates by experience range:** Bins were created from the total experience variable and the ranges were selected according to the most occurring experience ranges on job postings. (Refer to Appendix Figure 5)

2. **Descriptive Statistics on Total experience:**



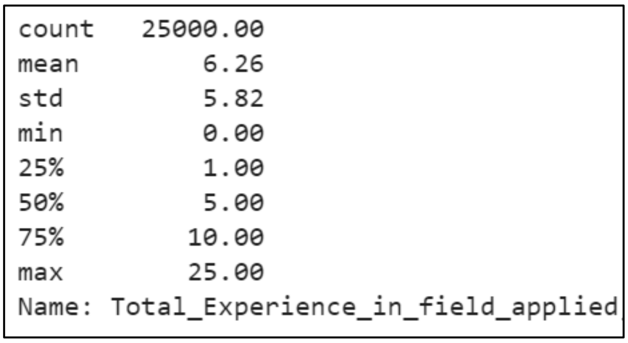
```
count    25000.00
mean        12.49
std          7.47
min          0.00
25%          6.00
50%         12.00
75%         19.00
max         25.00
Name: Total_Experience, dtype: float64
```

Table 3:Descriptive Statistics on Total experience

Insight: Average of the total experience of the candidates is ~**12.5 years**. This feature is approximately symmetric.

3. **Descriptive Statistics on Total experience in field applied:** Bins similar to the ranges of total experience feature were created and countplot was created. (Refer to Figure 6 on Appendix)

4. **Descriptive Statistics on Total experience in field applied:**



```
count    25000.00
mean         6.26
std          5.82
min          0.00
25%          1.00
50%          5.00
75%         10.00
max         25.00
Name: Total_Experience_in_field_applied
```

Table 4:Descriptive statistics of the total experience in field applied

Insight: Though, we had around **15%** of the candidates having total experience in the range of 21-25 years but only **1.8%** of the 21-25 years experienced candidates have this much of relevant experience. Let's investigate this further.

5. **Cross-tabulation of the total experience in field applied and total experience:** The cross-tab plot (refer to appendix table 5) explains the percentage of candidates across various categories of total experience ranges distributed across the categories of field experience ranges.

   For e.g., 47.81% of candidates with total experience range in 2-5 years have relevant field experience in the range of 0-1 years.

Note: All the remaining figures and tables have been added in the appendix

Insights from the Univariate and Bivariate analysis of the categorical variables:

1. Maximum number of candidates are from marketing department. It is ~10.3% of the total applicants which are from marketing.
2. Least proportion of candidates are being hired as 'Lab executives'
3. Educational level doesn't seem like much of a factor in hiring. This is because ~25% of applicants are from (PG, Doctorate, Graduation, and Under graduation)
4. Chemistry, Economics, and Mathematics are the major subjects in graduation, post-graduation, and PHD specialization for ~27% of these candidates.
5. Most of the applicants are preferring Kanpur as their work location. However, this number is not very high. 7%(maximum) of the applicants have Bangalore as their current location and 7% of the candidates prefer Kanpur as their work location.

Current location	Preferred location of maximum number of applicants	Percentage of applicants who wish to stay in their current location
Guwahati	Kanpur	6.60%
Bangalore	Kanpur	7.10%
Ahmedabad	Delhi	7.10%
Kanpur	Mangalore	7.70%
Pune	Surat	6.70%
Delhi	Delhi	7.50%
Surat	Bangalore	6.70%
Nagpur	Surat	6.60%
Jaipur	Ahmedabad	7.10%
Kolkata	Jaipur	7.20%
Bhubaneshwar	Kanpur	7.50%
Mangalore	Chennai	6.20%
Mumbai	Guwahati	6.20%
Lucknow	Lucknow	7.60%
Chennai	Ahmedabad	6.90%

Table 5:Analysis of current and preferred location

6. **70%** of the applicants have offer in-hand with them.
7. Going by department current CTC and consequently, expected CTC are highest for Top management, closely followed by engineering and healthcare
8. Education and accounts department have outliers in their current and expected CTC.

9. The median current CTC who has current roles as CEOs, Research scientist, Area sales manager, and head among the top 4. However, median current CTC who have current roles as professor, associate is among the lowest
10. There is no significant difference in current and expected CTC across different industries, different organizations, graduation specialization, city of university graduation, PG specialization, city of university PG specialization, PHD specialization subject, city of PHD specialization, and location.
11. Research scientist are being paid more than other designations. Doctorate are being paid highest.

12. Expected CTC if offer is in hand.

	Have an offer in hand	No offer in hand
Mean Current CTC	18,56,117	17,19,517
Mean Expected CTC	24,28,824	21,72,379

Table 6:Analysis of difference in expected/current CTC based on offer in-hand

13. **Pattern was seen**: Passing year of graduation/pg/phd were highly correlated with total experience and rightfully so. Therefore, I chose to drop *'passing year of graduation' ; 'passing year of PG' ; and 'passing year of PHD'*.

❖ Assessing the correlation:

- **Approach used:** As there is a mix of categorical and numerical columns, pearson's correlation would not be the direct choice of method to compute correlation. Therefore, I have opted to use a new library called as "dython". This library helped to find the correlation between all the features directly which can further help in feature selection process.
  - **Observations:**
    1. Current location and preferred location are giving little to no information about expected CTC.
    2. University grad and university pg are same for 17K+ records
    3. Grad specialization and PG specialization are same for 14K+ records
    4. This pattern of specialization and university is not same for PG and PHD
    5. Surprisingly, certifications have a negative impact on expected CTC which means rising number of certifications may not be good which can further mean company is looking for specialists not generalists assuming certifications are of different nature
    6. Relevant experience and total experience are closely correlated with each other so I have dropped that variable
    7. Current CTC and total experience are the best predictors for expected CTC
    8. Role that the candidate had in previous organization is more relevant predictor of expected ctc than designation

- **Action taken:** Current location, Preferred location, PG University, and PG Specialization will be dropped.
- **Correlation heatmap**: (Refer to figure 5 in the appendix

3. Data Cleaning and Pre-processing:

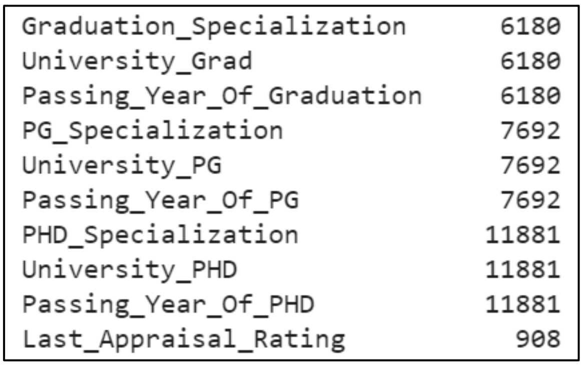
On investigating the nature of missing values, it was found that missing values were not missing at random for the features: ***Role, Designation, and Department***.

Approach used: It was observed that all these features had missing values for the applicants who had 0 years of total experience.

What was done? - 'Fresher' term was put in place of all those missing values.

Step #2 – A ML model based on CatBoostClassifier was created and one by one the missing values in Role, Department, and Designation were imputed. Post that, below mentioned was the distribution of the null values.

Step #3 – As it stands, given in the picture below is the state of the missing values now.



```
Graduation_Specialization      6180
University_Grad                6180
Passing_Year_Of_Graduation     6180
PG_Specialization              7692
University_PG                  7692
Passing_Year_Of_PG             7692
PHD_Specialization            11881
University_PHD                11881
Passing_Year_Of_PHD           11881
Last_Appraisal_Rating           908
```

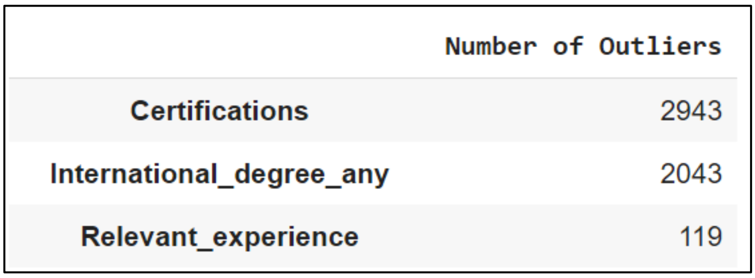
Table 7:Missing values status after first iteration

- **Imputation in Last_appraisal_rating:** It was observed that the value of this feature was missing for all the freshers. Hence, freshers were written in place of the missing values of this column
- **Imputation in graduation specialization, university graduation, and passing year of graduation:** For candidates where each of these three features are missing, education is under grad which means the candidate has not even graduated yet. Therefore, missing value was imputed by '*Did not do graduation*'.
- **Pattern was seen**: Passing year of graduation/pg/phd were highly correlated with total experience and rightfully so. Therefore, I chose to drop '*passing year of graduation' ; 'passing year of PG' ; and 'passing year of PHD'*.

Feature name	Correlation with expected CTC
Certifications	-0.173991964
Number_of_Publications	0.001517833
Curent_Location	0.020901888
Preferred_location	0.021430886
International_degree_any	0.074557037
Inhand_Offer	0.101581844
Graduation_Specialization	0.263417971
University_Grad	0.263461109
University_PG	0.269316706
PG_Specialization	0.26947235
Industry	0.280014114
Organization	0.280272284
Designation	0.293993246
University_PHD	0.294732533
PHD_Specialization	0.295065978
Department	0.322112863
Last_Appraisal_Rating	0.333040383
No_Of_Companies_worked	0.343150067
Education	0.360041448
Role	0.400620637
Relevant_experience	0.529115066
Total_Experience	0.816593165
Current_CTC	0.986718278

Table 8:Correlation of every feature with expected CTC in ascending order

3a. Outliers:



	Number of Outliers
Certifications	2943
International_degree_any	2043
Relevant_experience	119

Table 9: Outliers in the data

What to do with outliers? As certifications and international_degree_any can take distinct values only and none of the values are any anomaly. We can keep the outliers.

In addition to this, Relevant_experience also doesn't have any suspicious values. They have **24** and **25** years of experience are the outliers in this column which is possible.

3b. Dropping some columns:

As Passing year of PHD, passing year of PG, and Passing year of Graduation are very closely related to the variables Total experience and relevant experience, therefore, it has been decided to remove these variables.

In addition to this, as we saw, there was not much variance between the levels of current location and preferred location. We shall try and drop these columns as well to see if it helps in reducing the redundancy.

Also, since the place of university is a biased estimator of CTC, therefore, I have dropped these two variables.

In summary, following features were dropped.

1. Passing year of PHD
2. Passing year of graduation
3. Passing year of postgraduation
4. Current location
5. Preferred location
6. University location of Graduation
7. University location of post-graduation
8. University location of PHD

3d. Data pre-processing:

Categorical variable encoding: One-hot encoding was chosen to encode the nominal variables.

No. of columns	**127**
No. of rows	**25000**

Table 10:Feature matrix after one-hot

Feature selection: Features were selected using mutual information gain.

- **Mutual Information gain**: This parameter will let us know the information that a feature conveys about the predictor. We will eliminate the feature basis on the information conveyed by it.
- **Action taken**: Out of these 123 features 12 features were not conveying any information whatsoever about the target variable. So, those 12 features were eliminated.

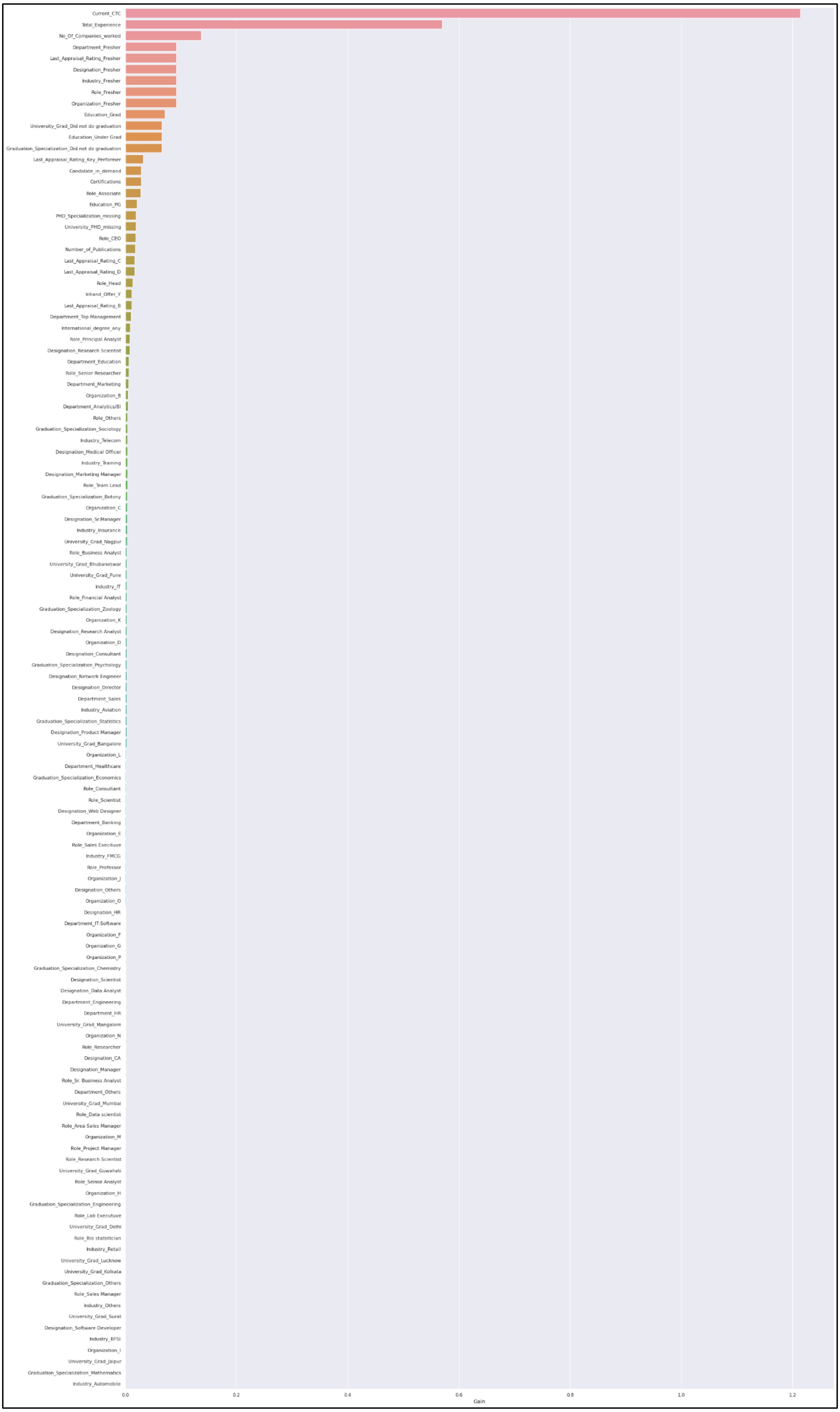
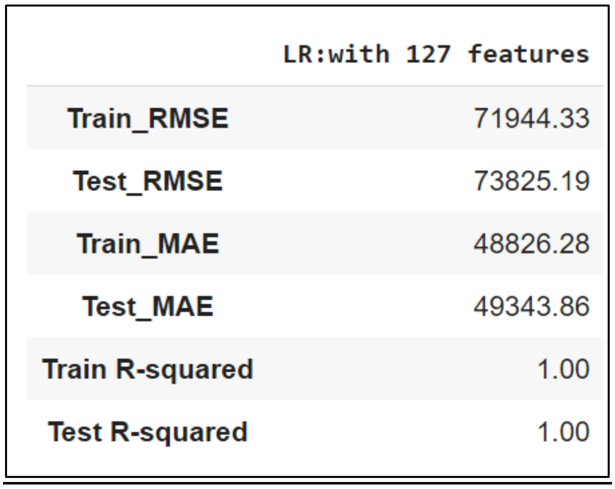


Figure 1:Information conveyed about CTC by each of the features

4. Model building:

As this is a regression problem, below mentioned models were trained and tested assess the suitability of the model. **The selection criteria** of choosing a model were **accuracy** as well as **explainability.**

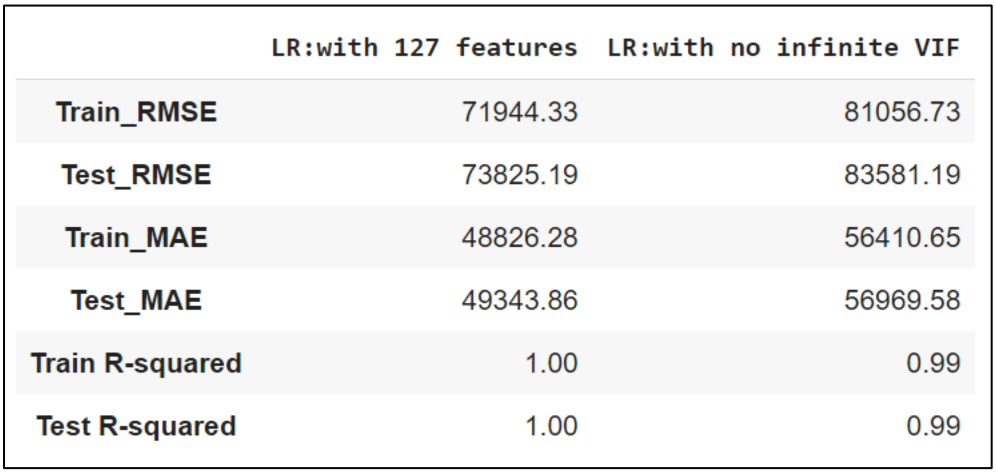
1. **Linear regression**: On regressing the Expected CTC with all the features, **31 features** had infinite VIF with each i.e., they were being completely explained by rest of the features.



	LR:with 127 features
Train_RMSE	71944.33
Test_RMSE	73825.19
Train_MAE	48826.28
Test_MAE	49343.86
Train R-squared	1.00
Test R-squared	1.00

Figure 2: Performance of Linear regression model with all the features

This clearly shows that the model is overfitted. On removing infinite VIF features, there were 109 **features** left. The performance then was:



	LR:with 127 features	LR:with no infinite VIF
Train_RMSE	71944.33	81056.73
Test_RMSE	73825.19	83581.19
Train_MAE	48826.28	56410.65
Test_MAE	49343.86	56969.58
Train R-squared	1.00	0.99
Test R-squared	1.00	0.99

Figure 3: Performance comparison of Linear regression models

OLS Significant features and their coefficients: Using statsmodels, following features were statistically significant and they are arranged according to their coefficients.

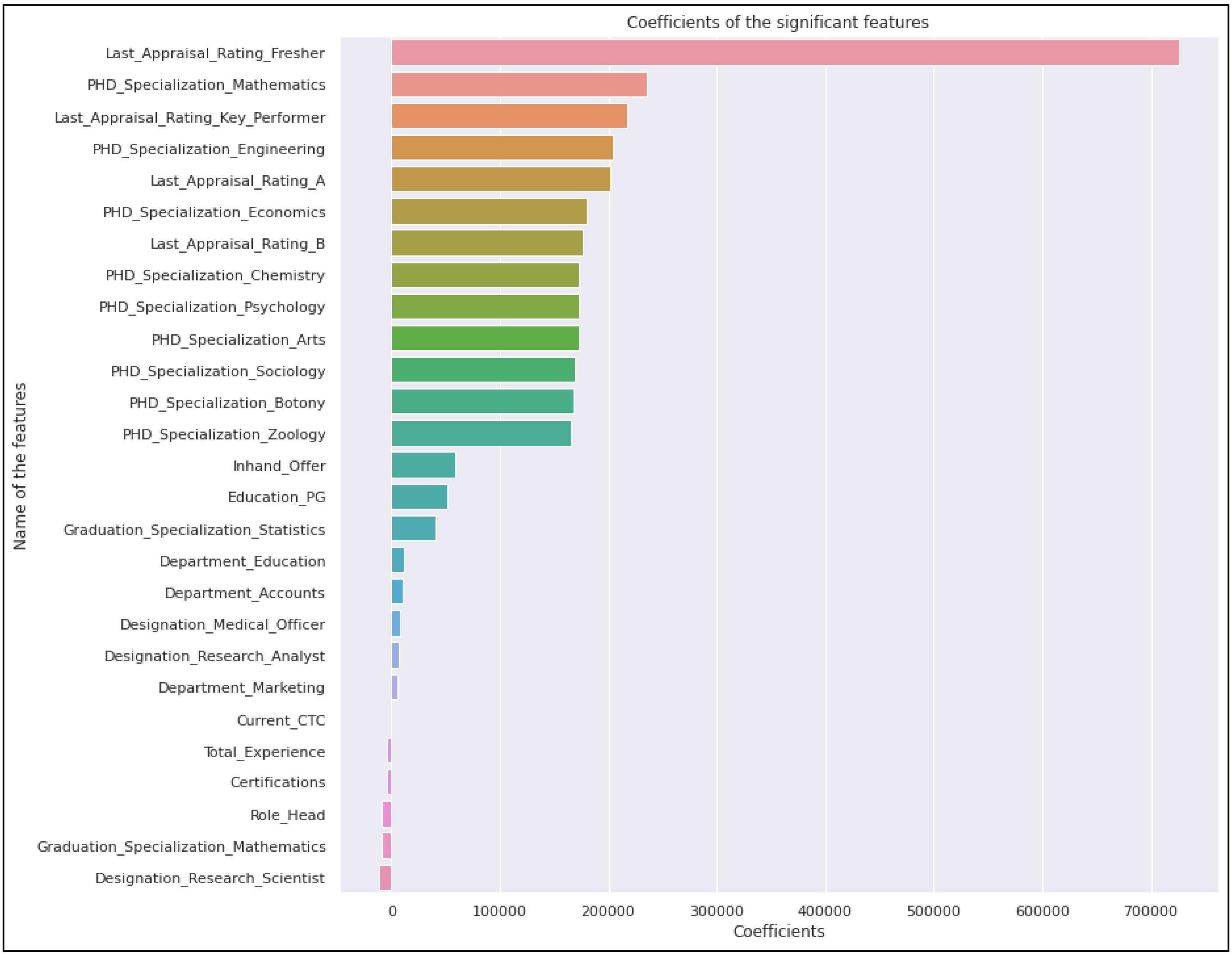
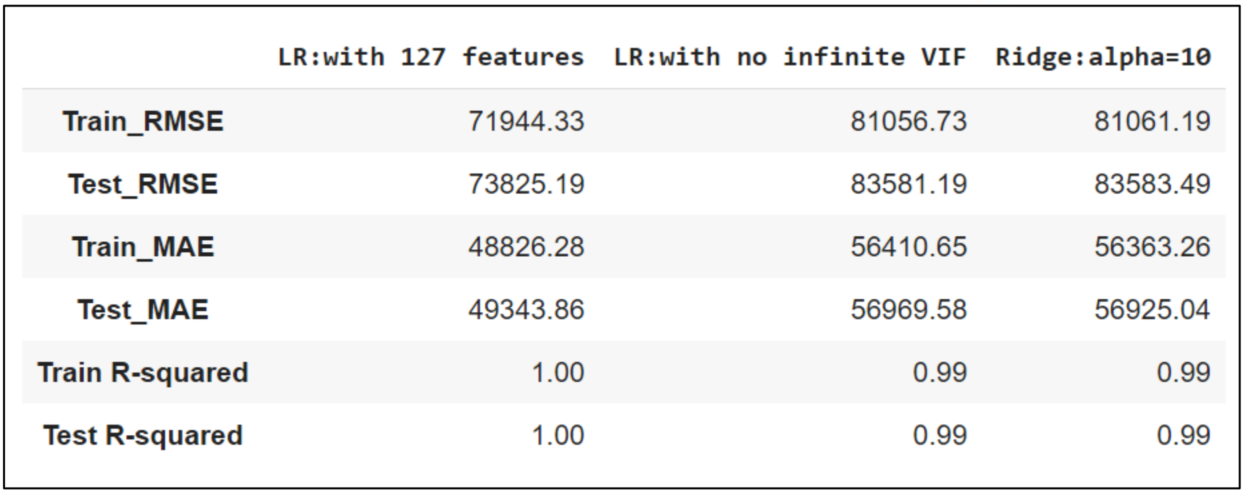


Figure 4: Coefficients of significant features

As the model was overfitted and had multicollinearity, **Ridge** and **lasso regression** were tried to tackle both the issues.

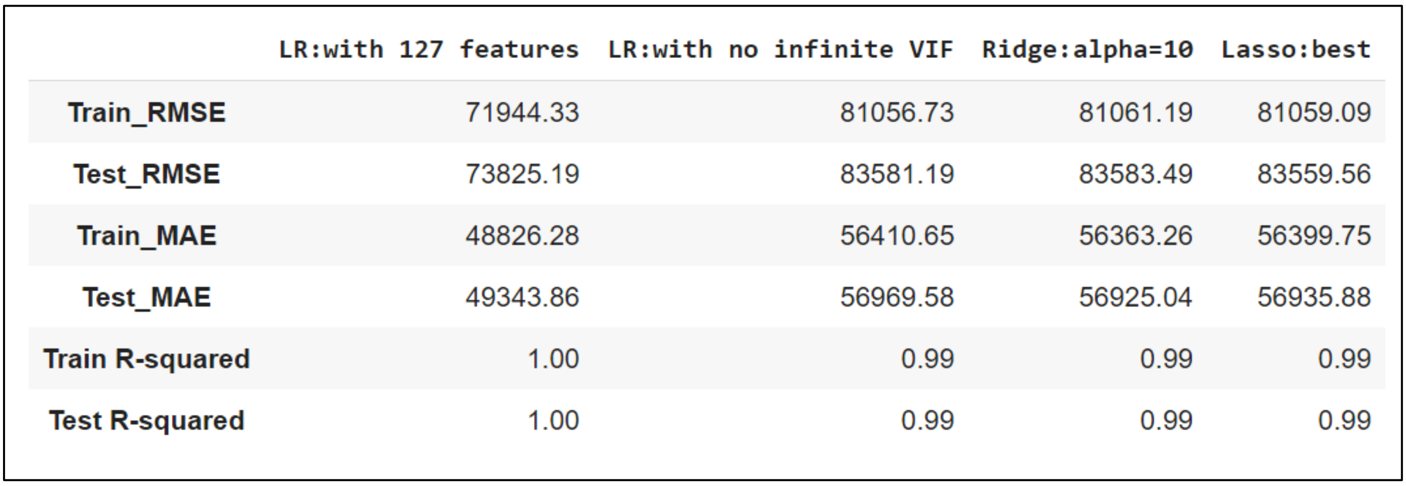
2. **Ridge regression**: Using the class RidgeCV, the suitable value of regularization parameter was found to be **1.01** but I chose a little higher (**10**) alpha so as to bring the difference between Training and testing RMSE down to a higher extent.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10
Train_RMSE	71944.33	81056.73	81061.19
Test_RMSE	73825.19	83581.19	83583.49
Train_MAE	48826.28	56410.65	56363.26
Test_MAE	49343.86	56969.58	56925.04
Train R-squared	1.00	0.99	0.99
Test R-squared	1.00	0.99	0.99

Figure 5: Linear regression and Ridge regression performance

3. **Lasso regression**: Using the class LassoCV, the suitable value of regularization parameter was found to be **29.82.**



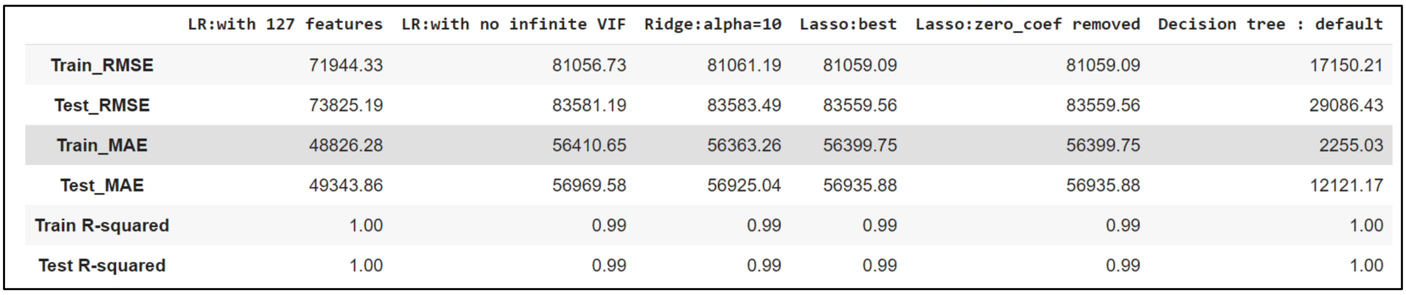
	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best
Train_RMSE	71944.33	81056.73	81061.19	81059.09
Test_RMSE	73825.19	83581.19	83583.49	83559.56
Train_MAE	48826.28	56410.65	56363.26	56399.75
Test_MAE	49343.86	56969.58	56925.04	56935.88
Train R-squared	1.00	0.99	0.99	0.99
Test R-squared	1.00	0.99	0.99	0.99

Figure 6: Linear, Ridge, and Lasso performance

Observation: Not much difference was seen between the predictive performance of linear, ridge, and lasso models.

The effect of multicollinearity is impeding the ability to interpret the coefficients, hence tree-based models were tried further.

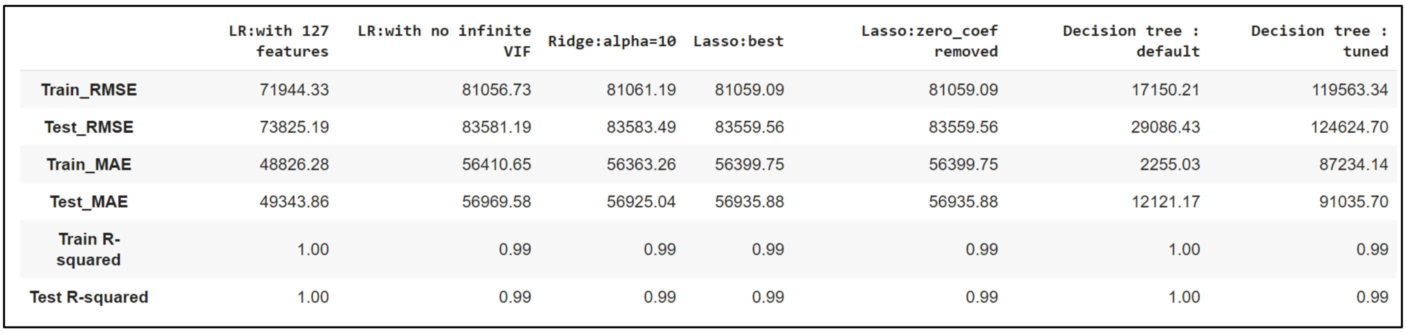
4. **Decision tree regressor**: With its default settings, the performance of the Decision tree model was as mentioned in the snippet below.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default
Train_RMSE	71944.33	81056.73	81061.19	81059.09	81059.09	17150.21
Test_RMSE	73825.19	83581.19	83583.49	83559.56	83559.56	29086.43
Train_MAE	48826.28	56410.65	56363.26	56399.75	56399.75	2255.03
Test_MAE	49343.86	56969.58	56925.04	56935.88	56935.88	12121.17
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00

Figure 7:Performance of Decision Tree Regressor against other models

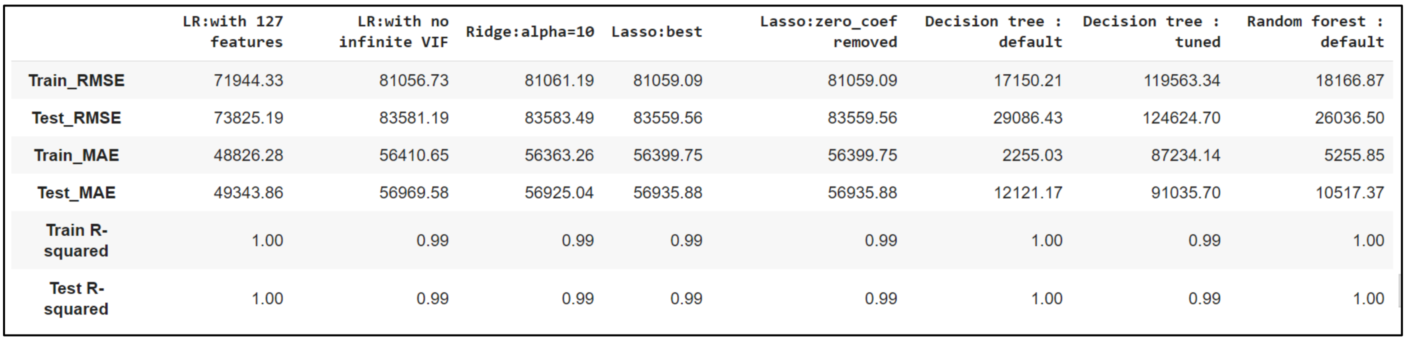
Clearly, the model is very overfitted. Therefore, pruning i.e., hyperparameter optimization was done using optuna.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default	Decision tree : tuned
Train_RMSE	71944.33	81056.73	81061.19	81059.09	81059.09	17150.21	119563.34
Test_RMSE	73825.19	83581.19	83583.49	83559.56	83559.56	29086.43	124624.70
Train_MAE	48826.28	56410.65	56363.26	56399.75	56399.75	2255.03	87234.14
Test_MAE	49343.86	56969.58	56925.04	56935.88	56935.88	12121.17	91035.70
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99

Figure 8: Effect of tuned decision tree

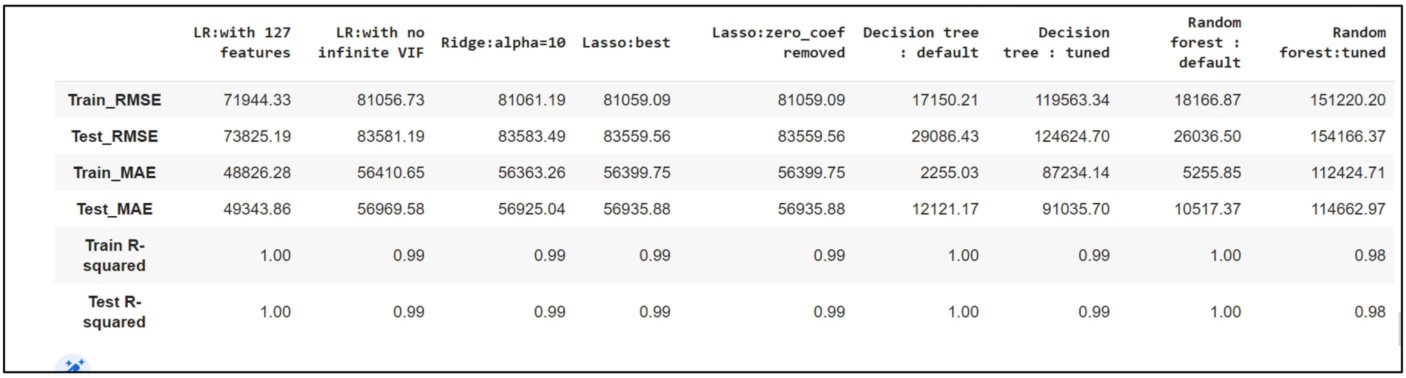
5. **Random forest**: With its default settings, the performance of the Random-forest model was as mentioned in the snippet below.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default	Decision tree : tuned	Random forest : default
Train_RMSE	71944.33	81056.73	81061.19	81059.09	81059.09	17150.21	119563.34	18166.87
Test_RMSE	73825.19	83581.19	83583.49	83559.56	83559.56	29086.43	124624.70	26036.50
Train_MAE	48826.28	56410.65	56363.26	56399.75	56399.75	2255.03	87234.14	5255.85
Test_MAE	49343.86	56969.58	56925.04	56935.88	56935.88	12121.17	91035.70	10517.37
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00

Figure 9: Performance of Random Forest with default settings

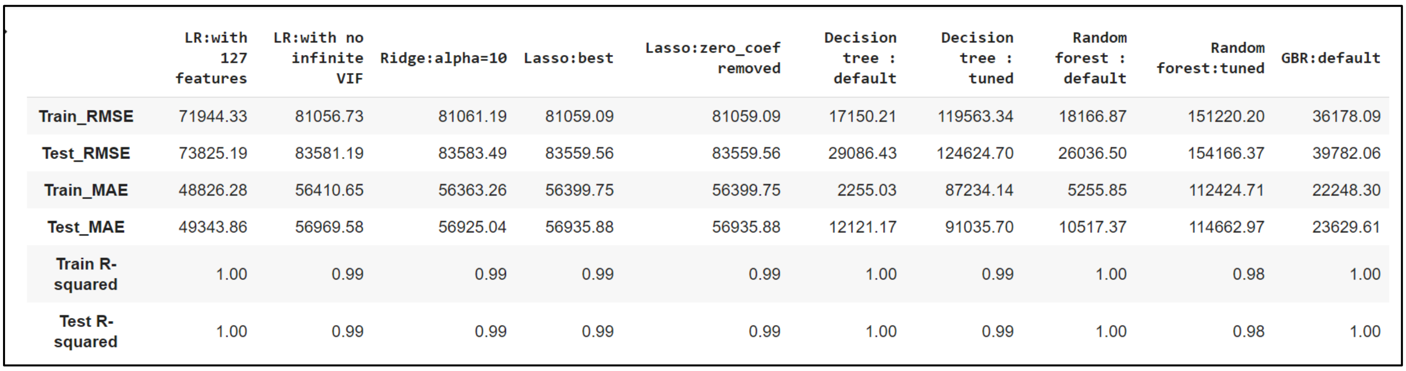
Observation: The difference between train and test RMSE went lower than it was with default Decision tree. Tuning the model further.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default	Decision tree : tuned	Random forest : default	Random forest:tuned
Train_RMSE	71944.33	81056.73	81061.19	81059.09	81059.09	17150.21	119563.34	18166.87	151220.20
Test_RMSE	73825.19	83581.19	83583.49	83559.56	83559.56	29086.43	124624.70	26036.50	154166.37
Train_MAE	48826.28	56410.65	56363.26	56399.75	56399.75	2255.03	87234.14	5255.85	112424.71
Test_MAE	49343.86	56969.58	56925.04	56935.88	56935.88	12121.17	91035.70	10517.37	114662.97
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98

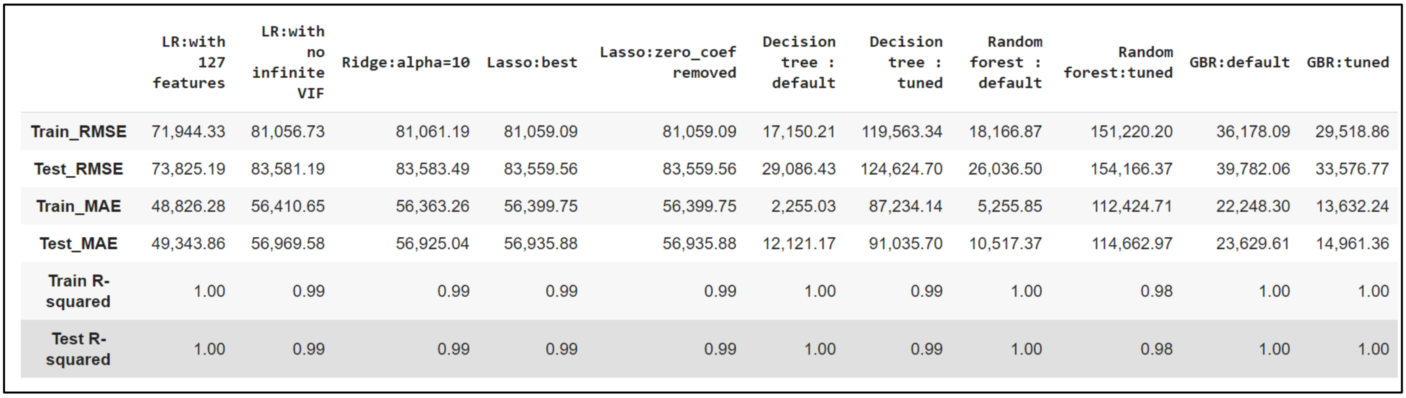
Figure 10: Tuned Random Forest performance

6. **Gradient Boosting Regressor**: With its default settings, the performance of the Gradient boosting regressor was as mentioned in the snippet below.



	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default	Decision tree : tuned	Random forest : default	Random forest:tuned	GBR:default
Train_RMSE	71944.33	81056.73	81061.19	81059.09	81059.09	17150.21	119563.34	18166.87	151220.20	36178.09
Test_RMSE	73825.19	83581.19	83583.49	83559.56	83559.56	29086.43	124624.70	26036.50	154166.37	39782.06
Train_MAE	48826.28	56410.65	56363.26	56399.75	56399.75	2255.03	87234.14	5255.85	112424.71	22248.30
Test_MAE	49343.86	56969.58	56925.04	56935.88	56935.88	12121.17	91035.70	10517.37	114662.97	23629.61
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98	1.00
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98	1.00

Figure 11:Performance of Gradient Boosting Regressor: default settings

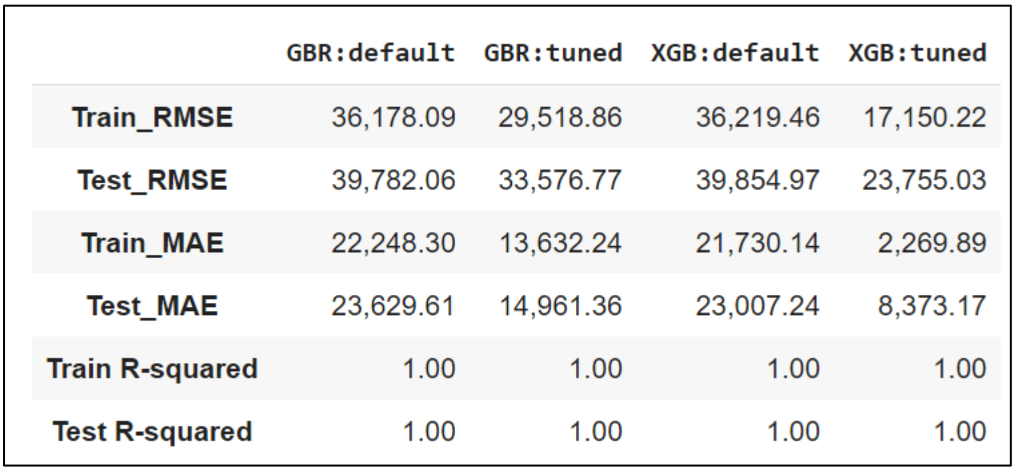


	LR:with 127 features	LR:with no infinite VIF	Ridge:alpha=10	Lasso:best	Lasso:zero_coef removed	Decision tree : default	Decision tree : tuned	Random forest : default	Random forest:tuned	GBR:default	GBR:tuned
Train_RMSE	71,944.33	81,056.73	81,061.19	81,059.09	81,059.09	17,150.21	119,563.34	18,166.87	151,220.20	36,178.09	29,518.86
Test_RMSE	73,825.19	83,581.19	83,583.49	83,559.56	83,559.56	29,086.43	124,624.70	26,036.50	154,166.37	39,782.06	33,576.77
Train_MAE	48,826.28	56,410.65	56,363.26	56,399.75	56,399.75	2,255.03	87,234.14	5,255.85	112,424.71	22,248.30	13,632.24
Test_MAE	49,343.86	56,969.58	56,925.04	56,935.88	56,935.88	12,121.17	91,035.70	10,517.37	114,662.97	23,629.61	14,961.36
Train R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98	1.00	1.00
Test R-squared	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	0.98	1.00	1.00

Figure 12: Performance of Gradient Boosting Regressor: Tuned

Observation: Tuned Gradient Boosting Regressor has performed quite well than the other models so far. Therefore, more advanced Gradient boosting models such as XGBoost, LightGBM, and CatBoost will be tried hereafter.

7. **XGBoost**: With its default settings and hyperparameter tuned settings, the performance of the XGBoost regressor was as mentioned in the snippet below.



	GBR:default	GBR:tuned	XGB:default	XGB:tuned
Train_RMSE	36,178.09	29,518.86	36,219.46	17,150.22
Test_RMSE	39,782.06	33,576.77	39,854.97	23,755.03
Train_MAE	22,248.30	13,632.24	21,730.14	2,269.89
Test_MAE	23,629.61	14,961.36	23,007.24	8,373.17
Train R-squared	1.00	1.00	1.00	1.00
Test R-squared	1.00	1.00	1.00	1.00

Figure 13:XGBoost Tuned and Default

Observation: The performance has increased but the model is very overfitted when tuned. Moving onto other high performing models i.e., LightGBM and CatBoost.

8. **LightGBM**: With its default settings, the performance of the XGBoost regressor was as mentioned in the snippet below.

Final conclusion about choosing the model:

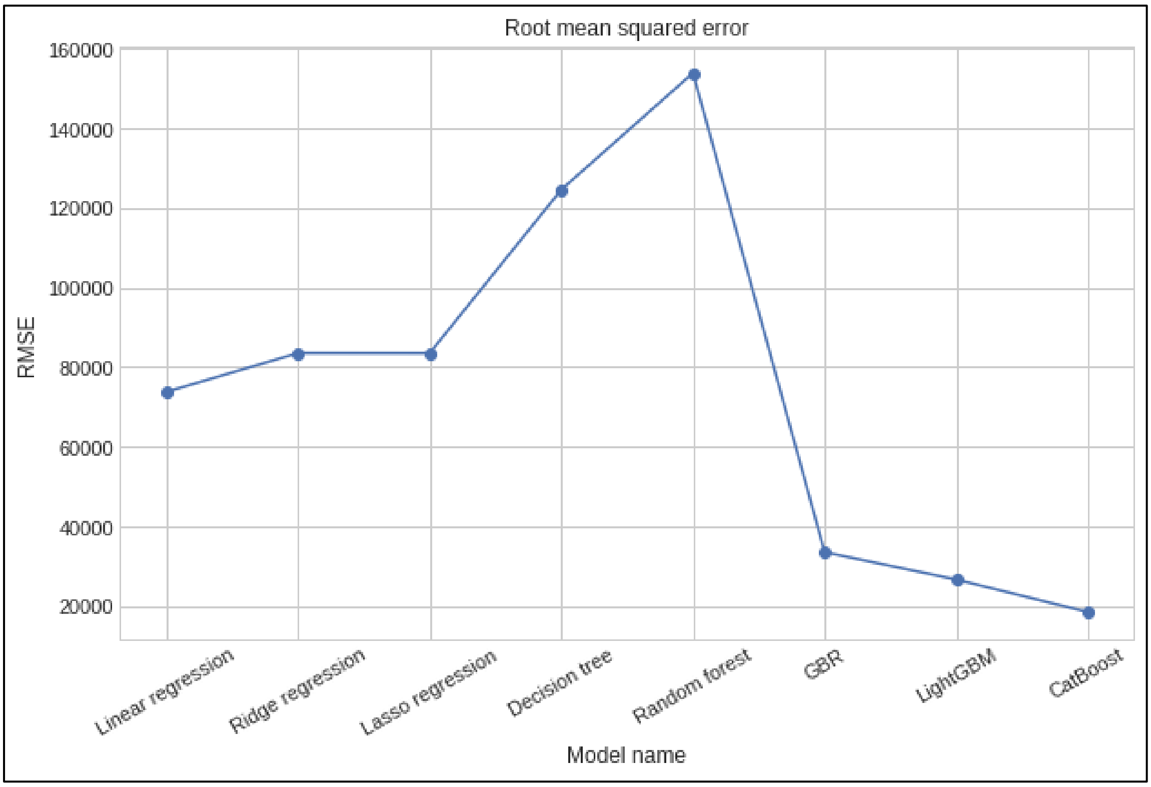


Figure 14: RMSE of test dataset for different models

As depicted in the graph above, the CatBoost model was chosen on the basis of its relatively better test RMSE.

The reason for choosing RMSE was the preference of penalizing large errors as the company would not want to have large errors however, smaller errors would not be of much problem. Other reason for choosing CatBoost or a tree-based model was the presence of multicollinearity.

Insights and Recommendations:

1. Analysis began with 28 features and the analysis was done with 16 features as most of them were dropped due to redundancy of the information given by them. 2 features were removed because of the presence of missing values in them. These were University PHD and PHD specialization
2. SHAP values were used to explain the impact of each of the predictor on the outcome

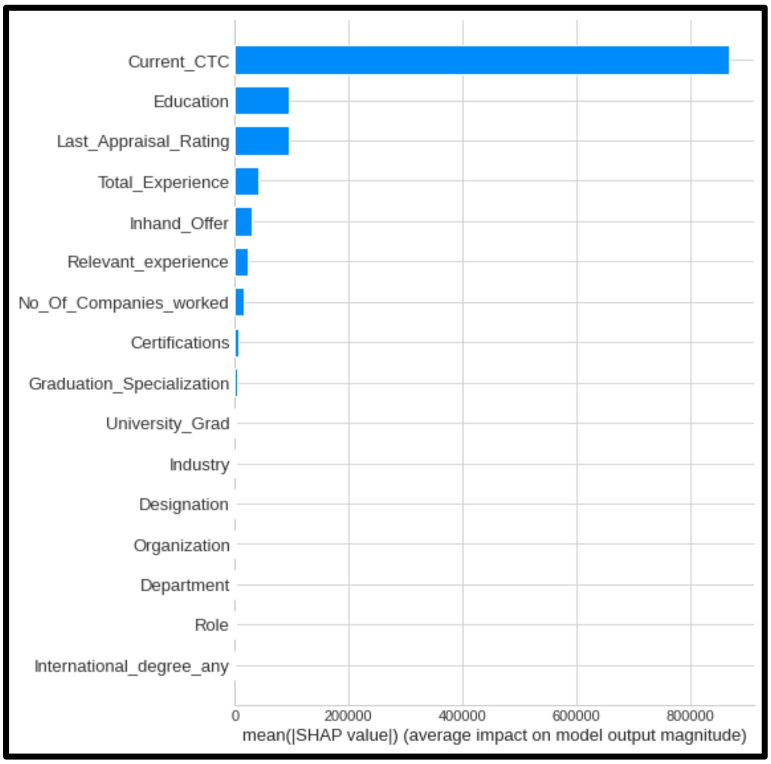


Figure 15:SHAP Values for the predictors

3. As expected, Current CTC was the most important predictor for CTC to be offered
4. Having an international degree doesn't have any impact on the expected CTC

Insights from Interaction effect between predictors:

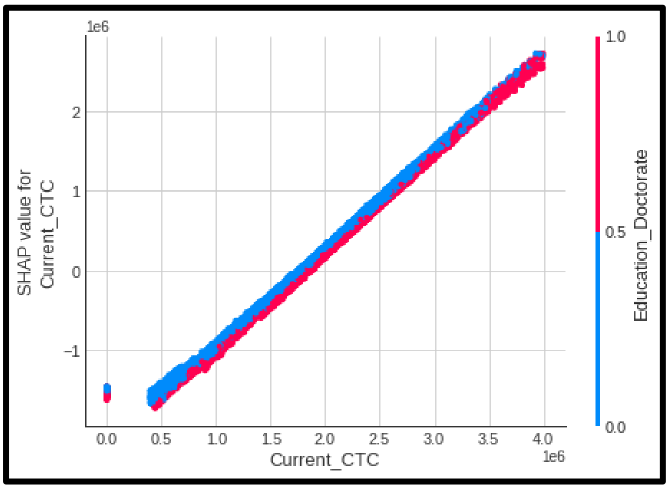


Figure 16:SHAP value plot of Current CTC with interaction effect of Education doctorate

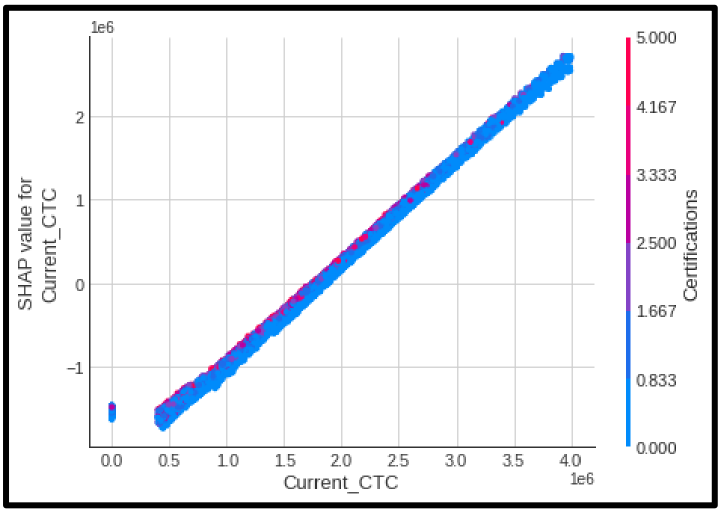


Figure 17:SHAP value plot of Current CTC with interaction effect of certifications

<u>How Current CTC is changing expected CTC for different set of candidates?</u>

- For PHD level candidates, the impact of current CTC on expected CTC is less for candidates having lower current CTC
- It is good to have done certification courses because having done certification courses brings more impact of the current CTC on the expected CTC
- It is not expected from high current CTC candidates and/or high total experience candidates to have last appraisal rating received as C or D

<u>How Relevant experience is changing expected CTC for different set of candidates?</u>

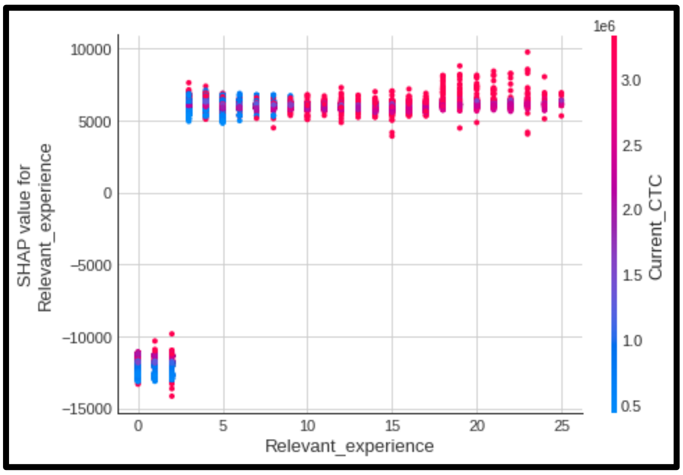


Figure 18:SHAP value plot of relevant experience with interaction effect of Current CTC

- Not having any relevant experience is not much of a concern and will not cause much negative impact on the increment on current CTC
- But to have significant positive impact the candidate must have at least 3 years of relevant experience

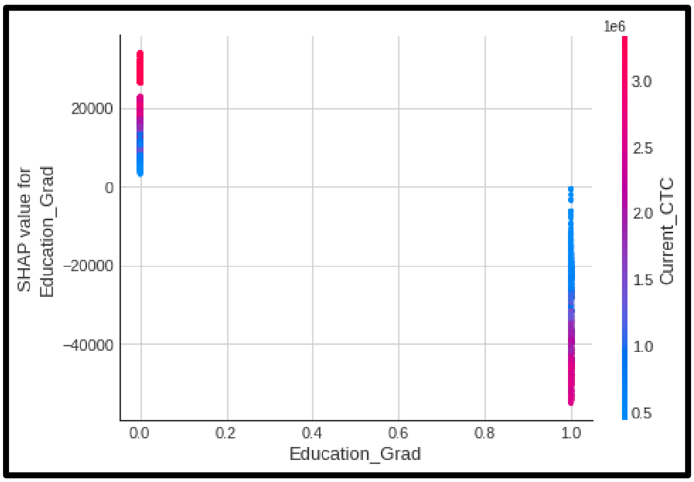


Figure 19:SHAP value plot of grad education with interaction effect of Current CTC

- If candidate is just a graduate even if the current CTC is very high the impact on expected CTC turns negative
- This impact reduces if the candidate has done some certifications

Inclination towards doctorate candidates:

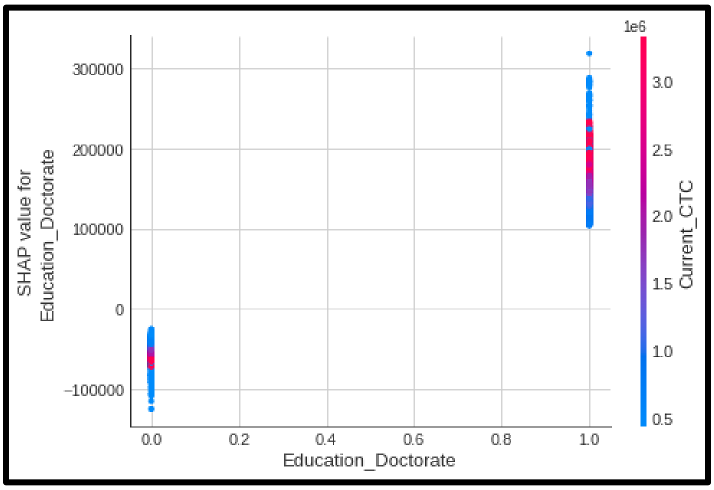


Figure 20:SHAP value plot of PHD education with interaction effect of Current CTC

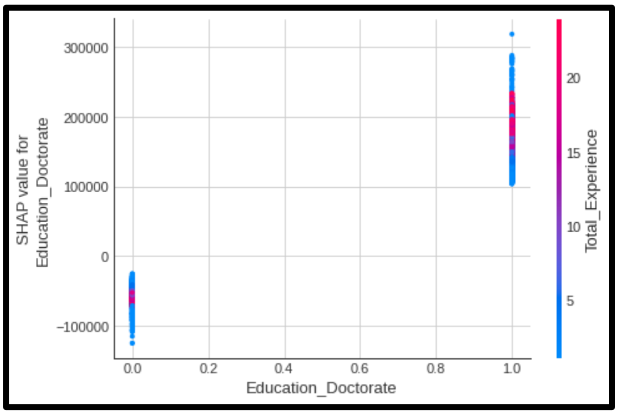


Figure 21:SHAP value plot of PHD education with interaction effect of Total experience

- PHD candidates are anticipated to have increment in the expected value of CTC to be offered from current CTC whereas the non-PHD candidates are expected to receive decrement in the expected value of CTC
- The total experience of the candidates is not of much importance if the candidate is PHD

Recommendations for the management:

1. Certifications could be taken as a potential parameter to judge candidates with relatively lower level of formal education. Evidently, freshers with certifications are able to secure good offer
2. The distribution of current location and preferred location is spread across many cities while not converging to just one city so company can offer remote location opportunities
3. Not much impact is being generated by learning the place of university of highest/all level of education. This might create some biasness basis the location so could get away with this feature
4. Interestingly, 24% of candidates with an offer in hand did not do graduation. Therefore, hiring team can be a bit flexible about the educational level requirements

Appendix:

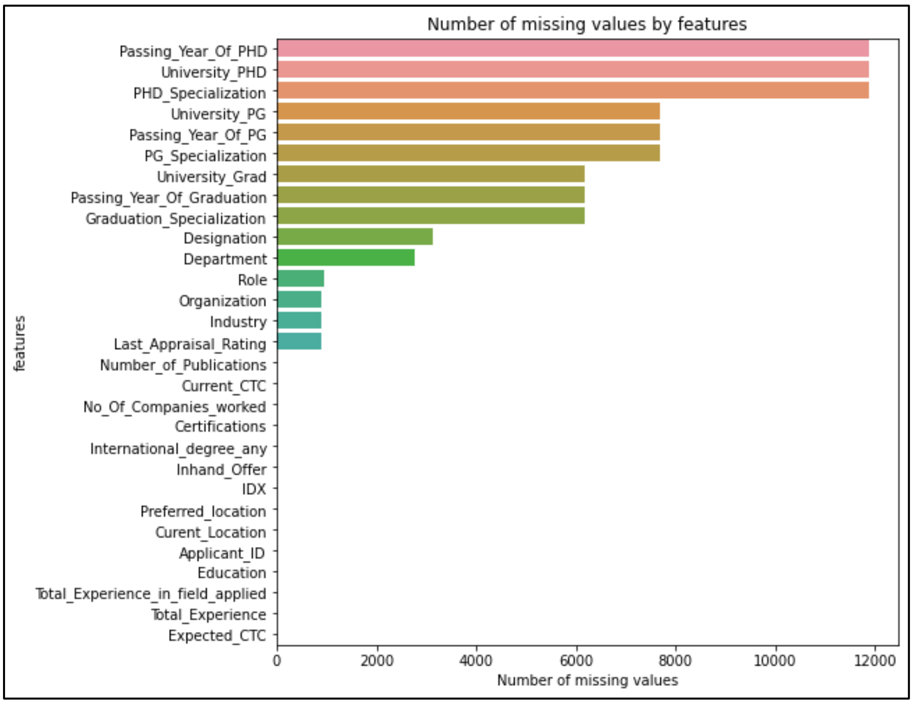


Figure 22:Null values in the dataset

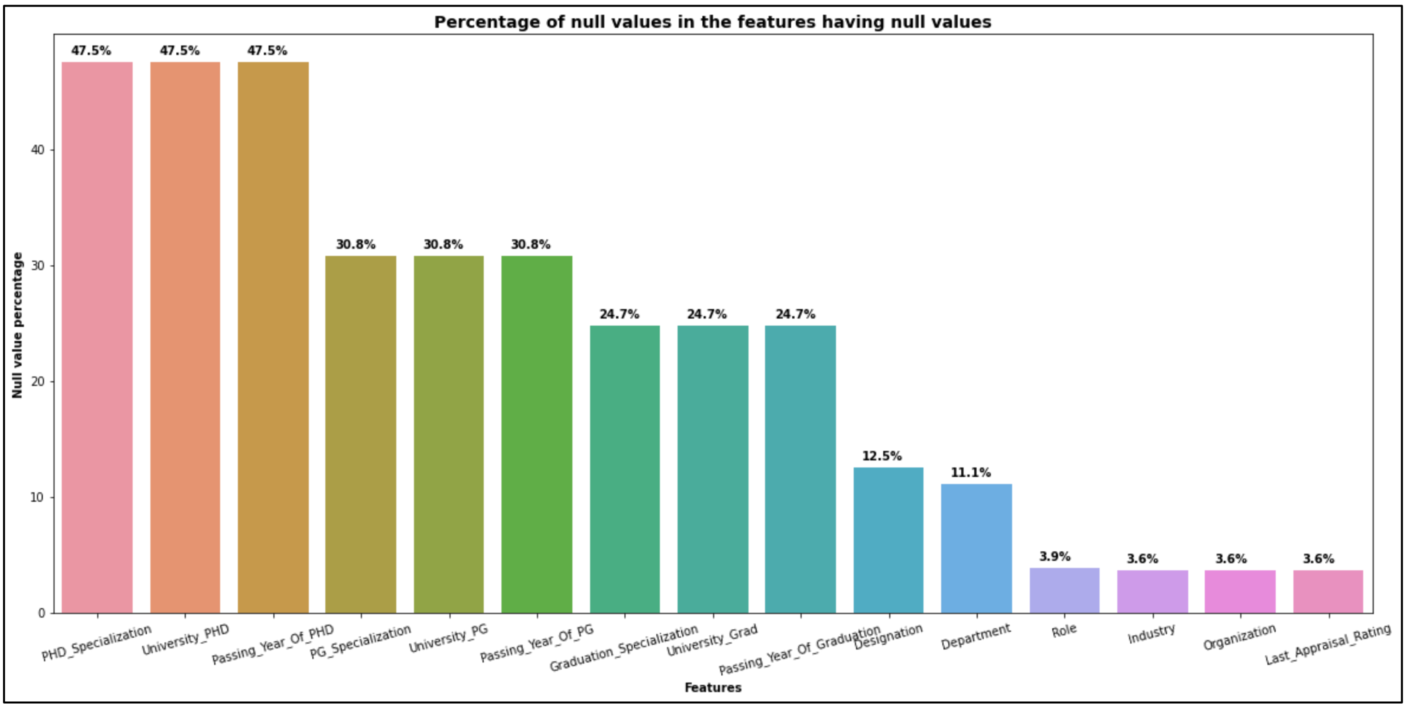
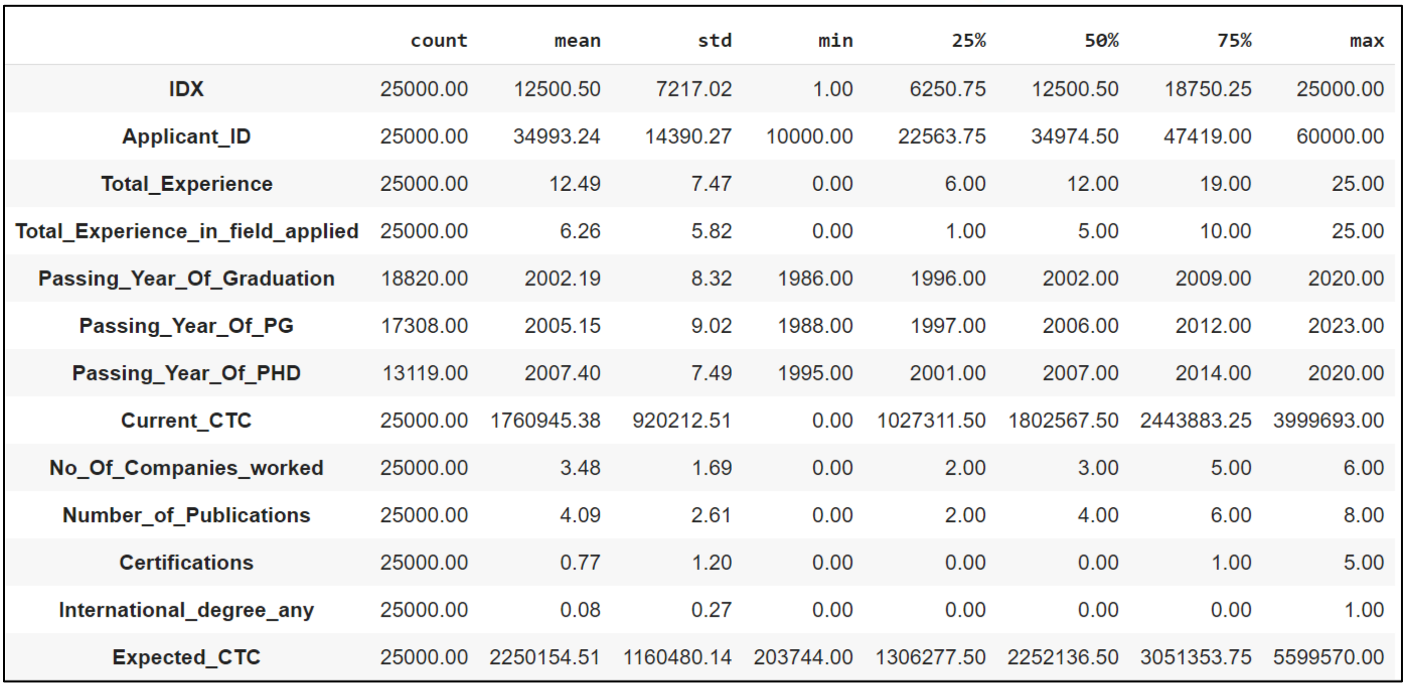
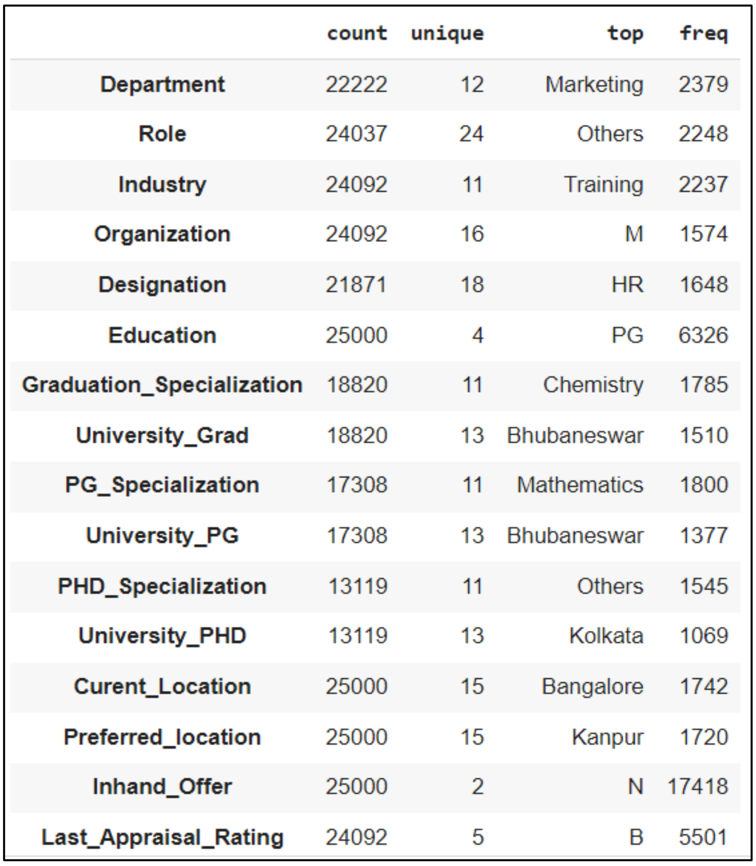


Figure 23:Percentage of null values in each null value feature



	count	mean	std	min	25%	50%	75%	max
IDX	25000.00	12500.50	7217.02	1.00	6250.75	12500.50	18750.25	25000.00
Applicant_ID	25000.00	34993.24	14390.27	10000.00	22563.75	34974.50	47419.00	60000.00
Total_Experience	25000.00	12.49	7.47	0.00	6.00	12.00	19.00	25.00
Total_Experience_in_field_applied	25000.00	6.26	5.82	0.00	1.00	5.00	10.00	25.00
Passing_Year_Of_Graduation	18820.00	2002.19	8.32	1986.00	1996.00	2002.00	2009.00	2020.00
Passing_Year_Of_PG	17308.00	2005.15	9.02	1988.00	1997.00	2006.00	2012.00	2023.00
Passing_Year_Of_PHD	13119.00	2007.40	7.49	1995.00	2001.00	2007.00	2014.00	2020.00
Current_CTC	25000.00	1760945.38	920212.51	0.00	1027311.50	1802567.50	2443883.25	3999693.00
No_Of_Companies_worked	25000.00	3.48	1.69	0.00	2.00	3.00	5.00	6.00
Number_of_Publications	25000.00	4.09	2.61	0.00	2.00	4.00	6.00	8.00
Certifications	25000.00	0.77	1.20	0.00	0.00	0.00	1.00	5.00
International_degree_any	25000.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
Expected_CTC	25000.00	2250154.51	1160480.14	203744.00	1306277.50	2252136.50	3051353.75	5599570.00

Table 11:Descriptive statistics of continuous variables



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

Table 12:Descriptive statistics of categorical variables

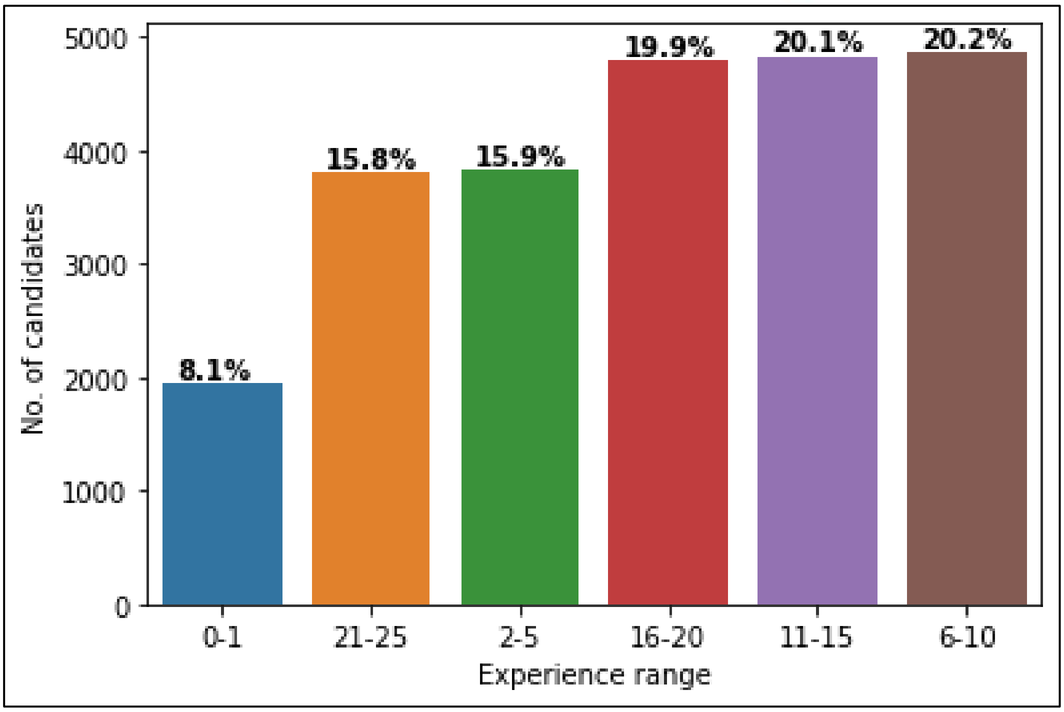


Figure 24:Percentage of candidates in each experience range

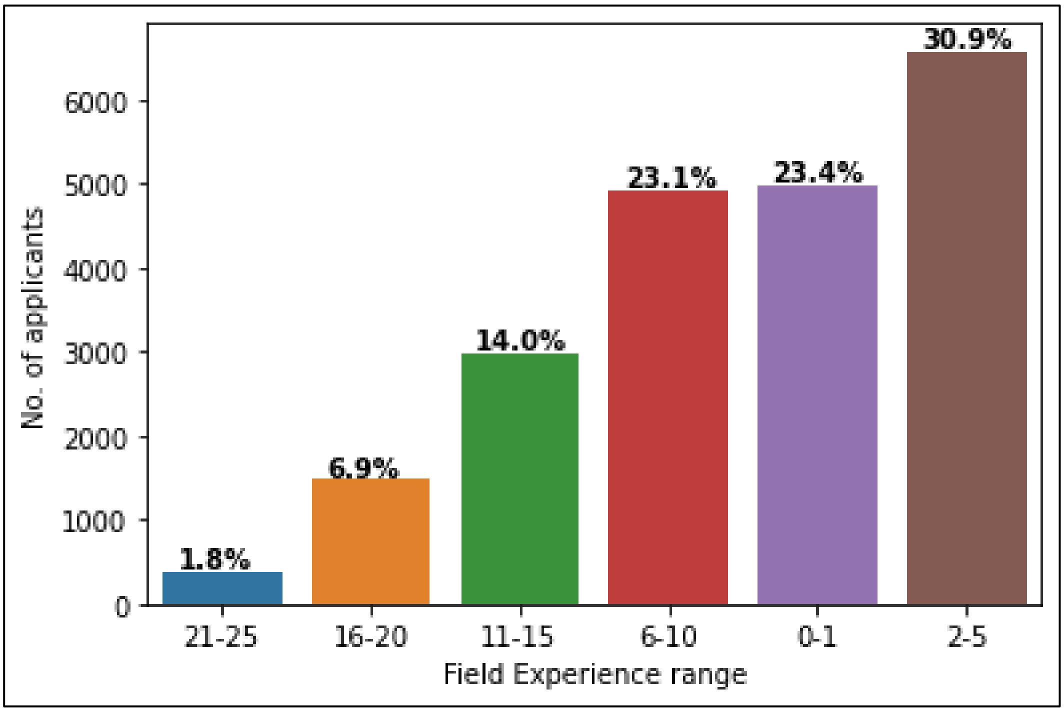
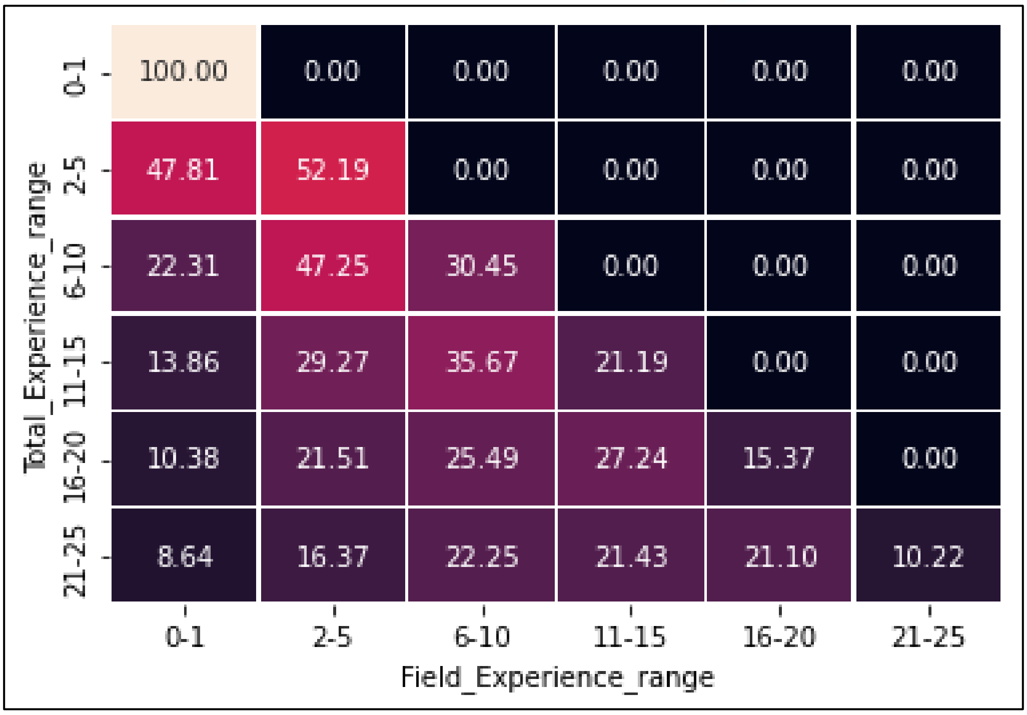


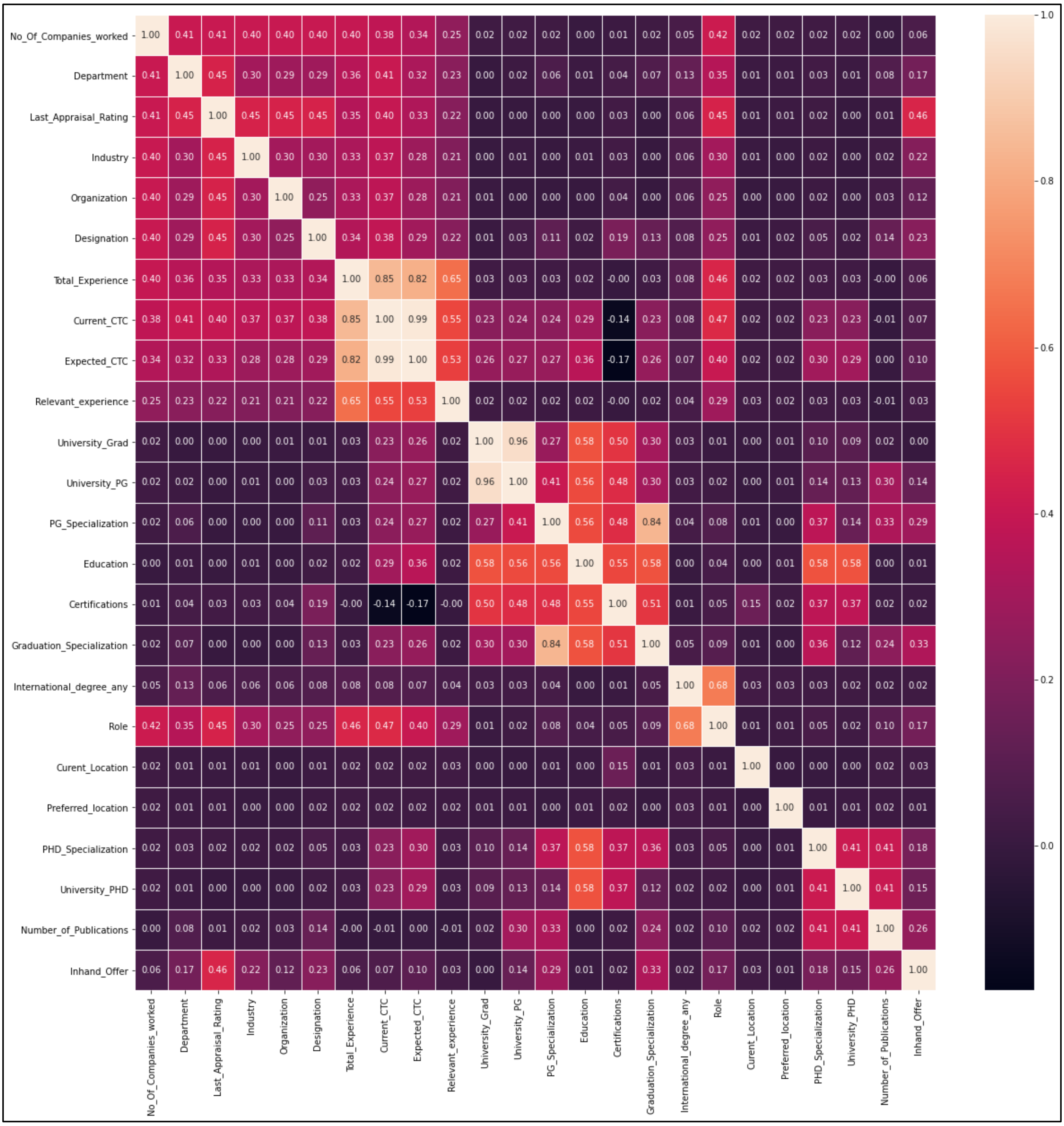
Figure 25:Percentage of candidates according to their relevant experience



Total_Experience_range	0-1	2-5	6-10	11-15	16-20	21-25
0-1	100.00	0.00	0.00	0.00	0.00	0.00
2-5	47.81	52.19	0.00	0.00	0.00	0.00
6-10	22.31	47.25	30.45	0.00	0.00	0.00
11-15	13.86	29.27	35.67	21.19	0.00	0.00
16-20	10.38	21.51	25.49	27.24	15.37	0.00
21-25	8.64	16.37	22.25	21.43	21.10	10.22

Table 13:Cross tab of total_experience_range and relevant_experience_range

➢



	No_Of_Companies_worked	Department	Last_Appraisal_Rating	Industry	Organization	Designation	Total_Experience	Current_CTC	Expected_CTC	Relevant_experience	University_Grad	University_PG	PG_Specialization	Education	Certifications	Graduation_Specialization	International_degree_any	Role	Curent_Location	Preferred_location	PHD_Specialization	University_PHD	Number_of_Publications	Inhand_Offer
No_Of_Companies_worked	1.00	0.41	0.41	0.40	0.40	0.40	0.40	0.38	0.34	0.25	0.02	0.02	0.02	0.00	0.01	0.02	0.05	0.42	0.02	0.02	0.02	0.02	0.00	0.06
Department	0.41	1.00	0.45	0.30	0.29	0.29	0.36	0.41	0.32	0.23	0.00	0.02	0.06	0.01	0.04	0.07	0.13	0.35	0.01	0.01	0.03	0.01	0.08	0.17
Last_Appraisal_Rating	0.41	0.45	1.00	0.45	0.45	0.45	0.35	0.40	0.33	0.22	0.00	0.00	0.00	0.00	0.03	0.00	0.06	0.45	0.01	0.01	0.02	0.00	0.01	0.46
Industry	0.40	0.30	0.45	1.00	0.30	0.30	0.33	0.37	0.28	0.21	0.00	0.01	0.00	0.01	0.03	0.00	0.06	0.30	0.01	0.00	0.02	0.00	0.02	0.22
Organization	0.40	0.29	0.45	0.30	1.00	0.25	0.33	0.37	0.28	0.21	0.01	0.00	0.00	0.00	0.04	0.00	0.06	0.25	0.00	0.00	0.02	0.00	0.03	0.12
Designation	0.40	0.29	0.45	0.30	0.25	1.00	0.34	0.38	0.29	0.22	0.01	0.03	0.11	0.02	0.19	0.13	0.08	0.25	0.01	0.02	0.05	0.02	0.14	0.23
Total_Experience	0.40	0.36	0.35	0.33	0.33	0.34	1.00	0.85	0.82	0.65	0.03	0.03	0.03	0.02	-0.00	0.03	0.08	0.46	0.02	0.02	0.03	0.03	-0.00	0.06
Current_CTC	0.38	0.41	0.40	0.37	0.37	0.38	0.85	1.00	0.99	0.55	0.23	0.24	0.24	0.29	-0.14	0.23	0.08	0.47	0.02	0.02	0.23	0.23	-0.01	0.07
Expected_CTC	0.34	0.32	0.33	0.28	0.28	0.29	0.82	0.99	1.00	0.53	0.26	0.27	0.27	0.36	-0.17	0.26	0.07	0.40	0.02	0.02	0.30	0.29	0.00	0.10
Relevant_experience	0.25	0.23	0.22	0.21	0.21	0.22	0.65	0.55	0.53	1.00	0.02	0.02	0.02	0.02	-0.00	0.02	0.04	0.29	0.03	0.02	0.03	0.03	-0.01	0.03
University_Grad	0.02	0.00	0.00	0.00	0.01	0.01	0.03	0.23	0.26	0.02	1.00	0.96	0.27	0.58	0.50	0.30	0.03	0.01	0.00	0.01	0.10	0.09	0.02	0.00
University_PG	0.02	0.02	0.00	0.01	0.00	0.03	0.03	0.24	0.27	0.02	0.96	1.00	0.41	0.56	0.48	0.30	0.03	0.02	0.00	0.01	0.14	0.13	0.30	0.14
PG_Specialization	0.02	0.06	0.00	0.00	0.00	0.11	0.03	0.24	0.27	0.02	0.27	0.41	1.00	0.56	0.48	0.84	0.04	0.08	0.01	0.00	0.37	0.14	0.33	0.29
Education	0.00	0.01	0.00	0.01	0.00	0.02	0.02	0.29	0.36	0.02	0.58	0.56	0.56	1.00	0.55	0.58	0.00	0.04	0.00	0.01	0.58	0.58	0.00	0.01
Certifications	0.01	0.04	0.03	0.03	0.04	0.19	-0.00	-0.14	-0.17	-0.00	0.50	0.48	0.48	0.55	1.00	0.51	0.01	0.05	0.15	0.02	0.37	0.37	0.02	0.02
Graduation_Specialization	0.02	0.07	0.00	0.00	0.00	0.13	0.03	0.23	0.26	0.02	0.30	0.30	0.84	0.58	0.51	1.00	0.05	0.09	0.01	0.00	0.36	0.12	0.24	0.33
International_degree_any	0.05	0.13	0.06	0.06	0.06	0.08	0.08	0.08	0.07	0.04	0.03	0.03	0.04	0.00	0.01	0.05	1.00	0.68	0.03	0.03	0.03	0.02	0.02	0.02
Role	0.42	0.35	0.45	0.30	0.25	0.25	0.46	0.47	0.40	0.29	0.01	0.02	0.08	0.04	0.05	0.09	0.68	1.00	0.01	0.01	0.05	0.02	0.10	0.17
Curent_Location	0.02	0.01	0.01	0.01	0.00	0.01	0.02	0.02	0.02	0.03	0.00	0.00	0.01	0.00	0.15	0.01	0.03	0.01	1.00	0.00	0.00	0.00	0.02	0.03
Preferred_location	0.02	0.01	0.01	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.02	0.00	0.03	0.01	0.00	1.00	0.01	0.01	0.02	0.01
PHD_Specialization	0.02	0.03	0.02	0.02	0.02	0.05	0.03	0.23	0.30	0.03	0.10	0.14	0.37	0.58	0.37	0.36	0.03	0.05	0.00	0.01	1.00	0.41	0.41	0.18
University_PHD	0.02	0.01	0.00	0.00	0.00	0.02	0.03	0.23	0.29	0.03	0.09	0.13	0.14	0.58	0.37	0.12	0.02	0.02	0.00	0.01	0.41	1.00	0.41	0.15
Number_of_Publications	0.00	0.08	0.01	0.02	0.03	0.14	-0.00	-0.01	0.00	-0.01	0.02	0.30	0.33	0.00	0.02	0.24	0.02	0.10	0.02	0.02	0.41	0.41	1.00	0.26
Inhand_Offer	0.06	0.17	0.46	0.22	0.12	0.23	0.06	0.07	0.10	0.03	0.00	0.14	0.29	0.01	0.02	0.33	0.02	0.17	0.03	0.01	0.18	0.15	0.26	1.00

Figure 26:Correlation matrix for all variables

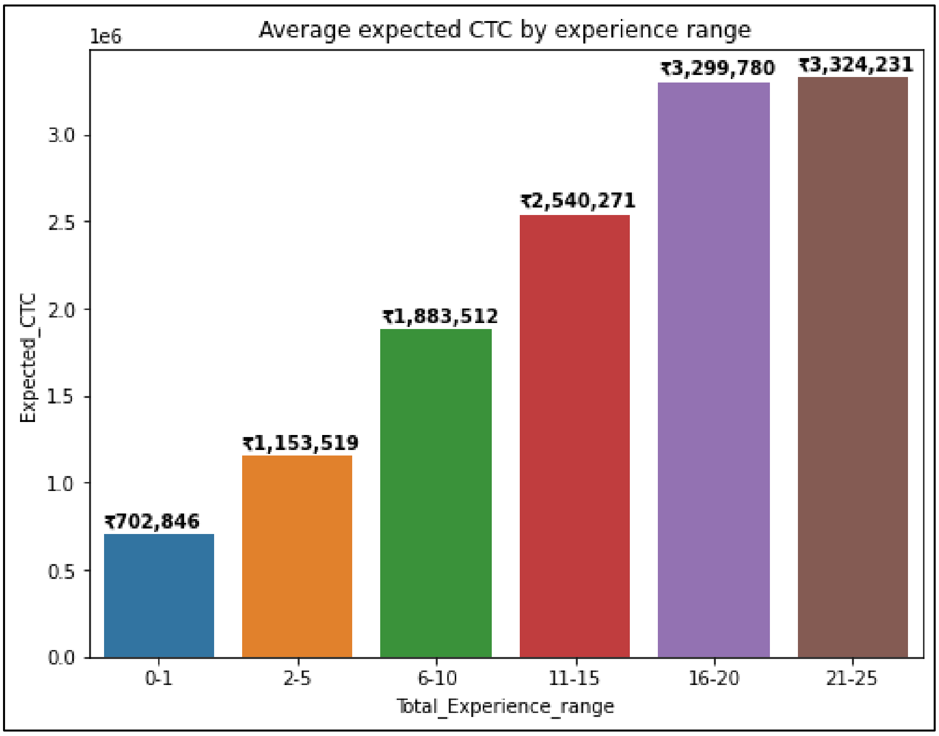


Figure 27:Average expected CTC by range of total experience

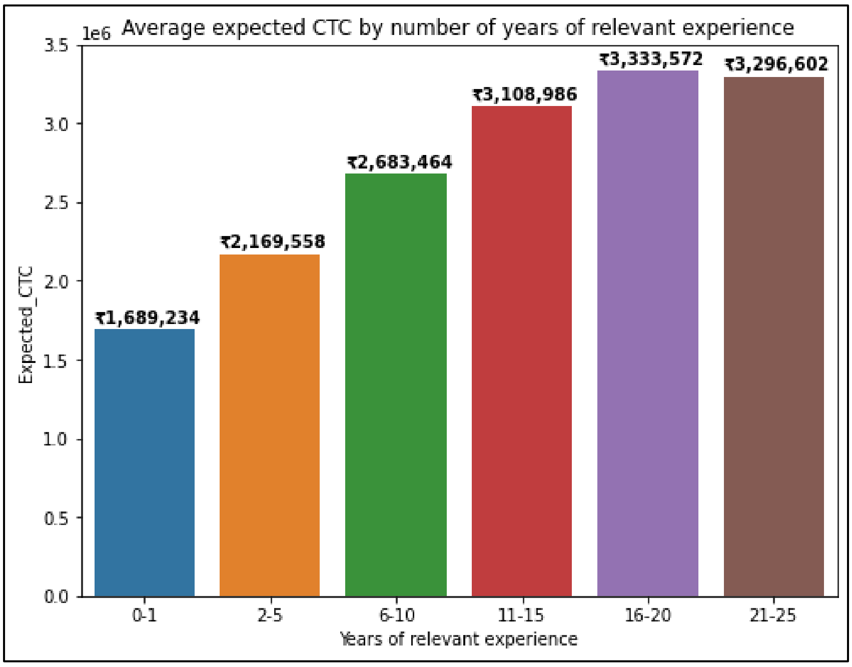


Figure 28:Average expected CTC by relevant experience range

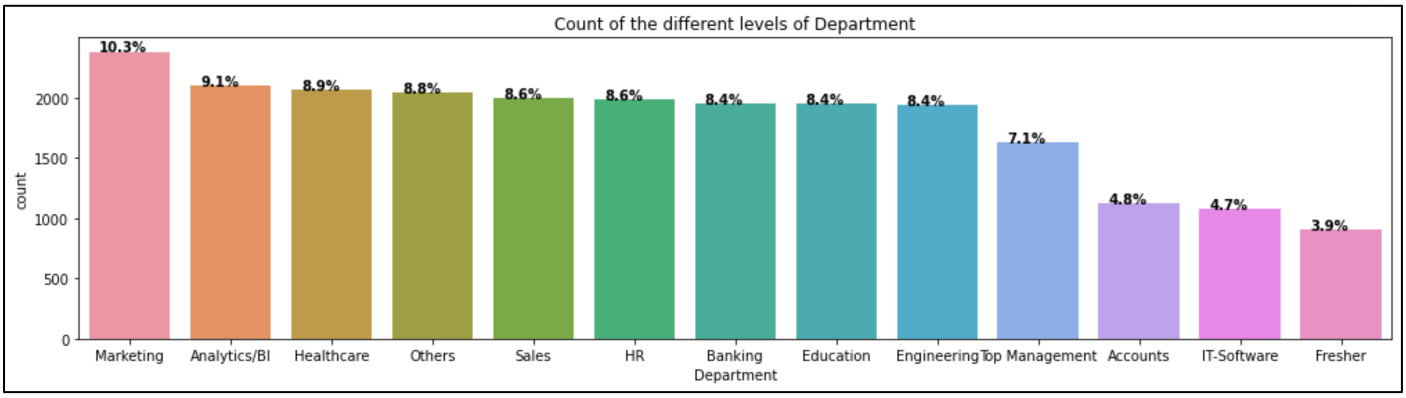


Figure 29:Percentage of applicants by department

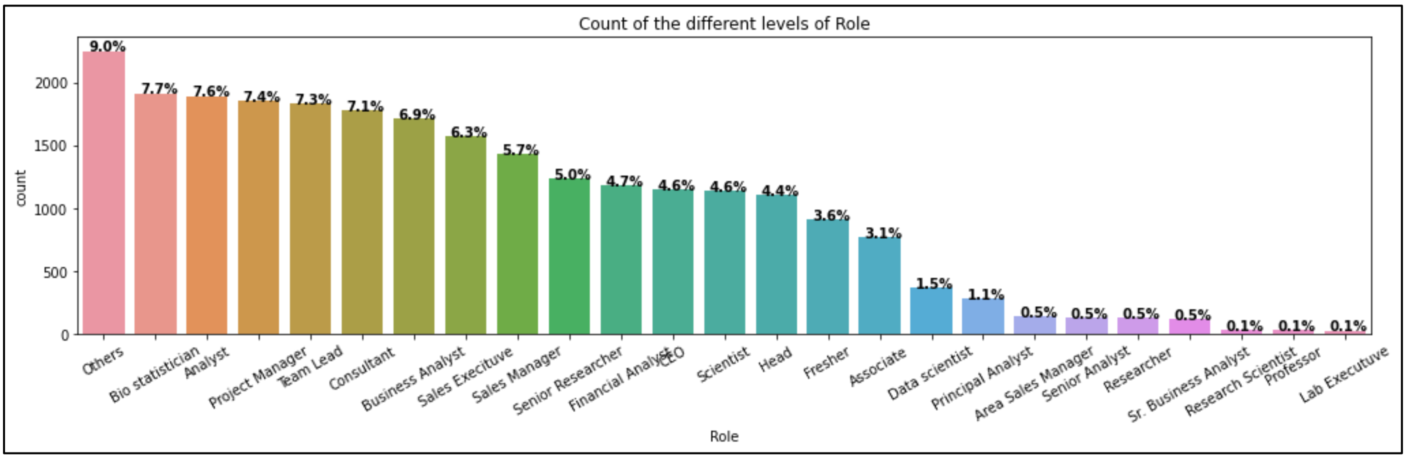


Figure 30:Percentage of applicants by role

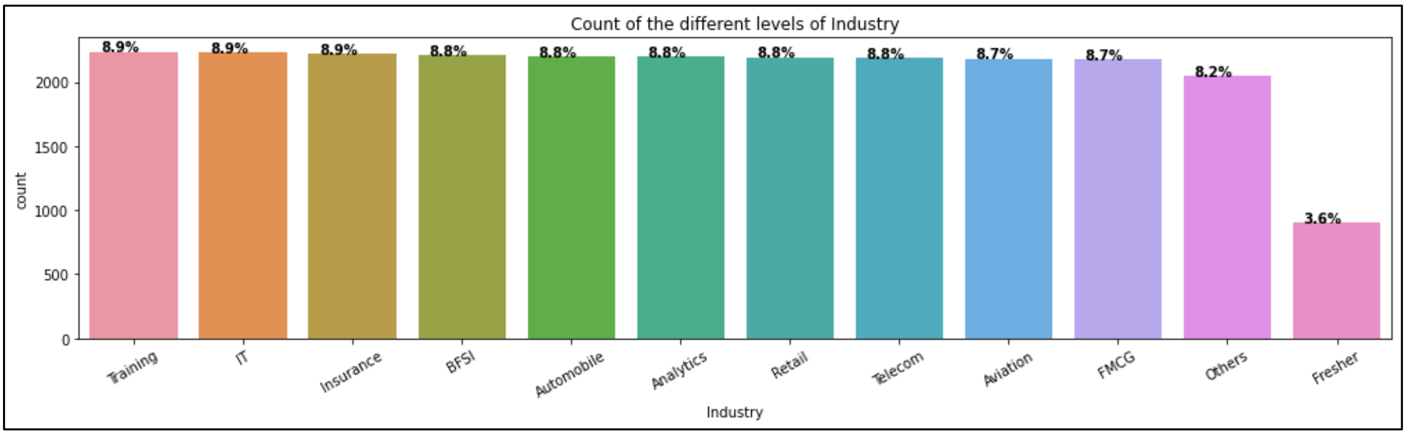


Figure 31:Percentage of applicants by industry

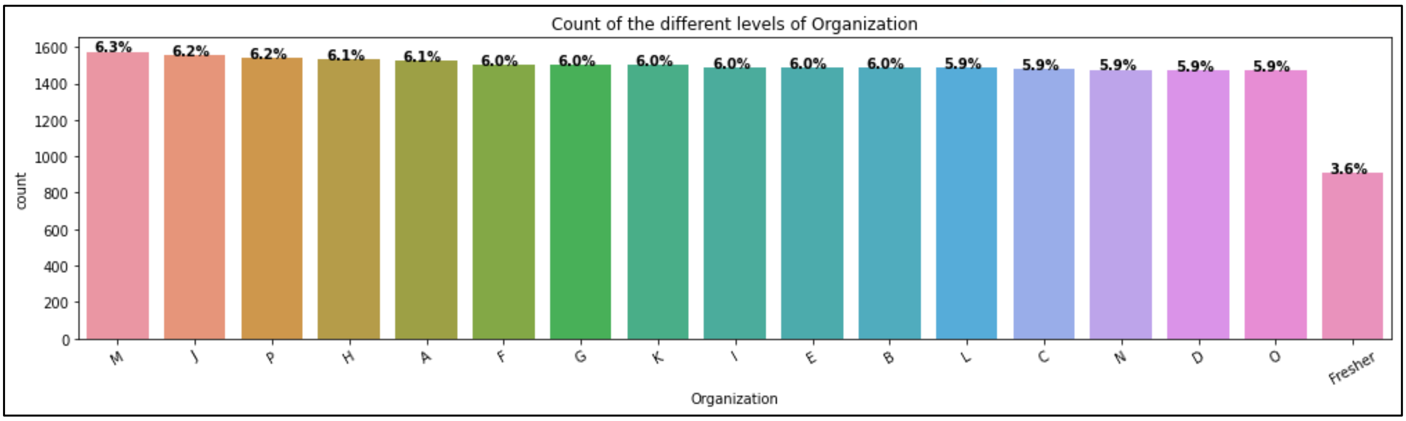


Figure 32:Percentage of applicants by organization

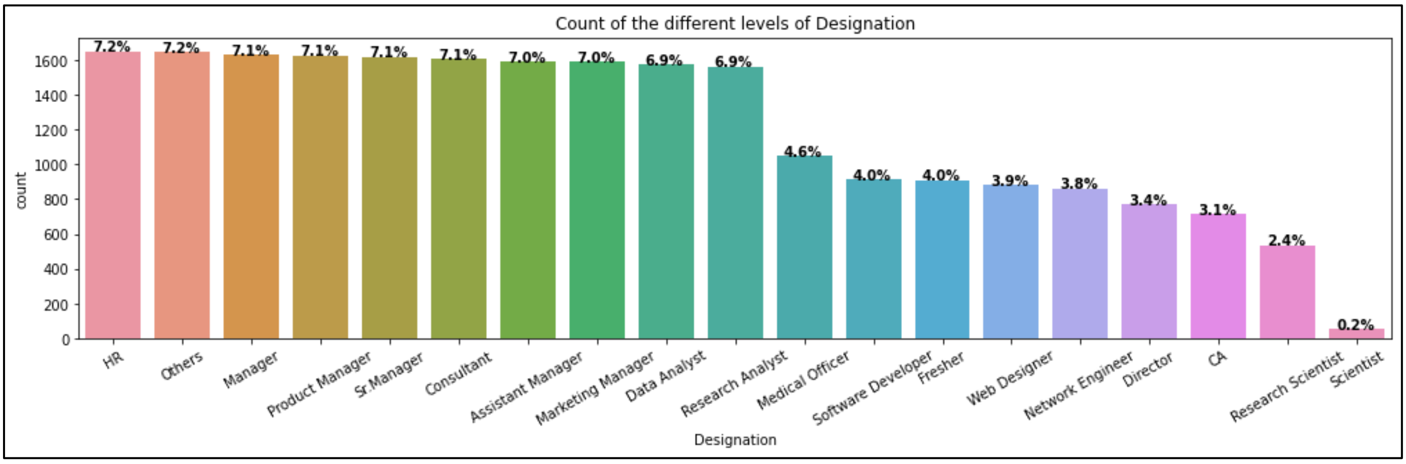


Figure 33:Percentage of applicants by designation

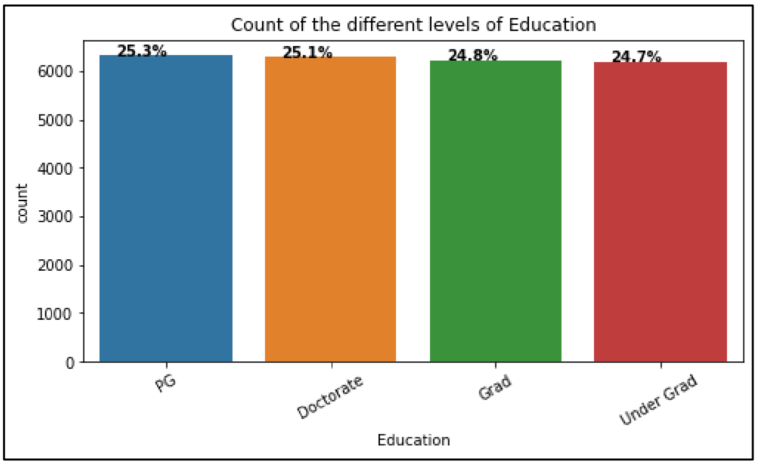


Figure 34:Percentage of applicants by level of education

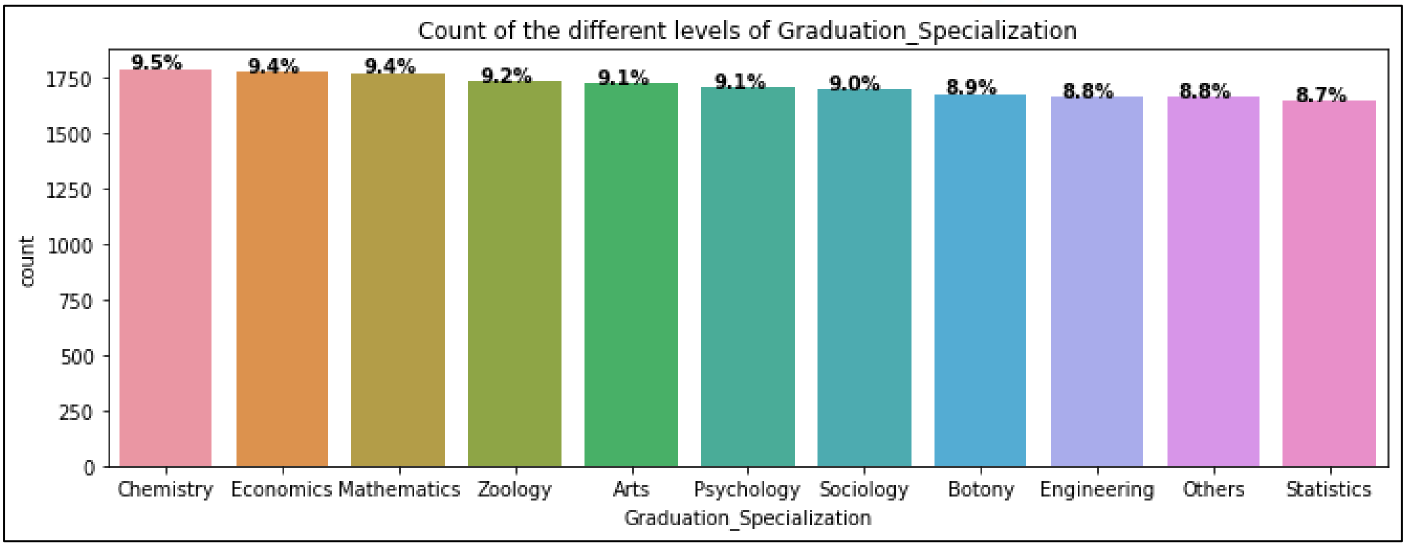


Figure 35:Percentage of applicants by graduation specialization

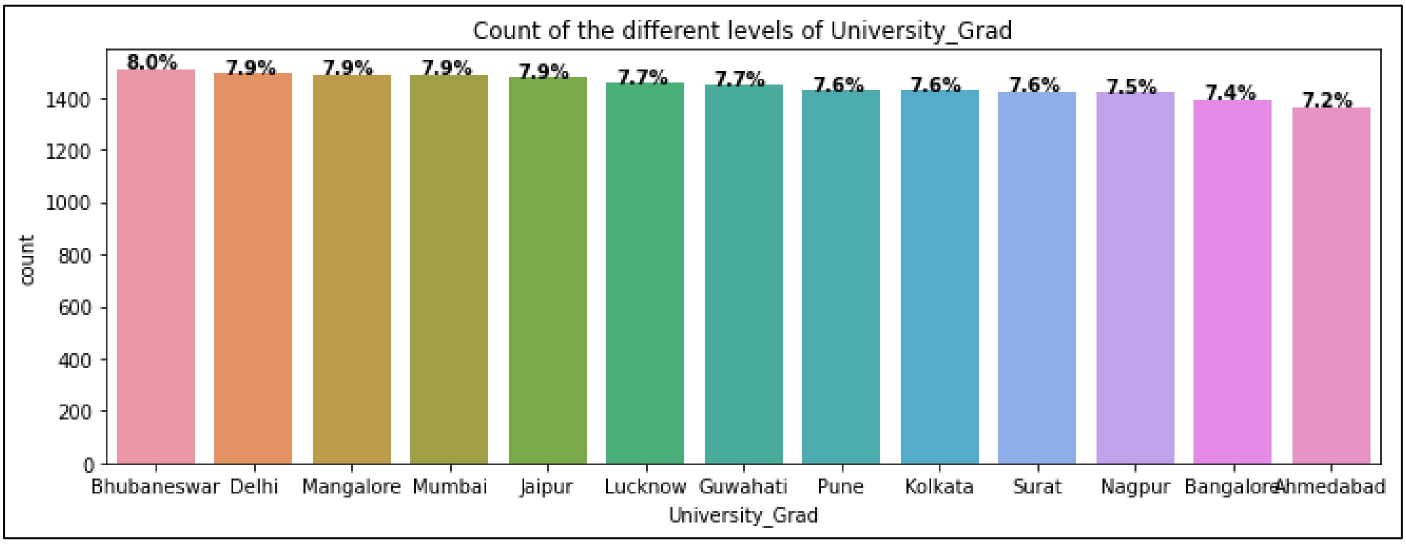


Figure 36:Percentage of applicants by graduation specialization

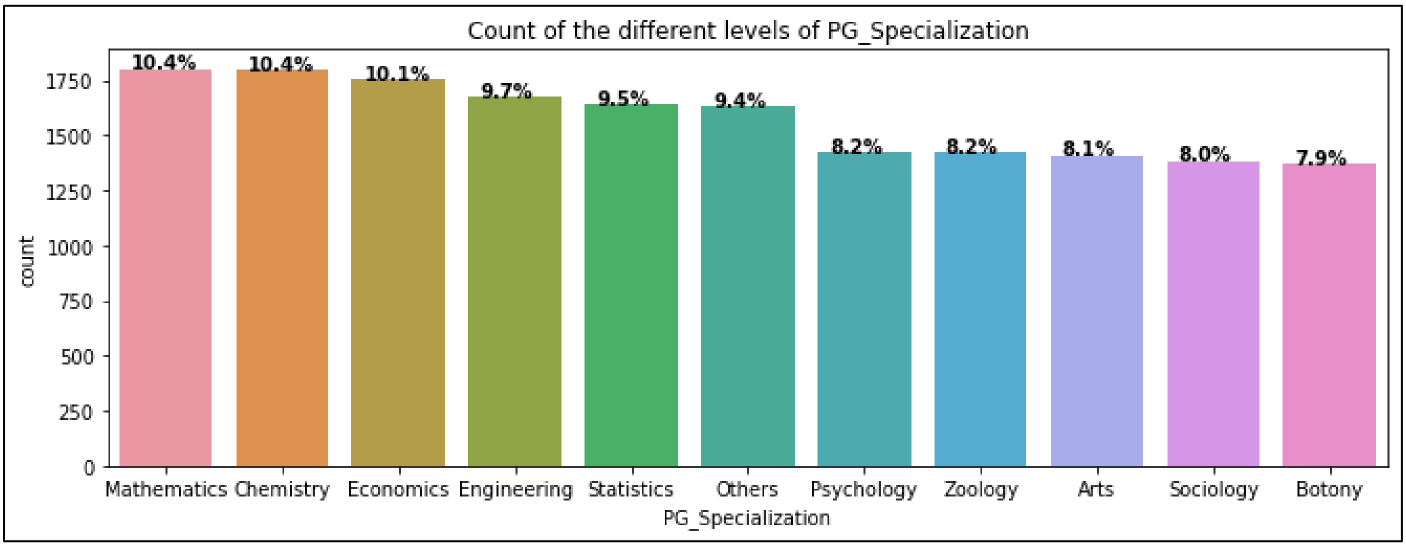


Figure 37:Percentage of applicants by subject of PG

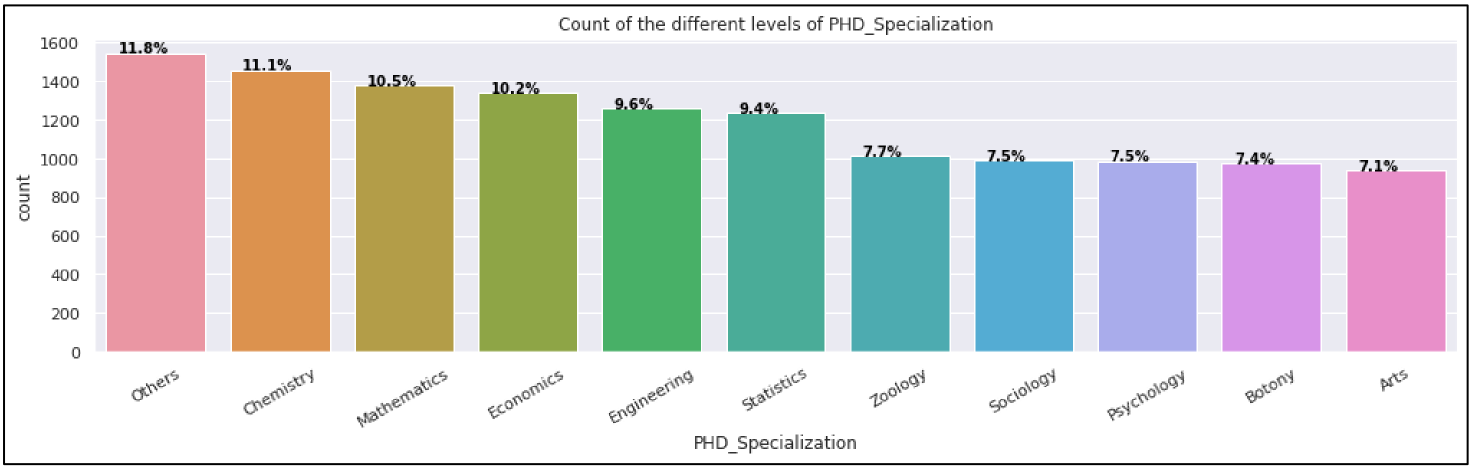


Figure 38:Percentage of applicants by subject of PHD

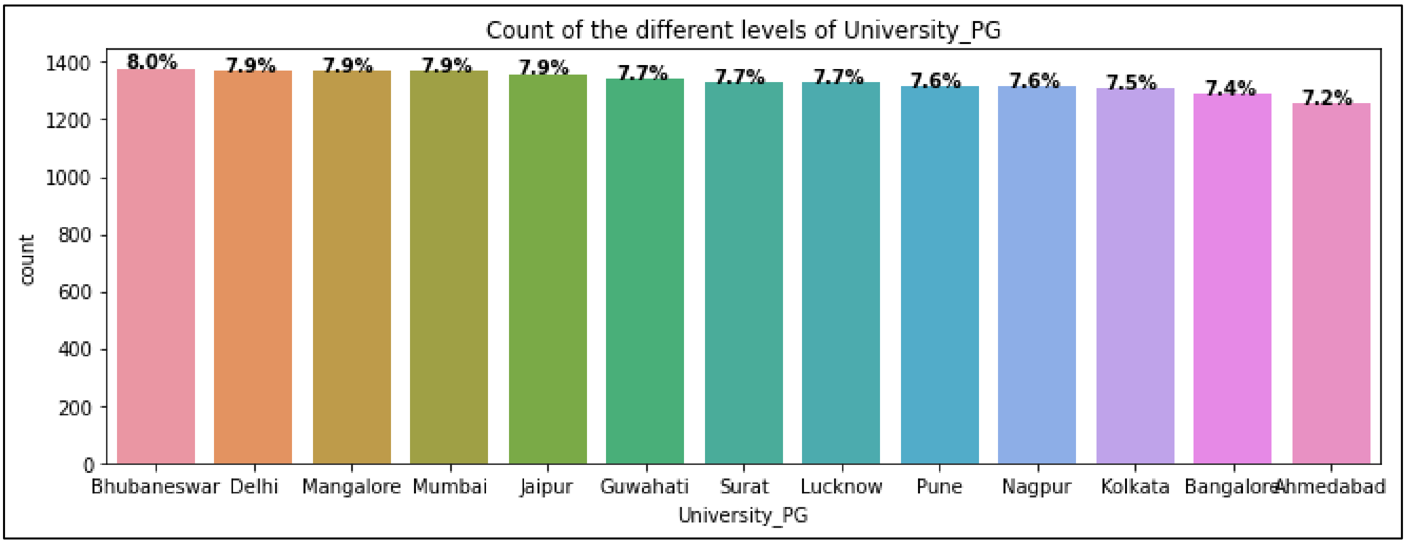


Figure 39:Percentage of applicants by subject of PHD

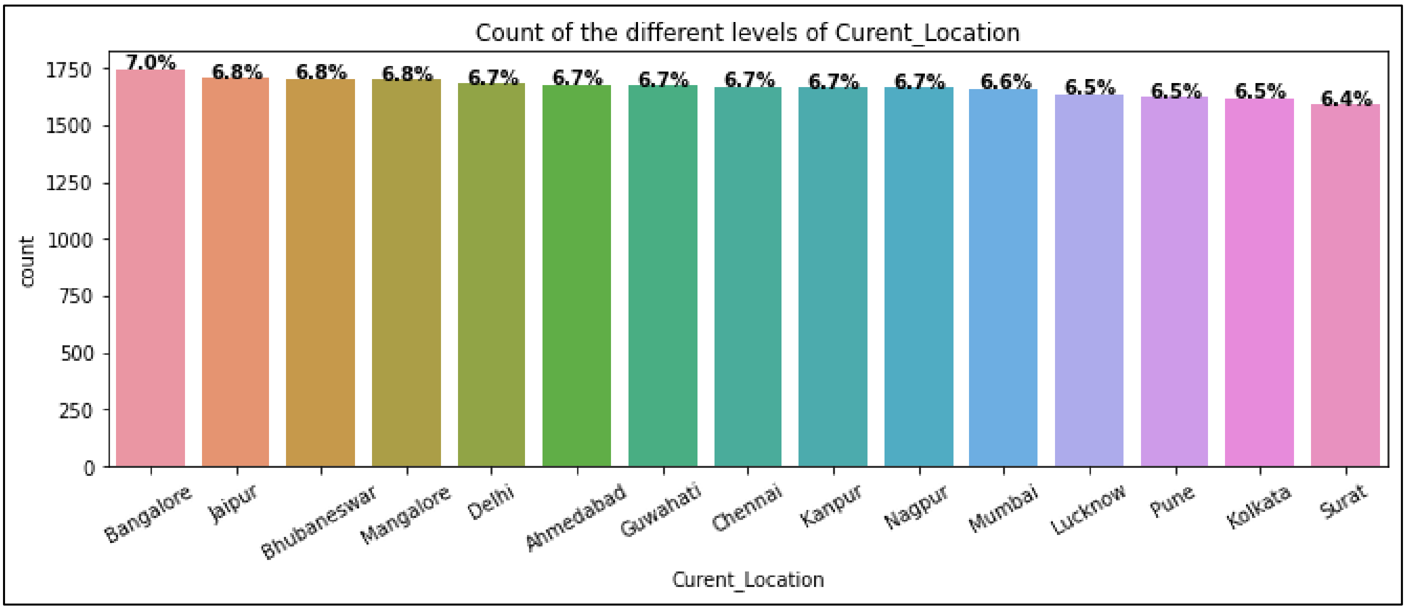


Figure 40:Percentage of applicants by current location

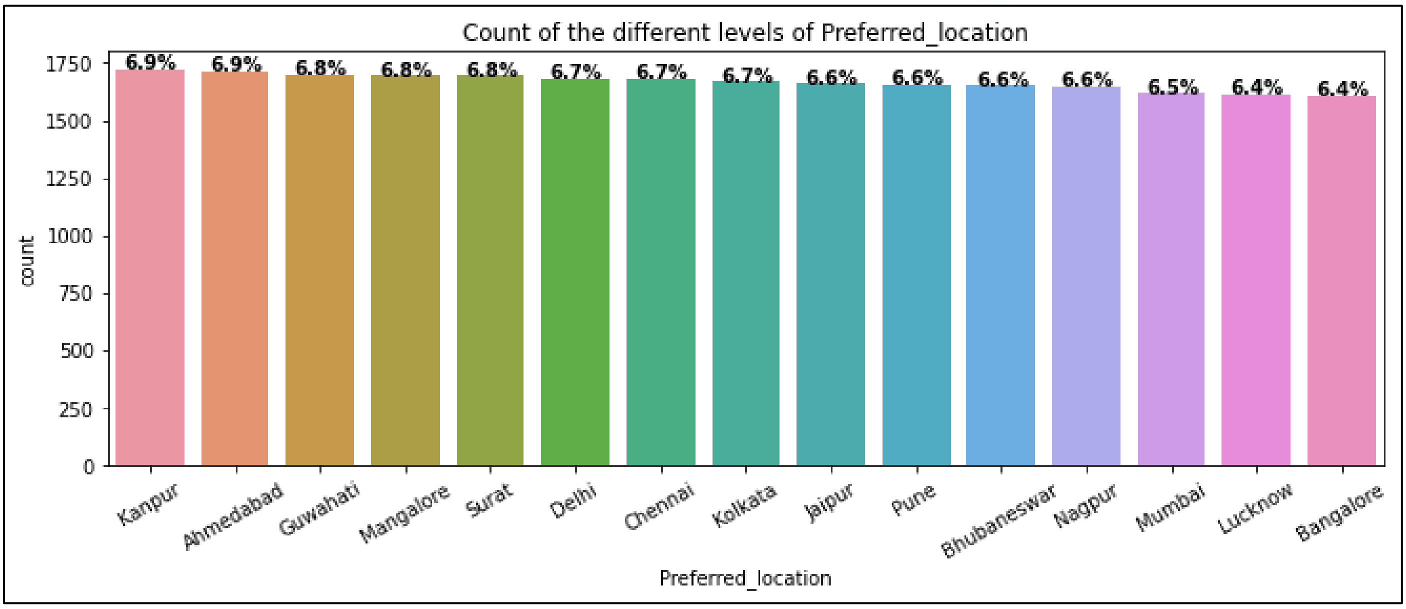


Figure 41:Percentage of applicants by preferred location

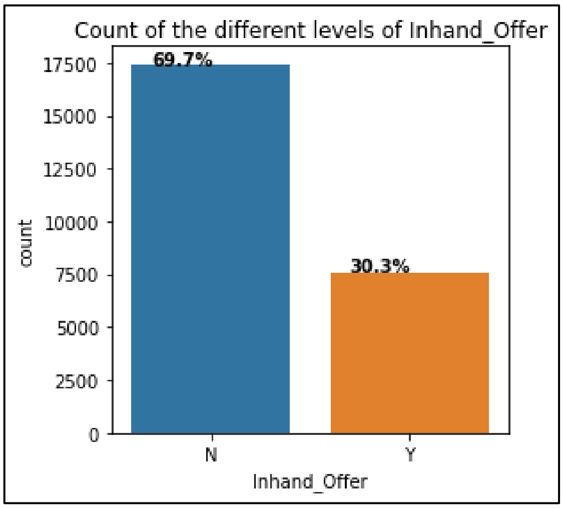


Figure 42:Percentage of applicants by availability of in-hand offer

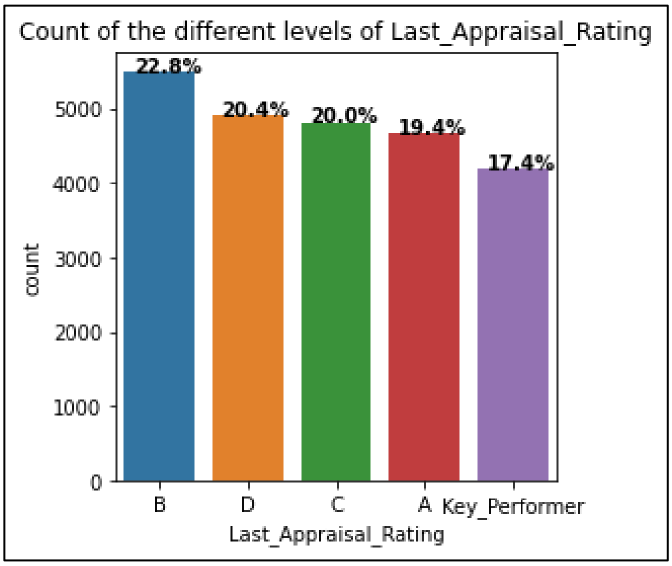


Figure 43:Percentage of applicants by last appraisal rating

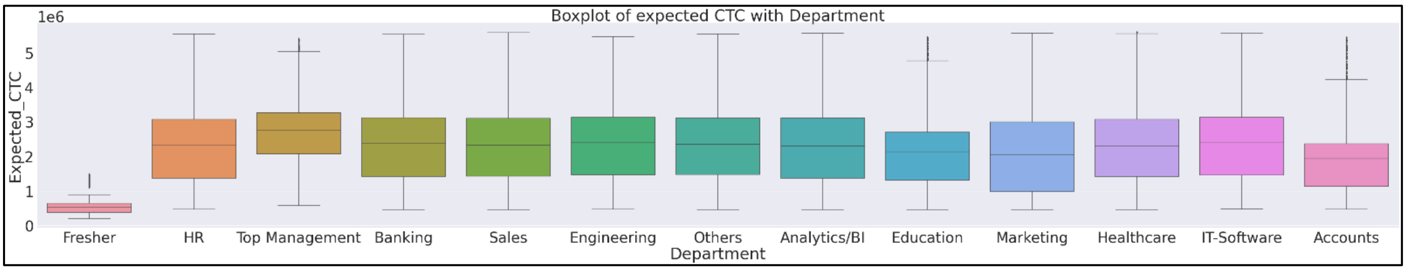


Figure 44:Distribution of expected CTC by department

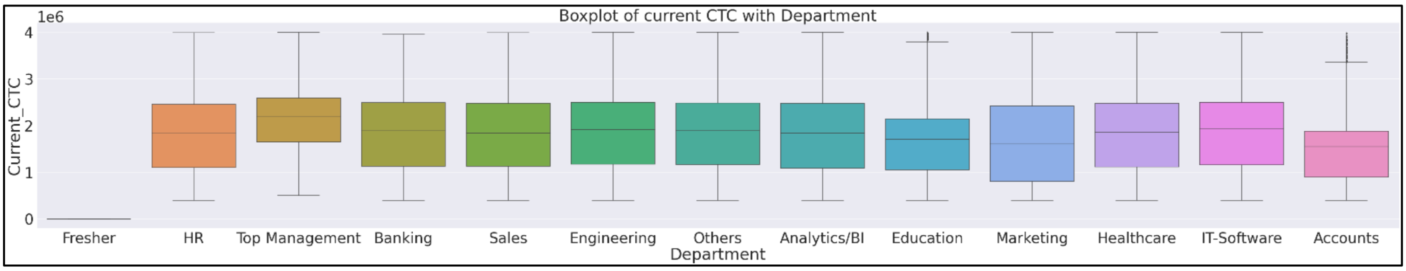


Figure 45:Distribution of current CTC by department

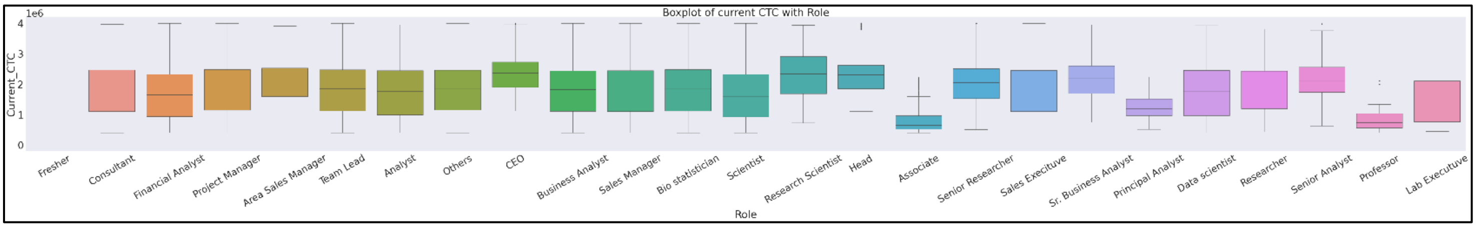


Figure 46:Distribution of expected CTC by role

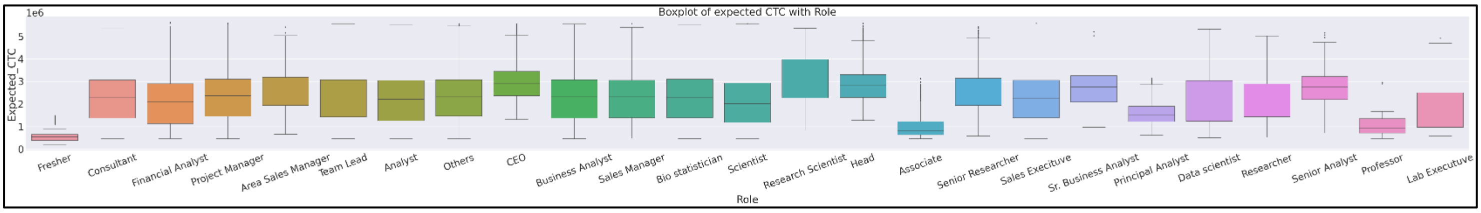


Figure 47:Distribution of current CTC by role

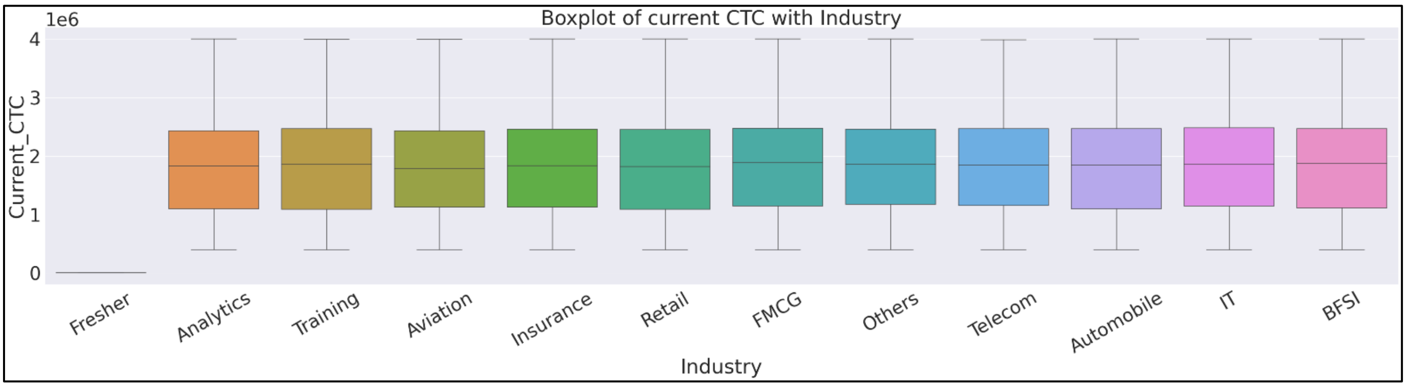


Figure 48:Distribution of current CTC by industry

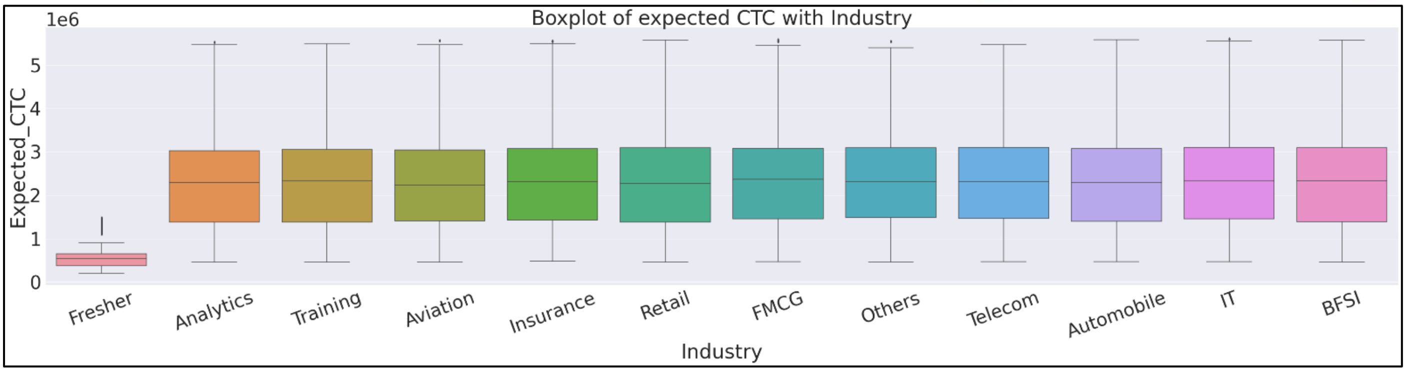


Figure 49:Distribution of expected CTC by industry

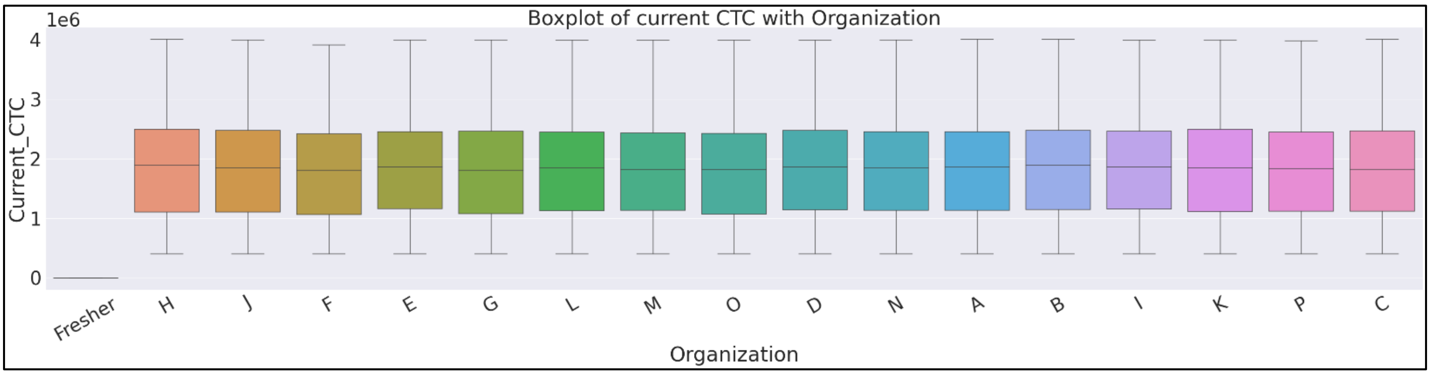


Figure 50:Distribution of current CTC by organization

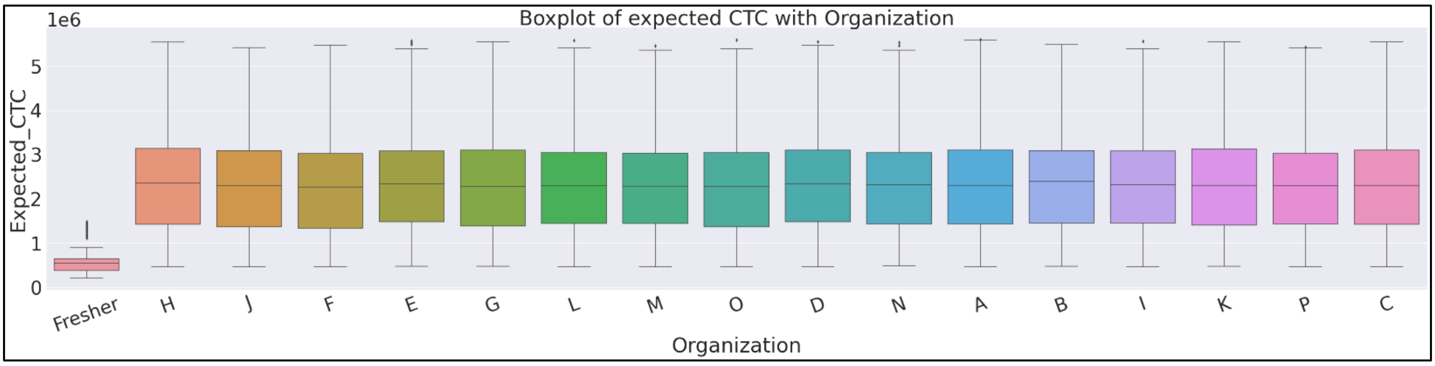


Figure 51:Distribution of expected CTC by organization

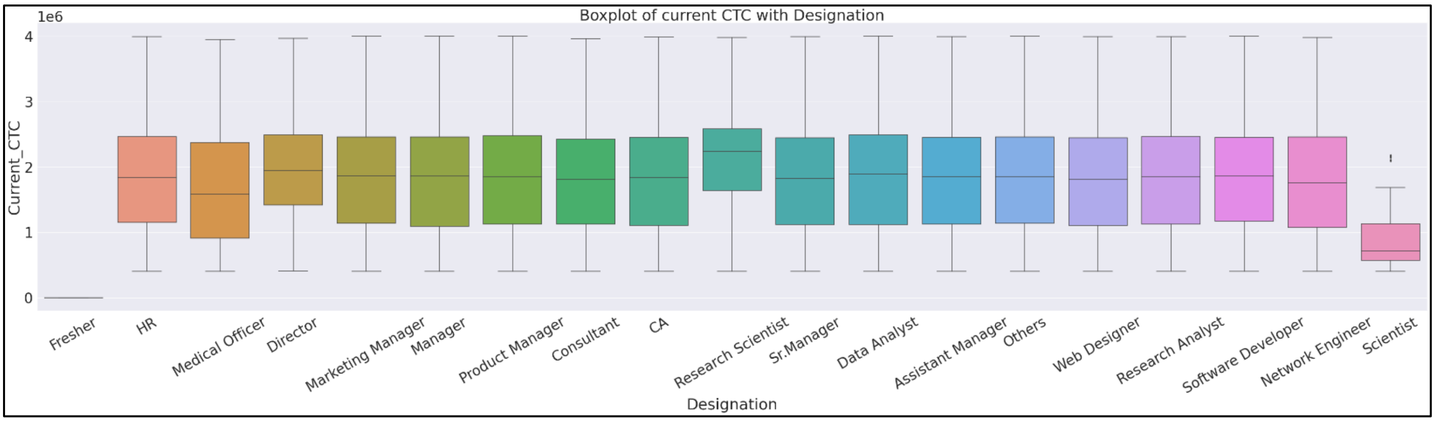


Figure 52:Distribution of current CTC by designation

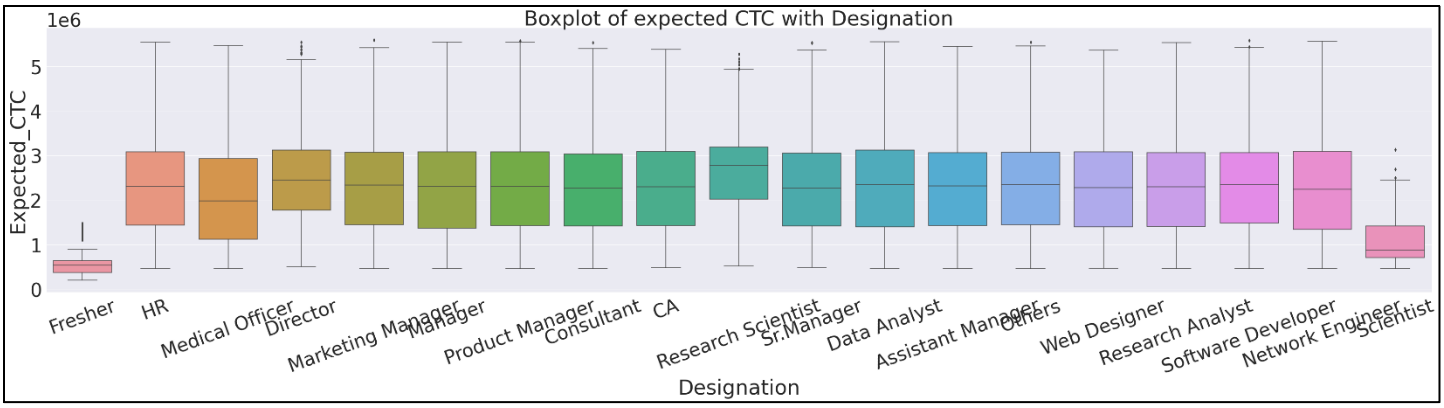


Figure 53:Distribution of expected CTC by designation

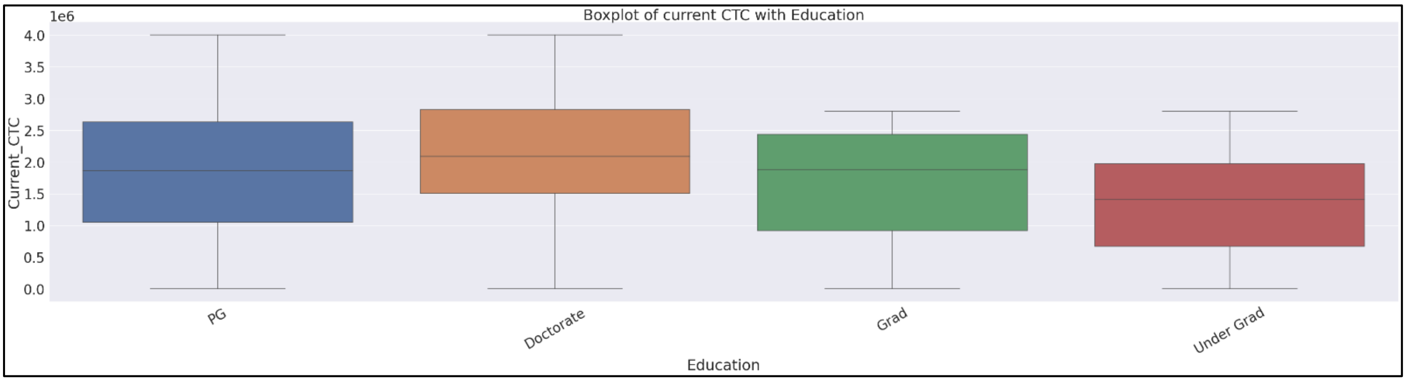


Figure 54:Distribution of current CTC by level of education

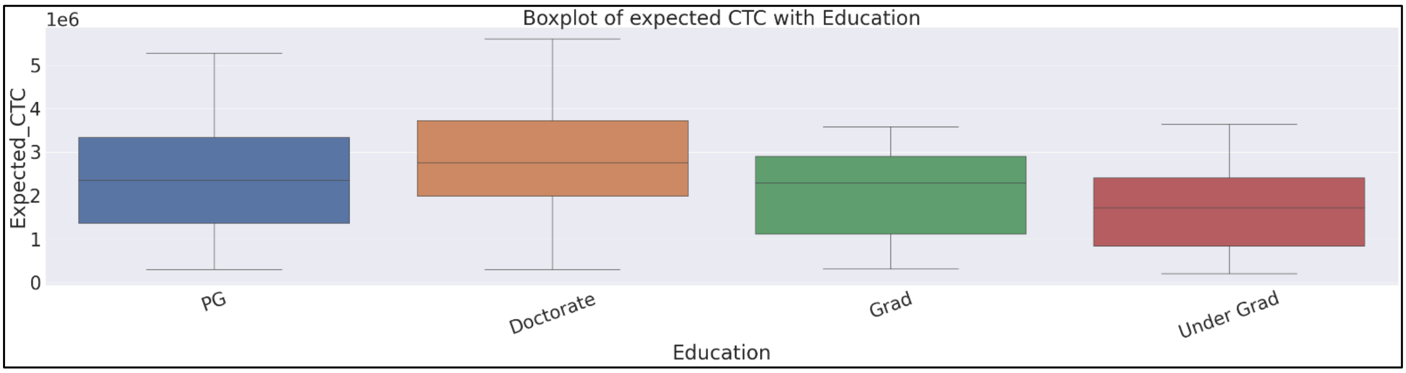


Figure 55:Distribution of expected CTC by level of education

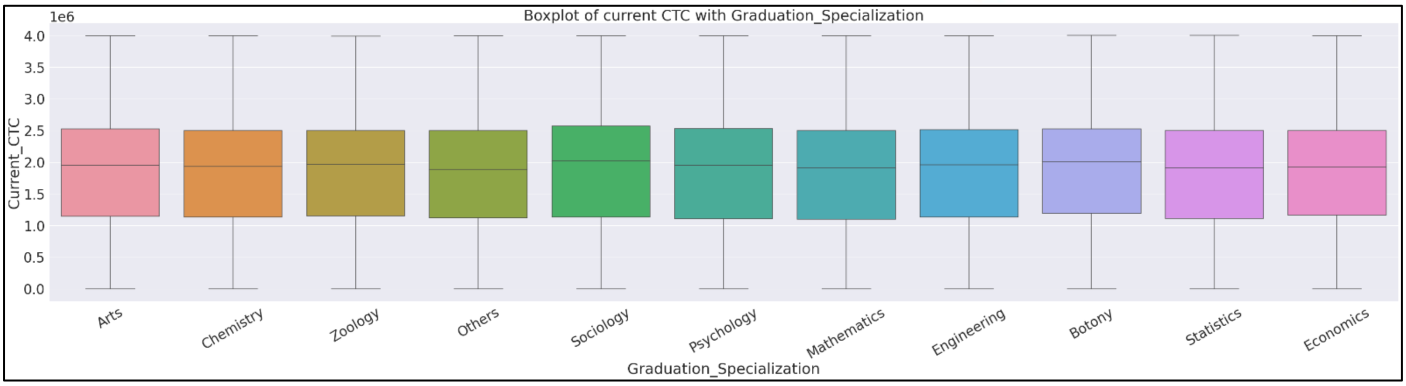


Figure 56:Distribution of current CTC by graduation specialization

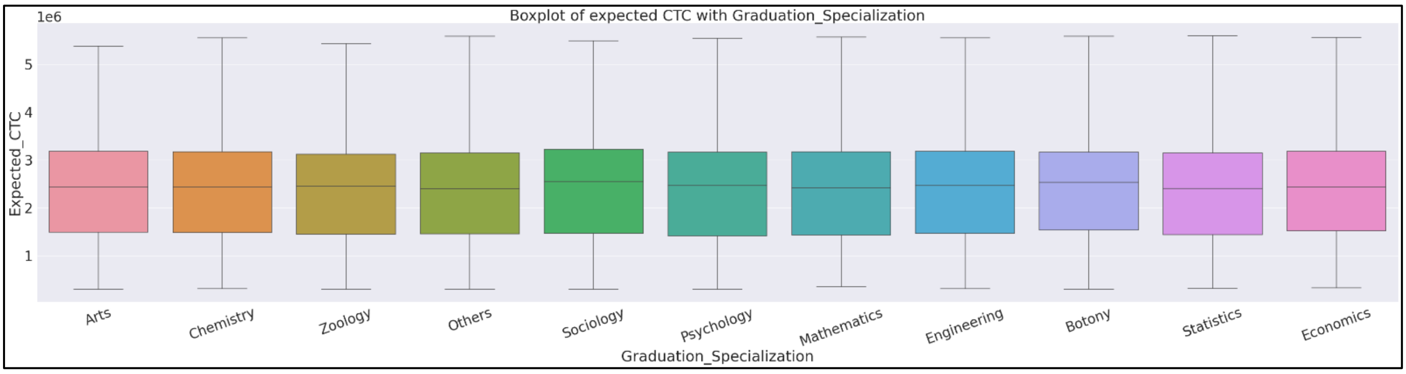


Figure 57:Distribution of expected CTC by graduation specialization

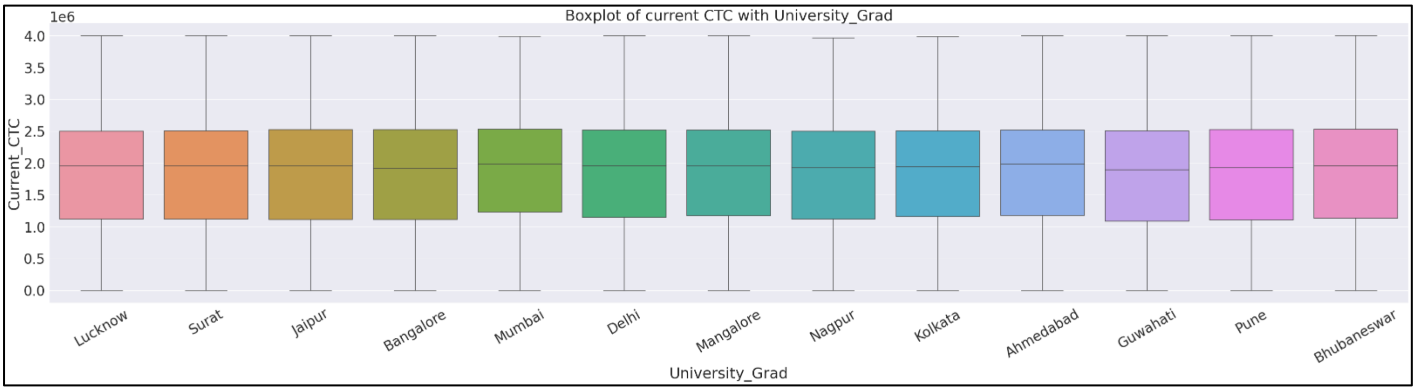


Figure 58:Distribution of current CTC by location of graduation university

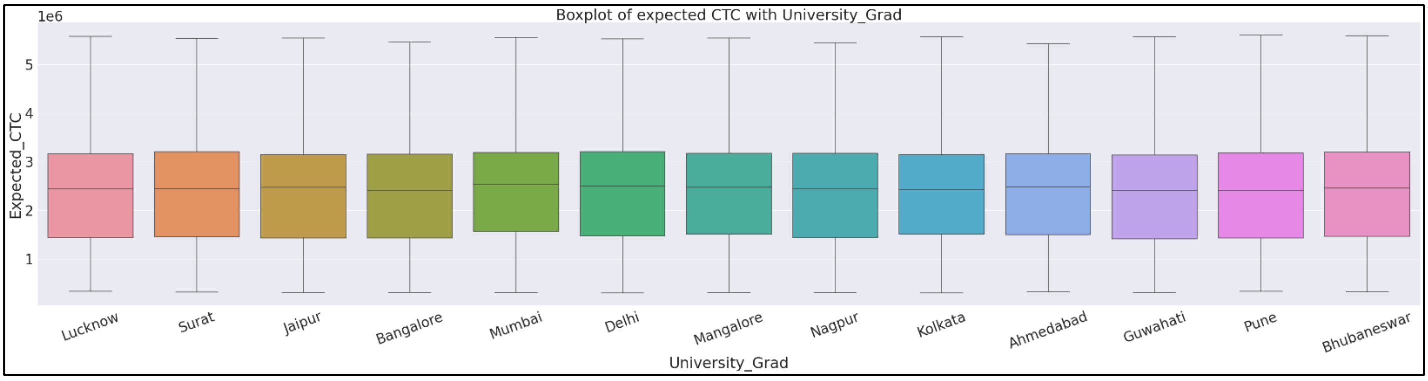


Figure 59:Distribution of expected CTC by location of graduation university

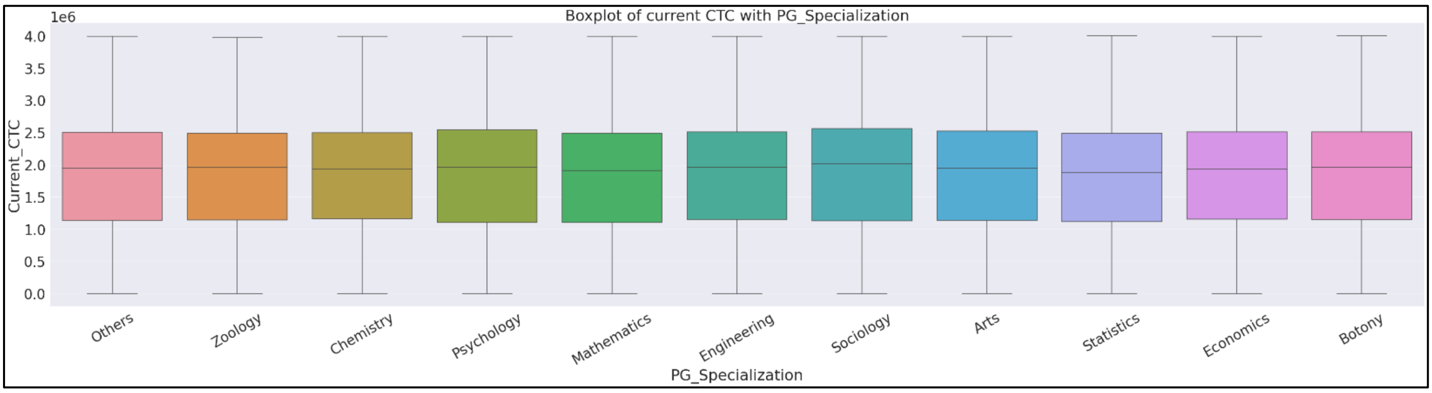


Figure 60:Distribution of expected CTC by location of graduation university

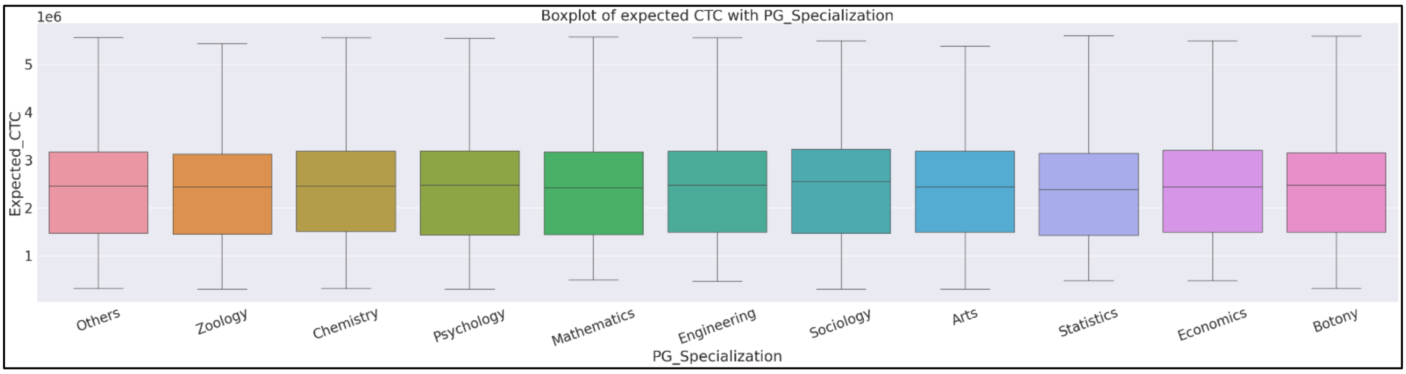


Figure 61:Distribution of expected CTC by post-graduation specialization

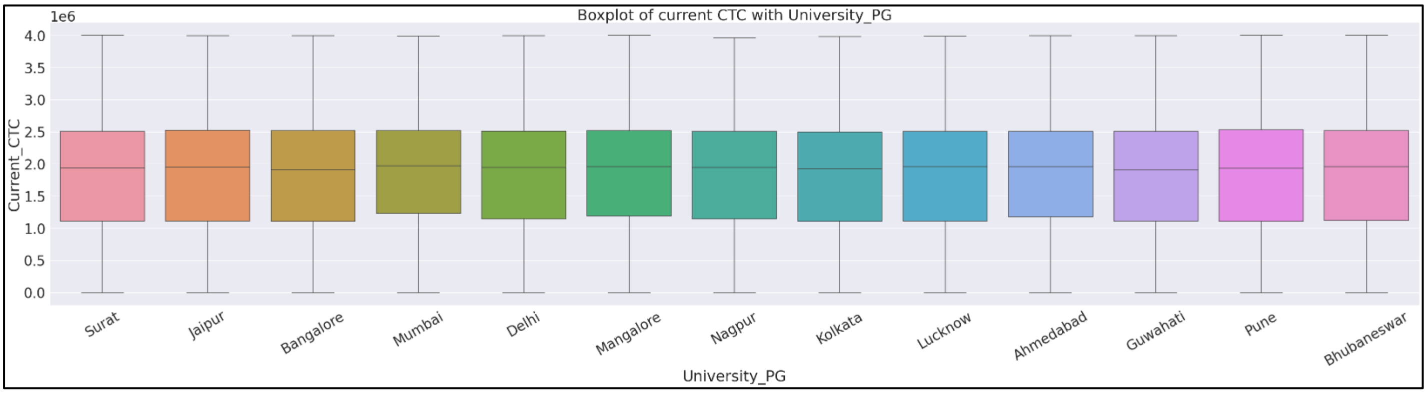


Figure 62:Distribution of current CTC by location of post-graduation university

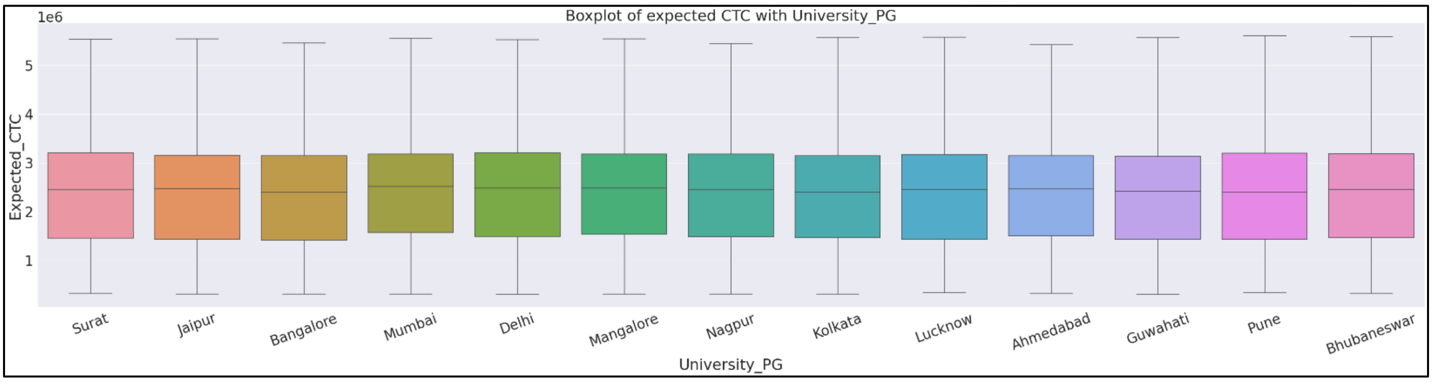


Figure 63:Distribution of expected CTC by location of post-graduation university

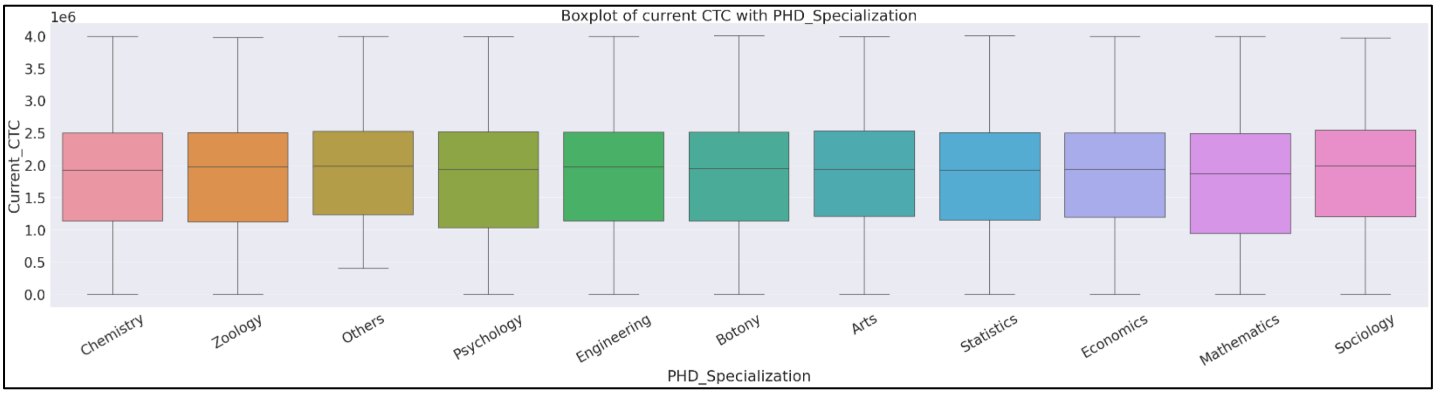


Figure 64:Distribution of current CTC by PHD specialization

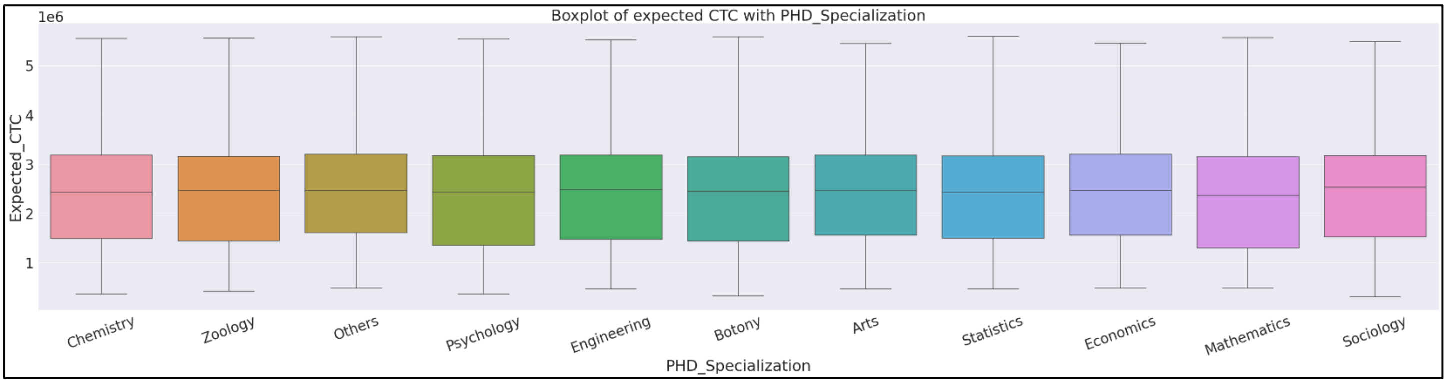


Figure 65:Distribution of expected CTC by PHD specialization

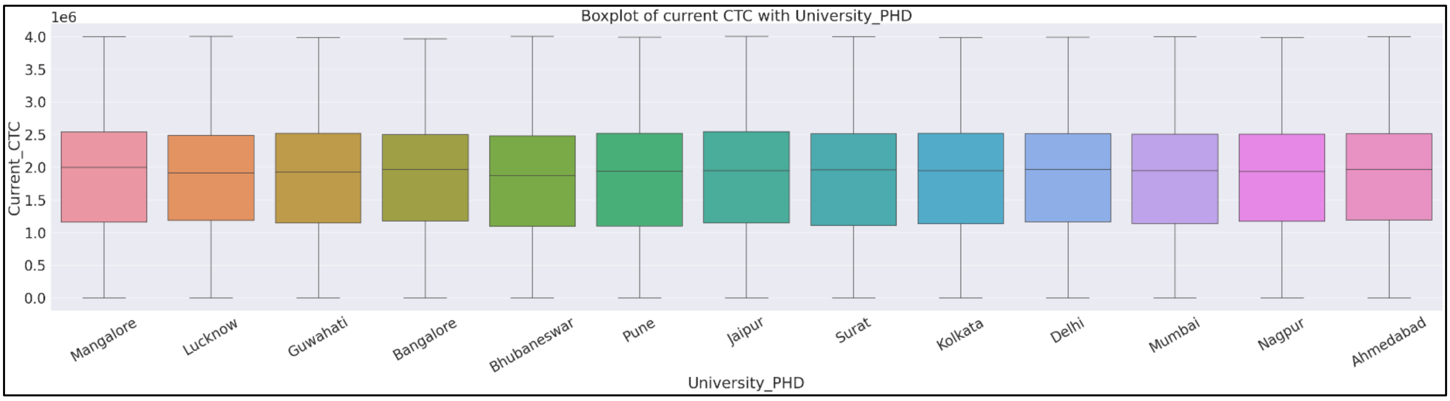


Figure 66:Distribution of current CTC by location of PHD university

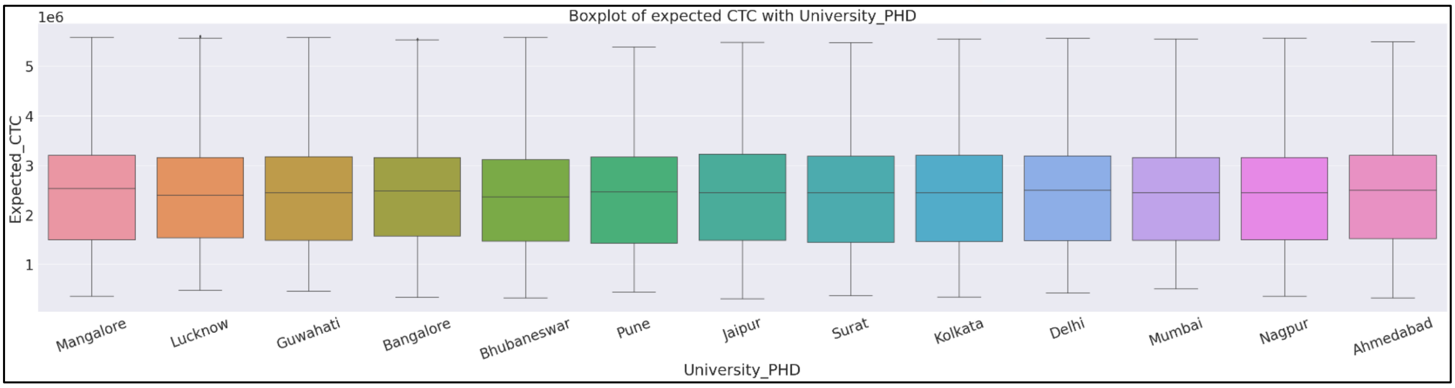


Figure 67:Distribution of expected CTC by location of PHD university

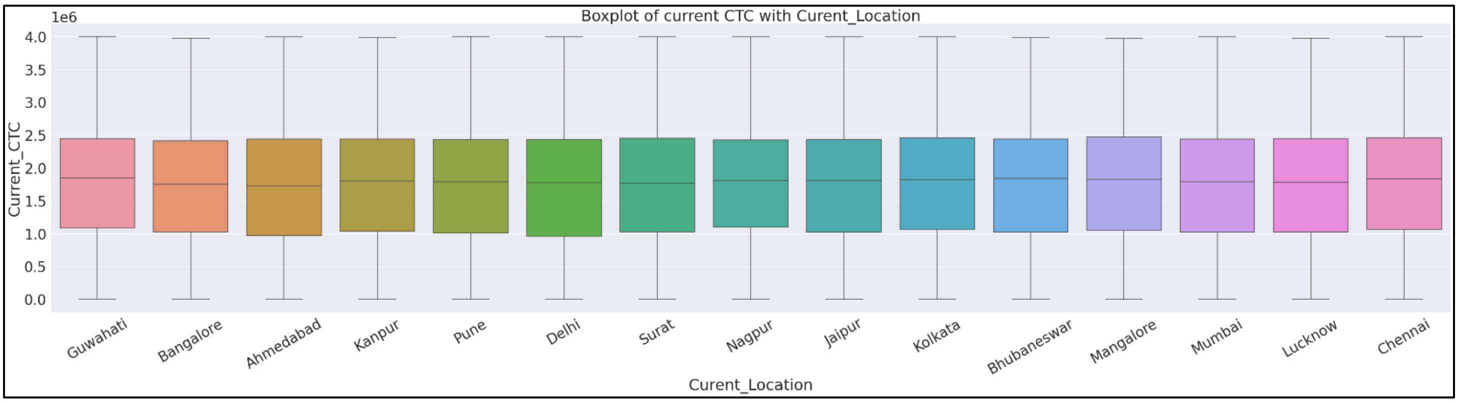


Figure 68:Distribution of current CTC by current location

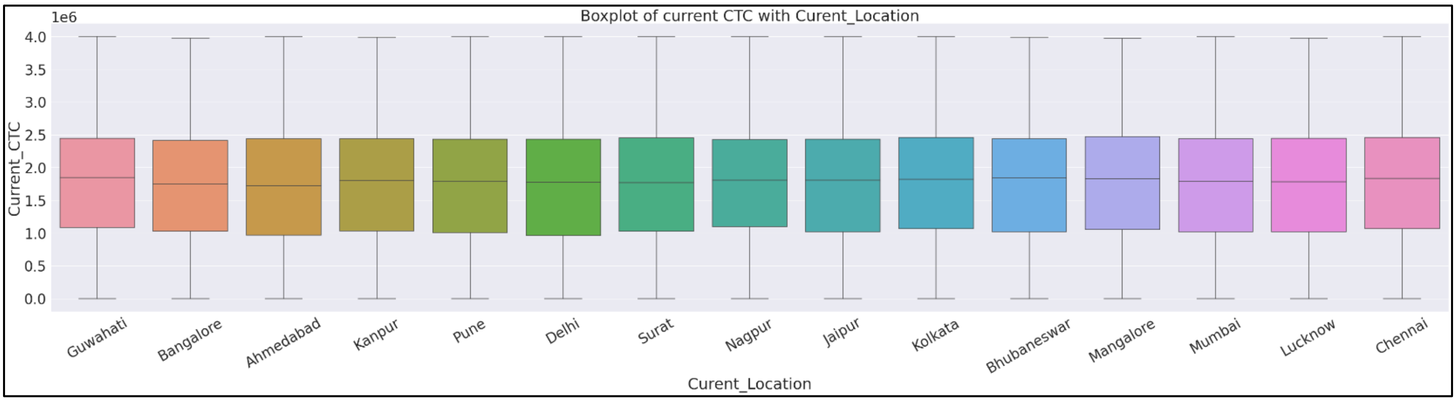


Figure 69:Distribution of expected CTC by current location

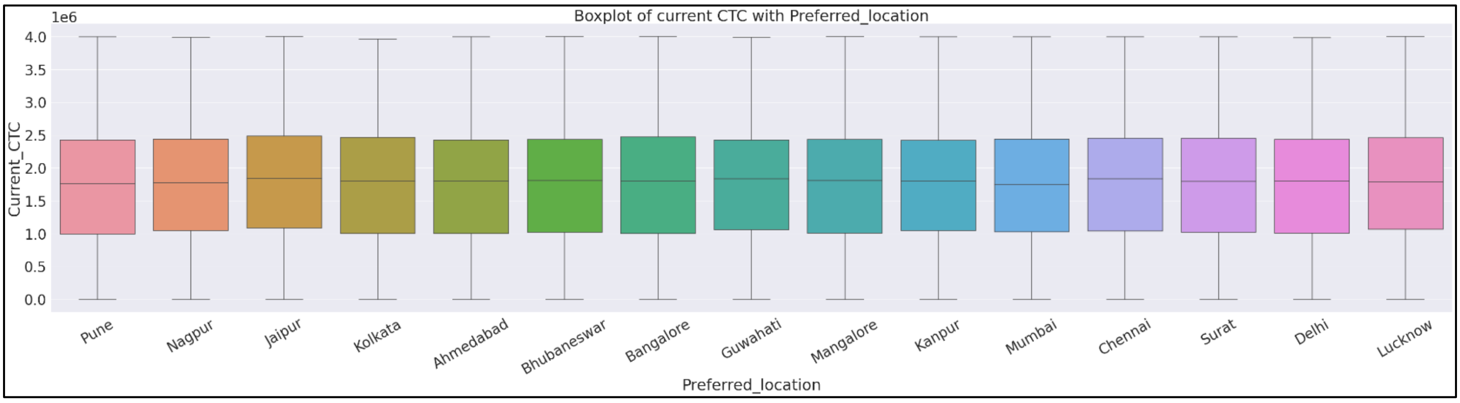


Figure 70:Distribution of current CTC by preferred location

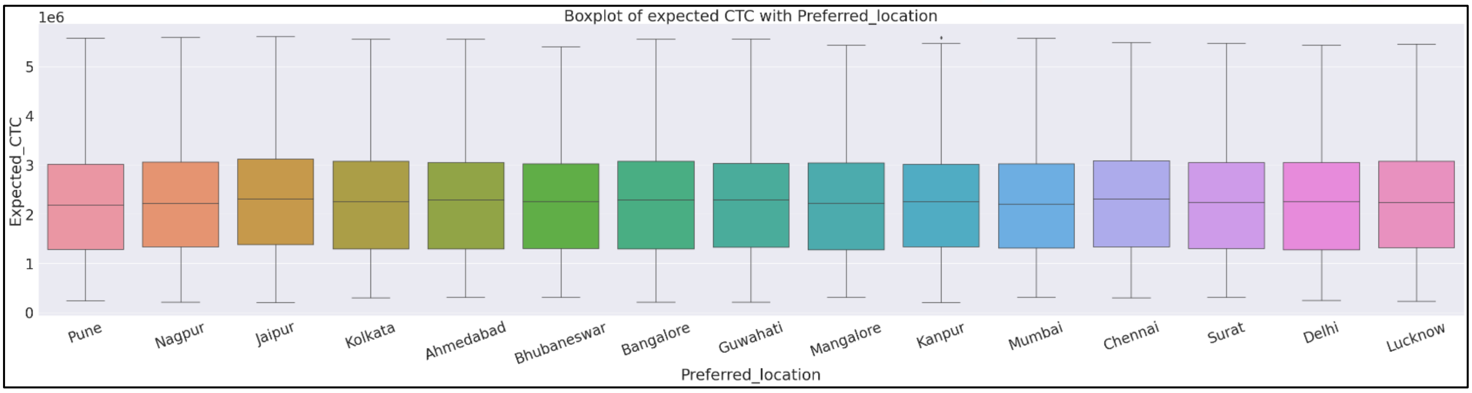


Figure 71:Distribution of expected CTC by preferred location

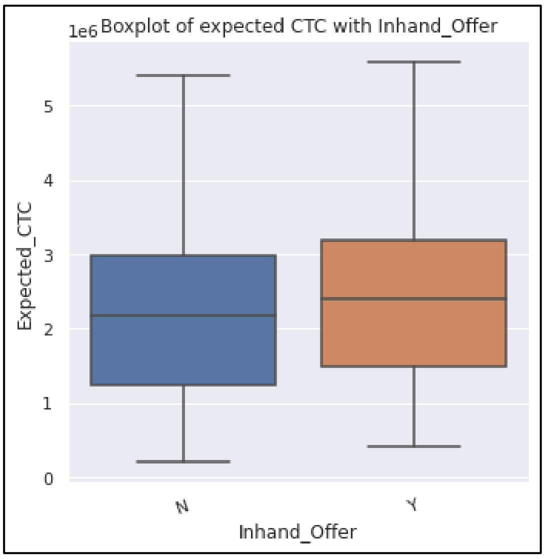


Figure 72:Distribution of expected CTC by availability of in-hand offer

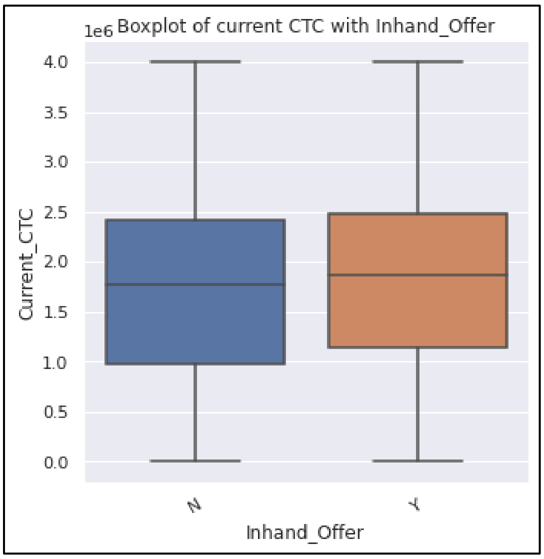


Figure 73:Distribution of current CTC by availability of in-hand offer

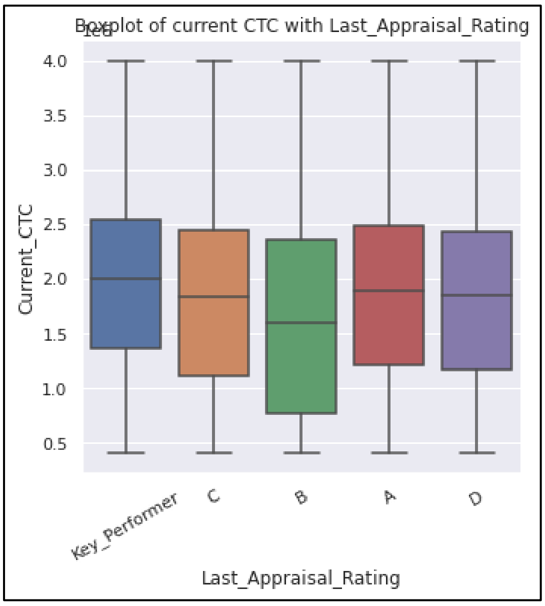


Figure 74:Distribution of current CTC by last appraisal rating

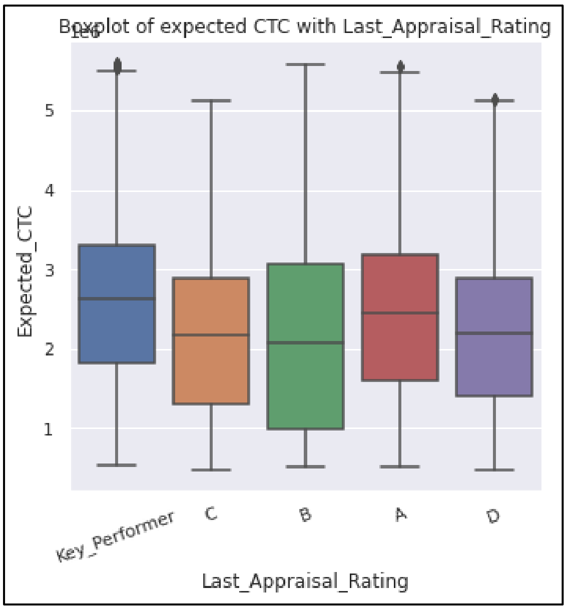


Figure 75:Distribution of current CTC by last appraisal rating