Capstone Final Report - HR CTC Prediction

28-May-2023

Ritu Utkarsha

Premkumar Raju

DSBA - LVC - May-2022 Batch

Table of Contents :

List of Figures :

List of Tables:

1. Executive Summary

Brief Description of the Problem:

- The HR team plays a very important role in predicting the salary of employees in an organization.The HR team conducts market research to determine industry trends and benchmark salaries for specific roles.
- It analyzes the job duties and responsibilities of different roles within an organization to determine the level of skill and experience required for each role. It develops a compensation structure for the organization that shows how salaries are determined and thus plays a crucial role in evaluating employee performance and making decisions about increase in salary,promotions,bonuses based on it.
- At present,people do move out frequently in search of better opportunities and in such conditions,organizations need more people as replacement.
- High employee turnover is a common challenge faced by many organizations. When employees frequently leave an Organization,it can also create a need for continuous recruitment to fill open positions.
- To overcome such a cumbersome process ,we can have a prediction tool which will use data analysis and statistical models to predict salaries based on a variety of factors,including job role,experience,education level and location.
- The problem at hand is to develop a model that can automatically determine the salary to be offered to prospective candidates applying for positions at Delta Ltd.
- The goal is to minimize human judgment and ensure fairness and consistency in the salary evaluation process.
- By utilizing historical data on applicants, the model will take into account various factors such as experience, education, certifications, publications, and other relevant attributes to predict an appropriate salary for each candidate.
- The dataset appears to contain information about job applicants, including their unique identification numbers, experience levels, education details, current salary, number of companies worked for, publications, certifications, international degree status, and expected salary.

Data Description

- We have data from the HR team of **25000** applicants who applied for jobs in Delta Limited in various departments and for various roles from many different organizations with different features from experiences, education qualifications, locations, roles, etc, including many features.
- Applicants are varying from freshers to 20+ years of experienced professionals with data till 2023 freshers who completed their higher education.
- The data consists of **25000 rows** and **29 columns** inclusive of the target variable **Expected_CTC**, which represents the final CTC offered by Delta Ltd and 11 other numerical data and 17 categorical data .There are **no duplicated** records.
- We have missing values in Department,role,Designation,education etc. Most of the missing values are due to freshers and undergraduates who will not have any data for

higher education. The **freshers** are **outliers** as they will be missing values for almost all the features except educational data.

Descriptive analysis of important variables:

- **Total_Experience**: The average total experience of applicants is around 12 years with a standard deviation of 7.47 years. The minimum and maximum values of total experience are 0 and 25 years, respectively.
- **Total_Experience_in_field_applied**: The average total experience in the field applied is around 5 years with a standard deviation of 5.82 years. The minimum and maximum values of this variable are 0 and 25 years, respectively.
- **Current_CTC**: The average current CTC of applicants is around 1.80 million with a standard deviation of 920,212.5. The minimum and maximum values of this variable are 0 and 3.99 million, respectively.
- **No_Of_Companies_worked**: The average number of companies worked is around 3 with a standard deviation of 1.69. The minimum and maximum values of this variable are 0 and 6, respectively.
- **Number_of_Publications**: The average number of publications is around 4 with a standard deviation of 2.61. The minimum and maximum values of this variable are 0 and 8, respectively.
- **Certifications**: The average number of certifications is around 0 with a standard deviation of 1.20. The minimum and maximum values of this variable are 0 and 5, respectively.
- **International_degree_any**: Around 8.17% of applicants have an international degree.
- **Expected_CTC**: The average expected CTC of applicants is around 2.25 million with a standard deviation of 1.16 million. The minimum and maximum values of this variable are 203,744 and 5.6 million, respectively.
- The majority of the applicants (80%) have not received any **in-hand offer**.
- The most common **education** level is PG, followed by Doctorate and Grad.
- The majority of the applicants (82%) do not have an **international degree**.

We are converting **No_Of_Companies_worked, Number_of_Publications, Certifications and International_degree_any** to the type categorical before proceeding with EDA.

2. Exploratory Data Analysis,Insights and Modelling

Exploratory Data Analysis

Univariate Analysis of Continuous variables - Inference based of Distribution and Boxplots :

- **Total Experience :** The boxplot shows that the median experience is 12 years, with a range from 0 to 25 years. There are few outliers beyond the upper whisker, indicating some employees with extensive experience.
- **Total experience in field applied :** The boxplot shows that the median experience in the field applied is 5 years, with a range from 0 to 25 years. There are few outliers beyond the upper whisker, indicating some employees with extensive experience.

- **Current CTC :** The boxplot shows that the median current CTC is around 18 lakhs, with a range from 0 to 40 lakhs. There are few outliers beyond the upper whisker, indicating some highly paid employees.
- **Expected CTC :** The average expected CTC of applicants is around 2.25 million .The minimum and maximum values of this variable are 1.5 and 5.5 million, respectively.

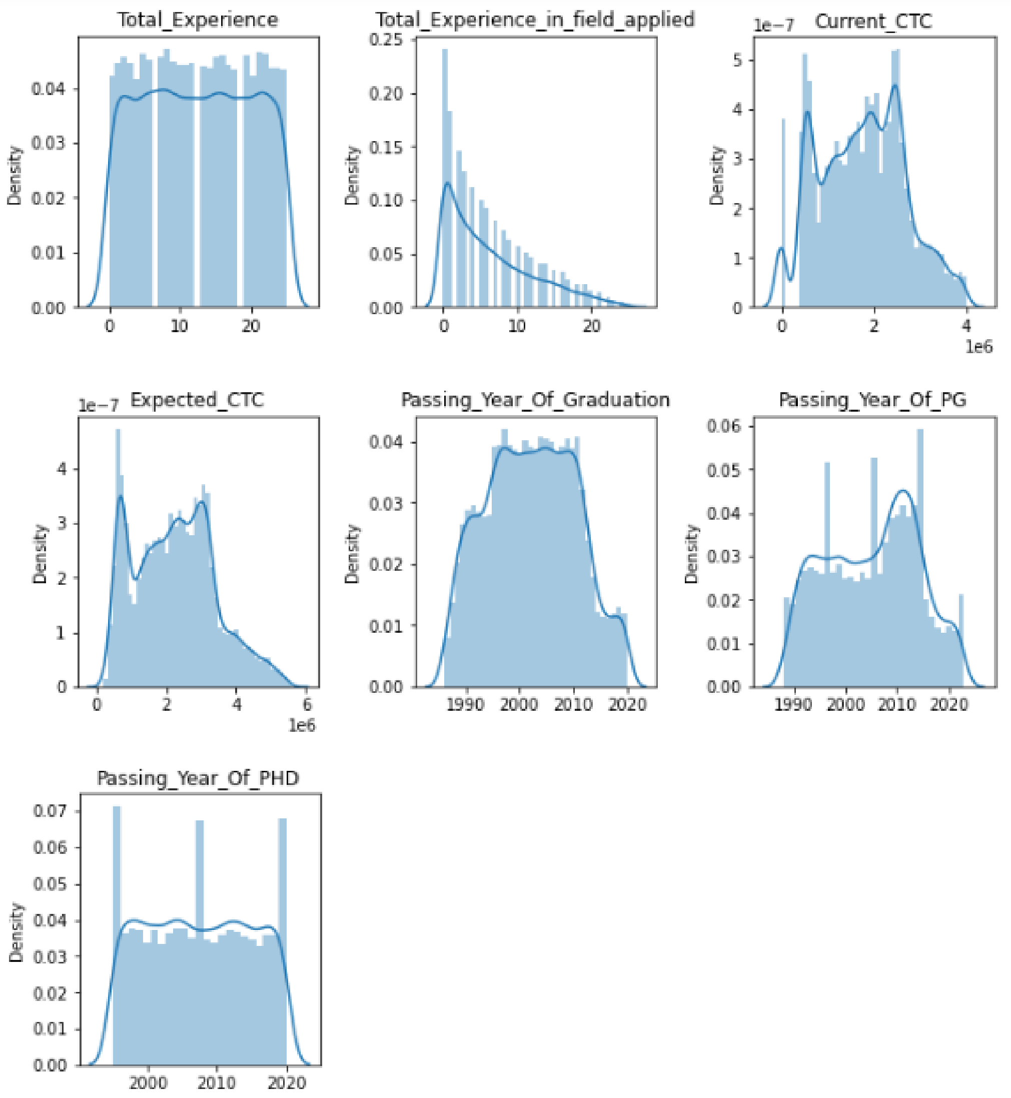


Figure 1 : Univariate analysis of each continuous variable using Histogram.

- **Passing_Year_Of_Graduation:** The average year of graduation seems to be in 2002 .The minimum and maximum values of this variable are 1986 and 2020, respectively. The curve is normally distributed with lots of missing values.
- **Passing_Year_Of_PG:** The average year of post-graduation is around 2005.The minimum and maximum values of this variable are 1988 and 2023, respectively. It contains more than 30% missing values

- **Passing_Year_Of_PHD**: The average year of PhD graduation is around 2007. The minimum and maximum values of this variable are 1995 and 2020, respectively. The curve is normally distributed with approx 50 percent missing values.

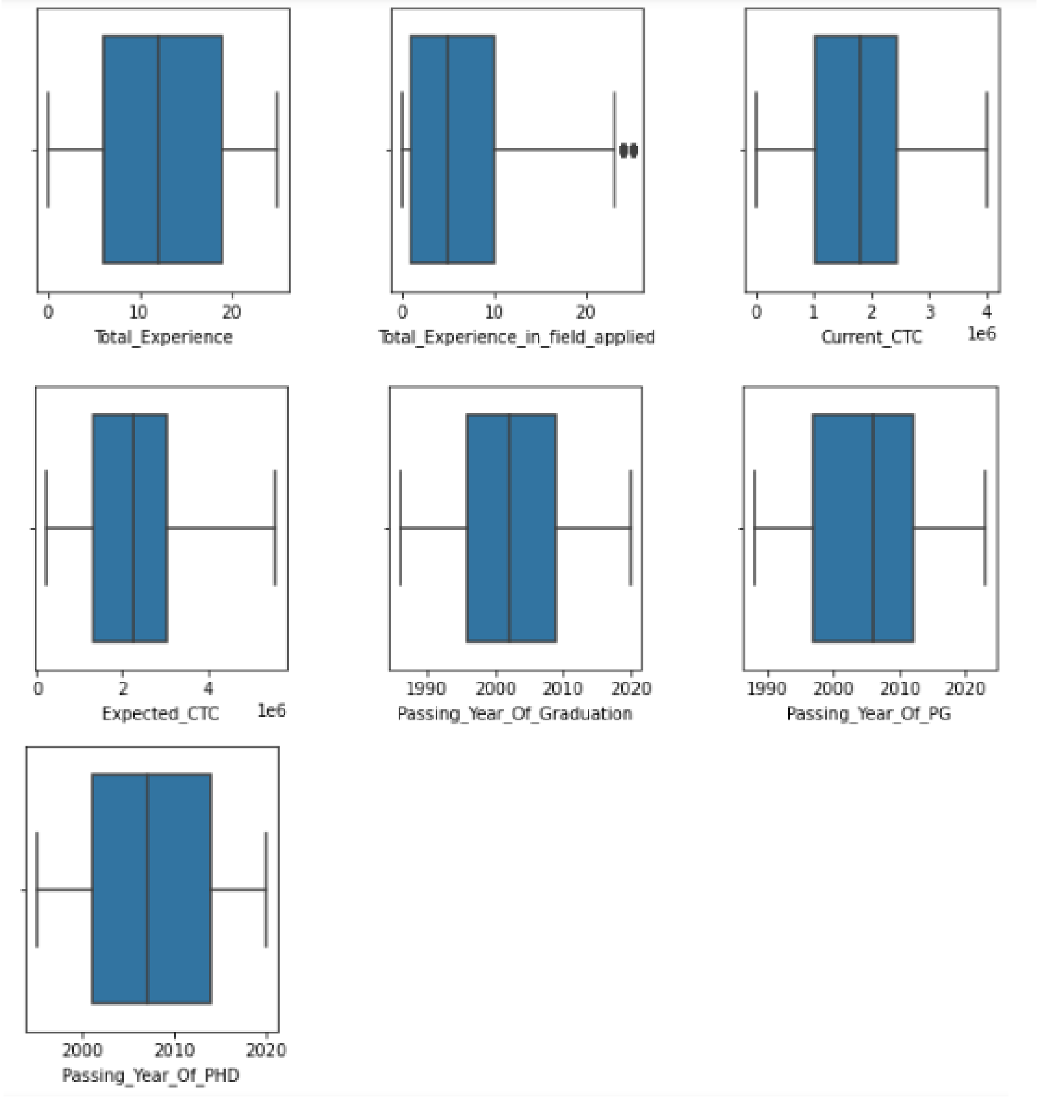


Figure 2 : Univariate analysis of each continuous variable using Boxplot.

Univariate Analysis of Categorical variables :

- **Department:** 12 unique values, with "Marketing" being the most frequent value and “Accounts” being the least frequent value.
- **Role:** 24 unique values, with "Others" being the most common(appearing 2248 times).
- **Industry:** 11 unique values, with "Training" being the most frequent value (appearing 2237 times)
- **Organization**: 16 unique values, with "M" being the most frequent value (appearing 1574 times)

- **Designation:** 18 unique values, with "HR" being the most frequent value (appearing 1648 times)
- **Education:** 4 unique values, with "PG" being the most frequent value (appearing 6326 times)

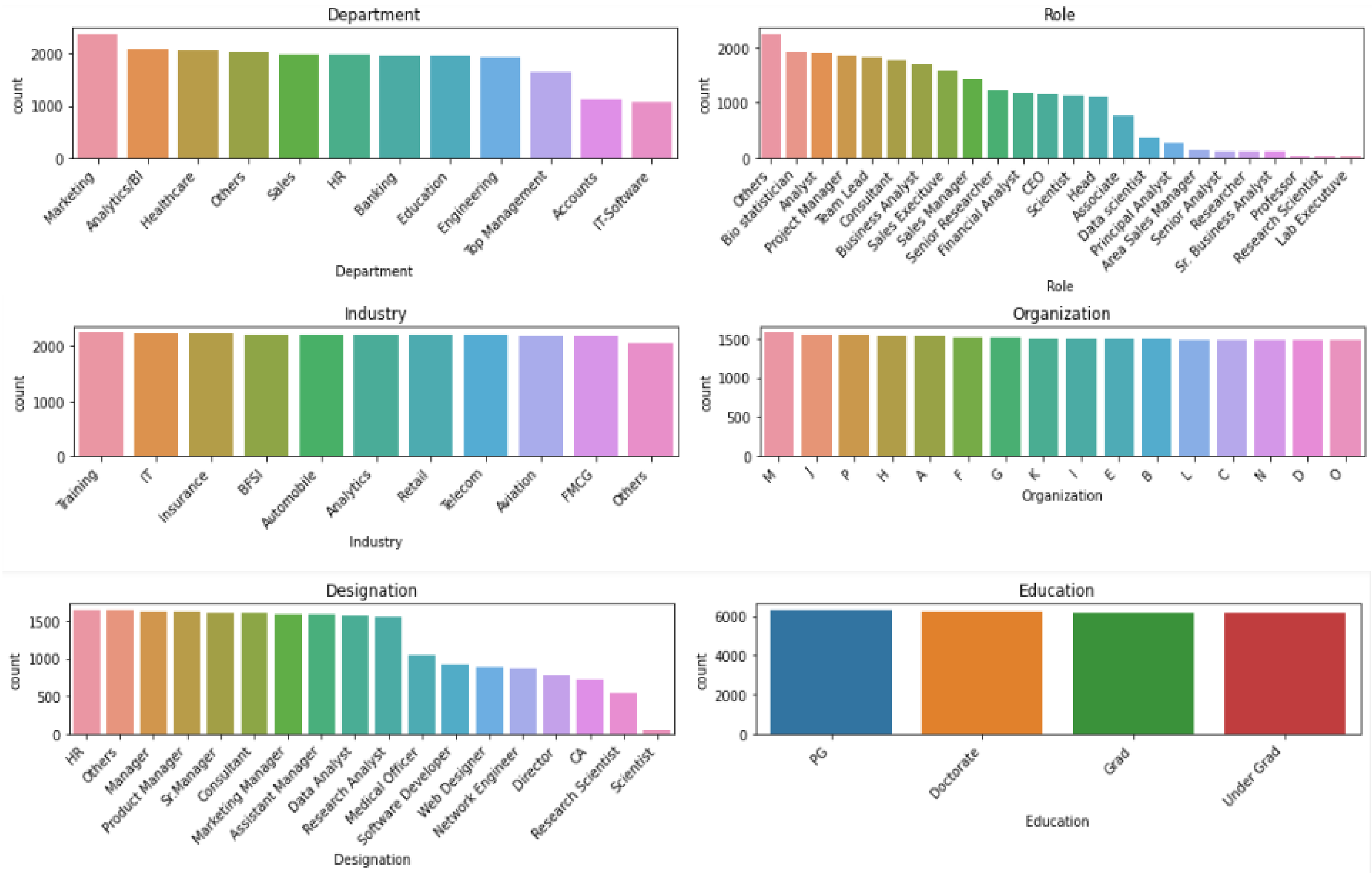


Figure 3.1 : Univariate analysis of each categorical variable using count plot.

- **Graduation_Specialization:** 11 unique specializations, with "Chemistry" being the most frequent value (appearing 1785 times)
- **University_Grad:** 13 unique universities, with "Bhubaneswar" being the most frequent value (appearing 1510 times)
- **PG_Specialization**: 11 unique specializations, with "Mathematics" being the most frequent value (appearing 1800 times)
- **University_PG:** 13 unique values, with "Bhubaneswar" being the most frequent value (appearing 1377 times) and Ahmedabad being the least frequent value.
- **PHD_Specialization:** 11 unique values, with "Others" being the most frequent value (appearing 1545 times) and the least common value is “Arts”
- **University_PHD:** 13 unique values, with "Kolkata" being the most frequent value (appearing 1069 times).The least common university for Phd is “Ahmedabad”.

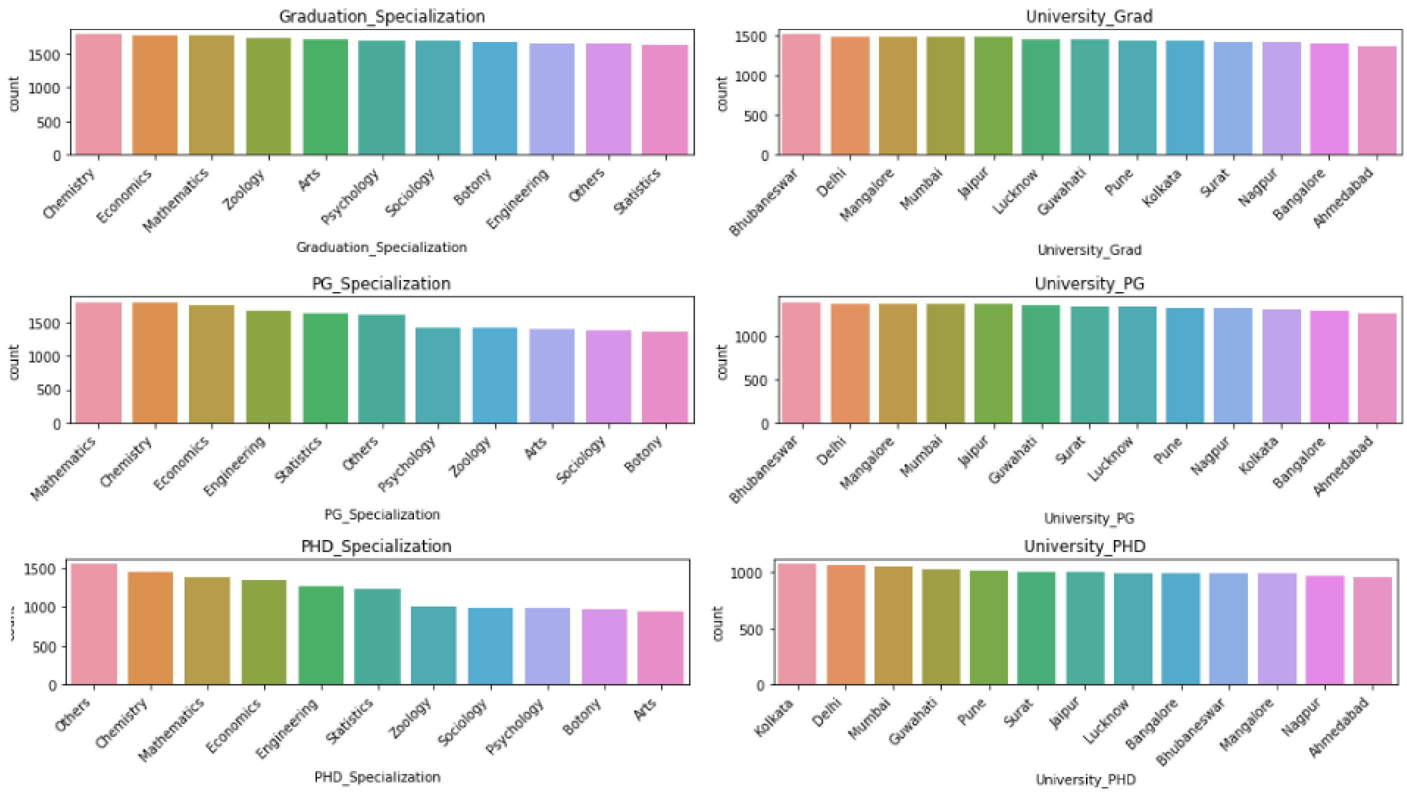


Figure 3.2 : Univariate analysis of each categorical variable using count plot.

- **Curent_Location:** 15 unique locations, with "Bangalore" being the most frequent value (appearing 1742 times) and Surat being the least frequent one.
- **Preferred_location**: 15 unique locations, with "Kanpur" being the most frequent value (appearing 1720 times) and Bangalore being the least frequent one.
- **Inhand_Offer:** 2 unique values (Yes/No), with "No" being the most frequent value (appearing 17418 times)
- **Last_Appraisal_Rating:** 5 unique ratings, with "B" being the most frequent value (appearing 5501 times) and Key performer being the least frequent value.
- **No_Of_Companies_worked:** Approx 5000 applicants have worked on 2 to 3 companies each and approx 3500 applicants have worked on 4,5 & 6 companies each before applying. Around 2000 applicants who might be juniors have worked only in 1 company and around 1000 applicants are freshers who never worked in any company.
- **Number_of_Publications :** Approx 3000 applicants have not published any publications and 2000 applicants have published 2 publications and we have approx 3000 applicants who have published in 1,3,4,5,6 & 7 publications.
- **Certifications :** Almost 15000 applicants do not have any certifications and we have a gradual number of applicants starting from 5000 with 1 certifications to lesser applicants with 2,3,4 and approx 400 applicants with 5 certifications.
- **International_degree_any :** Approx 2000 applicants have an international degree.

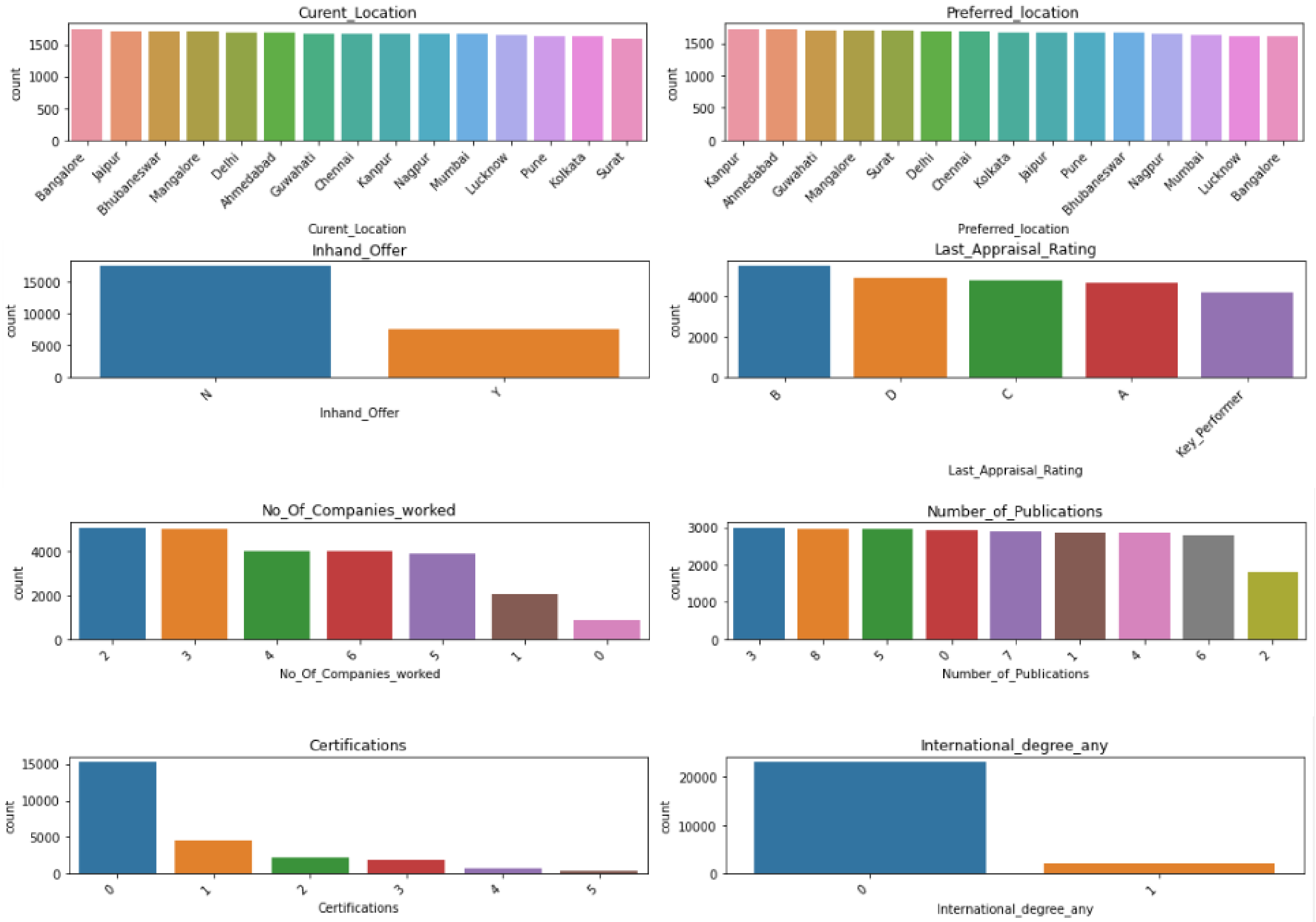


Figure 3.3 : Univariate analysis of each categorical variable using count plot.

Bivariate Analysis of Categorical variables by comparing relationship with the Expected_CTC :

- **Organization:** By plotting the expected CTC against each organization, we can observe the mean expected CTC does not vary much across organizations, with all the organizations offering approximately the same ctc.
- **Education:** We can see higher levels of education are associated with higher expected CTC as doctorates are getting the highest salary and undergraduates the least . Thus, we can conclude that education is a significant predictor of expected CTC.
- **Graduation Specialization:**We find that almost all the Graduation Specializations are associated with the same expected ctc. Thus,the plot of Graduation Specialization against Expected CTC provides insights that a candidate's field of study during graduation does not impact his expected ctc.

- **University Grad:**By analyzing the data, we find that graduates from all the universities tend to have about the same expected ctc .This information will not be valuable as there is no impact of the university where a candidate completed his graduation on his expected salary.

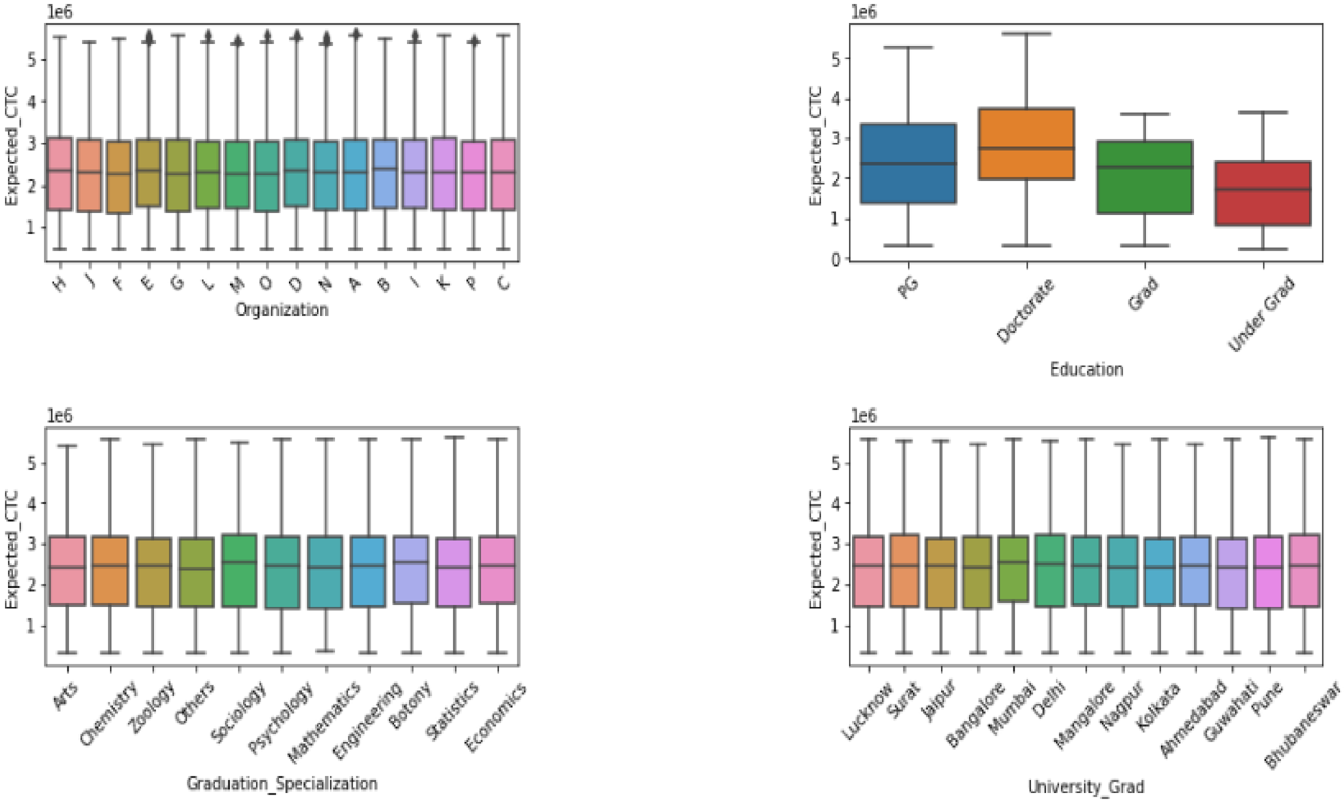


Figure 4.1 : Bivariate analysis of each categorical variable against Expected_CTC using boxplot.

- **PG specialization :** From the plot, we can see that the mean expected CTC for each PG specialization is nearly the same.Thus this variable doesn't influence the expected ctc of candidates.
- **University PG :** We can visualize that the average expected ctc for each university of postgraduate degree has a similar range. Thus,we can't use this information to make informed decisions when hiring employees and determining their salary.
- **PHD Specialization :** This analysis reveals that applicants with all the PhD specializations have the same expected ctc.Thus, this information can't be used by employers to make more informed decisions when recruiting job applicants with specific qualifications and specializations.
- **University PHD** : We determine there is not any significant difference in the expected ctc for candidates who obtained their PhD degree from different universities.Thus,the bivariate analysis of "University_PHD" against "Expected_CTC" does not provide valuable insights for prediction.

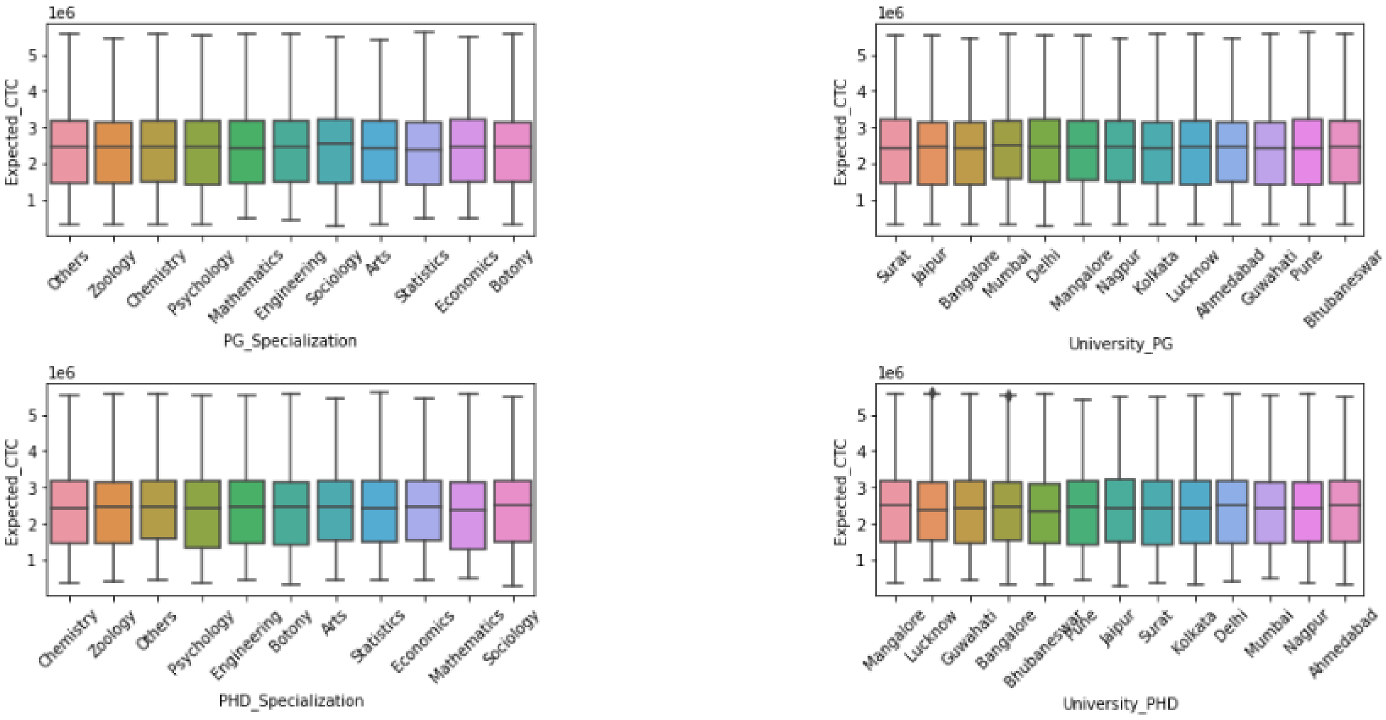


Figure 4.2 : Bivariate analysis of each categorical variable against Expected_CTC using boxplot.

- **Department:** From the plot, it can be observed that the "Top Management"department has the highest expected CTC, while the "Accounts" departments have the lowest expected CTC. The "IT",”Engineering”,"Analytics" and "Banking" departments have moderately high expected CTC. This analysis shows that the department a person works in has a significant impact on their expected CTC.
- **Role:** The box plot shows that the highest expected CTC is for the "Research Scientist" role, followed by the "CEO" and "Head" roles. The "Senior Analyst" and "Senior Business Analyst" roles also have high expected CTCs. On the other hand, the "Associate", and "Professor" roles have relatively very lower expected CTCs.
- **Industry:** The box plot of Industry against Expected_CTC shows that the distribution of expected CTC does not vary widely across industries.This analysis shows that industry is not an important factor that affects ctc expectations.
- **Designation:** From the analysis, we can observe that employees with job titles such as "Research Scientist" and "Director" have a significantly higher expected CTC compared to those in positions such as "Scientist". This suggests that job title plays a significant role in determining an employee's expected salary.
- **No of companies worked:**We observe that people who have worked for a higher number of companies have a slightly higher expected CTC.Thus, the number of companies worked can be a useful factor in predicting an individual's expected CTC, but it should be considered along with other factors such as experience, skills, and industry.
- **Number of Publications:**From the analysis, we can see that there is no visible relationship between the number of publications and expected CTC. This means that employees with more or less publications tend to have the same expected CTC.

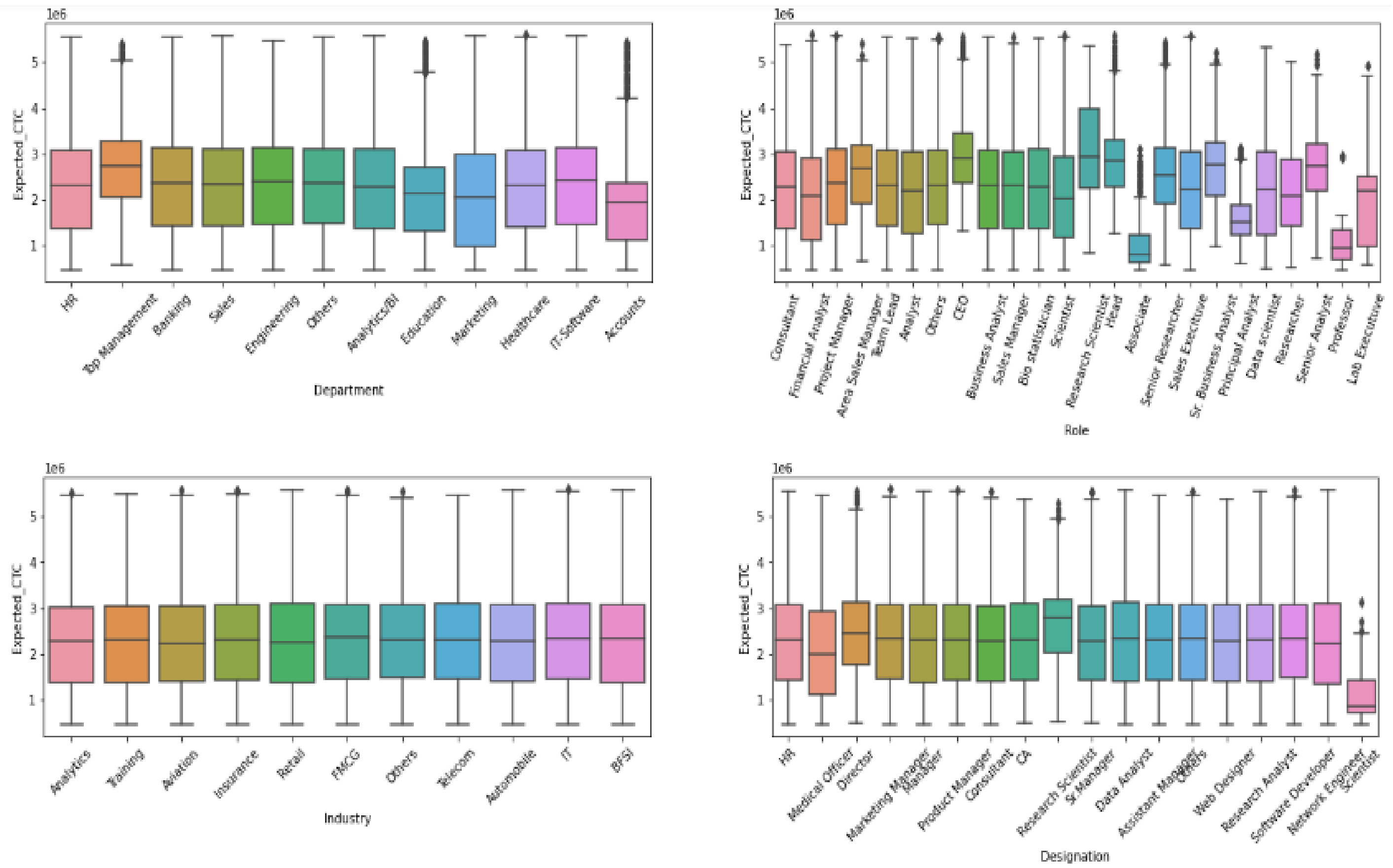


Figure 4.3 : Bivariate analysis of each categorical variable against Expected_CTC using boxplot.

- **Certifications:**The box plot indicates that individuals with the most number of certifications tend to have less expected CTC and those having 0 certifications have a higher expected CTC. Thus ,analyzing the relationship between certifications and expected CTC can provide valuable insights for both job seekers and employers.
- **International degree any:**From the plot, we can observe that applicants with an international degree tend to have a higher expected CTC compared to those without an international degree. This indicates that having an international degree may be perceived as a valuable asset by recruiters and may lead to higher salary expectations from the applicants.
- **Current_Location :** Irrespective of the current location of the applicant, there are no major differences in terms of mean Expected_CTC.
- **Preferred_location:** Similar to the current location, irrespective of the preferred location of the applicant, there are no major differences in terms of mean Expected_CTC.

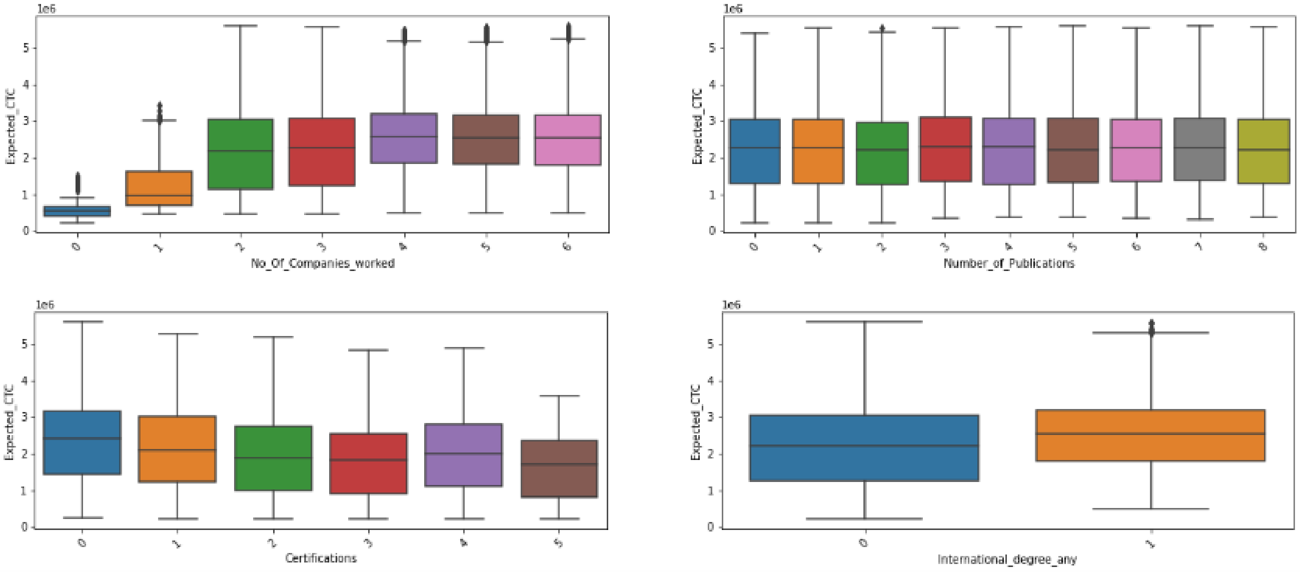


Figure 4.4 : Bivariate analysis of each categorical variable against Expected_CTC using boxplot.

- **Inhand_Offer:** We could see that when an applicant has an existing offer, their Expected_CTC is slightly more than those without any in hand offers.
- **Last_Appraisal_Rating:** We could see that the mean Expected_CTC of the key performers are way higher than that of other applicants followed by A. To our surprise C and D have better mean Expected_CTC than that of B.

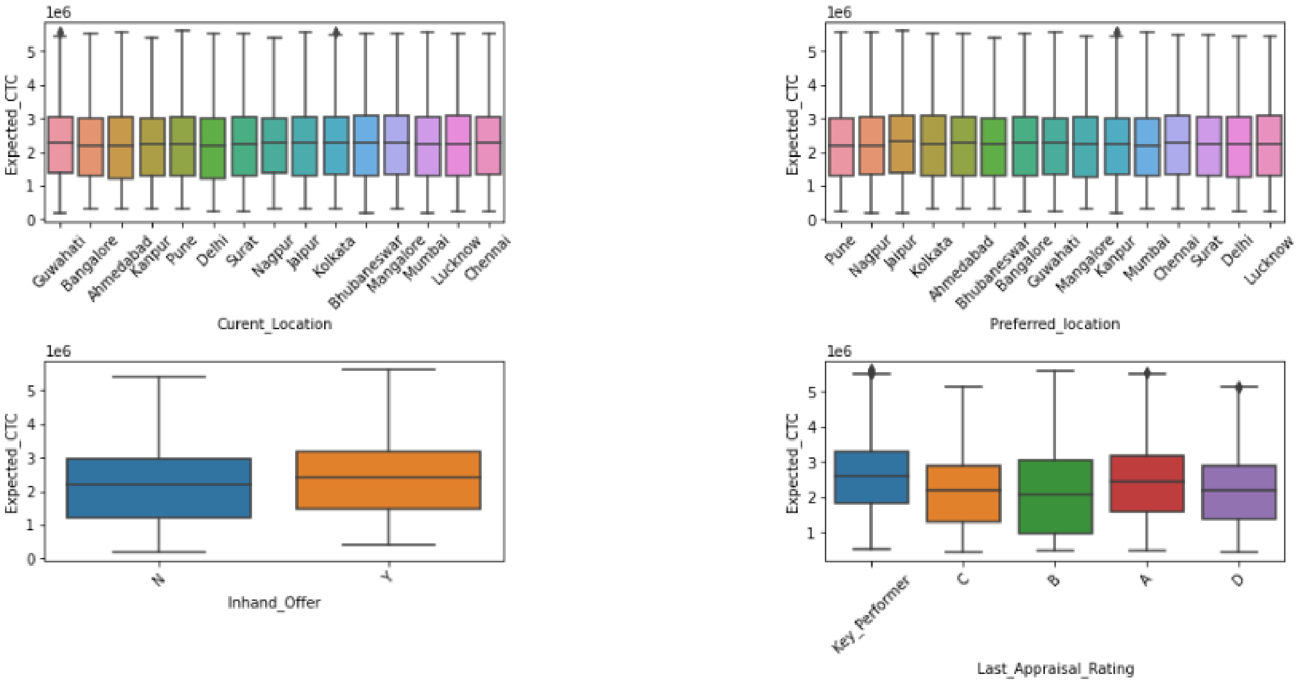


Figure 4.5 : Bivariate analysis of each categorical variable against Expected_CTC using boxplot.

Bivariate analysis based on Correlation heatmap :

- We could see that Current_CTC is very highly correlated with Expected_CTC.
- Experience is highly correlated with passing out years of education.
- Experience is pretty much correlated with Current and Expected CTC.

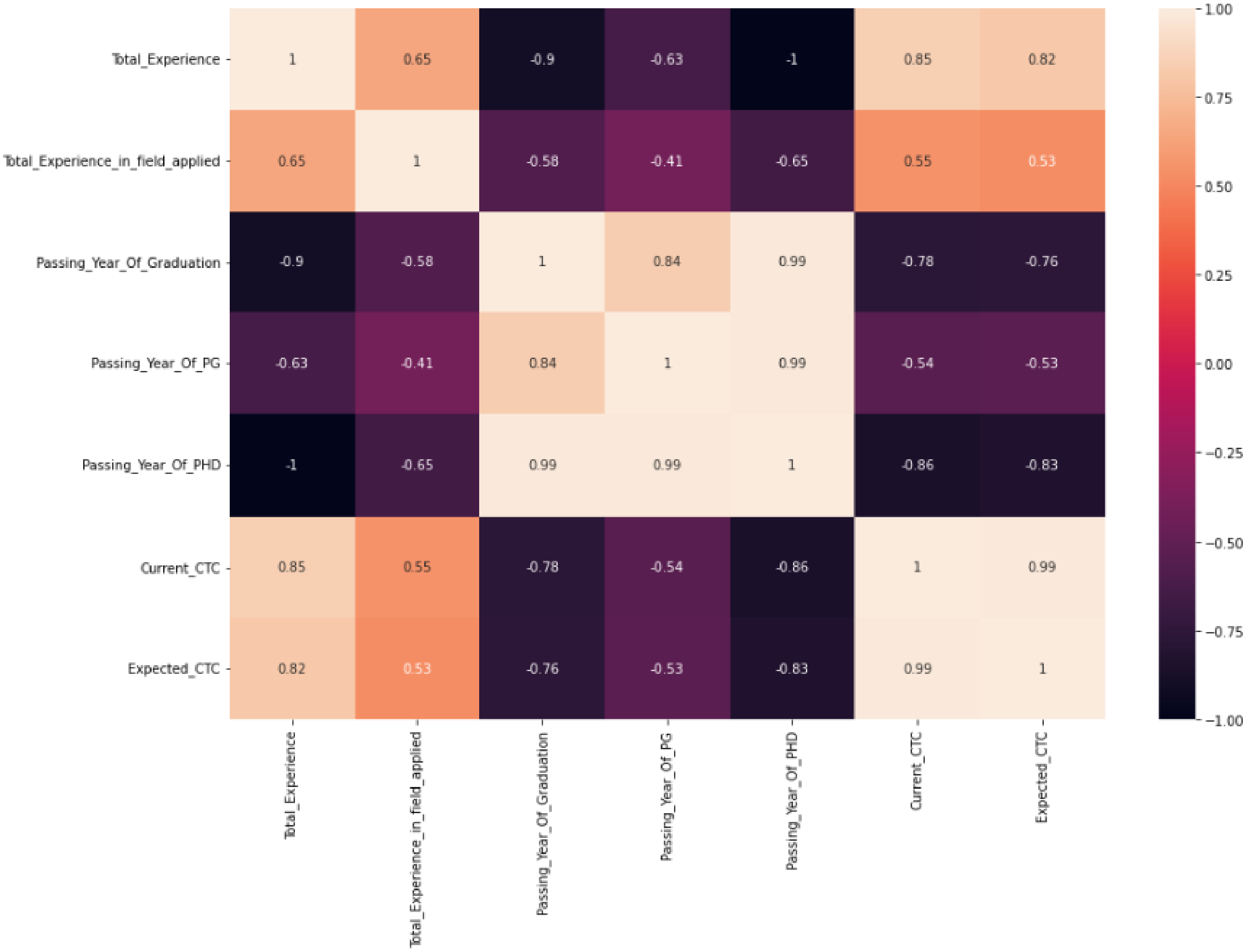


Figure 5 : Bivariate analysis of continuous variables using heatmap

Missing Value Treatment:

- We had a quite high number of missing values in education related variables, one major reason is that for Undergraduates all the higher education related features are null as expected similarly for graduates PG and PHD related columns had a null values. Based on the EDA we do not find much significance with these education related variables and dropped all of them except Education. So we have not performed any missing value treatments on those variables.

- We had a few missing values in the field of Designation, Department, Role, Organization, Industry, Last_Appraisal_Rating variables as freshers will not have any values for these fields and few experienced applicants too had missing values. So we have imputed them with Others as the value for the columns we kept from this list for our analysis.

Outlier Treatment :

- **Freshers** are the outliers as they will not have any values other than educational details and we have dropped all the educational related features. We cannot treat freshers with any outlier treatment as it will be meaningless. So we have dropped all of the fresher records from our data before analysis.

Variable transformation:

- We have converted **No_Of_Companies_worked, Number_of_Publications, Certifications and International_degree_any** to the categorical type.

- We have performed **One_hot encoding** for Education, Last_Appraisal_Rating, Inhand_Offer, Department, Designation variables and renamed a few of the variables to remove spaces in it.
- Based on the bivariate analysis of **Department** and Expected_CTC, we found an opportunity to reduce the number of categories based on the mean Expected_CTC. So we have reduced it from 12 to 3 categories.

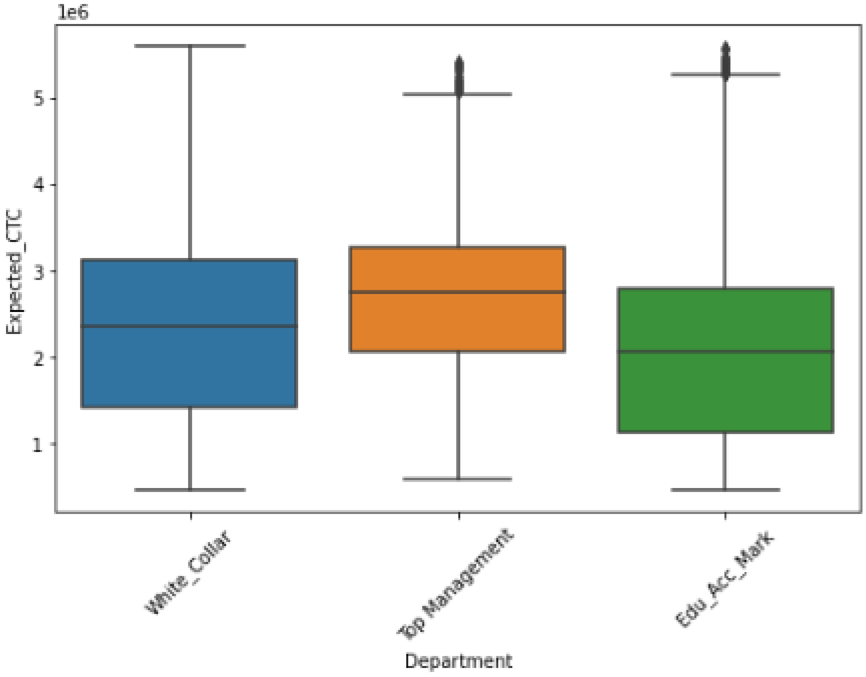


Figure 6 : Bivariate analysis of department against Expected_CTC using boxplot

Business insights from EDA

- We can get insights and recommendations from the EDA and important feature sets identified by the classifiers.
- The dataset is imbalanced based on the Experience for freshers and all other experienced applicants together. We have the same count of employees across all of the experience. So it is better to have a different method to predict expected CTC for freshers than including them among the experienced applicants.
- Except for Freshers, data is pretty much balanced in terms of all other important features.
- We have different categories of Roles, Designations and Departments from different Organizations, so instead of using them, we can try to focus on employee merits based on Experience, Education, Last appraisal rating, In Hand Offers, International degree, Number of Certifications and publications which differentiate the better employees among all the applicants.
- As Expected_CTC is highly correlated to Current_CTC, it is better to evaluate the applicant based on all other parameters else we will end up paying high salaries for few applicants who do not deserve the current CTC itself based on the actual merits and industrial standards. Similarly we might lose the eligible applicants who might draw a very lesser salary than his standard at present.

Initial model building and interpretation.

Target Variable : We are removing the test variable **Expected_CTC** before proceeding with the test train split.

Test & Train Split : The train-test split is a technique for evaluating the performance of a machine learning algorithm. The procedure involves taking a dataset and dividing it into 2 subsets and the objective is to estimate the performance of the machine learning model on new data (data not used to train the model).

Train dataset : Used to fit the machine learning model

Test dataset : Used to evaluate the fit machine learning model.

Here we have splitted the data into **70 : 30** data respectively using the **random state 0.**

Interpretation and Comparisons: Let us compare the metrics for all 4 models and then decide which is best for the given dataset.

Model \ Metric	MAE	MAPE	MSE	R2 Score	RMSE Score
Decision Tree Regression	291092.1042	0.1333298275	162023826310	0.8738294346	402521.8333
Random Forest Regression	289273.675	**0.1327722995**	158230674605	**0.876783223**	397782.1949
AdaBoost Regression	397729.8048	0.2324973444	236650627167	0.8157163418	486467.4986
Linear Regression	30560.53135	0.2181769414	1351185647	0.6399058042	367584.7722

Table 1 : Comparison of different metrics of the Models.

- Based on the metrics of all the 4 models with tuning, **Random Forest Regression** came as the best model in terms of R2 Score and MAPE value.
- R2 score is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data, so higher the value the better the model.
- The mean absolute percentage error (MAPE) is the most common measure used to forecast error, so the lesser the value, the better the model.

The analysis of the provided data has several potential impacts on the business:

Salary Planning and Budgeting:

The analysis provides insights into the distribution of salaries and various factors that influence them, such as experience, education, certifications, publications etc.This information can help the HR department and management in salary planning and budgeting for new hires and existing employees. It enables them to set appropriate salary ranges based on employee profiles, ensuring competitive yet fair compensation.

Equal Pay and Fairness:

The analysis helps in identifying any potential pay disparities among employees with similar profiles. It highlights the minimum, maximum, and quartile values for different salary components, enabling the business to ensure fairness and equal pay for employees in similar roles. Addressing any salary discrepancies based on objective data promotes a culture of equality and fairness within the organization.

Identifying Skill Gaps and Development Opportunities:

The analysis of certifications, publications, and international degrees can help the business identify skill gaps and development opportunities within the workforce. By analyzing the frequency and distribution of these variables, the business can identify areas where employees may need additional training or development t+o enhance their skills and knowledge.

Cost Management:

Understanding the distribution and statistics of current and expected CTC (Cost to Company) helps the business optimize its compensation budget. By analyzing the mean, standard deviation, and quartiles of CTC values, the business can identify areas where adjustments can be made to control costs while ensuring competitiveness in the job market. This analysis contributes to effective cost management and allocation of resources.

Compliance and Risk Mitigation:

The analysis assists the business in complying with legal requirements related to equal pay and non-discrimination. By examining salary data and identifying any potential biases or disparities, the business can take proactive measures to address and rectify any issues. This helps mitigate legal and reputational risks associated with discriminatory pay practices.

3. Model building

a. Technical details on the final model chosen

The final model chosen for the project is a **Random Forest algorithm**.

Random Forest Regression :

- The bootstrapping Random Forest algorithm combines ensemble learning methods with the decision tree framework to create multiple randomly drawn decision trees from the data, averaging the results to output a new result that often leads to strong predictions/classifications..
- Key strength of ensembling is that every model that we build should be independent of each other.

b. Step by step model building:

Data Preparation: The first step is to gather and preprocess the dataset. We collected the data and performed necessary data cleaning, feature engineering, and encoding categorical variables. The dataset is then split into training, validation, and testing sets.

Model Initialization: The Random Forest algorithm is then initialized with hyperparameters such as the number of trees, maximum tree depth, and the number of features to consider at each split.

Model Training: The Random Forest model is then trained on the training dataset. The algorithm creates multiple decision trees using a random subset of features and samples from the training data. Each tree is trained independently using a technique such as recursive partitioning to split the data based on the selected features**.**

Model Evaluation: After training, the model's performance is evaluated on the validation dataset. Evaluation metrics such as R squared,Mean Squared Error, Root Mean Squared Error are calculated to assess the model's predictive power and generalization ability**.**

Model Fine-tuning:

- Based on the validation results, the model may undergo fine-tuning. This involves adjusting hyperparameters such as the number of trees, maximum tree depth, or the number of features to consider at each split. Hyperparameter tuning techniques like grid search or random search can be used to find the optimal combination of hyperparameter values.
- To **optimize** the model, we are using Grid search and with wide values as parameters to max_depth, min_samples_split, min_samples_leaf, max_features and n_estimators.
- Based on the output of tuning with a wide range of parameters in multiple iterations, the best parameters came as **max_depth=10, max_features=28, min_samples_leaf=20, min_samples_split=40, n_estimators=289.**

c.Model Validation:

Once the model has been fine-tuned based on validation results, it is evaluated on a separate test dataset to obtain a final assessment of its performance. This helps in determining the model's effectiveness in real-world scenarios**.** Based on the below plot against actual and predicted values, we can see differences between segments and will discuss the same now.

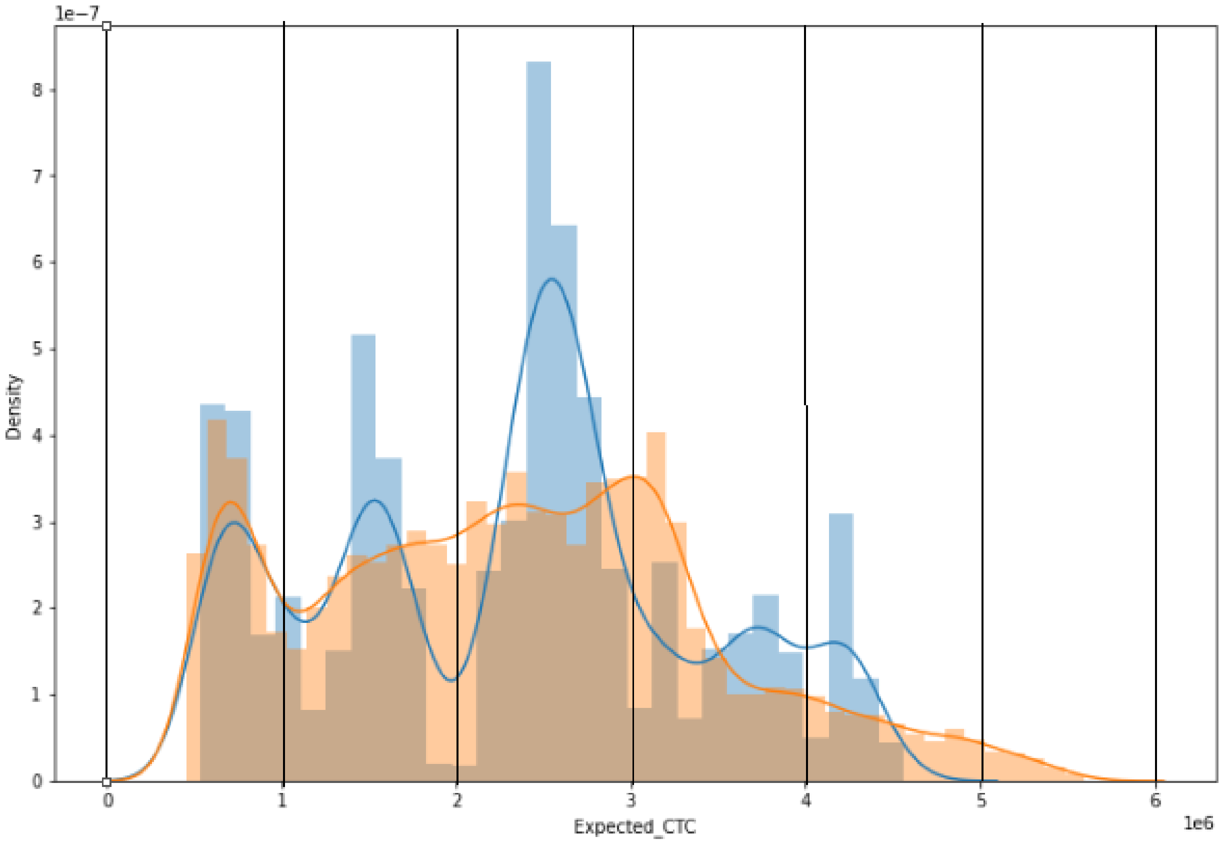


Figure 7 : Actual Vs Fitted values

Segment 1 - < 17% : We could see the actual and predicted values are almost matching in this segment without much deviation, showing that the model works very well for this segment.

Segment 2 - 17%-33% : We could see that we have little deviation in the middle of this segment and end of this segment, it might be due to few exceptional cases of the employees which caused it.

Segment 3 - 33%-50% : We could see that in this segment, the most of the predictions are towards the middle of the segment. and need to check the data further for this segment, might need to check the actual data or use a different model too.

Segment 4 - 50%-67% : This segment is similar to Segment 2 and might be due to few exceptional cases of employees.

Segment 5 - 67%-83% : In this segment, our model is predicting towards middle for most of the employees

Segment 6 - >83% : Model is predicting the expected salary of this segment in the earlier segment itself and might be due to senior employees who might be outliers and model might consider them for a little lesser salary.

4. Final recommendations

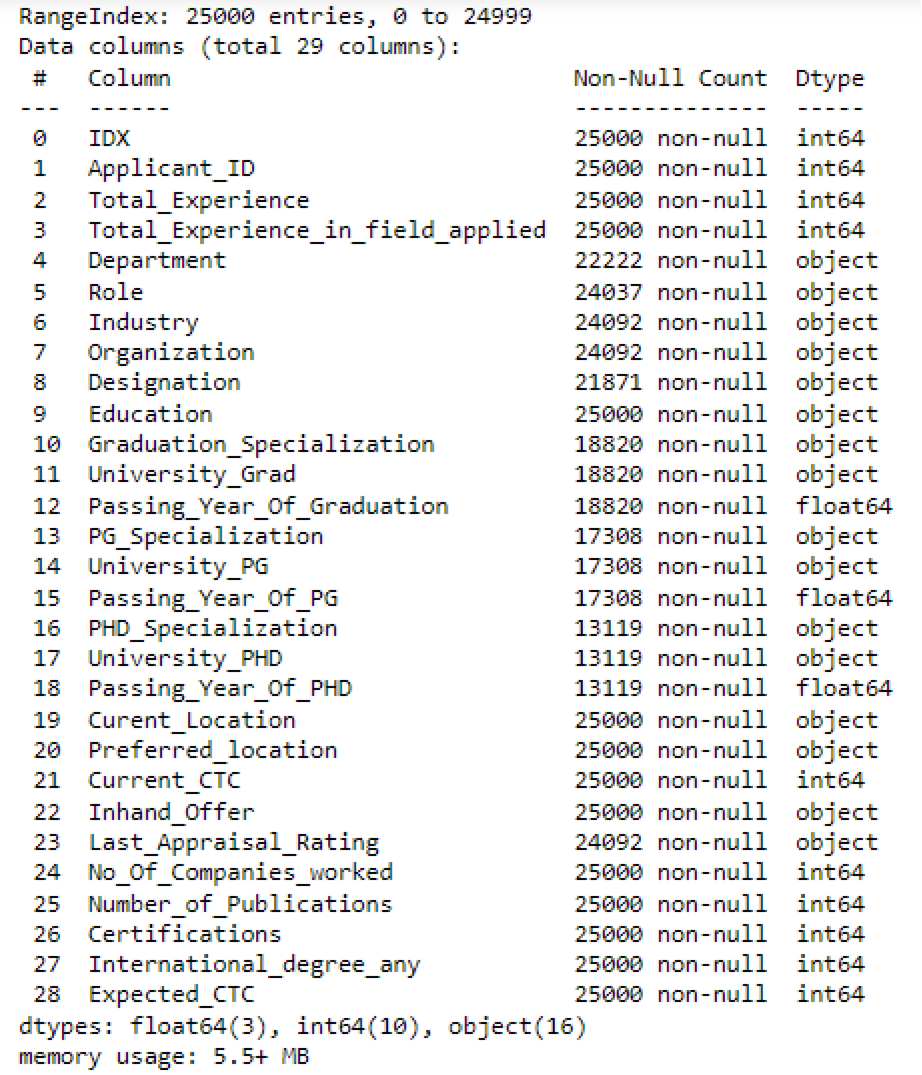
- **Prioritize continuous learning:** Offer training programs, workshops, and educational opportunities to enhance the skills and knowledge of the employees. This can contribute to their career growth as well as improve their performance in different roles.
- **Embrace Diversity and Inclusion:** Recognize the value of diversity and inclusion within the organization. Implement diversity recruitment strategies to attract candidates from different backgrounds, cultures, and experiences. Foster an inclusive work culture that values diverse perspectives and promotes equal opportunities for all the employees.
- **Offer competitive compensation :** Ensure that the compensation packages are competitive to attract and retain top talent. Regularly review salary ranges and benefits to align with industry standards and the qualifications and experience of the candidates you seek to attract.
- **Stay updated with education trends:** As the passing years of graduation, post-graduation, and Ph.D. indicate, it is important to stay updated with the latest trends in education.Regularly update the educational qualifications required for the positions you are hiring for, based on the passing years of graduation, post-graduation, and Ph.D. Stay informed about the latest educational trends and qualifications relevant to your industry to ensure you attract candidates with the right skills and knowledge.
- **Focus on Total Experience in Field Applied:** Emphasize the importance of relevant experience in the job postings and during the selection process. This will help ensure that candidates have the necessary expertise to contribute effectively to their roles.
- **Emphasize the importance of Research and Publications:** Encourage employees to pursue research projects, collaborate with academic institutions, and publish their work. This can enhance the company's reputation and attract candidates interested in contributing to their respective fields.
- When evaluating the applicants who already have an offer in hand, we must check their current CTC and merits to make sure if he will join our company or will just receive an offer and will not show up.
- Clustering can help us in identifying segments of applicants with similar characteristics, such as education, work experience and preferred location. This can be used to tailor recruitment strategies to each segment, improving the chances of attracting the best candidates.
- Clustering can also help in identifying pools of applicants who have specialized skills or knowledge in certain areas. This can be used to build relationships with these

candidates, providing them targeted opportunities for professional development and training.

- Clustering can also be used to group applicants who have similar qualifications and experience to past successful candidates. This can be used to develop a model that can predict the suitability of new candidates for specific roles based on their cluster membership.
- Companies can use this dataset to understand the average experience, education and skills of successful applicants in order to improve their recruitment process and attract more suitable candidates and to identify areas where employees need more training and development such as certifications and international degrees.
- Companies can also use this dataset to identify factors that contribute to higher appraisal ratings such as higher experience, education, and number of publications and to understand factors that contribute to employee satisfaction such as preferred location, industry, and department.

6. Appendix:

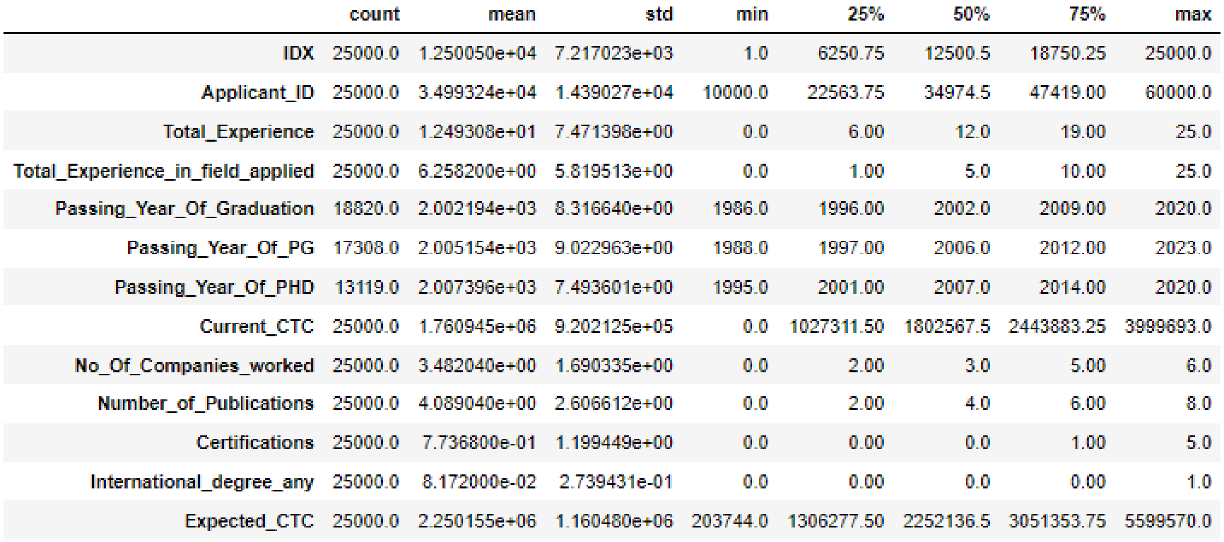
Info of the dataset:



```
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 29 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   IDX                                25000 non-null  int64  
 1   Applicant_ID                       25000 non-null  int64  
 2   Total_Experience                   25000 non-null  int64  
 3   Total_Experience_in_field_applied  25000 non-null  int64  
 4   Department                         22222 non-null  object 
 5   Role                               24037 non-null  object 
 6   Industry                           24092 non-null  object 
 7   Organization                       24092 non-null  object 
 8   Designation                        21871 non-null  object 
 9   Education                          25000 non-null  object 
 10  Graduation_Specialization          18820 non-null  object 
 11  University_Grad                    18820 non-null  object 
 12  Passing_Year_Of_Graduation         18820 non-null  float64
 13  PG_Specialization                  17308 non-null  object 
 14  University_PG                      17308 non-null  object 
 15  Passing_Year_Of_PG                 17308 non-null  float64
 16  PHD_Specialization                 13119 non-null  object 
 17  University_PHD                     13119 non-null  object 
 18  Passing_Year_Of_PHD                13119 non-null  float64
 19  Curent_Location                    25000 non-null  object 
 20  Preferred_location                 25000 non-null  object 
 21  Current_CTC                        25000 non-null  int64  
 22  Inhand_Offer                       25000 non-null  object 
 23  Last_Appraisal_Rating              24092 non-null  object 
 24  No_Of_Companies_worked             25000 non-null  int64  
 25  Number_of_Publications             25000 non-null  int64  
 26  Certifications                     25000 non-null  int64  
 27  International_degree_any           25000 non-null  int64  
 28  Expected_CTC                       25000 non-null  int64  
dtypes: float64(3), int64(10), object(16)
memory usage: 5.5+ MB
```

Table 3 : Info of the dataset

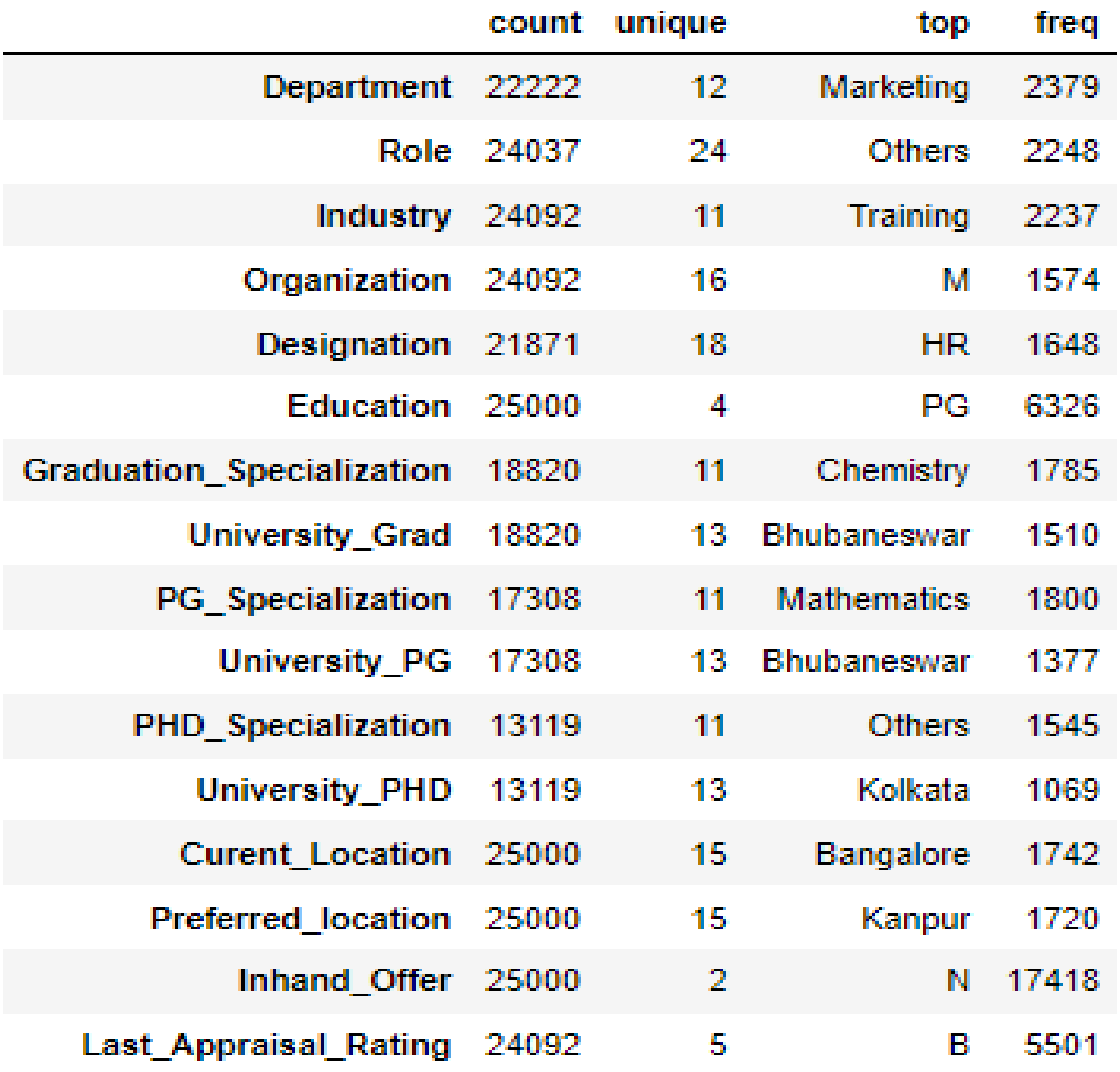
Description of Numerical variables:



	count	mean	std	min	25%	50%	75%	max
IDX	25000.0	1.250050e+04	7.217023e+03	1.0	6250.75	12500.5	18750.25	25000.0
Applicant_ID	25000.0	3.499324e+04	1.439027e+04	10000.0	22563.75	34974.5	47419.00	60000.0
Total_Experience	25000.0	1.249308e+01	7.471398e+00	0.0	6.00	12.0	19.00	25.0
Total_Experience_in_field_applied	25000.0	6.258200e+00	5.819513e+00	0.0	1.00	5.0	10.00	25.0
Passing_Year_Of_Graduation	18820.0	2.002194e+03	8.316640e+00	1986.0	1996.00	2002.0	2009.00	2020.0
Passing_Year_Of_PG	17308.0	2.005154e+03	9.022963e+00	1988.0	1997.00	2006.0	2012.00	2023.0
Passing_Year_Of_PHD	13119.0	2.007396e+03	7.493601e+00	1995.0	2001.00	2007.0	2014.00	2020.0
Current_CTC	25000.0	1.760945e+06	9.202125e+05	0.0	1027311.50	1802567.5	2443883.25	3999693.0
No_Of_Companies_worked	25000.0	3.482040e+00	1.690335e+00	0.0	2.00	3.0	5.00	6.0
Number_of_Publications	25000.0	4.089040e+00	2.606612e+00	0.0	2.00	4.0	6.00	8.0
Certifications	25000.0	7.736800e-01	1.199449e+00	0.0	0.00	0.0	1.00	5.0
International_degree_any	25000.0	8.172000e-02	2.739431e-01	0.0	0.00	0.0	0.00	1.0
Expected_CTC	25000.0	2.250155e+06	1.160480e+06	203744.0	1306277.50	2252136.5	3051353.75	5599570.0

Table 4 : Description of Numerical variables

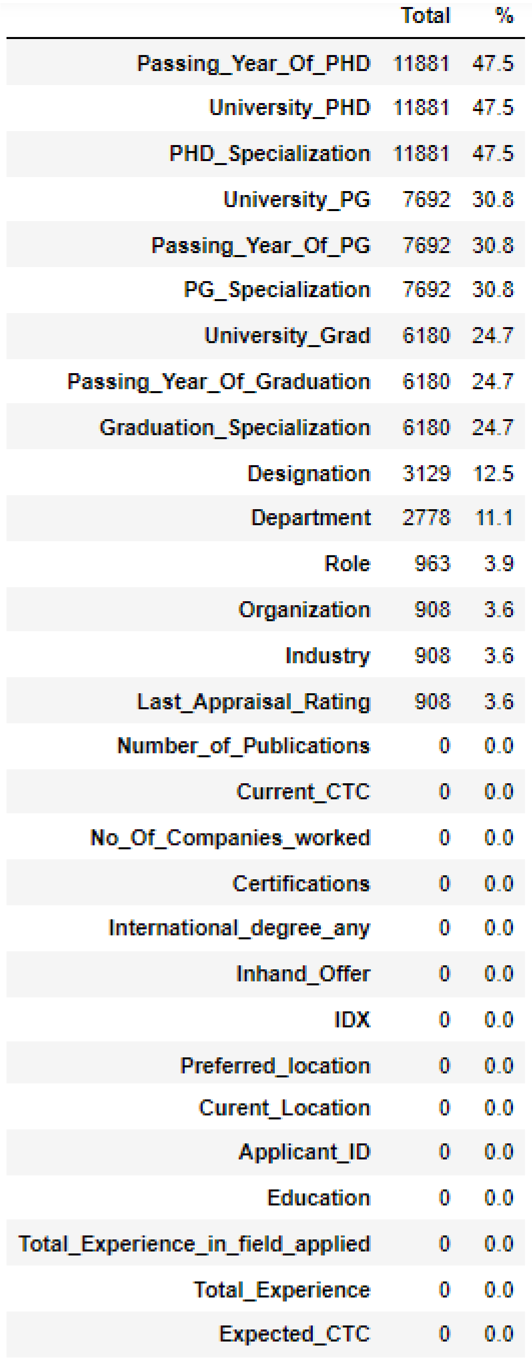
Description of Categorical variables:



	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 5 : Description of Categorical variables

% of Missing Values of the variables :



	Total	%
Passing_Year_Of_PHD	11881	47.5
University_PHD	11881	47.5
PHD_Specialization	11881	47.5
University_PG	7692	30.8
Passing_Year_Of_PG	7692	30.8
PG_Specialization	7692	30.8
University_Grad	6180	24.7
Passing_Year_Of_Graduation	6180	24.7
Graduation_Specialization	6180	24.7
Designation	3129	12.5
Department	2778	11.1
Role	963	3.9
Organization	908	3.6
Industry	908	3.6
Last_Appraisal_Rating	908	3.6
Number_of_Publications	0	0.0
Current_CTC	0	0.0
No_Of_Companies_worked	0	0.0
Certifications	0	0.0
International_degree_any	0	0.0
Inhand_Offer	0	0.0
IDX	0	0.0
Preferred_location	0	0.0
Curent_Location	0	0.0
Applicant_ID	0	0.0
Education	0	0.0
Total_Experience_in_field_applied	0	0.0
Total_Experience	0	0.0
Expected_CTC	0	0.0

Table 6 : Missing value percentages.

_________ END _________