Capstone_Project –Salary_Prediction_Part01

Raghavendra Kumar J R

PGP – DSBA Online

Date: 31/03/2024

Table of Contents:

Capstone Project - Part A

Business Problem:

To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles

Apart from the existing salary, there is a considerable number of factors regarding an employee's experience and other abilities to which they get evaluated in interviews. Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered.

Goal & Objective: The objective of this exercise is to build a model, using historical data that will determine an employee's salary to be offered, such that manual judgments on selection are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles.

Problem Understanding

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

Data Report

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

Exploratory Data Analysis

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) b) Bivariate analysis (relationship between different variables, correlations) a) Removal of unwanted variables (if applicable) b) Missing Value treatment (if applicable) d) Outlier treatment (if required) e) Variable transformation (if applicable) f) Addition of new variables (if required)

Business insights from EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business b) any business insights using clustering (if applicable) c) Any other business insights

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

Let's clarify the problem statement, state the necessity for the study/project, and comprehend the business/social opportunity in relation to Delta Ltd.'s effort to develop a predictive model for estimating salaries for prospective employees.

Defining problem statement:

Delta Ltd. is creating a predictive model that standardizes salary determination in an effort to address potential biases and inconsistencies in salary offerings. The objective of the salary negotiation process is to reduce subjective judgment by employing historical data on the salaries, experience, qualifications, and other pertinent factors of employees. Through reducing the possibility of discrimination and improving hiring process transparency, this model seeks to guarantee fair, equitable, and competitive compensation packages for all employees.

Need of the study/project:

Fairness and Equity: Promoting equity in the workplace requires making sure that salary offers are based on objective standards rather than arbitrary judgments or prejudices. This need is especially important in large, diverse organizations where disparities and unfair perceptions can result from inconsistent salary determinations.

Compliance and Risk Management: Strong procedures are needed to support salary decisions in order to comply with regulations pertaining to non-discrimination and equal pay. Subjective salary negotiations carry legal and compliance risks that can be reduced by using a data-driven approach.

Talent Attraction and Retention: Attracting and keeping top talent requires transparent, competitive pay structures. An organization's appeal as an employer can be greatly increased by implementing a transparent and equitable salary determination process.

Operational Efficiency: By standardizing the salary determination process, HR departments can operate more efficiently and spend less time and money resolving salary disparities and negotiating salaries.

Understanding business/social opportunity.

Improving Employer Branding: Delta Ltd. can reinforce its employer brand and establish itself as a moral and employee-focused company by advocating for equity and openness in compensation decisions. This can improve its standing with prospective workers as well as with the larger clientele and industry.

Creating Industry Standards: Establishing a strong, data-driven model for determining salaries creates a precedent in the sector and may persuade other businesses to follow suit. This promotes fairness and equity in the workplace on a larger scale by raising standards across the board.

Social Equity: By systematically addressing and eliminating gender, ethnicity, or other forms of pay gaps, such an initiative helps to promote social equity in addition to its immediate business benefits. It harmonizes the business's practices with more general societal ideals of equity and chance.

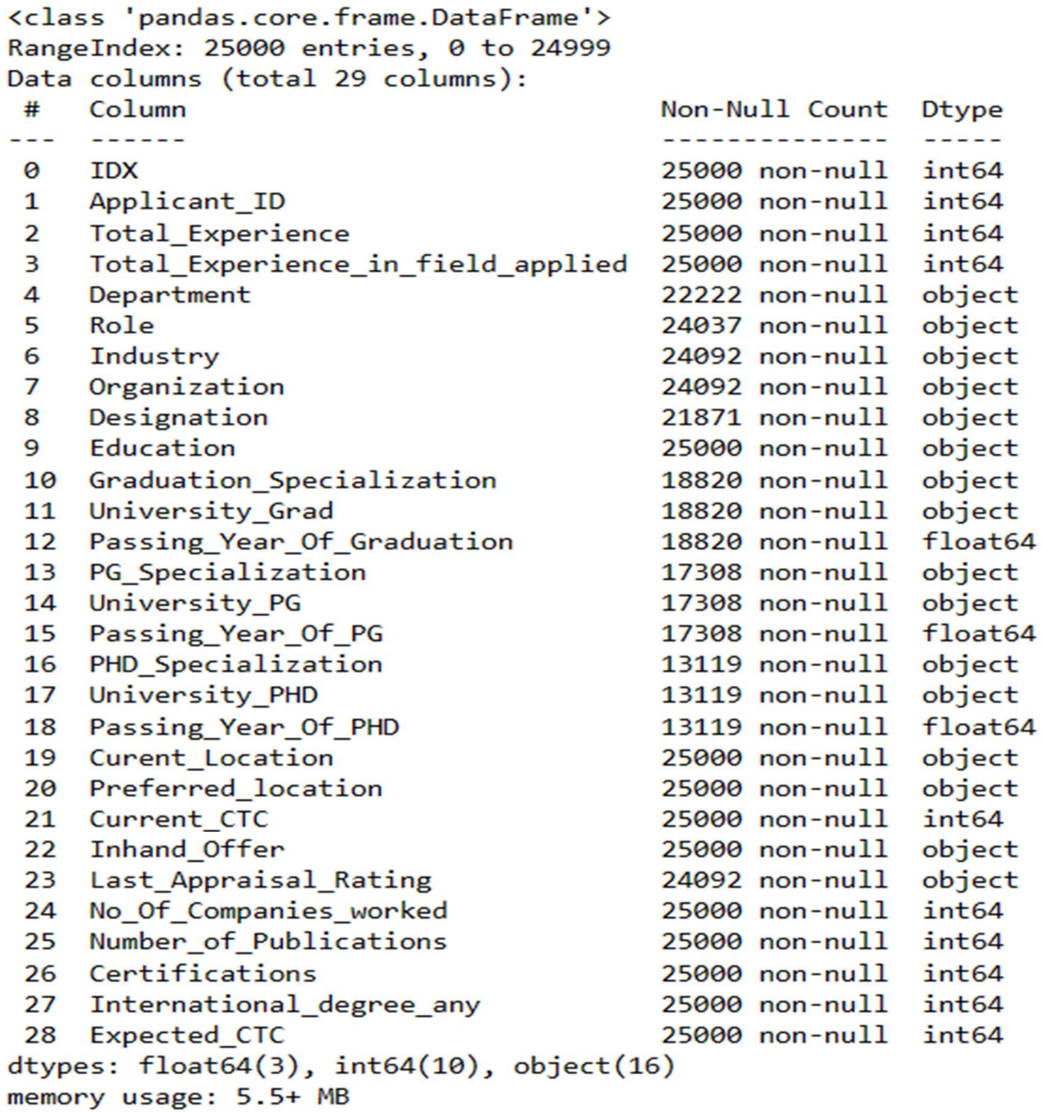
Data Utilization and Innovation: This project serves as an excellent example of how data and analytics can be used to address difficult HR problems, creating opportunities for more innovative approaches to personnel management and operational effectiveness.

In conclusion, Delta Ltd.'s creation of a predictive model for salary determination not only meets a crucial business need but also offers substantial chances to improve social equity, employer branding, and operational effectiveness. It highlights the business's dedication to equity, openness, and creativity.

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

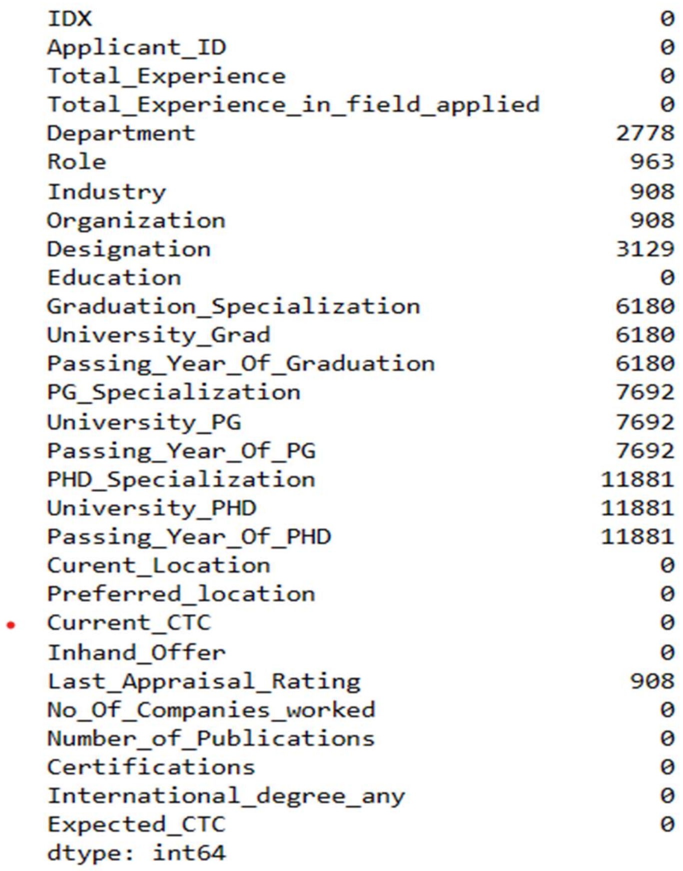
Data analysis - Solution:

The dataset contains 25,000 rows and 29 columns.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 29 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   IDX                                25000 non-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
```

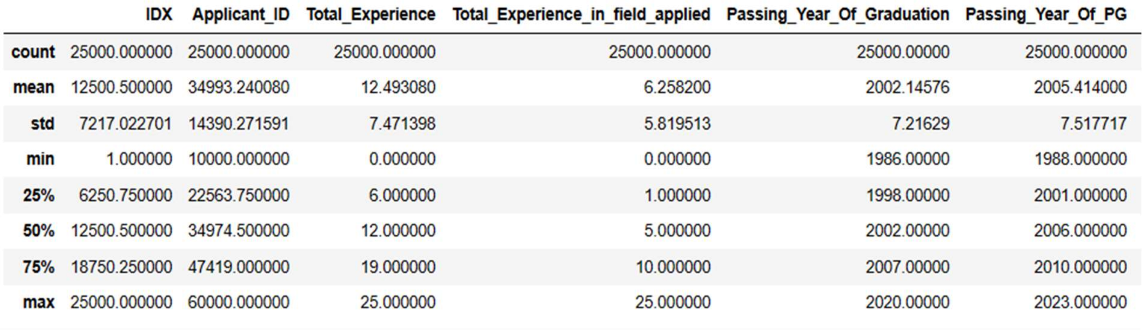
Several columns have missing values:



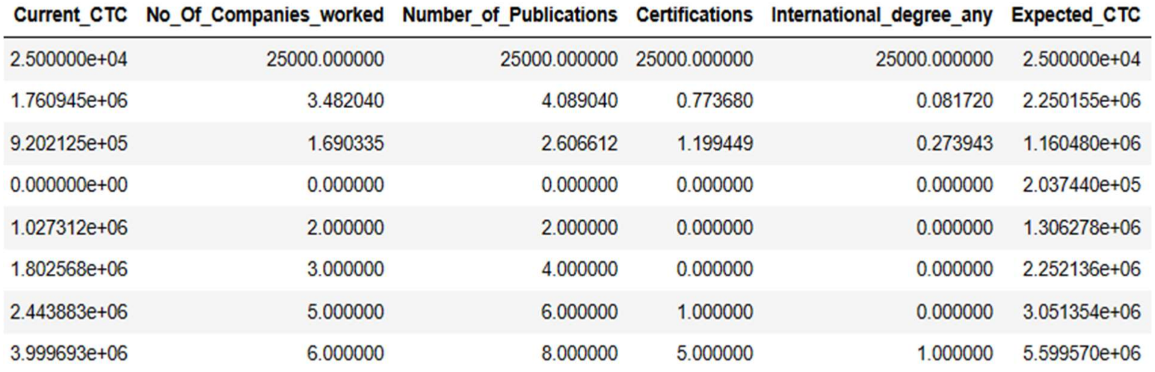
```
IDX                                  0
Applicant_ID                         0
Total_Experience                     0
Total_Experience_in_field_applied    0
Department                        2778
Role                               963
Industry                           908
Organization                       908
Designation                       3129
Education                            0
Graduation_Specialization         6180
University_Grad                   6180
Passing_Year_Of_Graduation        6180
PG_Specialization                 7692
University_PG                     7692
Passing_Year_Of_PG                7692
PHD_Specialization               11881
University_PHD                   11881
Passing_Year_Of_PHD              11881
Curent_Location                      0
Preferred_location                   0
Current_CTC                          0
Inhand_Offer                         0
Last_Appraisal_Rating              908
No_Of_Companies_worked               0
Number_of_Publications               0
Certifications                       0
International_degree_any             0
Expected_CTC                         0
dtype: int64
```

1. Department: 2,778 missing values
2. Designation: 3,129 missing values
3. Graduation_Specialization: 6,180 missing values
4. PG_Specialization, University_PG, and Passing_Year_Of_PG: Over 7,692 missing values each
5. PHD_Specialization, University_PHD, and Passing_Year_Of_PHD: Around 11,881 missing values each.

Data.describe:



	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Passing_Year_Of_Graduation	Passing_Year_Of_PG
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.00000	25000.000000
mean	12500.500000	34993.240080	12.493080	6.258200	2002.14576	2005.414000
std	7217.022701	14390.271591	7.471398	5.819513	7.21629	7.517717
min	1.000000	10000.000000	0.000000	0.000000	1986.00000	1988.000000
25%	6250.750000	22563.750000	6.000000	1.000000	1998.00000	2001.000000
50%	12500.500000	34974.500000	12.000000	5.000000	2002.00000	2006.000000
75%	18750.250000	47419.000000	19.000000	10.000000	2007.00000	2010.000000
max	25000.000000	60000.000000	25.000000	25.000000	2020.00000	2023.000000



Current_CTC	No_Of_Companies_worked	Number_of_Publications	Certifications	International_degree_any	Expected_CTC
2.500000e+04	25000.000000	25000.000000	25000.000000	25000.000000	2.500000e+04
1.760945e+06	3.482040	4.089040	0.773680	0.081720	2.250155e+06
9.202125e+05	1.690335	2.606612	1.199449	0.273943	1.160480e+06
0.000000e+00	0.000000	0.000000	0.000000	0.000000	2.037440e+05
1.027312e+06	2.000000	2.000000	0.000000	0.000000	1.306278e+06
1.802568e+06	3.000000	4.000000	0.000000	0.000000	2.252136e+06
2.443883e+06	5.000000	6.000000	1.000000	0.000000	3.051354e+06
3.999693e+06	6.000000	8.000000	5.000000	1.000000	5.599570e+06

Statistical Summary - Key numerical features include:

Total_Experience: Ranges from 0 to 25 years, with a mean of approximately 12.49 years.

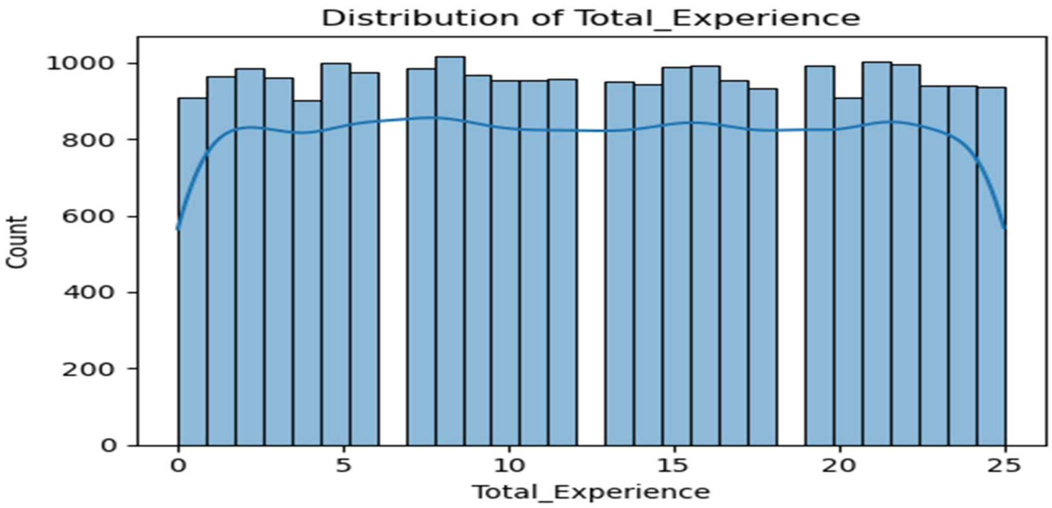
Current_CTC: The current salary ranges from 0 to approximately 3.999 million, with a mean of about 1.76 million.

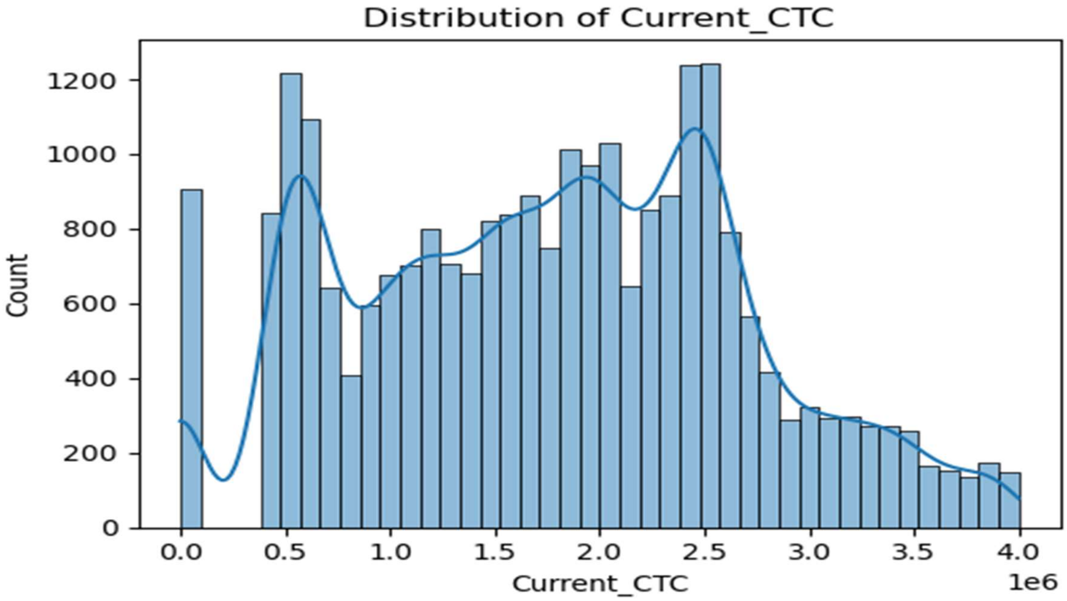
Expected_CTC: The target variable ranges from around 203,744 to 5.599 million, with a mean of approximately 2.25 million.

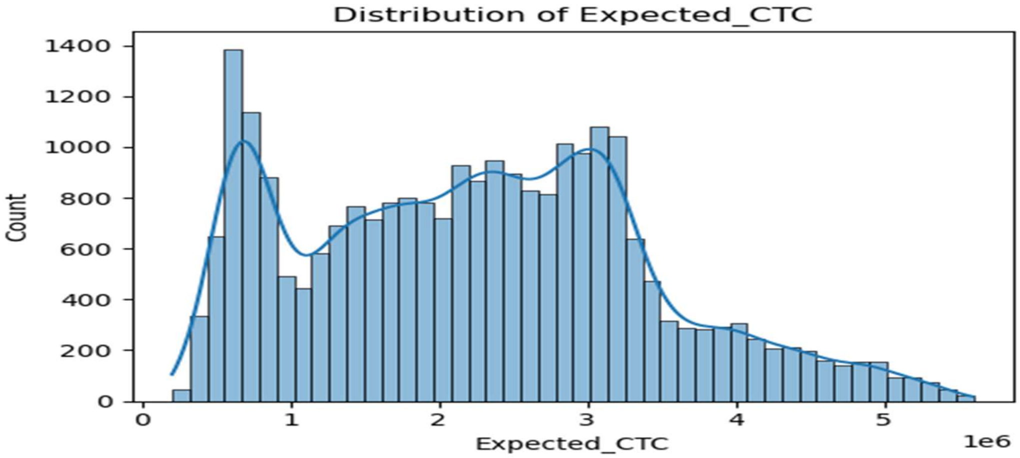
Number_of_Publications and Certifications also show a wide range, indicating diverse academic and professional achievements among applicants.

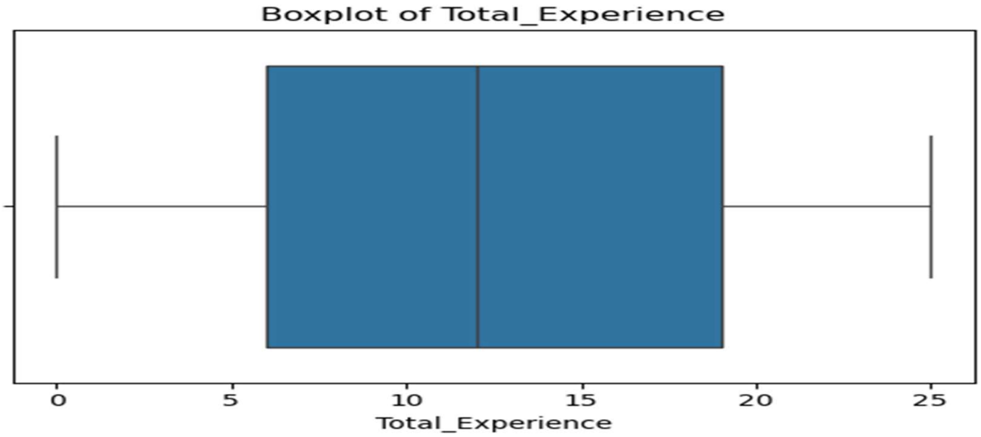
3. Exploratory Data Analysis - a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) b) Bivariate analysis (relationship between different variables, correlations) a) Removal of unwanted variables (if applicable) b) Missing Value treatment (if applicable) d) Outlier treatment (if required) e) Variable transformation (if applicable) f) Addition of new variables (if required)

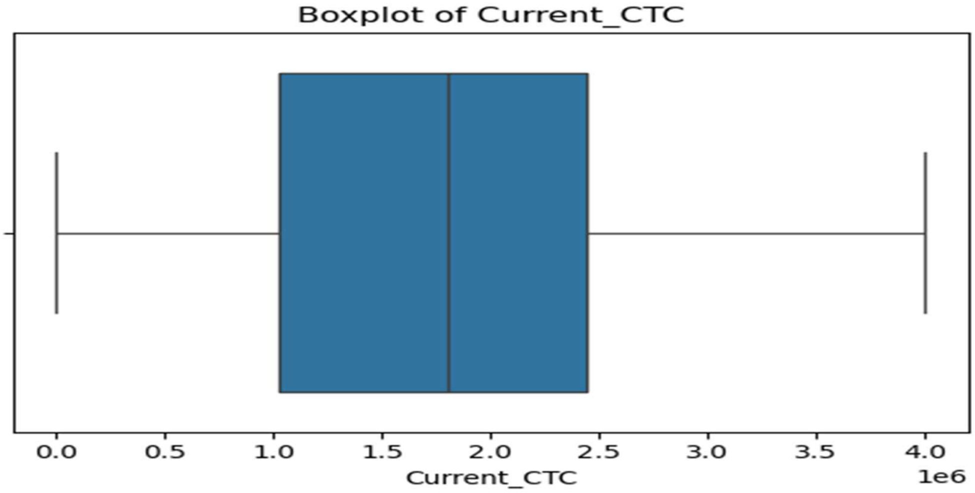
Univariate analysis: Continuous Attributes

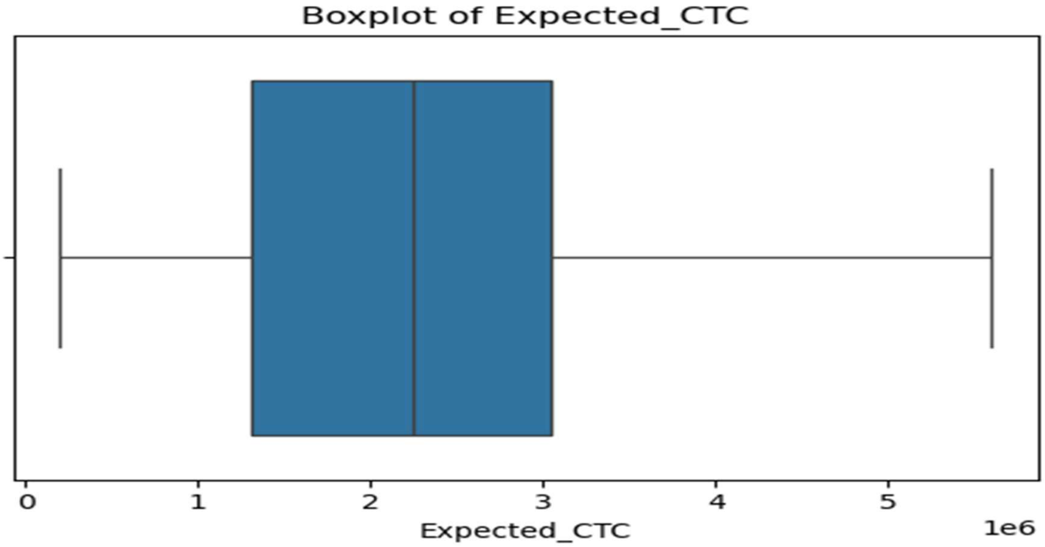




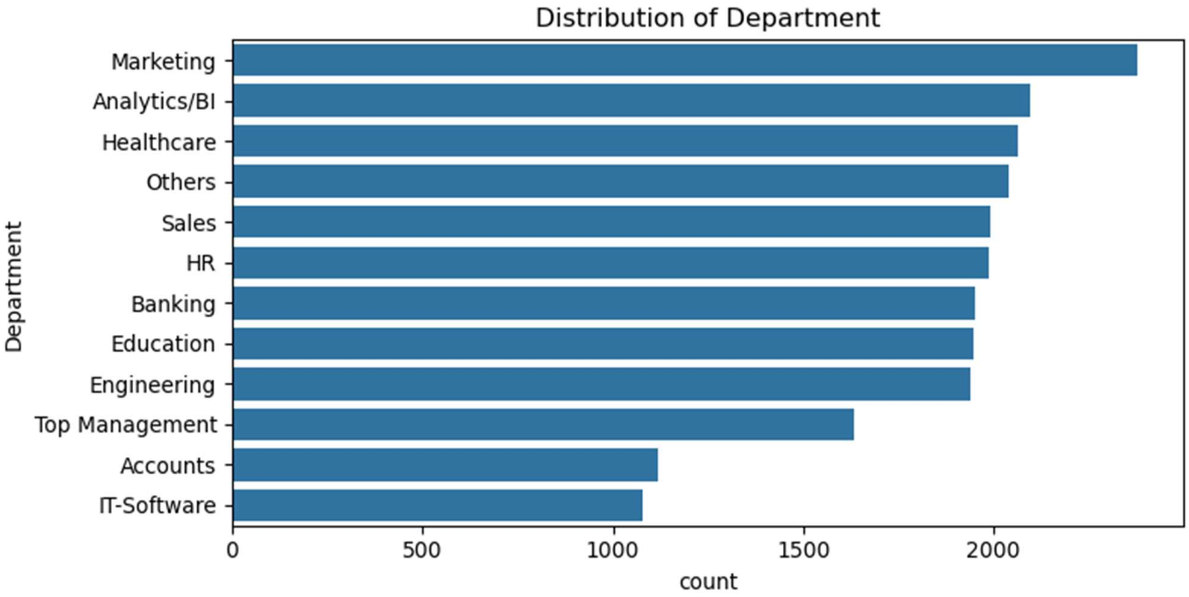

Distribution of Expected_CTC
Count
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
5
Expected_CTC
1e6

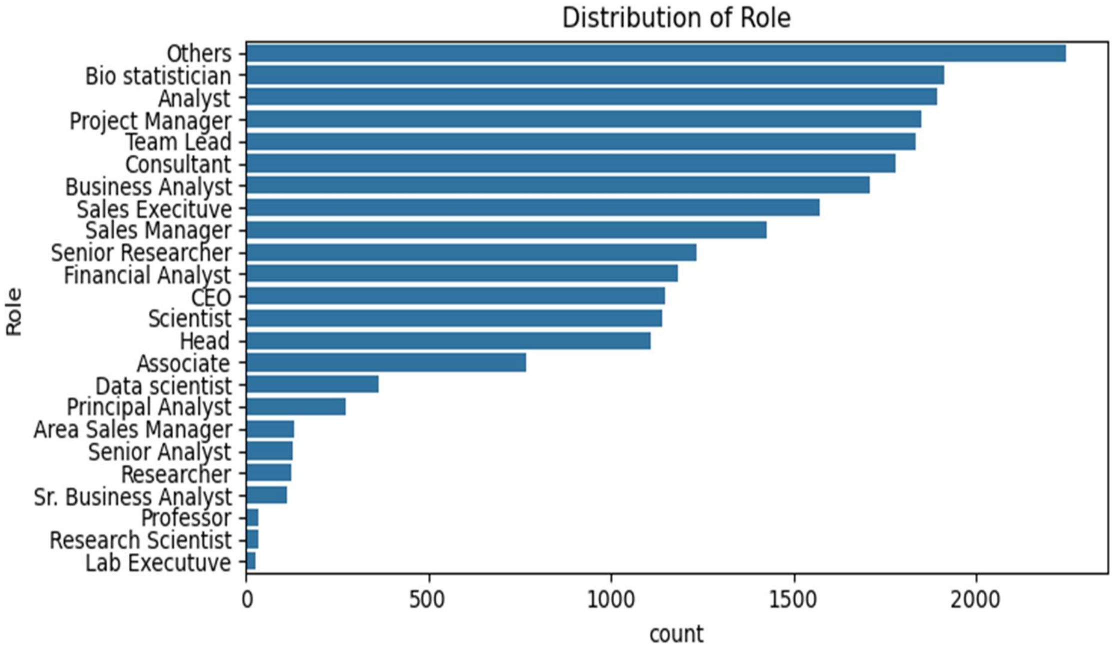

Boxplot of Total_Experience
0
5
10
15
20
25
Total_Experience

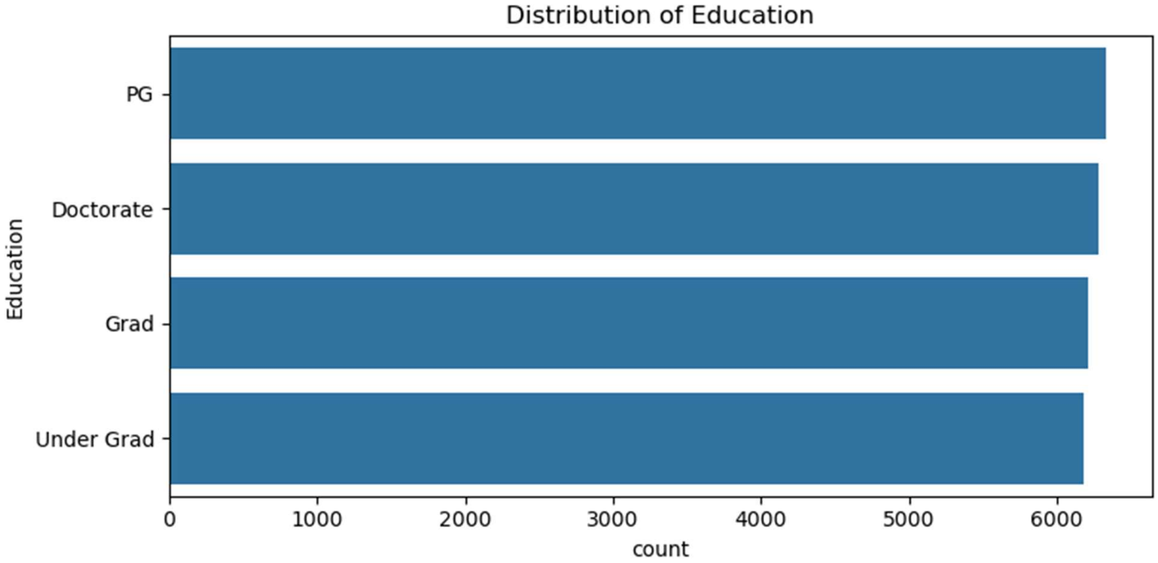

Boxplot of Current_CTC
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Current_CTC
1e6



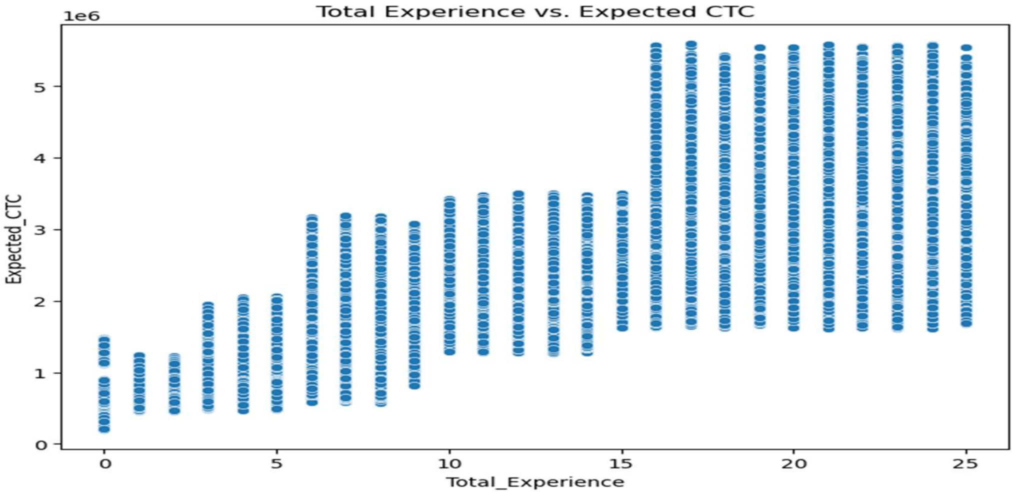
Univariate - Categorical attributes:

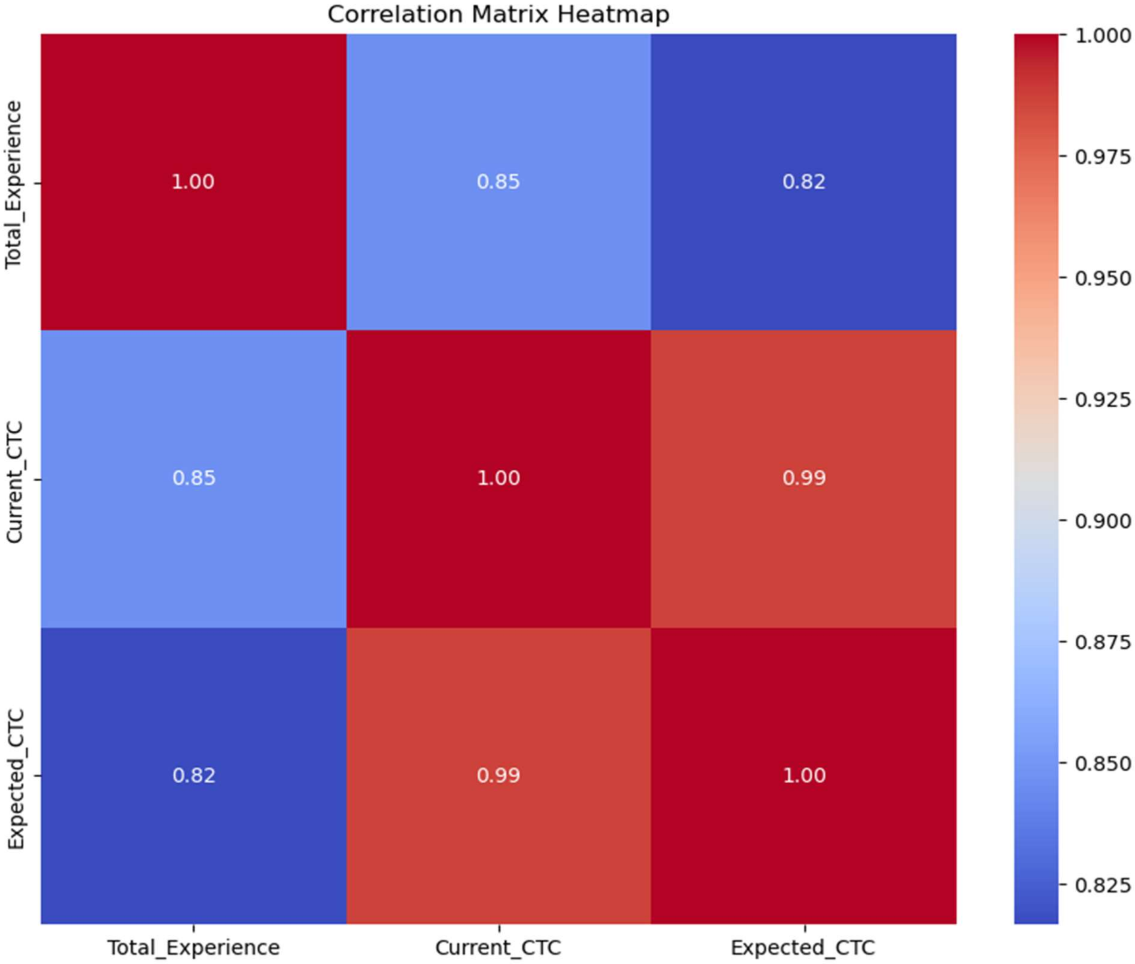


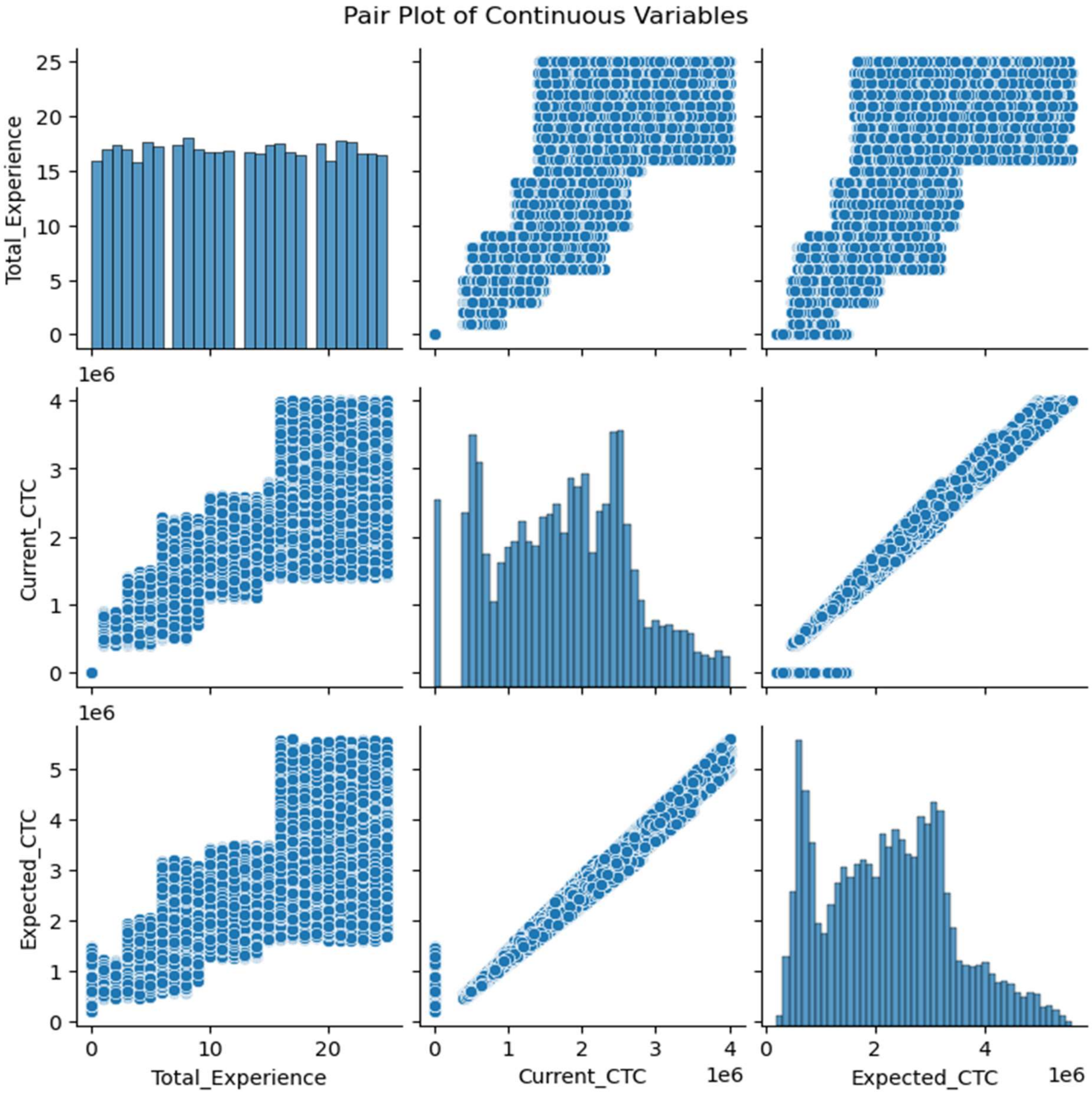

Distribution of Role
Role
Others
Bio statistician
Analyst
Project Manager
Team Lead
Consultant
Business Analyst
Sales Execituve
Sales Manager
Senior Researcher
Financial Analyst
CEO
Scientist
Head
Associate
Data scientist
Principal Analyst
Area Sales Manager
Senior Analyst
Researcher
Sr. Business Analyst
Professor
Research Scientist
Lab Executuve
0
500
1000
1500
2000
count


Distribution of Education
Education
PG
Doctorate
Grad
Under Grad
0
1000
2000
3000
4000
5000
6000
count

Bivariate analysis - Continuous Variables:



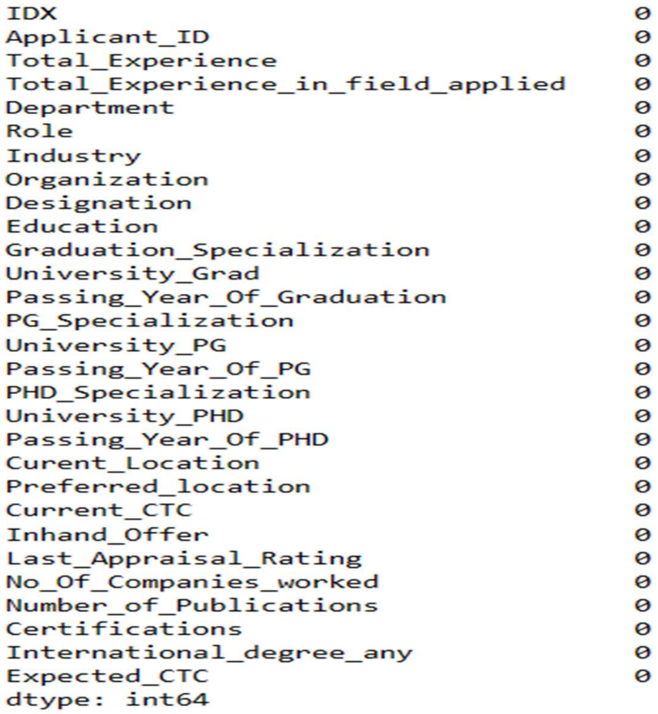




A thorough understanding of the relationships between variables is offered by these visualizations. The heatmap highlights the most strongly correlated pairs of variables and provides a brief summary of correlation coefficients. The scatter plots for variable pairs and the histograms for individual variables are visible in the pair plot, which provides a more thorough examination and is helpful in identifying non-linear relationships or variables that could benefit from transformations.

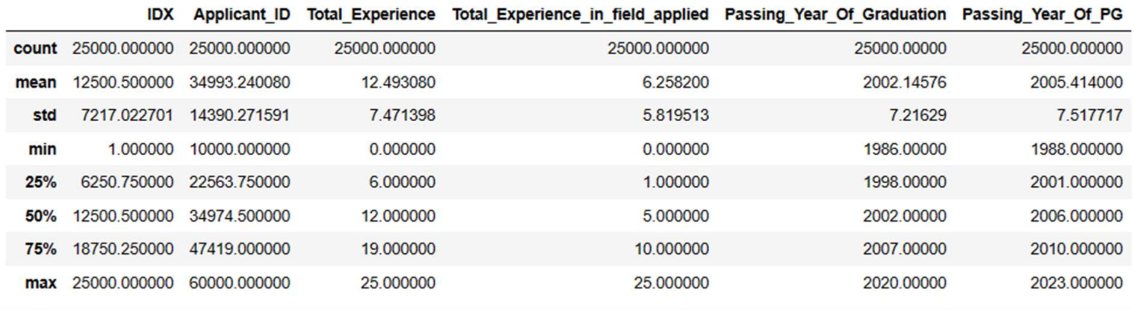
Recall that correlation does not imply causation, and more research may be required to completely comprehend the nature of these relationships. In order to preserve the pair plot's interpretability and clarity when working with a large number of continuous variables, it may be helpful to concentrate on a small number of important variables.

Missing Value treatment:

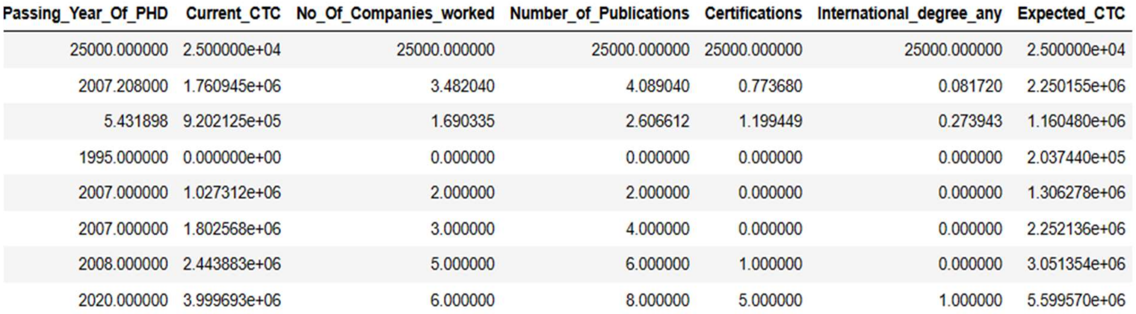


```
IDX                                  0
Applicant_ID                         0
Total_Experience                     0
Total_Experience_in_field_applied    0
Department                           0
Role                                 0
Industry                             0
Organization                         0
Designation                          0
Education                            0
Graduation_Specialization            0
University_Grad                      0
Passing_Year_Of_Graduation           0
PG_Specialization                    0
University_PG                        0
Passing_Year_Of_PG                   0
PHD_Specialization                   0
University_PHD                       0
Passing_Year_Of_PHD                  0
Curent_Location                      0
Preferred_location                   0
Current_CTC                          0
Inhand_Offer                         0
Last_Appraisal_Rating                0
No_Of_Companies_worked               0
Number_of_Publications               0
Certifications                       0
International_degree_any             0
Expected_CTC                         0
dtype: int64
```

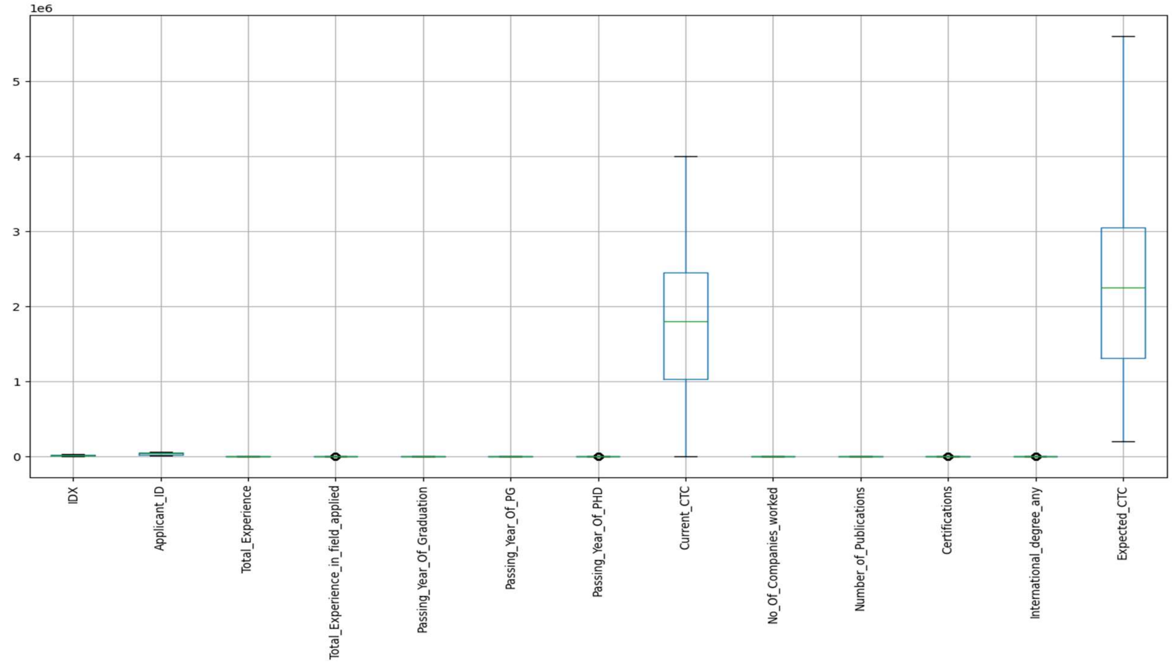
Outlier treatment:



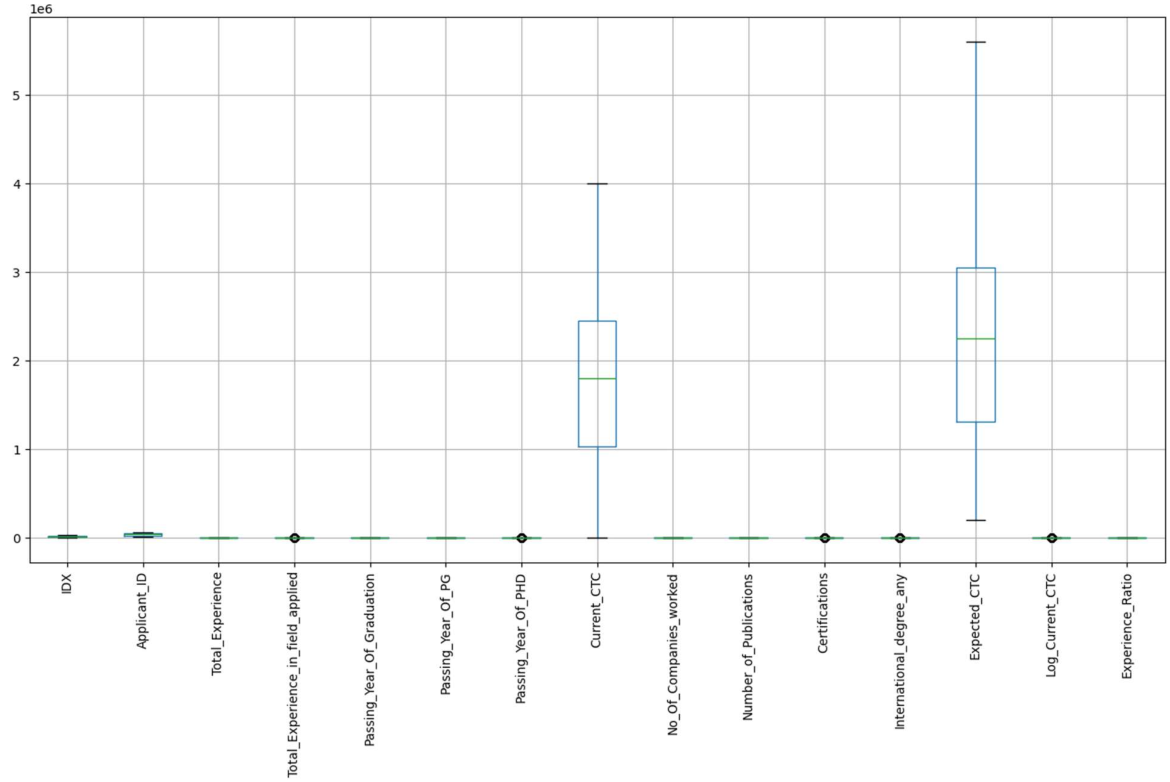
	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Passing_Year_Of_Graduation	Passing_Year_Of_PG
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.00000	25000.000000
mean	12500.500000	34993.240080	12.493080	6.258200	2002.14576	2005.414000
std	7217.022701	14390.271591	7.471398	5.819513	7.21629	7.517717
min	1.000000	10000.000000	0.000000	0.000000	1986.00000	1988.000000
25%	6250.750000	22563.750000	6.000000	1.000000	1998.00000	2001.000000
50%	12500.500000	34974.500000	12.000000	5.000000	2002.00000	2006.000000
75%	18750.250000	47419.000000	19.000000	10.000000	2007.00000	2010.000000
max	25000.000000	60000.000000	25.000000	25.000000	2020.00000	2023.000000



Passing_Year_Of_PHD	Current_CTC	No_Of_Companies_worked	Number_of_Publications	Certifications	International_degree_any	Expected_CTC
25000.000000	2.500000e+04	25000.000000	25000.000000	25000.000000	25000.000000	2.500000e+04
2007.208000	1.760945e+06	3.482040	4.089040	0.773680	0.081720	2.250155e+06
5.431898	9.202125e+05	1.690335	2.606612	1.199449	0.273943	1.160480e+06
1995.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	2.037440e+05
2007.000000	1.027312e+06	2.000000	2.000000	0.000000	0.000000	1.306278e+06
2007.000000	1.802568e+06	3.000000	4.000000	0.000000	0.000000	2.252136e+06
2008.000000	2.443883e+06	5.000000	6.000000	1.000000	0.000000	3.051354e+06
2020.000000	3.999693e+06	6.000000	8.000000	5.000000	1.000000	5.599570e+06

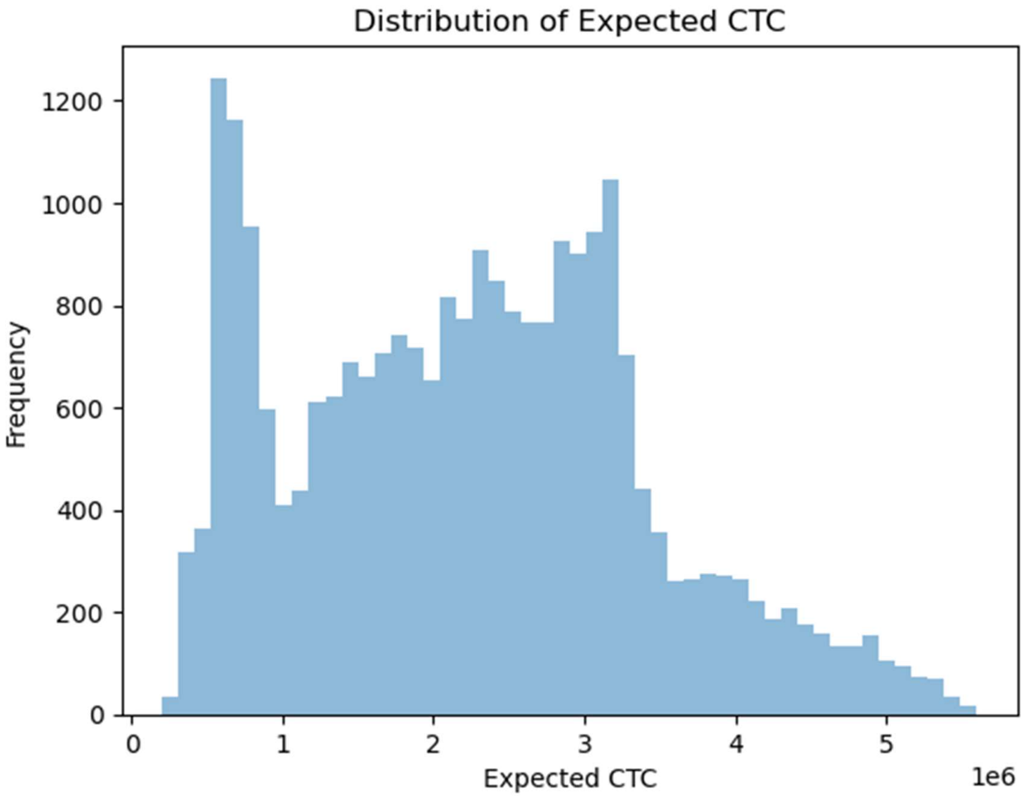


Variable transformation (if applicable) - Addition of new variables (if required)



Business insights from EDA - a) is the data unbalanced? If so, what can be done? Please explain in the context of the business b) any business insights using clustering (if applicable) c) Any other business insights

Is the data unbalanced?

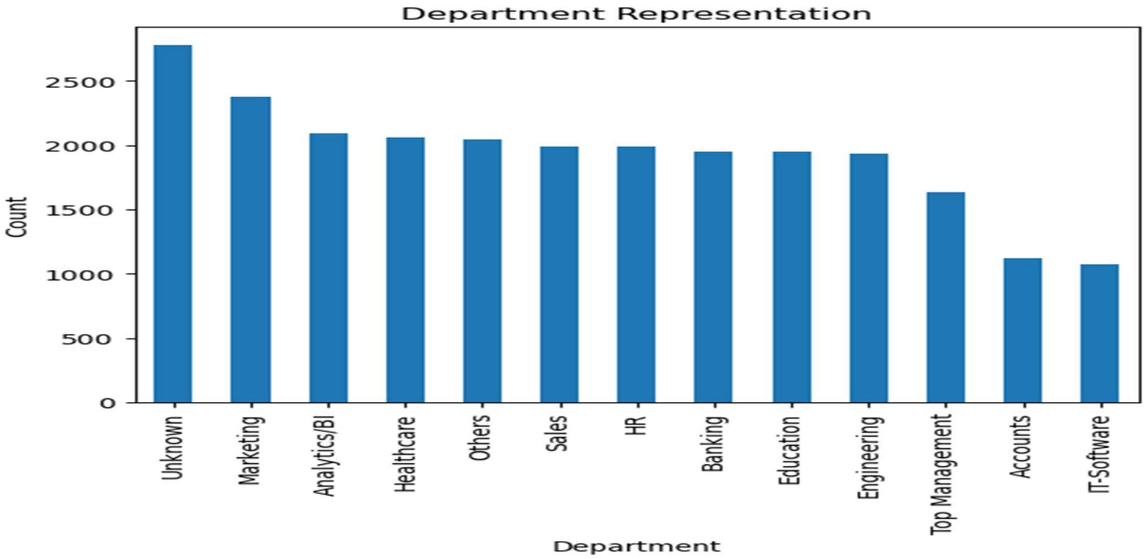


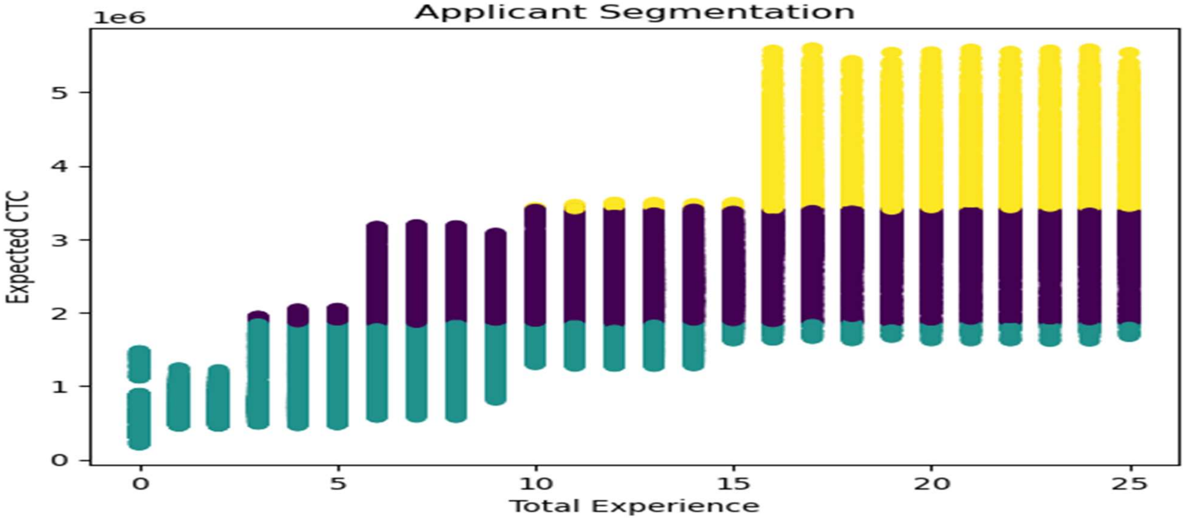
Checking for Imbalance in the Dataset:

Expected_CTC Distribution: By plotting the histogram of Expected_CTC, you sought to understand its spread across different salary ranges. A skewed distribution could indicate that most applicants fall into specific salary brackets, suggesting potential imbalances in salary expectations.

Department Representation: The bar plot of Department counts was intended to reveal any disproportionate representation of departments within the applicant pool. A balanced representation is crucial for building a fair and unbiased salary determination model.

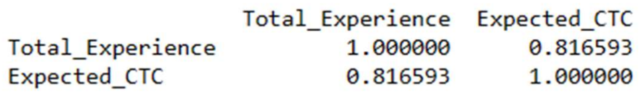
Any business insights using clustering:



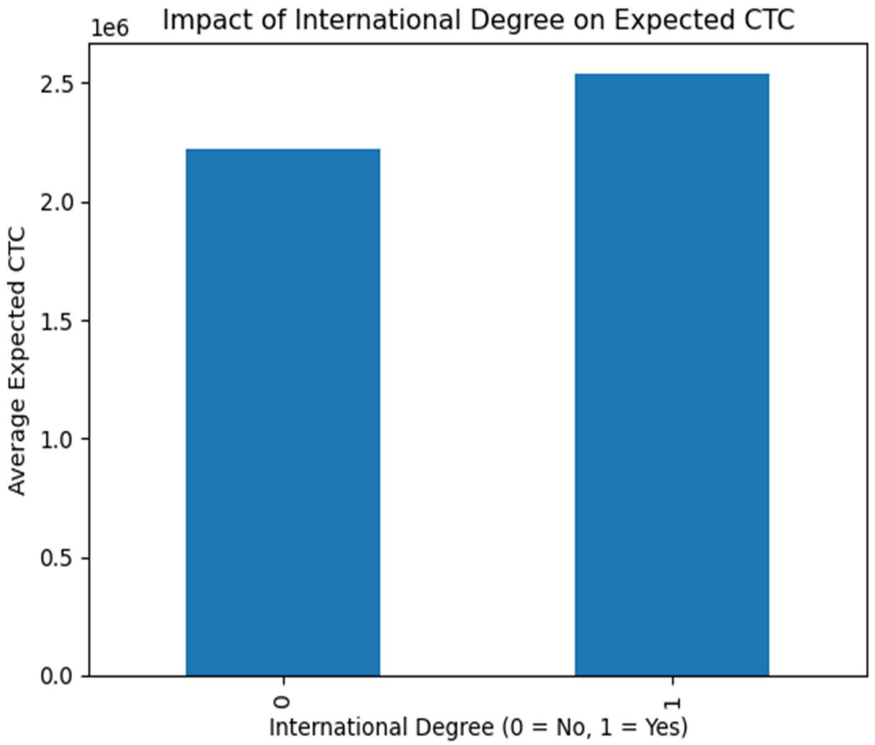


Clustering applicants based on their total experience and expected CTC could help identify distinct groups within the dataset, such as high-experience, high-expectation applicants versus those with less experience and lower salary expectations. This segmentation offers insights into the workforce's composition and can guide targeted salary structures.

Other Business Insights - Exploring correlations between numerical features, for instance, between Total_Experience and Expected_CTC



```
                  Total_Experience  Expected_CTC
Total_Experience          1.000000      0.816593
Expected_CTC              0.816593      1.000000
```



Correlation Analysis: Exploring the correlation between Total_Experience and Expected_CTC might have revealed how strongly these two factors are related, indicating the extent to which experience influences salary expectations.

Impact of International Degrees: By comparing the average Expected_CTC between those with and without international degrees, you investigated whether holding an international degree influences salary expectations. This could inform whether such qualifications are valued differently in salary considerations.