Capstone Project Business Report

Submitted by:

Dev Kumar

Batch-PG-DSBA – Apr'21

TABLE OF CONTENTS:

LIST OF TABLES:

LIST OF FIGURES:

1.	Introduction
	Brief introduction about the problem statement and the need of solving it.

Defining business problem statement:

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.

- A system to screen candidates and score them, thus enabling candidate filtering and reducing the workload of recruitment officers and Company cost.

Need of the study/project:

Recruitment is a very major challenge to take right employee for the company or an Organization. In many research and several studies the impact in the industry. On the recruiter side, it is very important to take a right employee for the company who will help to grow the organization in terms of social and monetary value.

- The goal of this project is to predict salary of a person using multiple features like total experience , educational background , total experience in relevant field , Rating of employee and many other features that will help to solve this particular problem.
- With the help Exploratory data analysis and fetaure selection methods and right approach according to the domain of that business problem will led to achieved this type of Regression Problem to achecive our goal to predict the right Expected_CTC to hire a person for the company with good accuracy that will helps to grow the organization and economy of the country.

- Accurate and proper recruitment of employees is a key element in the business strategy of every company due to its impact on companies productivity and competitiveness.

- In the data driven world, recruitment processes have evolved into complex tasks involving rigorous evaluations and interviews of candidates, with the goal of hiring the best suited professionals for the company.

- With the advancement of Internet and the web development , online Recruitment has become an essential element of all hiring strategies. Many websites, such as Naukri.com, Indeed.com and like LinkedIn, help companies and job seekers to find the best possible matches.

2.	EDA and Business Implication
	Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?
	Both visual and non-visual understanding of the data.

First of all we check the data dictionary of the dataset in which brief description of column/variables is given.

Data dictionary:

IDX	Index
Applicant_ID	Application ID
Total_Experience	Total industry experience
Total_Experience_in_field_applied	Total experience in the field applied for (past work experience that is relevant to the job)
Department	Department name of current company
Role	Role in the current company
Industry	Industry name of current field
Organization	Organization name
Designation	Designation in current company
Education	Education
Graduation_Specialization	Specialization subject in graduation
University_Grad	University or college in Graduation
Passing_Year_Of_Graduation	Year of passing Graduation
PG_Specialization	Specialization subject in Post-Graduation
University_PG	University or college in Post-Graduation
Passing_Year_Of_PG	Year of passing Post Graduation
PHD_Specialization	Specialization subject in Post-Graduation
University_PHD	University or college in Post Doctorate
Passing_Year_Of_PHD	Year of passing PHD
Curent_Location	Curent Location
Preferred_location	Preferred location to work in the company applied
Current_CTC	Current CTC
Inhand_Offer	Holding any offer in hand (Y: Yes, N:No)
Last_Appraisal_Rating	Last Appraisal Rating in current company
No_Of_Companies_worked	No. of companies worked till date
Number_of_Publications	Number of papers published
Certifications	Number of relevant certifications completed
International_degree_any	Hold any international degree (1: Yes, 0: No)
Expected_CTC	Expected CTC (Final CTC offered by Delta Ltd.)

Table 1 Data Dictionary for above problem statement

Lets have a look on the top 5 and last 5 observation/rows of the dataset

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	...	Curent_L
0	1	22753	0	0	NaN	NaN	NaN	NaN	NaN	PG	...	G
1	2	51087	23	14	HR	Consultant	Analytics	H	HR	Doctorate	...	Ba
2	3	38413	21	12	Top Management	Consultant	Training	J	NaN	Doctorate	...	Ahm
3	4	11501	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	...	
4	5	58941	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	...	Ahm

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	...	Cı
24995	24996	25550	18	13	Engineering	Project Manager	Automobile	I	Assistant Manager	PG	...	
24996	24997	53442	12	8	HR	Others	Analytics	B	Sr.Manager	Under Grad	...	
24997	24998	15777	22	8	Banking	Head	Insurance	D	Software Developer	Under Grad	...	
24998	24999	57616	25	8	Marketing	CEO	BFSI	D	Marketing Manager	PG	...	
24999	25000	20788	8	0	Banking	Consultant	Automobile	P	Sr.Manager	Grad	...	

5 rows × 29 columns

Table 2 Head (Top 5) rows and Tail(Last 5) rows of the dataset

Note: For More details kindly go to python jupyter code file that is also share.

With the help of shape function we can check the the rows and columns in the dataset.

- Shape attribute tells us number of observations and variables we have in the data set. It is used to check the dimension of data.
- The data set has 25000 observations (rows) and 29 variables (columns) in the dataset.

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	...
count	25000.000000	25000.000000	25000.000000	25000.000000	22222	24037	24092	24092	21871	25000	...
unique	NaN	NaN	NaN	NaN	12	24	11	16	18	4	...
top	NaN	NaN	NaN	NaN	Marketing	Others	Training	M	HR	PG	...
freq	NaN	NaN	NaN	NaN	2379	2248	2237	1574	1648	6326	...
mean	12500.500000	34993.240080	12.493080	6.258200	NaN	NaN	NaN	NaN	NaN	NaN	...
std	7217.022701	14390.271591	7.471398	5.819513	NaN	NaN	NaN	NaN	NaN	NaN	...
min	1.000000	10000.000000	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	...
25%	6250.750000	22563.750000	6.000000	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	...
50%	12500.500000	34974.500000	12.000000	5.000000	NaN	NaN	NaN	NaN	NaN	NaN	...
75%	18750.250000	47419.000000	19.000000	10.000000	NaN	NaN	NaN	NaN	NaN	NaN	...
max	25000.000000	60000.000000	25.000000	25.000000	NaN	NaN	NaN	NaN	NaN	NaN	...

11 rows × 29 columns

Table 3 Five point Description summary of the dataset

- From the above table we can infer the count,mean, std , 25% , 50% ,75% and min & max values of the all numeric variables present in the dataset.
- From the above table we can infer the count,unique,top,freq of all the categorical variables present in 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  Current_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 4 Info summary of the column of the dataset

Insights

- From the above results we can see that there is null values present in the many column of the dataset.
- Their are total 25000 rows & 29 columns in this dataset,indexed from 0 to 24999.
- Out of 10 variables 3 are float64 , 16 variables are object and 10 variable is int64.
- Memory used by the dataset: 5.5+ MB.

Note:-There is a miss spell word of Curent_Location so we have to change the column name to Current_Location

Let have a look on the Skewness of the continuous variable of the Dataset -

Number_of_Publications	-0.075217
No_Of_Companies_worked	-0.068026
Passing_Year_Of_PG	-0.066166
Total_Experience	0.004109
Percentage_in_Relevant_Field	0.005018
Passing_Year_Of_PHD	0.014436
Passing_Year_Of_Graduation	0.061408
Current_CTC	0.097643
Expected_CTC	0.331972
Total_Experience_in_field_applied	0.961951
Certifications	1.610907
International_degree_any	3.054017
dtype: float64	

Table 5 Info Skewness stats of the dataset

Insights

- The variables with skewness value greater than 1 or less than -1 indicates a highly skewed distribution.
- The variables value between 0.5 and 1 or -0.5 and -1 is moderately skewed.
- The variables with value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.

Univariate Analysis on Numerical/continuous variable: Histogram & Boxplot

- A histogram provides a visual representation of the distribution of a dataset: location, spread and skewness of the data; it also helps to visualize whether the distribution is symmetric or skewed left or right.

- A box plot also known as Five Number Summary, summarizes data using the median, upper quartile, lower quartile, and the minimum and maximum values. It allows you to see important characteristics of the data. This also help us to visualize outliers in the data set.

Variable Name :Total_Experience, dtype: float64

count	25000.000000
mean	12.493080
std	7.471398
min	0.000000
25%	6.000000
50%	12.000000
75%	19.000000
max	25.000000

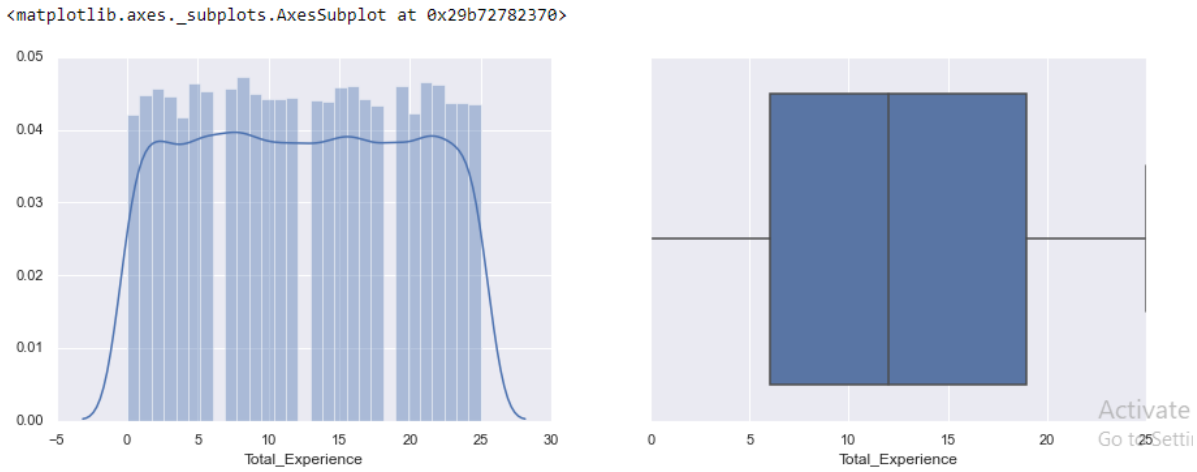


Figure 1 Histogram and Boxplot of Total_Experience variable

Insights

- Total Experience ranges from a minimum of 0 years to maximum of 25 years.
- The average Total Experience is around 12 years.
- The standard deviation of Total Experience is around 7.47.
- 25% , 50% (median) and 75 % of the Total_Experience: are 6 , 12 and 19 years.
- Skewness indicating that the ditribution is slightly right skewed.
- Total Experience does not have any outliers.

Variable Name: Total_Experience_in_field_applied, dtype: float64

count	25000.000000
mean	6.258200
std	5.819513
min	0.000000
25%	1.000000
50%	5.000000
75%	10.000000
max	25.000000

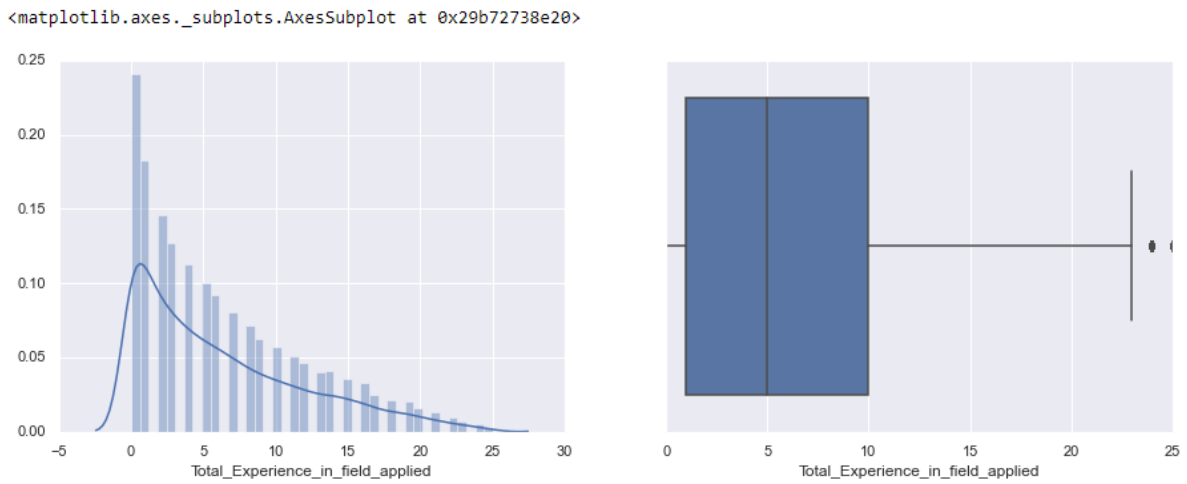


Figure 2 Histogram and Boxplot of Total_Experience_in_field_applied variable

Insights

- Total_Experience_in_field_applied ranges from a minimum of 0 years to maximum of 25 years.

- The average of Total_Experience_in_field_applied is around 5 years.
- The standard deviation of Total_Experience_in_field_applied is around 5.819.
- 25% , 50% (median) and 75 % of the Total_Experience_in_field_applied: are 1 , 5 and 10 years.
- Skewness indicating that the ditribution is right skewed.
- Total_Experience_in_field_applied does have outliers.

Variable Name: Passing_Year_Of_Graduation, dtype: float64

count	18820.000000
mean	2002.193624
std	8.316640
min	1986.000000
25%	1996.000000
50%	2002.000000
75%	2009.000000
max	2020.000000

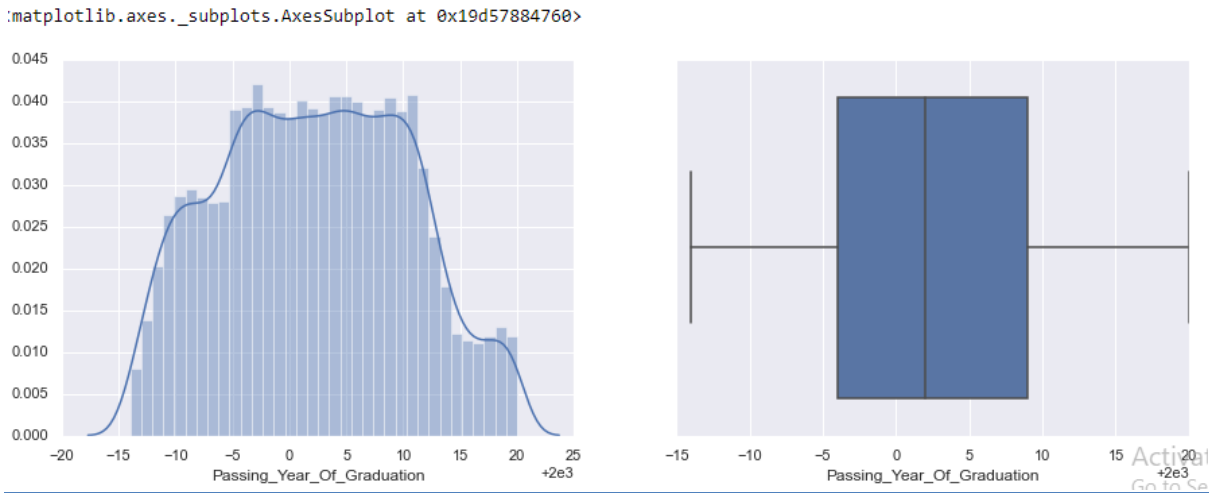


Figure 3 Histogram and Boxplot of Passing_Year_Of_Graduation variable

Insights

- Passing_Year_Of_Graduation ranges from a 1986 year to 2020 year.
- The average year of Passing_Year_Of_Graduation is around 2002.
- The standard deviation of Passing_Year_Of_Graduation is around 8.31
- 25% , 50% (median) and 75 % of the year are : are 1996 , 2002 and 2009 respectively.
- Skewness indicating that it is normally ditribution.

- Passing_Year_Of_Graduation does not have outliers.

Variable Name: Passing_Year_Of_PG, dtype: float64

count	17308.000000
mean	2005.153571
std	9.022963
min	1988.000000
25%	1997.000000
50%	2006.000000
75%	2012.000000
max	2023.000000

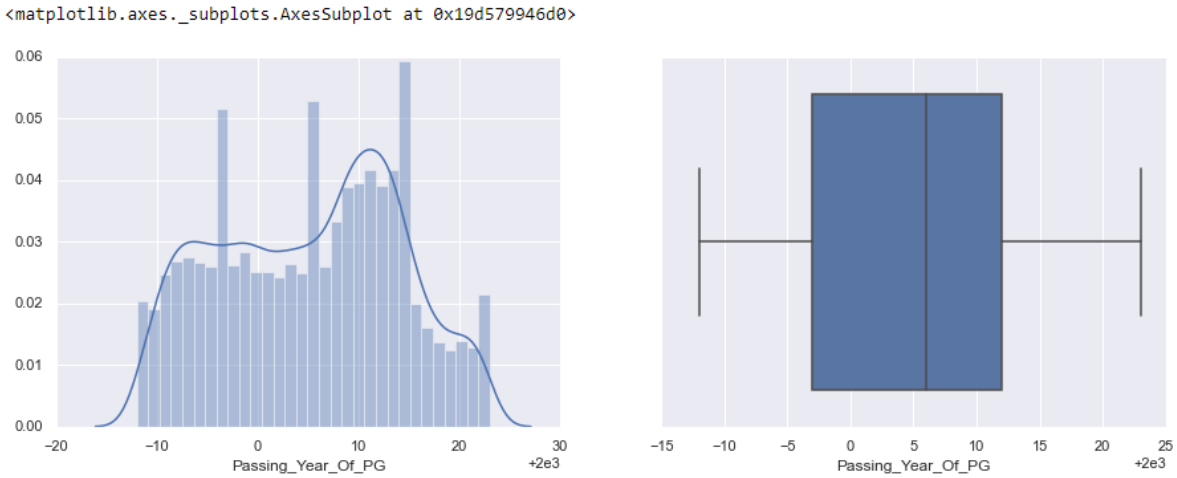


Figure 4 Histogram and Boxplot of Passing_Year_Of_PG variable

Insights

- Passing_Year_Of_PG ranges from a 1988 year to 2023 year.
- The average year of Passing_Year_Of_PG is around 2006.
- The standard deviation of Passing_Year_Of_PG is around 9.02
- 25% , 50% (median) and 75 % of the year are : are 1997 , 2006 and 2012 respectively.
- Skewness indicating that it is normally ditribution.
- Passing_Year_Of_PG does not have outliers.

Variable Name: Passing_Year_Of_PHD, dtype: float64

count	13119.000000

mean	2007.396372
std	7.493601
min	1995.000000
25%	2001.000000
50%	2007.000000
75%	2014.000000
max	2020.000000

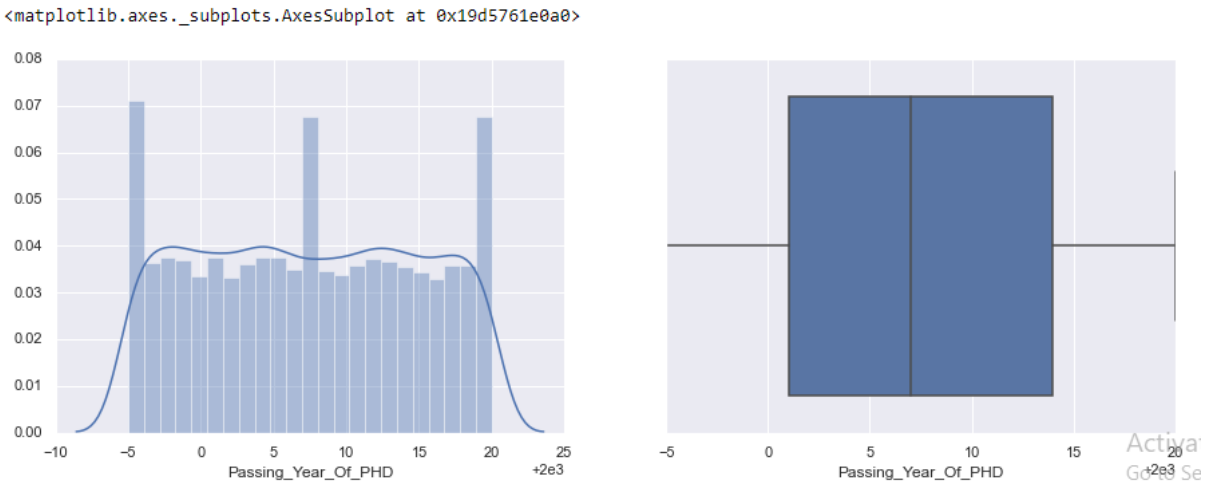


Figure 5 Histogram and Boxplot of Passing_Year_Of_PHD variable

Insights

- Passing_Year_Of_PHD ranges from a 1995 year to 2020 year.
- The average year of Passing_Year_Of_PHD is around 2007.
- The standard deviation of Passing_Year_Of_PHD is around 7.49
- 25% , 50% (median) and 75 % of the year are : are 2001 , 2007 and 2014 respectively.
- Skewness indicating that it is normally ditribution.
- Passing_Year_Of_PHD does not have outliers.

Variable Name: Current_CTC, dtype: float64

count	2.500000e+04
mean	1.760945e+06
std	9.202125e+05
min	0.000000e+00
25%	1.027312e+06
50%	1.802568e+06
75%	2.443883e+06
max	3.999693e+06

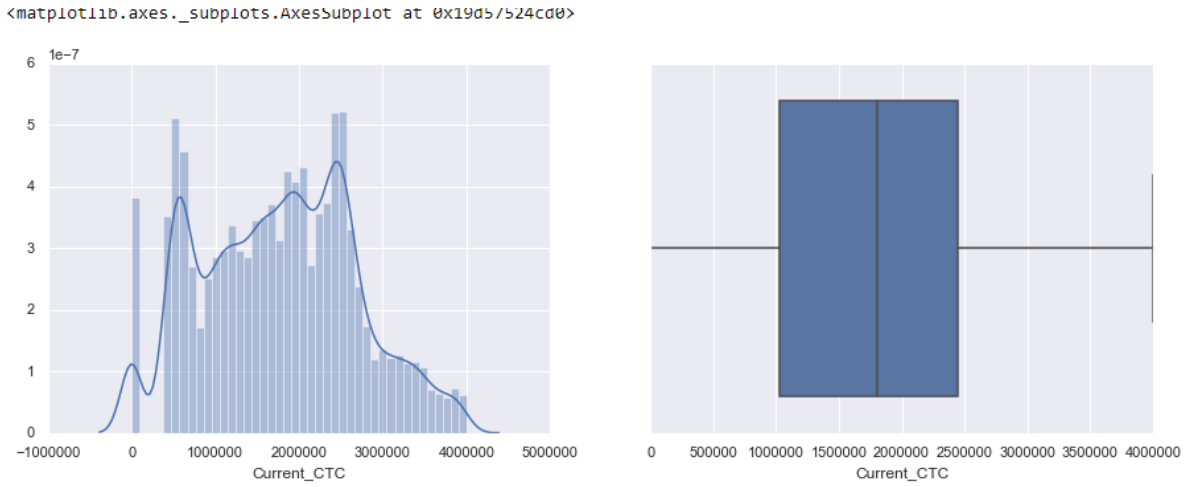


Figure 6 Histogram and Boxplot of Current_CTC variable

Insights

- Current_CTC ranges from 0 to 3999693 .
- The average Current CTC is around 1802568.
- The standard deviation of Current_CTC is around 920212.5
- 25% , 50% (median) and 75 % of the Current_CTC are :1027312, 1802568 and 2443883 respectively.
- Skewness indicating that it is slightly right distribution.
- Current_CTC does not have outliers.

Univariate Analysis on Discrete Continuous variable:-

Variable Name: No_Of_Companies_worked, dtype: float64

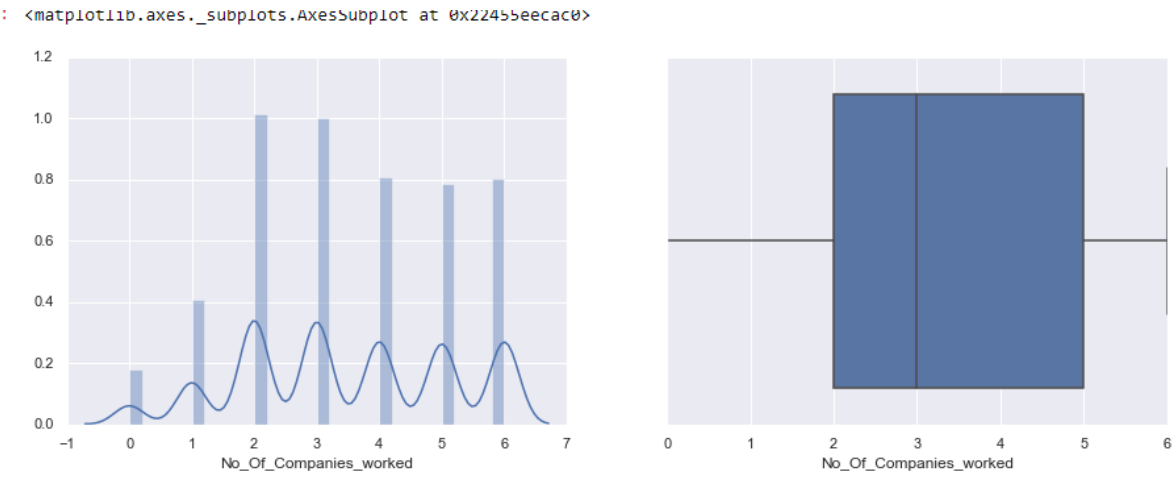


Figure 7 Histogram and Boxplot of No_Of_Companies_worked variable

Insights -

- No. of companies worked till date ranges from a minimum of 0 to maximum of 6.
- Mean No. of companies worked till date is near by 3.48.
- 25% , 50% (median) and 75 % of No. of companies worked till date are 2 , 3 and 5.

Variable Name: Number_of_Publications, dtype: float64

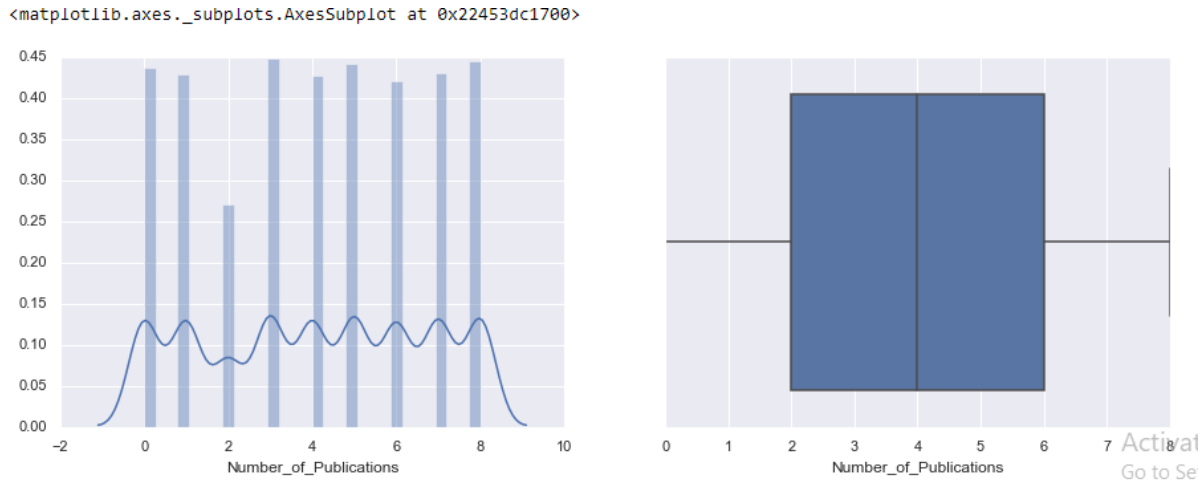


Figure 8 Histogram and Boxplot of Number_of_Publications variable

Insights -

- Number_of_Publications range is from a min of 0 to max of 8.
- Average Number_of_Publications is near by 4.
- 25% , 50% and 75 % of Number_of_Publications are 2 , 4 and 6.
- Number_of_Publications does not have any outliers.

Variable Name: Certifications, dtype: float64

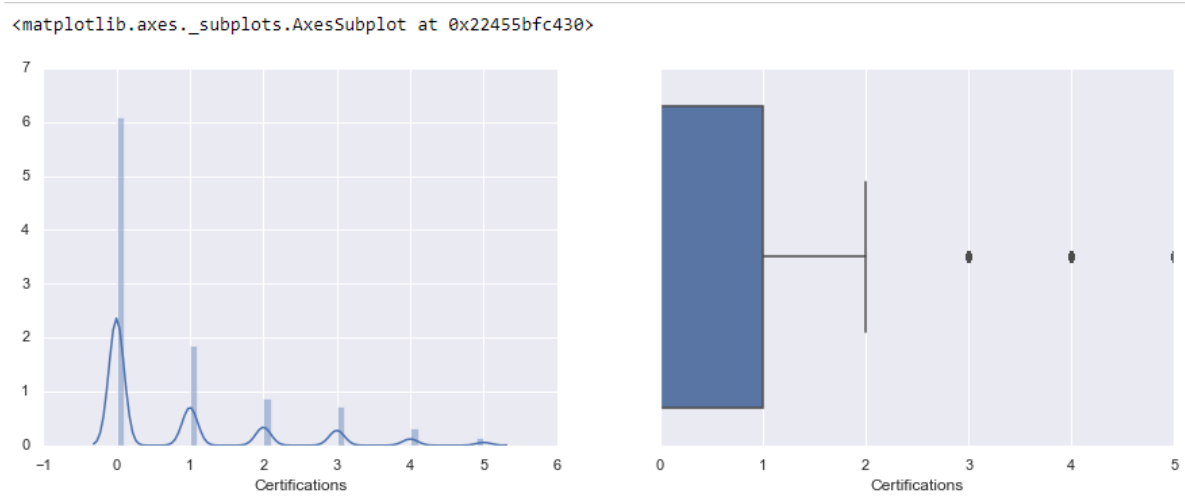


Figure 9 Histogram and Boxplot of Certifications variable

Insights -

- Number of relevant certifications completed ranges from a minimum of 0 to maximum of 5.
- Average Number of relevant certifications completed is around 1.
- 25% , 50% and 75 % of Number of relevant certifications completed are 0 , 0 and 1.
- Distribution is right-skewed.
- certifications have some outliers.

Univariate Analysis on categorical variable-For this we can use countplot or piechart.

A countplot is kind of like a histogram or a bar graph for categorical variables.

Note:-For More Details Explanation in terms of Contibution of each label go to Python Code File.

Analysis on Some Categorical Variable with the help of Countplot:

Count plot of " Department " Categorical Variable shown below:

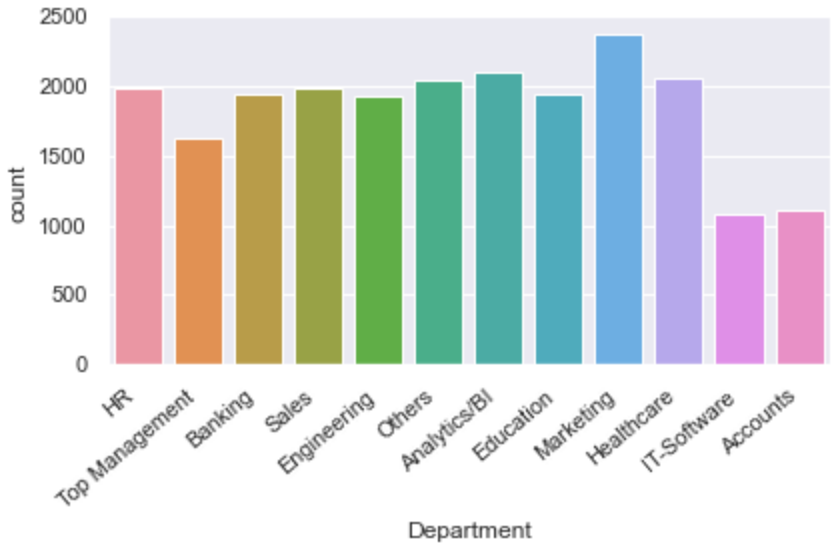


Figure 10 Countplot of Department variable

- There are twelve types of Department present in the data named as 'Marketing' , 'Top Management' ,'Accounts' and 'IT-Software'. 'Analytics/BI' , 'Healthcare' , 'Others' , 'Sales' , 'HR' ,'Banking' , 'Education' ,'Engineering' ,
- Mostly people or observations belongs to Marketing Department (10.70%) ,Analytics/BI (9.43%) , Health-care (9.27%) and others (9.18%)
- Marketing label has the most no of frequency in Department Column where as IT-Software has least no of Frequency.

Count plot of " Role " Categorical Variable shown below:

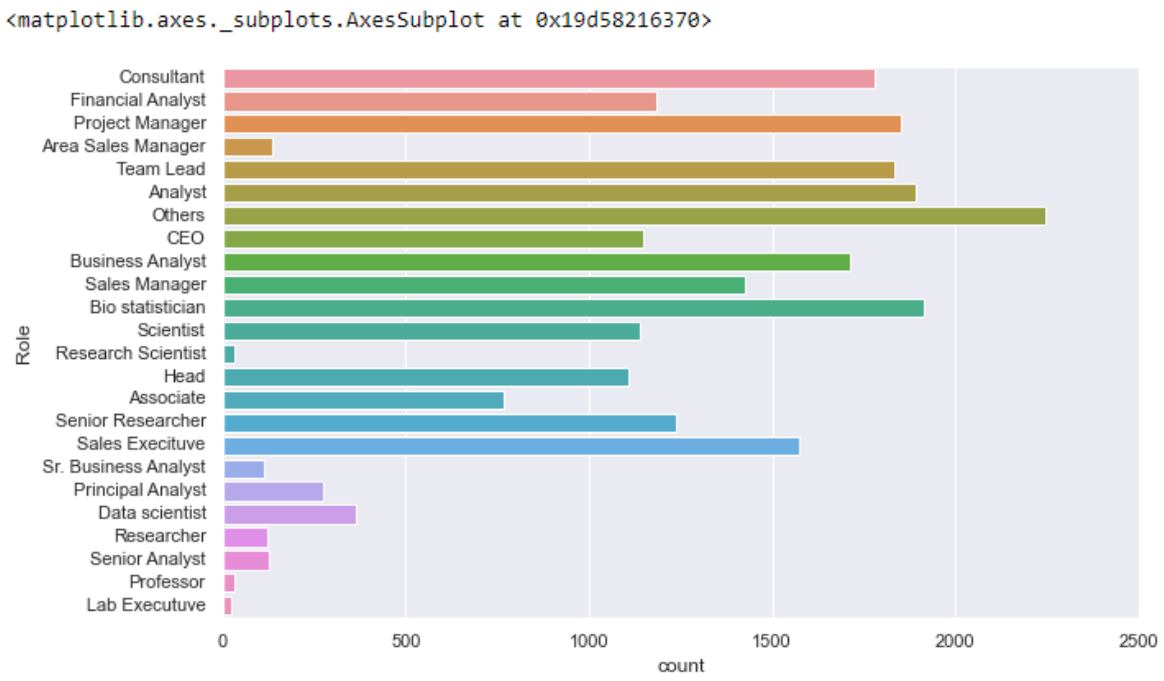


Figure 11 Countplot of Role variable

Insights

- Others label has highest no of count where as Lab Executive has least no of count in terms of frequency.
- Research Scientist, Sr. Business Analyst, Area Sales Manager, Professor, Researcher,Senior Analyst are on the lower side in terms of frequency count.

Count plot of " Industry " Categorical Variable shown below:

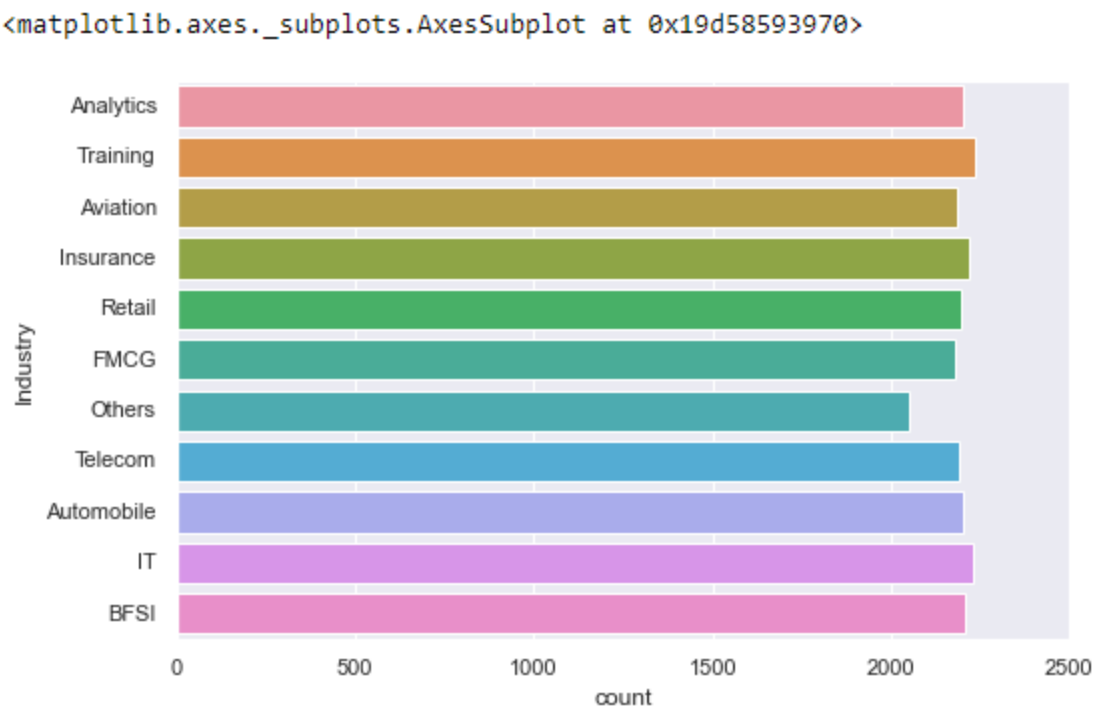


Figure 12 Countplot of Industry variable

Insights

- Most of the Labels have the similar kind of frequency .
- Training is on topmost in frequency where as Others has least count.
- Most labels have around 9 % of contribution in terms of percentage.

Count plot of " Designation " Categorical Variable shown below:

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d58659790>
```

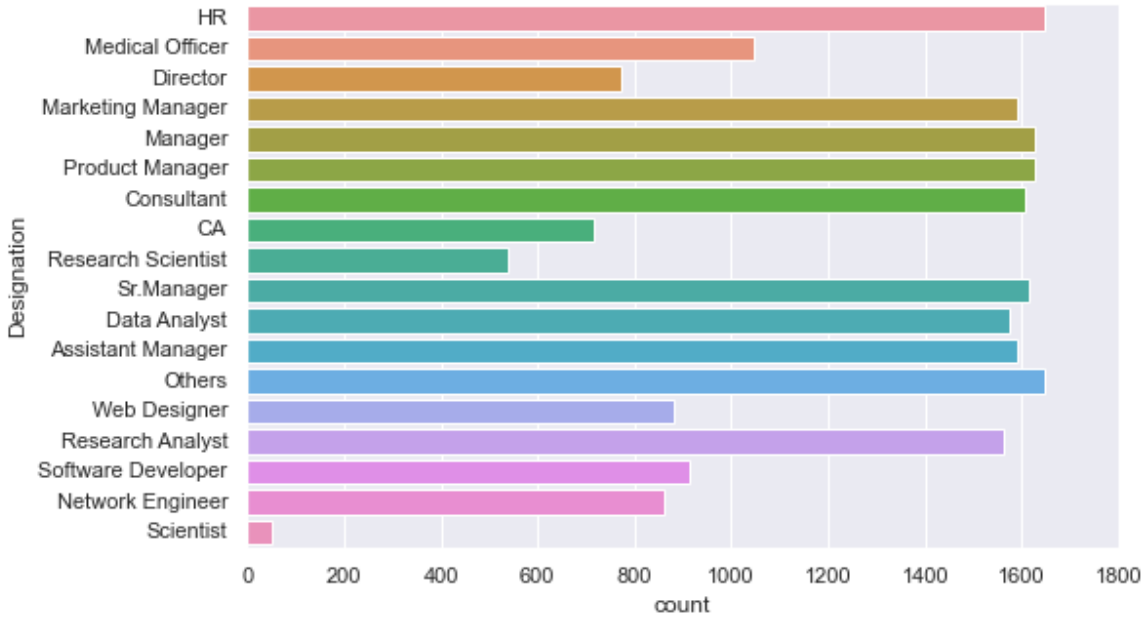


Figure 13 Countplot of Designation variable

Insights

- Most of the Labels have the similar kind of frequency like Marketing Manager,Manager,Product Manager,Sr. Manager, Data Analyst , Assistant Manager ,Others .Reasearch Analyst each contibute around 7 % of total.

- HR and Others label is on topmost in frequency where as Scientist label has least count.

Count plot of " Education " Categorical Variable shown below:

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d587032e0>
```

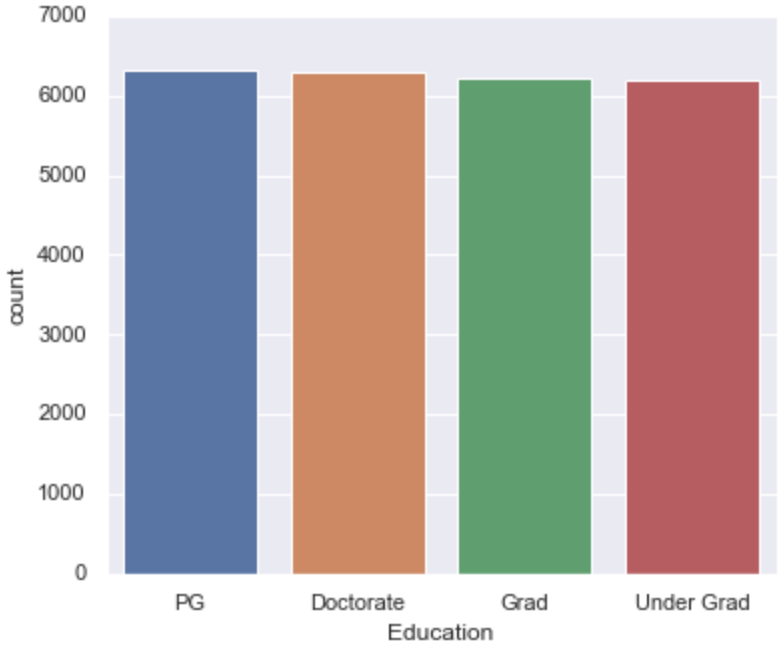


Figure 14 Countplot of Education variable

Insights

- All 4 labels have near by equivalent no of frequency count around 25 % each or we can say there are near by 6000 no of observations of each labels in the dataset.

Count plot of " Inhand_Offer " Categorical Variable shown below:

```
: <matplotlib.axes._subplots.AxesSubplot at 0x21e22048790>
```

Figure 15 Countplot of Inhand_Offer variable

Insights

- N label has highest frequency count greater than 17000 around 70% where as Y label has more than 7000 frequency count around 30%.

Count plot of " Last_Appraisal_Rating " Categorical Variable shown below:

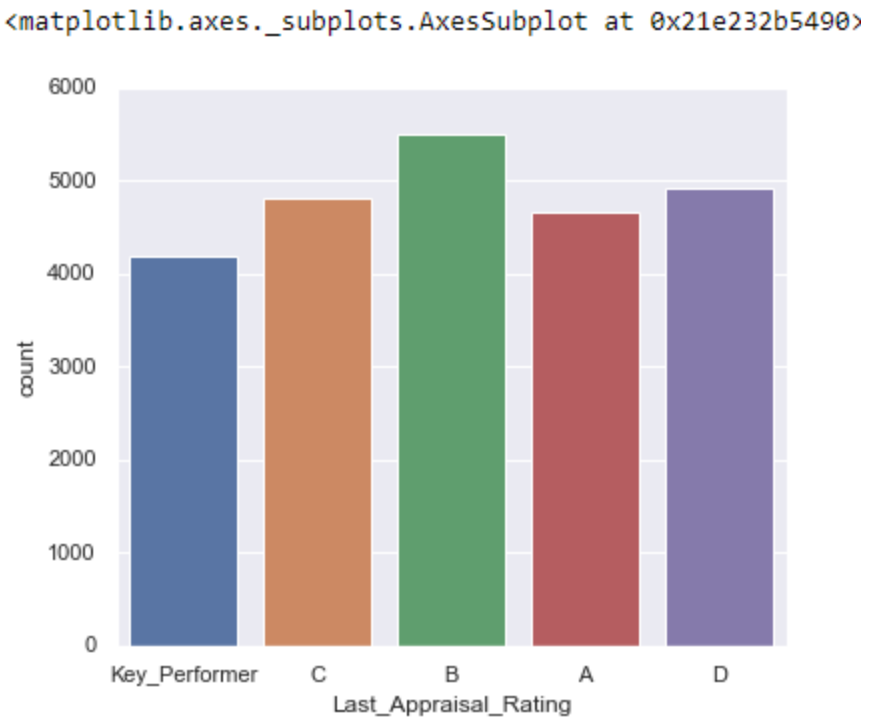


Figure 16 Countplot of Last_Appraisal_Rating variable

Insights

- B labels has maximum frequency count where as Key_Performer has least number of count.
- C and D labels have quite similar kind of frequency count.

b) Bivariate analysis (relationship between different variables , correlations)

Scatter Plot-For two Continuous Variable

Scatter (XY) Plot has points that show the relationship between two sets of data. In this example, each dot shows one person's weight versus their height. (The data is plotted on the graph as "Cartesian (x,y) Coordinates")

Scatter plot of Total_Experience vs Expected_CTC

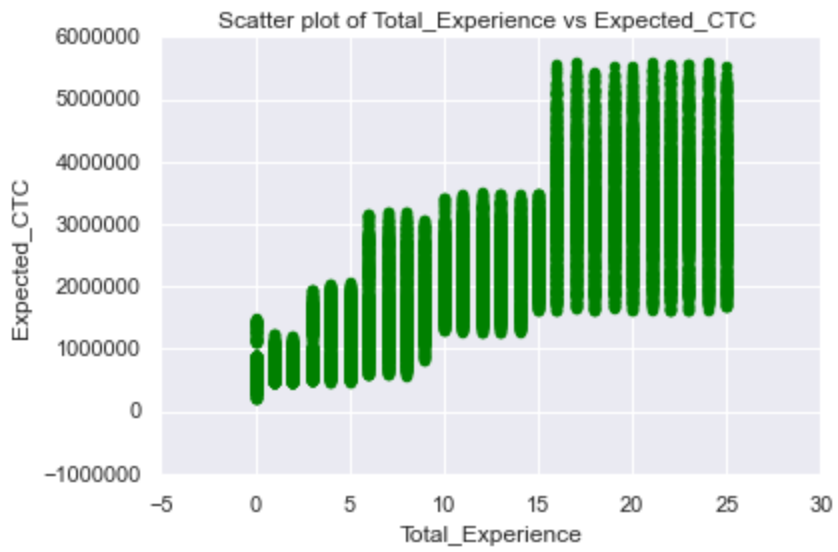


Figure 17 Scatter plot of Total_Experience vs Expected_CTC

Insights:

- Total_Experience having strong positive relationship with respect to Expected_CTC as the Total_Experience increases the Expected_CTC will also increases.

Scatter plot of Total_Experience_in_field_applied vs Expected_CTC

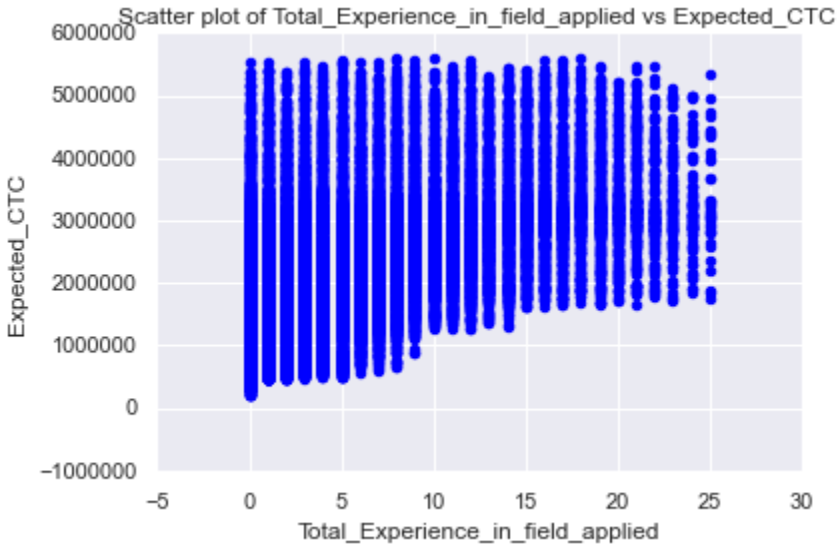


Figure 18 Scatter plot of Total_Experience_in_field_applied vs Expected_CTC

Insights:

- Total_Experience_in_field_applied having quite cloudy relationship with respect to Expected_CTC.
- we can infer only that as the Total_Experience_in_field_applied is increases the Expected_CTC will also get slightly increases.

Scatter plot of Passing_Year_Of_Graduation vs Expected_CTC

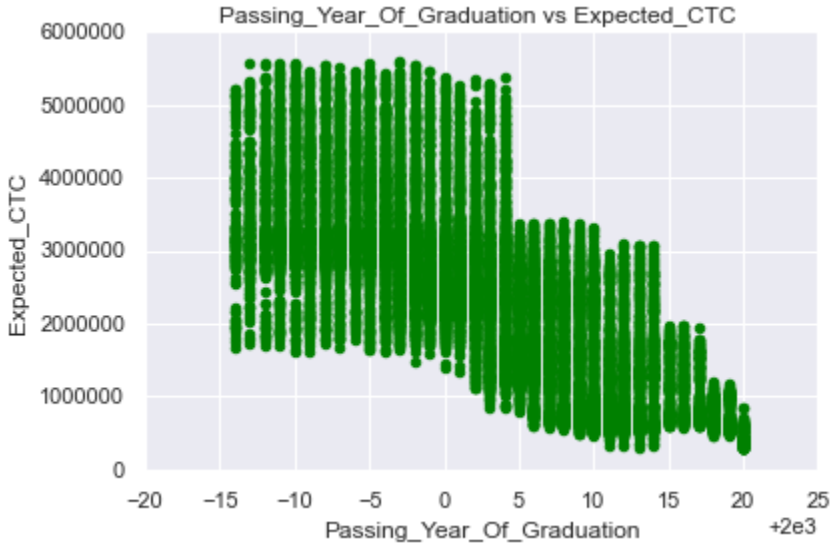


Figure 19 Scatter plot of Passing_Year_Of_Graduation vs Expected_CTC

Insights

- Passing_Year_Of_Graduation have negative relation with Expected_CTC as the oldest year having higher Expected_CTC where as the latest year has lowest Expected_CTC.

Scatter plot of Passing_Year_Of_PG vs Expected_CTC

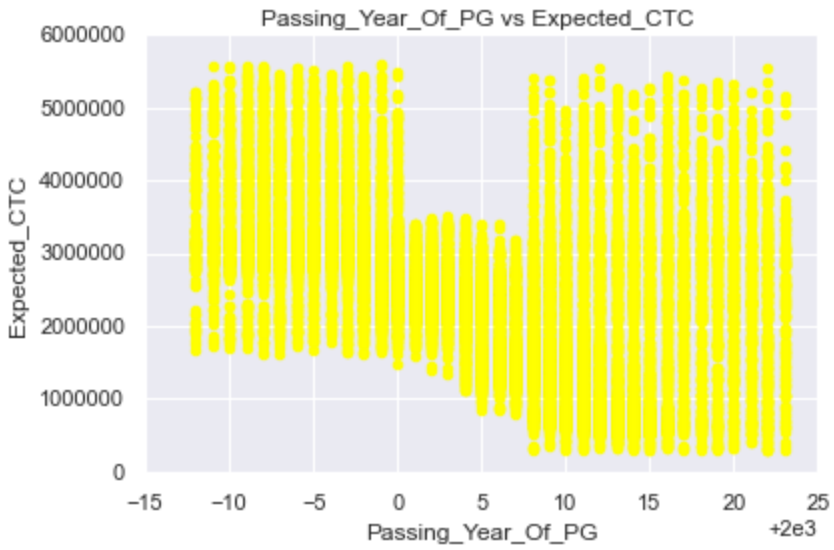


Figure 20 Scatter plot of Passing_Year_Of_PG vs Expected_CTC

Insights

- Passing_Year_Of_PG has no clearly no such clear relationship with respect to Expected_CTC.

Scatter plot of Passing_Year_Of_PHD vs Expected_CTC

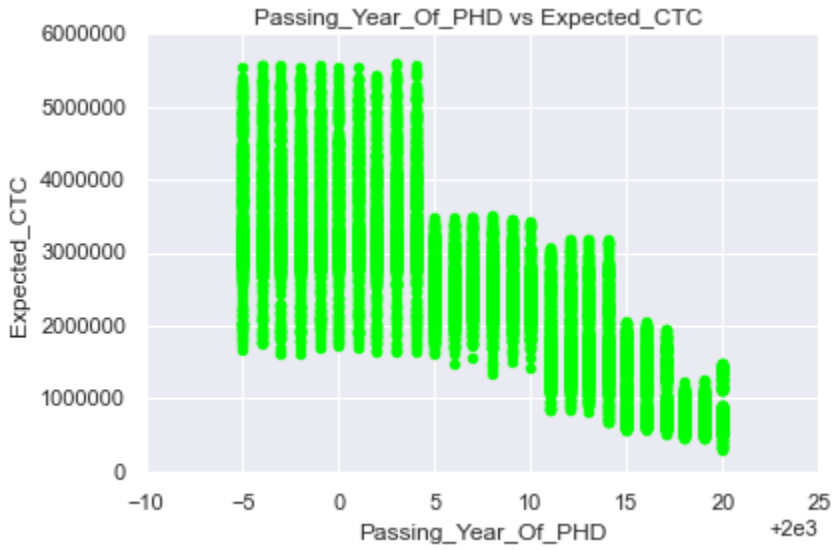


Figure 21 Scatter plot of Passing_Year_Of_PHD vs Expected_CTC

Insights

- Passing_Year_Of_PHD having negative corelation with the respect of Expected_CTC as the latest years having more Expected_CTC compared to past no of years.

Scatter plot of Current_CTC vs Expected_CTC

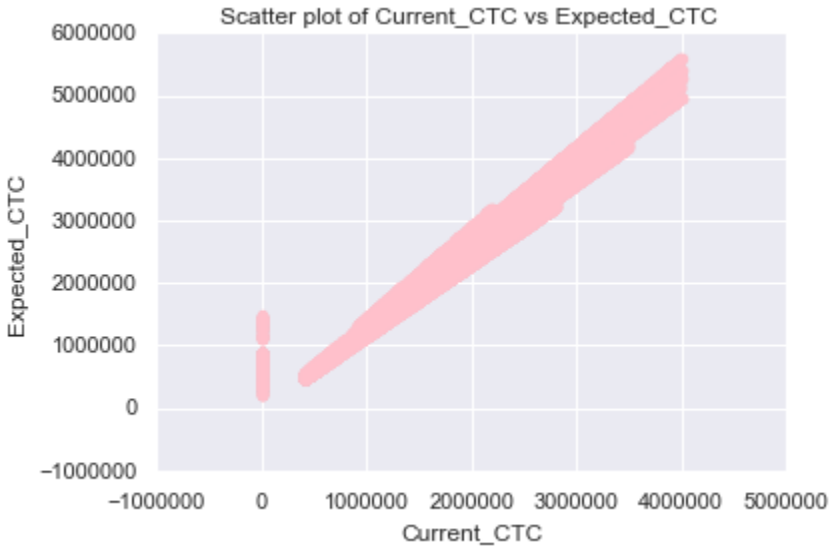


Figure 22 Scatter plot of Current_CTC vs Expected_CTC

Insights

- Current_CTC having positive corelation with the respect of Expected_CTC as the Current_CTC is increasing the Expected_CTC is also increasing.

Now we have to show the relationship of one categorical feature to the dependent variable and fetch insights from the graph using boxplot.

Department wise Expected CTC

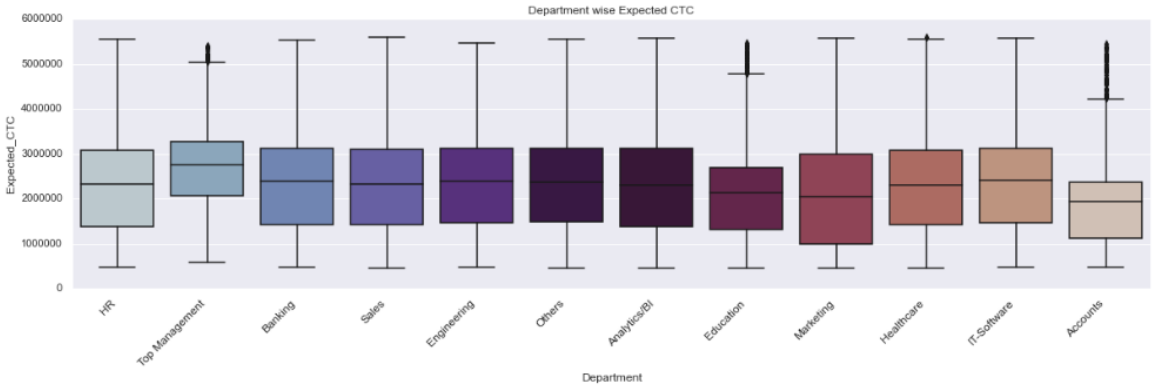


Figure 23 Boxplot of Department wise Expected CTC

Insights

- Top Management has the highest median value more than 2500000 rupees it means the higher number of Expected_CTC is associated with Top Management in terms of money.
- Accounts has least median value in terms of Expected_CTC
- Top Management , Education , Healthcare and Accounts have outliers.

Role wise Expected CTC

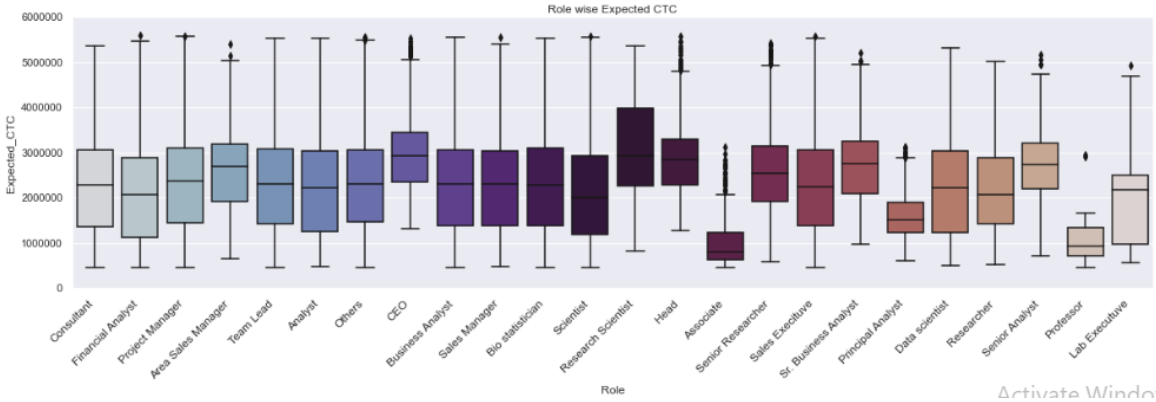


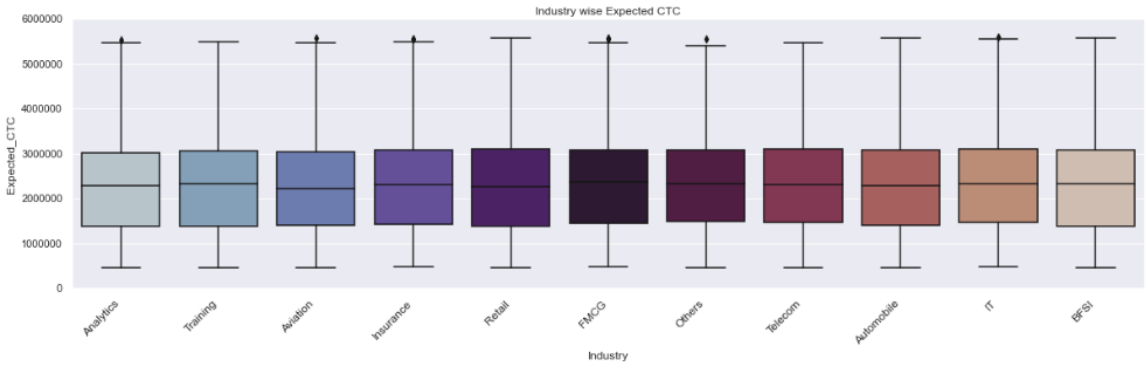
Figure 24 Boxplot of Role wise Expected CTC

Insights

- CEO and Head have highest median value that means salary range are high with respect to Expected_CTC.

- Associate has least median value means Expected_CTC has low salary with respect to Associate.
- Most of the label type of Role Feature has Outliers in terms of Expected_CTC.

Industry wise Expected CTC



Insights

- FMCG has highest median value that means salary range median value are high with respect to FMCG label.
- Training, Retail, Telecom, AutoMobile, BFSI have no outliers expect other labels have outliers with respect to Expected_CTC.
- The range of all the Industry Label have same kind of inter-quartile range.

Figure 25 Boxplot of Industry wise Expected CTC

Designation wise Expected CTC

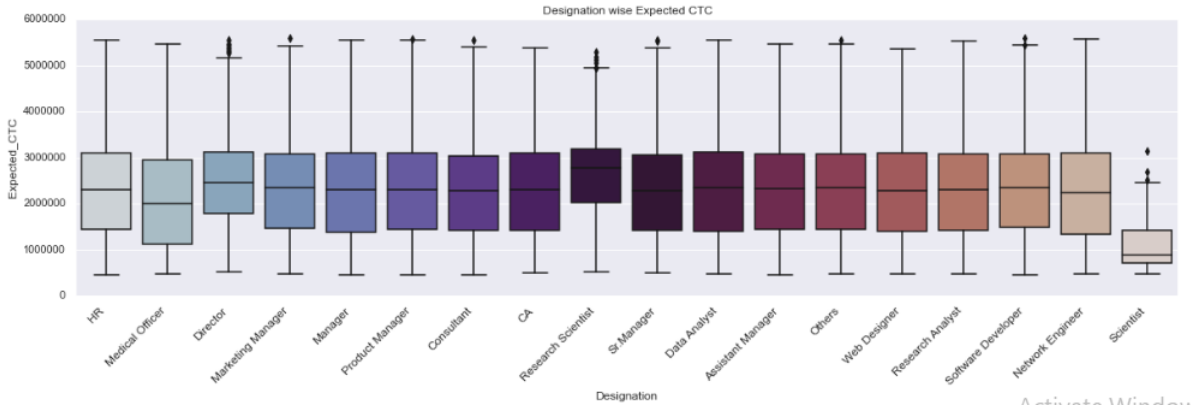


Figure 26 Boxplot of Designation wise Expected CTC

Insights

- Research Scientist has highest median value that means Expected_CTC range median value are high with respect to Research Scientist label.
- HR ,Medical Officer , Manager , CA , Data Analyst , Assistant Manager ,Web Designer, Research Analyst ,Network Engineer have no outliers expect other labels have outliers with respect to Expected_CTC.
- The Scientist has lowest median value across all labels.

Education wise Expected CTC

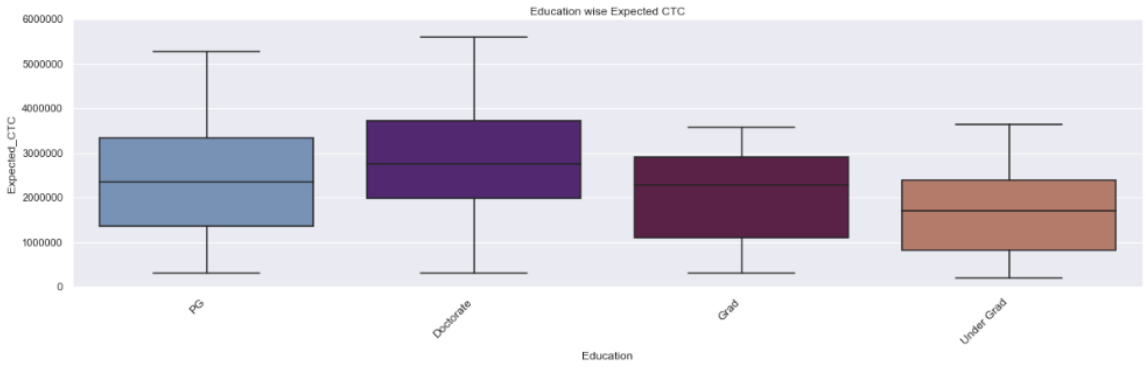


Figure 27 Boxplot of Education wise Expected CTC

Insights

- Doctorate has highest median value that means Expected_CTC range median value are high with respect to Doctorate label.
- PG, Doctorate , Grad , Under Grad have no outliers expect other labels have outliers with respect to Expected_CTC.
- The Under Grad has lowest median value across all labels.

Graduation_Specialization wise Expected CTC

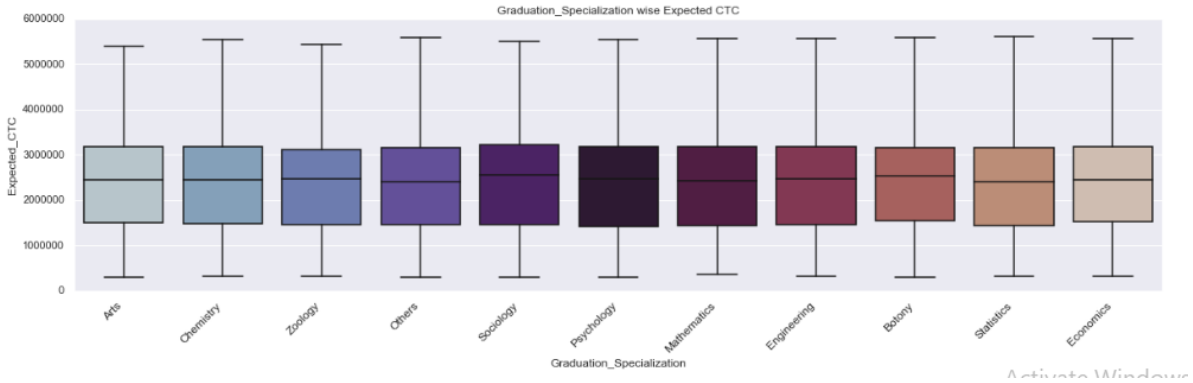


Figure 28 Boxplot of Graduation_Specialization wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of Graduation_Specialization with respect to Expected_CTC.

University_Grad wise Expected CTC

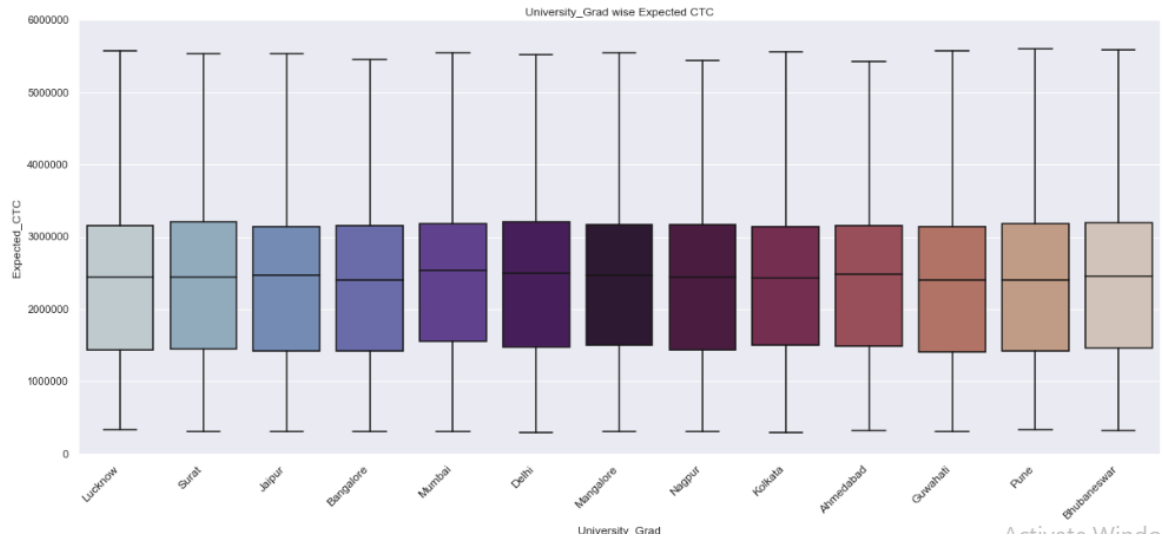


Figure 29 Boxplot of University_Grad wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of University_Grad with respect to Expected_CTC.

PG_Specialization wise Expected CTC

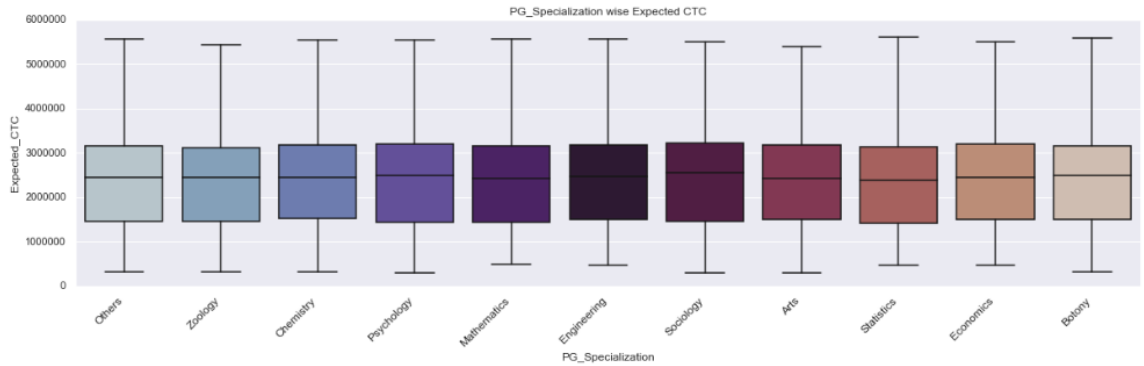


Figure 30 Boxplot of PG_Specialization wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of PG_Specialization with respect to Expected_CTC.

University_PG wise Expected CTC

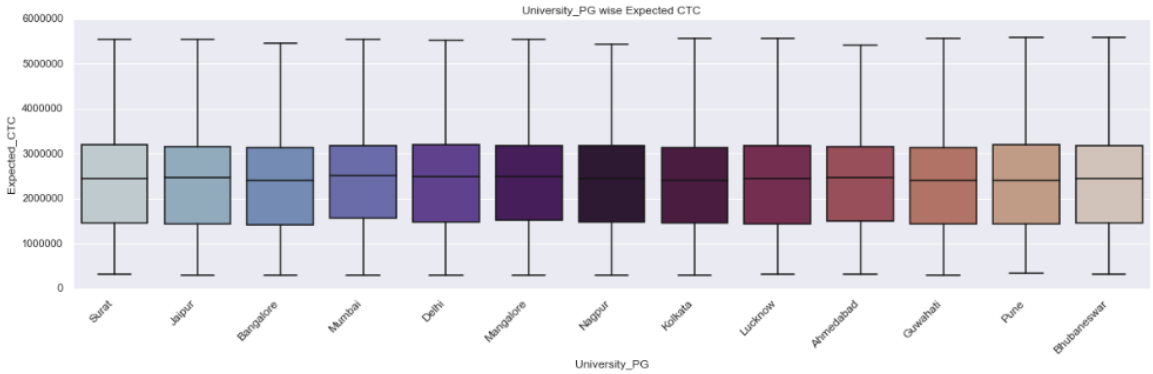


Figure 31 Boxplot of University_PG wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of University_PG with respect to Expected_CTC.

- There is no variations among all the labels i.e similar kind of distribution across Expected_CTC.

PHD_Specialization wise Expected CTC

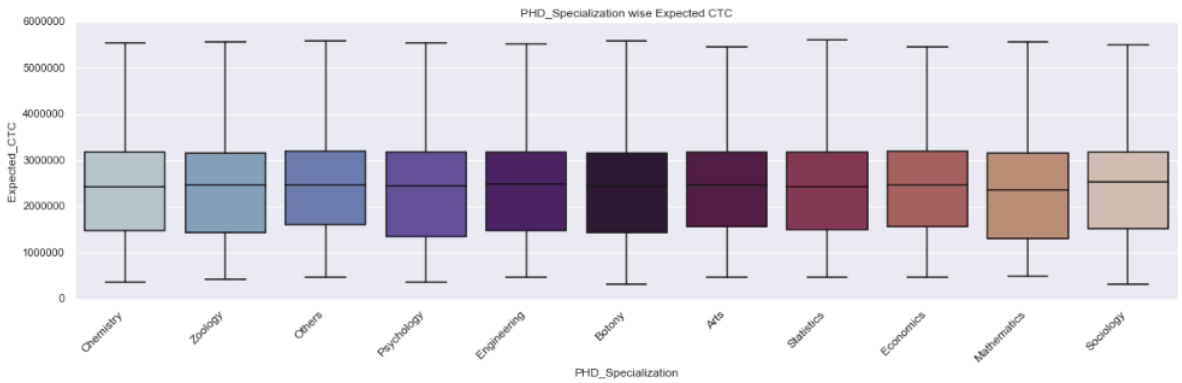


Figure 32 Boxplot of PHD_Specialization wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of PHD_Specialization with respect to Expected_CTC.
- There is no variations among all the labels i.e similar kind of distribution across Expected_CTC.

University_PHD wise Expected CTC

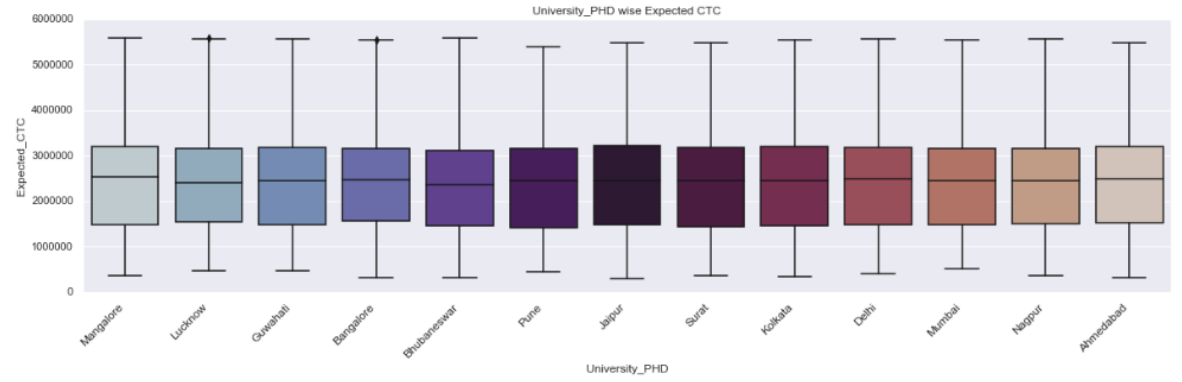


Figure 33 Boxplot of University_PHD wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of University_PHD with respect to Expected_CTC.
- There is no variations among all the labels i.e similar kind of distribution across Expected_CTC.

Current_Location wise Expected CTC

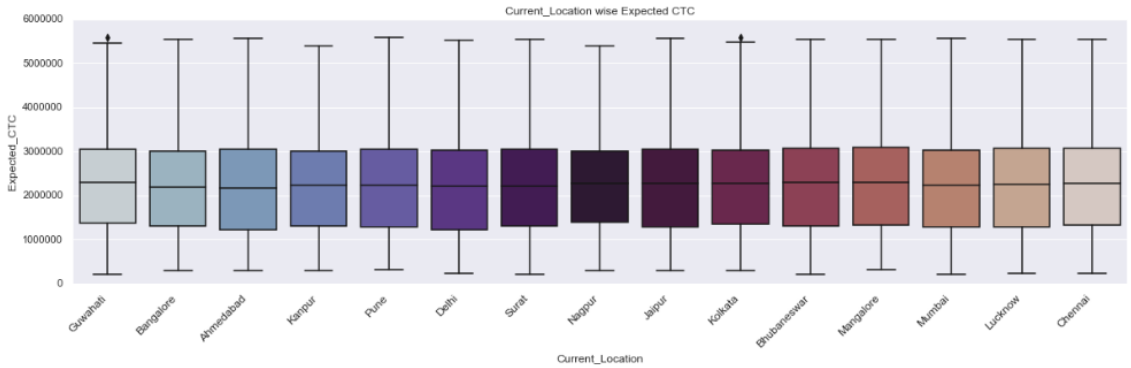


Figure 34 Boxplot of Current_Location wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of Curent_Location with respect to Expected_CTC.
- There is no variations among all the labels i.e similar kind of distribution across Expected_CTC.

Preferred_location wise Expected CTC

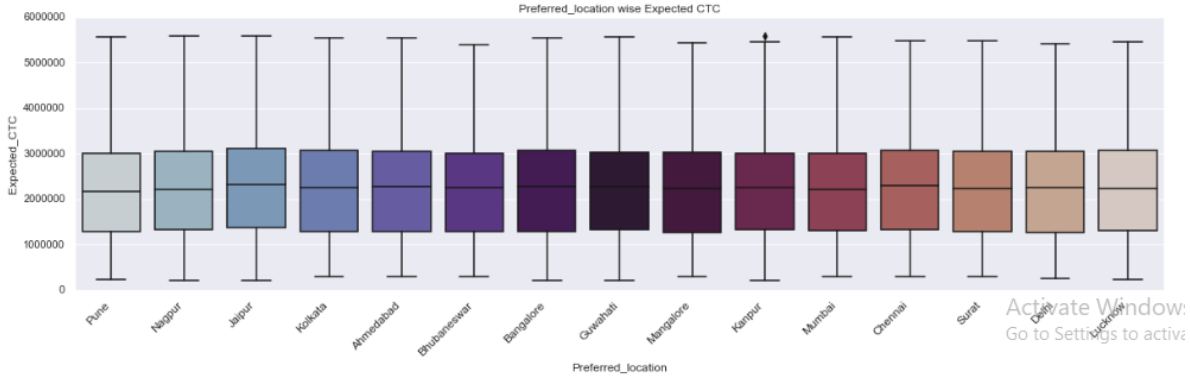


Figure 35 Boxplot of Preferred_location wise Expected CTC

Insights

- Median value and IQR range of all the labels are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of Preferred_location with respect to Expected_CTC.
- There is no variations among all the labels i.e similar kind of distribution across Expected_CTC.

Inhand_Offer wise Expected CTC

Figure 36 Boxplot of Inhand_Offer wise Expected CTC

Insights

- IQR range of Inhand_Offer are quite similar with respect to Expected_CTC.
- There is no outlier present in the labels of Inhand_Offer with respect to Expected_CTC.
- The median value of Y label of Inhand_Offer is greater than N label.

Last_Appraisal_Rating wise Expected CTC

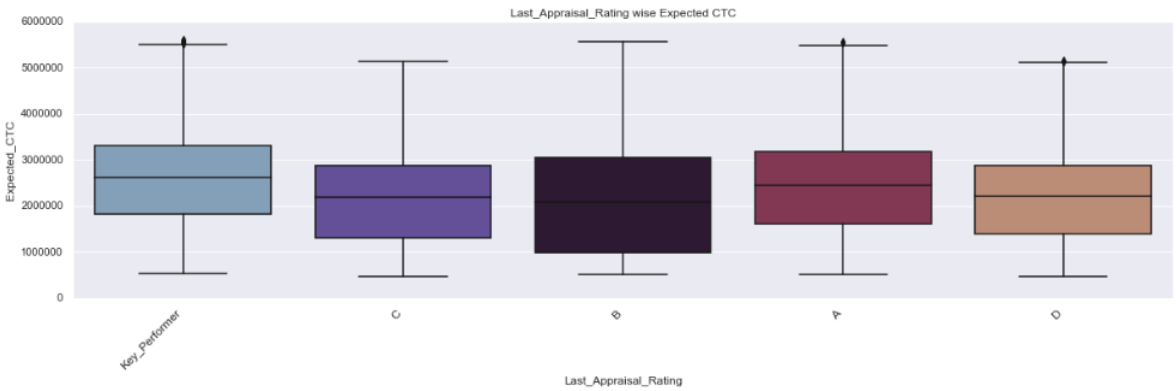


Figure 37 Boxplot of Last_Appraisal_Rating wise Expected CTC

Insights

- There is no outlier present C and B labels with respect to Expected_CTC ,rest all labels have outliers.
- The median value of Key_Performer label is highest among all labels.

Multivariate Analysis

Correlation Matrix:

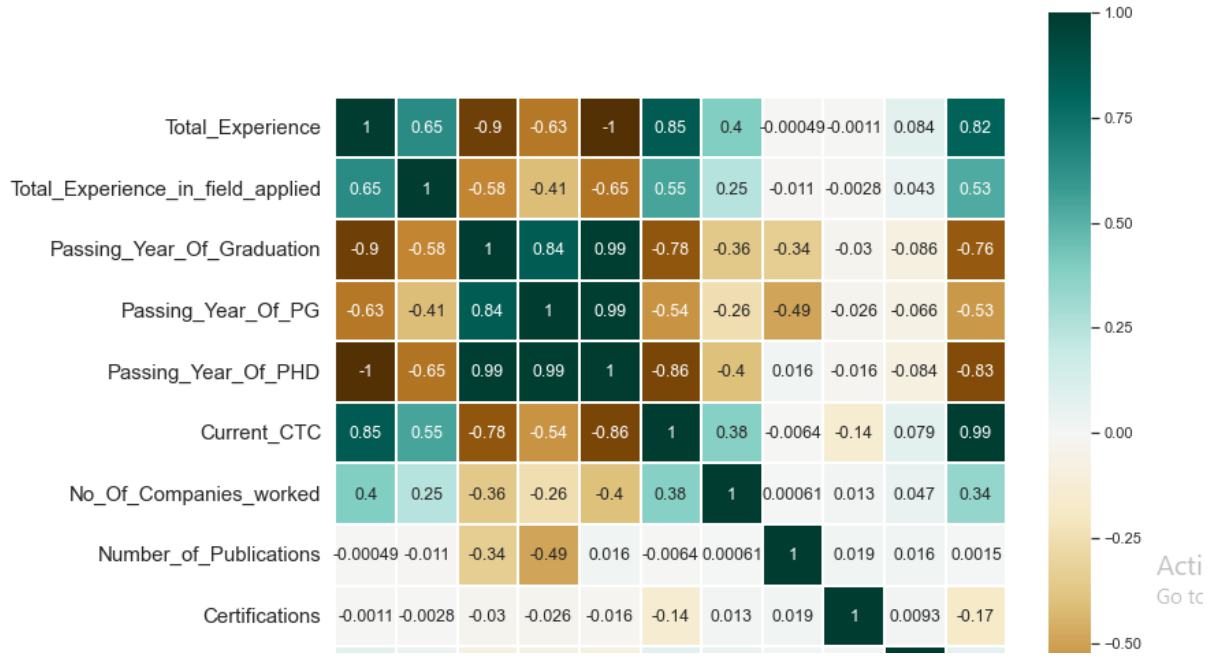
A correlation matrix is simply a table which displays the correlation. It is best used in variables that demonstrate a linear relationship between each other. coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table.

	Total_Experience	Total_Experience_in_field_applied	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Passing_Year_Of_PHD
Total_Experience	1.000000	0.645135	-0.902931	-0.634718	-1.000000
Total_Experience_in_field_applied	0.645135	1.000000	-0.581495	-0.410642	-0.648457
Passing_Year_Of_Graduation	-0.902931	-0.581495	1.000000	0.841074	0.989101
Passing_Year_Of_PG	-0.634718	-0.410642	0.841074	1.000000	0.989101
Passing_Year_Of_PHD	-1.000000	-0.648457	0.989101	0.989101	1.000000
Current_CTC	0.846476	0.548017	-0.778366	-0.544691	-0.863459
No_Of_Companies_worked	0.398135	0.249045	-0.362545	-0.255205	-0.402878
Number_of_Publications	-0.000494	-0.010663	-0.336380	-0.491231	0.015752
Certifications	-0.001130	-0.002814	-0.030236	-0.026095	-0.015784
International_degree_any	0.084072	0.043070	-0.085648	-0.066140	-0.083883
Expected_CTC	0.816593	0.529115	-0.758694	-0.530964	-0.834222
Percentage_in_Relevant_Field	0.006853	0.656567	-0.003310	-0.000663	-0.014550

Table 6 Coorelation Matrix Table of the dataset

Heatmap

- A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors.
- It represents the collinearity of the multiple variables in the dataset. data.corr() was used in the code to show the correlation between the values. This is where we want to set our independent or target variable.



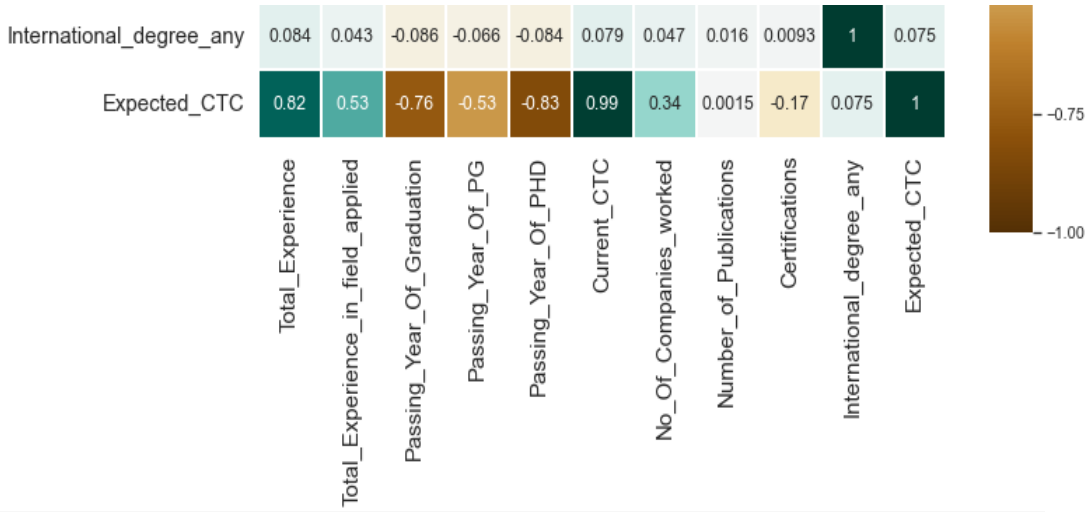


Figure 38 Heatmap of Given DataSet

Insights

- Total_Experience with Total_Experience_in_field_applied having quite good correlation (0.65).
- Total_Experience with Passing_Year_Of_Graduation having strong negative correlation (-0.90).
- Total_Experience with Passing_Year_Of_PG having negative correlation (-0.63).
- Total_Experience with Current_CTC , Expected_CTC having strong correlation (0.85 and 0.82) respectively.
- Total_Experience with No_Of_Companies_worked having weak correlation i.e.(0.40).
- Total_Experience_in_field_applied with Expected_CTC having moderate type of correlation (0.53).
- Passing_Year_Of_Graduation with Expected_CTC having negative correlation .(-0.76).
- Passing_Year_Of_PG with Expected_CTC having negative correlation .(-0.53).
- Passing_Year_Of_PHD with Expected_CTC having negative correlation .(-0.83).
- Current_CTC with Expected_CTC having very strong correlation (0.99).
- International_degree_any with Expected_CTC having very weak correlation. (0.07).
- Percentage_Relevant_Exp_in_Field with Expected_CTC having very weak correlation .(0.01).
- Passing_Year_Of_Graduation with Current_CTC and Expected_CTC having negative correlation . (-0.78 and -0.76) respectively.

Paiplot:

Pairplot visualizes given data to find the relationship between them where the variables can be continuous or categorical. Plot pairwise relationships in a data-set. Pairplot is a module of seaborn library which provides a high-level interface for drawing attractive and informative statistical graphics.

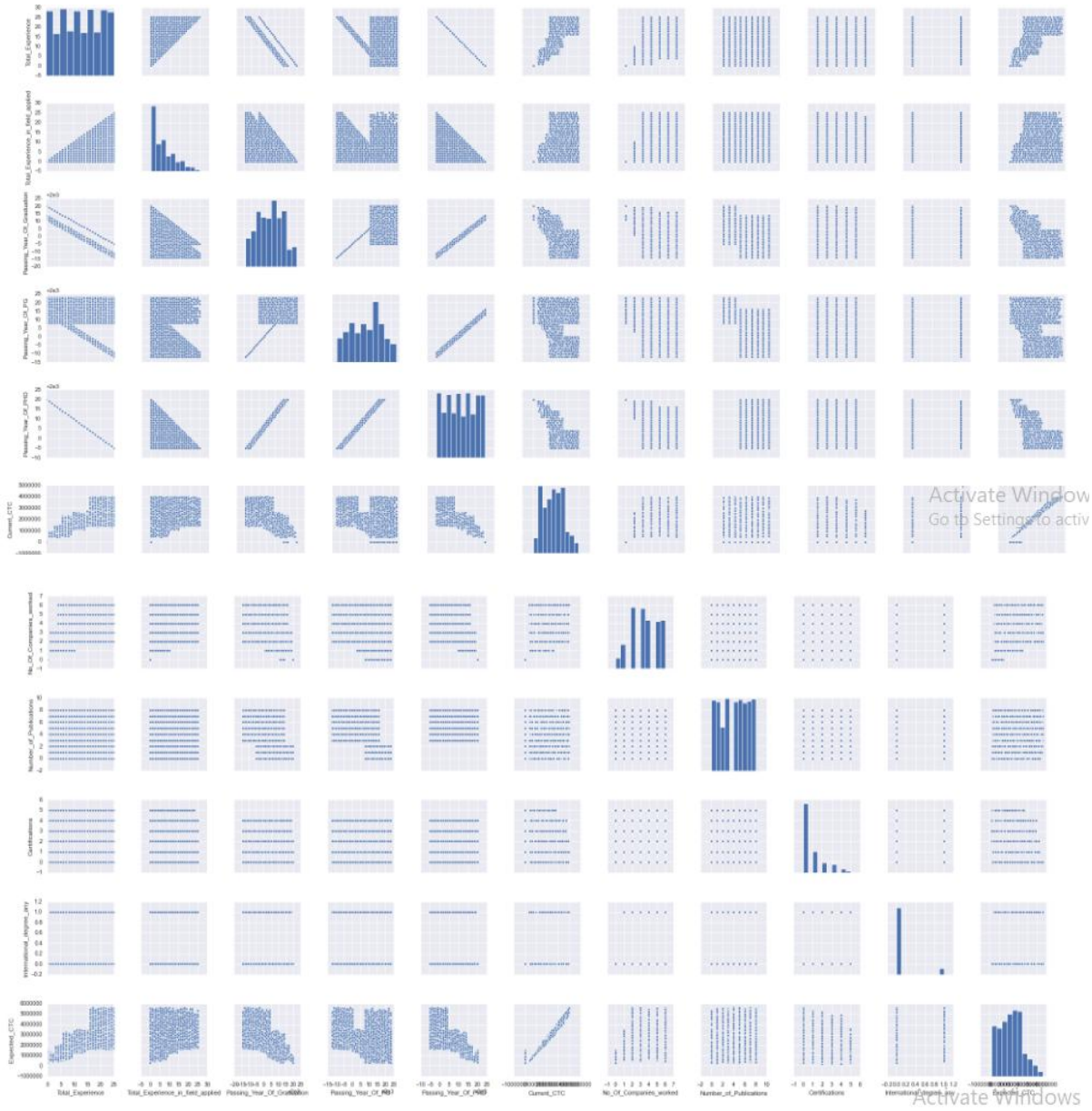


Figure 39 Pairplot of Given DataSet

Insights

- Total_Experience having strong positive relationship with respect to Expected_CTC as the Total_Experience increases the Expected_CTC will also increases.
- Total_Experience_in_field_applied having quite cloudy relationship with respect to Expected_CTC.

- we can infer only that as the Total_Experience_in_field_applied is increases the Expected_CTC will also get slightly increases.
- Passing_Year_Of_Graduation have negative relation with Expected_CTC as the oldest year having higher Expected_CTC where as the latest year has lowest Expected_CTC.
- Passing_Year_Of_PG has no clearly no such clear relationship with respect to Expected_CTC.
- Passing_Year_Of_PHD having negative corelation with the respect of Expected_CTC as the latest years having more Expected_CTC compared to past no of years.
- Current_CTC having positive corelation with the respect of Expected_CTC as the Current_CTC is increasing the Expected_CTC is also increasing.

3.	Data Cleaning and Pre-processing
	Approach used for identifying and treating missing values and outlier treatment (and why)
	Need for variable transformation (if any)
	Variables removed or added and why (if any)

	Approach used for identifying and treating missing values and outlier treatment (and why)

Checking Null value/missing values:

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
Current_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
Percentage_in_Relevant_Field	908
dtype: int64	

Table 7 Checking Null Values/Missing Values of the dataset

Insights

- we observe that PG_Specialization , University_PG and Passing_Year_Of_PG have same number of missing values (7692) which indicates that data is not available or these applicants have not PG education.

- From above result we observe that University_Grad , Passing_Year_Of_Graduation and Graduation_Specialization have same number of missing values (6180) which indicates that that data is not available.

- we observe that Passing_Year_Of_PHD, PHD_Specialization and University_PHD have same number of missing values (11881) which indicates that data is not available or these applicants have not PG education.

we have derived Null Values in form of % in ascending order -

Total_Experience	0.00000
Total_Experience_in_field_applied	0.00000
International_degree_any	0.00000
Certifications	0.00000
Number_of_Publications	0.00000
No_Of_Companies_worked	0.00000
Education	0.00000
Inhand_Offer	0.00000
Current_CTC	0.00000
Preferred_location	0.00000
Current_Location	0.00000
Expected_CTC	0.00000
Last_Appraisal_Rating	0.03632
Percentage_in_Relevant_Field	0.03632
Organization	0.03632
Industry	0.03632
Role	0.03852
Department	0.11112
Designation	0.12516
Passing_Year_Of_Graduation	0.24720
University_Grad	0.24720
Graduation_Specialization	0.24720
PG_Specialization	0.30768
University_PG	0.30768
Passing_Year_Of_PG	0.30768
PHD_Specialization	0.47524
University_PHD	0.47524
Passing_Year_Of_PHD	0.47524
dtype: float64	

Table 8 Null Values in form of % in ascending order of the dataset

Insights

- Total Experience, Total Experience in field applied, Education, Current_Location, Preferred_location, Current_CTC, Inhand_Offer, No_of_Companies_Worked, Certifications, International_degree_any, Expected_CTC has no null values apart from that all other feaures have null values.

- Maximum no of Null Values are present in the PHD_Specialization, University_PHD , Passing_Year_Of_PHD.

Note:-We have to impute the missing values with the help of fillna function depending on the distribution of the feature/column and the observations that are distributed in that particular feature/column.

There is nan value present in department column we have to change it with the "unidentified"
We name the Categorical missing values with the name unidentified we can not clearly fill the value with the some KNN Imputation Method approach or any other method like mode is generally used for categorical feature if many data points/observations are missing because it is not a right approach according to the context of the business problem or with ot knowing the complete domain knowledge so it is better to mark those missing observation which are blank or missing with indentifed label apart from that there is also a labels name others so we can further grouping them to a similar cluster or group so that will be really helpful for the model accuracy.

For eg: In the given dataset problem if a applicant is not done graduation so it not a right approach to fill the particular field or any other feature like department column have many observations is missing so it not a good approach to fill with the mode or other method according to the context of the business problem, if it is according to the domain and matches the context then we can impute with the help of mode or KNN method.

For eg:In "Department "column 2778 observations , In "Industry" column 908 observations , In "Designation" column 3129 observations are missing, In "Graduation Specialization" column 6180 observations are missing and In "Role" 963 observations are missing so we can label as unidentified because it is not a right approach to change the observations with the mode labels or with KNN imputation method or any other method. There is others label also present in these three categorical columns so we can group the others label with unidentified label.

Missing Value Treatment for Numerical Features -we can impute the missing value according to the distribution of the particular feature with the help of mean or median or depending upon the context of the business problem or may with zero.

we have successfully imputed the blank values that is not given in the Passing_Year_Of_PG and Passing_Year_Of_PHD Column with 0 because in this approach if the data point is missing in terms of education qualification we can not impute it with the diatance based techinque like Decision Tree or KNN Imputation Method or median or mean value that will led to against the business prospective and domain of the business problem because it does not mean that if someone is not completed his/her education in terms of PHD or PG we can not fill it.so we imputed with 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
Current_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
Percentage_in_Relevant_Field	0
dtype: int64	

Table 9 Null Values imputed after imputation

Note:After that we can club or group the similar labels with the help of encoding techniques.

Outlier treatment (if required)

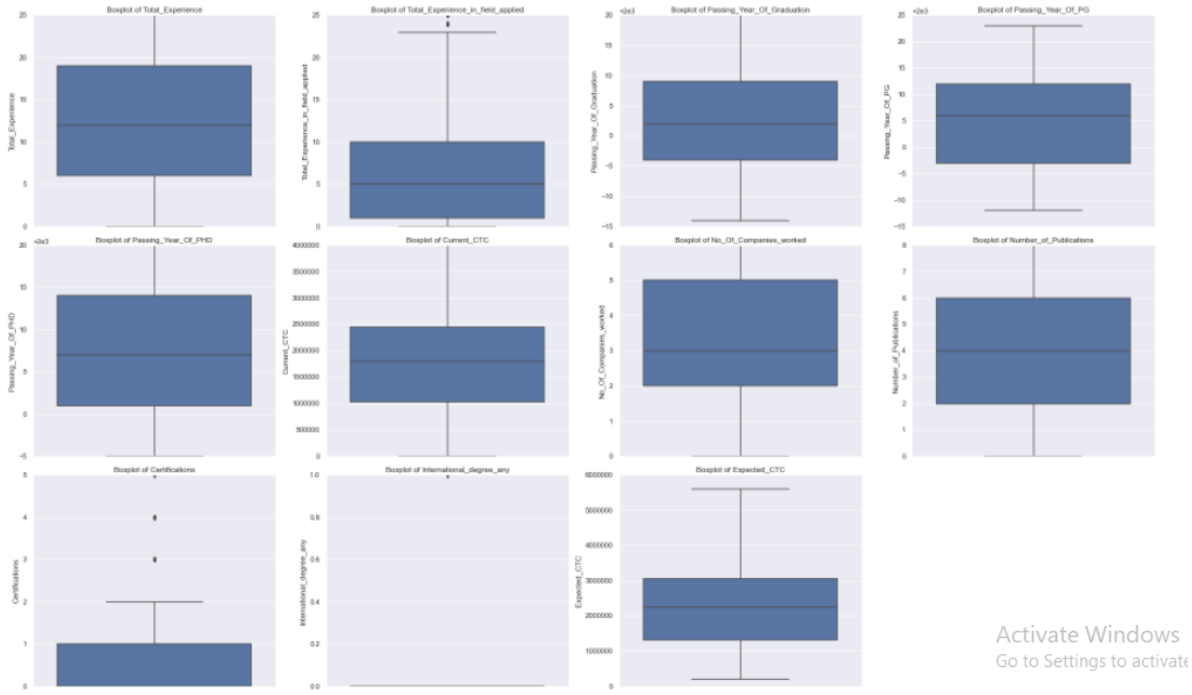


Figure 40 Show outliers using Boxplot of Given DataSet

Insights and Recommendation:

- From the above box plots we can infer that there is only few columns in which outliers is present such as Total Experience in Field Applied , Certifications contains outliers so as per the context of the business problem we do not need the requirement to treat them.

	Need for variable transformation (if any)

Variable transformation (if applicable)

- Transformation is a mathematical operation that changes the measurement scale of a variable. This is usually done to make a set of useable with a particular statistical test or method.

- Many statistical methods require data that follow a particular kind of distribution, usually a normal distribution. All of the observations must come from a population that follows a normal distribution. Groups of observations must come from populations that have the same variance or standard deviation. Transformations that normalize a distribution commonly make the variance more uniform and vice versa.

Transformation is a mathematical operation that changes the measurement scale of a variable. This is usually done to make a set of useable with a particular statistical test or method. Many statistical methods require data that follow a particular kind of distribution, usually a normal distribution.

For Independent Variable:-Depending upon the skewness Some Independent variable like we have current_CTC in high magnitude so we have to scale it with repect to other variable,so scaling is required for this and some other variable have skewness greater than 0.5.

Total_Experience_in_field_applied	0.961951
Certifications	1.610907
International_degree_any	3.054017

Anderson Darling Test for Normality - Here we apply this test on the target variable Expected_CTC to check weather the distribution is normal or not?

Hypothesis For Anderson Darling Test for Normality

H0 : Expected_CTC is normally Distributed.(Null Hypothesis)

H1 : Expected_CTC is not normally distributed.(Alternate Hypothesis)

AndersonResult(statistic=145.17631491729117, critical_values=array([0.576, 0.656, 0.787, 0.918, 1.092]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))

Conclusion : As per the test hypothesis if p_value < 0.05 ,we have to reject the null hypothesis ,or if the p_value > 0.05 , then fail to reject null hyphothesis or we have to accept null hypothesis.

As we found that the p_value > 0.05. We fail to reject the null hypothesis and conclude that Expected_CTC distribution is normal, so we do not need right now any kind of transformation if in further process in model building process if required we can go for this.

	Variables removed or added and why (if any)

- Now we drop unwanted column/Variable from the dataframe named IDX and Applicant_ID which are not useful for further model building process because thsee variable are in form of uniqueness and there is high cardinality in these two variable which is not good for model.

After removing the unwanted column the head (top 5 rows)of the dataset shown below:-

:

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Graduation_Specialization	Unive
0	0	0	NaN	NaN	NaN	NaN	NaN	PG	Arts	
1	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Chemistry	
2	21	12	Top Management	Consultant	Training	J	NaN	Doctorate	Zoology	
3	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Others	
4	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Zoology	

5 rows × 27 columns

Table 10 Top five Rows After removing the unwanted column of the dataset

Addition of new variables (if required)

Featuring Engineering -We have add a new column on the basis of Total_Experience and Total_Experience_in_Field_applied named "Percentage_in_Relevant_Field".

Note: This new column will tells the information about the percentage of an applicant in his/her relevant field.

4.	Model building
	Clear on why was a particular model(s) chosen.
	Effort to improve model performance.

Now lets Checking the Dataset after Encoding- For more details in Encoding part kindly refer to the code file that is shared.

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Designation	Education	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Passing_Ye
0	0	0	0	0	0	2	2020.0	0.0	
1	23	14	2	1	1	3	1988.0	1990.0	
2	21	12	3	1	0	3	1990.0	1992.0	
3	15	8	2	1	1	3	1997.0	1999.0	
4	10	5	2	1	1	1	2004.0	2006.0	

Table 11 Top five Rows After Encoding of the dataset

Converting the Datatypes of encoded variables(object) into int64

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 18 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Total_Experience                    25000 non-null  int64
 1   Total_Experience_in_field_applied   25000 non-null  int64
 2   Department                          25000 non-null  int64
 3   Role                                25000 non-null  int64
 4   Designation                         25000 non-null  int64
 5   Education                           25000 non-null  int64
 6   Passing_Year_Of_Graduation          25000 non-null  float64
 7   Passing_Year_Of_PG                  25000 non-null  float64
 8   Passing_Year_Of_PHD                 25000 non-null  float64
 9   Current_CTC                         25000 non-null  int64
```

```
 10   Inhand_Offer                      25000 non-null   int64
 11   Last_Appraisal_Rating             25000 non-null   int64
 12   No_Of_Companies_worked            25000 non-null   int64
 13   Number_of_Publications            25000 non-null   int64
 14   Certifications                    25000 non-null   int64
 15   International_degree_any          25000 non-null   int64
 16   Expected_CTC                      25000 non-null   int64
 17   Percentage_in_Relevant_Field      25000 non-null   float64
dtypes: float64(4), int64(14)
memory usage: 3.4 MB
```

Table 12 Info After Encoding of the dataset

Model building and interpretation.

- a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes).
- b. Test your predictive model against the test set using various appropriate performance metrics.
- c.Interpretation of the model(s).

Now we have to perform Linear Regression Model-1 using Sklearn.

LinearRegression()

Let us explore the coefficients for each of the independent attributes

In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one.

```
The coefficient for Total_Experience is -7875.7006988648345
The coefficient for Total_Experience_in_field_applied is 10331.296058151447
The coefficient for Department is -38555.332840246694
The coefficient for Role is -76588.19499331282
The coefficient for Designation is -97100.79148970016
The coefficient for Education is 92821.4193490013
The coefficient for Passing_Year_Of_Graduation is -3782.6905138905327
The coefficient for Passing_Year_Of_PG is -26.514071003045565
```

The coefficient for Passing_Year_Of_PHD is -17.25954826464948
The coefficient for Current_CTC is 1.245143148699813
The coefficient for Inhand_Offer is 81329.3083211779
The coefficient for Last_Appraisal_Rating is 83404.77691671914
The coefficient for No_Of_Companies_worked is -14005.82673701161
The coefficient for Number_of_Publications is 3363.2220575701826
The coefficient for Certifications is 51.77055637779263
The coefficient for International_degree_any is 33802.22612299101
The coefficient for Percentage_in_Relevant_Field is -1704.181242307113

Table 13 Coefficient Table(using Linear Regression) of the dataset

Lets check the Intercept for the Model

The intercept for our model is [7791186.37783288]

R square on training data

0.9854357636298562

R square on testing data

0.9861489871486351

RMSE on Training data

139510.23524866343

RMSE on Testing data

137786.14281034478

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

Note-R square is quite similar and good on both training as well as testing data this may be due to the data itself that is given .

RMSE is quite similar for both on training and test data set.Least RMSE better will be the model.

Note-we can perform various type of ensemble technique to perform the more accurate results in terms of accuracy and performance metrices.

Linear Regression Model-2 Using statsmodels.

Let us explore the coefficients for each of the independent attributes using statsmodel

Intercept	**7.791186e+06**
Total_Experience	**-7.875701e+03**
Total_Experience_in_field_applied	**1.033130e+04**
Department	**-3.855533e+04**
Role	**-7.658819e+04**
Designation	**-9.710079e+04**
Education	**9.282142e+04**
Passing_Year_Of_Graduation	**-3.782691e+03**
Passing_Year_Of_PG	**-2.651407e+01**
Passing_Year_Of_PHD	**-1.725955e+01**
Current_CTC	**1.245143e+00**
Inhand_Offer	**8.132931e+04**
Last_Appraisal_Rating	**8.340478e+04**
No_Of_Companies_worked	**-1.400583e+04**
Number_of_Publications	**3.363222e+03**
Certifications	**5.177056e+01**
International_degree_any	**3.380223e+04**
Percentage_in_Relevant_Field	**-1.704181e+03**
dtype: float64	

Table 14 Coefficient Table(using Linear Regression using stats model) of the dataset

OlS summary on train dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.985
Method:	Least Squares	F-statistic:	6.958e+04
Date:	Sun, 24 Apr 2022	Prob (F-statistic):	0.00
Time:	14:16:21	Log-Likelihood:	-2.3213e+05
No. Observations:	17500	AIC:	4.643e+05
Df Residuals:	17482	BIC:	4.644e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.791e+06	6.23e+05	12.515	0.000	6.57e+06	9.01e+06
Total_Experience	-7875.7007	420.123	-18.746	0.000	-8699.184	-7052.218
Total_Experience_in_field_applied	1.033e+04	414.382	24.932	0.000	9519.066	1.11e+04
Department	-3.856e+04	1546.674	-24.928	0.000	-4.16e+04	-3.55e+04
Role	-7.659e+04	2984.888	-25.659	0.000	-8.24e+04	-7.07e+04
Designation	-9.71e+04	3051.978	-31.816	0.000	-1.03e+05	-9.11e+04
Education	9.282e+04	1504.023	61.715	0.000	8.99e+04	9.58e+04
Passing_Year_Of_Graduation	-3782.6905	309.273	-12.231	0.000	-4388.897	-3176.484
Passing_Year_Of_PG	-26.5141	2.003	-13.235	0.000	-30.441	-22.587
Passing_Year_Of_PHD	-17.2595	2.161	-7.987	0.000	-21.495	-13.024
Current_CTC	1.2451	0.003	472.187	0.000	1.240	1.250
Inhand_Offer	8.133e+04	2510.797	32.392	0.000	7.64e+04	8.63e+04
Last_Appraisal_Rating	8.34e+04	2219.057	37.586	0.000	7.91e+04	8.78e+04
No_Of_Companies_worked	-1.401e+04	702.053	-19.950	0.000	-1.54e+04	-1.26e+04
Number_of_Publications	3363.2221	546.874	6.150	0.000	2291.294	4435.150
Certifications	51.7706	1036.443	0.050	0.960	-1979.761	2083.302
International_degree_any	3.38e+04	4090.898	8.263	0.000	2.58e+04	4.18e+04
Percentage_in_Relevant_Field	-1704.1812	55.678	-30.608	0.000	-1813.316	-1595.046

Omnibus:	5186.950	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32298.537
Skew:	1.278	Prob(JB):	0.00
Kurtosis:	9.145	Cond. No.	1.17e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 15 OLS Summary of Train Data(using Linear Regression using stats model) of the dataset

Hypothesis Testing

- H0 : There is no relationship between independent and the dependent variable.
- H1 : There is a relationship between independent and the dependent variable.

We observe that the pvalue of Certifications variable is 0.960. We know that if pvalue < 0.5 reject the null hypohesis & if pvalue > 0.5 we fail to reject the null hypothesis or accept the null hypothesis..Here we found that pvalue of Certifications is 0.960 which greater than 0.05. So, we fail to reject the null hypothesis.

Note:-Certifications is not good variable to predict the expected_ctc so we can drop it and build he model again and check the statsmodel summary and rmse and accuracy of the model again.

OlS summary on test dataset.

OLS Regression Results

Dep. Variable:	Expected_CTC	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	3.146e+04
Date:	Sun, 24 Apr 2022	Prob (F-statistic):	0.00
Time:	14:16:21	Log-Likelihood:	-99378.
No. Observations:	7500	AIC:	1.988e+05
Df Residuals:	7482	BIC:	1.989e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.774e+06	9.33e+05	8.334	0.000	5.95e+06	9.6e+06
Total_Experience	-7328.9754	614.120	-11.934	0.000	-8532.822	-6125.128
Total_Experience_in_field_applied	9911.9678	627.176	15.804	0.000	8682.527	1.11e+04
Department	-3.895e+04	2321.142	-16.781	0.000	-4.35e+04	-3.44e+04
Role	-7.709e+04	4438.986	-17.367	0.000	-8.58e+04	-6.84e+04
Designation	-8.914e+04	4678.931	-19.052	0.000	-9.83e+04	-8e+04
Education	9.252e+04	2260.553	40.929	0.000	8.81e+04	9.7e+04
Passing_Year_Of_Graduation	-3784.1832	463.516	-8.164	0.000	-4692.805	-2875.561
Passing_Year_Of_PG	-22.1458	2.995	-7.393	0.000	-28.018	-16.274
Passing_Year_Of_PHD	-20.7594	3.290	-6.309	0.000	-27.209	-14.309
Current_CTC	1.2455	0.004	320.794	0.000	1.238	1.253
Inhand_Offer	9.205e+04	3775.364	24.382	0.000	8.47e+04	9.95e+04
Last_Appraisal_Rating	7.99e+04	3309.425	24.144	0.000	7.34e+04	8.64e+04
No_Of_Companies_worked	-1.212e+04	1066.702	-11.362	0.000	-1.42e+04	-1e+04
Number_of_Publications	3214.7838	831.585	3.866	0.000	1584.644	4844.924
Certifications	-753.3559	1537.814	-0.490	0.624	-3767.904	2261.192
International_degree_any	2.789e+04	6145.700	4.538	0.000	1.58e+04	3.99e+04
Percentage_in_Relevant_Field	-1629.5139	84.618	-19.257	0.000	-1795.389	-1463.639

Omnibus:	2203.600	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13323.577
Skew:	1.269	Prob(JB):	0.00
Kurtosis:	9.016	Cond. No.	1.17e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 16 OLS Summary of Test Data(using Linear Regression using stats model) of the dataset

Hypothesis Testing

H0 : There is no relationship between independent and the dependent variable.

H1 : There is a relationship between independent and the dependent variable.

we observe that the p value of Certifications variable is 0.624. We know that if pvalue < 0.5 reject the null hypohesis & if pvalue > 0.5 we fail to reject the null hypothesis or accept the null hypothesis..Here we found that pvalue of Certifications is 0.624 which greater than 0.05. So, we fail to reject the null hypothesis

Calculate MSE of Train Data

139510.23524866364

Calculate MSE of Test Data

137505.7997142062

Prediction on Train data

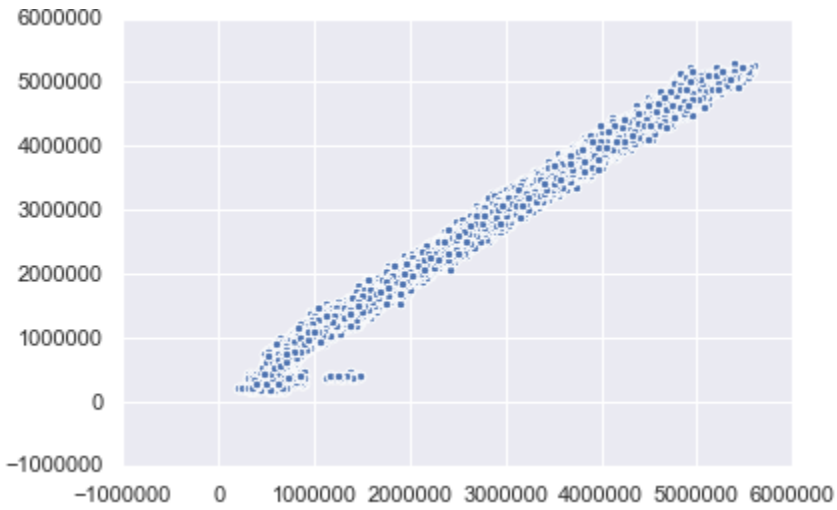


Figure 41 Prediction on Train Data using Linear Regression

Prediction on Test data

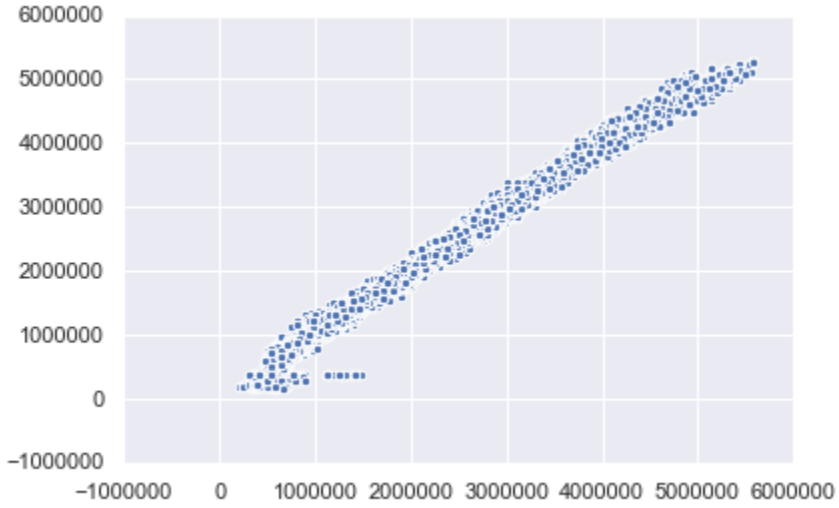


Figure 42 Prediction on Test Data using Linear Regression

Note-

From the above result Certifications variable/feature has p-value more than 0.05 which indicates us that Certifications feature is not important for predicting the expected_ctc our target variable.

Note:-If the training set's R-squared is higher and the R-squared of the validation set is much lower, it indicates overfitting. If the same high R-squared translates to the validation set as well, then we can say that the model is a good fit.

Note:This completely depends on the type of the problem being solved. In some problems which are hard to model, even an R-squared of 0.5 may be considered a good one. There is no rule of thumb to confirm the R-squared to be good or bad. However, a very low R-squared indicates underfitting and adding additional relevant features or using a complex model might help.

Note:-In the above result in both training and test result R-squared is nearly same.so it is a very good model.

Linear Rregression Model 3 using : (Z-Score)

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Designation	Education	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Pa
4289	0.471308	-0.048497	1.720631	2.250312	0.281815	0.431962	0.253142	0.681995	
19621	-0.063826	0.804631	0.377707	-0.056146	0.281815	-1.357019	-0.023395	-1.503906	
14965	1.675361	1.145882	0.377707	2.250312	0.281815	1.326453	-0.991274	0.673342	
12321	0.203741	-1.072250	0.377707	-0.056146	0.281815	-1.357019	-0.023395	-1.503906	
6269	1.006443	1.828385	0.377707	-0.056146	0.281815	1.326453	-1.544348	0.651710	

Table 17 Rows after Linear Regression using z-score of the dataset

lets perform the LinearRegression function.

```
LinearRegression()
```

In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one.

The coefficient for Total_Experience is -0.05092418163596045
The coefficient for Total_Experience_in_field_applied is 0.05237794668530744
The coefficient for Department is -0.024835381454620005
The coefficient for Role is -0.02872460870435031
The coefficient for Designation is -0.03127023856689118
The coefficient for Education is 0.089765641548738
The coefficient for Passing_Year_Of_Graduation is -0.02366549207875398
The coefficient for Passing_Year_Of_PG is -0.021205579313856648
The coefficient for Passing_Year_Of_PHD is -0.014965789712132005
The coefficient for Current_CTC is 0.9882208458359613
The coefficient for Inhand_Offer is 0.03233039380381115
The coefficient for Last_Appraisal_Rating is 0.03885095999869427
The coefficient for No_Of_Companies_worked is -0.020483440011908105
The coefficient for Number_of_Publications is 0.007606086886444415
The coefficient for Certifications is 5.3575938286870876e-05
The coefficient for International_degree_any is 0.008030161802295911
The coefficient for Percentage_in_Relevant_Field is -0.04996218250393585

Table 18 Coefficient Table (Linear Regression using z-score) of the dataset

Intercept for the model
-2.653867058572929e-16

R square on training data
0.9854357636298562

R square on test data
0.986154596005444

RMSE on Training data
139510.23524866343

RMSE on Test data
137505.79971420625

Note:-

From the above result we can observe that Linear Regression base model and Linear Regression with Zscore has similar kind of results. There is no such major changes in the RMSE and R square .

Now let us perform XG Boost Model 4:-

```
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
             early_stopping_rounds=None, enable_categorical=False,
             eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
             importance_type=None, interaction_constraints='',
             learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
             max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
             missing=nan, monotone_constraints='()', n_estimators=10, n_jobs=0,
             num_parallel_tree=1, objective='reg:linear', predictor='auto',
             random_state=123, reg_alpha=0, ...)
```

Model Name	Training RMSE	Test RMSE	R-squared(Train Data)	R-squared(Test Data)	Adj-R Square(Train Data)	Adj-R Square (Test Data)
Linear Regression(Model 1)	139510.23	137786.14	0.985	0.986	0.98	0.98
Linear Regression(Model 2-Z-Score Scaling)	137505.7997	137786.14	0.985	0.986	0.98	0.98
XG-Boost Regressor	36872.64	35874.54	0.99	0.99	0.99	0.99
Decision Tree Regressor	42336.55	39587.22	0.98	0.98		

Table 19 Comparing Model Accuracy Table of the dataset

1. Interpretation of the model(s)

From the below table that all regression models performs good and not having any issues of Overfitting and Underfitting plus model scores are also good and almost similar for all models on train and test.

On comparing the Model Scores from the Linear Regression models and other regression models XG-Boost Regressor well performed as we get lowest RMSE on XG-Boost Regressor for train and test and have very good model score of 0.99 on train and test.

We can also use Decision Tree Regressor and it also have very good model score and least differnece in between RMSE of train and test set.

XG_Boost Regressor –

Before running XGBoost, we must set three types of parameters: general parameters, booster parameters and task parameters. Learning task parameters decide on the learning scenario. For example, regression tasks may use different parameters with ranking tasks.

We can choose XG_Boost Regressor and it also have very good model score and least difference in between RMSE of train and test set.

This model gives us accuracy 99% on train and test data.There is no problem of underfit or overfit in this model. We build XG_Boost Regressor model on standard parameter like

xg(objective ='reg:linear' ,n_estimators = 10, seed = 123)

The random seed most likely effects the resulted xgboost model through the sampling of train data in the fitting process. So a random seed will determine a 'path' of trees that focus on different part (e.g. subset of columns) of training data.

The number of trees (or rounds) in an XGBoost model is specified to the XGBRegressor class in the n_estimators argument.

The most common loss functions in XGBoost for regression problems is reg:linear . Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoost is one of the ensemble learning methods.

Note:- we can see from the above result the RMSE of the Xg Boost Model is quite good as compare to the the other model.

Note:-Some of the important fetures that will helps to predict the expected_CTC(dependent variable) in a better way having good positive cofficient value.

The coefficient for Current_CTC is 0.9882208458359613
The coefficient for Inhand_Offer is 0.03233039380381115
The coefficient for Last_Appraisal_Rating is 0.03885095999869427

Table 20 Important Features of XG Boost Moel of the dataset

Tuned Decision Tree parameters

These are non-parametric decision tree learning techniques that provide regression or classification trees, relying on whether the dependent variable is categorical or numerical respectively. This algorithm deploys the method of Gini Index to originate binary splits. Both Gini Index and Gini Impurity are used interchangeably.
Decision trees have influenced regression models in machine learning. While designing the tree, developers set the nodes' features and the possible attributes of that feature with edges.

Calculation

The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and are very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.
The Gini Index varies between 0 and 1, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes. A Gini Index of 0.5 shows that there is equal distribution of elements across some classes.

Tuned Decision Tree Parameters: {'criterion': 'gini', 'max_depth': None, 'max_features': 7, 'min_samples_leaf': 3}

According to the industry standard we can set max_depth ,max_depth and max_features. parameters in tunning process.

2.Model Tuning

Ensemble means a group of elements viewed as a whole rather than individually. An Ensemble method creates multiple models and combines them to solve it. Ensemble methods help to improve the robustness/generalizability of the model.

Basic ensemble methods

1. Averaging method: It is mainly used for regression problems. The method consists of building multiple models independently and returning the average of the prediction of all the models. In general, the combined output is better than an individual output because variance is reduced.

In the below example, three regression models (linear regression, xgboost, and random forest) are trained and their predictions are averaged. The final prediction output is pred_final.

```
# predicting the output on the test dataset
```

pred_1 Result

```
LinearRegression()
```

Result

```
array([[3913010.64893546],
       [1483970.44374886],
       [ 520769.58479667],
       ...,
       [4359452.62425181],
       [1356846.38401941],
       [2588345.57702589]])
```

pred_2 Result

```
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
       colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
       early_stopping_rounds=None, enable_categorical=False,
       eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
       importance_type=None, interaction_constraints='',
       learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
       max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
       missing=nan, monotone_constraints='()', n_estimators=100, n_jobs=0,
       num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
       reg_lambda=1, ...)
```

```
array([3823191.  , 1558187.5 ,  592585.44, ..., 4384405.5 , 1443844.1 ,
```

```
       2529341.  ], dtype=float32)
```

pred_3 Result

RandomForestRegressor()

```
array([3854841.63, 1538089.12,  588275.21, ..., 4384563.21, 1444918.23,
       2482882.64])
```

```
# final prediction after averaging on the prediction of all 3 models
```

```
array([[3863681.09297849, 2336429.08964515, 1697957.09881182, ...,
        4227326.45297849, 2267257.66797849, 2975078.09631182],
       [3054001.02458295, 1526749.02124962,  888277.03041629, ...,
        3417646.38458295, 1457577.59958295, 2165398.02791629],
       [2732934.07159889, 1205682.06826556,  567210.07743222, ...,
        3096579.43159889, 1136510.64659889, 1844331.07493222],
       ...,
       [4012495.0847506 , 2485243.08141727, 1846771.09058394, ...,
        4376140.4447506 , 2416071.6597506 , 3123892.08808394],
       [3011626.33800647, 1484374.33467314,  845902.3438398 , ...,
        3375271.69800647, 1415202.91300647, 2123023.3413398 ],
       [3422126.06900863, 1894874.0656753 , 1256402.07484196, ...,
        3785771.42900863, 1825702.64400863, 2533523.07234196]])
```

Hyperparameter tuning

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters.

However, there is another kind of parameters, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

The aim of this article is to explore various strategies to tune hyperparameter for Machine learning model.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two best strategies for Hyperparameter tuning are:

- GridSearchCV
- RandomizedSearchCV

Lets perform one example of Random Search Cv

```
RandomizedSearchCV(cv=2, estimator=DecisionTreeClassifier(),
                   param_distributions={'criterion': ['gini', 'entropy'],
                                        'max_depth': [3, None],
```

5.	Model validation
	How was the model validated? Just accuracy, or anything else too?

From the below table that all regression models performs good and not having any issues of Overfitting and Underfitting plus model scores are also good and almost similar for all models on train and test.

On comparing the Model Scores from the Linear Regression models and other regression models XG-Boost Regressor well performed as we get lowest RMSE on XG-Boost Regressor for train and test and have very good model score of 0.99 on train and test.

Most optimimum model is taken on the basis of the RMSE and the Statistical summary/OLS Summary of the model that describe many things in which which variable is most important and which variable is not suited for predicting the accuracy of the target variable.

In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one.

Note:-Some of the important fetures that will helps to predict the expected_CTC(dependent variable) in a better way having good positive cofficient value.

The coefficient for Current_CTC is 0.9882208458359613
The coefficient for Inhand_Offer is 0.03233039380381115
The coefficient for Last_Appraisal_Rating is 0.03885095999869427

The Durbin Watson statistic is a test for autocorrelation in a regression model's output. The DW statistic ranges from zero to four, with a value of 2.0 indicating zero autocorrelation.

Values below 2.0 mean there is positive autocorrelation and above 2.0 indicates negative autocorrelation.

In training OLS summary Durbin Watson is 1.991 and in testing summary 2.014

Note:-Certifications is not good variable to predict the expected_ctc so we can drop it and build he model again and check the statsmodel summary and rmse and accuracy of the model again.

So,we can also perform PCA to reduce the no of features according to the most important features selection method that will reduce the dimensionality in the dataset.

Note:-XG boost model will be a good model having less RMSE compared to Linear Regression and Z-score with linear regression apart from that we can consider the Linear Regression Model will be a good model because.

Note:-

If the training set's R-squared is higher and the R-squared of the test set is much lower, it indicates overfitting. If the same high R-squared translates to the test set as well, then we can say that the model is a good fit.

6.	Final interpretation / recommendation

	Detailed recommendations for the management/client based on the analysis done.

Later on We can make a Expected_CTC range like for fresher applicants(for eg:0to 5 lakhs), less experience applicants(5 lakhs to 10 lakhs), moderate experience applicants(10 to 20 lakhs) and extreme experience applicants (more than 20 lakhs)with the help of clustering techniques with respect to expected_CTC .

Statistically This can be done done by group by function finding features importance of dataset with respected target column by giving conditions.

As we know clustering is a unsupervise technique When you have a set of unlabeled data, it's very likely that you'll be using some kind of unsupervised learning algorithm.

Clustering is especially useful for exploring data you know nothing about. It might take some time to figure out which type of clustering algorithm works the best, but when you do, you'll get invaluable insight on your data.

We can use K-Means algorithm: The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of O(n).

Other Business Insights

- Total_Experience having strong positive relationship with respect to Expected_CTC as the Total_Experience increases the Expected_CTC will also increases.
- Total_Experience_in_field_applied having quite cloudy relationship with respect to Expected_CTC.
- we can infer only that as the Total_Experience_in_field_applied is increases the Expected_CTC will also get slightly increases.
- Passing_Year_Of_Graduation have negative relation with Expected_CTC as the oldest year having higher Expected_CTC where as the latest year has lowest Expected_CTC.

- Passing_Year_Of_PG has no clearly no such clear relationship with respect to Expected_CTC.
- Passing_Year_Of_PHD having negative corelation with the respect of Expected_CTC as the latest years having more Expected_CTC compared to past no of years.
- Current_CTC having positive corelation with the respect of Expected_CTC as the Current_CTC is increasing the Expected_CTC is also increasing.

- CEO and Head have highest median value that means salary range are high with respect to Expected_CTC.
- Associate has least median value means Expected_CTC has low salary with respect to Associate.
- Most of the label type of Role Feature has Outliers in terms of Expected_CTC.

Note:-

- Total_Experience and Current_CTC will make a imapct on predicting the right Expected_CTC for company as well as applicant.
- Features that have negative correlation with the target variable can play a crucial role to predict the Expected_CTC of an applicant.
- Apart from that We can consider the high qualification out of total qualification details that is given like Graduation, PG, PHD and can make a seprate column for higher education that can be a good predictor to attain better accuracy of the model.

Recommendations

➢ Instead of putting all kind of educational qualification of particular applicant we have to more focus on the higher education of that particular applicant because there is no such high impact on the target variable that we want to predict.

- We can rate a particular applicant experience level in that specific domain using the profile of that particular applicant in terms of how much years he or she has been worked so far.

- Good Key Performer in terms of rating will be a good suggestion for selecting the better person for an organization.

- Person having no gap in terms of education with decent amount of current_CTC will be a good person for the company.

- Organization should check there should be no fraud done by the employee or person in the working history that will become a major variable to predict the expected_CTC(the target column).

Note:-As we infer from our analysis fresher applicants have 0 CTC and having Total Experience 0 for such applicants we need to build separate model for predicting the CTC.

Thankyou