SALARY PREDICTION

ARPITA BAYEN

1) Introduction of the business problem

DEFINING PROBLEM STATEMENT

The Dataset consists of The Dataset Contains historical datas of various salary ranges of old employees.The purpose of this project is to Predict the salary of the new employees.The dataset contains various factors which determines the salary of an individual.

NEED OF THE STUDY/PROJECT

The purpose of this research is to create a strong machine learning model that can predict future employee wages . This project will help to avoid biasness in offering salary to candidates.It will keep the salary of the future employees standardised.

To ensure there is no discrimination between employees.

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

From the perspective of recruiters, a salary prediction model is beneficial for improving recruitment and salary standards, as well as for providing more reasonable salaries to attract and discover talents.

2)Data Report

UNDERSTANDING HOW DATA WAS COLLECTED IN TERMS OF TIME, FREQUENCY AND METHODOLOGY.

Ans :As No date is mentioned in the dataset, it is not possible to say for how long has been collected and what was the frequency .

The data should not be very old as the standard of living and value of rupees also matters in determining the salary.

The methodology used was to collect all the details of any employee during their joining in the company.

This can be taken by asking them or by telling them to fill the tracker containing these variables.

The experience of employee ranges from 0 years of exp to 25 years of exp.The data was collected from different department ,different hierarchy and of different experience.

VISUAL INSPECTION OF DATA (ROWS, COLUMNS, DESCRIPTIVE DETAILS).

1)The Data Consists of 25000 employees and 29 variables which determines the salary of an employee.

2)There are 3 float type,10 integer type and 16 object type variables.

3)There are no duplicate variables.

4)There are many missing values in the dataset.

Table 1: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 2:First few row of the dataset

D X	pplicant_ID	otal_Experience	otal_Experience_in_field_applied	epartment	Role	ndustry	rganization	esignation	ducation	raduation_Specialization
1	22753	0	0	aN	aN	aN	aN	aN	G	ts
2	51087	23	14	R	onsultant	nalytics		R	octorate	emistry
3	38413	21	12	op Management	onsultant	aining		aN	octorate	ology
4	11501	15	8	anking	nancial Analyst	viation		R	octorate	thers
5	58941	10	5	les	oject Manager	surance		edical Officer	rad	ology
D X	niversity_Grad	assing_Year_Of_Graduation	G_Specialization	niversity_PG	assing_Year_Of_PG	HD_Specialization	niversity_PHD	assing_Year_Of_PHD	urent_Location	
1	cknow	2020	aN	aN	aN	aN	aN	aN	uwahati	
2	rat	1988	thers	rat	1990	emistry	angalore	1997	angalore	
3	ipur	1990	ology	ipur	1992	ology	cknow	1999	nmedabad	
4	angalore	1997	ology	angalore	1999	emistry	uwahati	2005	anpur	
5	umbai	2004	ology	umbai	2006	ology	angalore	2010	nmedabad	
D X	referred_location	urrent_CTC	nhand_Offer	ast_Appraisal_Rating	lo_Of_Companies_worked	umber_of_Publications	ertifications	ternational_degree_any	xpected_CTC	
1	une	0		aN	0	0	0	0	84551	
2	agpur	2702664		ey_Performer	2	4	0	0	783729	
3	ipur	2236661		ey_Performer	5	3	0	0	131325	
4	olkata	2100510			5	3	0	0	608833	
5	nmedabad	1931644			2	3	0	0	221390	

UNDERSTANDING OF ATTRIBUTES (VARIABLE INFO, RENAMING IF REQUIRED)

Table 3:Statistical description of the dataset

	IDX	Applicant_ID	Total_Experience	Total_Experience_in_field_applied	Passing_Year_Of_Graduation	Passing_Year_Of_PG	Passing_Year_Of_PHD
count	25000	25000	25000	25000	18820	17308	13119
mean	12500.5	34993.24	12.49308	6.2582	2002.194	2005.154	2007.396
std	7217.023	14390.27	7.471398	5.819513	8.31664	9.022963	7.493601
min	1	10000	0	0	1986	1988	1995
25%	6250.75	22563.75	6	1	1996	1997	2001
50%	12500.5	34974.5	12	5	2002	2006	2007
75%	18750.25	47419	19	10	2009	2012	2014
max	25000	60000	25	25	2020	2023	2020

	Current_CTC	No_Of_Companies_worked	Number_of_Publications	Certifications	International_degree_any	Expected_CTC	
count	2.50E+04	25000	25000	25000	25000	2.50E+04	
mean	1.76E+06	3.48204	4.08904	0.77368	0.08172	2.25E+06	
std	9.20E+05	1.690335	2.606612	1.199449	0.273943	1.16E+06	
min	0.00E+00	0	0	0	0	2.04E+05	
25%	1.03E+06	2	2	0	0	1.31E+06	
50%	1.80E+06	3	4	0	0	2.25E+06	
75%	2.44E+06	5	6	1	0	3.05E+06	
max	4.00E+06	6	8	5	1	5.60E+06	

Total Experience ranges from 0 to 25 .Expected Salary ranges upto 60 LPA.

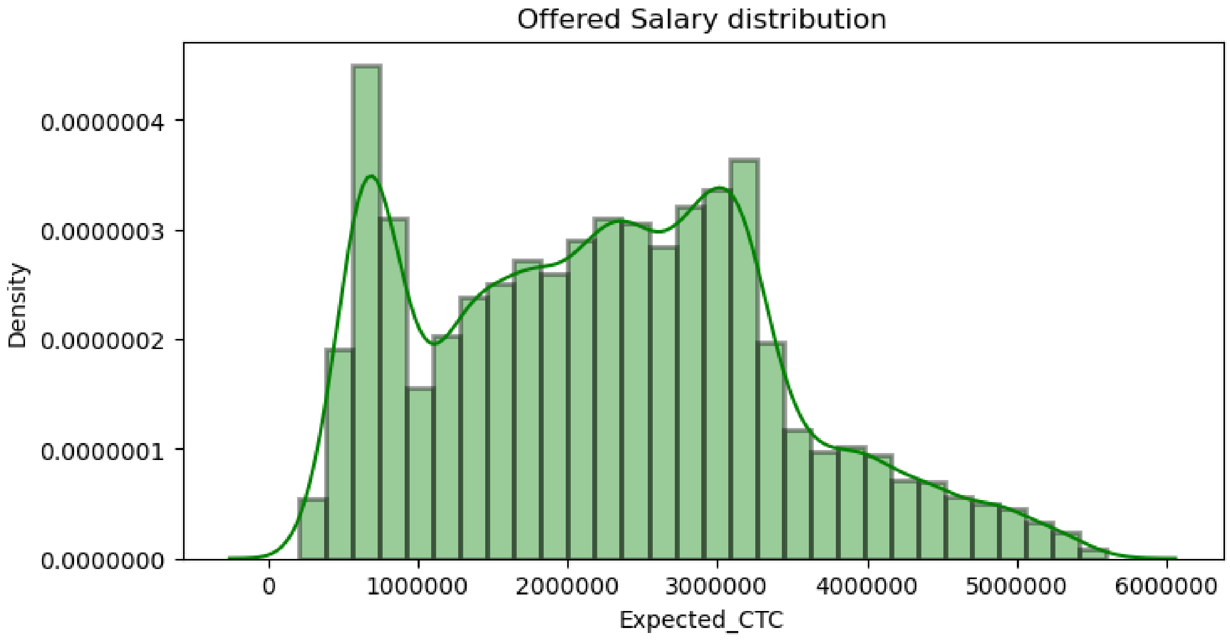
Table 4:Unique vales of the categorical variables

Department	**12**	**PF Specialization**	**11**
Role	24	University PG	13
Industry	11	PHD Specialization	11
Organisation	16	University PHD	13
Designation	18	Current Location	15
Education	4	Preferred Location	15
Graduation Specialization	11	Inhand Offer	2
University Grad	13	Last Appraisal Rating	5

3) Exploratory data analysis

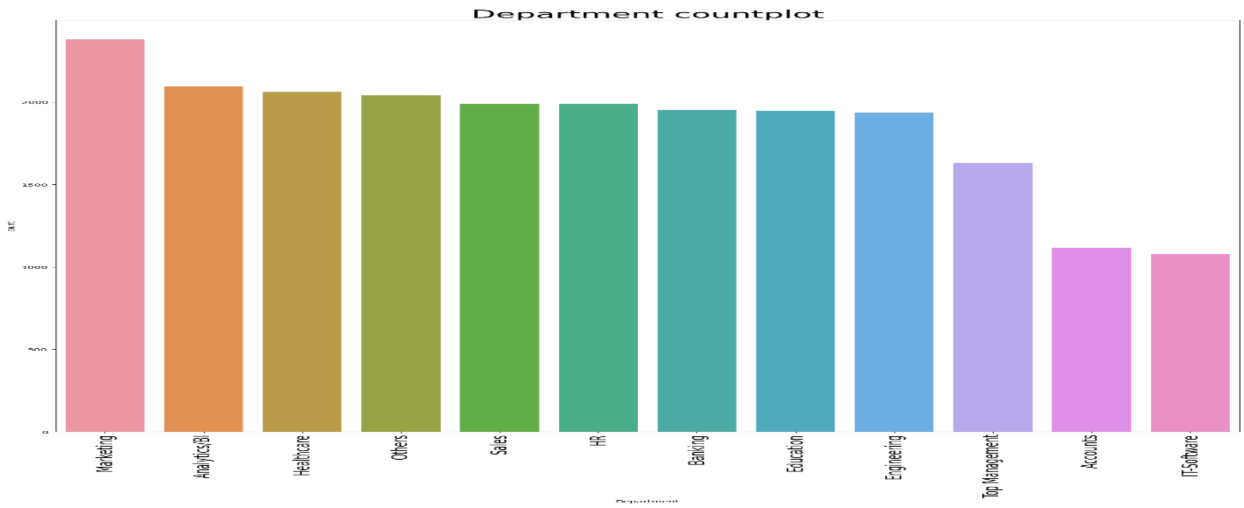
UNIVARIATE ANALYSIS (DISTRIBUTION AND SPREAD FOR EVERY CONTINUOUS ATTRIBUTE, DISTRIBUTION OF DATA IN CATEGORIES FOR CATEGORICAL ONES)

Fig1:Offered Salary Distribution

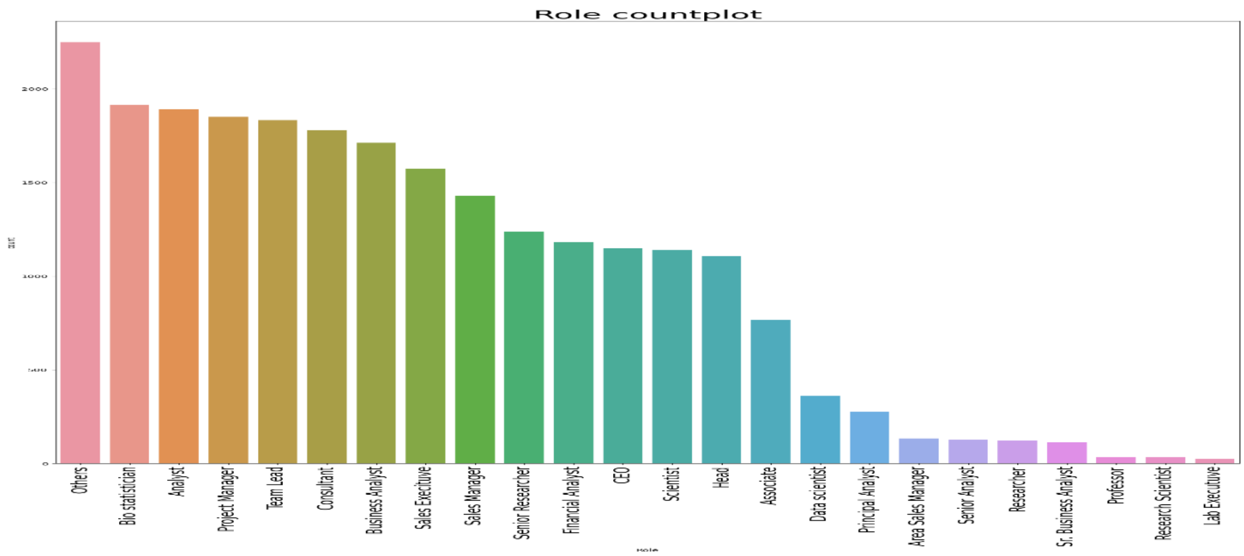


Offered salary ranges upto 60 LPA

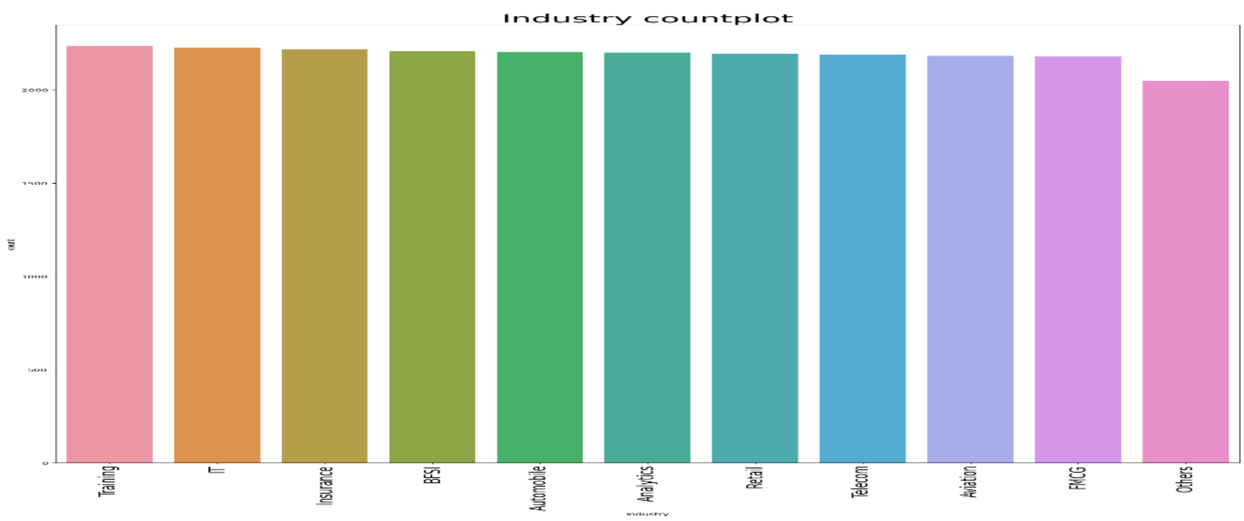
Fig 2:Countplot of All Categorical Variable



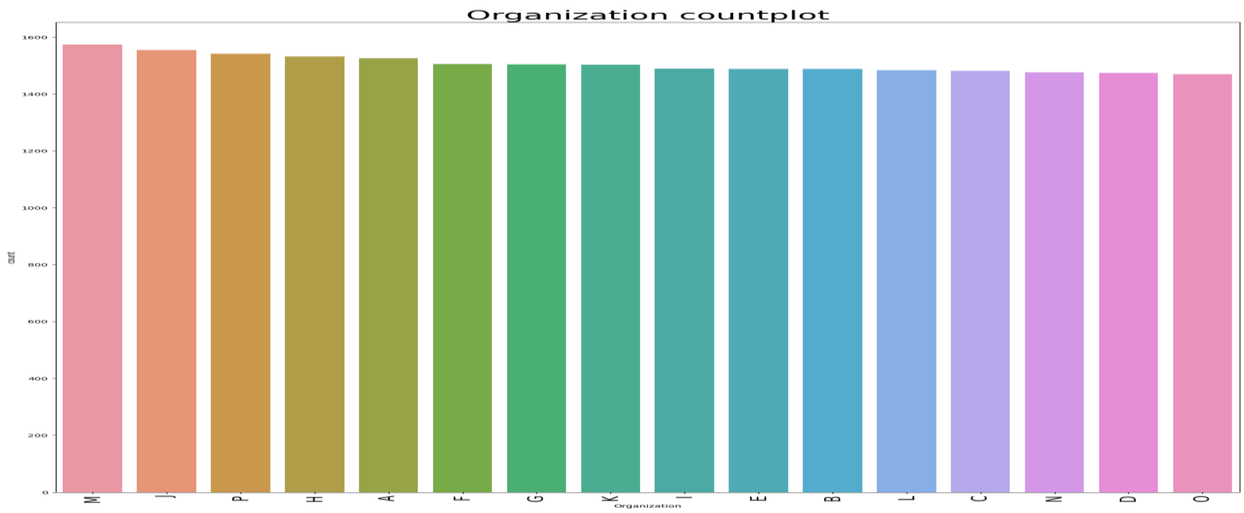
Marketing Department has the highest employee number .IT software department has the lowest number of employee in the dataset.



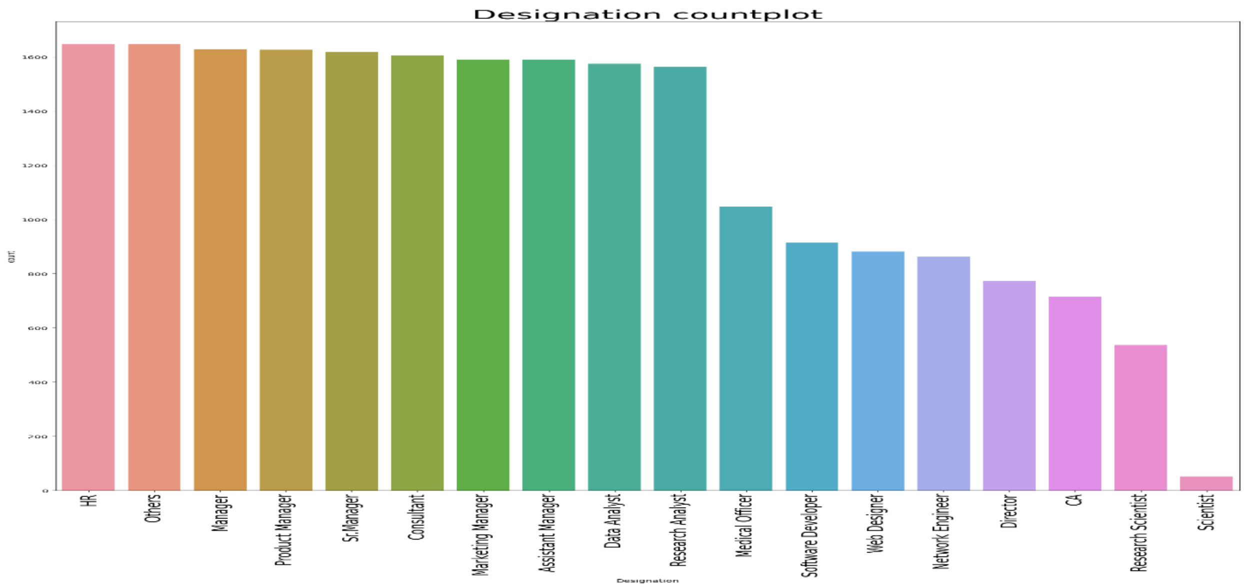
Researcher,Lab executives ,Research Scientist,Professor are the lowest among all the roles in the dataset



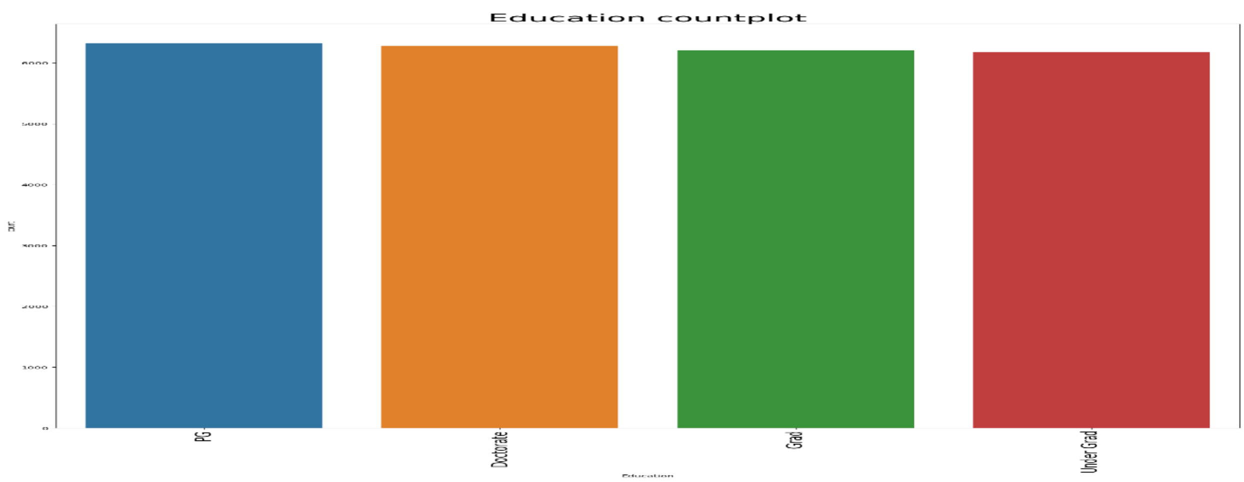
There are 11 industry taken into account.



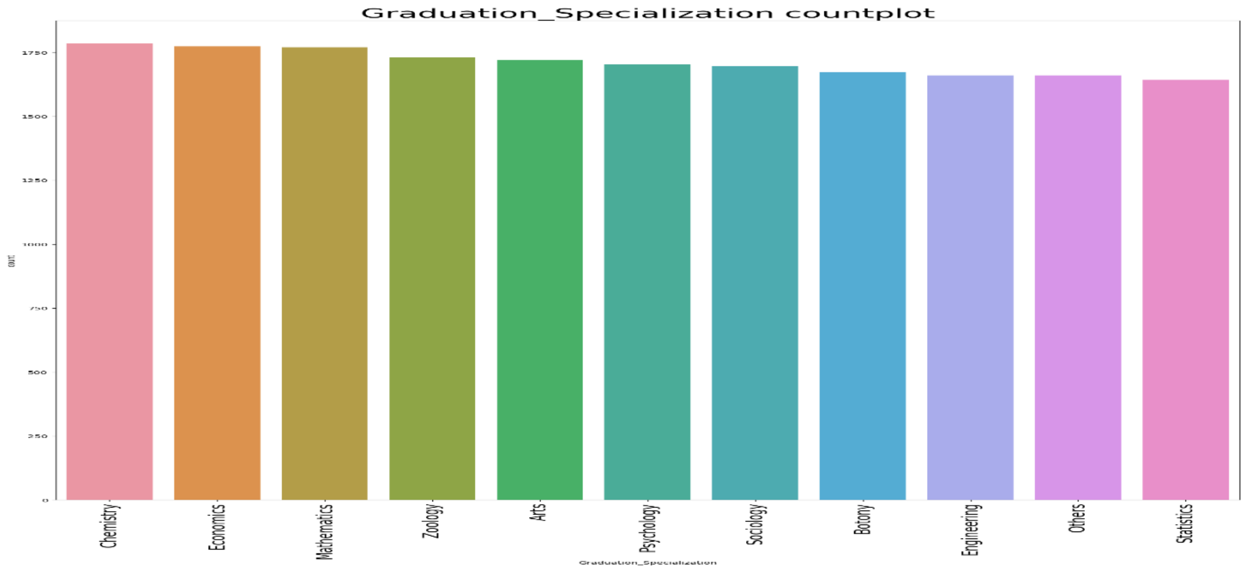
There are 16 Organisations taken into account.



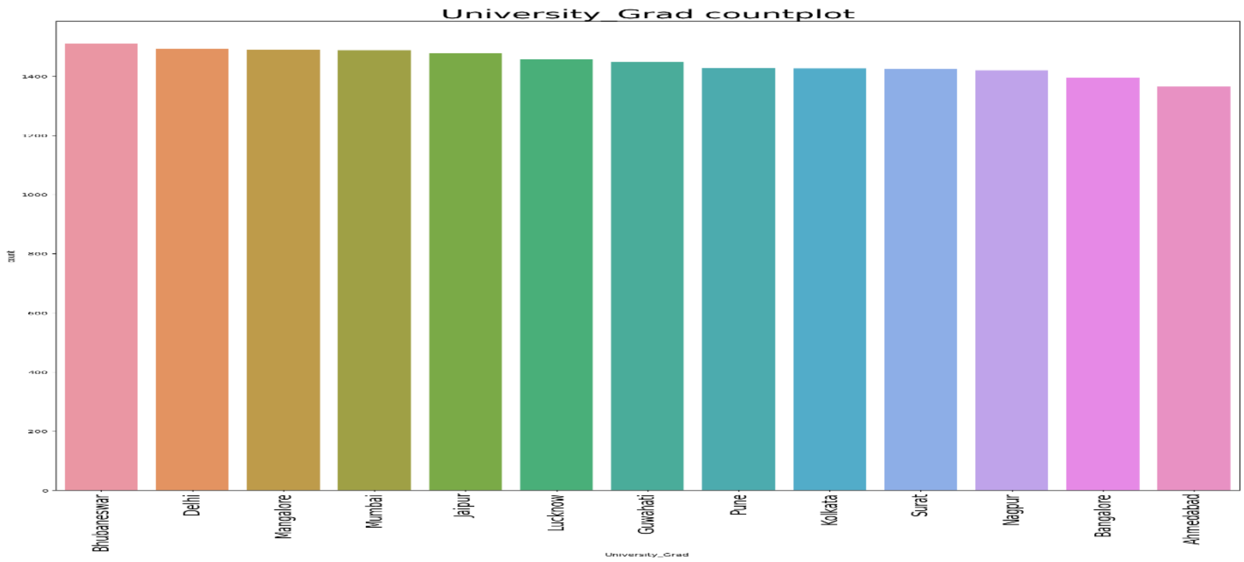
There are 18 Designations



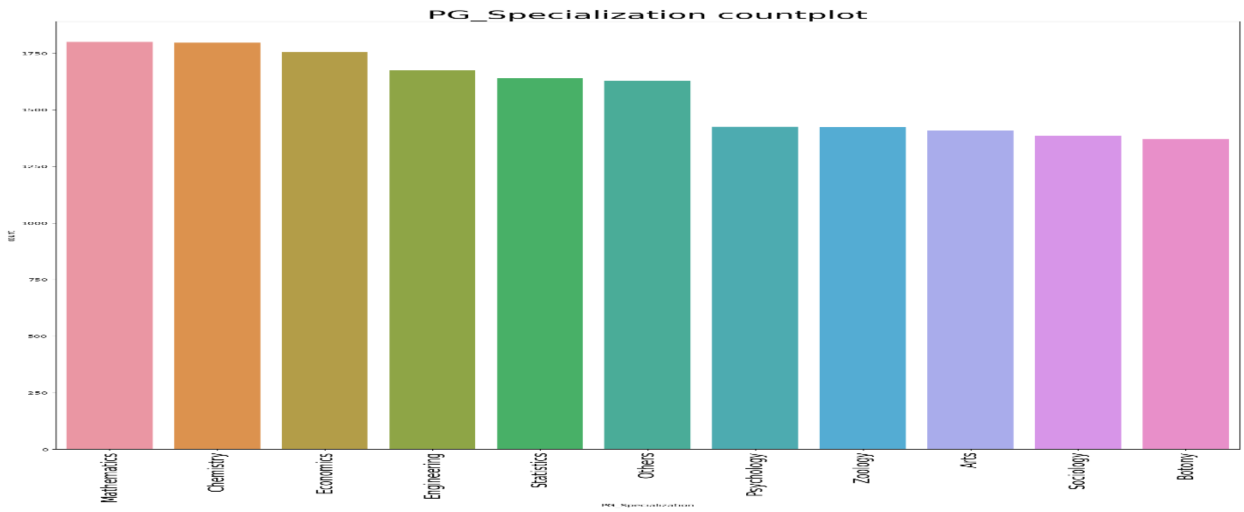
There are 4 Education and Everyone has equally participated



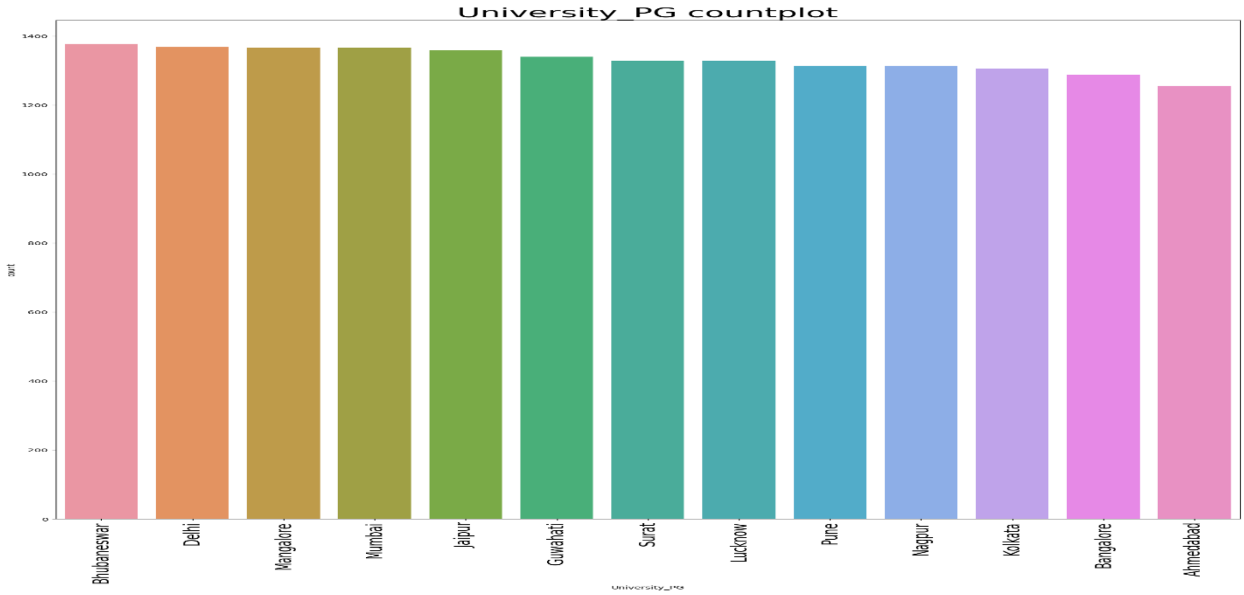
There are 11 Graduation specializations



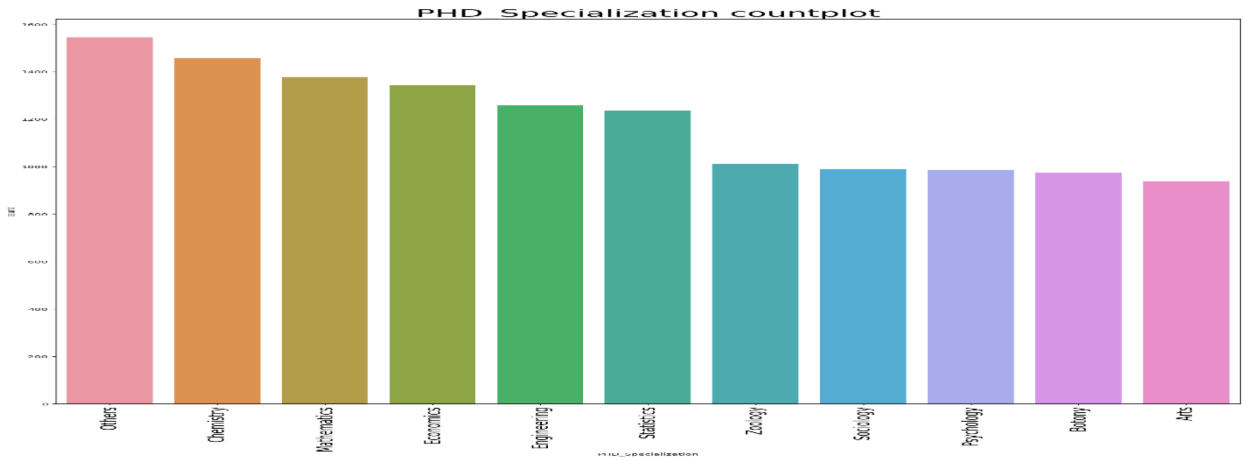
There are 13 state universities.



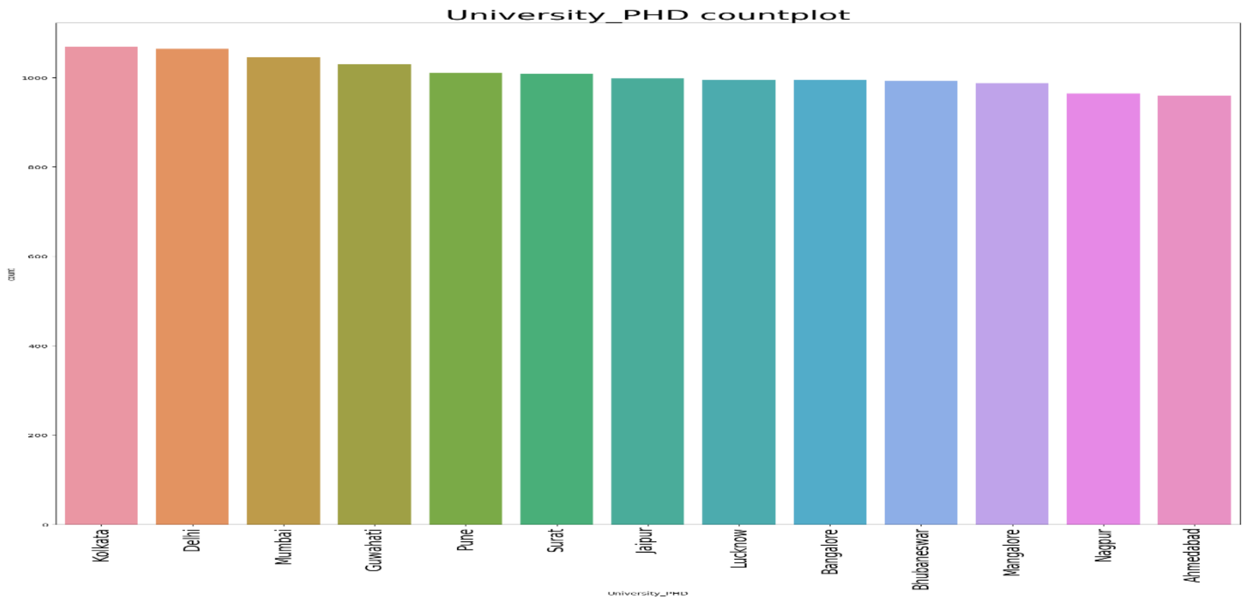
There are 11 PG Specialisation



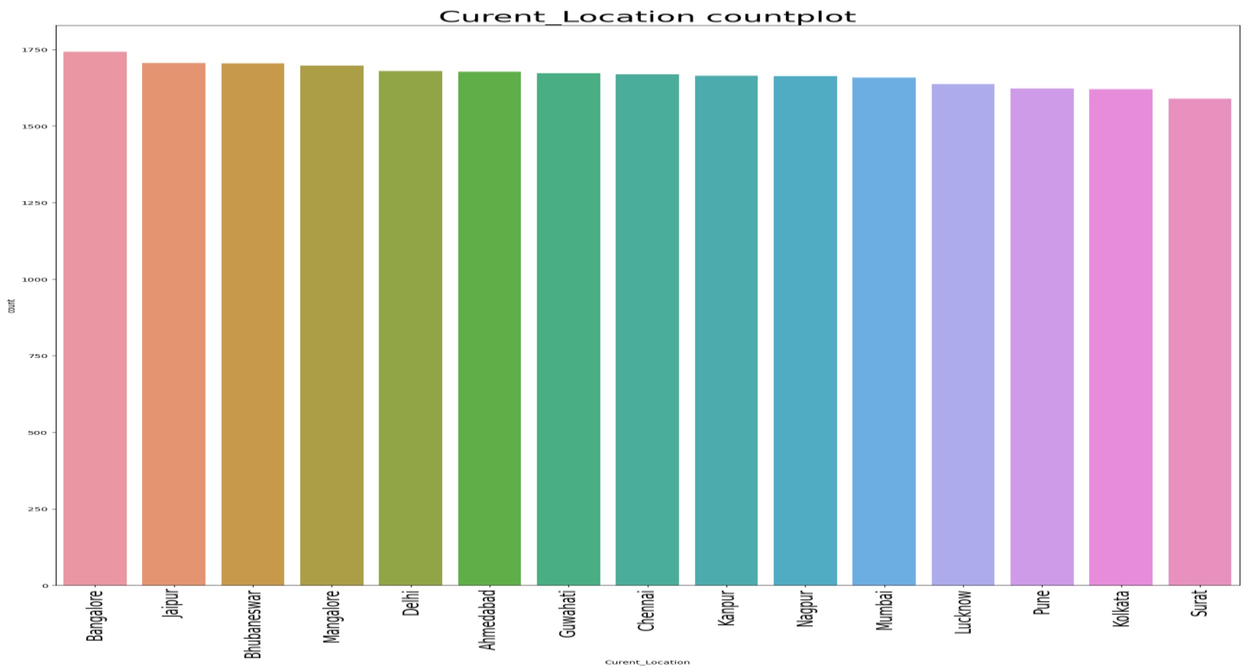
There are 13 state PG universities.



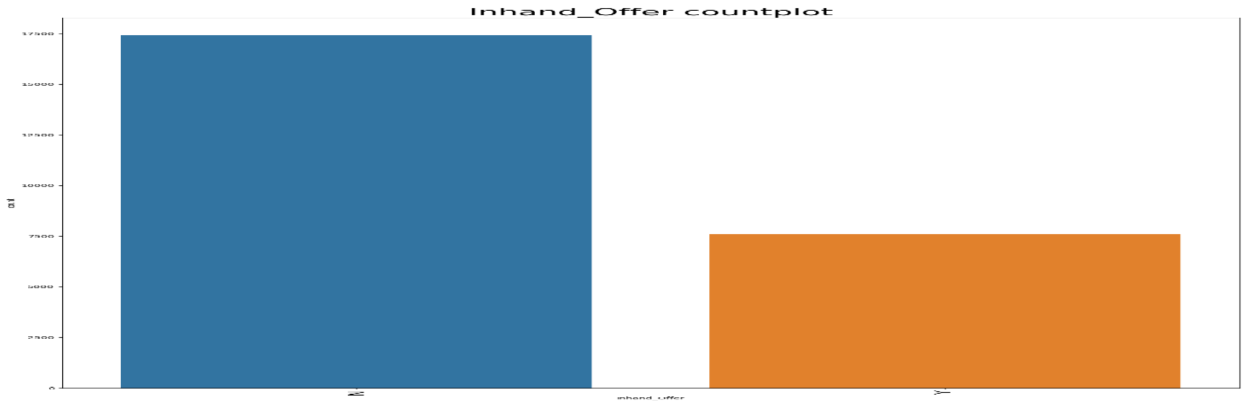
There are 11 PHD Specialization



There are 13 PHD State Universities



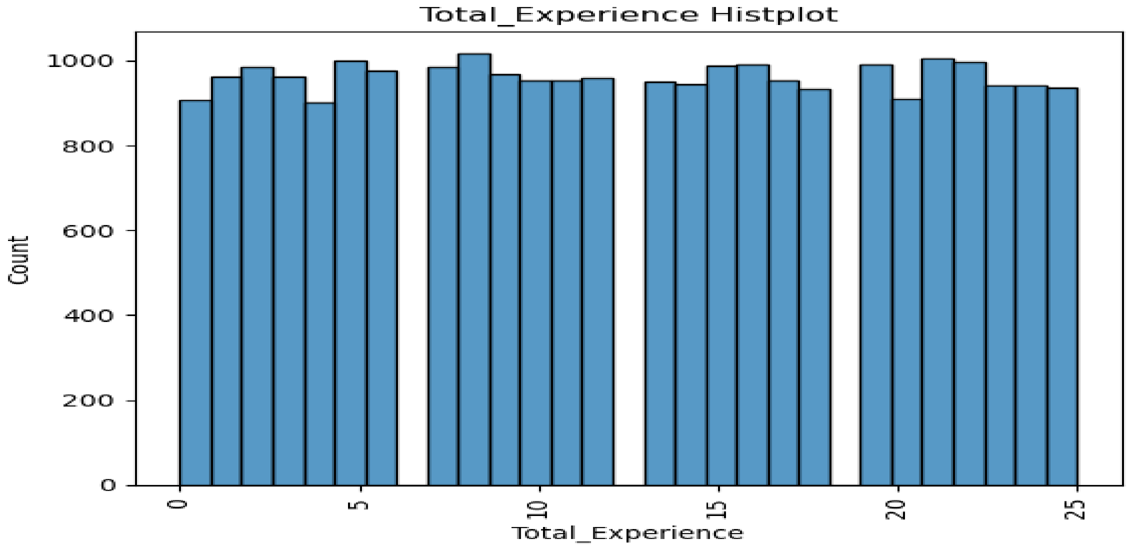
There are 15 Current and Preferred Locations.



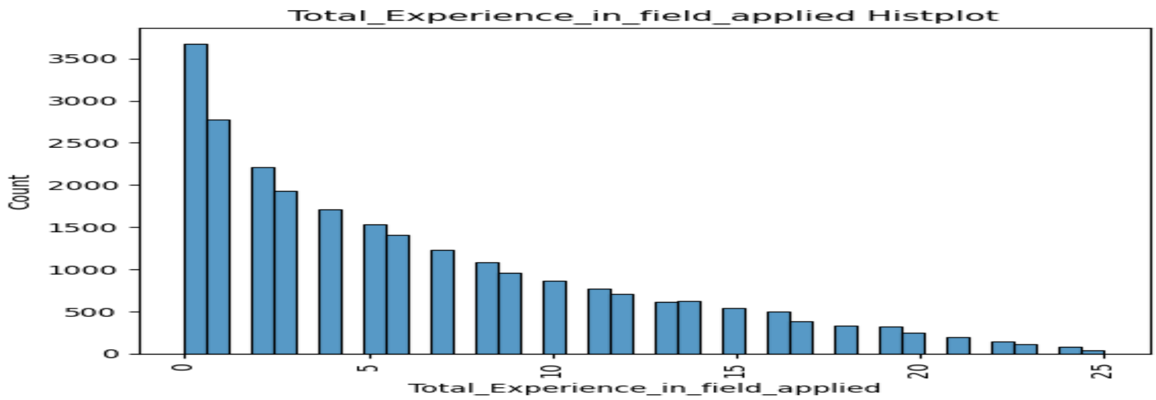
Some Candidates already had Offer in hand before getting selected in this company.

Last Appraisal rating is also taken into consideration

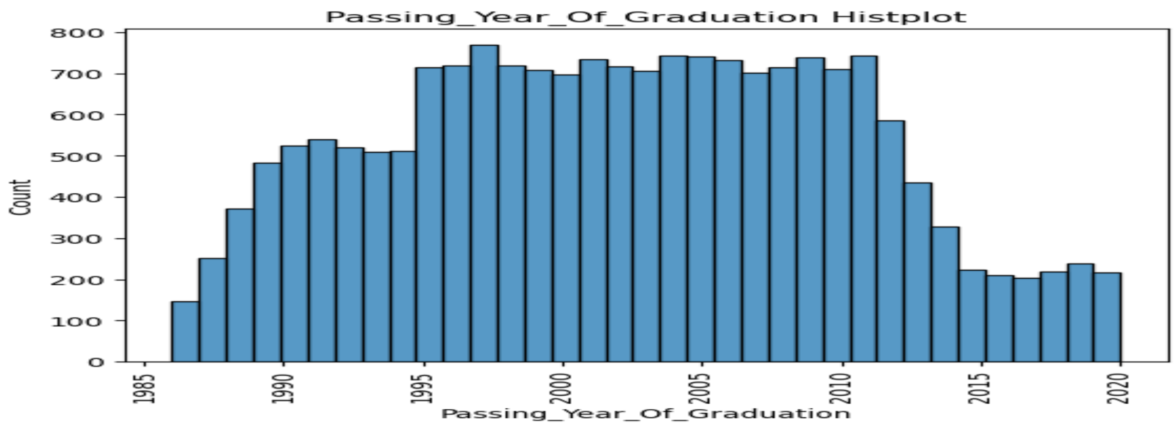
Fig 3:Histplot of all Numerical Variables



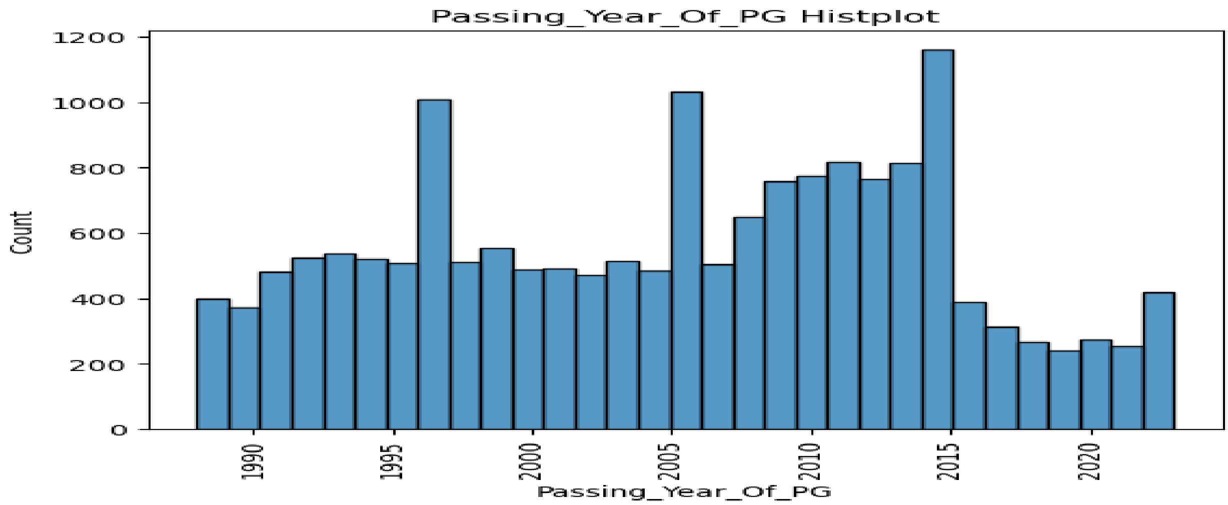
The Experience range is from 0 years to 25 years.



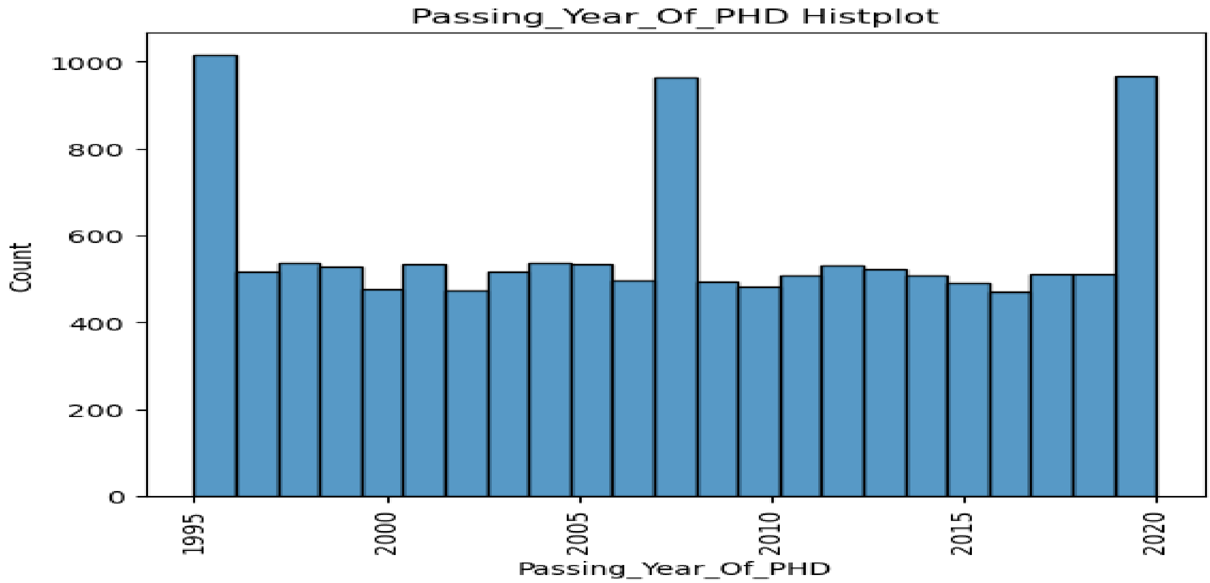
Relevant experience is not same as total experience as the person may be having less relevant experience but he is experienced in other domain.



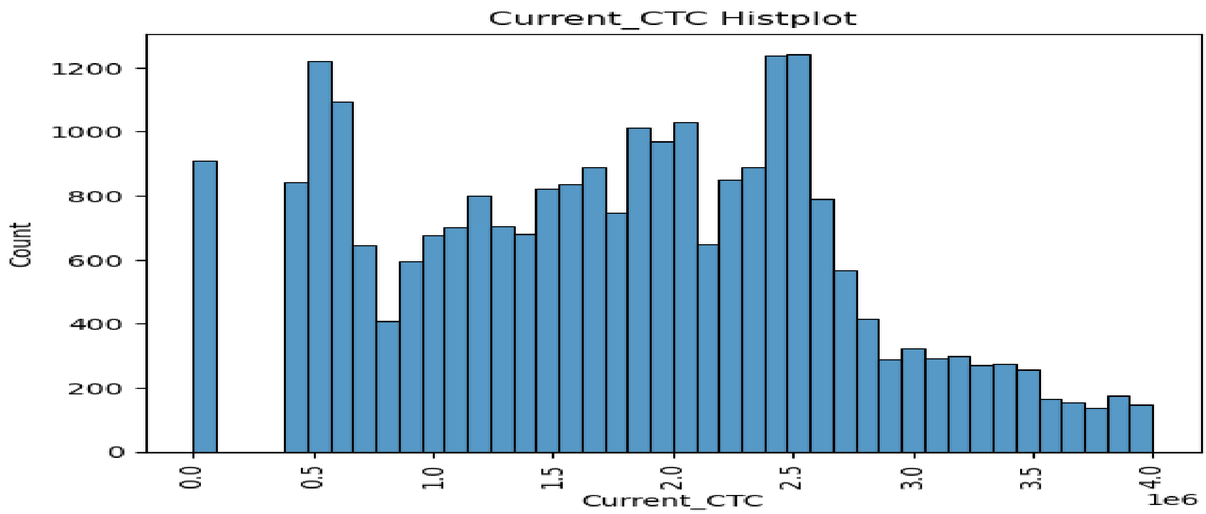
The passing year of graduation is from 1985 to 2020.

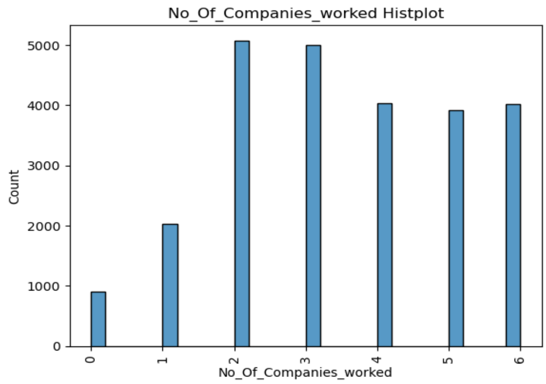


The passing year of graduation is from 1990 to 2023

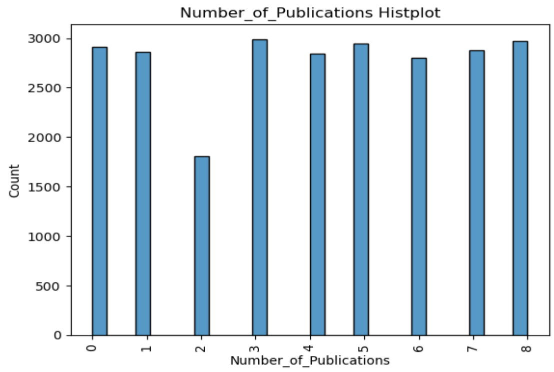


The passing year of PHD is from 1995 to 2020.

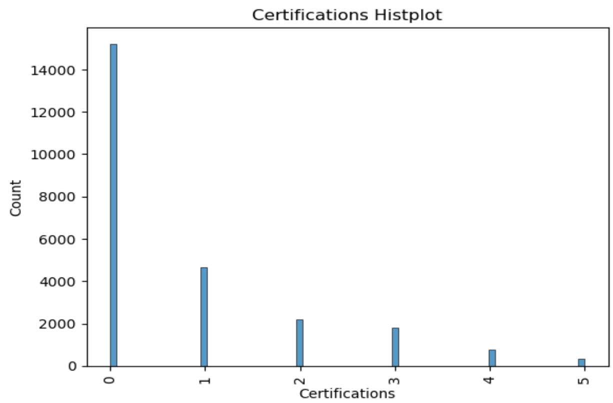




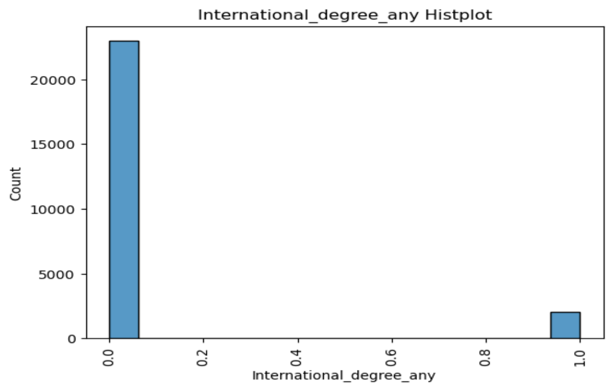
The data set also contains how many companies an employee has changed .This will give the stability of the employee



Number of publications will be helpful in recruiting a scientist or researcher .



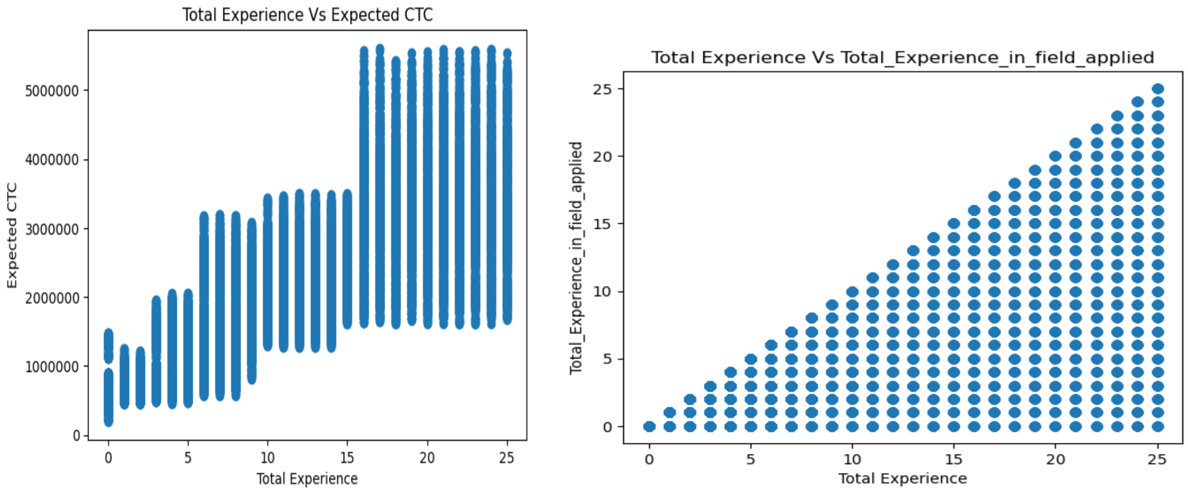
Candidates have also show their certificates.



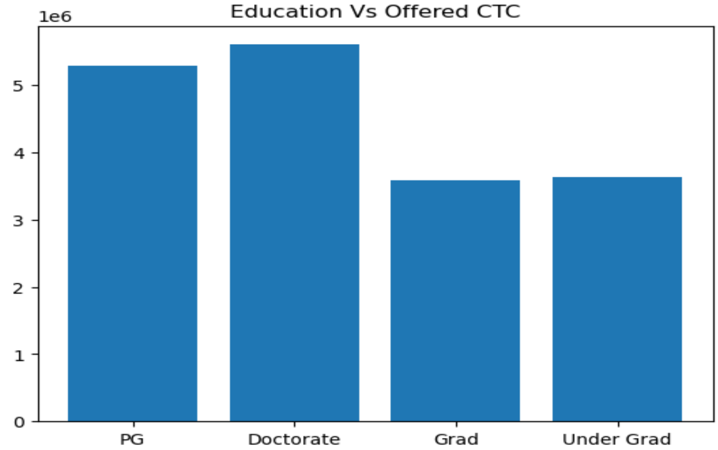
Candidates have given if they have any international degree or not.

BIVARIATE ANALYSIS (RELATIONSHIP BETWEEN DIFFERENT VARIABLES , CORRELATIONS)

Fig 4:Bivariate Graph

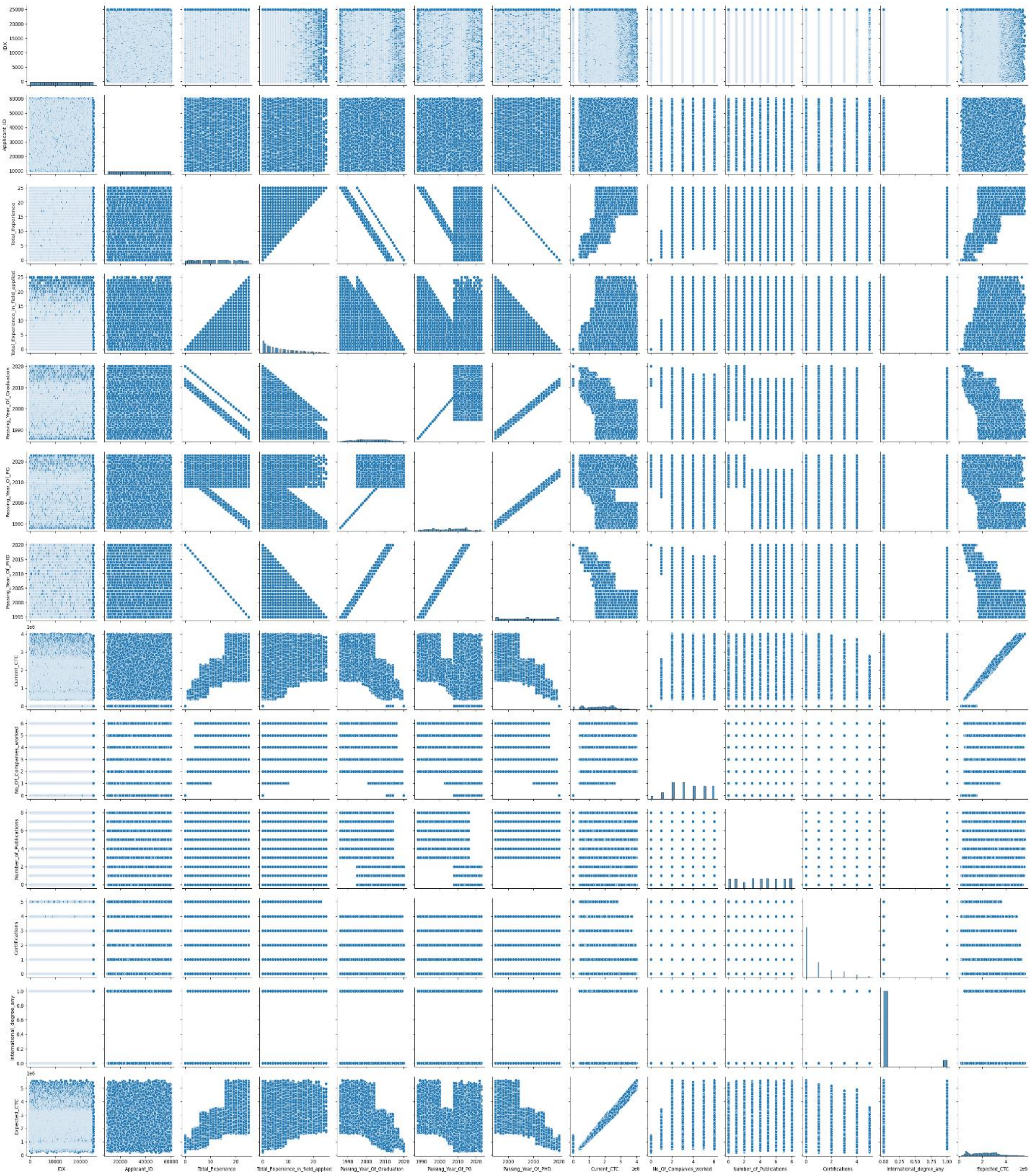


Expected CTC is increasing with total experience but many less experienced candidates are getting higher than more experienced candidate.Total Experience and relavant experience are not related.



Graduate and under Graduate candidates are getting less than PG .Doctorate employees are getting more than PG employees.

Fig 5:Pairplot



Observations:

There are positive relations and negatve relations among variables,but many are obvious relations like 1)earlier the passing year more is the experience.

2)the current ctc and expected ctc will increase at a same rate.

3)Relevant Experience is not related to expected CTC.

REMOVAL OF UNWANTED VARIABLE

The variables Index and Applicant ID are removed from the dataset.

MISSING VALUE TREATMENT

Table 5:Missing values in percentage

```
Total_Experience                       0.000
Total_Experience_in_field_applied      0.000
Department                            11.112
Role                                   3.852
Industry                               3.632
Organization                           3.632
Designation                           12.516
Education                              0.000
Graduation_Specialization             24.720
University_Grad                       24.720
Passing_Year_Of_Graduation            24.720
PG_Specialization                     30.768
University_PG                         30.768
Passing_Year_Of_PG                    30.768
PHD_Specialization                    47.524
University_PHD                        47.524
Passing_Year_Of_PHD                   47.524
Curent_Location                        0.000
Preferred_location                     0.000
Current_CTC                            0.000
Inhand_Offer                           0.000
Last_Appraisal_Rating                  3.632
No_Of_Companies_worked                 0.000
Number_of_Publications                 0.000
Certifications                         0.000
International_degree_any               0.000
Expected_CTC                           0.000
```

There are so many columns where the percentage of missing values are equal to or more than 30%.3 variables has 24% of missing values which are almost close to 30%. It is better to drop those variables as imputing those missing values will create a synthetic data.

After removing variables containing more than 24 % missing values the dataset contains 25000rows and 18 variables.

Dropping all the rows containing null values.-The dataset contains 20307 rows and 18 variables.

Around 18.72 % of data are removed from the dataset.

ADDITION OF NEW VARIABLE

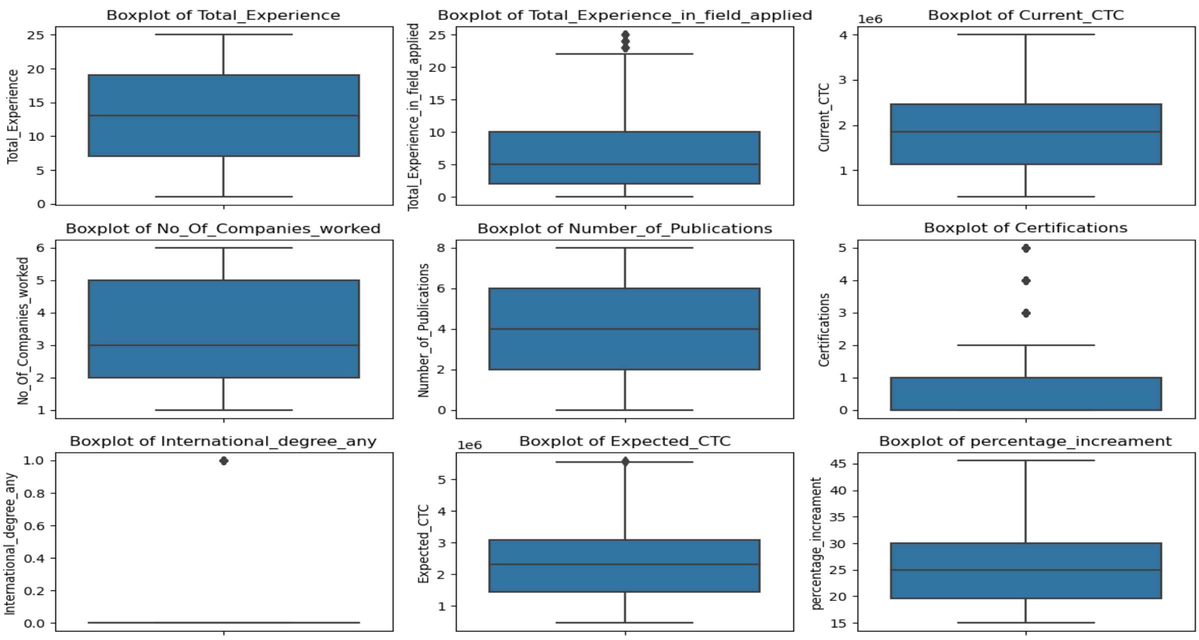
One new variables is added –Percentage increament which is the percentage of increament given on the Current CTC .

The formula used is –Percentage_Increament=[(Expected CTC-Current CTC)/Current CTC]*100

OUTLIER TREATMENT

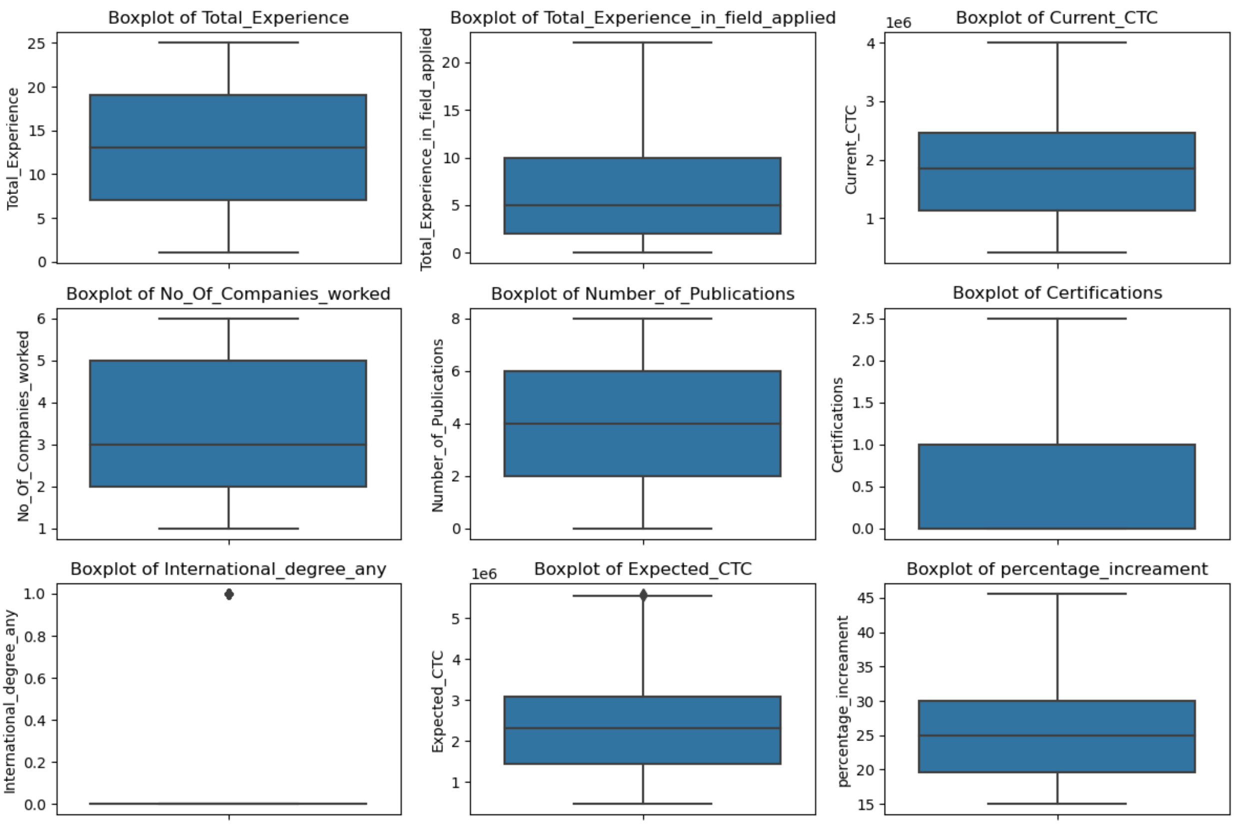
There are outliers present in some variables

Fig 6:Outliers of Numerical Variable



Outliers are removed from the dataset by replacing them by 0.25 or0.75 of the values.

Fig 7:After removing Outliers



SCALING OF DATA

As every numerical variable except Current CTC and Expected CTC have 1 or 2 digit value so they do not need to be scaled.

So only Current CTC and Expected CTC has been scaled by using standard scaler.

Table 6:First few rows of the dataset after scaling

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Curent_Location
1	23	14	HR	Consultant	Analytics	H	HR	Doctorate	Bangalore
3	15	8	Banking	Financial Analyst	Aviation	F	HR	Doctorate	Kanpur
4	10	5	Sales	Project Manager	Insurance	E	Medical Officer	Grad	Ahmedabad
5	16	3	Top Management	Area Sales Manager	Retail	G	Director	Doctorate	Pune
6	1	1	Engineering	Team Lead	FMCG	L	Marketing Manager	Grad	Delhi

Preferred_location	Current_CTC	Inhand_Offer	Last_Appraisal_Rating	No_Of_Companies_worked	Number_of_Publications	Certifications	International_degree_any	Expected_CTC	percentage_increament
Nagpur	1.007712	Y	Key_Performer	2	4	0	0	1.295859	39.99998
Kolkata	0.313262	N	C	5	3	0	0	0.257249	24.19998
Ahmedabad	0.118513	N	C	2	3	0	0	-0.08525	14.99997
Bhubaneswar	1.940138	Y	C	5	4	0	0	1.94883	28.8
Pune	-1.54074	Y	B	3	3	0	0	-1.49122	27.99985

LABEL ENCODING

As there are many categorical variables so they needs to be changed to numerical for easiness in making models.

Here Label encoder is used

Table 7: First few rows of the dataset after label encoding

	Total_Experience	Total_Experience_in_field_applied	Department	Role	Industry	Organization	Designation	Education	Curent_Location
1	23	14	5	6	0	7	5	0	1
3	15	8	2	8	2	5	5	0	7
4	10	5	10	14	6	4	8	1	0
5	16	3	11	1	8	6	4	0	13
6	1	1	4	23	4	11	7	1	4

Preferred_location	**Current_CTC**	**Inhand_Offer**	**Last_Appraisal_Rating**	**No_Of_Companies_worked**	**Number_of_Publications**	**Certifications**	**International_degree_any**	**Expected_CTC**	**percentage_increament**
12	1.007712	1	4	2	4	0	0	1.295859	39.99998
8	0.313262	0	2	5	3	0	0	0.257249	24.19998
0	0.118513	0	2	2	3	0	0	-0.08525	14.99997
2	1.940138	1	2	5	4	0	0	1.94883	28.8
13	-1.54074	1	1	3	3	0	0	-1.49122	27.99985

All the categorical variables are changed to numerical label.

SPLITTING DATA INTO TRAIN AND TEST

The dataset is splitted with test set 70 %.

```
X_train Shape- (6092, 18)
X_test Shape- (14215, 18)
Y-train Shape- (6092,1)
Y_test Shape- (14215,1)
```

Table 8-X_train first few rows of datasets

	Total_Exp	Total_Exp	Departme	Role	Industry	Organizat	Designati	Education	Curent_L	Preferred	Current_C TC	Inhand_O	Last_Appr	No_Of_C	Number_	Certificati	Internatio	percentag e_increa
20708	12	4	5	21	9	14	13	1	8	3	0.21	0	3	2	7	0	0	15.00
11551	7	3	9	11	10	0	10	3	0	3	-1.04	0	2	1	3	2	0	15.00
1630	20	20	11	22	3	3	10	2	10	4	0.88	1	0	5	1	1	0	32.00
19873	15	3	3	23	3	8	0	2	3	14	-0.28	1	0	2	3	0	0	32.00
17295	20	12	7	23	7	14	16	2	11	9	1.52	1	1	3	3	2	0	32.00

Table 9 X_test first few rows of datasets:

	Total_Exp	Total_Exp	Departme	Role	Industry	Organizat	Designati	Education	Curent_L	Preferred	Current_C TC	Inhand_O	Last_Appr	No_Of_C	Number_	Certificati	Internatio	percentag e_increa
16352	25	6	0	17	1	11	5	1	1	14	-0.38	0	2	4	8	0	0	15.00
8940	22	9	8	17	3	5	2	3	11	0	0.51	0	2	2	5	2.5	0	15.00
14511	23	0	1	6	10	8	6	1	9	5	0.68	1	4	3	7	0	0	28.00
18995	17	9	2	18	6	2	5	2	7	4	0.54	0	0	6	6	0	0	30.00
15304	20	1	10	3	6	10	7	0	13	10	0.67	1	2	3	7	0	0	28.80

Table 10-Y_train first few rows of dataset:

20708	-0.004609
11551	-1.110210
1630	0.972763
19873	-0.194590
17295	1.620325

Table 11-Y_test first few rows of Dataset:

16352	-0.525811
8940	0.256427
14511	0.684480
18995	0.587614
15304	0.698102

IMPORTANT FEATURE SELECTION

As there are 19 variables in the dataset ,it will lead to curse of dimensionality.To reduce the number of features RFE method is used .It will reduce the multicollinearity.

Feature selection is done to speed up the model execution time and make the process easy to handle.

With estimator Random Forest-Important variables are:

Table 12-

```
1 Total_Experience
1 Department
1 Organization
1 Designation
1 Education
1 Curent_Location
1 Preferred_location
1 Current_CTC
1 Inhand_Offer
1 percentage_increament
2 Role
3 Total_Experience_in_field_applied
4 Industry
5 Number_of_Publications
6 Last_Appraisal_Rating
7 No_Of_Companies_worked
8 Certifications
9 International_degree_any
```

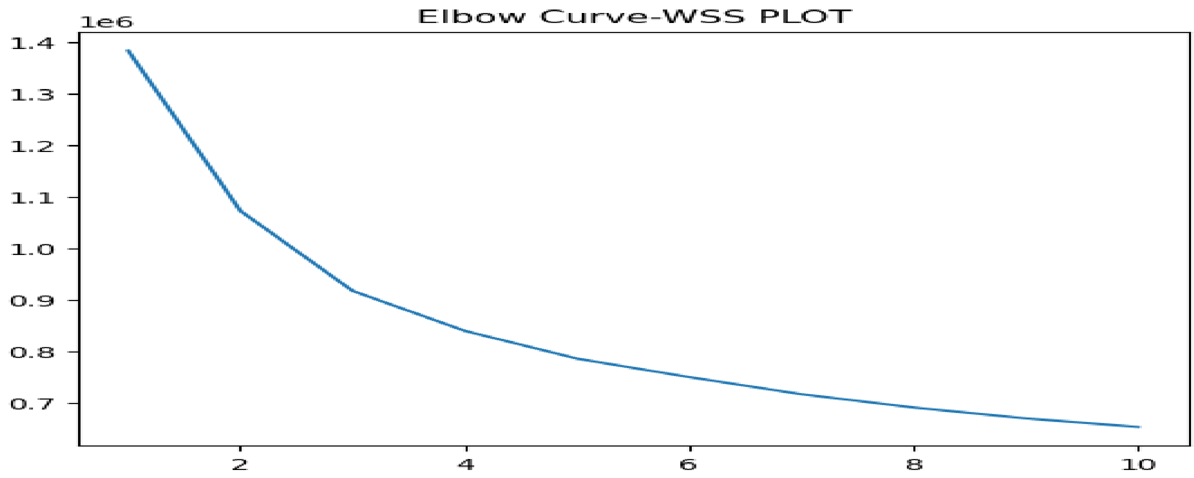
CLUSTERING(K-means)

The dataset has been clustered by K-Means.

WSS Value of clusters 1- 10 are:

```
Cluster1 - 1383934.51,
Cluster2 - 1072401.27,
 Cluster3 - 916831.38,
 Cluster4 - 839085.96,
 Cluster5 - 785194.76,
 Cluster6 - 749169.30,
 Cluster7 - 716021.24,
 Cluster8 - 690068.85,
 Cluster9 - 669021.15,
 Cluster10 - 652407.63.
```

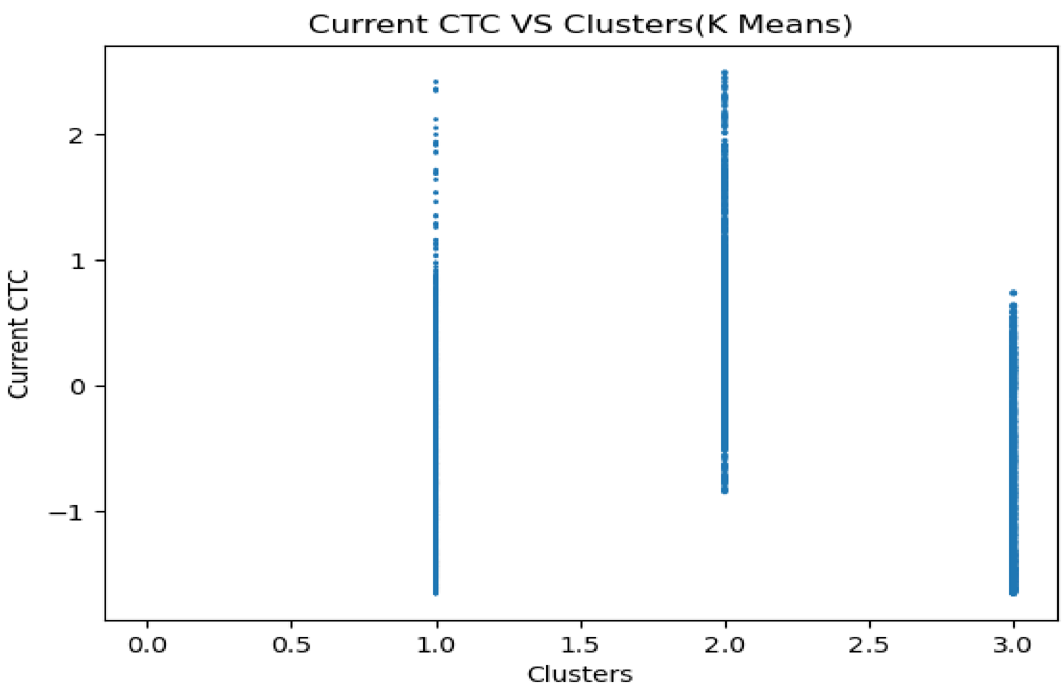
Fig 8-Elbow Curve WSS plot:



Elbow Curve is not clearly showing the break .So I will try the silhouette score.

Silhoutte Score of Cluster 2-0.2017
Silhoutte Score of Cluster 3- 0.1780
Silhoutte Score of Cluster 4- 0.1508
Silhoutte Score of Cluster 5- 0.1478
Silhoutte Score of Cluster 6- 0.1457

Cluster 4 is the breaking point.4 clusters are suitable here.
A new files is prepared with cluster labelling data named ‘ ArpitaBayen_Salary Predictions_K-Means.csv’
Fig 9-Current CTC VS Cluster (Kmeans)



It shows the number of clusters and Current CTC.

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)

The Dataset is preprocessed .

MODEL BUILDING:

Reason to use Regression-It is a Supervised Learning model.It means it has a labelled datasets and a target output.As we have to predict the target variables so Regression Models are used.

This is Regression Model . Regression models describe the relationship between variables by fitting a line to the observed data.

Regression is a tool that allows you to estimate how the dependent variable changes as the independent variable(s) change.

Regression models can be used for many purposes:

- Evaluating the effect of an independent variable on a dependent variable.
- Forecasting future values of the dependent variable based on prior observations of both variables.

SIMPLE LINEAR REGRESSION:

Simple linear regression is a statistical method for establishing the relationship between two variables using a straight line. The line is drawn by finding the slope and intercept, which define the line and minimize regression errors.

Reason-One of the main advantages of using linear regression for predictive analytics is that it is easy to understand and interpret.

$y = \beta 0 + \beta 1x + \varepsilon$ is the formula used for simple linear regression.

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- B0 is the intercept, the predicted value of y when the x is 0.
- B1 is the regression coefficient – how much we expect y to change as x increases.
- x is the independent variable (the variable we expect is influencing y).

- e is the error of the estimate, or how much variation there is in our regression coefficient estimate.

The Simple Linear Regression library is imported from scikit Learn module

The data is fitted to both test set and train set.

<u>The coefficients for each of the independent attributes of the data:</u>

The coefficient for Total_Experience is -0.0016283949877708979
The coefficient for Total_Experience_in_field_applied is -1.229695122647198e-05
The coefficient for Organization is 3.773289454316433e-06
The coefficient for Designation is 1.254819215709172e-05
The coefficient for Education is -0.001016816768916053
The coefficient for Curent_Location is -0.00015226460764011964
The coefficient for Preferred_location is -0.00011115522661619285
The coefficient for Current_CTC is 0.9847694559882453
The coefficient for Inhand_Offer is 0.0005962556758171019
The coefficient for percentage_increament is 0.015582788554075647

<u>The intercept for the set model:</u>

The intercept for our model is -0.3850670231797762

we can write our Linear model as:

Y=-0.385-0.0016*(Total Experience)-(1.2296e-05) *(Total Experience in field applied)+(3.77e-06)*(Organisation)+(1.25e-05)*(Designation)-0.00101*(Education)-0.000152*(Current Location)-0.00011*(Preferred Location)+0984*(Current CTC)+0.005*(Inhand Offer)+0.0155*(percentage increament)

<u>Fitting and predicting the Model on Test dataset</u>

<u>The coefficients for each of the independent attributes in test dataset</u>

The coefficient for Total_Experience is -0.0016837313142405393
The coefficient for Total_Experience_in_field_applied is 5.177701944692249e-05
The coefficient for Organization is 5.677609206226164e-05
The coefficient for Designation is -5.9238794124578736e-05
The coefficient for Education is -0.003661168331945236
The coefficient for Curent_Location is 2.6473697327251974e-05
The coefficient for Preferred_location is 7.00570291232028e-05
The coefficient for Current_CTC is 0.9834772526046177
The coefficient for Inhand_Offer is 0.0029165304793320797

The coefficient for percentage_increament is 0.015503556696892072

we can write our linear model(test set) as:

Y=-0.385-0.0016*(Total Experience)+(5.177e-05)*(Total Experience in field applied)+(5.677e-05)*(Organisation)-(5.92e-05)*(Designation)-0.003*(Education)+(2.647e-05)*(Current Location)+(7.005e-05)*(Preferred Location)+0.984*(Current CTC)+0.0029*(Inhand Offer)+0.0155*(percentage increament)

Table 13: R^2 and RMSE values of Linear Regression

Linear Regression		
	R Square	RMSE
Train Set	0.9969665363118088	0.054989256584532215
Test Set	0.9968765513845881	0.05596938424534416.

Inferences: R^2 is almost equal to 1 RMSE of both set are almost equal ,which means the train set and test set are equally distributed.

ORDINARY LEAST SQUARES (OLS)

Ordinary least squares (OLS) regression is an optimization strategy that helps to find a straight line as close as possible to your data points in a linear regression model.

Reason:-OLS is considered the most useful optimization strategy for linear regression models as it can help you find unbiased real value estimates for your alpha and beta. To be more precise, the model will minimize the squared errors.

OLS is imported using statsmodel module.

OLS Regression result on Train set:

Table 14: OLS Regression result on Train set

OLS Regression Results			
Dep. Variable:	Expected_CTC	**R-squared (uncentered):**	0.994
Model:	OLS	**Adj. R-squared (uncentered):**	0.994
Method:	Least Squares	**F-statistic:**	9.500e+04
Date:	Sun, 22 Oct 2023	**Prob (F-statistic):**	0.00
Time:	10:17:14	**Log-Likelihood:**	6781.7
No. Observations:	6092	**AIC:**	-1.354e+04
Df Residuals:	6082	**BIC:**	-1.348e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Total_Experience	-0.0098	0.000	-38.188	0.000	-0.010	-0.009
Total_Experience_in_field_applied	7.468e-05	0.000	0.328	0.743	-0.000	0.001
Organization	-0.0031	0.000	-14.580	0.000	-0.004	-0.003
Designation	-0.0025	0.000	-12.682	0.000	-0.003	-0.002
Education	-0.0144	0.001	-13.899	0.000	-0.016	-0.012
Curent_Location	-0.0033	0.000	-14.408	0.000	-0.004	-0.003
Preferred_location	-0.0034	0.000	-15.005	0.000	-0.004	-0.003
Current_CTC	1.0347	0.002	570.018	0.000	1.031	1.038
Inhand_Offer	0.0253	0.002	11.013	0.000	0.021	0.030
percentage_increament	0.0091	0.000	81.276	0.000	0.009	0.009

Omnibus:	**15.750**	**Durbin-Watson:**	**1.988**
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18.919
Skew:	0.045	**Prob(JB):**	7.79e-05
Kurtosis:	3.258	**Cond. No.**	77.5

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 15: OLS Regression result on Test set

Dep. Variable:	Expected_CTC	**R-squared (uncentered):**	0.994
Model:	OLS	**Adj. R-squared (uncentered):**	0.994
Method:	Least Squares	**F-statistic:**	2.211e+05
Date:	Sun, 22 Oct 2023	**Prob (F-statistic):**	0.00
Time:	23:08:11	**Log-Likelihood:**	15731.
No. Observations:	14215	**AIC:**	-3.144e+04
Df Residuals:	14205	**BIC:**	-3.137e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Total_Experience	-0.0097	0.000	-57.727	0.000	-0.010	-0.009
Total_Experience_in_field_applied	-6.183e-05	0.000	-0.415	0.678	-0.000	0.000
Organization	-0.0031	0.000	-22.053	0.000	-0.003	-0.003
Designation	-0.0027	0.000	-20.725	0.000	-0.003	-0.002
Education	-0.0175	0.001	-25.679	0.000	-0.019	-0.016
Curent_Location	-0.0029	0.000	-19.135	0.000	-0.003	-0.003
Preferred_location	-0.0033	0.000	-21.619	0.000	-0.004	-0.003
Current_CTC	1.0339	0.001	863.361	0.000	1.032	1.036
Inhand_Offer	0.0294	0.002	19.348	0.000	0.026	0.032

percentage_increament	0.0091	7.5e-05	121.426	0.000	0.009	0.009

Omnibus:	14.305	**Durbin-Watson:**	1.980
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	16.301
Skew:	0.016	**Prob(JB):**	0.000289
Kurtosis:	3.163	**Cond. No.**	77.7

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Get the value of coefficient of determination

The variation in the independent variable which is explained by the dependent variable is 99.3638 %

Table 16:R^2 and RMSE values of OLS

Ordinary least squares (OLS)		
	R Square	RMSE
Train Set	0.994	0.07948752886451363
Test Set	0.994	0.08013799514078769

Inference: R^2 is almost equal to 1 and RMSE is higher than Linear Regression.Linear regression is better model than OLS.

The graph is linear.

LASSO REGRESSION:

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

The word "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection.

Reason- The main advantage of a LASSO regression model is that it has the ability to set the coefficients for features it does not consider interesting to zero. This means that the model does some automatic feature selection to decide which features should and should not be included on its own.

Mathematical equation of Lasso Regression

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients)

Where,

- λ denotes the amount of shrinkage.
- λ = 0 implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model
- λ = ∞ implies no feature is considered i.e, as λ closes to infinity it eliminates more and more features
- The bias increases with increase in λ
- variance increases with decrease in λ

Lasso Coefficients:

```
Lasso model: [ 0.02898297  0.          0.         -0.         -0.         -0.
  0.          0.70031259 -0.          0.01984192]
```

Observe, many of the coefficients have become 0 indicating drop of those dimensions from the model.

Table 17:R^2 and RMSE values of LASSO

LASSO Regression		
	R Square	RMSE
Train Set	0.9717537590492116	0.17060677652170506
Test Set	0.9713596902266868	0.16948146197751796

Inference: R^2 is almost equal to 1 here too and RMSE is higher than Linear Regression.Linear regression is better model than LASSO Regression.

POLYNOMIAL REGRESSION

Reason-A polynomial regression model is a machine learning model that can capture non-linear relationships between variables by fitting a non-linear regression line, which may not be possible with simple linear regression.

First of all both the test data and train data was transformed to polynomial features.

Shape of train dataset after polynomial transformation:

Before -(6092, 10)
After-(6092, 56)

Shape of test dataset after polynomial transformation:

Before-(14215, 10)
After-(14215, 56)

Table 18:R^2 and RMSE values of Polynomial Regression

Polynomial Regression		
	R Square	RMSE
Train Set	1	5.069148711350037e-14
Test Set	1	5.0380742515126526e-14

Inference: R^2 is equal to 1 and RMSE is lesser than Linear Regression.Polynomial regression is better model than Linear Regression.

RANDOM FOREST REGRESSION

Random forest regression is an invaluable tool in data science. It enables us to make accurate predictions and analyze complex datasets with the help of a powerful machine-learning algorithm.
A Random forest regression model combines multiple decision trees to create a single model. Each tree in the forest builds from a different subset of the data and makes its own independent prediction. The final prediction for input is based on the average or weighted average of all the individual trees' predictions.
Reason- The random forest technique can also handle big data with numerous variables running into thousands. It can automatically balance data sets when a class is more infrequent than other classes in the data. The method also handles variables fast, making it suitable for complicated tasks.

Random forest is imported using scikit learning:
Table 19:R^2 and RMSE values of Random Forest Regression

Random Forest Regression		
	R Square	RMSE
Train Set	0.9866661512352104	0.11507734224295219
Test Set	0.9859365306699769	0.11876256843496513

Inference: R^2 is almost equal to 1 and RMSE is higher than polynomial Regression.Polynomial regression is better model than Random Forest Regression.

BAYESIAN REGRESSION

Bayesian linear regression is a statistical technique that utilizes Bayesian methods to estimate the parameters of a linear regression model. In Bayesian linear regression, we assume that the regression coefficients have a prior probability distribution, which is updated based on the observed data to produce a posterior probability distribution.

Reason:-The primary distinction between Bayesian linear regression and traditional linear regression is that Bayesian linear regression enables the incorporation of prior knowledge or assumptions about the data into the model. This can be especially useful when data is limited or when we want to incorporate expert knowledge into the model.

Table 20:R^2 and RMSE values of Bayesian linear Regression

Bayesian Regression		
	R Square	RMSE
Train Set	0.9969665362118418	0.054888513350202665
Test Set	0.9968873603611399	0.05587245675686155

Inference: R^2 is almost equal to 1 and RMSE is higher than polynomial Regression.Polynomial regression is better model than Bayesian Regression.

Comparing all the RMSE value It is found that Polynomial Regression model is the best one among all the regression model.

MODEL VALIDATION:

Each model is tested with testing dataset and their statistical value of R^2 and RMSE are checked .From the R^2 and RMSE values .

Table 21:R^2 and RMSE values (Consolidated)

Linear Regression		
	R Square	RMSE
Train Set	0.997	0.055
Test Set	0.997	0.055
Ordinary least squares (OLS)		
	R Square	RMSE
Train Set	0.994	0.079
Test Set	0.994	0.080
LASSO Regression		
	R Square	RMSE
Train Set	0.972	0.171
Test Set	0.971	0.169
Polynomial Regression		
	R Square	RMSE
Train Set	1	5.07E-14
Test Set	1	5.04E-14
Random Forest Regression		
	R Square	RMSE
Train Set	0.987	0.115
Test Set	0.986	0.119
Bayesian Regression		
	R Square	RMSE
Train Set	0.997	0.055
Test Set	0.997	0.056

Polynomial regression has the highest R^2 and lowest RMSE among all the models.

INSIGHTS FROM ANALYSIS

The salary is depending on so many factors.More the number of total experience more will be the offered salary.Similarly it also depends on department,organization etc.

Our Evaluation Metric is R^2 and RMSE(Root Mean Squared Error).

As we can see that R^2 of Polynomial Regression is 1 and more than the other model R^2.So Polynomial Regression model is the best model here.

RMSE Value of Polynomial Regression is the least RMSE value among other models ,which makes the Polynomial Regression as the best model again.

RECOMMENDATIONS

This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better results.

More factors like number of company changed in the last 3 years should be added to check the stability of the candidate.

Fixed Salary and Variable Salary should be properly mentioned in the dataset to reduce discrepancies.