2022



DEBSMITA CHAKRABORTY

JULY-‘C’

TABLE OF CONTENTS

**HEADERS:**

1. Introduction to businessProblem ………………………………………………………………..…. Pg-2
2. Data report………………………………………………………………………………………………...Pg-3-4
3. Exploratory Data Analysis……………………………………………………………………….......…Pg-5-11
4. Business Insights………………………………………………………………………………………… Pg-11-12

**PLOTS and GRAPHS:**

1. Data Dictionary and Heading (Fig1) …………………………………………………………….………………. Pg-2
2. Data Details (Fig2) …………………………………………………………………………………………………...Pg-3-4
3. Exploratory Data Analysis (Fig3)…………………………………………………………………………………..Pg-5-11

1) Introduction of the business problem

**Business Problem:**

a) 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.

b) Need of the study/project-The idea is to minimise human intervention to calculate salaries of an employee who is joining the company but automate it with data of the past with reference to his interview. To keep all employees with similar profile under one umbrella pricing.

c) Understanding business/social opportunity- Helping business reduce manual labour and minimize human judgement which at times may be less practical.

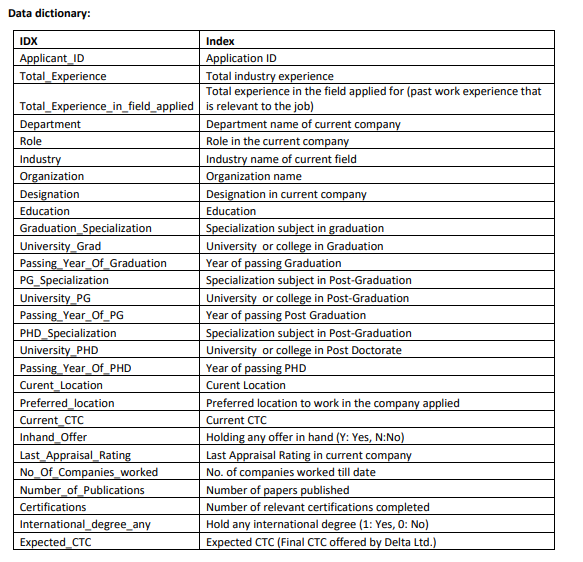


Fig 1.1- Data Dictionary

* The data looks like:

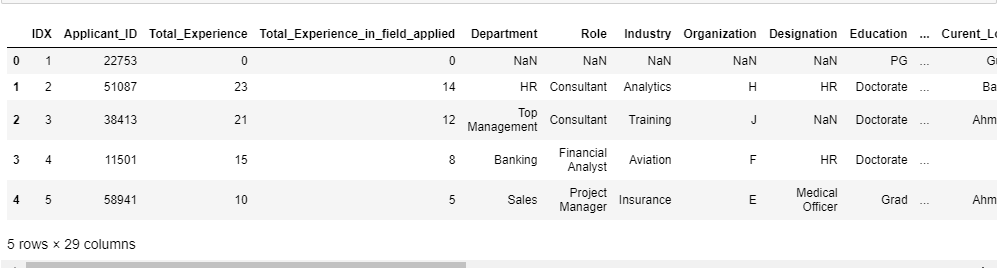


Fig 1.2- Data Heading

2)Data Report

* The column headings being:

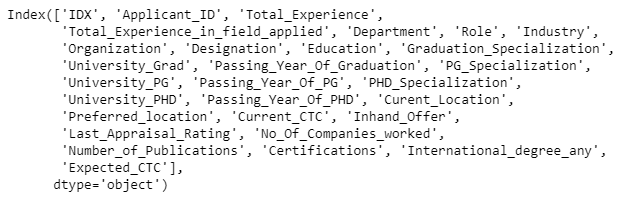


Fig 2.1- Data Column Headings

* The data has 29 columns and 2500 rows. We see below a lot of missing values least being University PHD, PHD specialization and passing year of PHD which is almost 10000 entries missing:

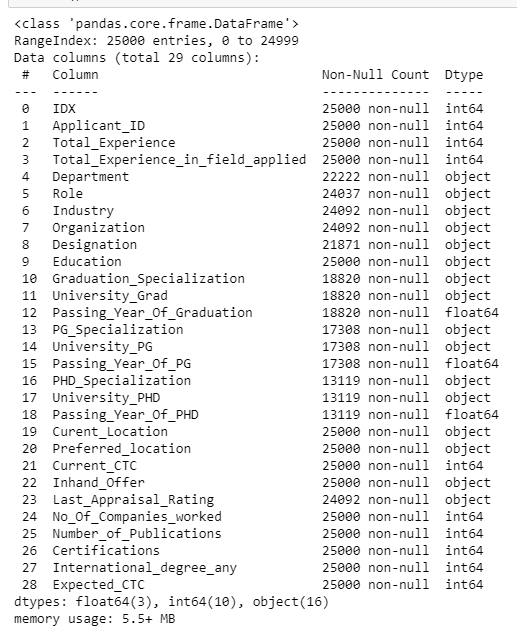


Fig 2.2 – Data Information

* The data shows descriptive analysis wherein we do see a lot of outliers as the difference in mean and maximum is high. A lot of fluctuations in the values can be seen:

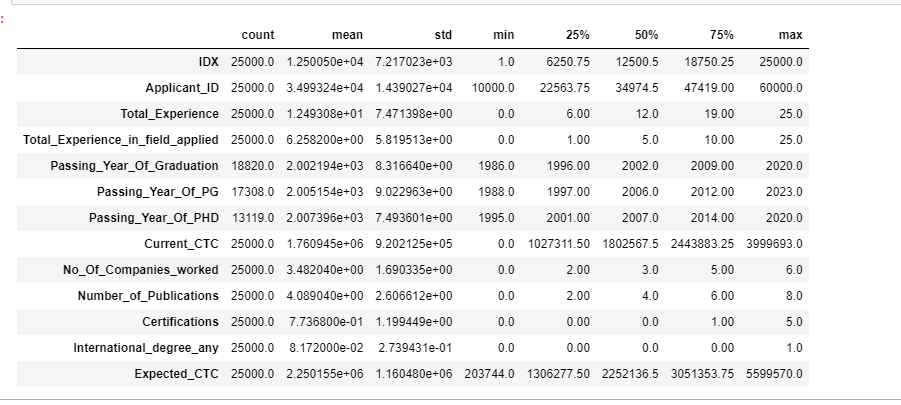


Fig 2.3 Data Described

* Data definitely from its describe function seemed to have outliers as missing values are high but number of duplicate rows is zero.

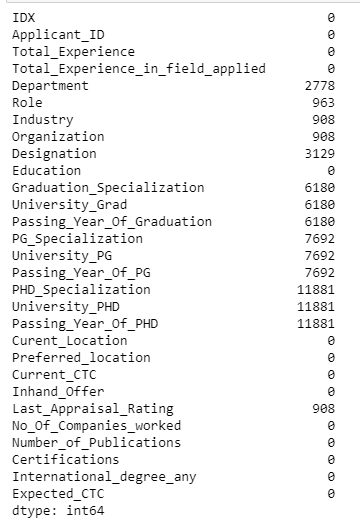
 

Fig 2.4 Null Values and Duplicate rows

* Getting the summary statistics of object data type:

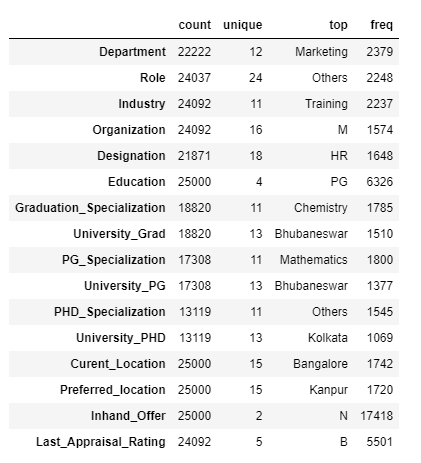
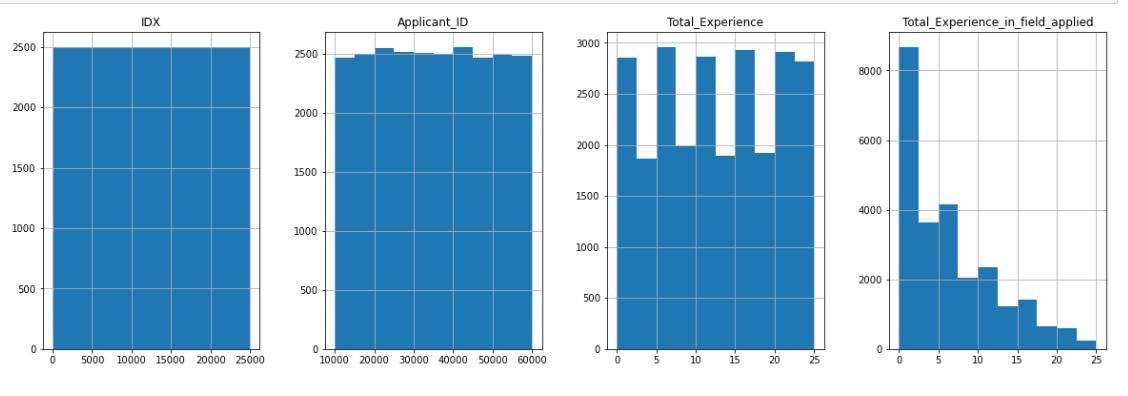


Fig 2.4 Data Summary

3) Exploratory data analysis

**Our target variable is Expected CTC since the entire modelling is about predicting the same.**

**UNIVARIATE ANALYSIS:**

****

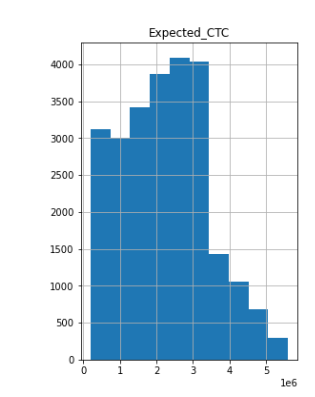
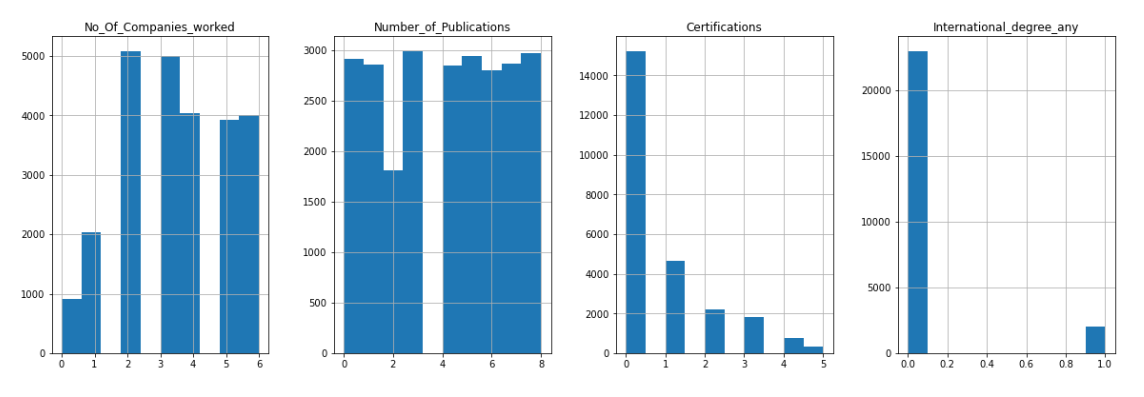
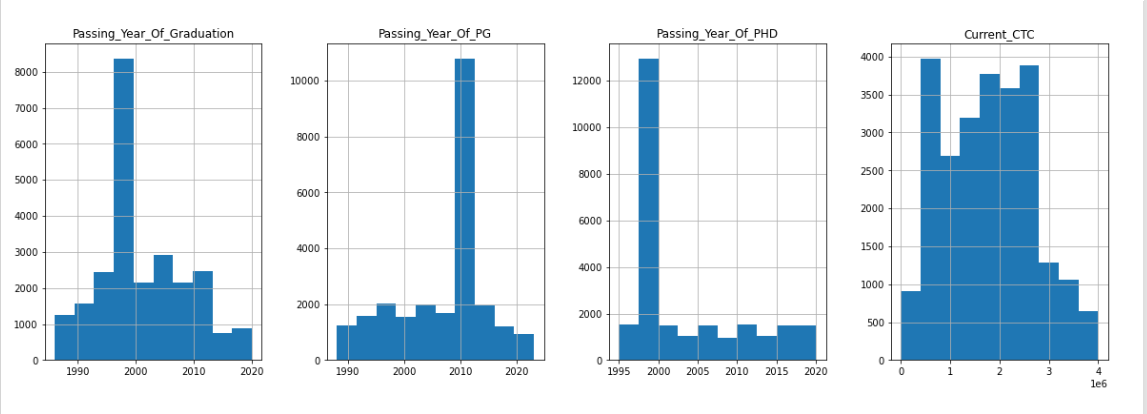
****

Fig 3.1 Histogram

INFERENCE:

* We see a lot of pass outs in the year 2000 and the least in 2014-2015.
* Passing year of PG is the maximum in 2010 and least post 2020.
* Total years of experience is maximum 10, 5 and 25. A uniform distribution of data.
* The total experience in the field applied is 0.
* More people have worked in 2 to 3 companies amongst applicates with 0 certifications Internationally.
* We see Expected CTC at a uniform distribution.

**CATEGORICAL DATA ANALYSIS:**

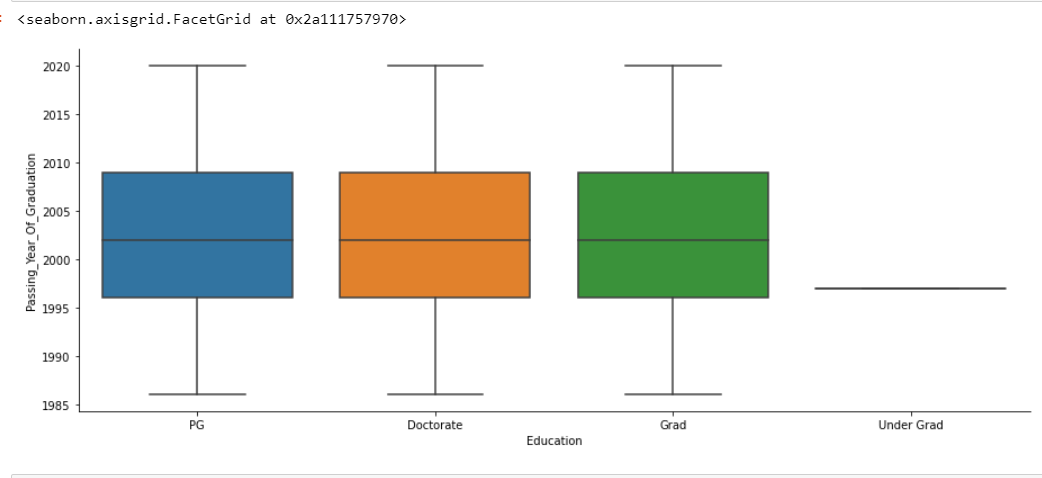
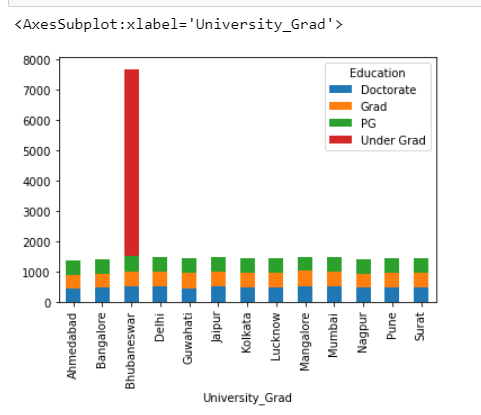
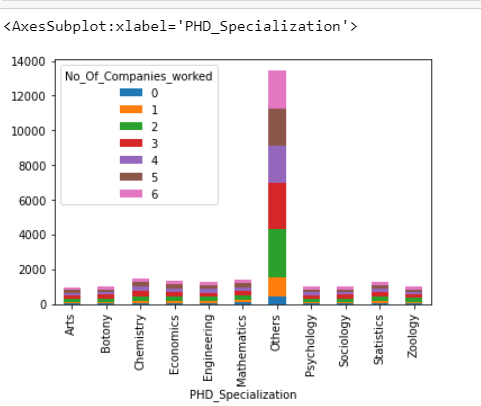
****

Fig 3.2 Bivariate Analysis

**Inference:**

* **From the categorical representation it is applicants who have 3 years of experience with other qualifications.**
* **Maximum applicants are undergraduates.**
* **The Education column with the year of passing has no outliers as such.**

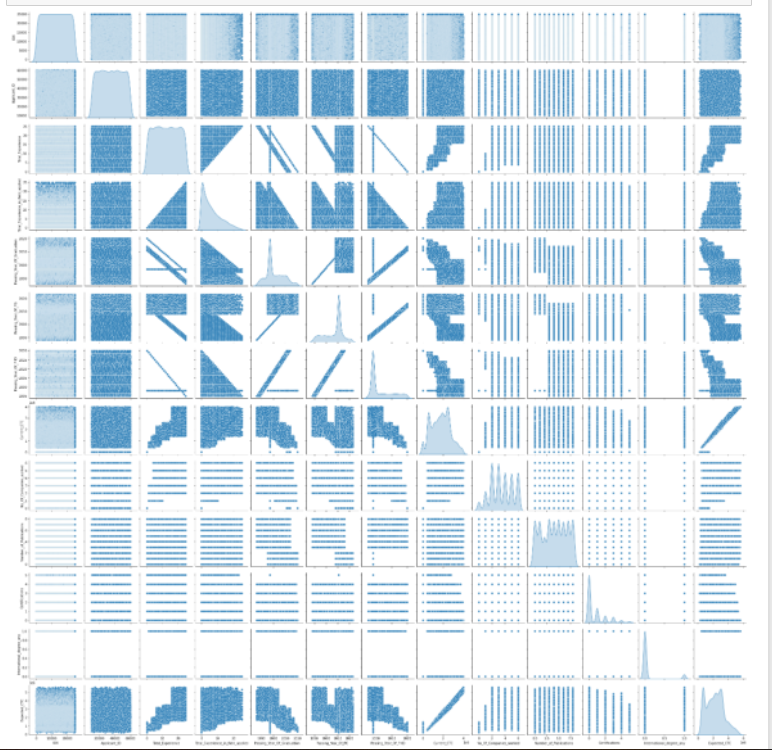
**MULTIVARIATE ANALYSIS** 

Fig 3.3 Pair Plot

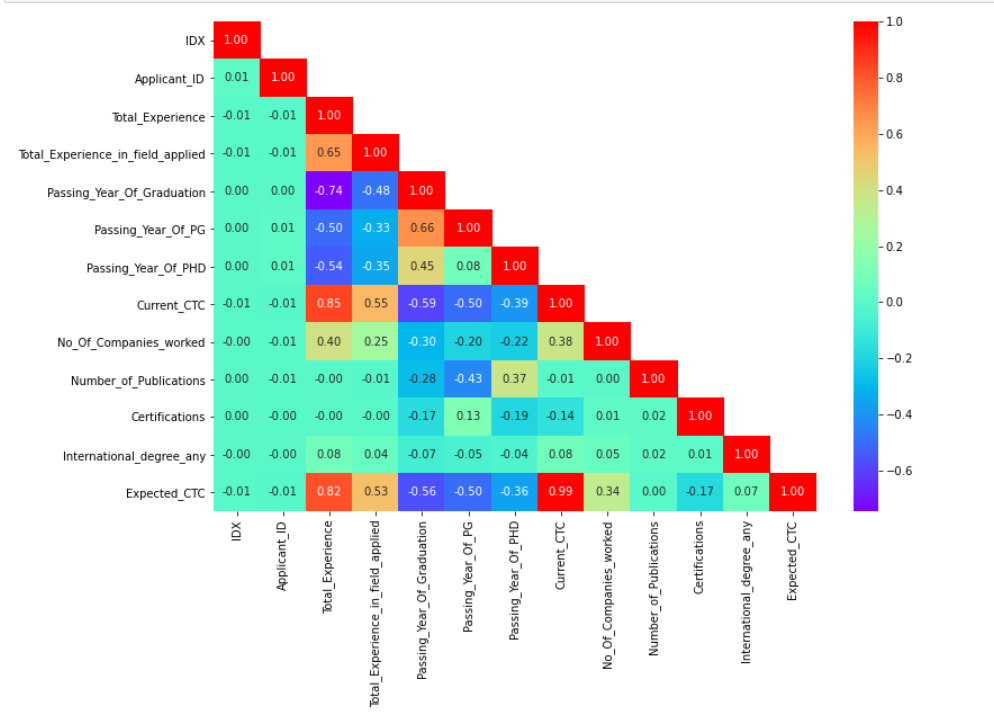
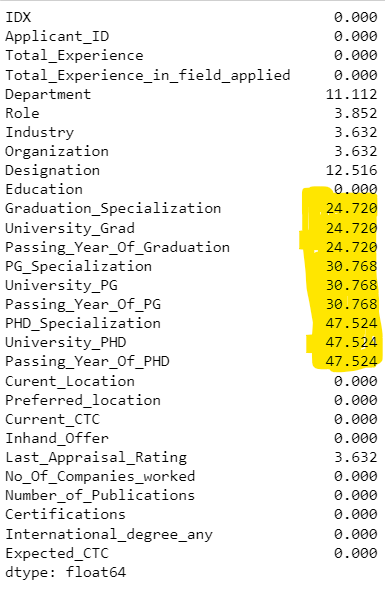


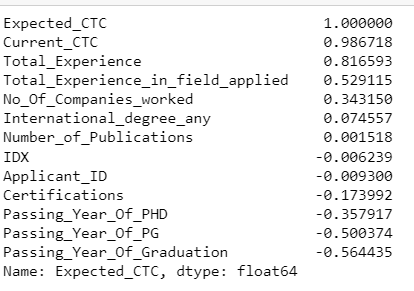
Fig 3.4 Correlation Matrix

**INFERENCE:**

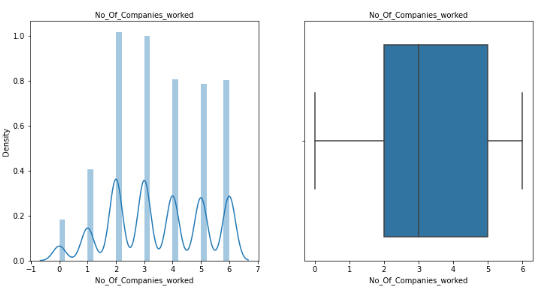
* **The data is not following normal distribution indicating missing values in the pair plot. The number of missing data contribution of each field is as:**

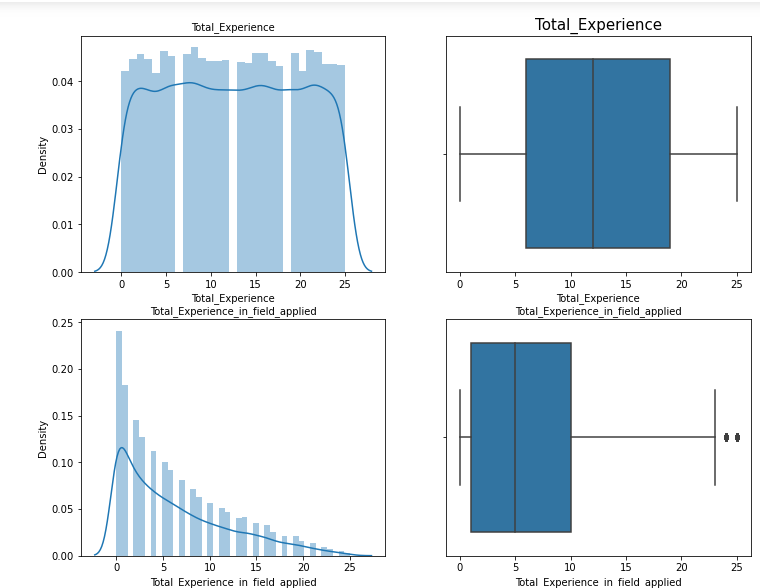
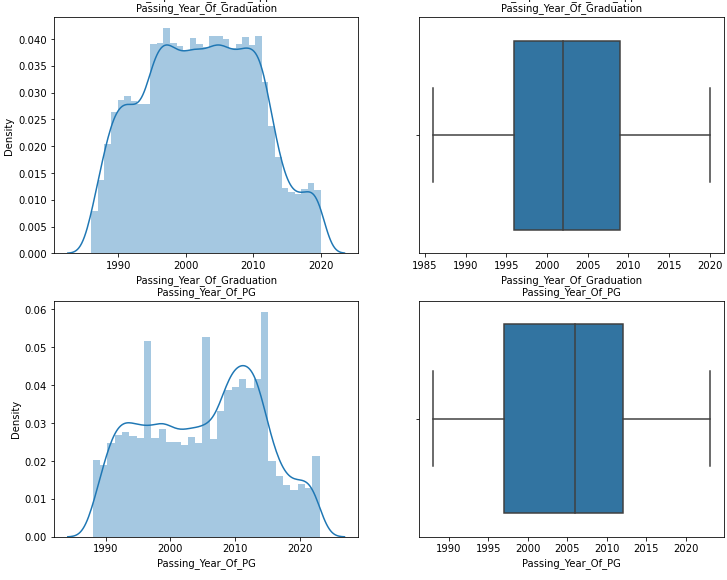


* **The data having more than 10%-15% missing values we will impute them later.**
* **Correlation between -1 to +1 are the best which is valid for fields – Expected CTC to Current CTC, Total experience to current CTC and expected CTC to Total experience.**
* **The correlations of all fields to Expected CTC is as in descending to ascending order with respect to Expected CTC our target variable:**

****

**UNIVARIATE ANALYSIS**



****

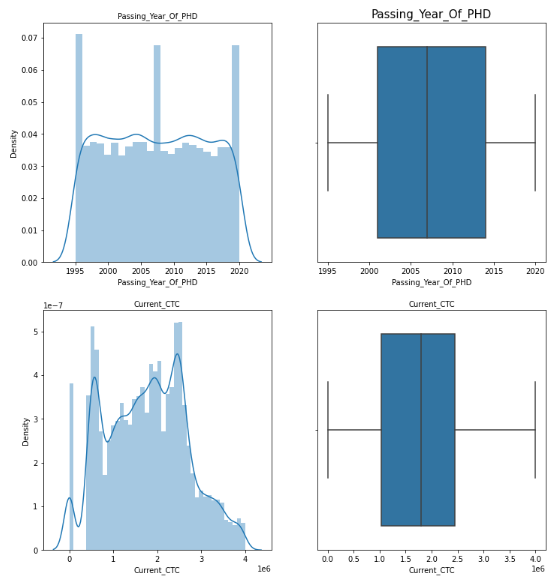
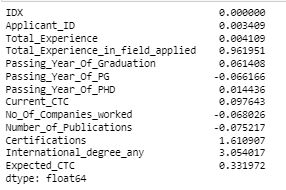


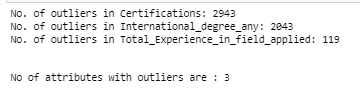
Fig 3.5 Univariate Analysis

INFERENCE:

* Data seems to be normally distributed with the skewness in the fields as below:



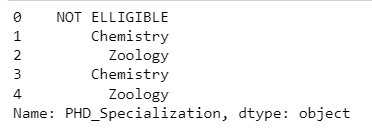
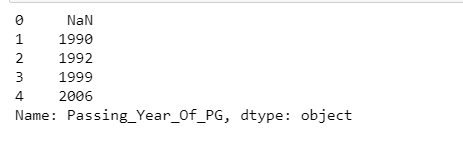
* Data has outliers in the fields:



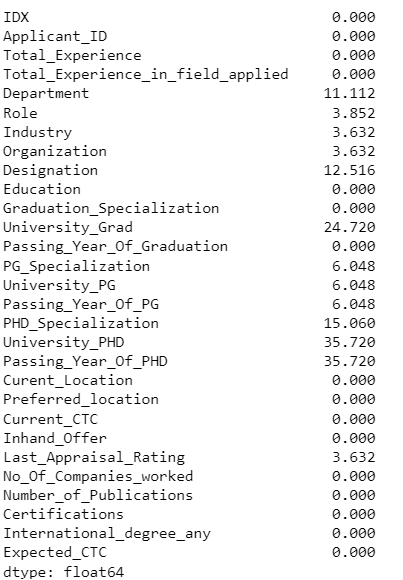
* IDX column adds no benefit hence dropping that:



* People who are not a Graduate or are in Post Graduate wont do PHD specialization hence making a special feature in the column as “NOT ELLIGIBLE”. Applicants who are undergrads will not have Graudation Year, Passing year of Graduation, PG Specialization, University\_PG, Passing Year of PG or PHD Specialization, University PHD or passing year of PHD.

* We can see the contribution of missing values prior to applying ‘not eligible and dropping further NA values’ is :



* Treating outliers wont make sense since data variations is not much and has been imputed with mode where outliers were present.

4. Business insights from EDA

* The maximum CTC currently and Expected is of Marketting and the least is for NA. This NA columns will be dropped.
* For PHD\_specialization we again see Post Graduates not eligible for PHD and null columns have maximum salaries. We have listed the top six.

1)anyone with Post Graduate is not Elligible for PHD

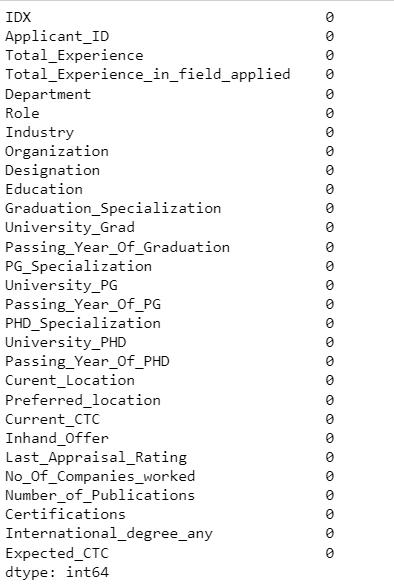
2)anyone doing under graduate is not applicable for PHD

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Current\_CTC** | **Sum of Expected\_CTC** |
| NOT APPLICABLE | 8586164272 | 10615927964 |
| (blank) | 7124359270 | 9114879409 |
| NOT ELLIGIBLE | 3631319721 | 4626091219 |
| Others | 3011632385 | 3842847540 |
| Chemistry | 2751078517 | 3559270228 |
| Arts | 1785389163 | 2302221807 |
| **Grand Total** | **44023634597** | **56253862760** |

* Location wise we can see Ahmedabad the top most paid and Surat the least:

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Current\_CTC** | **Sum of Expected\_CTC** |
| Ahmedabad | 2904173639 | 3706384890 |
| Bangalore | 3027091674 | 3859247837 |
| Pune | 2837067122 | 3628982145 |
| Surat | 2787404313 | 3573136269 |

* We have applied mode and median since mean cannot be applied with outliers and then dropped the remaining null values:



* The data is now a balanced data currently no clustering is needed.