

Ml Capstone project report

BY: ROHAN BANERJEE



August 13, 2025

**Machine Learning Capstone Project – Final Report**

**1. Problem Statement**

The goal of this project is to predict the target variable from the given dataset using two models:

* **Basic Algorithm**: Linear Regression
* **Advanced Algorithm**: Random Forest Regressor

We will compare their performances and ensure that there is no overfitting.

**2. Data Preprocessing**

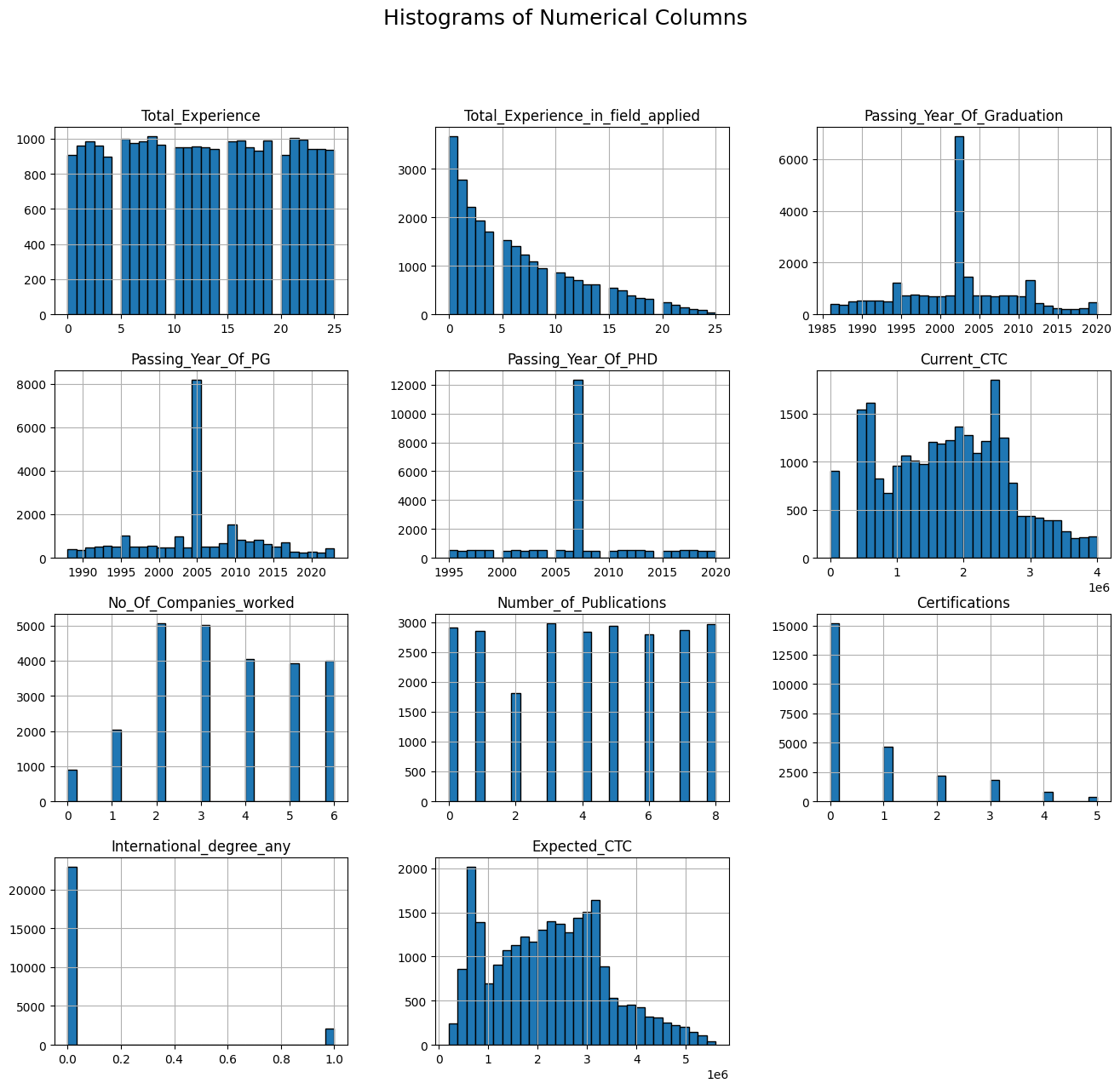
**Steps performed:**

1. **Missing Values Handling**
   * For features with ≤50% missing → imputed with mean/median/mode.
   * No features with >50% missing, so none were dropped.
2. **Duplicate Removal**
   * Dropped duplicate rows.
3. **Outlier Removal**
   * Used **iterative outlier removal** based on IQR until no outliers remained.
4. **Encoding**
   * Applied **One-Hot Encoding** for categorical variables → Shape increased from (5520, 26) to (5520, 118).
5. **Scaling & Normalization**
   * Applied **StandardScaler** to standardize numerical features (mean=0, std=1).

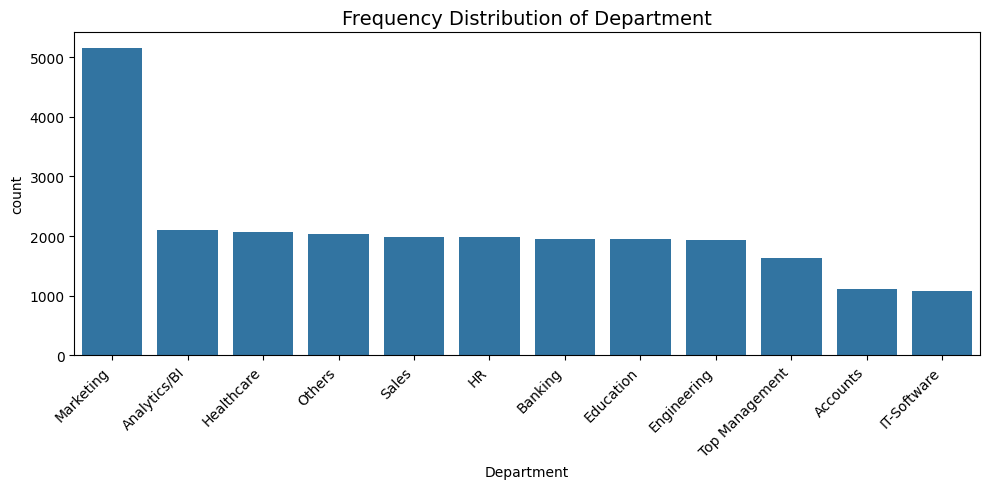
**3. Exploratory Data Analysis (EDA)**

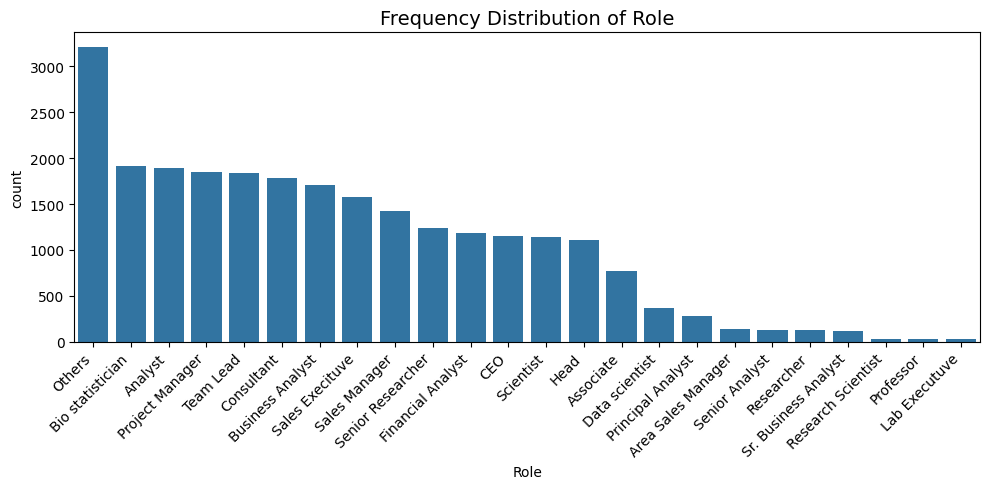
Generated and analyzed:

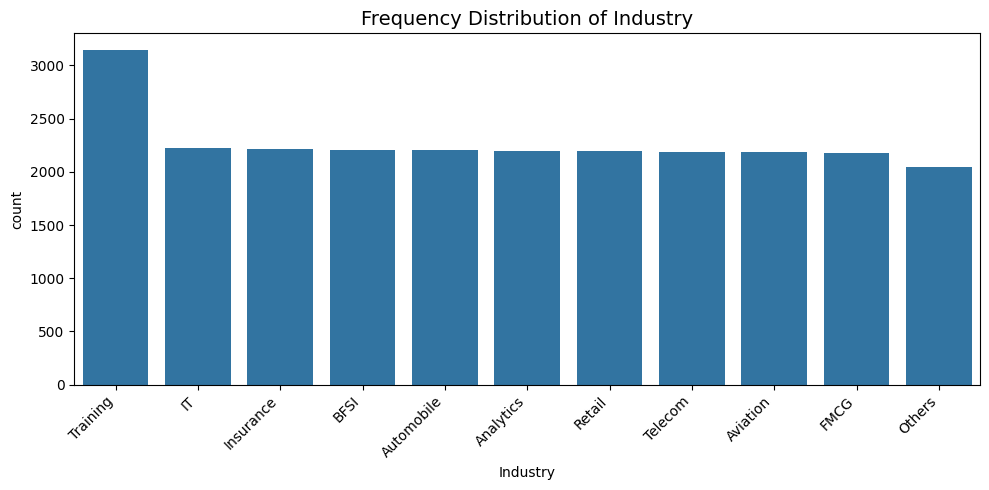
* **Histograms** for each numerical column

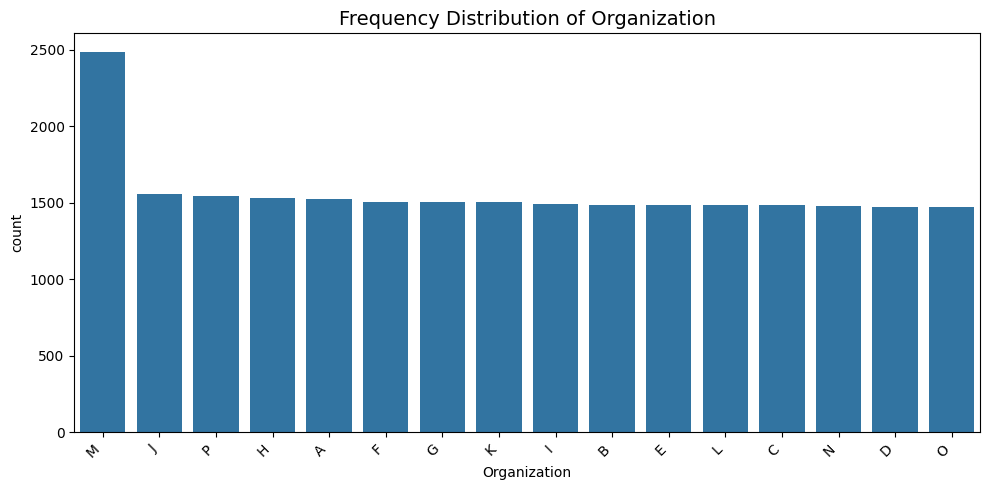


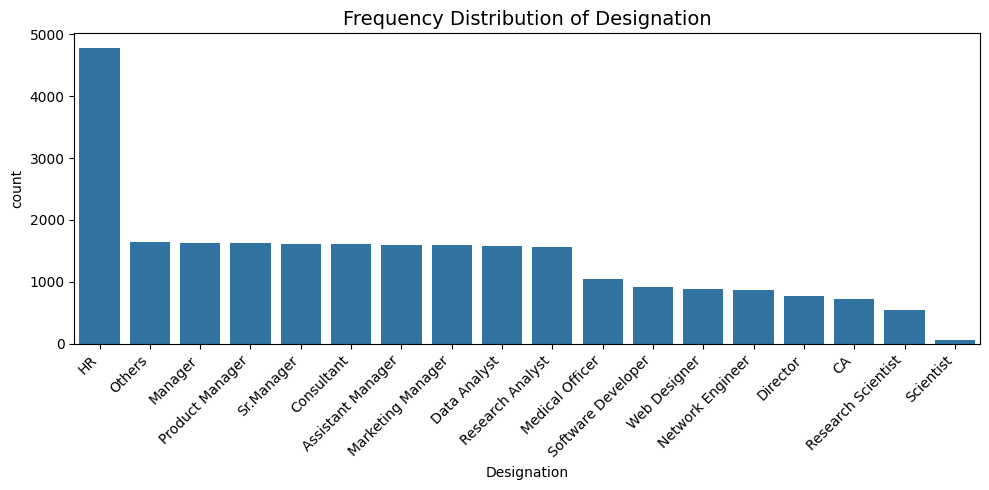
* **Frequency distribution** for categorical variables

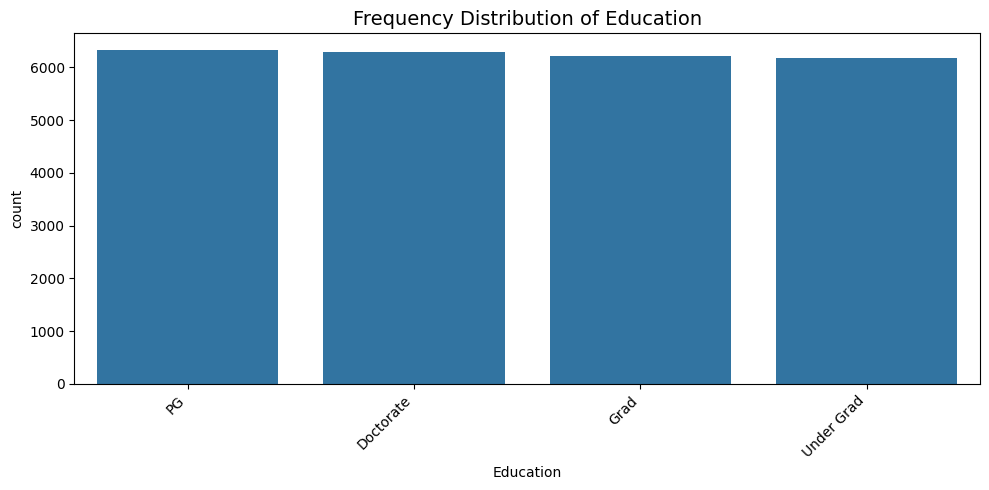


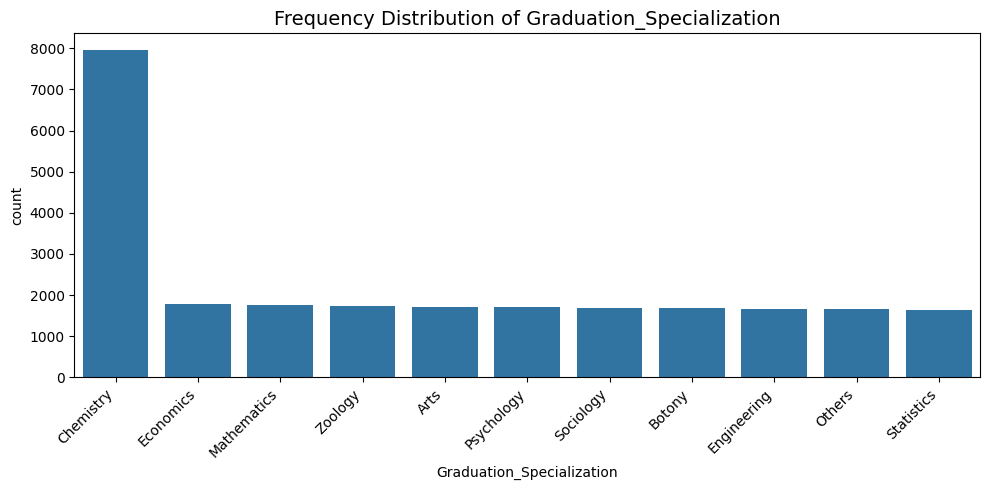


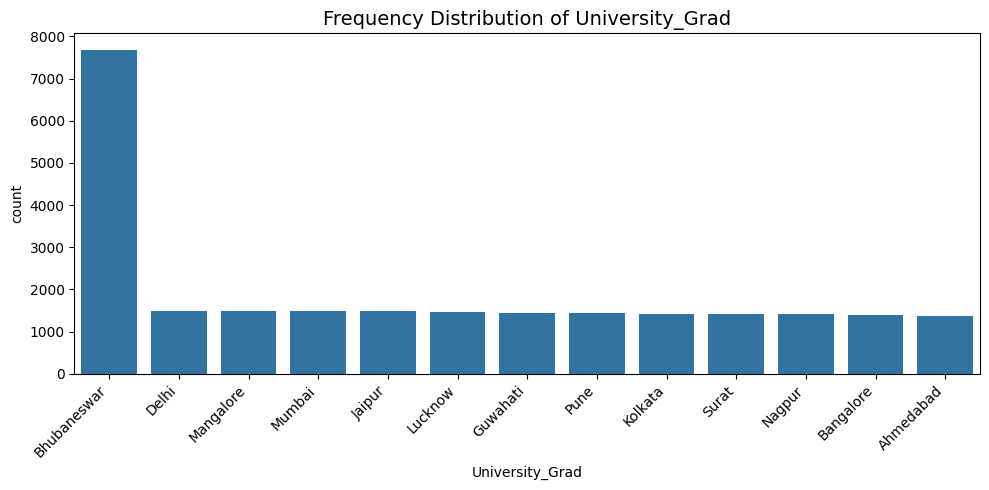


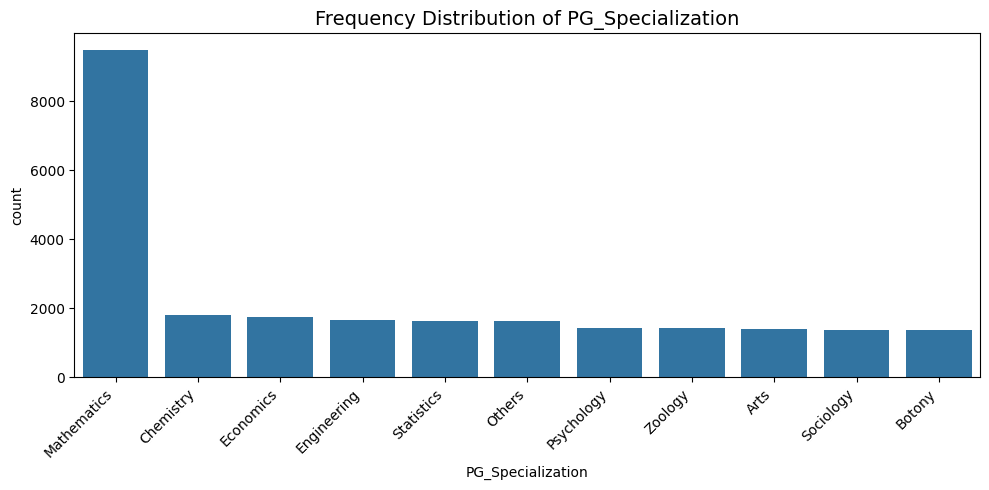


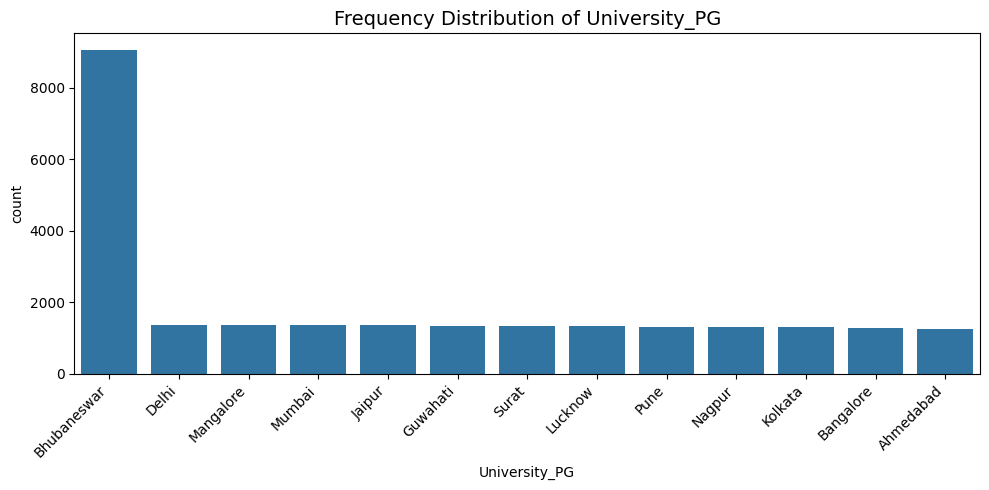


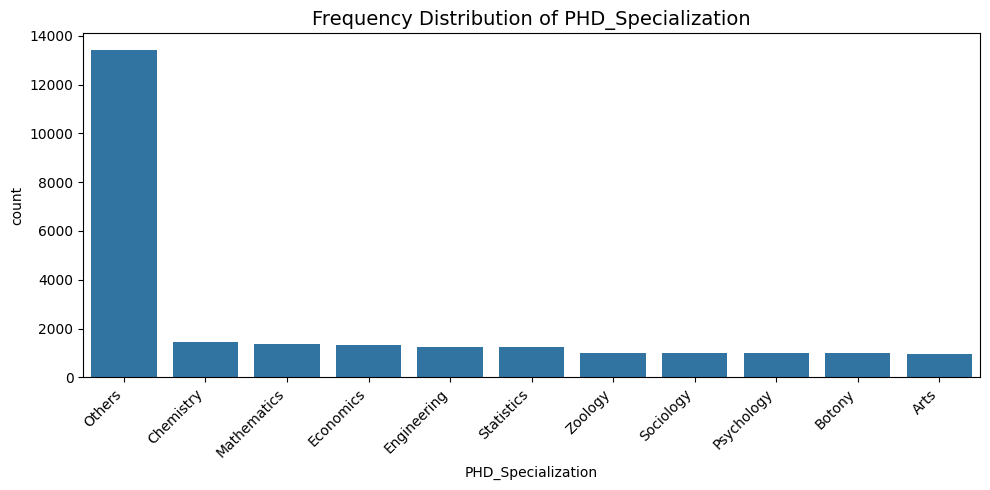


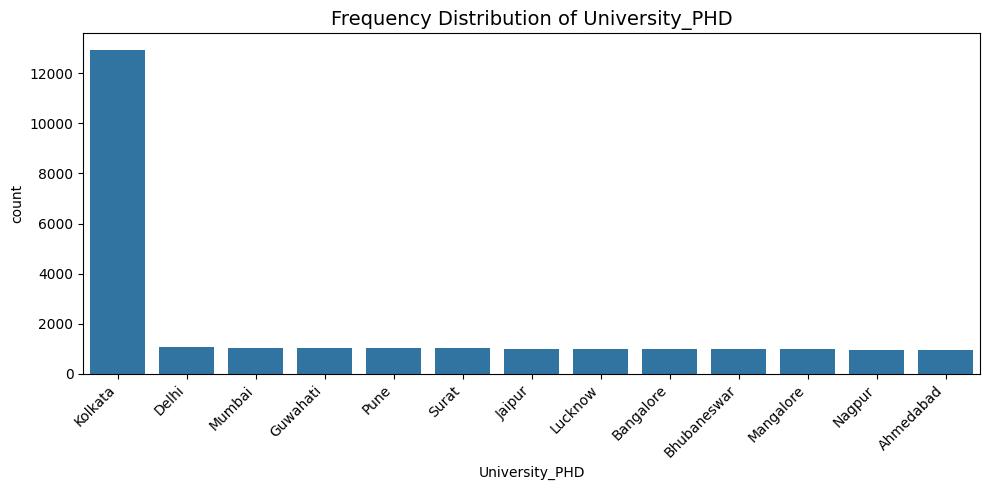


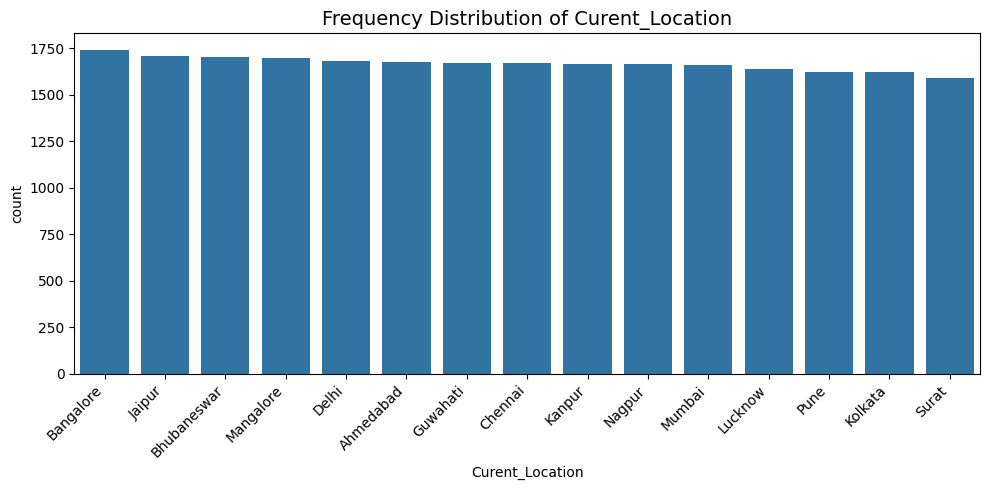


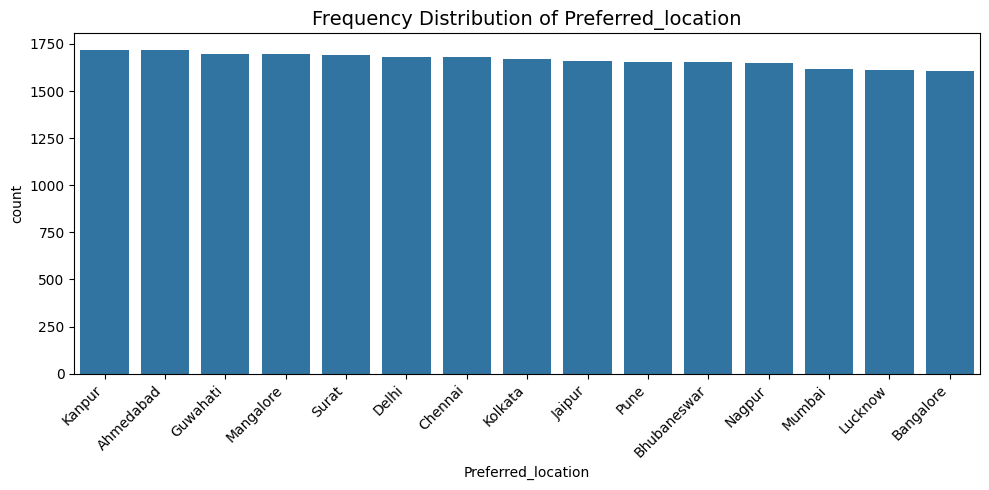


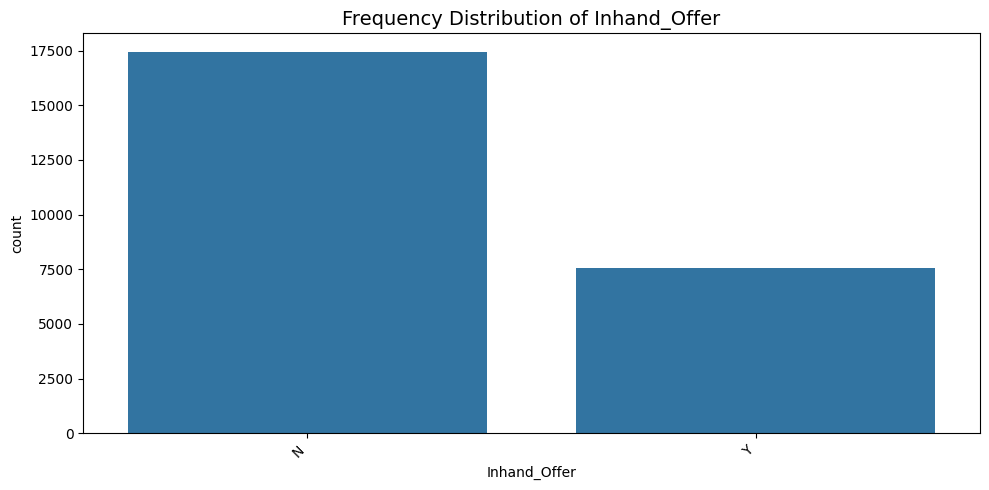


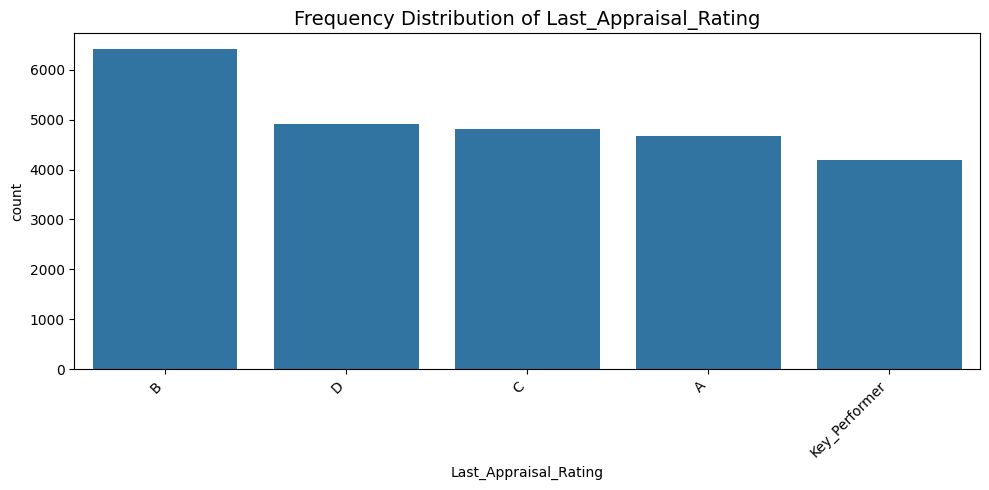


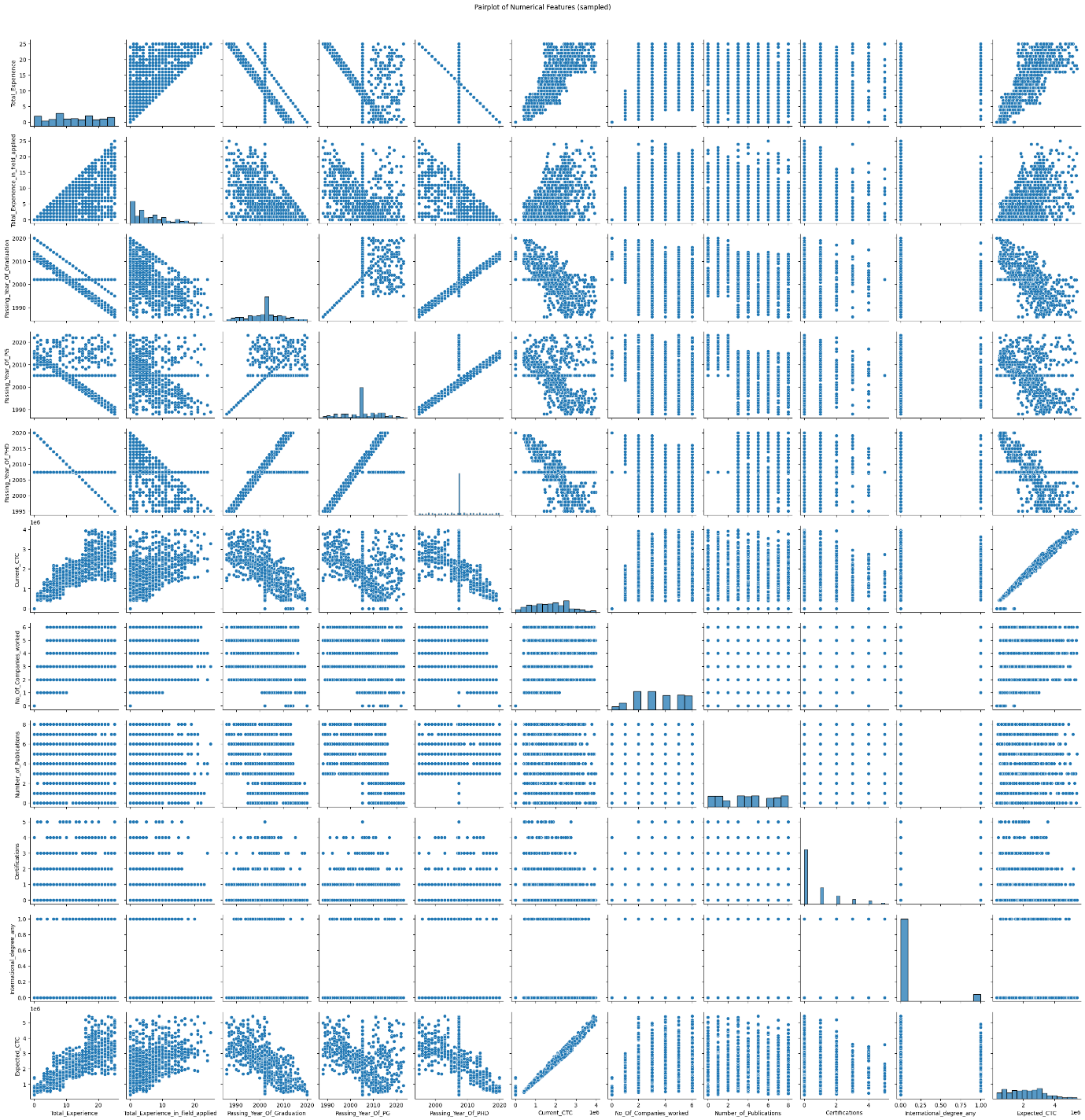
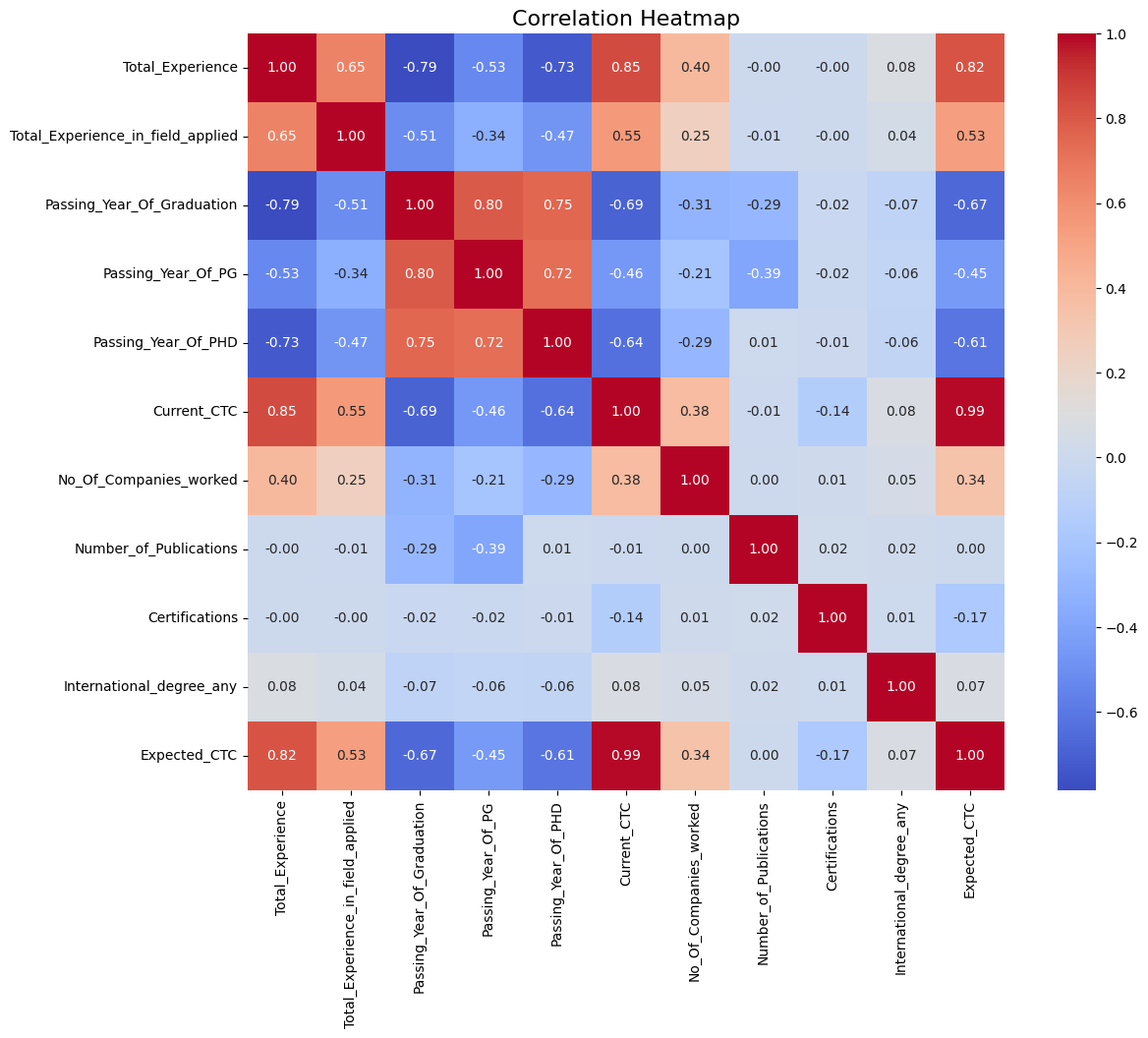
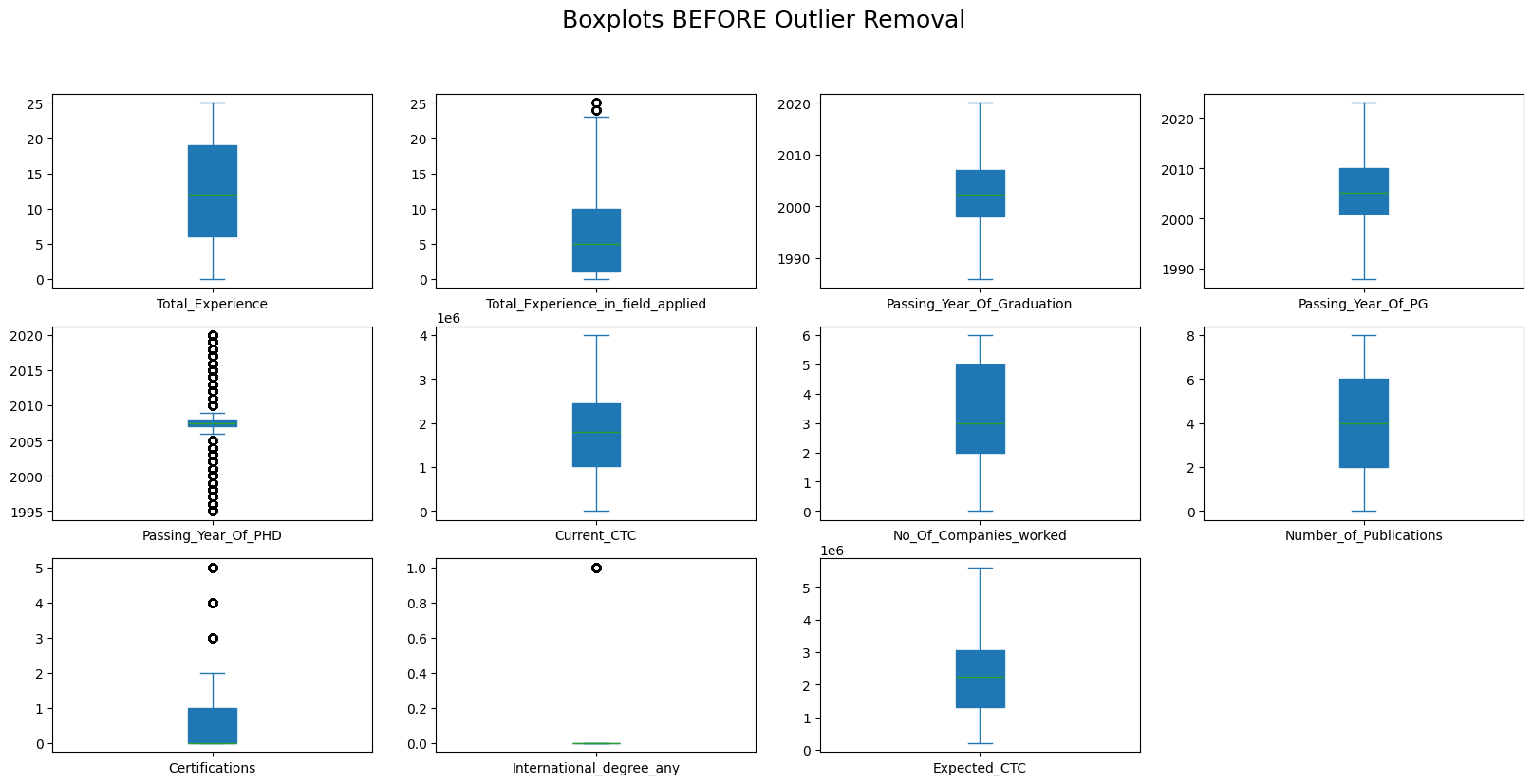


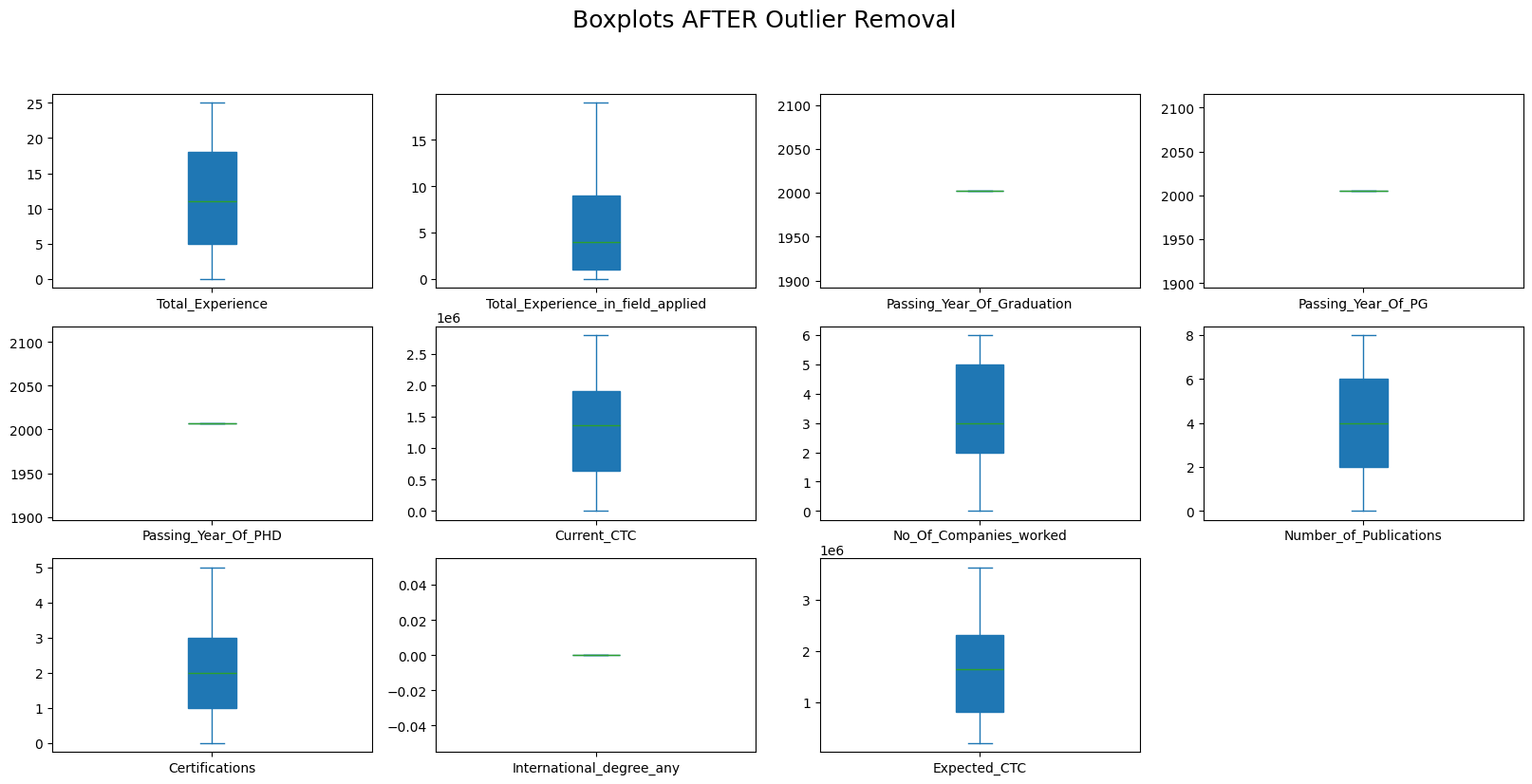








* **Pairplots**
* **Heatmap** (correlation)
* **Boxplots** for outlier inspection (before & after removal)



**4. Model Training**

**Train/Test Split** → 80/20 split.

* **Linear Regression**
* **Random Forest Regressor** (n\_estimators=100, random\_state=42)

**5. Model Evaluation**

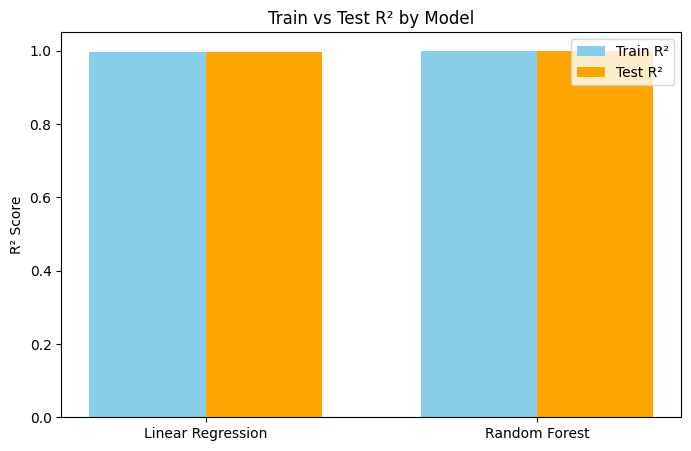
| **Metric** | **Linear Regression** | **Random Forest** |
| --- | --- | --- |
| Train R² | 0.9965 | 1.0000 |
| Test R² | 0.9963 | 0.9998 |
| Train RMSE | 52,798.98 | 5298.42 |
| Test RMSE | 54,221.69 | 13,466.51 |

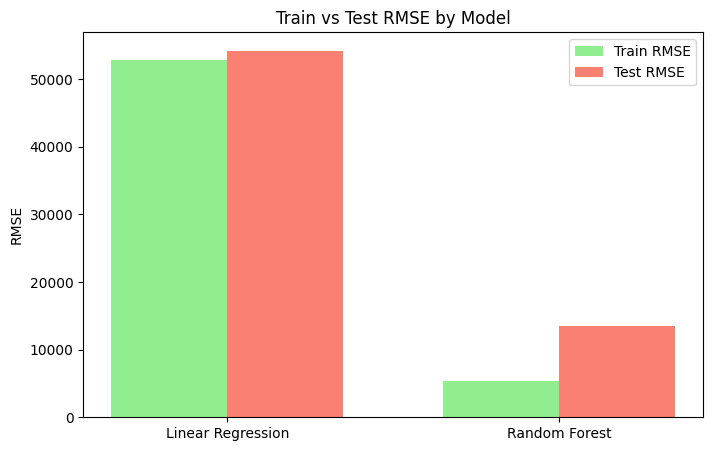
**6. Interpretation**

* **Random Forest significantly outperformed Linear Regression** in RMSE while both models had very high R².
* Train and Test scores were almost identical → **No overfitting detected**.
* Feature importance from Random Forest revealed the most impactful predictors.

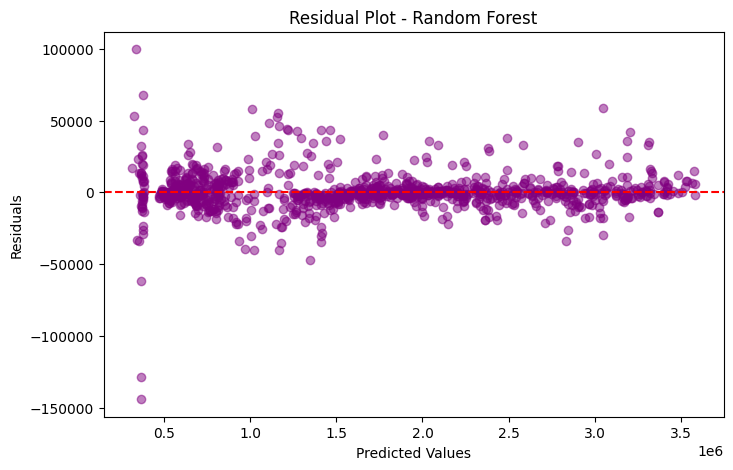
**7. Final Visualizations**

* **Bar charts**: Train vs Test R² and RMSE.

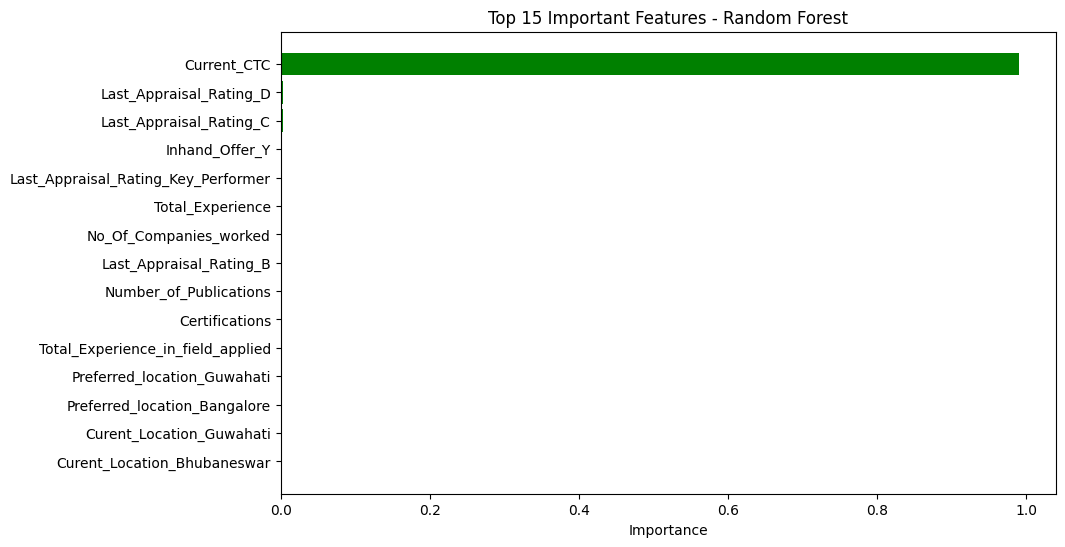




* **Residual plot** for Random Forest → No major patterns → good fit.



* **Feature importance chart** from Random Forest.



**8. Deliverables**

* **Best model saved**: random\_forest\_model.pkl
* **Cleaned dataset saved**: cleaned\_preprocessed\_data.csv
* **Performance comparison table saved**: model\_performance\_comparison.csv

**9. Conclusion**

* The dataset was clean, well-structured, and had strong predictive power.
* Random Forest captured complex, non-linear patterns that Linear Regression could not.
* Final recommendation: **Use Random Forest** for deployment.