Project 5

- Machine Learning -

(AI-Powered Predictive Modeling for Workforce Visa Approvals)

1.Introduction

1.1.Problem Definition

Introduction:

Business communities in the United States are facing a growing demand for skilled labor, making the identification and attraction of the right talent a constant challenge. Companies across the U.S. require hard-working, talented, and qualified individuals, not only locally but also from abroad, to maintain their competitive edge in the global economy. This demand has led to a greater reliance on foreign workers to fill critical gaps in the workforce.

The Immigration and Nationality Act (INA) of the U.S. allows foreign workers to come to the United States to work, either on a temporary or permanent basis. This act also provides protections for U.S. workers by ensuring that the wages and working conditions of American workers are not adversely affected by the hiring of foreign workers. It mandates that U.S. employers comply with specific statutory requirements when hiring foreign workers to fill workforce shortages, ensuring that wages for foreign workers are competitive and comparable to those of domestic workers.

These immigration programs are managed by the Office of Foreign Labor Certification (OFLC), which processes job certification applications from employers who wish to hire foreign workers. The OFLC grants certifications only if the employers can demonstrate that there is an insufficient pool of U.S. workers available to perform the work, and that the wages for the foreign worker meet or exceed the wages paid for similar occupations in the region of intended employment.

The OFLC's responsibility of processing a growing number of applications each year poses a significant challenge, with the increasing volume of visa applications making manual review a tedious and time-consuming process. As the number of applications continues to rise, the need for data-driven solutions to facilitate the processing of these applications becomes ever more critical.

Business Problem

The business problem arises from the increasing volume of visa applications processed by the Office of Foreign Labor Certification (OFLC), which has made the manual review process inefficient and time-consuming. With the growing number of applicants, the OFLC struggles to manage and assess each case effectively. This leads to delays and potential errors in decision-making, impacting the overall efficiency of the visa certification process. A machine learning solution is needed to automate and streamline the process, enabling OFLC to focus resources on the most critical cases while improving the speed and accuracy of visa approvals.

Objective

The objective of this project is to develop a machine learning model that assists the Office of Foreign Labor Certification (OFLC) in predicting visa application outcomes. With the increasing volume of applications, the goal is to streamline the process of visa approvals by building a classification model that can accurately determine which applicants are likely to have their visas certified or denied. The model will analyze various factors such as the applicant's education, job experience, prevailing wages, and other relevant features to identify the key drivers influencing the approval process. By automating this aspect of the visa certification, the project aims to improve the efficiency and accuracy of decision-making, reduce the workload on OFLC, and ultimately facilitate faster processing of visa applications.

1.2.Data Background and Contents

1.2.1.Dataset Overview

The dataset consists of 25,480 entries with 12 columns, capturing various attributes of visa applications processed by the Office of Foreign Labor Certification (OFLC). The columns include a mix of categorical and numerical features:

Numerical Columns:

1. no_of_employees: The number of employees in the employer's company.
2. yr_of_estab: The year the employer's company was established.
3. prevailing_wage: The average wage paid to similarly employed workers in the area of employment.

Categorical Columns:

1. case_id: A unique identifier for each visa application (this column is dropped as it has unique values that do not contribute to analysis).
2. continent: The continent of the applicant (e.g., Asia, Africa).
3. education_of_employee: The education level of the applicant (e.g., High School, Bachelor's, Master's).
4. has_job_experience: Whether the applicant has job experience (Y = Yes, N = No).
5. requires_job_training: Whether the applicant requires job training (Y = Yes, N = No).
6. region_of_employment: The U.S. region where the applicant intends to work (e.g., West, Northeast).
7. unit_of_wage: The unit of the prevailing wage (e.g., Hourly, Yearly).
8. full_time_position: Whether the position is full-time (Y = Yes, N = No).
9. case_status: The target variable that indicates whether the visa application was certified or denied.

This dataset is well-structured, with no missing values, and contains a combination of categorical and numerical data. It offers valuable insights into the factors that influence visa approval decisions and will serve as the basis for predictive modeling to streamline the visa certification process.

1.2.2.Statistical Summary

The dataset consists of 25,480 entries, with a mix of categorical and numerical columns. Below is the statistical summary, which provides insights into the distribution of numerical columns and the frequency of categorical values:

Numerical Columns:

- no_of_employees: The number of employees in the employer's company ranges from -26 (which may indicate an outlier or erroneous value) to 602,069, with a mean of 5,667 employees. The standard deviation is quite large (22,877), indicating significant variability in company sizes.
- yr_of_estab: The year the employer's company was established ranges from 1800 to 2016, with a mean year of 1979.4. The standard deviation of 42 years suggests a broad range of company establishment years.
- prevailing_wage: The prevailing wage varies from 2.14 to 319,210.27, with a mean value of 74,455.81. The high standard deviation (52,815.94) reflects significant variability in wage values.

Categorical Columns:

- continent: The most frequent continent in the dataset is Asia (16,861 occurrences), followed by Africa and Europe.
- education_of_employee: The most common education level is Bachelor's (10,234 occurrences), followed by High School and Master's.
- has_job_experience: The majority of applicants (14,802) have job experience (Y = Yes).
- requires_job_training: Most applicants do not require job training (22,525 N = No).
- region_of_employment: The Northeast region is the most common (7,195 occurrences), followed by West, South, and Midwest.
- unit_of_wage: The majority of applicants are paid yearly (22,962 occurrences), followed by hourly, monthly, and weekly wages.

- full_time_position: The majority of positions are full-time (22,773 occurrences).

Target Variable (case_status):

- case_status: This is the target variable, indicating whether the visa application was Certified or Denied. The dataset is imbalanced, with Certified being the most frequent outcome (17,018 occurrences), while Denied applications occur 8,462 times. This disparity is important for modeling, as the classification model will need to account for the imbalance to make accurate predictions.

This statistical summary reveals significant variability in some columns (e.g., number of employees, wages) and provides a clear view of the distribution of applicants across different characteristics. Understanding the target variable case_status and the imbalance in its distribution will be crucial in model selection and evaluation.

1.3.Univariate Analysis :

Categorical Columns:

1. Visa Applicants Distribution Across Continents:

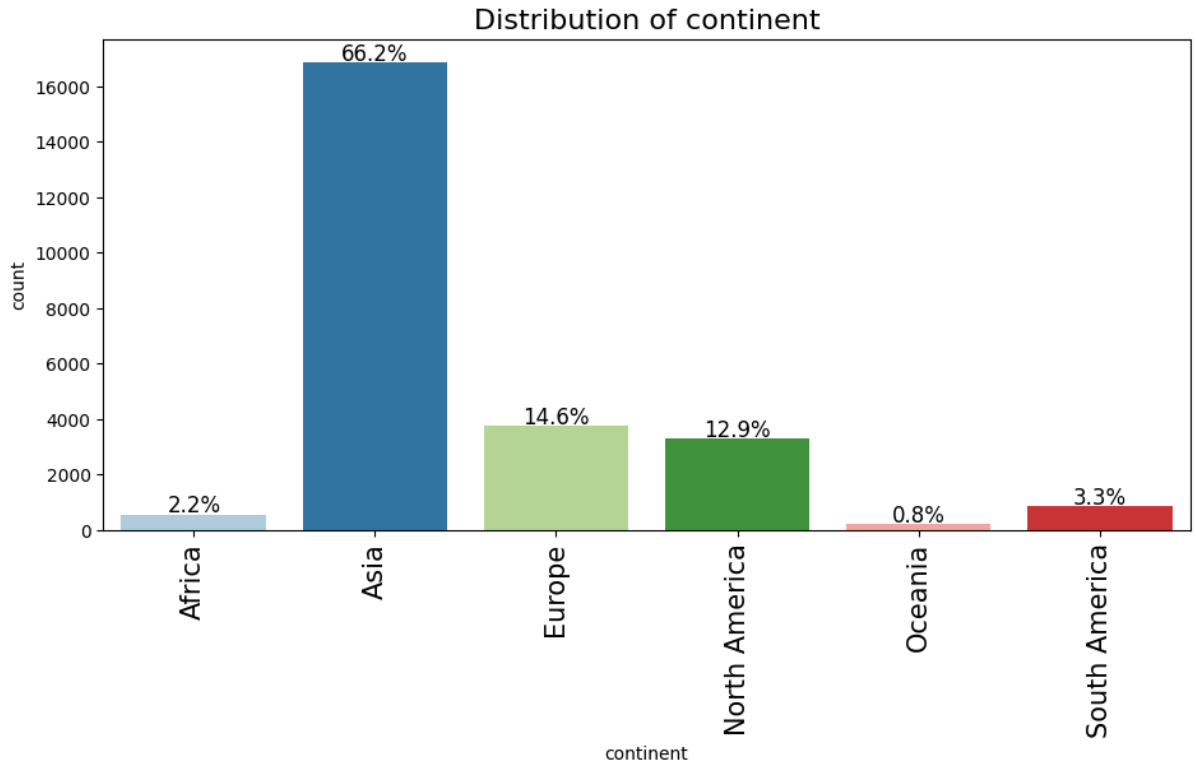

Figure 1. Visa Applicants Distribution Across Continents

Interpretation:

The distribution of visa applicants shows that Asia dominates with 66.2% of the total applicants, followed by Europe with 14.6% and North America with 12.9%. South America accounts for 3.3%, while Africa and Oceania have minimal representation at 2.2% and 0.8%, respectively. This indicates a significant concentration of applicants from Asia, with other continents contributing much smaller proportions.

2. Distribution of Employee Education Levels:

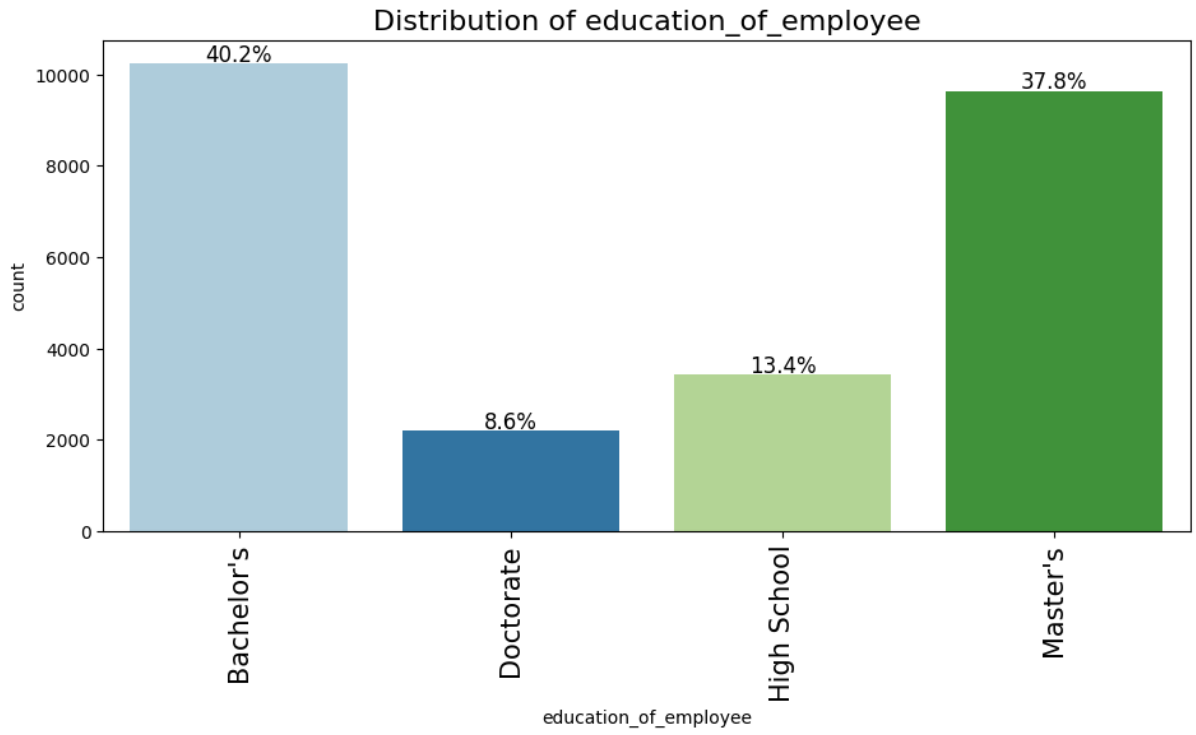

Figure 2. Distribution of Employee Education Levels

Interpretation:

The distribution of education levels among employees shows that the majority hold a Bachelor's degree (40.2%), followed closely by those with a Master's degree (37.8%). A smaller proportion of employees have completed High School (13.4%), while only 8.6% possess a Doctorate. This highlights a strong preference for applicants with higher educational qualifications.

3. Distribution of Job Experience Among Applicants:

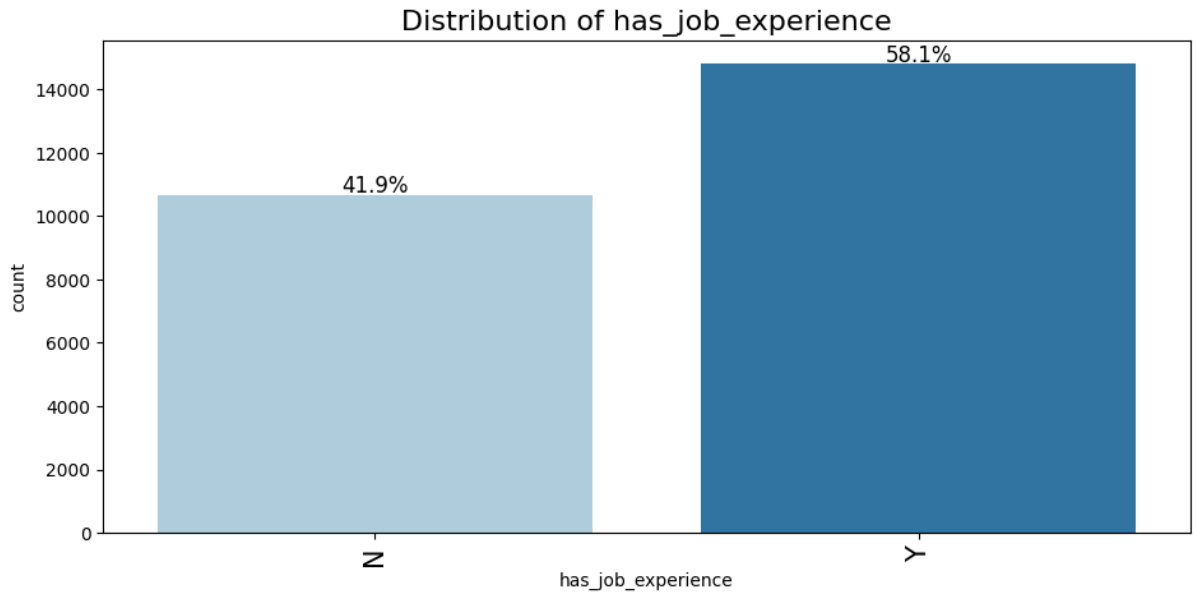

Figure 3. Distribution of Job Experience Among Applicants

Interpretation:

The distribution of job experience among applicants reveals that 58.1% of applicants have prior job experience, while 41.9% do not. This indicates that a majority of the applicants possess relevant work experience, which could play a significant role in the visa approval process.

4. Distribution of Required Job Training:

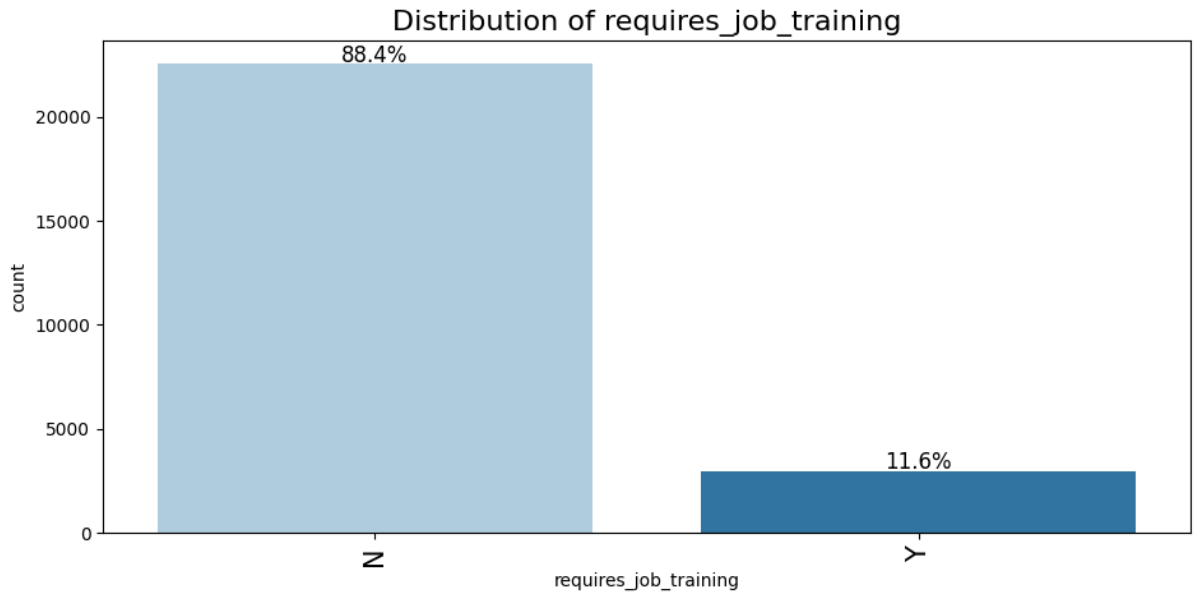

Figure 4. Distribution of Required Job Training

Interpretation:

The bar graph illustrates the distribution of job postings based on the requirement for job training. A clear majority, 88.4%, of the postings do not require specific training, while a smaller percentage, 11.6%, necessitate for training. This suggests that most job opportunities in the dataset do not have job training requirements.

5. Distribution of Employment by Region:

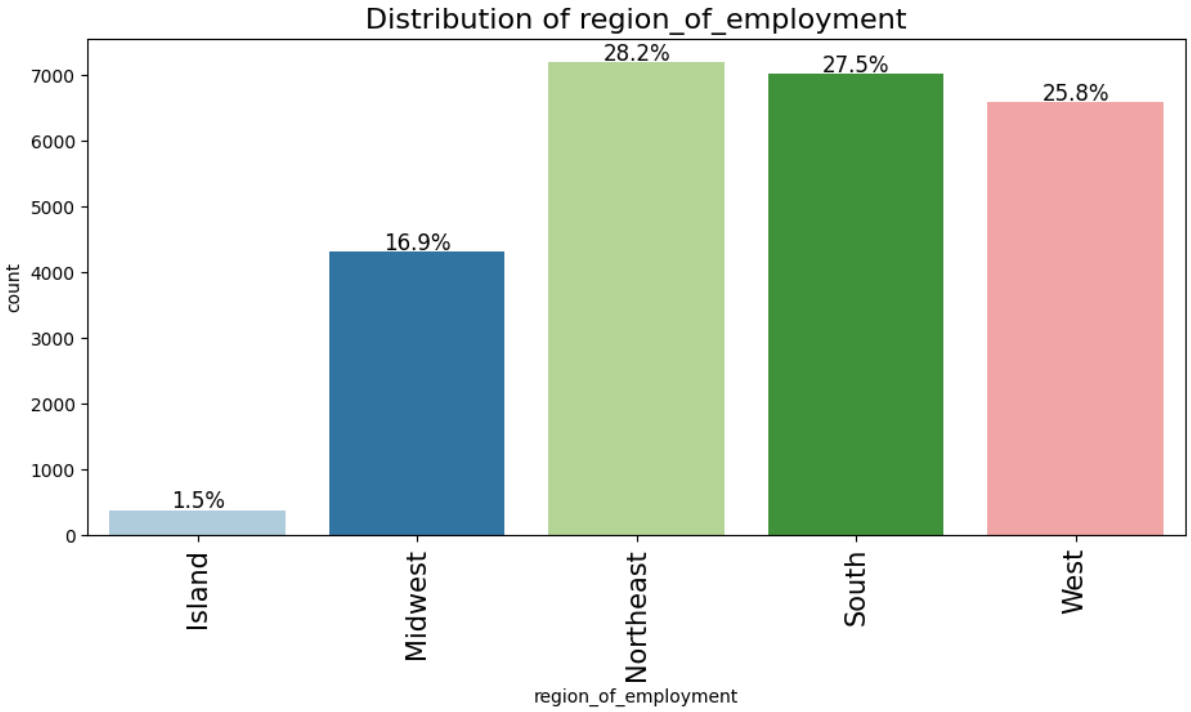

Figure 5. Distribution of Employment by Region

Interpretation:

The bar graph displays the distribution of job postings across various regions of employment. The Northeast and South regions lead with the highest number of job postings, representing approximately 28.2% and 27.5% of the total, respectively. The West follows closely with about 25.8% of the postings, while the Midwest accounts for 16.9%. The Island region has the smallest share, with only 1.5% of the job postings. Overall, the graph highlights a concentration of job opportunities in the Northeast, South, and West, with the Midwest also having a notable presence, and the Island region offering the fewest opportunities.

6. Distribution of Wage Unit:

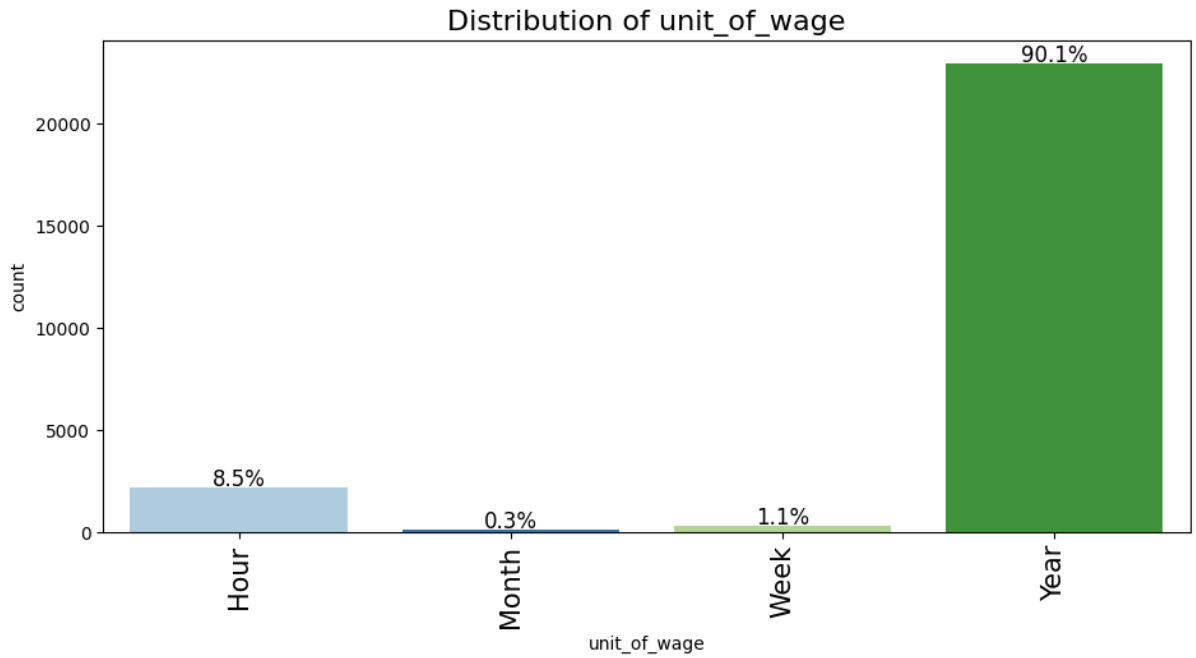

Figure 6: Distribution of Wage Unit

Interpretation:

The bar graph illustrates the distribution of job postings based on the unit of wage. The majority, 90.1%, of job postings offer a yearly wage, while a smaller proportion, 8.5%, specify an hourly wage. Very few postings list weekly (1.1%) or monthly (0.3%) wage units. Overall, the graph shows that most job opportunities are advertised with an annual salary, with hourly wages being less common.

7. Distribution of Full-Time Positions:

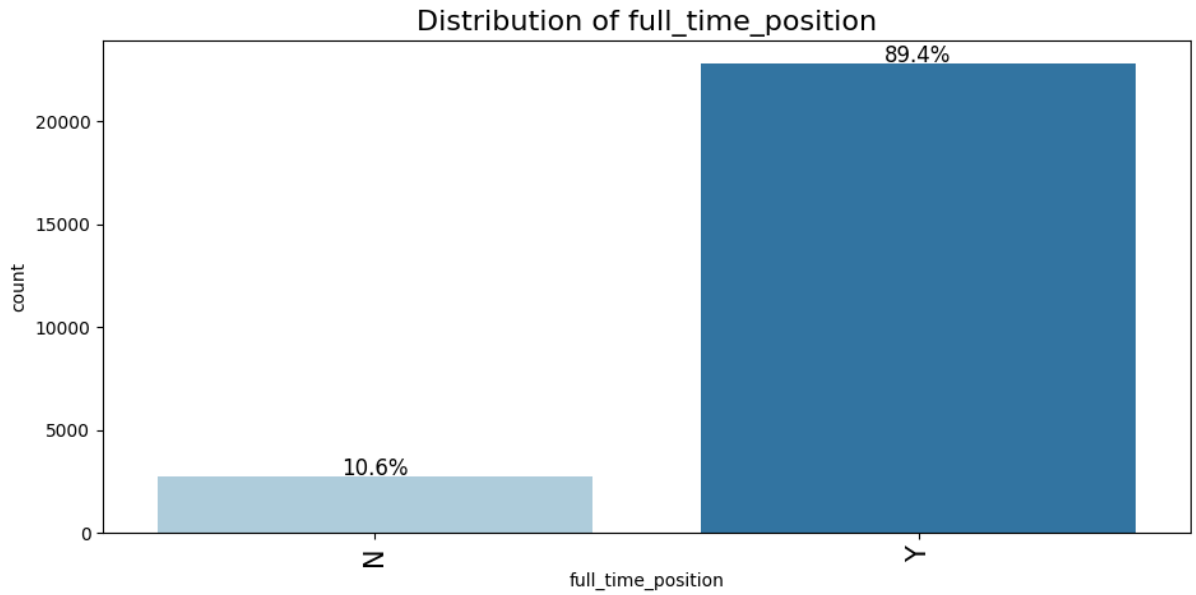

Figure 7. Distribution of Full-Time Positions

Interpretation:

The bar graph shows the distribution of job postings based on whether they are full-time positions. A significant majority, 89.4%, of the job postings are for full-time positions, while a smaller proportion, 10.6%, are for part-time or other non-full-time roles. Overall, the graph reveals that most job opportunities in the dataset are full-time positions.

8. Distribution of Visa Case Status:

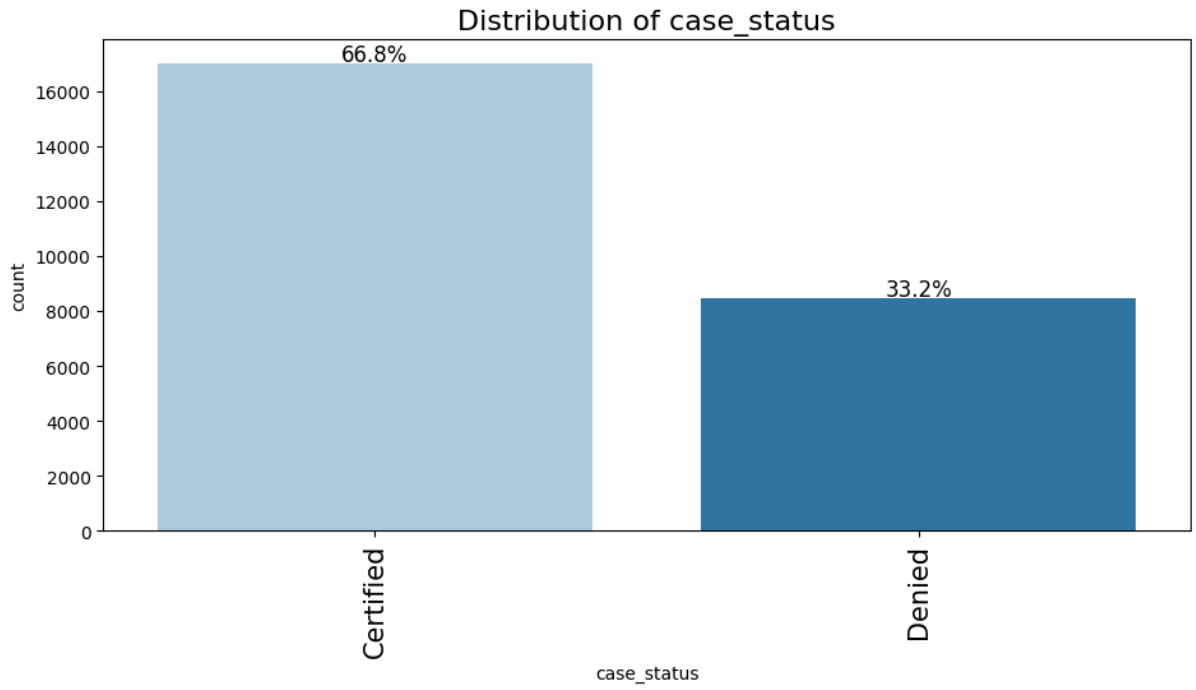

Figure 8. Distribution of Visa Case Status

Interpretation:

The bar graph illustrates the distribution of visa cases based on their status. A significant majority, 66.8%, of the visa cases were certified, while a smaller proportion, 33.2%, were denied. Overall, the graph indicates that most visa applications were approved.

Numerical Columns:

9. Distribution of Number of Employees:

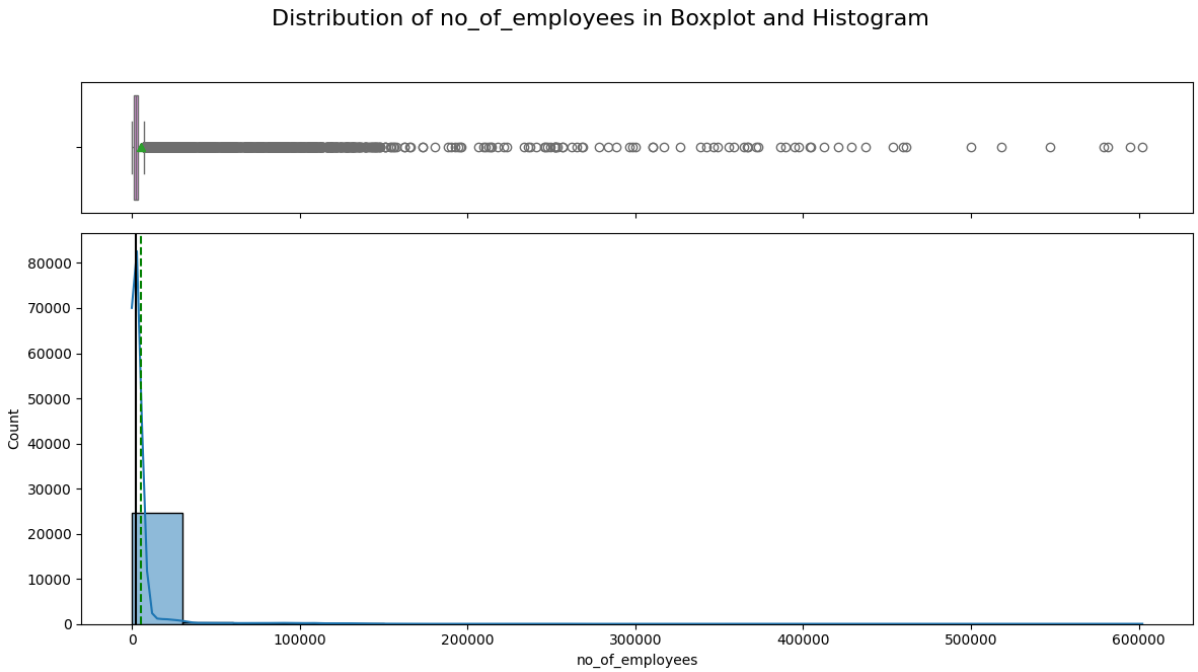

Figure 9. Distribution of Number of Employees

Interpretation:

The distribution of the number of employees is highly skewed, with most organizations having a small number of employees. The histogram shows a dense clustering below 10,000 employees, while the boxplot highlights the presence of significant outliers extending up to 602,069 employees. The mean number of employees is 5,667, and the median is 2,109, indicating that the data is heavily influenced by outliers. The interquartile range (IQR) lies between 1,022 (25th percentile) and 3,504 (75th percentile), further emphasizing the concentration of organizations with smaller workforce sizes. This stark disparity suggests that a few companies dominate in terms of workforce size, while the majority remain relatively small.

10. Distribution of yr_of_estab:

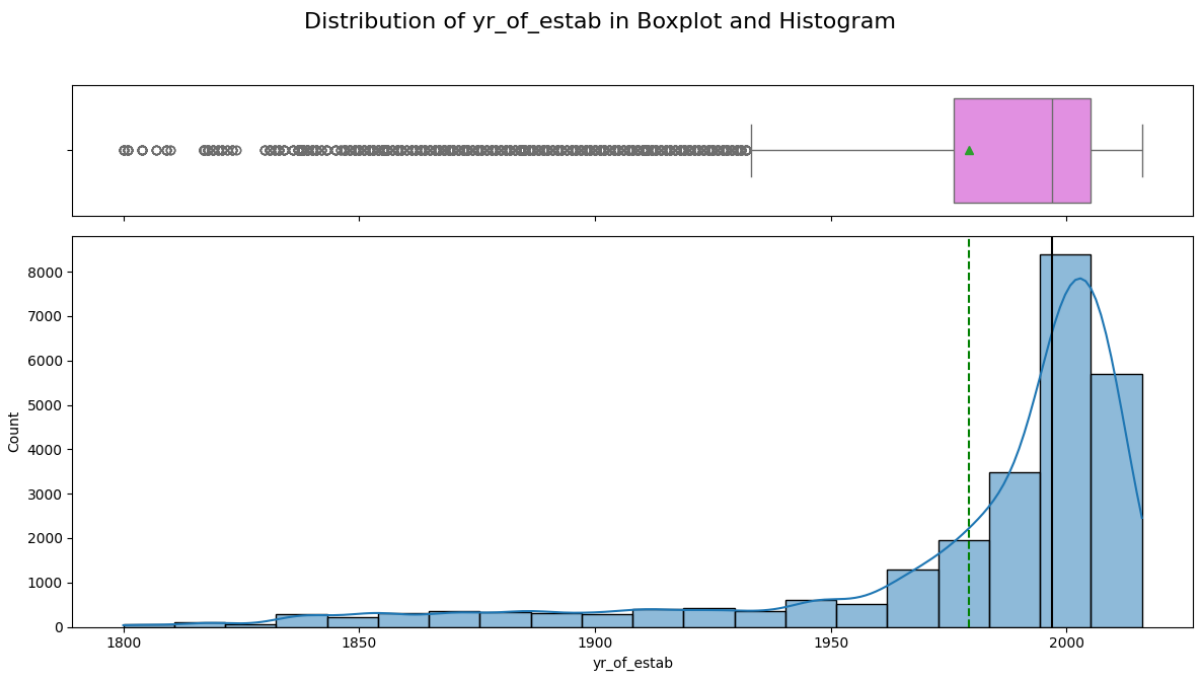

Figure 10. Distribution of yr_of_estab

Interpretation:

The distribution of the year of establishment shows a clear right-skewed pattern, with a significant number of companies being established after 1950. The histogram indicates a gradual increase in the count of establishments over time, with a peak in the late 20th century. The mean year of establishment is approximately 1980, while the median is around 1995, signifying that most organizations were established in recent decades. The boxplot highlights a few extreme outliers representing companies established before 1800, which deviate significantly from the rest of the data. This pattern reflects the trend of increasing business formation in modern times.

11. Distribution of prevailing_wage:

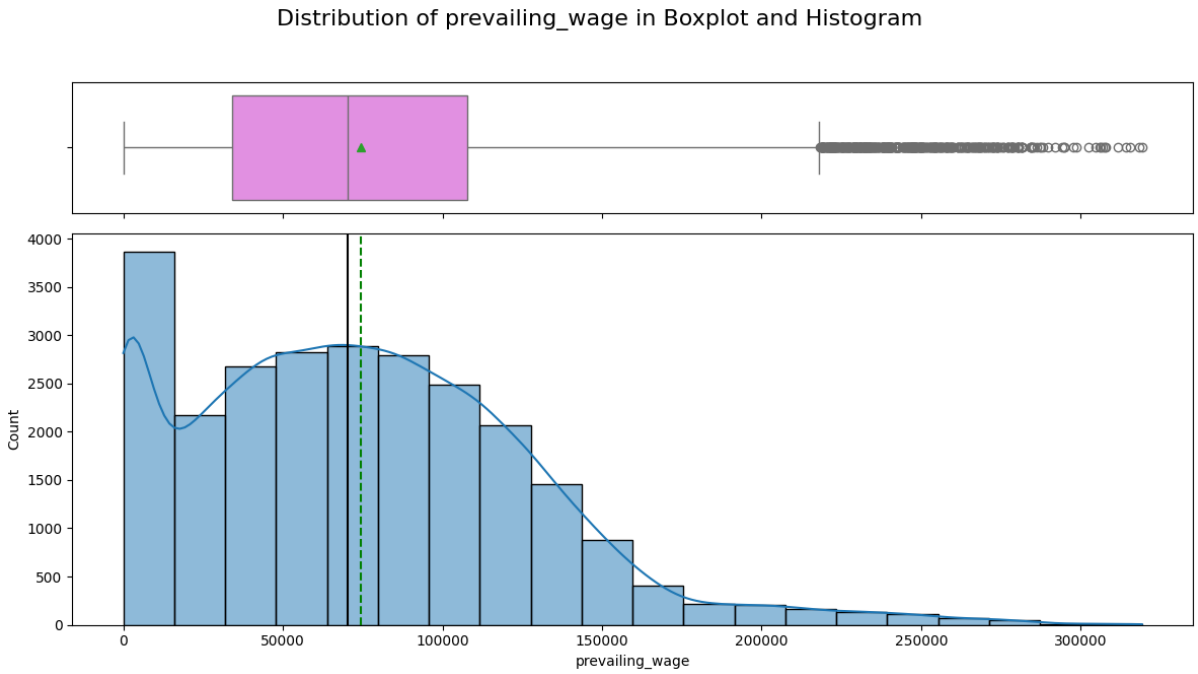

Figure 11. Distribution of prevailing_wage

Interpretation:

The distribution of prevailing wages is positively skewed, with most wages concentrated below $100,000. The histogram reveals a gradual decline in frequency as wages increase, with the majority of values clustering around the median of approximately $60,000. The mean wage is slightly higher at about $70,000, indicating the influence of higher wage outliers.

The boxplot highlights the presence of numerous outliers approximately beyond $220,000, stretching up to over $300,000. These outliers signify a small subset of high-paying jobs, which considerably impact the overall distribution. This pattern reflects significant wage disparities within the dataset.

1.4.Bivariate Analysis:

12. Correlation Between Numeric Columns and Case Status:

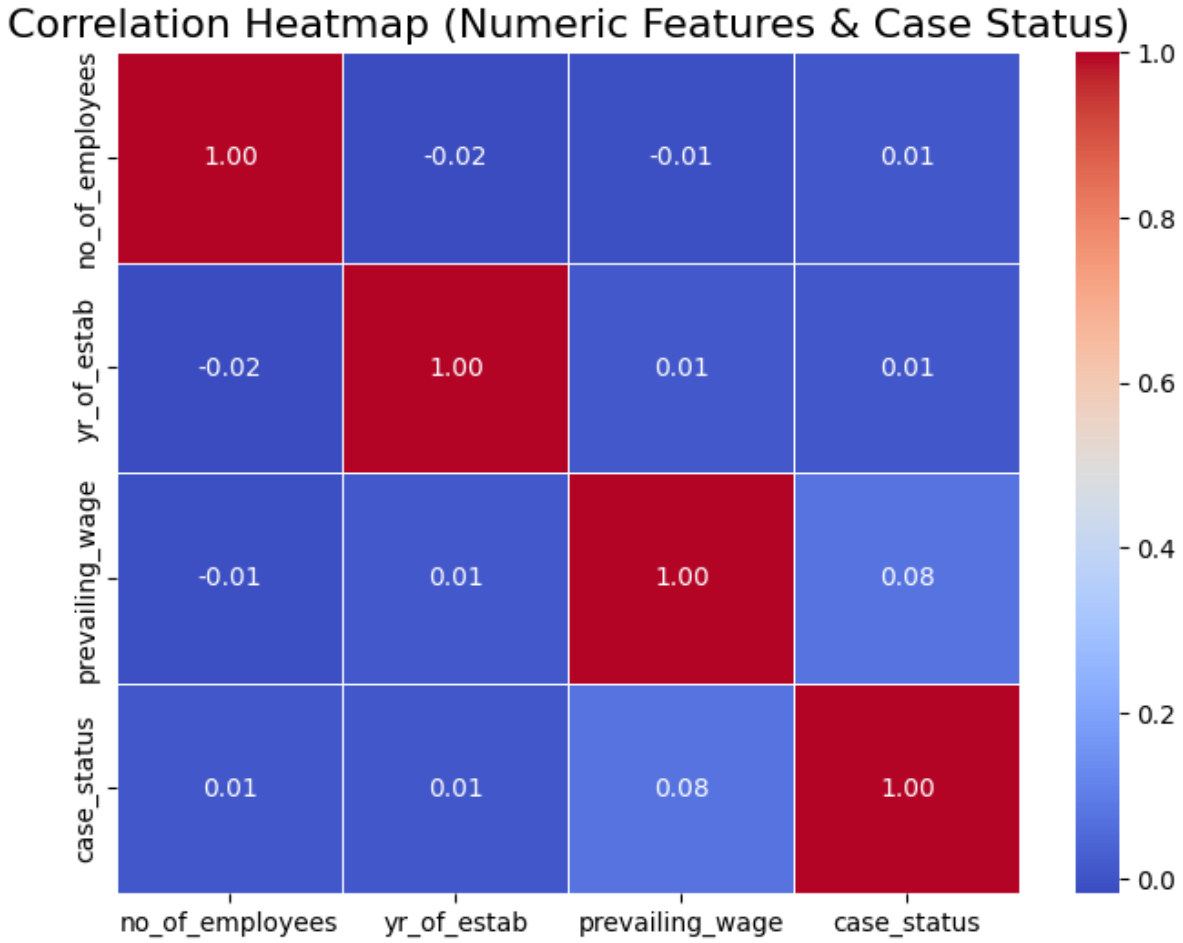

Figure 12. Correlation Between Numeric Columns and Case Status

Interpretation:

The correlation heatmap shows weak correlations between numeric features (no_of_employees, yr_of_estab, prevailing_wage) and the categorical feature (case_status), with values close to 0, indicating minimal impact on case status. There is a slight positive correlation between prevailing_wage and case_status (0.08), suggesting a marginal increase in the chance of certification with higher wages. However, no strong relationships are observed between the numeric features, indicating their relative independence.

Categorical Columns :

13. Visa Approval Distribution by Continent and Case Status:

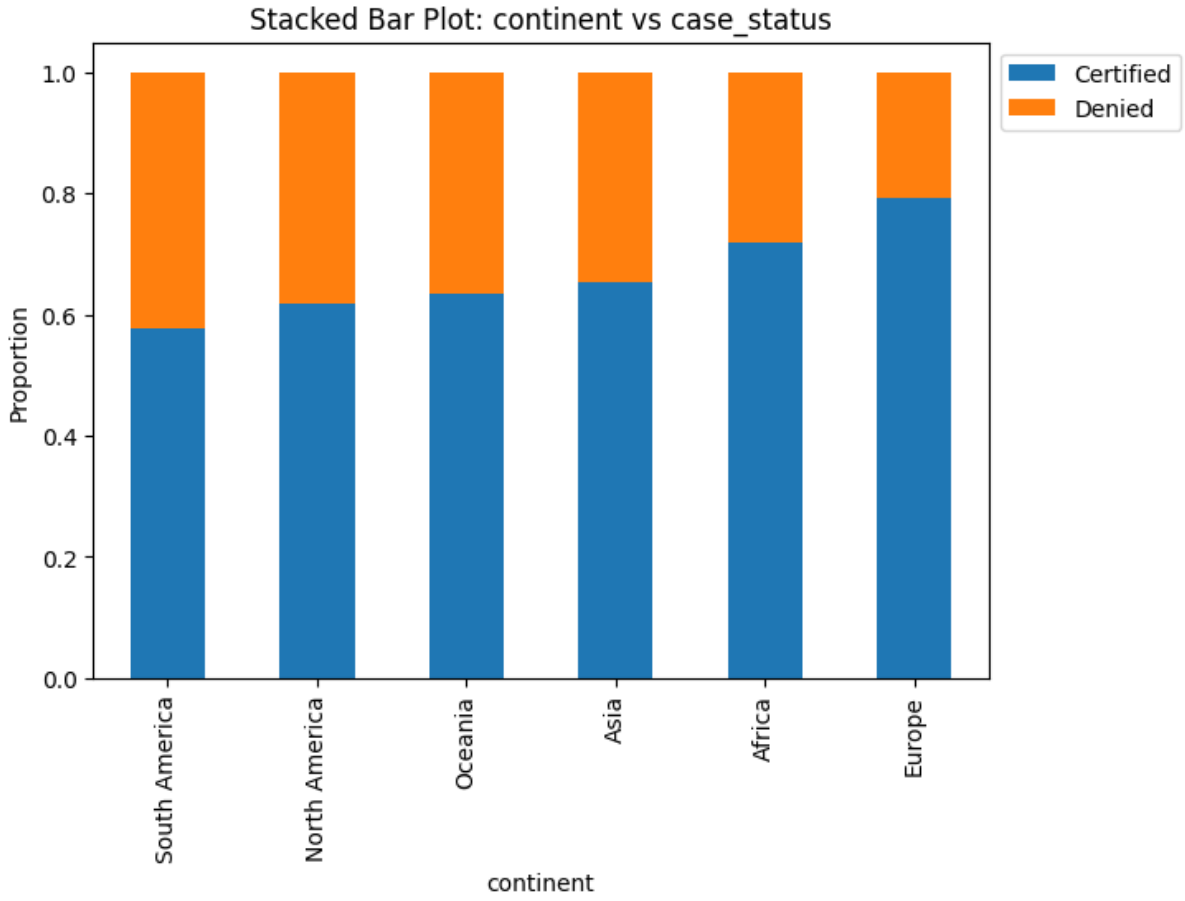

Figure 13. Visa Approval Distribution by Continent and Case Status

Interpretation:

The stacked bar plot illustrates that the majority of visa applications across all continents are certified, with certified cases consistently outnumbering denied ones. Europe has the highest certification rate at approximately 80%, followed by Africa at 75%. Oceania and Asia show around 65% certified cases, while North America has about 60% certified. South America has the lowest certification rate, with less than 60% of cases certified. These trends highlight a generally high approval rate for visa applications across continents, with minor variations in approval rates between regions.

14. Visa Approval Distribution by education_of_employee vs case_status:

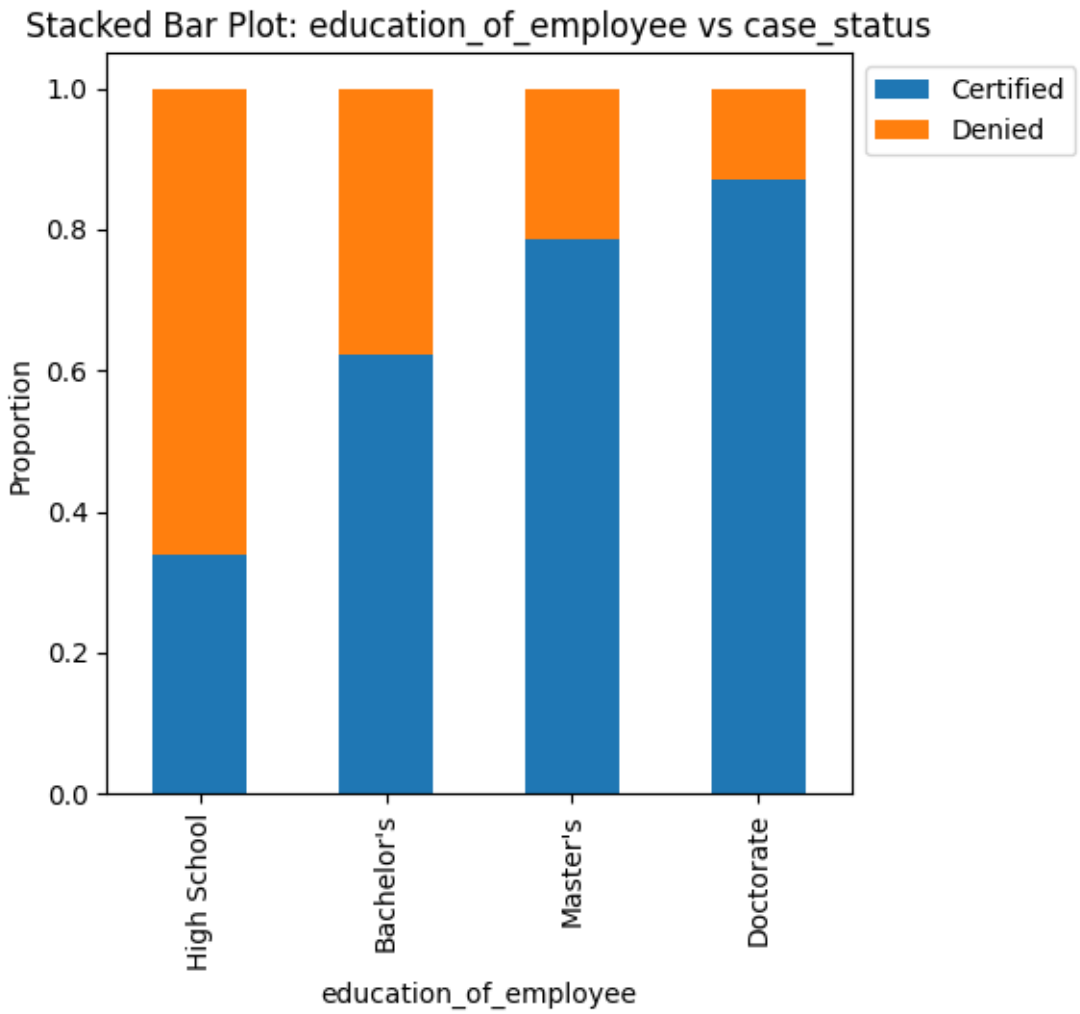

Figure 14: Visa Approval Distribution by education_of_employee vs case_status

Interpretation:

The stacked bar plot shows a clear trend where higher education levels are associated with higher visa approval rates. Individuals with a doctorate degree have the highest certification rate at approximately 90%, followed by master's degree holders at 80%. Bachelor's degree holders have about 60% certified cases, while high school graduates have the lowest certification rate at around 35%. Overall, the data suggests that higher education levels, particularly advanced degrees like master's and doctorate, are positively correlated with higher chances of visa approval.

15. Visa Approval Distribution by has_job_experience vs case_status:

Figure 15. Visa Approval Distribution by has_job_experience vs case_status

Interpretation:

The stacked bar plot shows a strong positive correlation between job experience and visa approval. Applicants with prior job experience have a much higher proportion of certified cases around 75% compared to those without job experience below 60%. Conversely, applicants without job experience are more likely to have their visa denied, with above 40% denial rate compared to only 25% for those with job experience. Overall, the data suggests that having job experience significantly increases the likelihood of visa approval.

16. Visa Approval Distribution by requires_job_training vs case_status:

Figure 16. Visa Approval Distribution by requires_job_training vs case_status

Interpretation:

The stacked bar plot shows the distribution of certified and denied visa cases based on whether the job requires prior training. Both categories—jobs requiring training and jobs not requiring training— approximately display similar approval rates. In both cases, around 65% of visa applications were certified, while around 35% were denied. This suggests that the requirement for job training does not have a significant impact on visa approval, as both categories have nearly identical certification and denial rates. The overall trend indicates that the job training requirement does not substantially influence the likelihood of visa approval.

17. Visa Approval Distribution by region_of_employment vs case_status:

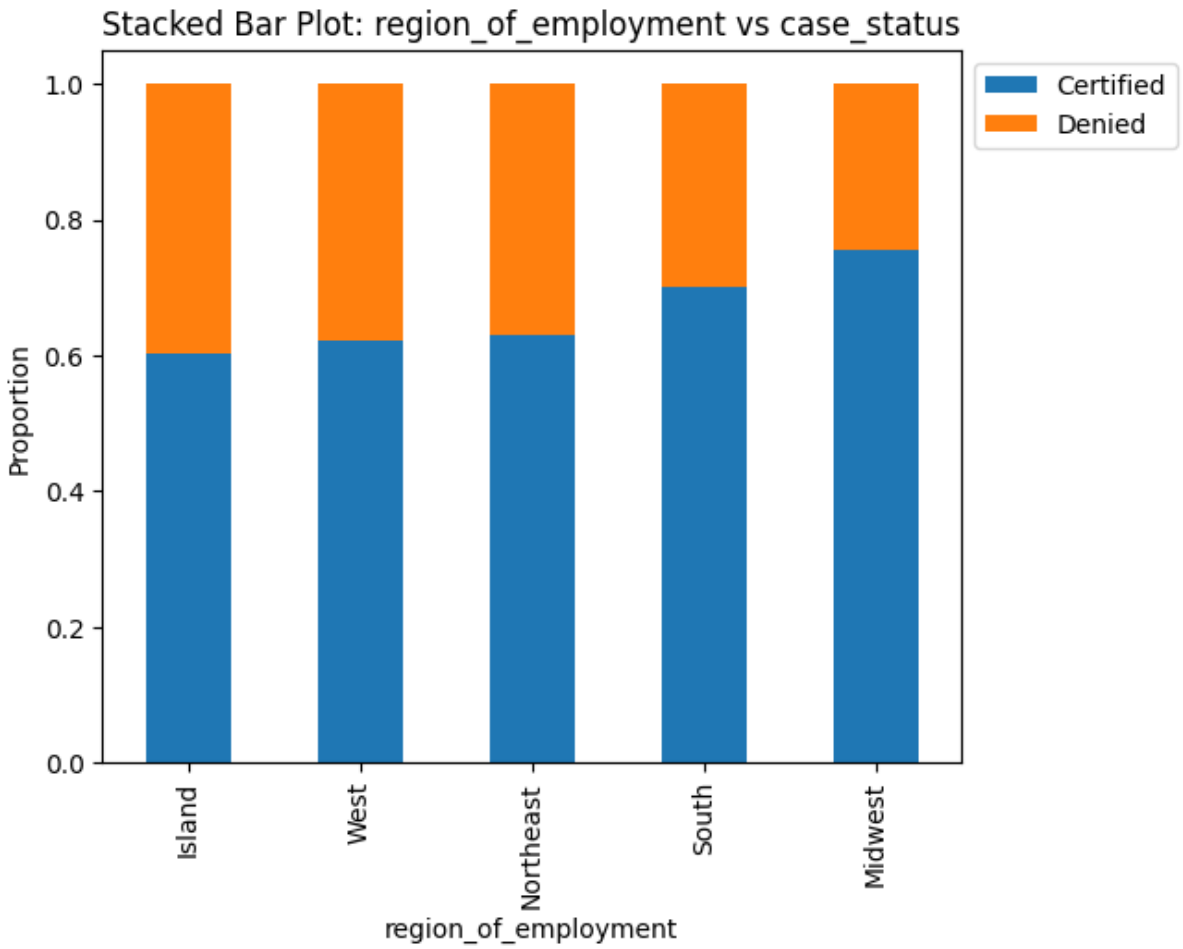

Figure 17. Visa Approval Distribution by region_of_employment vs case_status

Interpretation:

The stacked bar plot shows the distribution of certified and denied visa cases across different regions of employment. The approval rates are generally consistent across regions, with slight variations. The Midwest region has the highest certification rate, around 75%, while other regions like the South show approximately 70% certified cases, and the Northeast and West show approximately 65% certified cases. The Island region has the lowest, around 60%. Overall, the region of employment appears to have a minimal effect on visa approval, as most regions exhibit similar trends in certification and denial rates.

18. Visa Approval Distribution by unit_of_wage vs case_status:

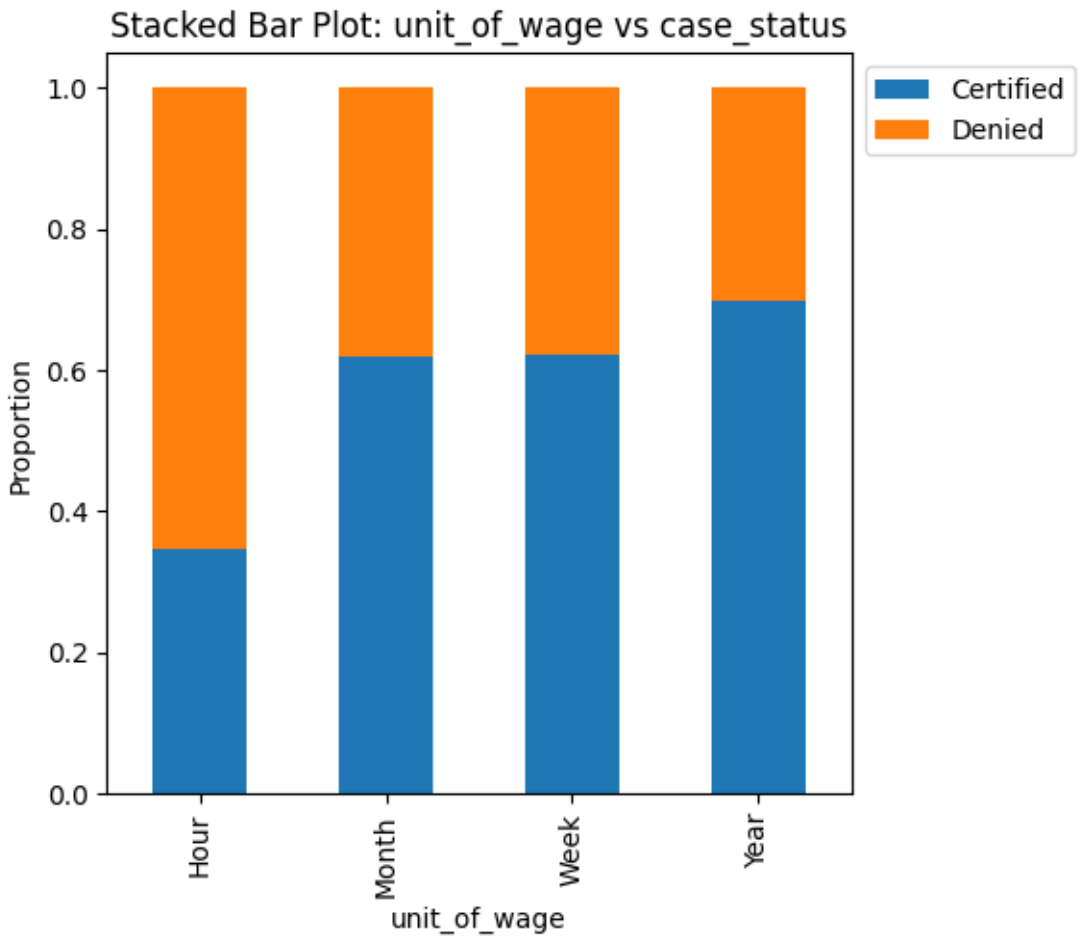

Figure 18. Visa Approval Distribution by unit_of_wage vs case_status

Interpretation:

The stacked bar plot shows the distribution of certified and denied visa cases across different wage units. Overall, the approval rates are fairly consistent across the units, with slight variations. The "Hour" unit of wage has a higher proportion of denied cases (approximately 65%), while the "Year" unit has the highest certification rate, around 70%. "Month" and "Week" units have similar certification rates of about 60%.

19. Visa Approval Distribution by full_time_position vs case_status:

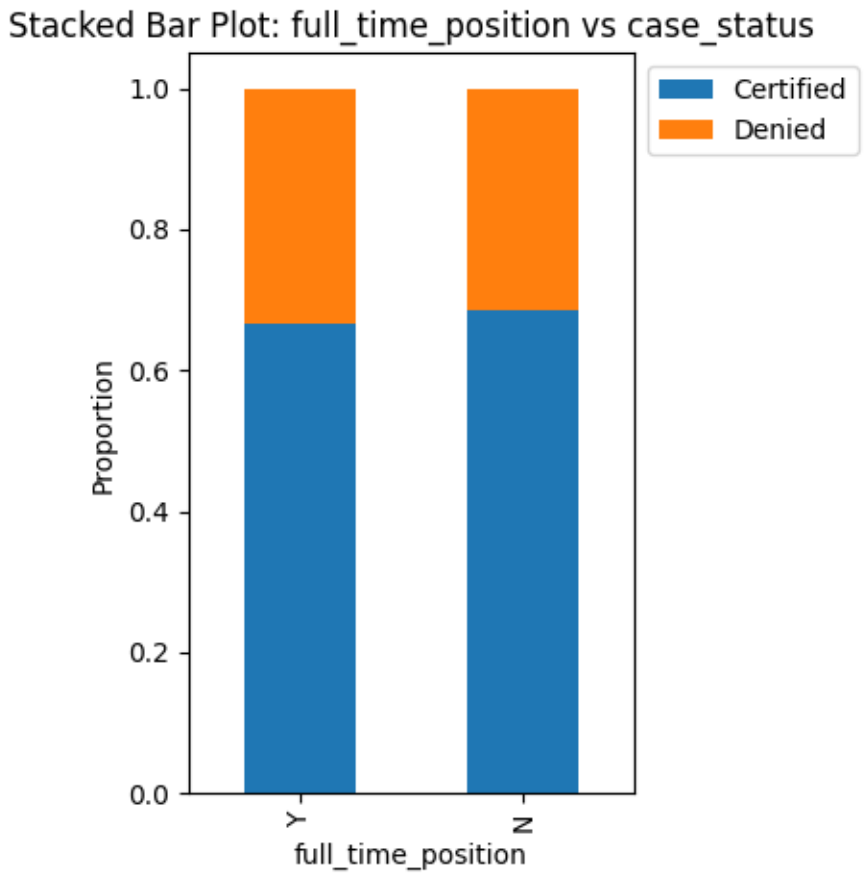

Figure 19. Visa Approval Distribution by full_time_position vs case_status

Interpretation:

The stacked bar plot shows the distribution of certified and denied visa cases based on whether the position is full-time or not. Overall, the approval rates are similar for both full-time and non-full-time positions, with full-time positions having a slightly higher proportion of certified cases (approximately 70%) compared to non-full-time positions (approximately 65%). In conclusion, the plot indicates that the employment status (full-time or non-full-time) has a minimal impact on the likelihood of visa approval, with only slight variations in the approval rates.

Numerical Columns :

20. Visa Approval Distribution of no_of_employees by case_status:

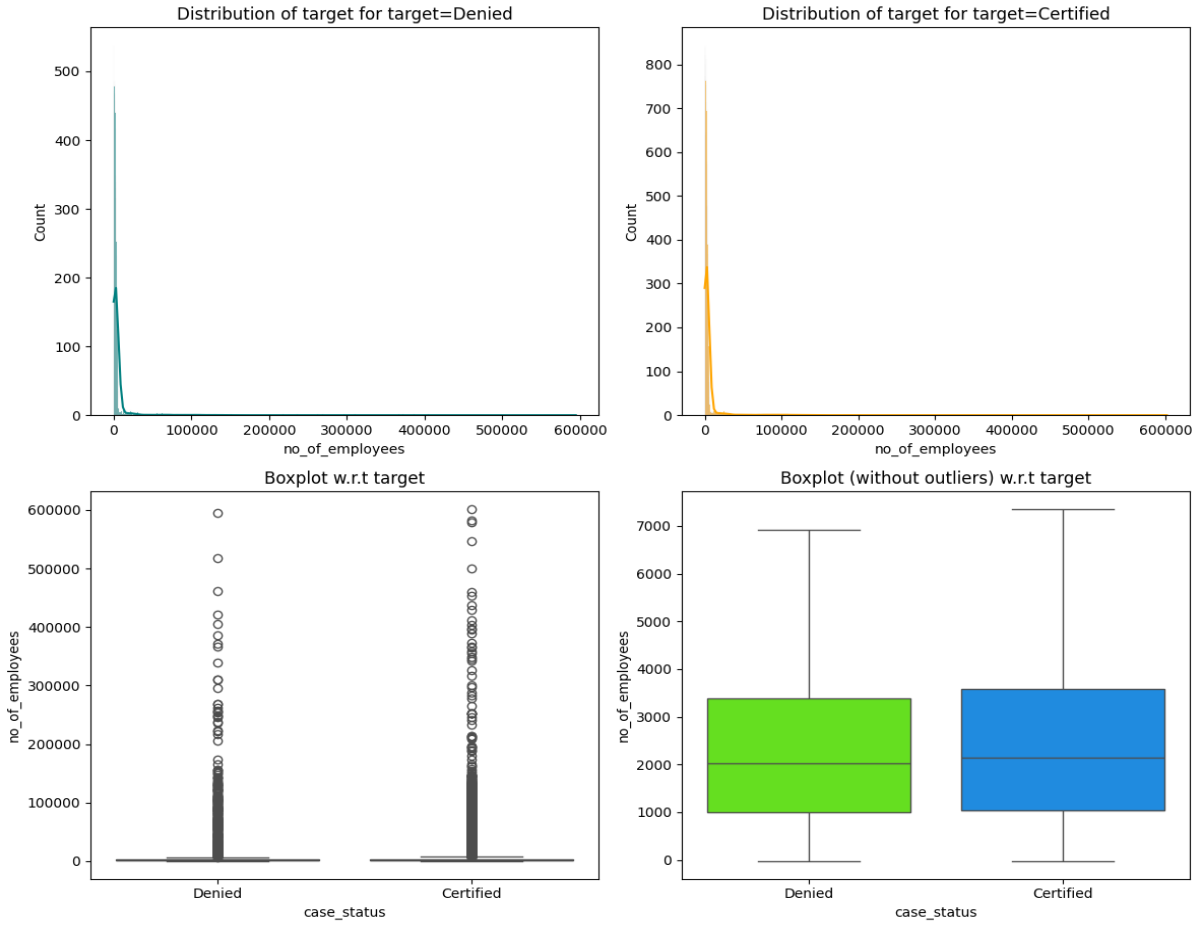

Figure 20. Visa Approval Distribution of no_of_employees by case_status

Interpretation:

The Histogram reveal the distribution of employee numbers for both certified and denied visa cases, showing right-skewed patterns for both. Denied cases are primarily from companies with fewer employees, though a few larger companies are present, while certified cases, though also right-skewed, tend to have a slightly higher proportion of companies with more employees.

Boxplots further show that denied cases have a lower median employee count with more outliers, whereas certified cases have a slightly higher median. Overall, the number of employees appears to have a minor impact on visa case outcomes, with larger companies slightly more likely to have their cases certified.

21. Visa Approval Distribution of no_of_employees by case_status:

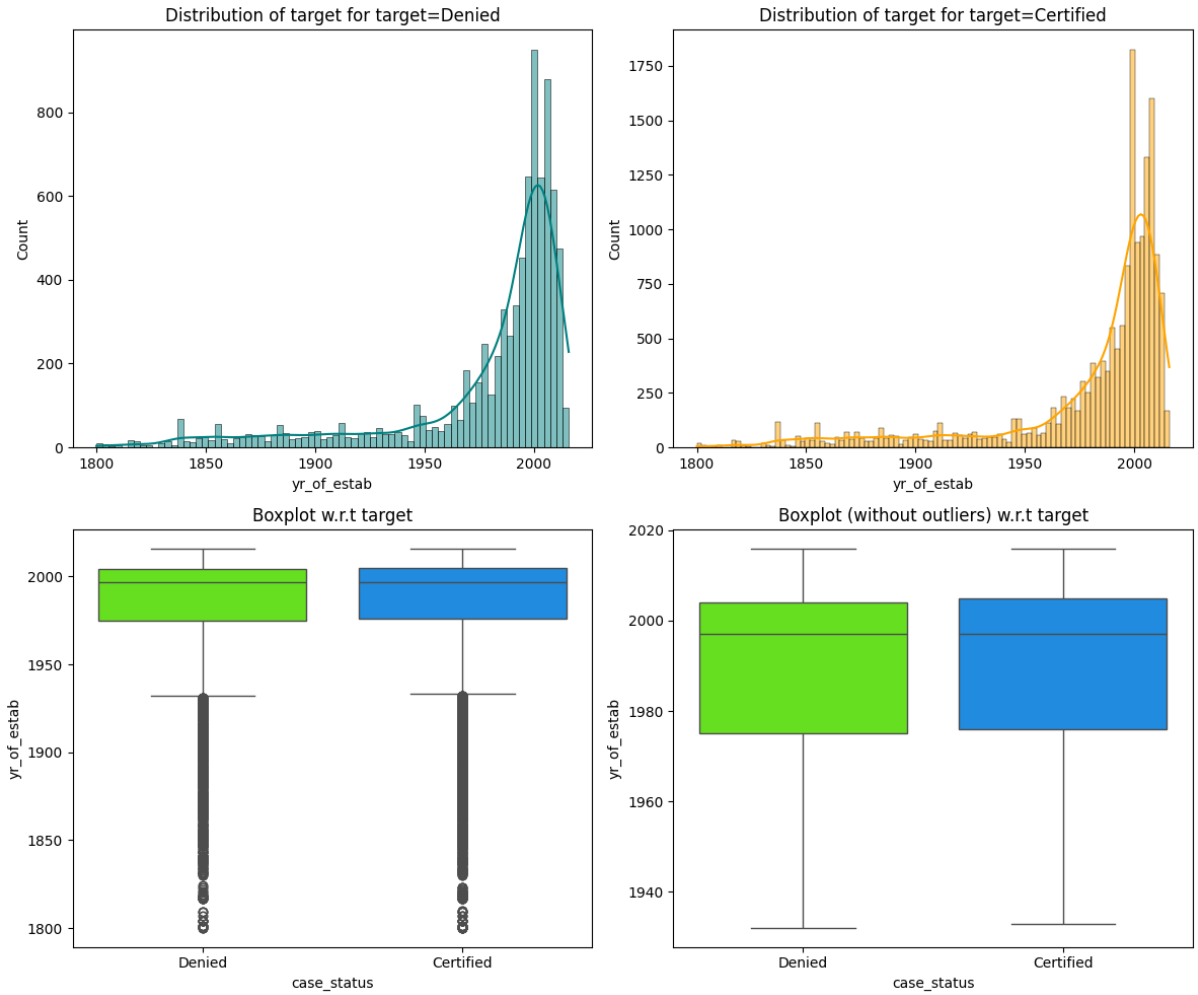

Figure 21. Visa Approval Distribution of no_of_employees by case_status

Interpretation:

The plots reveal that both denied and certified cases exhibit similar left-skewed distributions, with tails extending toward more older companies. The boxplots for both denied and certified cases are nearly identical, with comparable ranges and medians, suggesting that the year of establishment has little correlation with the case status. This indicates that the establishment year of a company does not significantly influence whether a visa case is certified or denied.

22. Visa Approval Distribution of Prevailing Wage by case_status

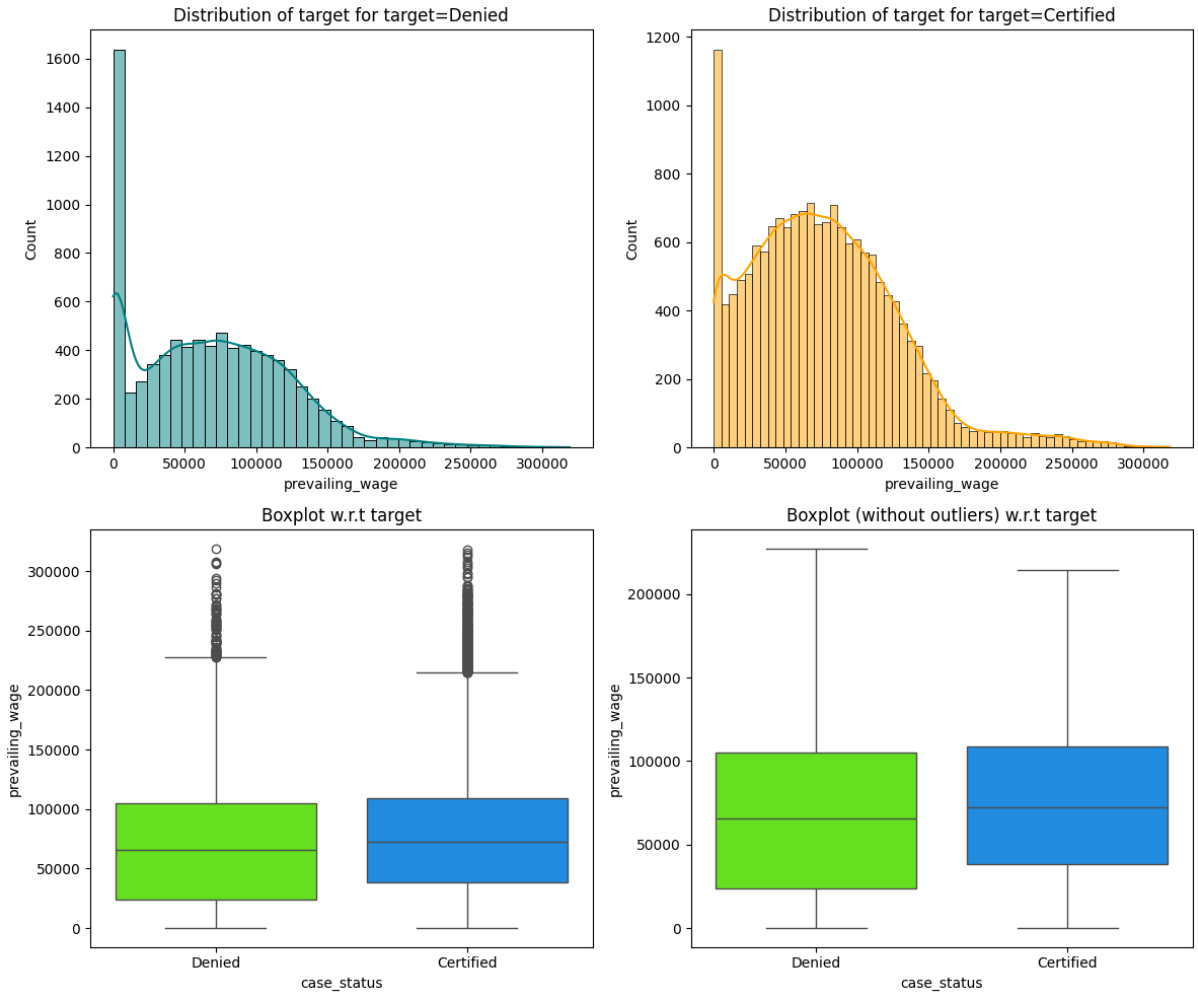

Figure 22. Visa Approval Distribution of Prevailing Wage by case_status

Interpretation:

The plots illustrate the distribution of prevailing wages for both certified and denied visa cases. Both the histograms and boxplots show right-skewed distributions, with denied cases having a longer tail toward higher wages. Denied cases have a wider range of wages and a median around $70,000, suggesting that they are more likely to involve jobs with lower prevailing wages. The certified cases show a comparatively narrower range with a slightly higher median of around $80,000, indicating that jobs with higher prevailing wages have a slightly higher likelihood of being certified.

1.5. Key Insights on Variable Relationships:

1.Observations on Individual Variables:

Categorical Columns:

The categorical variables highlight significant trends in visa applications and job characteristics. Asia dominates as the primary source of visa applicants, reflecting regional disparities in migration patterns. Education levels of applicants are predominantly high, with most holding Bachelor's or Master's degrees, aligning with the preference for skilled workers. Job experience is common among applicants, and most job postings do not require additional training. Furthermore, full-time positions and annual wage offers dominate the dataset. These patterns suggest a strong demand for skilled, experienced, and full-time employees across key regions of employment in the U.S.

Categorical variables emphasize the concentration of skilled visa applicants and job postings in specific regions and sectors, highlighting the demand for educated and experienced professionals in full-time roles with annual wages.

Numerical Columns:

The numerical variables reveal diverse patterns within the dataset. The majority of organizations have small workforces, while a few large companies significantly influence the data. Businesses established after 1950 dominate, reflecting recent economic trends. The prevailing wage distribution shows significant disparities, with most wages clustering below $100,000, while outliers reflect a subset of high-paying jobs. These observations highlight the diversity in company sizes, establishment timelines, and compensation structures.

The analysis of numerical variables underscores the dominance of small organizations and recent business establishments, along with significant wage disparities influenced by a few high-paying jobs.

2.Observations on Relationships Between Variables:

Categorical Columns:

Examining relationships between categorical variables provides deeper insights. Higher visa approval rates are observed in Europe and among applicants with advanced degrees or prior job experience. Education and experience play crucial roles in determining visa success, while factors like job training requirements and employment type have minimal impact. Geographical variations in approval rates also suggest regional influences on visa outcomes.

Relationships among categorical variables reinforce the importance of education and job experience in visa approval, with regional differences adding complexity to the approval dynamics.

Numerical Columns:

Relationships between numerical variables and visa approval outcomes reveal nuanced patterns. Larger organizations are slightly more likely to have visa applications approved, while the year of establishment has minimal impact. Higher prevailing wages show a slight correlation with approval rates, indicating a preference for higher-paying roles. However, outliers in workforce size and wages contribute to variability in these trends.

Numerical relationships suggest that while factors like company size and wages may marginally influence visa approvals, their impact is not pronounced, with overall trends reflecting broader workforce and wage dynamics.

2.Data preprocessing

2.1. Data Cleaning and Preparation:

The initial step of data preprocessing focused on addressing data inconsistencies to ensure the dataset's readiness for analysis. Negative values in the "no_of_employees" column were identified as a key issue, with the column's minimum value initially recorded as -26. To address this, the median value of the column, calculated as 2,109, was used to replace all negative values. Following this treatment, the minimum value increased to 12, and the column's descriptive statistics were updated, reflecting a mean of 5,669.80 and a standard deviation of 22,877.37. This adjustment eliminated negative values, resulting in a more accurate and meaningful representation of employee numbers.

In addition to handling negative values, the dataset was evaluated for duplicate entries, and no duplicate rows were found. This confirmed the dataset's integrity, ensuring a consistent foundation for further analysis. These preparatory actions have refined the dataset, making it suitable for subsequent preprocessing steps.

2.2.Feature Engineering:

In the Feature Engineering step, a new feature, company_age, was created to represent the age of the company by calculating the difference between the current year and the year of establishment. Following the creation of this feature, the redundant column, yr_of_estab, was dropped to streamline the dataset and eliminate unnecessary repetition of information.

Manual mapping was applied to the categorical variables that have ordered values. The levels of education_of_employee were mapped to numerical values, with Bachelor's mapped to 2, Master's to 3, High School to 1, and Doctorate to 4. Similarly, binary categorical variables such as has_job_experience, requires_job_training, and full_time_position were mapped, where "Yes" was assigned a value of 1 and "No" was assigned a value of 0. The unit_of_wage variable was mapped to numerical values to represent the frequency of wages: Year was mapped to 4, Hour to 1, Week to 2, and Month to 3. The case_status has 2 values where certified is mapped to 1 and denied is mapped to 0.

For the categorical variables with unordered values, continent and region_of_employment, one-hot encoding was applied. This transformation resulted in multiple binary columns representing each category, with the first category in each variable

being dropped to avoid multicollinearity. Lastly, all boolean values in the dataset were converted to float data types to ensure consistency and compatibility with machine learning models. These transformations have effectively prepared the dataset for subsequent analysis and model development.

Multicollinearity Assessment and Findings:

There is no significant multicollinearity observed among the features in the dataset. Most of the correlation values between the numerical variables are relatively low, with the highest correlations observed between certain regions of employment and the continent columns, which is expected given the geographical categories. However, the correlation values do not exceed the threshold that would indicate problematic multicollinearity (typically around 0.8 or higher). Therefore, multicollinearity is not a major concern for this dataset, and there is no immediate need for feature elimination or transformation due to multicollinearity.

2.3.Missing Value Treatment:

Observation:

No columns in the dataset have missing values. All features have a count of zero for missing values.

1. **No Missing Data:**
    - Since no missing values exist, there is no need for imputation or removal of records at this stage.
2. **Verification:**
    - It is essential to validate the absence of missing data to ensure data completeness and accuracy. This ensures no errors arise from unobserved data gaps during subsequent analysis.
3. **Future Considerations:**
    - If any missing values are introduced during feature engineering or derived computations, appropriate handling techniques such as mean/median

imputation, mode imputation, or advanced methods like KNN imputation may be applied based on the nature of the data.

No further action is required for missing value treatment at this stage since the dataset is already complete.

2.4.Outlier Treatment:

During the outlier treatment process, the dataset initially had 25,480 datas. After applying outlier removal techniques, the dataset was reduced significantly to 20,531, indicating a substantial data loss of nearly 20%. Given this considerable reduction in data size, it was decided not to proceed with outlier removal. Instead, the analysis will move forward without treating the outliers, opting to use machine learning models that are less sensitive to the presence of outliers. This approach preserves the integrity and diversity of the dataset while ensuring that the analysis remains robust.

2.5. Ensuring Clean Split of Training, Testing, and Validation Data

To prevent data leakage and ensure robust model evaluation, the dataset was split into training, validation, and test sets. First, the features (X) and target variable (y) were separated, with the target variable being the numerical representation of case_status. A stratified splitting strategy was employed to maintain the class proportions across all subsets.

The training set was allocated 80% of the total data, while the remaining 20% was split evenly between the validation and test sets. Stratified sampling was applied again during the splitting of the validation and test sets to preserve class balance. This process ensures that the proportions of Certified and Denied cases remain consistent across all subsets.

The resulting shapes of the splits were as follows: the training set consisted of 80% of the data, while the validation and test sets each held 10%. This structured approach effectively minimizes the risk of data leakage while enabling reliable model evaluation during the training and tuning phases.

3. Model Building - Original Data

3.1. Selection of Evaluation Metric:

In the context of visa certification, where false positives (denied cases predicted as approved) carry a higher cost due to regulatory violations and their potential to affect U.S. workers' wages and conditions, Recall for the denied class is the key metric. This ensures that the model correctly identifies as many denied cases as possible, minimizing false positives. At the same time, overall performance on the validation set is measured to ensure the model is generalizing well. Thus, Recall, along with the difference between training and validation Recall, is used as the primary evaluation metric for this task.

3.2. Development of Classification Models:

Five classification models were built and evaluated:

- Decision Tree: A simple tree-based model, useful for understanding decision paths but prone to overfitting.
- Random Forest: A bagging method that combines multiple decision trees for improved performance and robustness.
- Gradient Boosting: A boosting method that iteratively minimizes errors, providing high accuracy on validation data.
- AdaBoost: Another boosting method that focuses on correcting previous errors, typically robust to overfitting.
- XGBoost: An advanced boosting algorithm known for its speed and efficiency, often outperforming other tree-based methods.

Each model was trained and evaluated using Recall as the key metric. The difference between training and validation Recall was also examined to assess overfitting.

3.3. Analysis of Model Performance:

The models performed as follows:

- Decision Tree:
  Training Recall: 1.0000, Validation Recall: 0.7368, Difference: 0.2632.
  The Decision Tree model overfit the training data significantly, leading to poor generalization on the validation set.

- Random Forest:
  Training Recall: 1.0000, Validation Recall: 0.8331, Difference: 0.1669.
  Random Forest improved generalization compared to the Decision Tree, but a moderate gap between training and validation Recall remained.

- Gradient Boosting:
  Training Recall: 0.8717, Validation Recall: 0.8660, Difference: 0.0056.
  Gradient Boosting showed excellent performance, striking a balance between training and validation Recall with minimal overfitting.

- AdaBoost:
  Training Recall: 0.8873, Validation Recall: 0.8749, Difference: 0.0125.
  Similar to Gradient Boosting, AdaBoost exhibited strong performance with low overfitting and high Recall on the validation set.

- XGBoost:
  Training Recall: 0.9274, Validation Recall: 0.8613, Difference: 0.0660.
  XGBoost demonstrated competitive performance but had a slightly larger gap between training and validation Recall compared to Gradient Boosting and AdaBoost.

Gradient Boosting and AdaBoost emerged as the best-performing models due to their high validation Recall and minimal overfitting. These models are better suited for the task, as they balance generalization and the ability to minimize false positives, which carry a higher cost.

4. Model Building - Oversampled Data

4.1. Oversampling of Training Data:

To address the imbalance in the training dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generates synthetic samples for the minority class (denied cases) to achieve a balanced class distribution. Before oversampling, the class distribution was heavily skewed, with 13,614 instances of the majority class (approved) and only 6,770 instances of the minority class. After applying SMOTE, the class distribution was equalized with 13,614 instances for each class. This ensured the models were trained on balanced data, improving their ability to detect the minority class effectively.

4.2. Development of Classification Models:

Using the oversampled data, five classification models were developed to predict visa application outcomes:

- Decision Tree: A simple and interpretable tree-based model, known for its propensity to overfit.
- Random Forest: A robust bagging technique that builds an ensemble of decision trees to improve generalization.
- Gradient Boosting: A powerful boosting technique that iteratively reduces errors, optimizing performance on validation data.
- AdaBoost: A boosting method that focuses on correcting misclassified instances, providing resilience to overfitting.
- XGBoost: An advanced gradient boosting algorithm that is computationally efficient and typically outperforms other tree-based models.

Each model was trained on the oversampled data and evaluated based on Recall for the denied class as the primary metric. The difference between training and validation Recall was also calculated to assess overfitting.

4.3. Analysis of Model Performance:

The models yielded the following results:

- Decision Tree:
    - Training Recall: 1.0000, Validation Recall: 0.7309, Difference: 0.2691
      The Decision Tree model overfitted significantly, achieving perfect Recall on the training data but generalizing poorly to the validation data.
- Random Forest:
    - Training Recall: 1.0000, Validation Recall: 0.8161, Difference: 0.1839
      Random Forest reduced overfitting compared to the Decision Tree but still exhibited a moderate gap between training and validation Recall.
- Gradient Boosting:
    - Training Recall: 0.8371, Validation Recall: 0.8308, Difference: 0.0063
      Gradient Boosting demonstrated outstanding performance with minimal overfitting, achieving a balance between training and validation Recall.
- AdaBoost:
    - Training Recall: 0.8352, Validation Recall: 0.8237, Difference: 0.0115
      Similar to Gradient Boosting, AdaBoost performed exceptionally well, showing high validation Recall with low overfitting.
- XGBoost:
    - Training Recall: 0.9052, Validation Recall: 0.8349, Difference: 0.0703
      XGBoost achieved competitive results but exhibited a slightly larger gap between training and validation Recall compared to Gradient Boosting and AdaBoost.

Gradient Boosting and AdaBoost performed best on oversampled data, achieving high validation recall with minimal overfitting, making them ideal for identifying denied visa applications. XGBoost showed strong performance but had slightly more overfitting. These results highlight the effectiveness of boosting methods in handling class imbalance and ensuring reliable predictions.

5. Model Building - Undersampled Data

5.1. Undersample the Train Data:

In Undersample the Train Data, the training data was undersampled using RandomUnderSampler to balance the class distribution. Prior to undersampling, the dataset had a skewed class distribution with 13,614 instances of the "approved" class and 6,770 instances of the "denied" class. After undersampling, the classes were balanced with 6,770 instances each for "approved" and "denied" cases.

5.2. Development of Classification Models:

Five models were trained on the undersampled dataset, using a variety of decision tree, bagging, and boosting methods:

- Decision Tree: A simple yet interpretable model that tends to overfit, especially on small datasets.
- Random Forest: A bagging method that combines multiple decision trees to improve accuracy and reduce overfitting.
- Gradient Boosting: A boosting technique that builds sequential models to reduce errors from previous ones, typically offering strong performance.
- AdaBoost: Another boosting method that adjusts the weight of misclassified samples to improve model performance.
- XGBoost: A highly optimized implementation of gradient boosting known for its speed and accuracy.

5.3. Analysis of Model Performance:

The models were evaluated based on Recall, which is the key evaluation metric for identifying "denied" cases in visa certification (where false negatives are costly).

- Decision Tree:
    - Training Recall: 1.0000
    - Validation Recall: 0.6093
    - Recall Difference: 0.3907
      The Decision Tree model showed significant overfitting, with perfect recall on the training set but a much lower recall on the validation set, indicating poor generalization.
- Random Forest:
    - Training Recall: 1.0000
    - Validation Recall: 0.6563
    - Recall Difference: 0.3437
      Like the Decision Tree, Random Forest also overfitted the training data but to a lesser extent, with a somewhat improved recall on the validation set.
- Gradient Boosting:
    - Training Recall: 0.7409
    - Validation Recall: 0.7180
    - Recall Difference: 0.0229
      Gradient Boosting performed well, with a balanced recall between the training and validation sets, indicating minimal overfitting and good generalization.
- AdaBoost:
    - Training Recall: 0.7180
    - Validation Recall: 0.7051

- Recall Difference: 0.0130
  AdaBoost performed similarly to Gradient Boosting, showing low overfitting and a high recall value on the validation set.

- XGBoost:
  - Training Recall: 0.8412
  - Validation Recall: 0.6692
  - Recall Difference: 0.1720
    XGBoost showed strong training recall but had a larger recall gap between the training and validation sets compared to Gradient Boosting and AdaBoost.

Among the five models, Gradient Boosting and AdaBoost emerged as the best performers, demonstrating a strong ability to generalize to unseen data while maintaining high recall values on the validation set. These models balance model complexity and performance, making them ideal for identifying "denied" cases where minimizing false negatives is critical.

6. Model Performance Improvement Using Hyperparameter Tuning:

6.1. Selection of Models for Tuning with Justification:

Random Forest, Gradient Boosting, and XGBoost were selected for hyperparameter tuning due to their effectiveness with imbalanced datasets and improving recall scores. Random Forest was chosen for its diverse decision trees and overfitting reduction. Gradient Boosting's iterative learning enhances model strength, while XGBoost offers computational efficiency, regularization, and better handling of class imbalances. These models were expected to significantly improve recall post-tuning.

6.2.Hyperparameter Optimization Using Randomized Search:

The selected models were fine-tuned using randomized search, focusing on key hyperparameters. For Random Forest, parameters like the number of estimators (300), maximum depth (10), and minimum samples per split (5) were optimized. Gradient Boosting showed optimal performance with 200 estimators, a maximum depth of 3, and a learning rate of 0.1. XGBoost achieved its best results with 200 estimators, a maximum depth of 10, a learning rate of 0.001, a subsample ratio of 0.7, and a column sample by tree ratio of 0.7. The recall metric was prioritized to enhance the models' sensitivity to detecting positive cases. This tuning process yielded configurations tailored to maximize performance on the oversampled dataset.

6.3.Evaluation of Performance for Optimized Models:

The tuned models were evaluated on the validation dataset using recall as the primary metric. XGBoost achieved the highest recall score of 0.8625, demonstrating its superior ability to identify positive cases effectively. Gradient Boosting followed with a recall of 0.8361, showcasing a consistent and reliable improvement in its performance. Random Forest, with a recall score of 0.8278, also exhibited a significant enhancement compared to its untuned counterpart. These results indicate that hyperparameter tuning significantly improved the models’ effectiveness, with XGBoost emerging as the top-performing model in terms of recall. The improvements underscore the importance of tuning in refining model performance for imbalanced datasets.

7. Model Performance Comparison and Final Model Selection:

7.1.Analyze the Performance Metrics of Tuned Models:

The performance of the three tuned models was compared using the recall metric on both the validation and test datasets. On the validation dataset, the recall scores were 0.8278, 0.8361, and 0.8625 for the Random Forest, Gradient Boosting, and XGBoost models, respectively. When evaluated on the test dataset, the recall scores improved slightly for all three models. The Random Forest achieved a recall of 0.8425, Gradient Boosting attained 0.8420, and XGBoost stood out with a recall of 0.8696. These results indicate that all models generalize well to unseen data, with XGBoost consistently outperforming the others.

7.2.Select the Optimal Model for Deployment:

Based on the test dataset performance, XGBoost was identified as the best-performing model. It achieved the highest recall score of 0.8696, which reflects its ability to correctly classify the majority of positive cases in the dataset. This consistent performance across both validation and test datasets demonstrates the model's robustness and effectiveness, making it the most suitable choice for deployment.

7.3.Evaluate and Interpret the Final Model's Test Results:

The selected XGBoost model was further analyzed to assess its performance. With a recall score of 0.8696 on the test set, the model effectively minimizes false negatives, which is critical for the given problem. This performance metric aligns with the project's objective of accurately predicting the target variable while ensuring minimal misclassification of positive cases. The XGBoost model's fine-tuned parameters, such as a learning rate of 0.001, a maximum depth of 10, and a subsample rate of 0.7, contributed to its superior performance, demonstrating the impact of hyperparameter tuning.

8. Actionable Insights & Recommendations

8. 1. Insights from the Analysis Conducted:

Through the data analysis and model-building phases, several key insights were uncovered that can significantly impact the decision-making process for visa approval. The data revealed that certain factors, such as the applicant's job experience, education level, and wage offered, are strongly correlated with the approval status of their visa applications. Applicants from regions with higher wages and stronger job experience tend to have higher approval rates. Additionally, the analysis indicated that companies hiring for positions with a larger number of employees or those offering full-time positions have a better chance of receiving visa approvals. Further, we observed that candidates with higher wages and those applying for specialized roles were more likely to receive certifications. The machine learning models, particularly the XGBoost classifier, demonstrated high accuracy in predicting the approval status based on these insights, highlighting the importance of these features in determining outcomes.

8. 2. Actionable Business Recommendations:

Based on the analysis and the predictive models, the following actionable recommendations can be made for the Office of Foreign Labor Certification (OFLC) to streamline the visa approval process and ensure better outcomes:

1. Focus on High-Wage Roles and Specialized Job Titles: Since applicants with higher wages and specialized roles show higher approval rates, it is recommended that the OFLC prioritize processing cases for positions that offer competitive wages and demand specialized skill sets. This can help reduce bottlenecks and ensure that highly qualified candidates are not delayed.

2. Emphasize Job Experience and Education Requirements: The analysis suggests that applicants with more job experience and higher educational qualifications are more likely to be approved. Therefore, companies should be encouraged to provide detailed profiles of the applicants' qualifications and experience, which could improve the chances of approval.

3. Optimize Full-Time and Large Employer Applications: Companies with larger workforces and those hiring for full-time positions generally have a higher chance of

successful certification. The OFLC could expedite processing for these cases to prioritize companies that are committed to long-term, stable employment opportunities.

4. Utilize Predictive Models for Case Prioritization: By implementing the machine learning models to predict the likelihood of visa approval, the OFLC can automatically flag high-priority cases, allowing the agency to focus resources on applications with higher chances of success. This would also enable a more efficient and timely review process.

5. Regular Review of Policy Changes: It is essential for the OFLC to periodically update the model as external factors like wage trends, employment patterns, and policy changes evolve. This will ensure the decision-making process remains relevant and responsive to the changing dynamics of the labor market.

By following these recommendations, the OFLC can improve the efficiency of the visa approval process, reduce processing times, and increase the accuracy of decisions, benefiting both businesses and potential foreign workers.