Visa Approval Prediction Report

Presented by :
Sanjay Rajan J

TABLE OF CONTENTS

LIST OF FIGURES

1.DATA OVERVIEW

CONTEXT

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

OBJECTIVE

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

1. Facilitate the process of visa approvals.
2. Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

DATA DICTIONARY:

- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? Y = Yes; N = No
- no_of_employees: Number of employees in the employer's company
- yr_of_estab: Year in which the employer's company was established

- region_of_employment: Information of foreign worker's intended region of employment in the US.
- prevailing_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- full_time_position: Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
- case_status: Flag indicating if the Visa was certified or denied

➢ Shape:

There are 25480 rows and 12 columns in this dataset.

➢ Duplicates:

There are no duplicate entries in this dataset.

➢ Basic Info:

Fig.1 Data Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25480 entries, 0 to 25479
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   continent              25480 non-null  object
 1   education_of_employee  25480 non-null  object
 2   has_job_experience     25480 non-null  object
 3   requires_job_training  25480 non-null  object
 4   no_of_employees        25480 non-null  int64
 5   yr_of_estab            25480 non-null  int64
 6   region_of_employment   25480 non-null  object
 7   prevailing_wage        25480 non-null  float64
 8   unit_of_wage           25480 non-null  object
 9   full_time_position     25480 non-null  object
 10  case_status            25480 non-null  object
dtypes: float64(1), int64(2), object(8)
memory usage: 2.1+ MB
```

- Null values Check:

Fig.2 Null values check

```
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
no_of_employees          0
yr_of_estab              0
region_of_employment     0
prevailing_wage          0
unit_of_wage             0
full_time_position       0
case_status              0
dtype: int64
```

- Numerical Statistics:

Fig.3 Numerical Statistics

	no_of_employees	yr_of_estab	prevailing_wage	case_status
count	25447.000000	25447.000000	25447.000000	25447.000000
mean	5674.415334	1979.394506	74468.281479	0.668094
std	22891.842245	42.385932	52822.177370	0.470907
min	12.000000	1800.000000	2.136700	0.000000
25%	1025.000000	1976.000000	34039.210000	0.000000
50%	2112.000000	1997.000000	70312.500000	1.000000
75%	3506.500000	2005.000000	107739.505000	1.000000
max	602069.000000	2016.000000	319210.270000	1.000000

- No. of Unique values:

Fig.4 Unique Values

```
case_id : 25480
continent : 6
education_of_employee : 4
has_job_experience : 2
requires_job_training : 2
no_of_employees : 7105
yr_of_estab : 199
region_of_employment : 5
prevailing_wage : 25454
unit_of_wage : 4
full_time_position : 2
case_status : 2
```

2. EXPLORATORY DATA ANALYSIS

2.1 UNIVARIATE ANALYSIS:

Fig.5 No_of_employees distribution

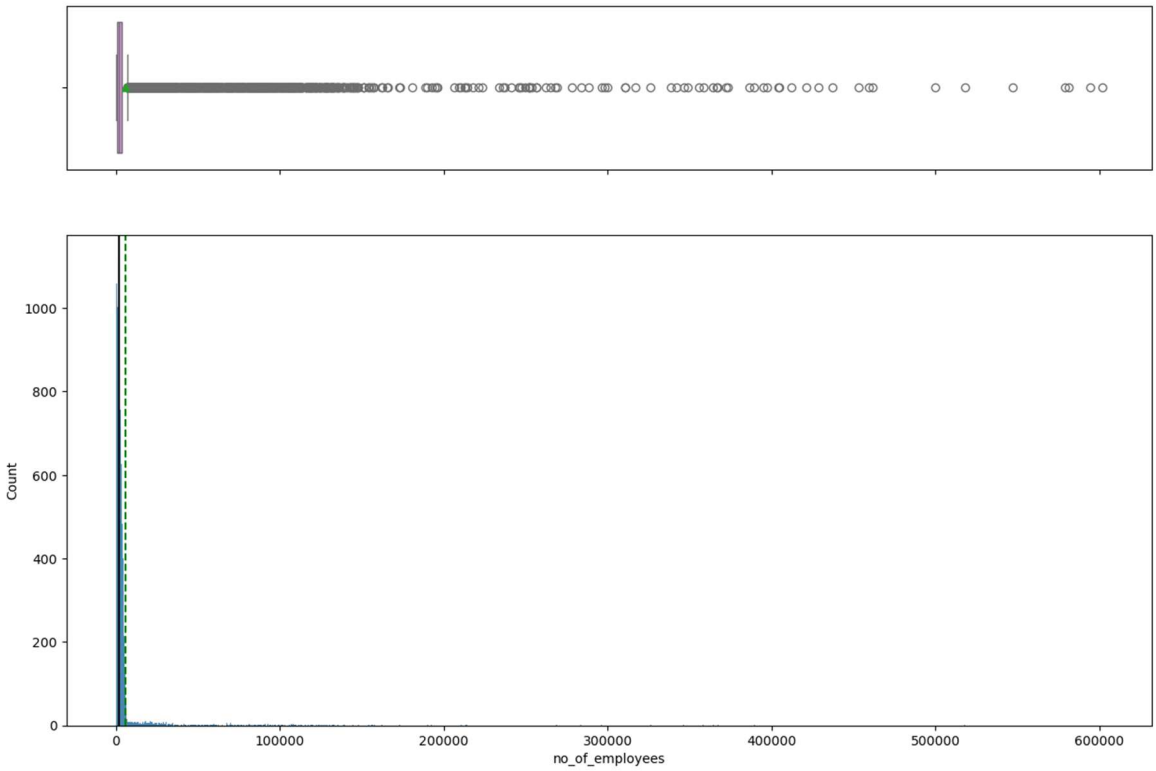


Fig.6 yr_of_estab distribution

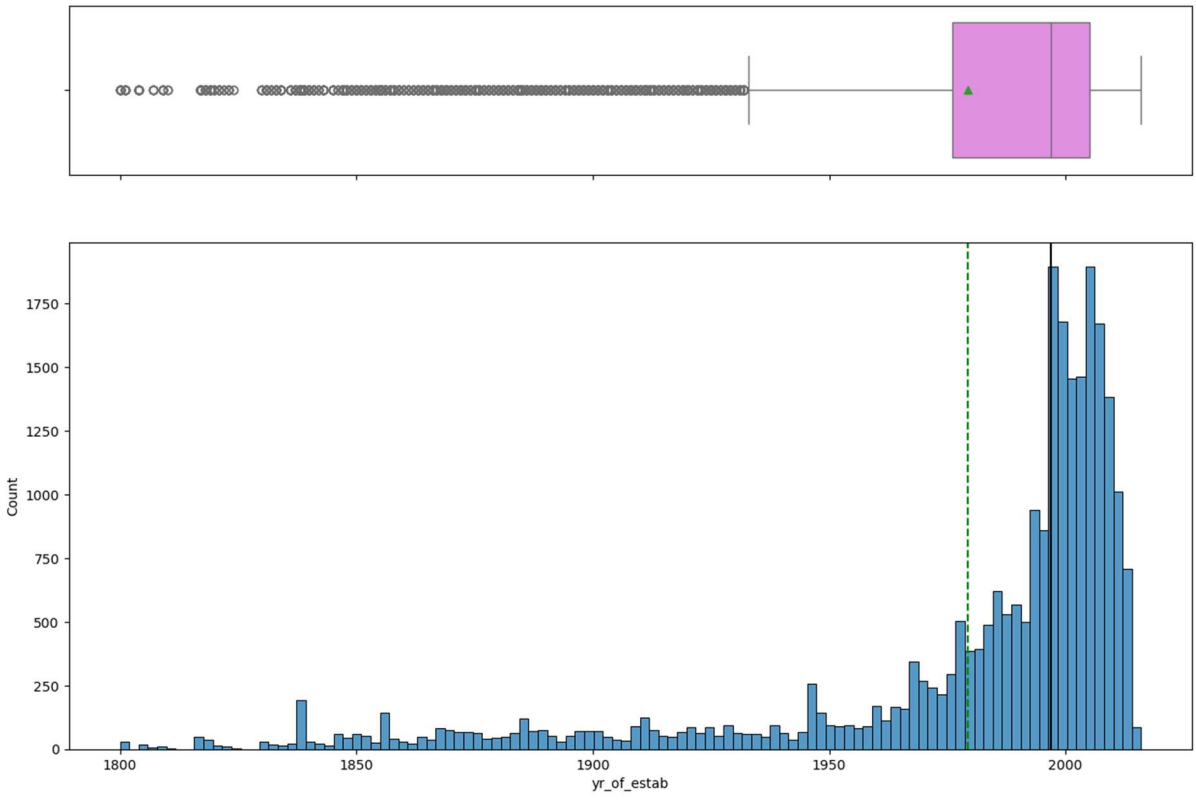


Fig.7 prevailing_wage distribution

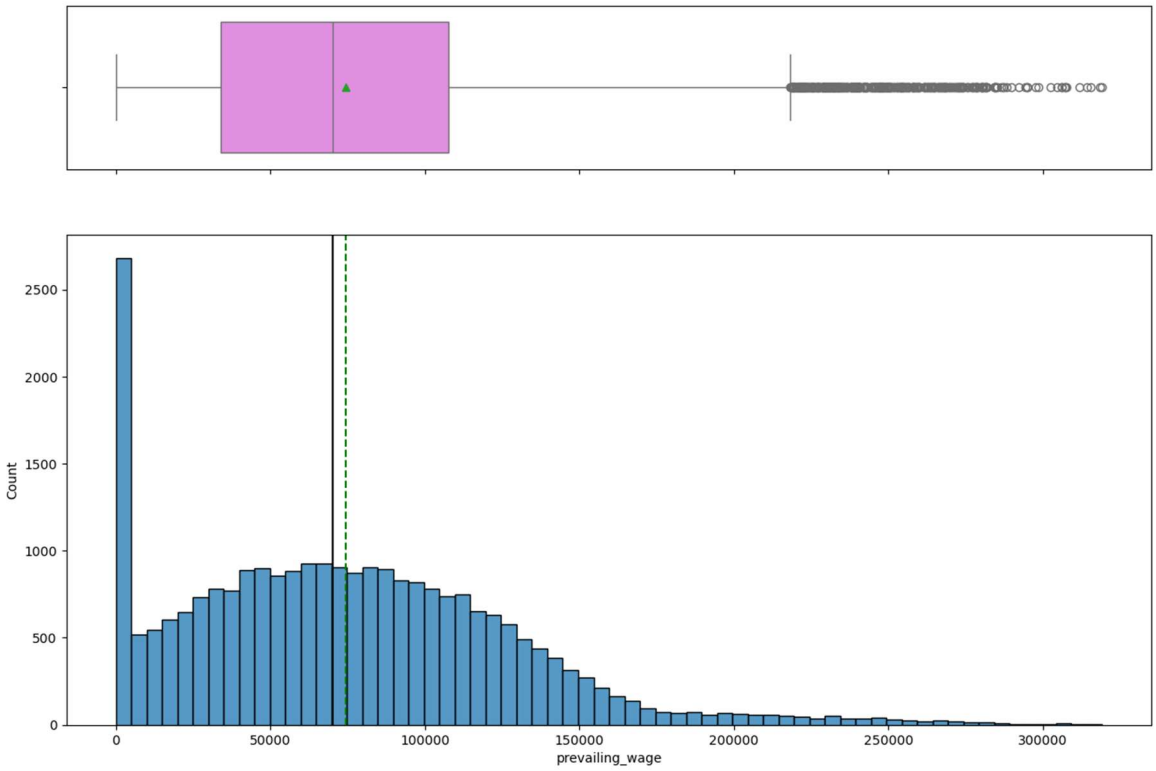


Fig.8 continent distribution

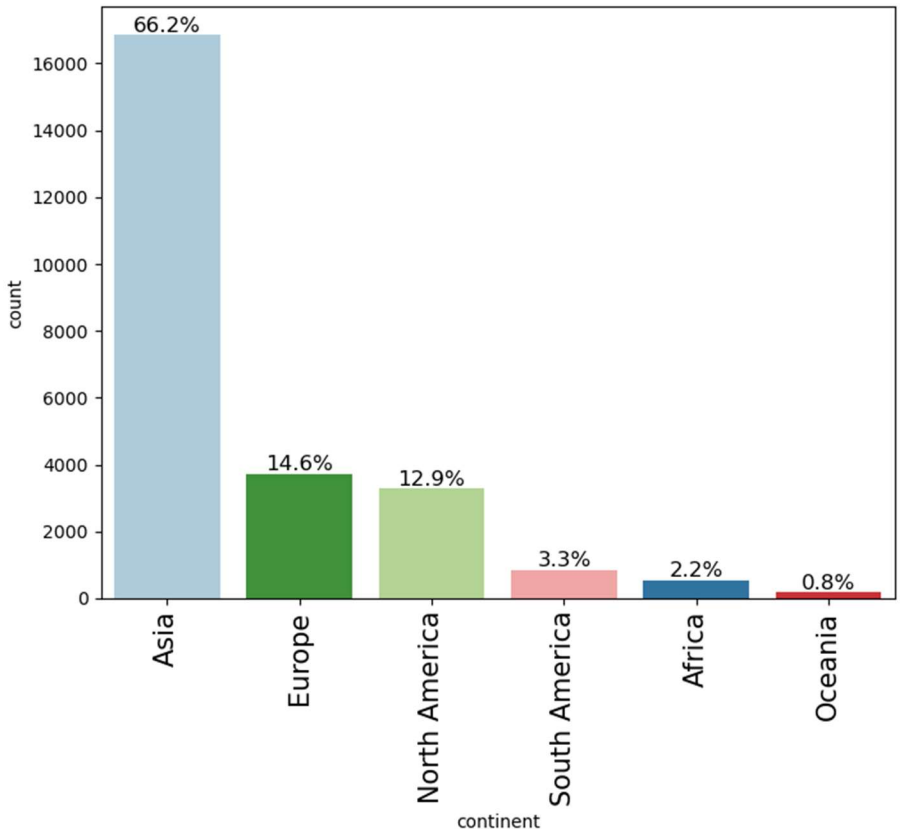


Fig.9 education_of_employee distribution

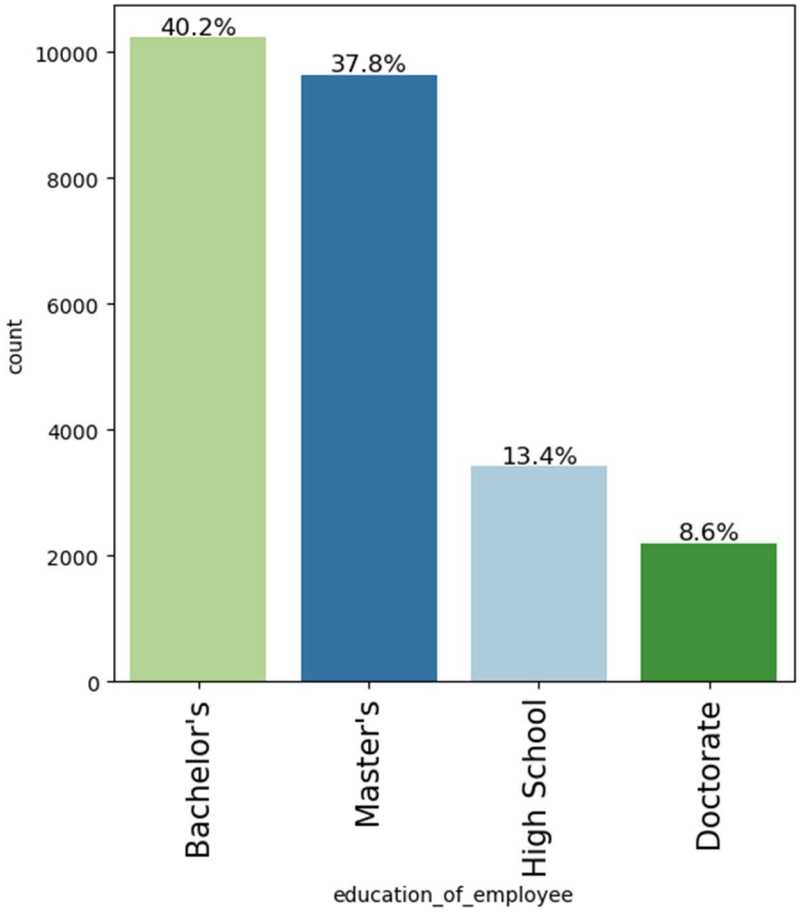


Fig.10 has_job_experience distribution

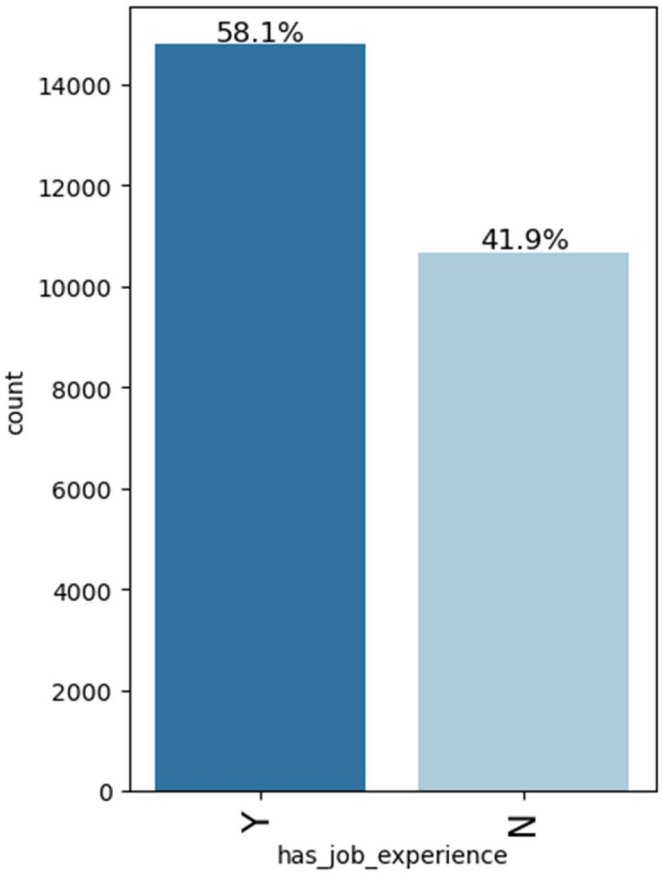


Fig.11 requires_job_training distribution

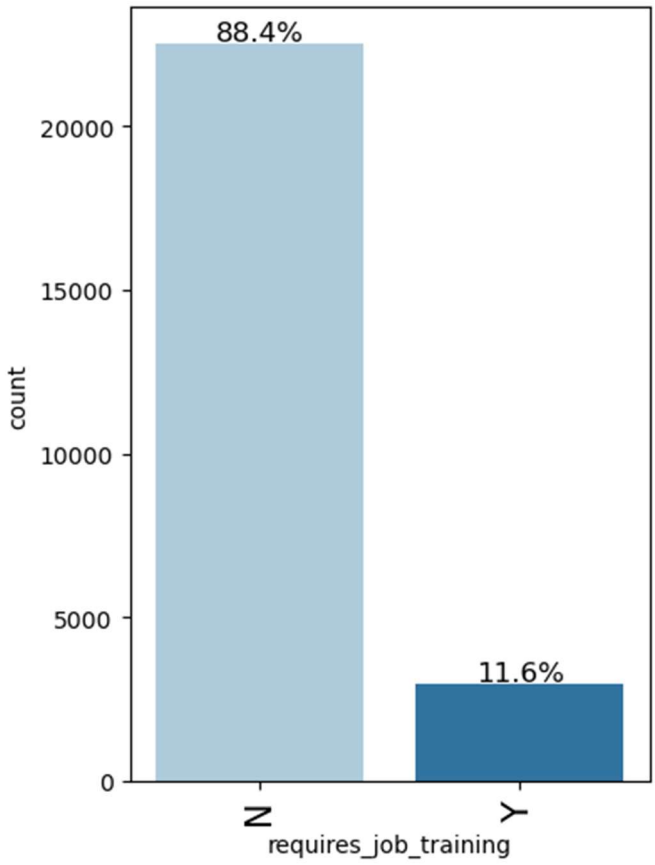


Fig.12 region_of_employment distribution

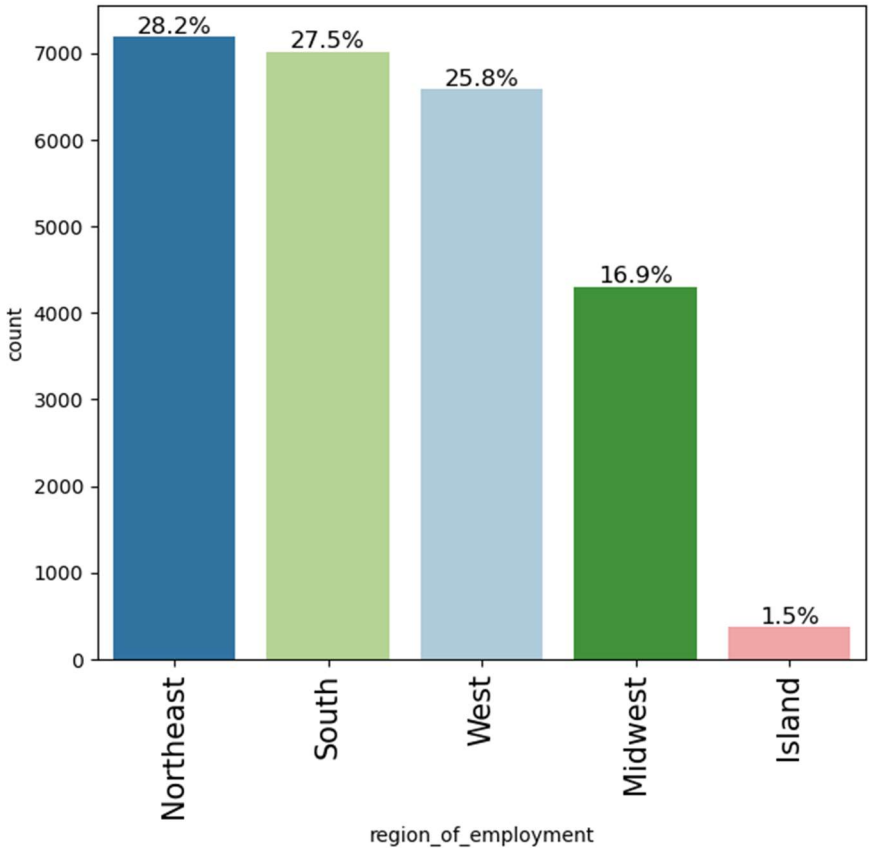


Fig.13 unit_of_wage distribution

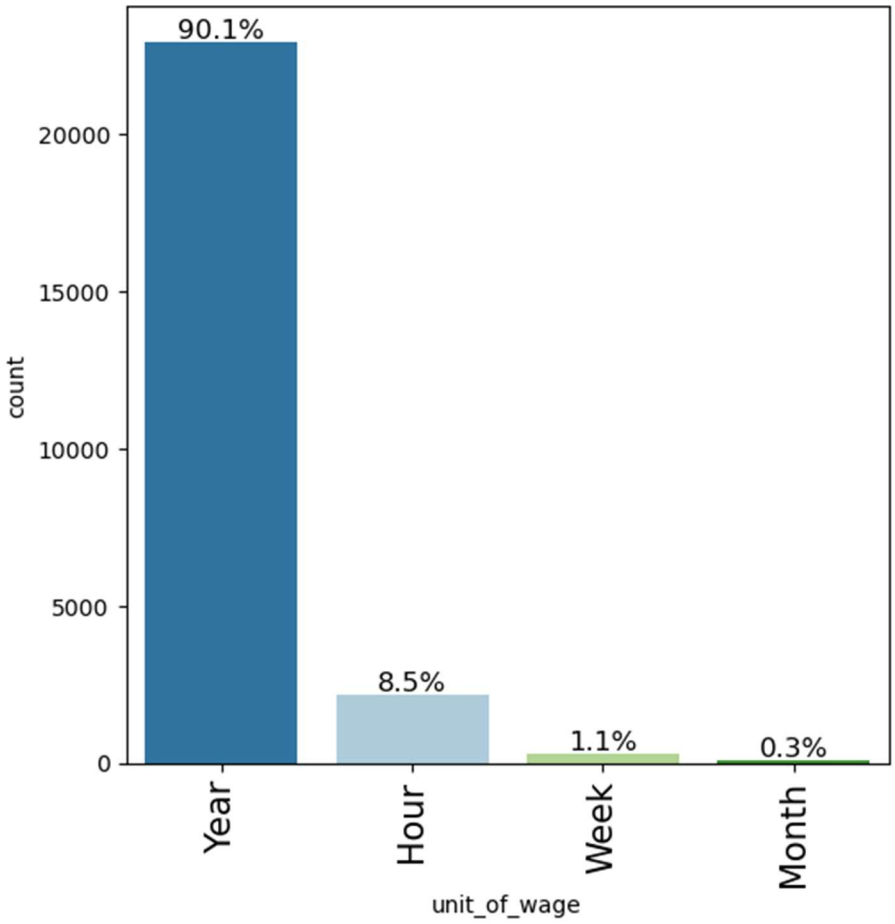


Fig.14 full_time_position distribution

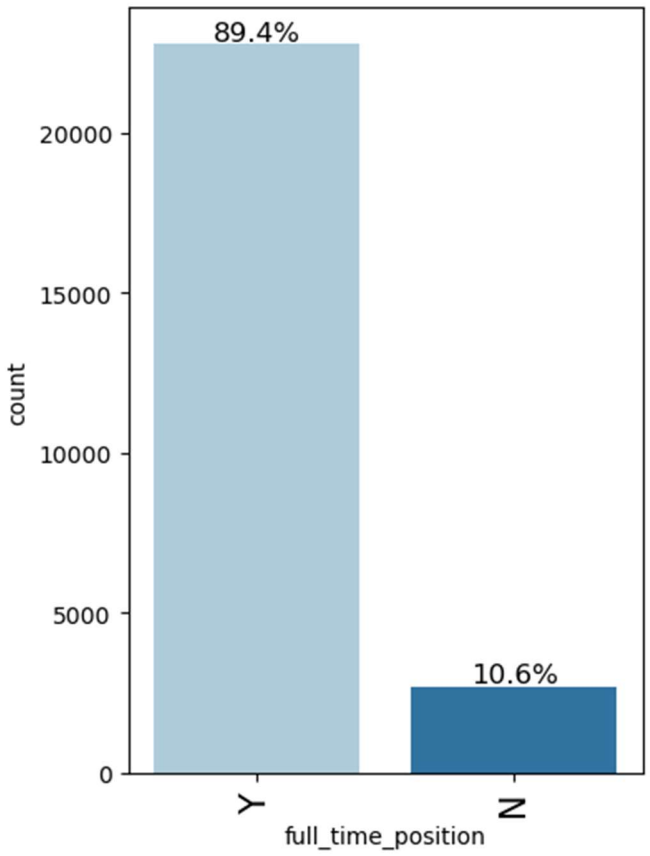


Fig.15 case_status distribution

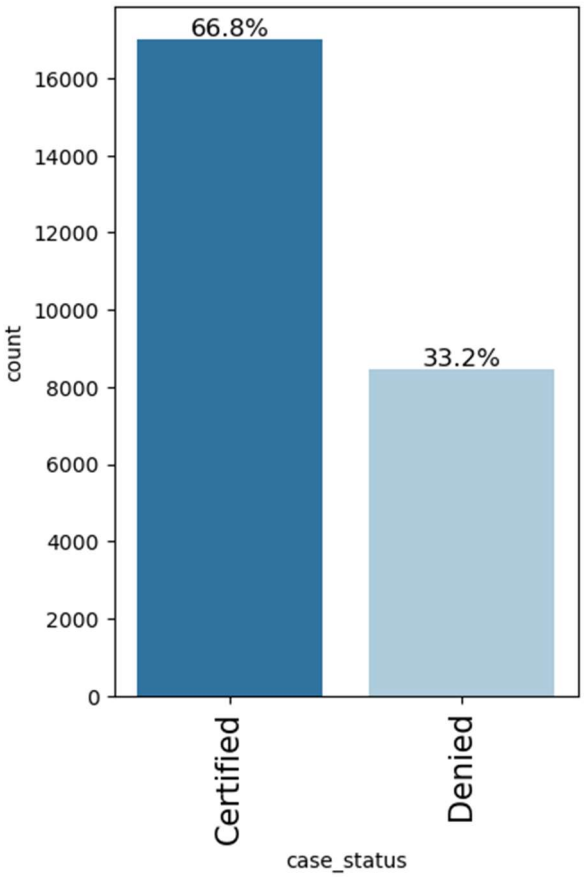


Fig.16 Estb. Yr Bin distribution

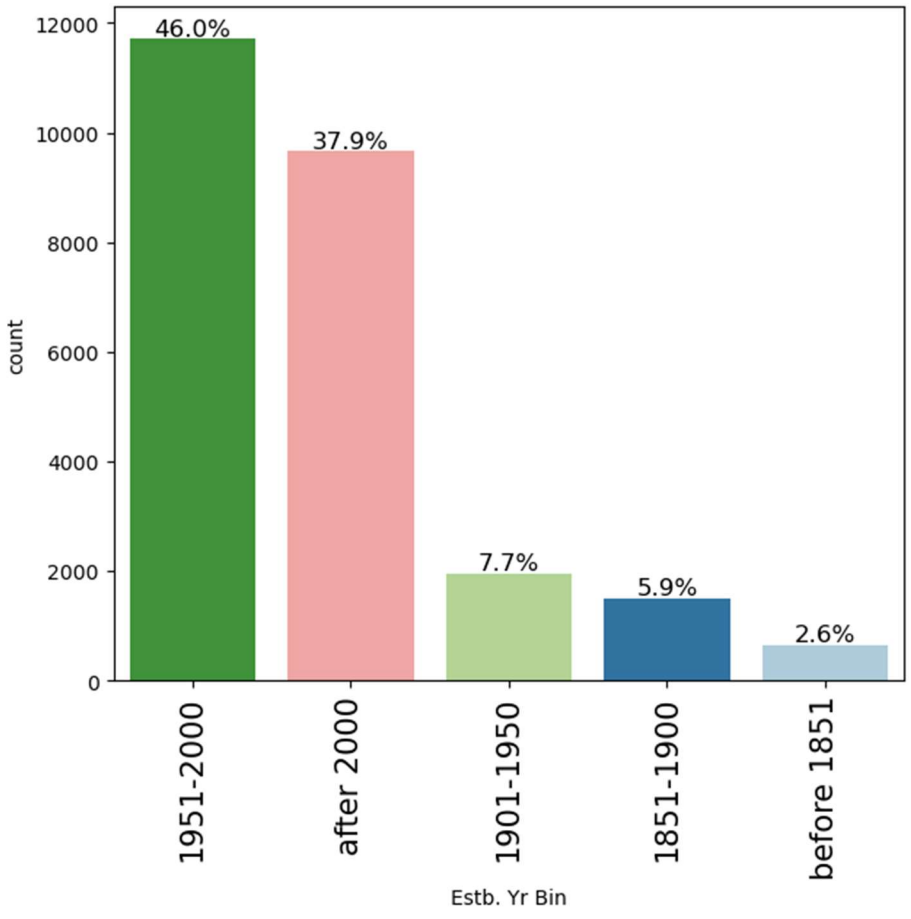
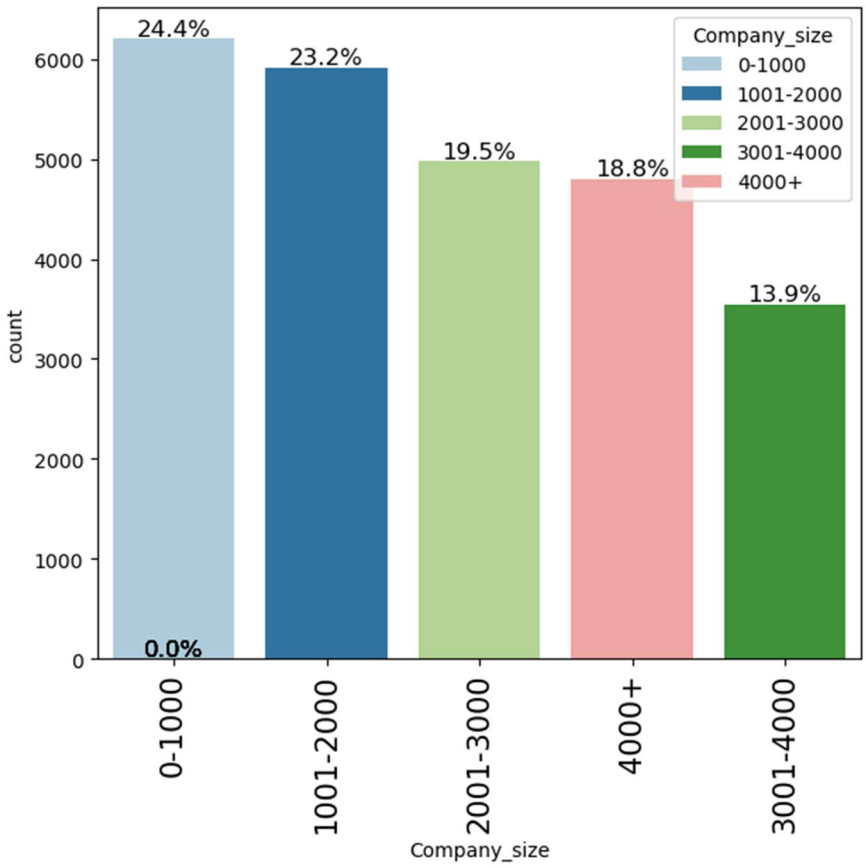


Fig.17 Company Size distribution



COMMENTS:

- 66.8% of the cases are certified and 33.2% are denied.
- Majority of the employees have a Bachelor's degree (40.2%) followed by Master's degree.
- The Northeast region has the highest employment share at 28.2%, followed closely by the South (27.5%) and West (25.8%), while the Island region has the lowest share at just 1.5%.
- The majority of wages are measured on a yearly basis, accounting for 90.1%.

- Majority of the employees are from Asia (66.2%) followed by Europe(14.6%) and North America(12.9%).
- The prevailing wage distribution is right-skewed, indicating that most wage values are concentrated on the lower end.
- The yr_of_Estab distribution is heavily left-skewed.
- Majority of the employees (46%) came from the companies established between 1951-200.
- The continent Oceania has the least employee count with only 0.8%

2.2 BIVARIATE ANALYSIS:

Fig.18 Heatmap

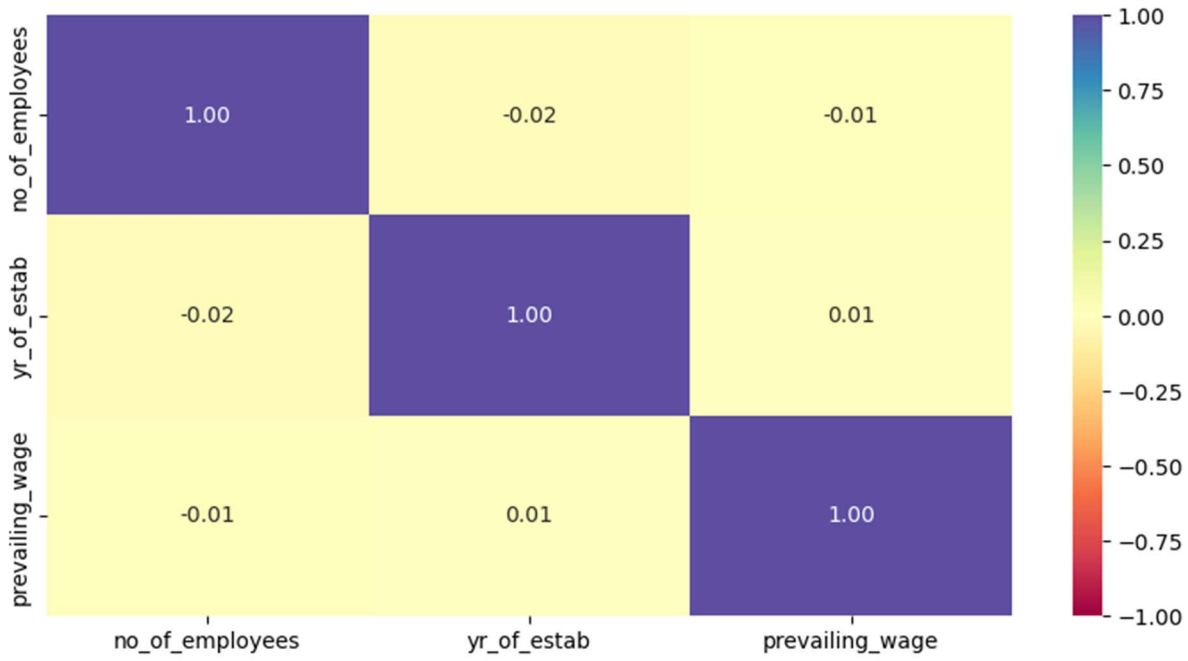


Fig.19 region_of_employment vs prevailing_wage

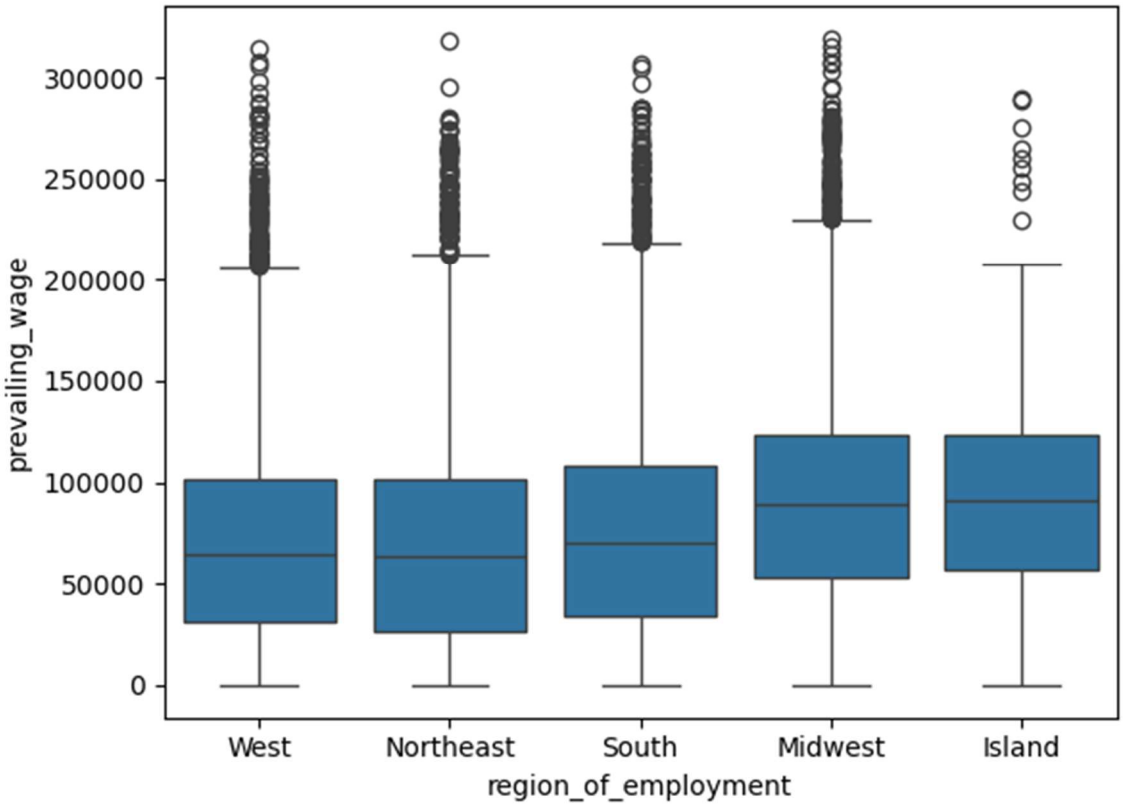


Fig.20 continent vs prevailing_wage

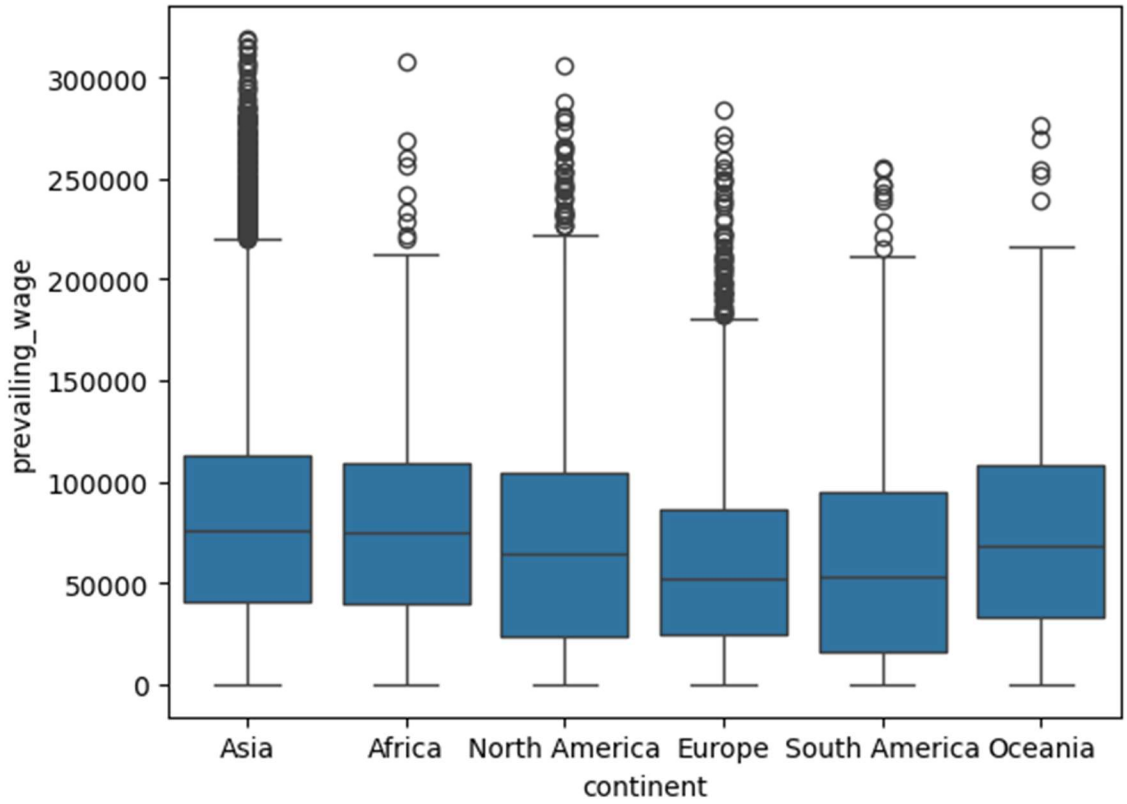


Fig.21 education_of_employee vs prevailing_wage

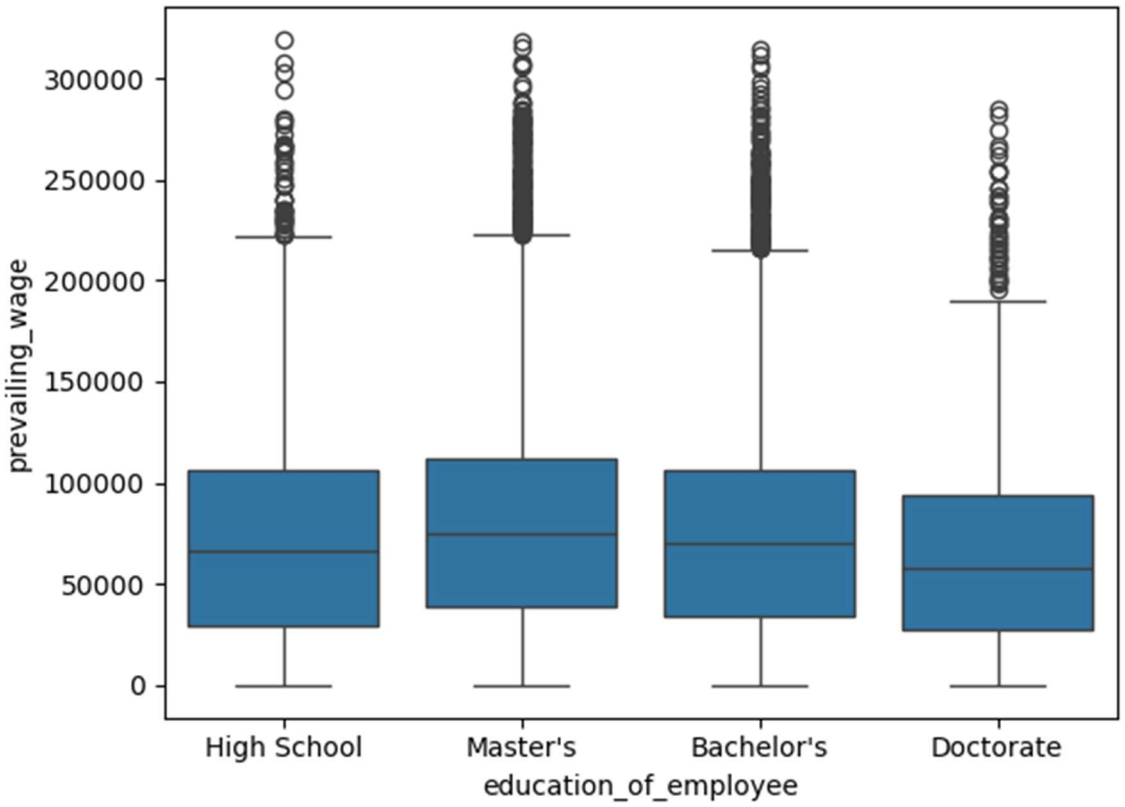


Fig.22 unit_of_wage vs prevailing_wage

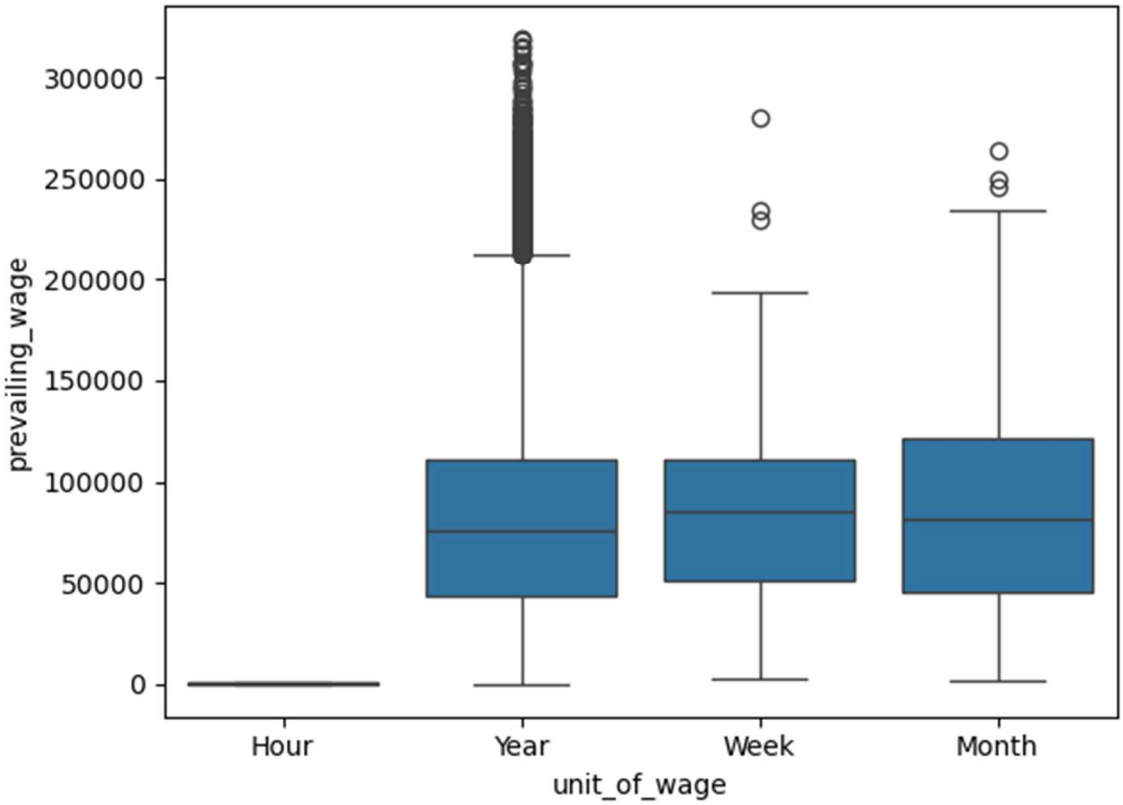


Fig.23 Case_status vs continent

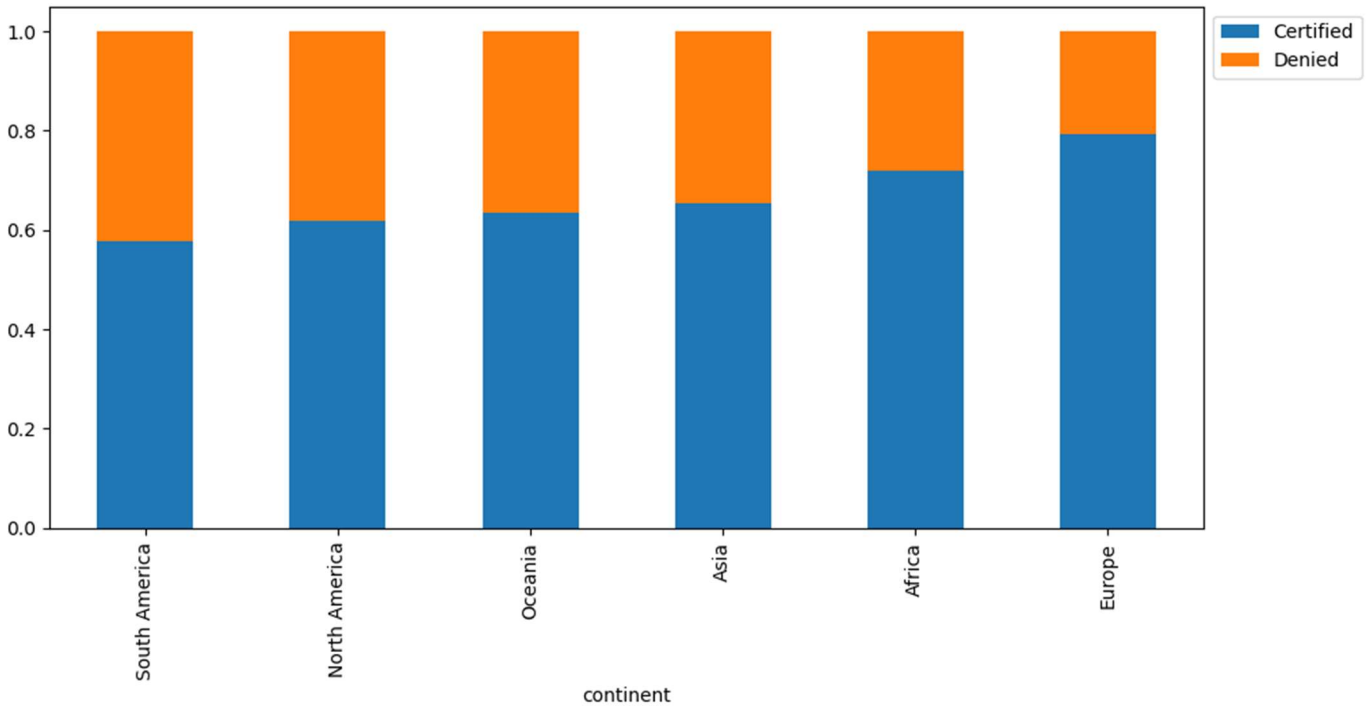


Fig.24 Case_status vs education_of_employee

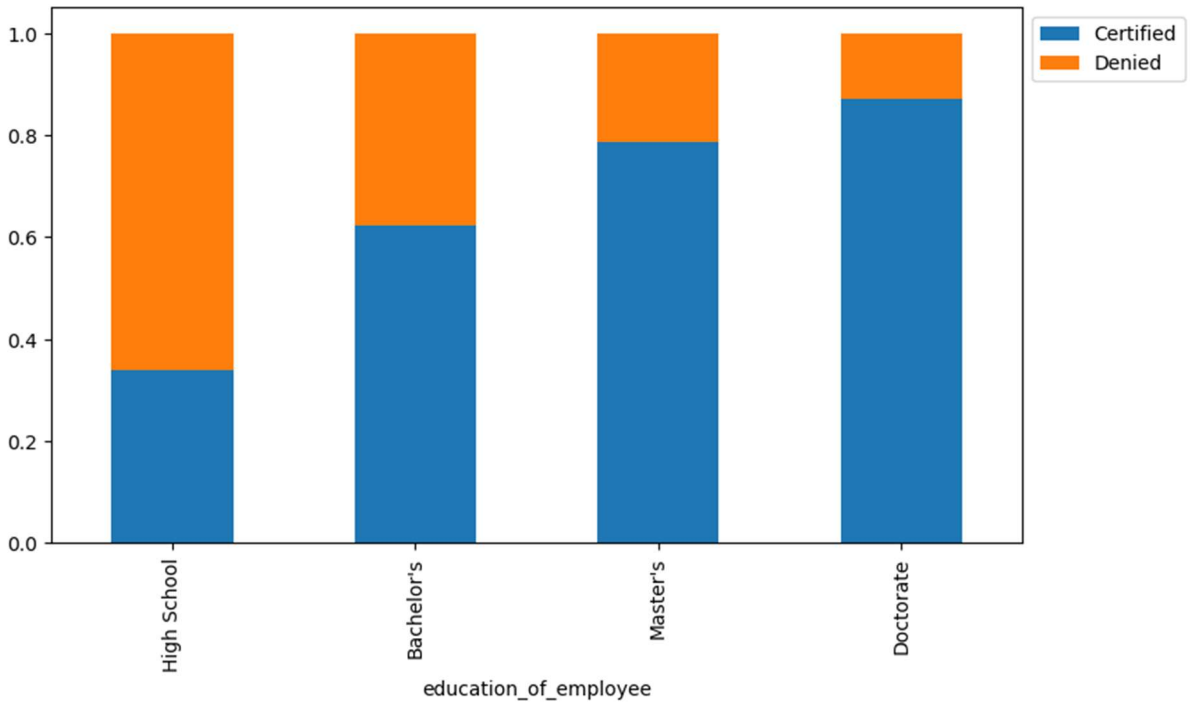


Fig.25 Case_status vs has_job_experience

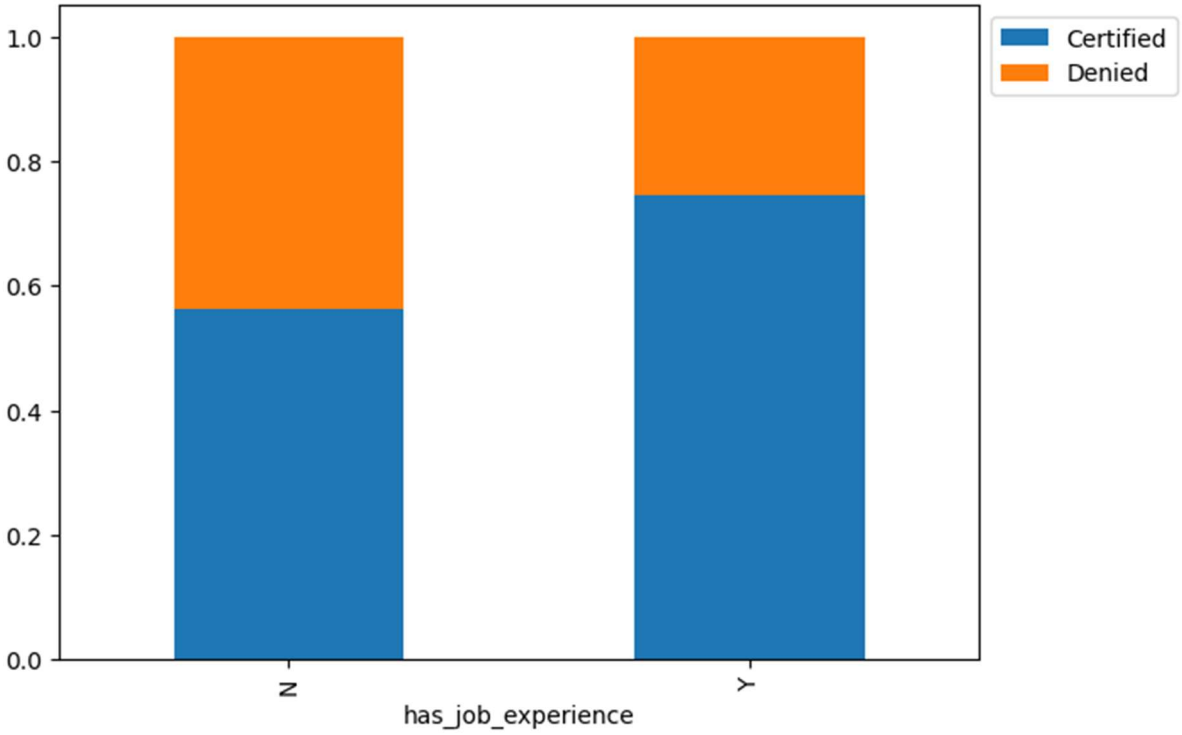


Fig.26 Case_status vs requires_job_training

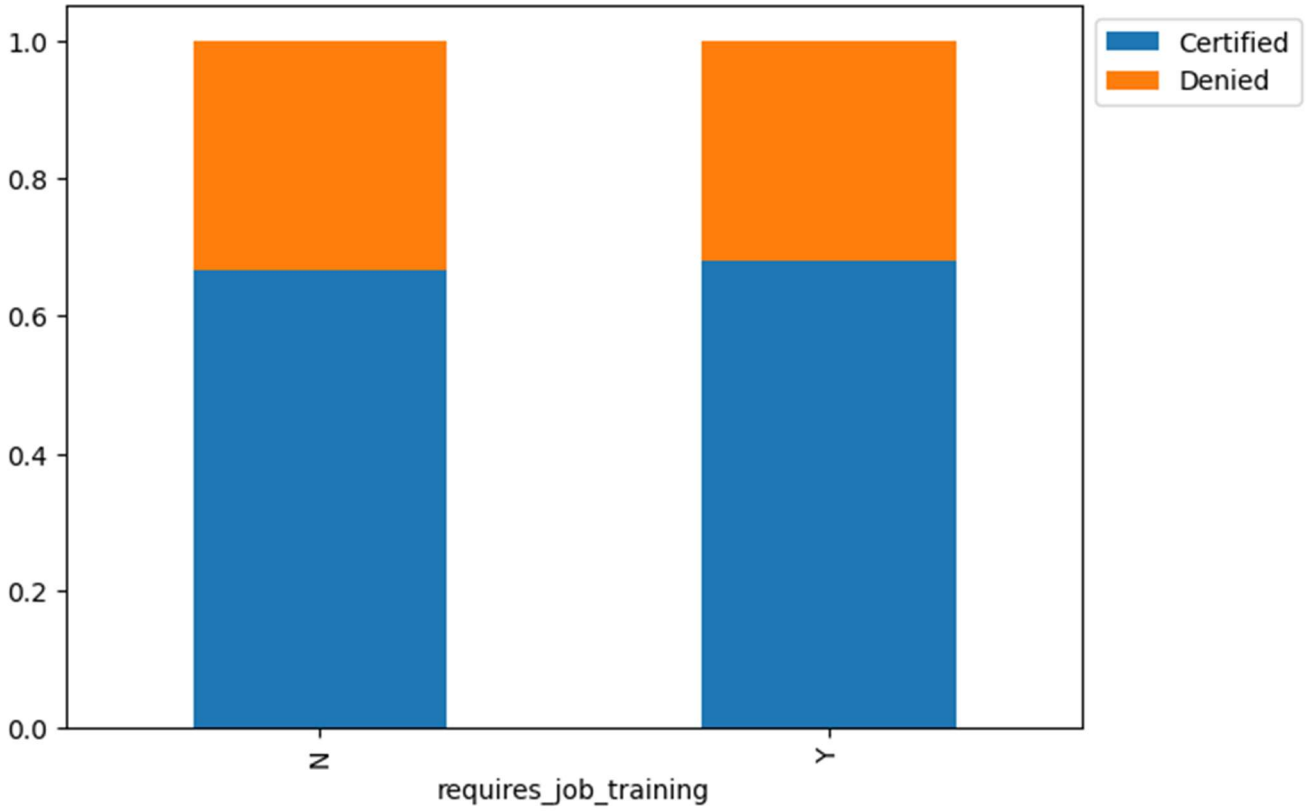


Fig.27 Case_status vs region_of_employment

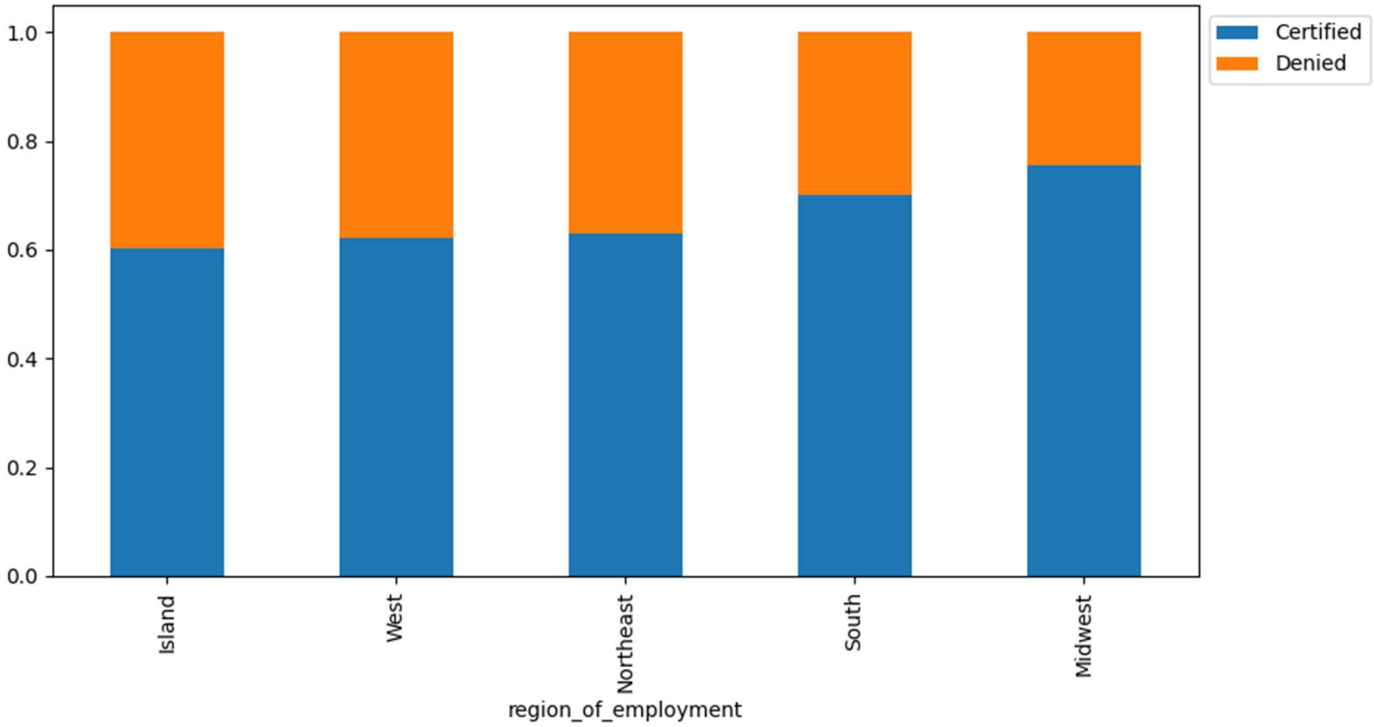


Fig.28 Case_status vs unit_of_wage

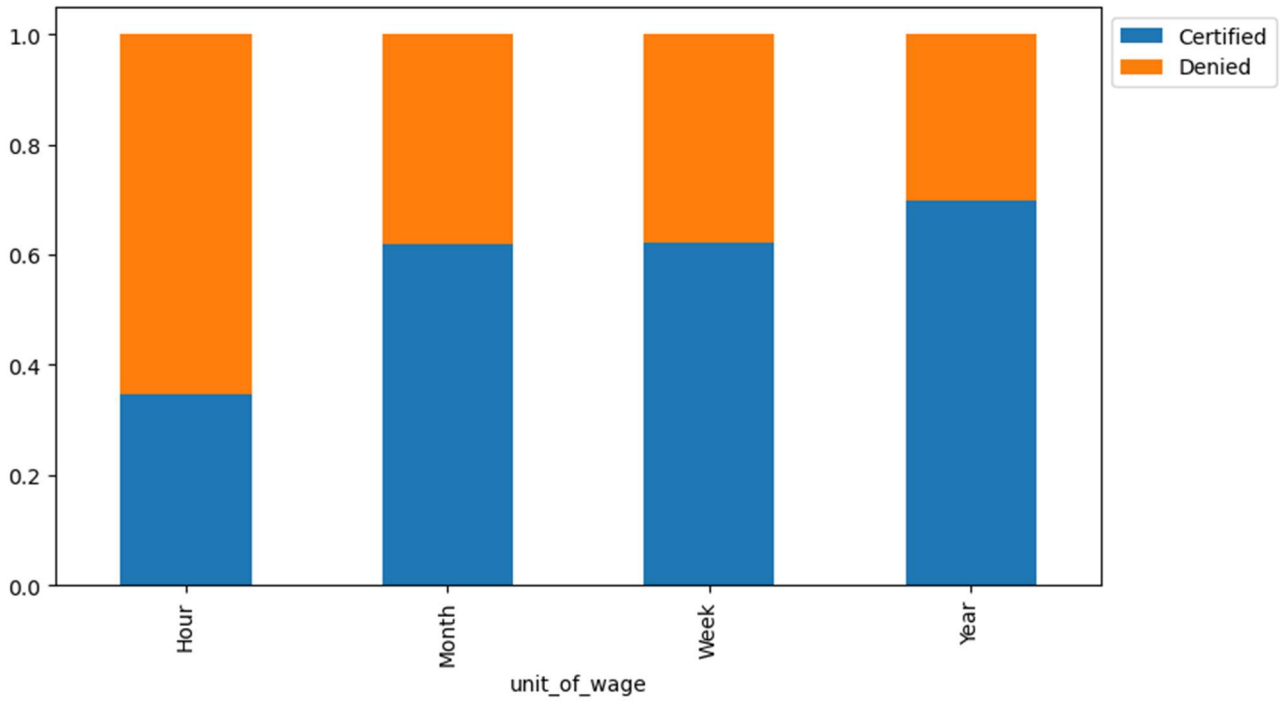


Fig.29 Case_status vs full_time_position

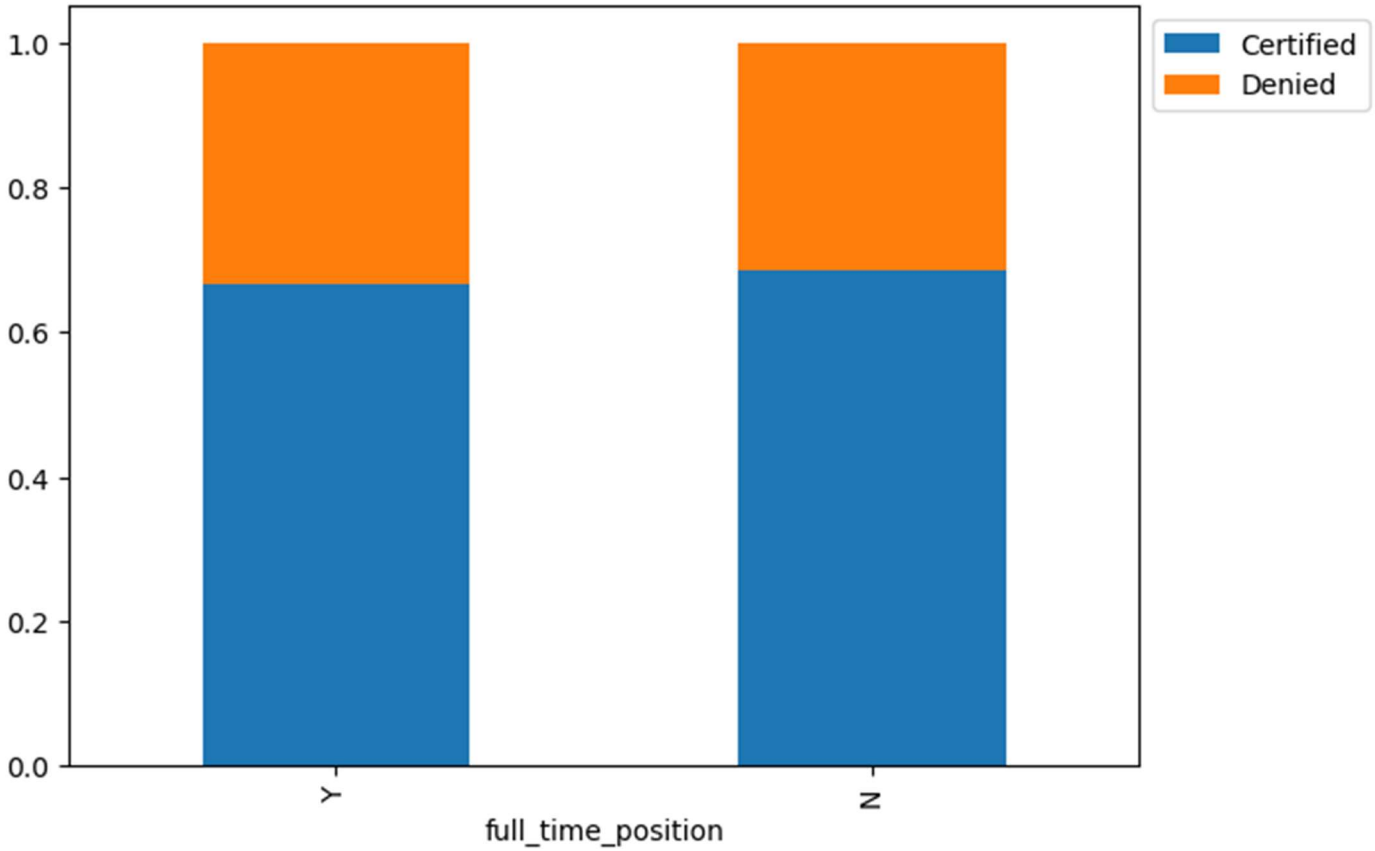


Fig.30 No_of_employees w.r.t Case_status

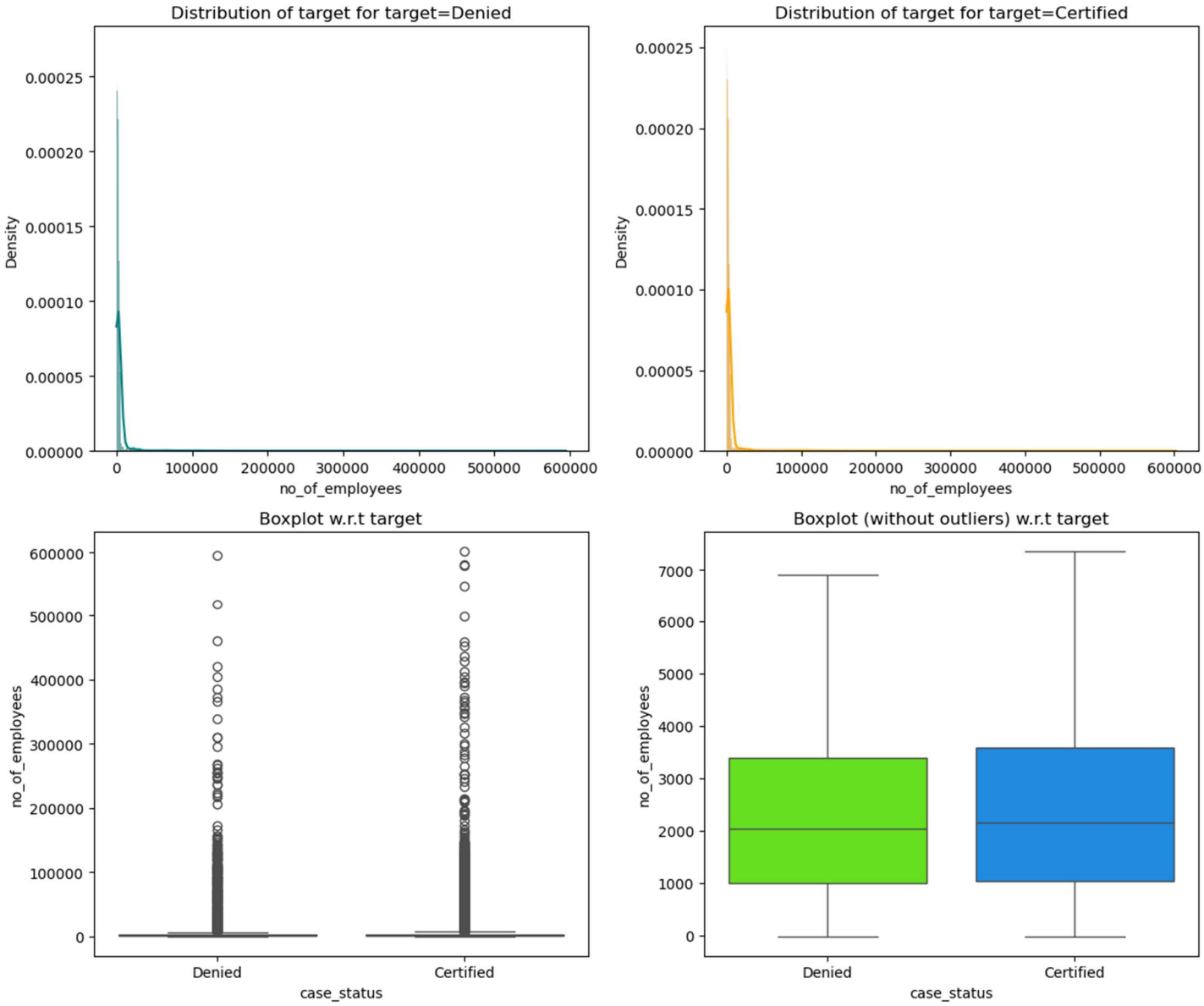
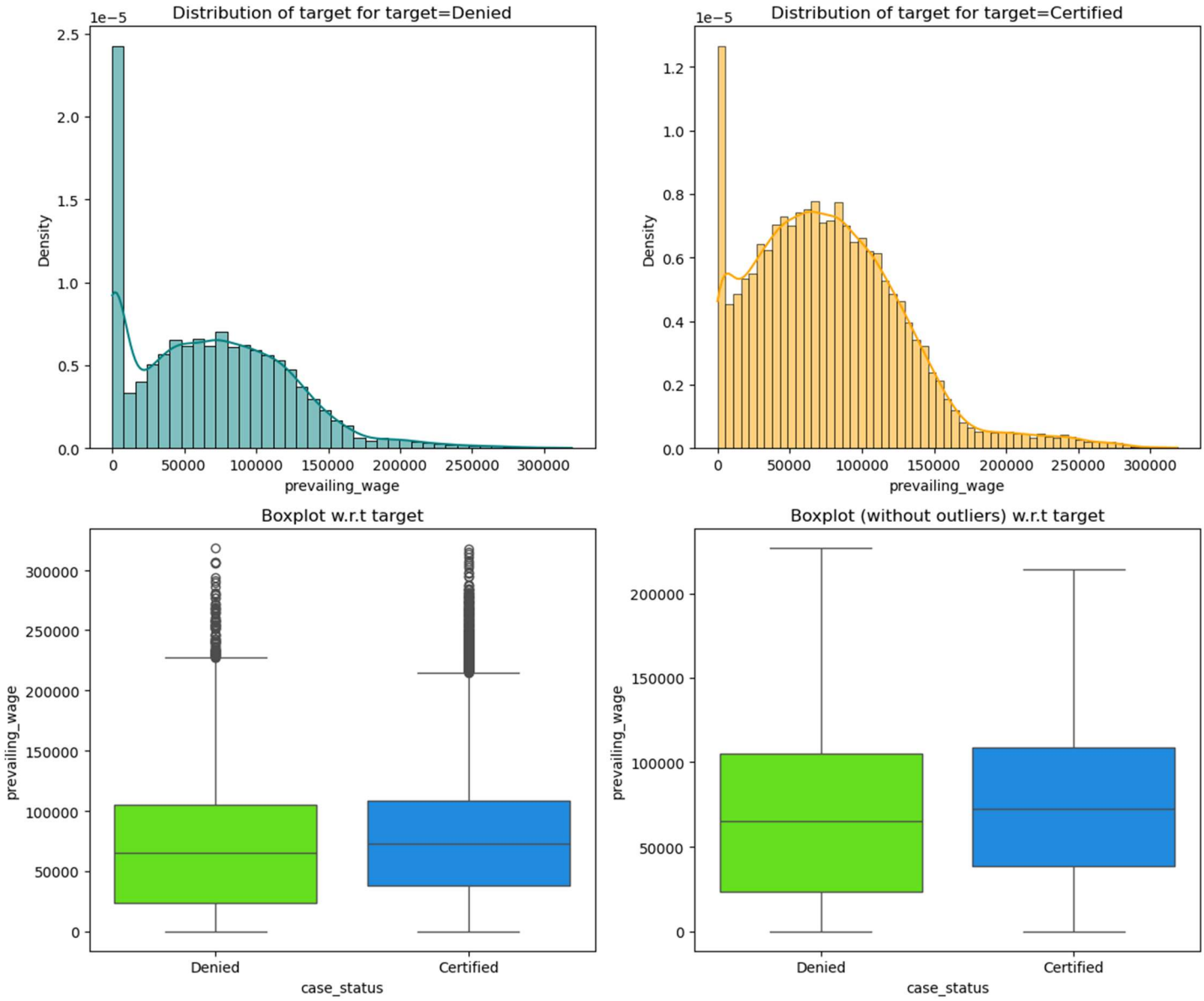


Fig.31 Prevailing_wage w.r.t Case_status



COMMENTS:

- There isn’t any correlation between the numerical variables.
- The median prevailing wage is similar across education levels, with a slight increase for Master's and Bachelor's degree holders compared to High School and Doctorate levels.
- The Midwest and Island have higher median of prevailing wages.
- Asia and North America have slightly higher prevailing wage median.
- As the education of the employee increases, the chance of getting certified is increased.
- Europe has the highest proportion of Certified cases.
- South America has the lowest proportion of Certified cases.
- Employees with job experience are higher chance of getting certified.
- The proportions of cases is similar between full-time position cases.
- Employees with hour wages have lower chances of getting certified.
- Employees who preferred Midwest as their work region has the highest proportion of Certified cases.
- The median prevailing wage of the certified employees is around 65k.

3. DATA PREPROCESSING

Missing Value Treatment:

Fig.32 Null value checks

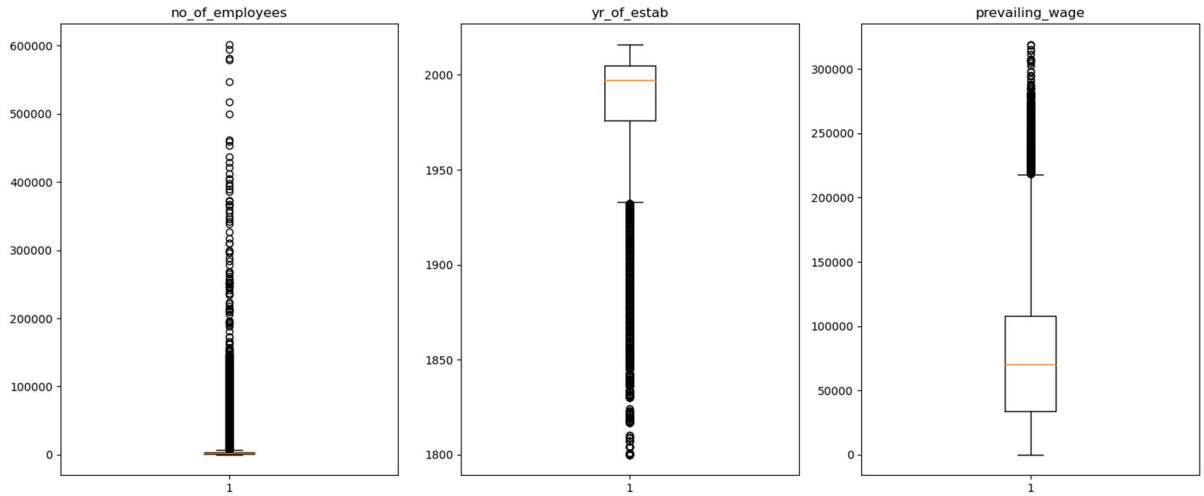
```
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
no_of_employees          0
yr_of_estab              0
region_of_employment     0
prevailing_wage          0
unit_of_wage             0
full_time_position       0
case_status              0
dtype: int64
```

There are no null values or missing values in this dataset.

Outlier Treatment:

- As we can clearly see, there are many outliers in the numerical columns of this dataset and they are meaningful.
- The columns “No_of_employees” and “yr_of_estab” are categorized and “prevailing_wage” is not treated.

- **Fig.33 Outlier checks**



Feature Engineering:

- The “case_id column” is dropped, as it is not useful.
- There are 33 rows which has negative entries of no_of_employees and are dropped.
- The numerical variable “**yr_of_estab**” was categorized into a new feature, “**Estb. Yr Bin**”, by grouping years into meaningful historical periods: **before 1851**, **1851-1900**, **1901-1950**, **1951-2000**, and **after 2000**. This binning simplifies analysis and provides intuitive groupings for better interpretation.
- The “**no_of_employees**” variable was transformed into a categorical feature, “**Company_size**”, by binning employee counts into ranges: **0-1000**, **1001-2000**, **2001-3000**, **3001-4000**, and **4000+**. This simplifies analysis by grouping companies based on size.
- The target variables are labelled as 0 and 1 (Denied: 0, Certified: 1)
- Data is splitted in the ratio of 6:2:2 for training, validation and testing.
- One-hot Encoding is used for 'continent', 'education_of_employee', 'has_job_experience', 'requires_job_training', 'unit_of_wage', 'region_of_employment', 'full_time_position', 'Estb. Yr Bin', 'Company_size'.

4. MODEL BUILDING – ORIGINAL DATA

Metric of choice:

Minimize Missed Approvals: High recall ensures capturing nearly all potential visa approvals, reducing the risk of missing eligible candidates. Prioritizing recall helps in identifying true positives, improving the approval process, and optimizing resource usage.

Fig.34 Model performances (Original data)

```
Training Performance:

Bagging: 0.9804882831650161
Random forest: 1.0
GBM: 0.8721443278752818
Adaboost: 0.8857731150112756
XGB: 0.9210706932052162

Validation Performance:

Bagging: 0.7741840635107321
Random forest: 0.7862393413701853
GBM: 0.8762128785651279
Adaboost: 0.8912084680976183
XGB: 0.8579829461922964
```

Fig.35 Difference in Recall Scores 1

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9805, Validation Score: 0.7742, Difference: 0.2063
Random forest: Training Score: 1.0000, Validation Score: 0.7862, Difference: 0.2138
GBM: Training Score: 0.8721, Validation Score: 0.8762, Difference: -0.0041
Adaboost: Training Score: 0.8858, Validation Score: 0.8912, Difference: -0.0054
XGB: Training Score: 0.9211, Validation Score: 0.8580, Difference: 0.0631
```

- The Adaboost model and Gradient Boosting model performs better compared to other models with good recall scores and low differences.
- The Random Forest model tends to overfit and performs poorly in validation set.

5. MODEL BUILDING – OVERSAMPLED DATA

Fig.36 Data Count after Oversampling

```
Before Oversampling, counts of label 'Cerified': 10199
Before Oversampling, counts of label 'Denied': 5068

After Oversampling, counts of label 'Cerified': 10199
After Oversampling, counts of label 'Denied': 10199

After Oversampling, the shape of train_X: (20398, 27)
After Oversampling, the shape of train_y: (20398,)
```

Fig.37 Model performances (Oversampled)

```
Training Performance:

Bagging: 0.9804882831650161
Random forest: 0.9999019511716835
GBM: 0.8680262770859888
Adaboost: 0.8517501715854495
XGB: 0.9123443474850476

Validation Performance:

Bagging: 0.7650690973243164
Random forest: 0.7827109673625404
GBM: 0.8709203175536607
Adaboost: 0.8576889150249927
XGB: 0.8571008526903852
```

Fig.38 Difference in Recall Scores 2

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9805, Validation Score: 0.7651, Difference: 0.2154
Random forest: Training Score: 0.9999, Validation Score: 0.7827, Difference: 0.2172
GBM: Training Score: 0.8680, Validation Score: 0.8709, Difference: -0.0029
Adaboost: Training Score: 0.8518, Validation Score: 0.8577, Difference: -0.0059
XGB: Training Score: 0.9123, Validation Score: 0.8571, Difference: 0.0552
```

- The dataset is over-sampled by using SMOTE (Synthetic Minority Over Sampling Technique).
- The Gradient Boosting model performs slightly better compared to other models with best recall score and low difference.

6. MODEL BUILDING – UNDERSAMPLED DATA

Fig.39 Data Count after Undersampling

```
Before Under Sampling, counts of label 'Cerified': 10199
Before Under Sampling, counts of label 'Denied': 5068

After Under Sampling, counts of label 'Cerified': 5068
After Under Sampling, counts of label 'Denied': 5068

After Under Sampling, the shape of train_X: (10136, 27)
After Under Sampling, the shape of train_y: (10136,)
```

Fig.40 Model performances (undersampled)

```
Training Performance:

Bagging: 0.9636937647987371
Random forest: 0.999802683504341
GBM: 0.7407261247040252
Adaboost: 0.7063930544593529
XGB: 0.8196527229676401

Validation Performance:

Bagging: 0.6177594825051456
Random forest: 0.6489267862393414
GBM: 0.7236107027344899
Adaboost: 0.7162599235518965
XGB: 0.69920611584828
```

Fig.41 Difference in Recall Scores 3

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9637, Validation Score: 0.6178, Difference: 0.3459
Random forest: Training Score: 0.9998, Validation Score: 0.6489, Difference: 0.3509
GBM: Training Score: 0.7407, Validation Score: 0.7236, Difference: 0.0171
Adaboost: Training Score: 0.7064, Validation Score: 0.7163, Difference: -0.0099
XGB: Training Score: 0.8197, Validation Score: 0.6992, Difference: 0.1204
```

- The dataset is under-sampled by using RandomUnderSampler() function.
- The Gradient Boosting model performs slightly better compared to other models with best recall score and low difference.

7. MODEL BUILDING – USING HYPERPARAMETER TUNING

Models selected for hyperparmeter tuning:

- Adaboost Model trained with original data.
- Gradient Boosting model trained with original data.
- Adaboost Model trained with oversampled data.

These 3 models are the top performing models in the initial model building stage. Let's improve the models by tuning with hyperparameters.

➢ **Adaboost Model trained with original data:**

Parameters:

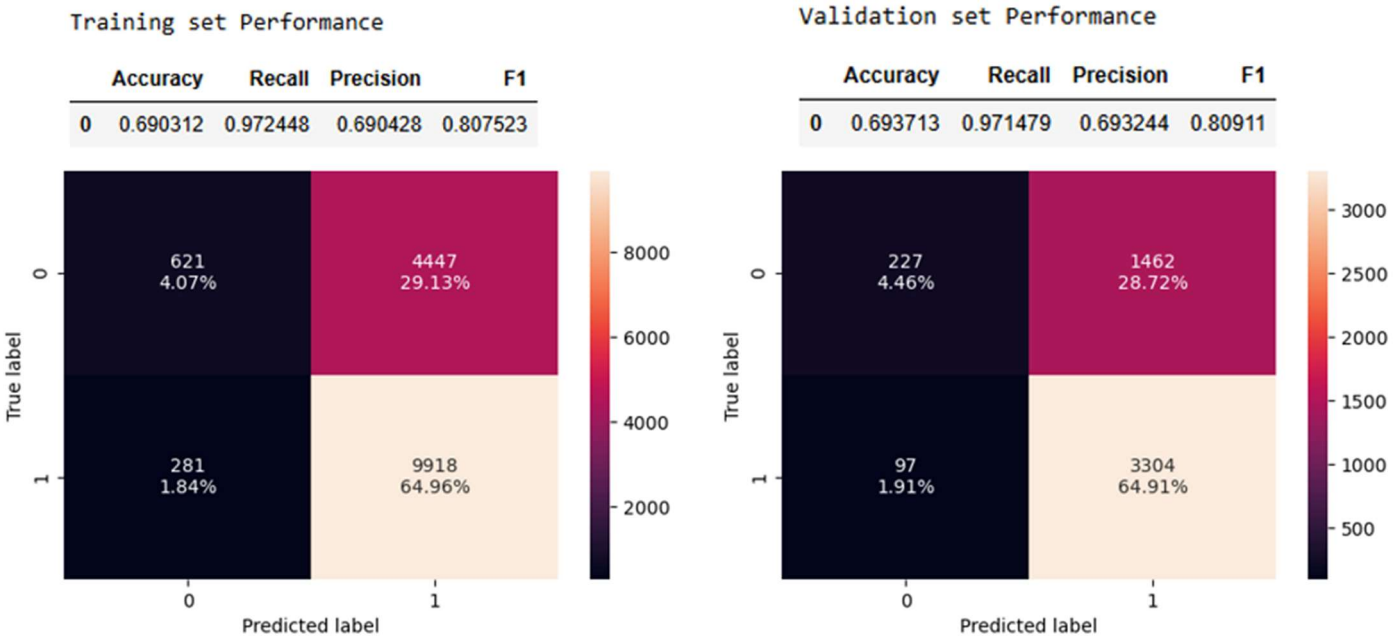
random_state = 1,

n_estimators = 30,

learning_rate = 0.05,

estimator = DecisionTreeClassifier()

Fig.42 Tuned Adaboost Performance 1



- **Gradient Boosting model trained with original data:**

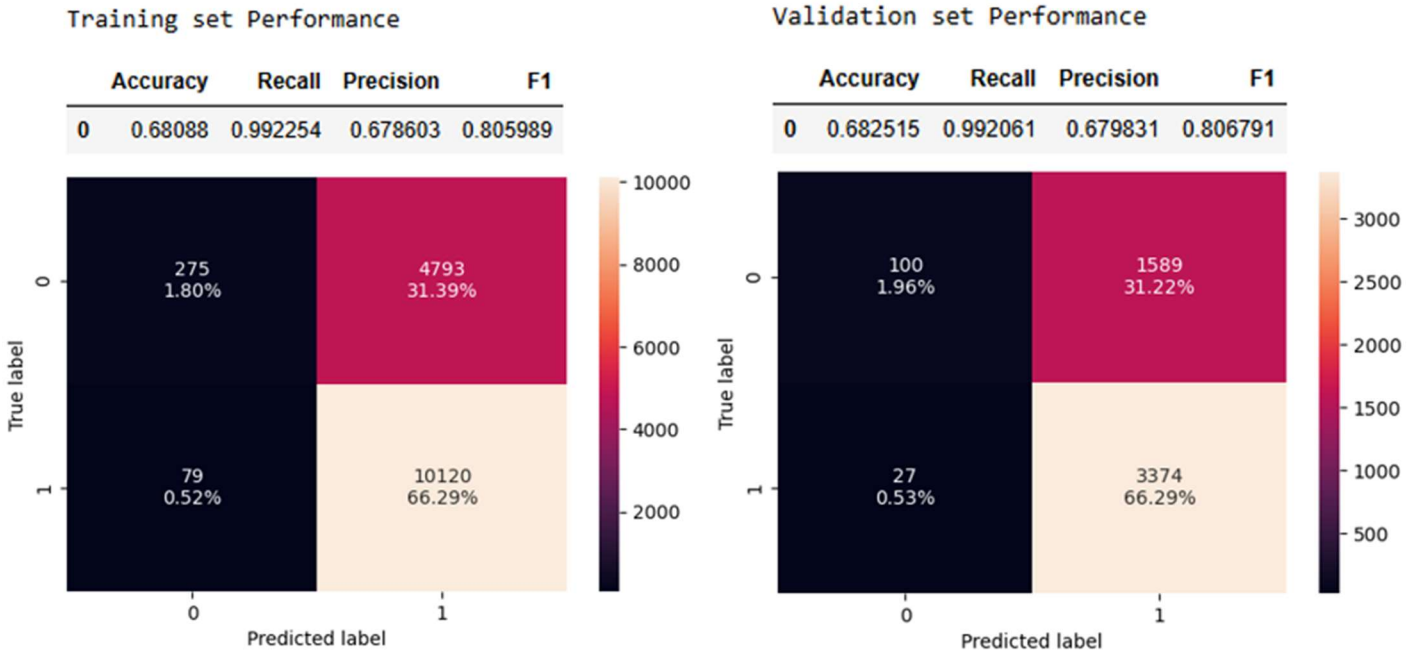
  Parameters:

  random_state = 1, subsample = 0.5, n_estimators = 125,

  max_features = 1, learning_rate = 0.01,

  init = AdaBoostClassifier(random_state=1)

Fig.43 Tuned GBM Val Perf



- **Adaboost Model trained with oversampled data:**

  Parameters:

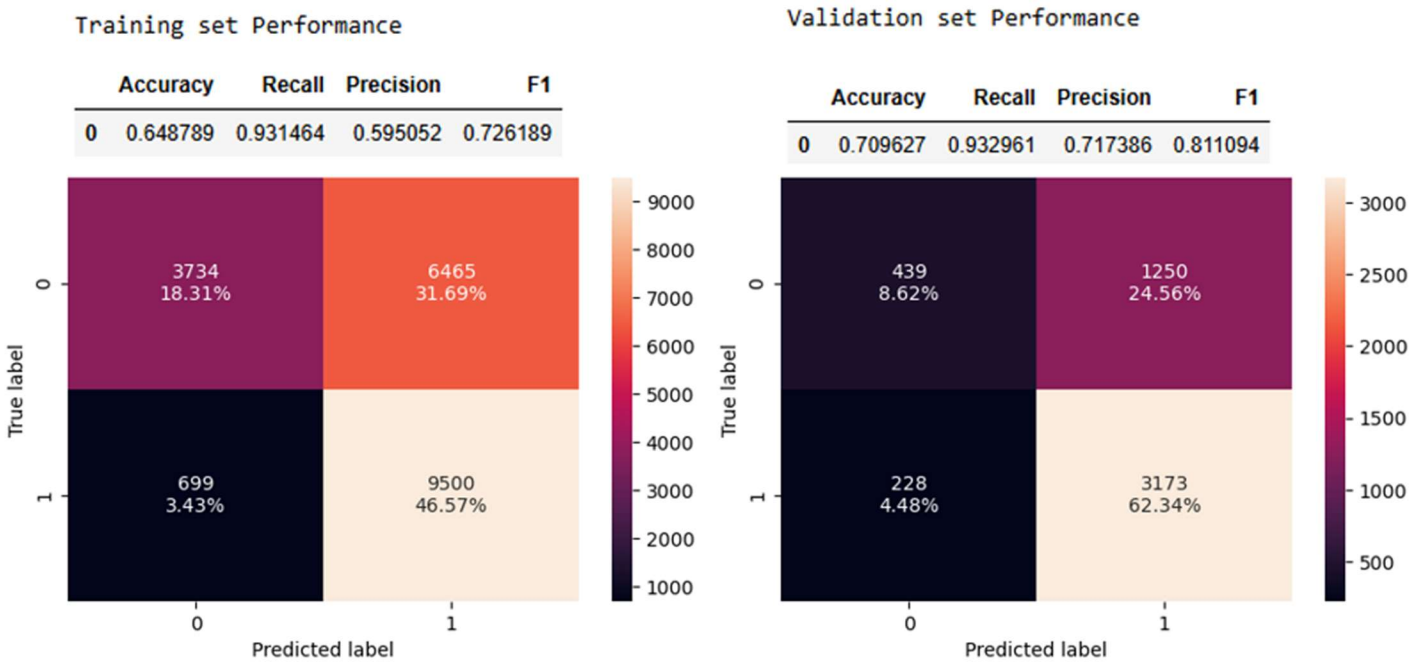
  random_state=1,

  n_estimators=10,

  learning_rate=0.01,

  estimator=DecisionTreeClassifier(max_depth=1, random_state=1)

Fig.44 Tuned Adaboost Performance 2



COMMENTS:

- The parameters for the models are selected by using RandomizedSearchCV method.
- Recall score is used to compare the parameter combinations.
- After Hyper-parameter tuning, there is a significant increase in the recall scores of all models.
- Gradient Boosting Models trained with original data performs very well in both training and validation sets after hyperparameter tuning.

8. MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

Tuned Model Performances:

Fig.45 Train Set Perf

Training performance comparison:

	AdaBoost trained with Original data	Gradient boosting trained with Original data	AdaBoost trained with Oversampled data
Accuracy	0.690312	0.680880	0.648789
Recall	0.972448	0.992254	0.931464
Precision	0.690428	0.678603	0.595052
F1	0.807523	0.805989	0.726189

Fig.46 Validation Set Perf

Validation performance comparison:

	AdaBoost trained with Original data	Gradient boosting trained with Original data	AdaBoost trained with Oversampled data
Accuracy	0.693713	0.682515	0.709627
Recall	0.971479	0.992061	0.932961
Precision	0.693244	0.679831	0.717386
F1	0.809110	0.806791	0.811094

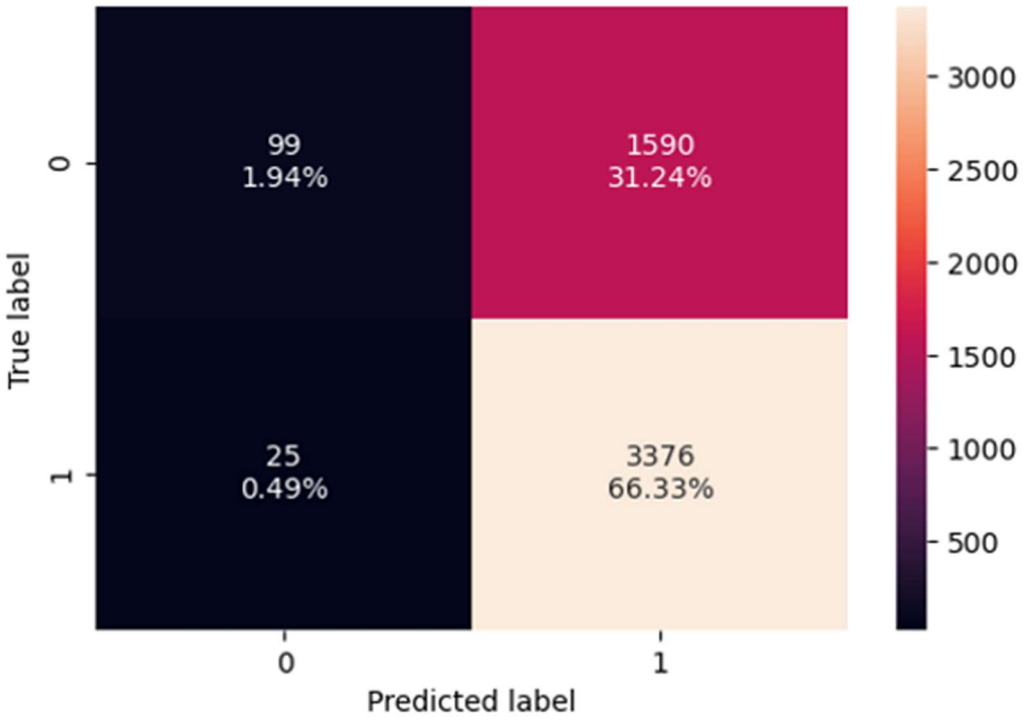
- As we can clearly see, the tuned Gradient Boosting Model trained with original data has generalised performance with 99% recall on both training and validation sets.
- So, the tuned Gradient Boost model trained with original data is considered as the best model for predicting the case_status.

➢ **Final Model Performance on the Test Set:**

Fig.47 Final Model Performance

Test set Performance

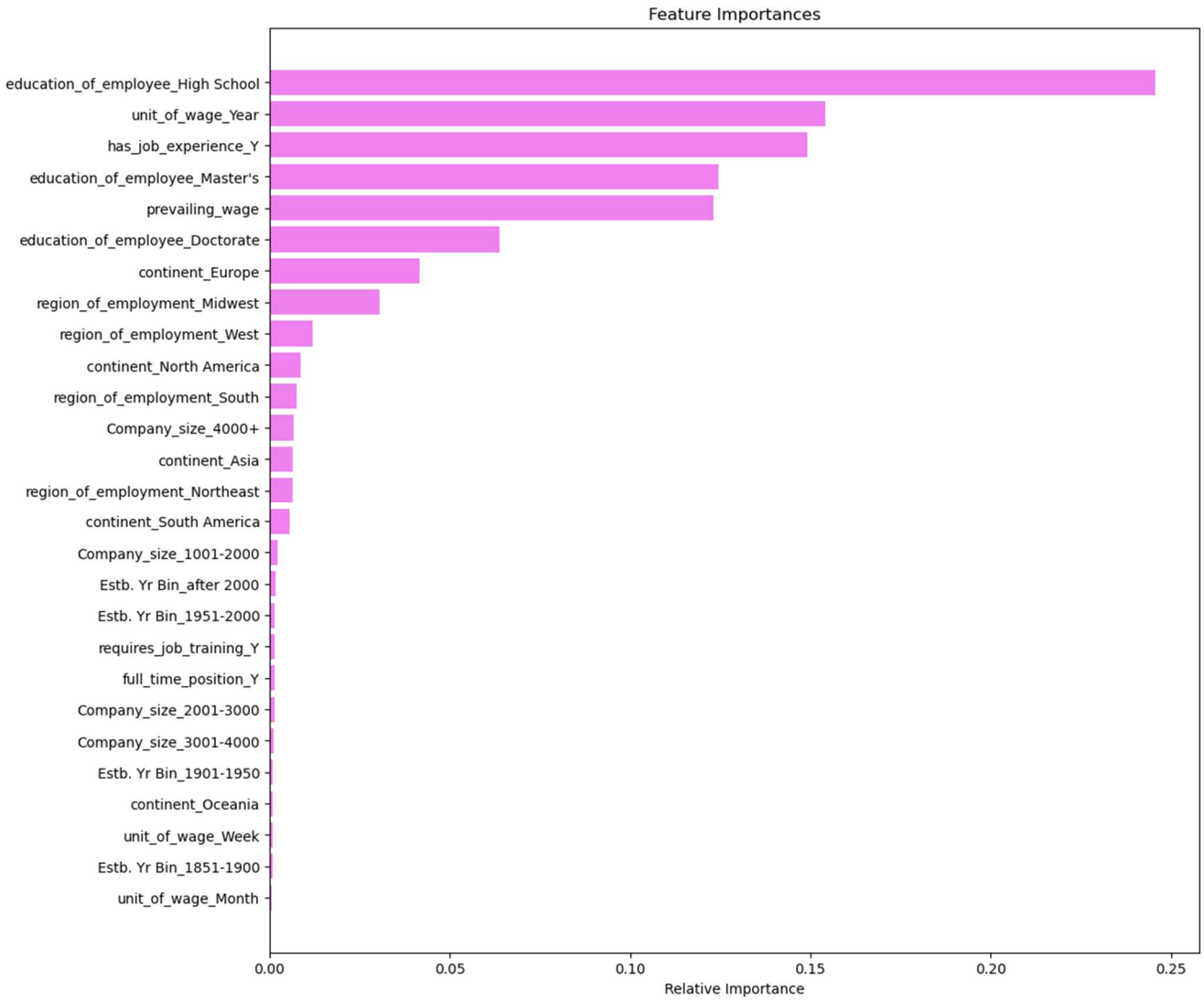
	Accuracy	Recall	Precision	F1
0	0.682711	0.992649	0.679823	0.80698



- The Gradient boosting model trained on original data has given ~99% recall on the test set.
- This model is highly effective at predicting almost all actual positive cases (approved visas).
- The performance is in line with what we achieved with this model on the train and validation sets.
- So, this is a generalized model.

➢ **Feature Importances:**

Fig.48 Feature Importances



➢ According to this model, “education_of_employee_High school” is the most important variable for predicting the case_status followed by “unit_of_wage_Year” and “has_job_experience_Y”.

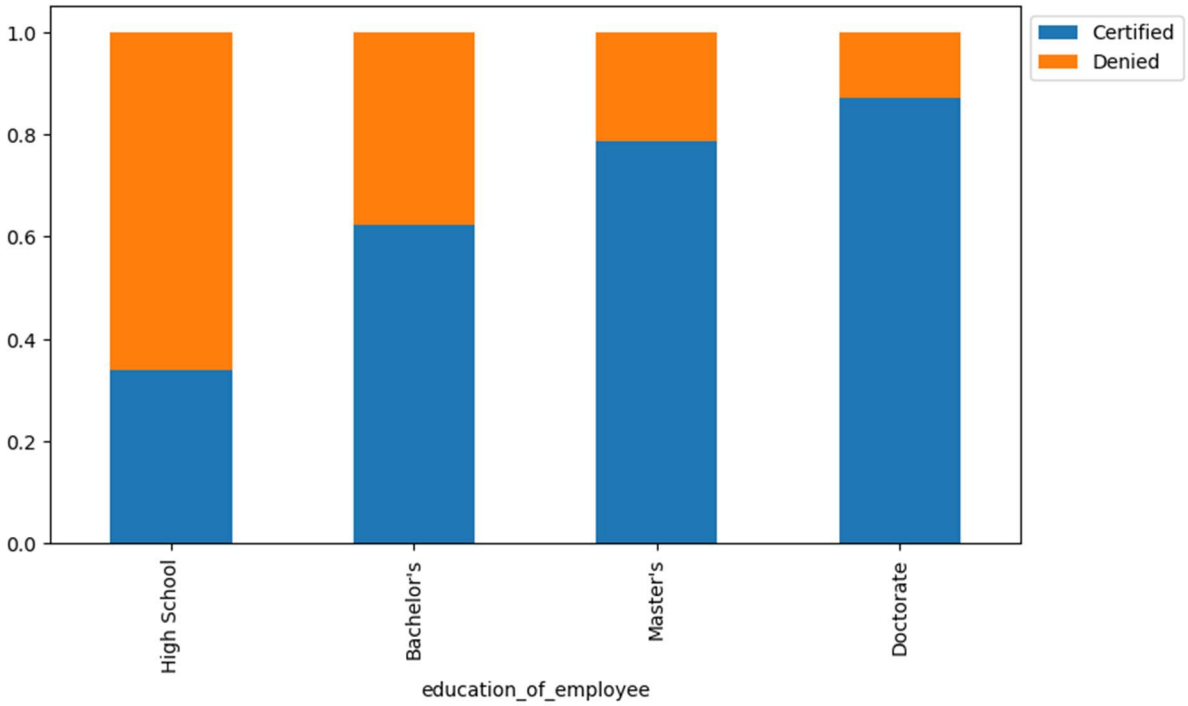
9. ACTIONABLE INSIGHTS & RECOMMENDATIONS

- Based on the model, the applicants are certified for visa on the basis of:
    1. Education of the applicant
    2. Unit of wage
    3. Job experience.
    4. Prevailing Wages

  These are top 4 features of the model.

- **Education of applicant:**

Fig.49 Education of Employee vs Case status



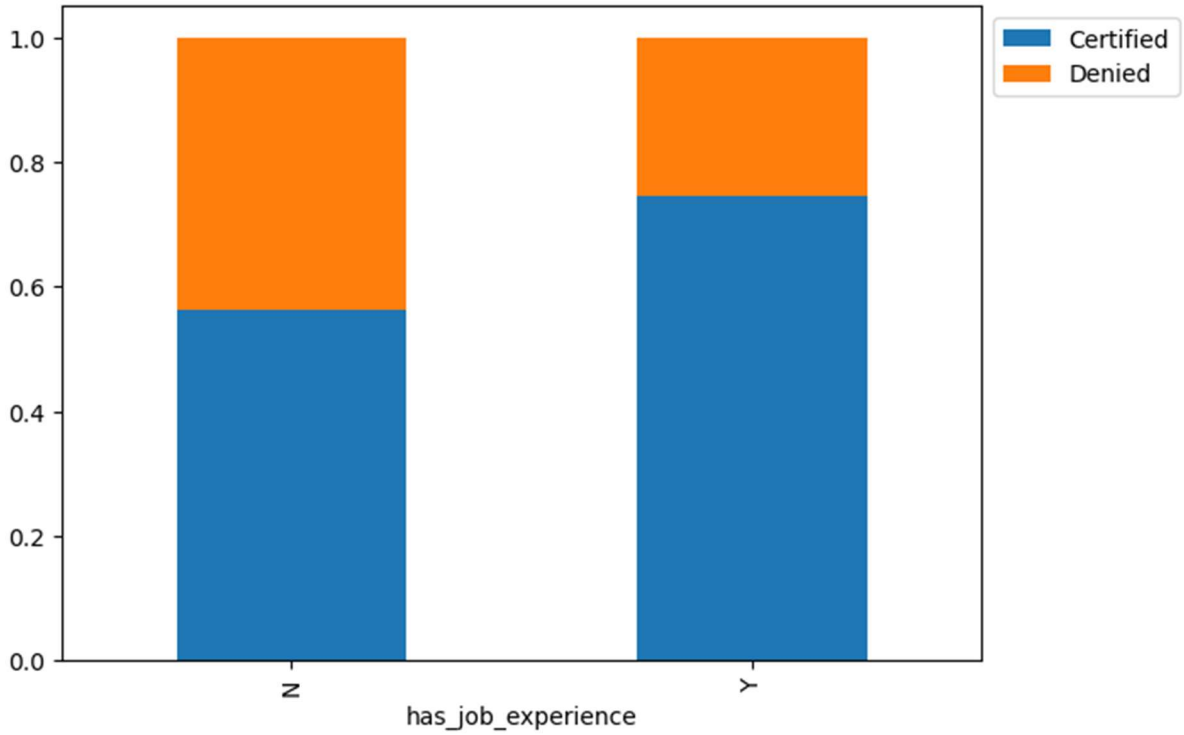
As we can clearly see, the chances of getting certified increases with the higher education. 78% of the employees having Master degree are certified and 87% of the employees having Doctorate are certified. 65% of the High school applicants are denied.

- **Unit of wage:**

  Applicants having yearly unit of wage are mostly selected (69%)

- **Job experience:**

Fig.50 Job Experience vs Case status

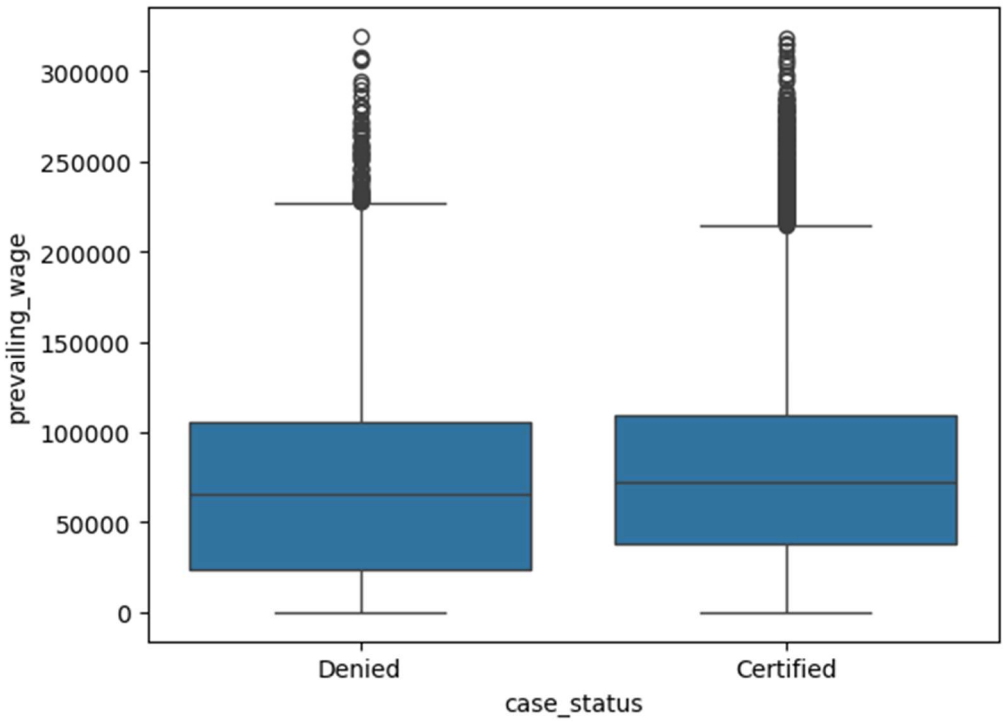


74% of the applicants having job experience are certified for visa.

Job experience is a crucial factor for the approval.

- **Prevailing Wages:**

Fig.51 Case status vs Prevailing wage



The median prevailing wage of the certified employees is around 65k.

RECOMMENDATIONS:

- Office of Foreign Labor Certification (OFLC) should consider the applicants' level of education, their job experience, their prevailing wages and their unit in reviewing its visa certification.
- The applicants who have a higher education, have job experience, and their US employment's wage unit is year are more likely to be certified for a work visa.
- In few cases, applicants from Europe also have higher chances of visa certification.
- Applicants choosing Midwest has higher chance of visa approval.