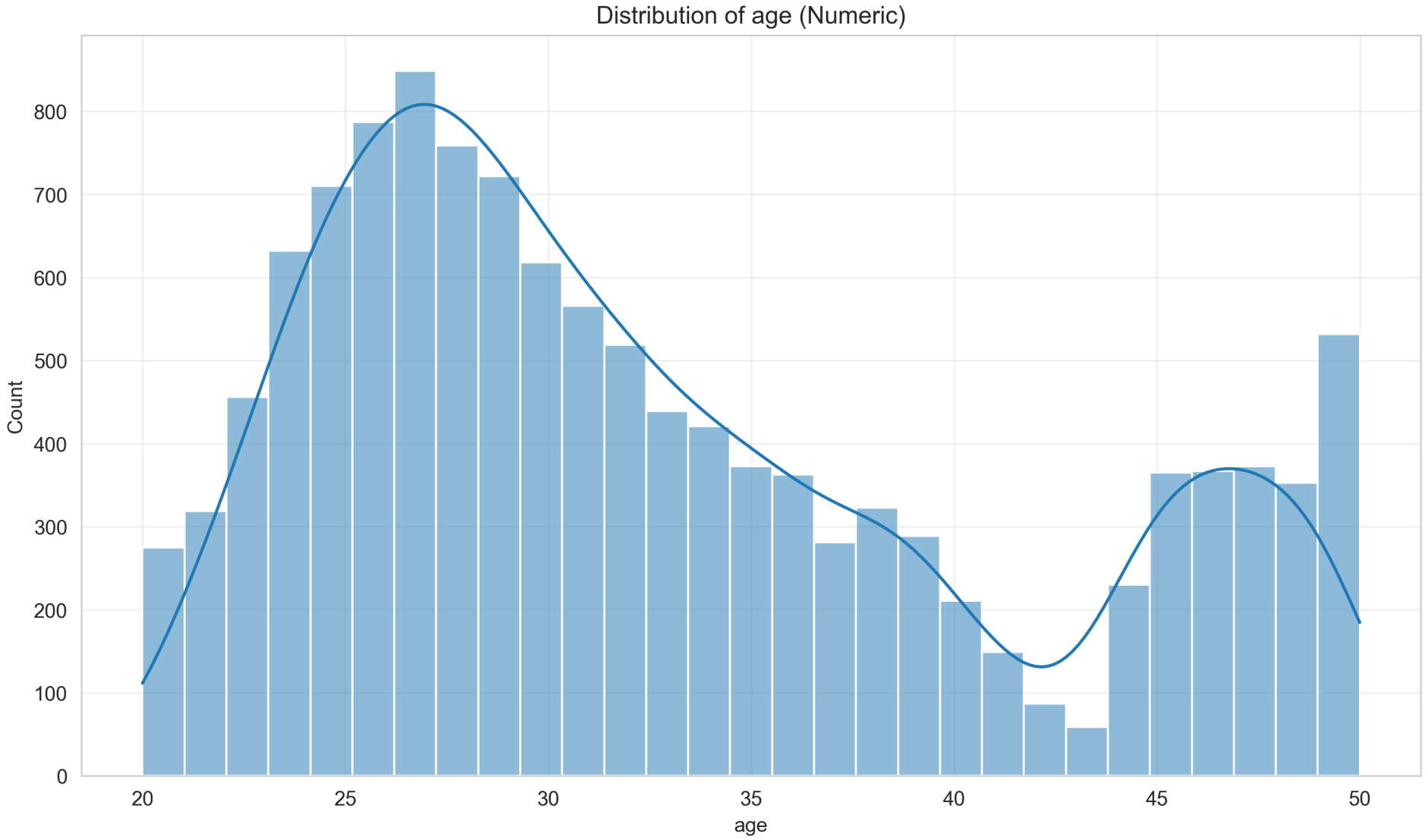
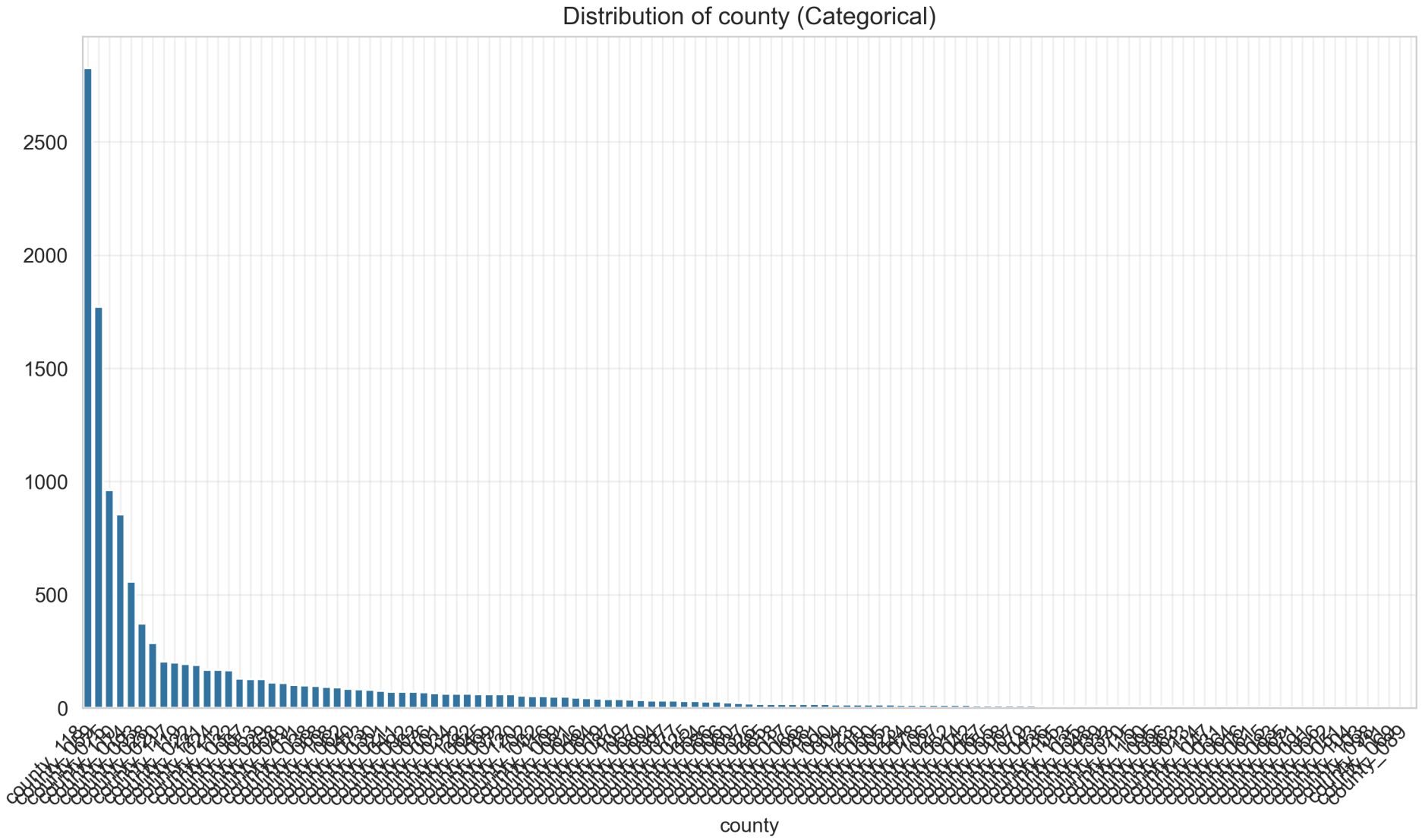
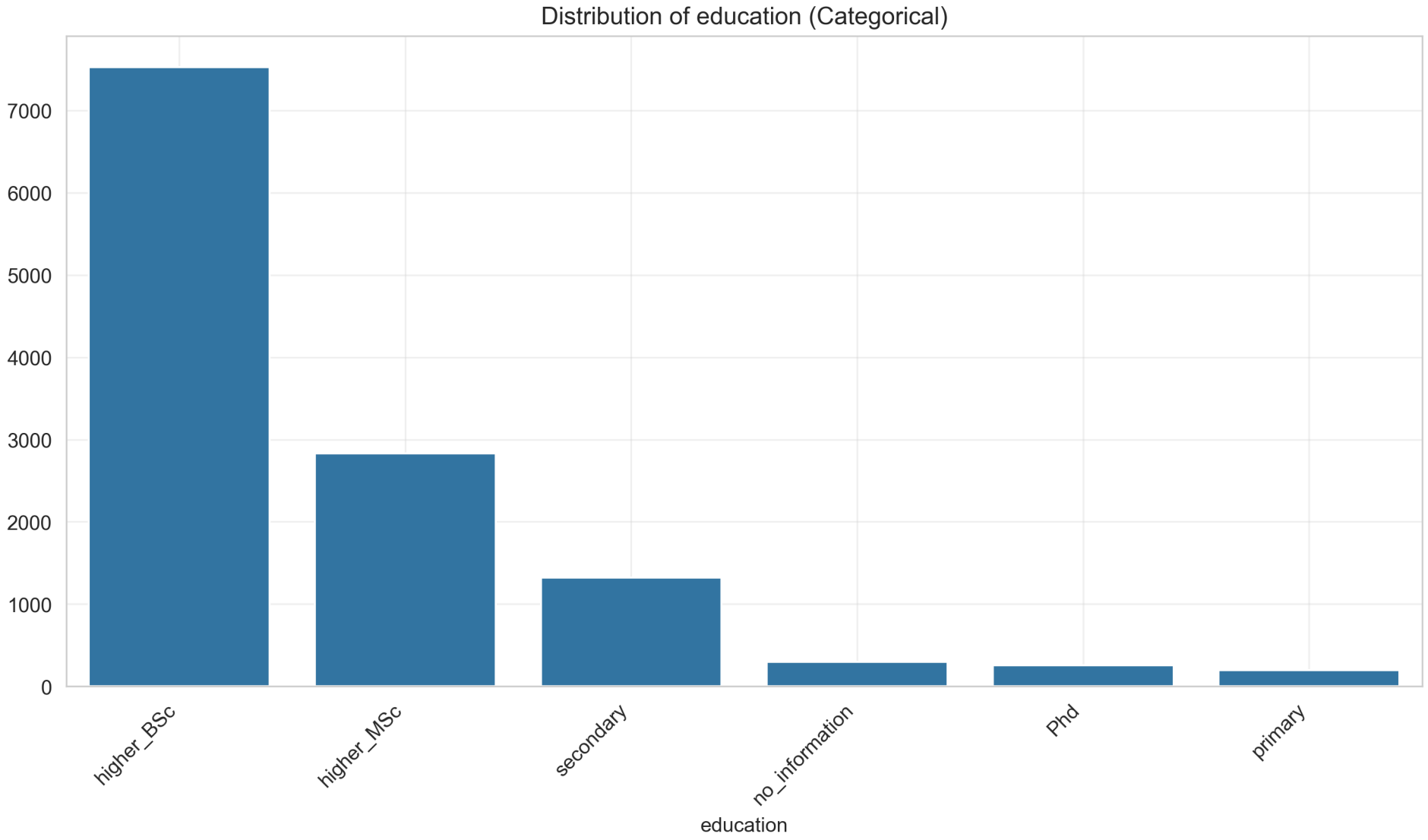
Job Change Prediction Model Analysis

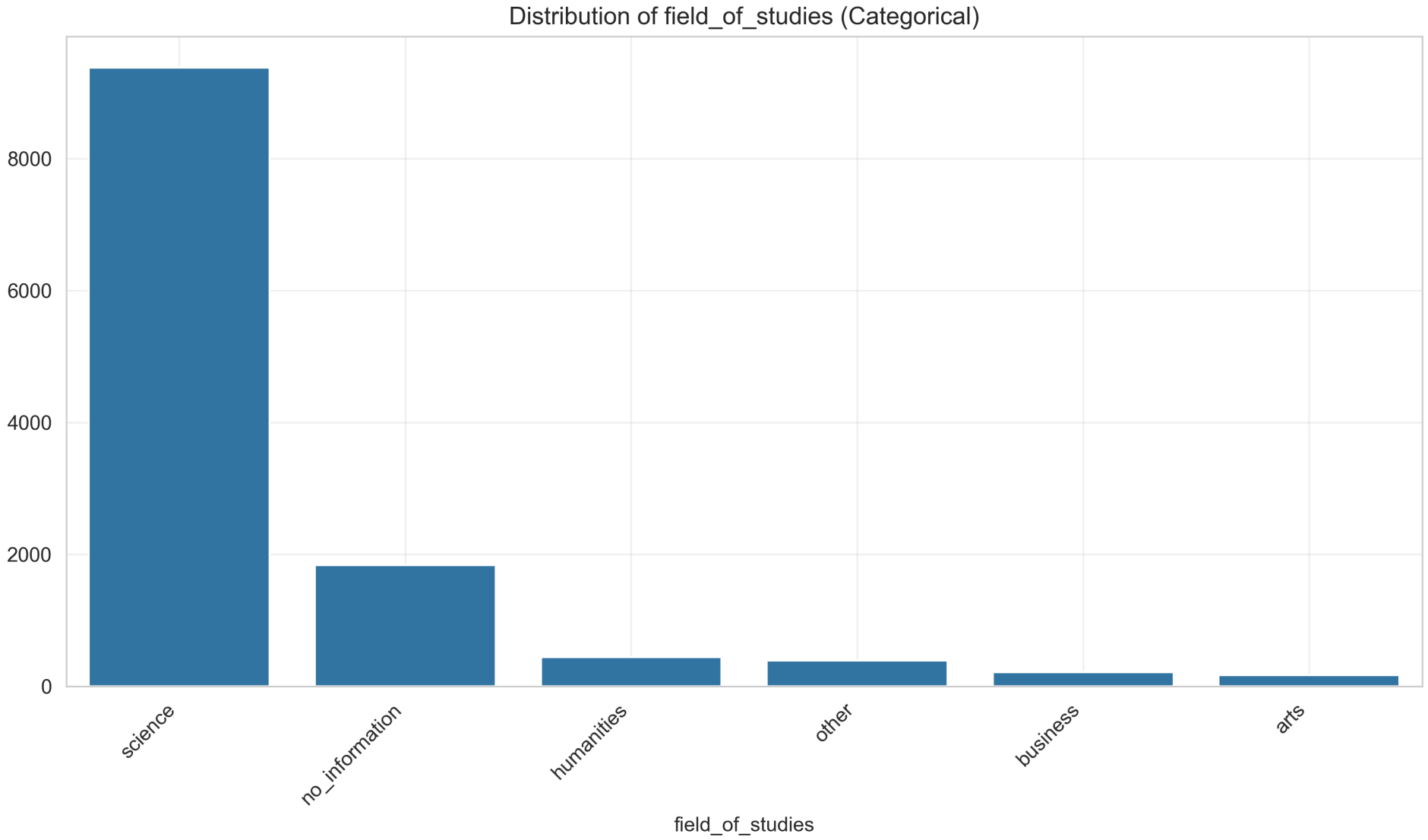
EDA

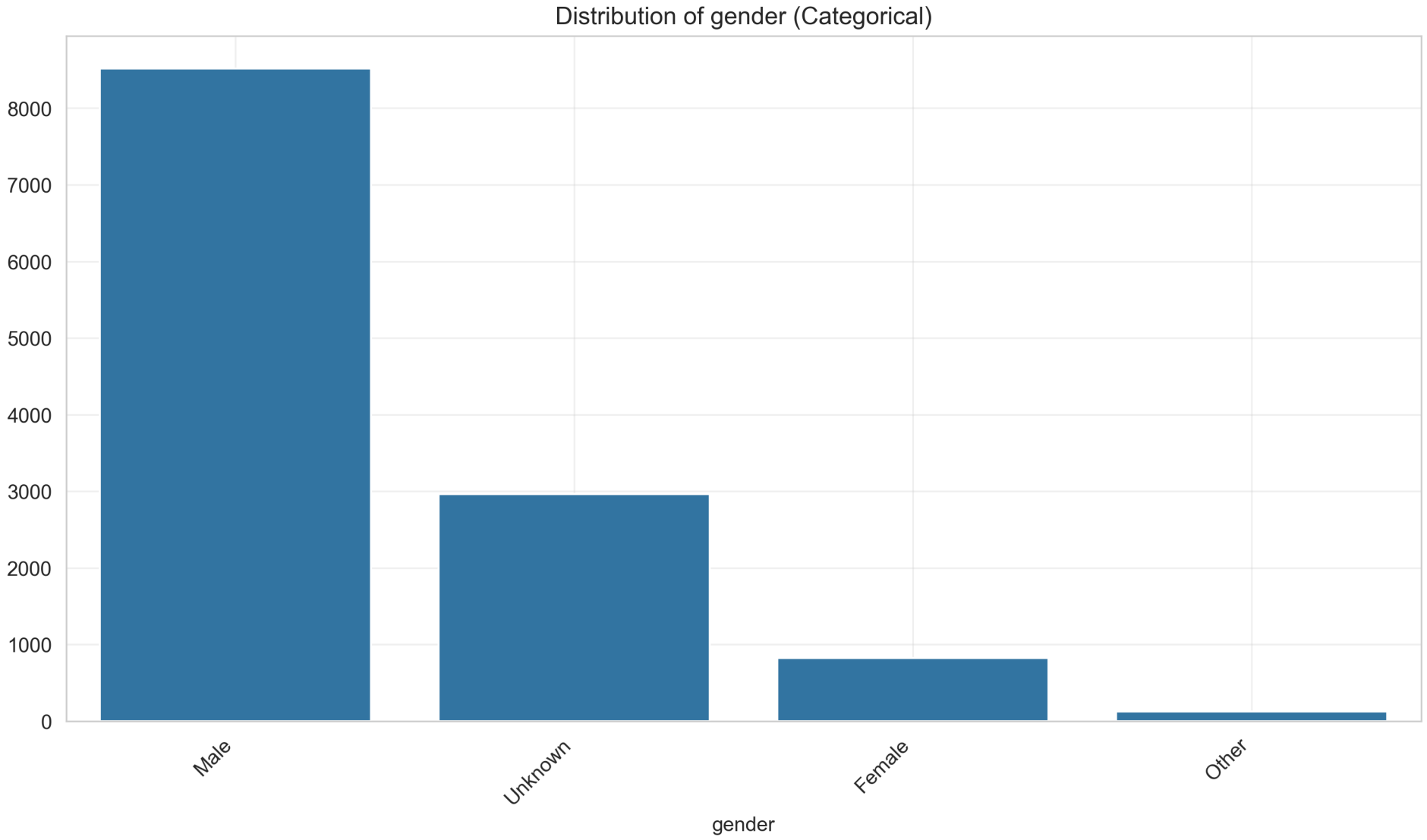
- No missing values
- Features histograms:

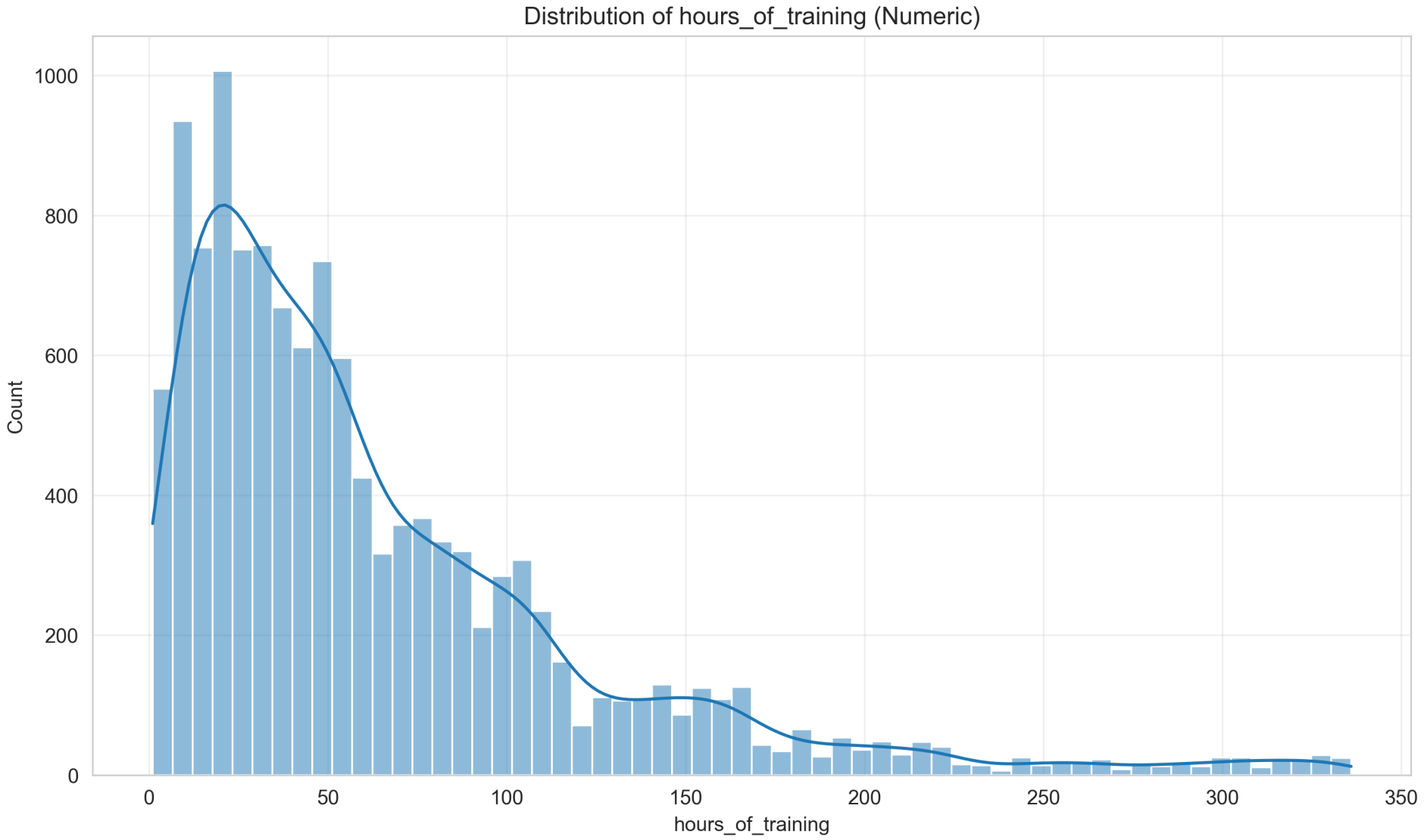


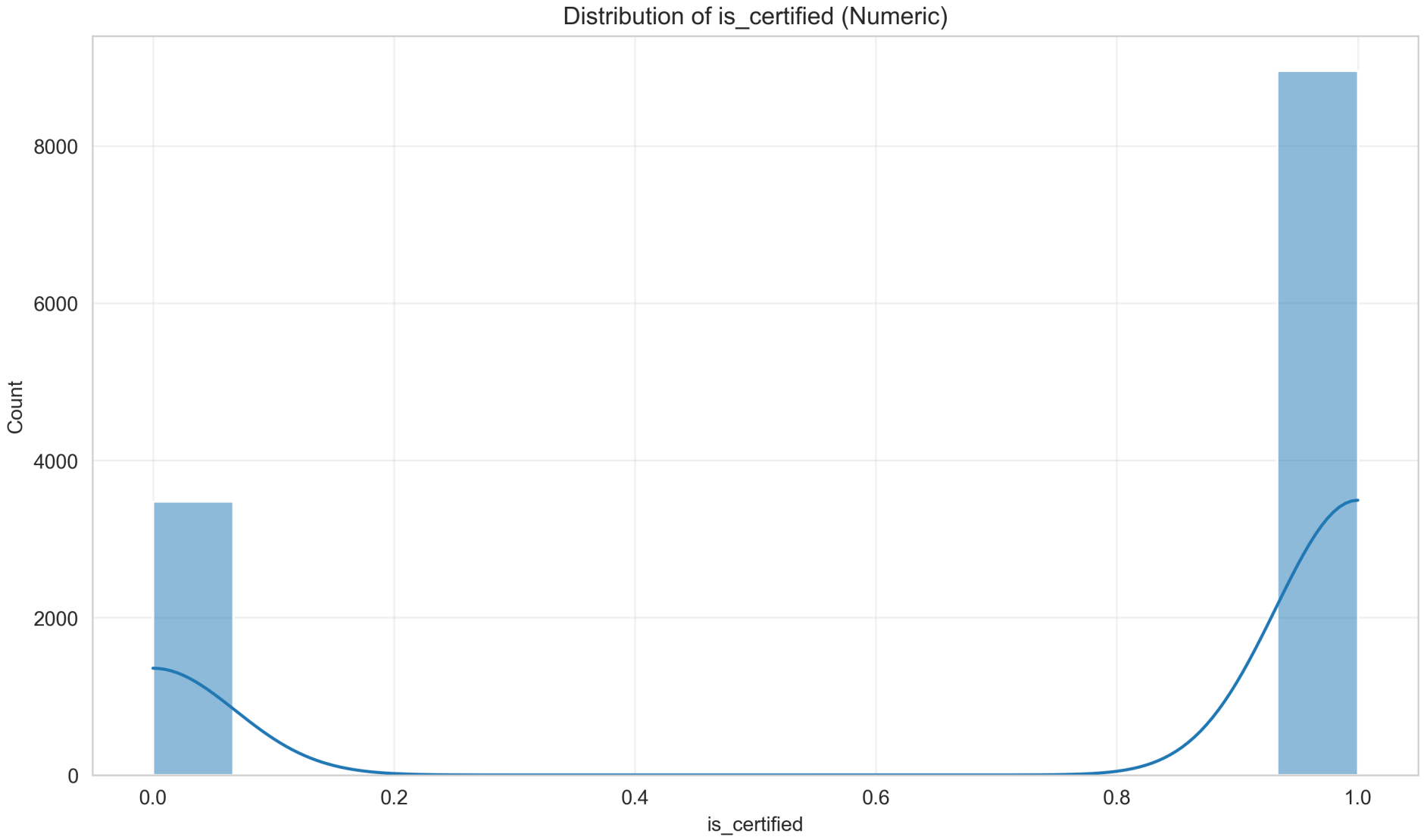

Distribution of county (Categorical)
2500
2000
1500
1000
500
0
county

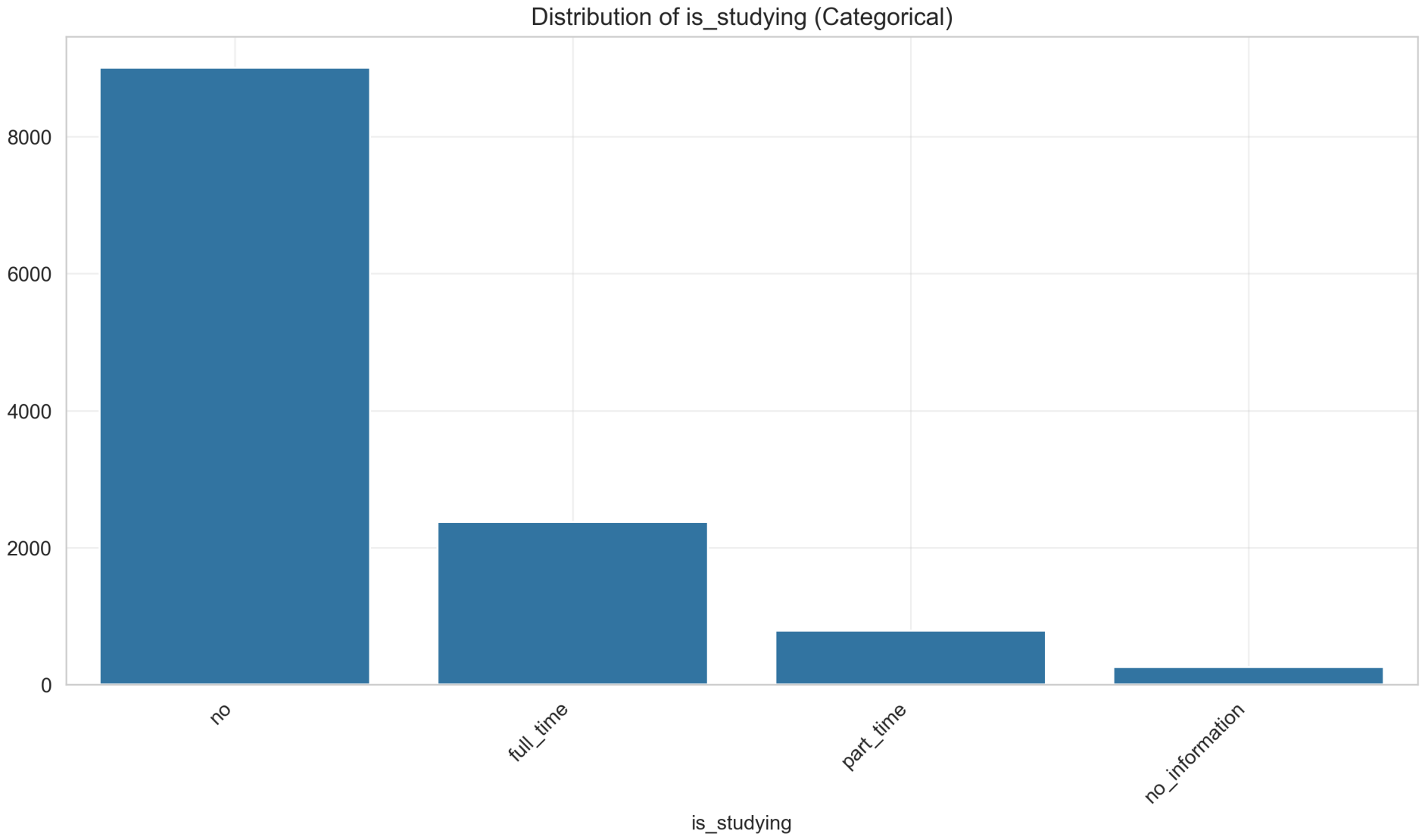

Distribution of education (Categorical)
7000
6000
5000
4000
3000
2000
1000
0
higher_BSc
higher_MSc
secondary
no_information
Phd
primary
education

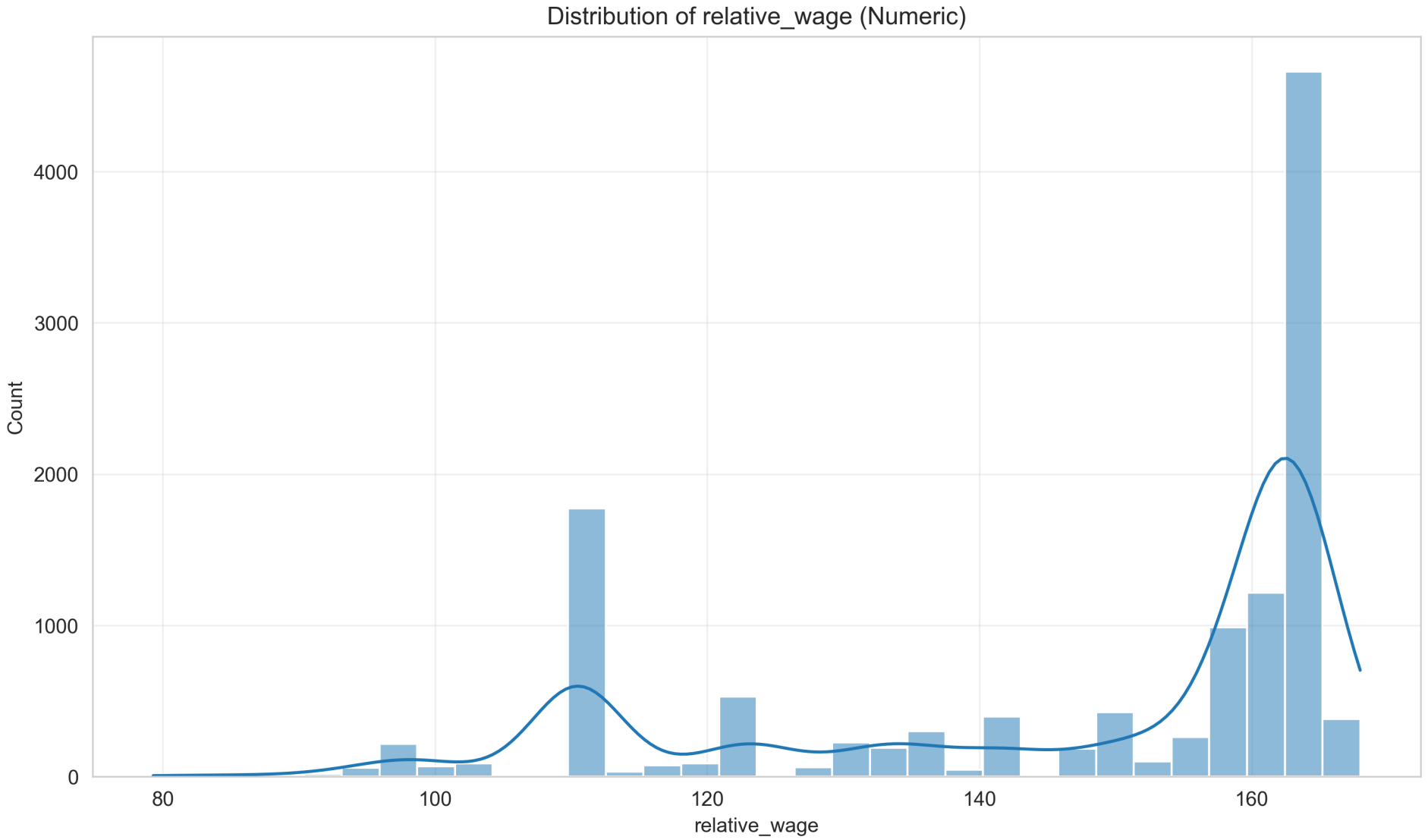

Distribution of field_of_studies (Categorical)
8000
6000
4000
2000
0
science
no_information
humanities
other
business
arts
field_of_studies

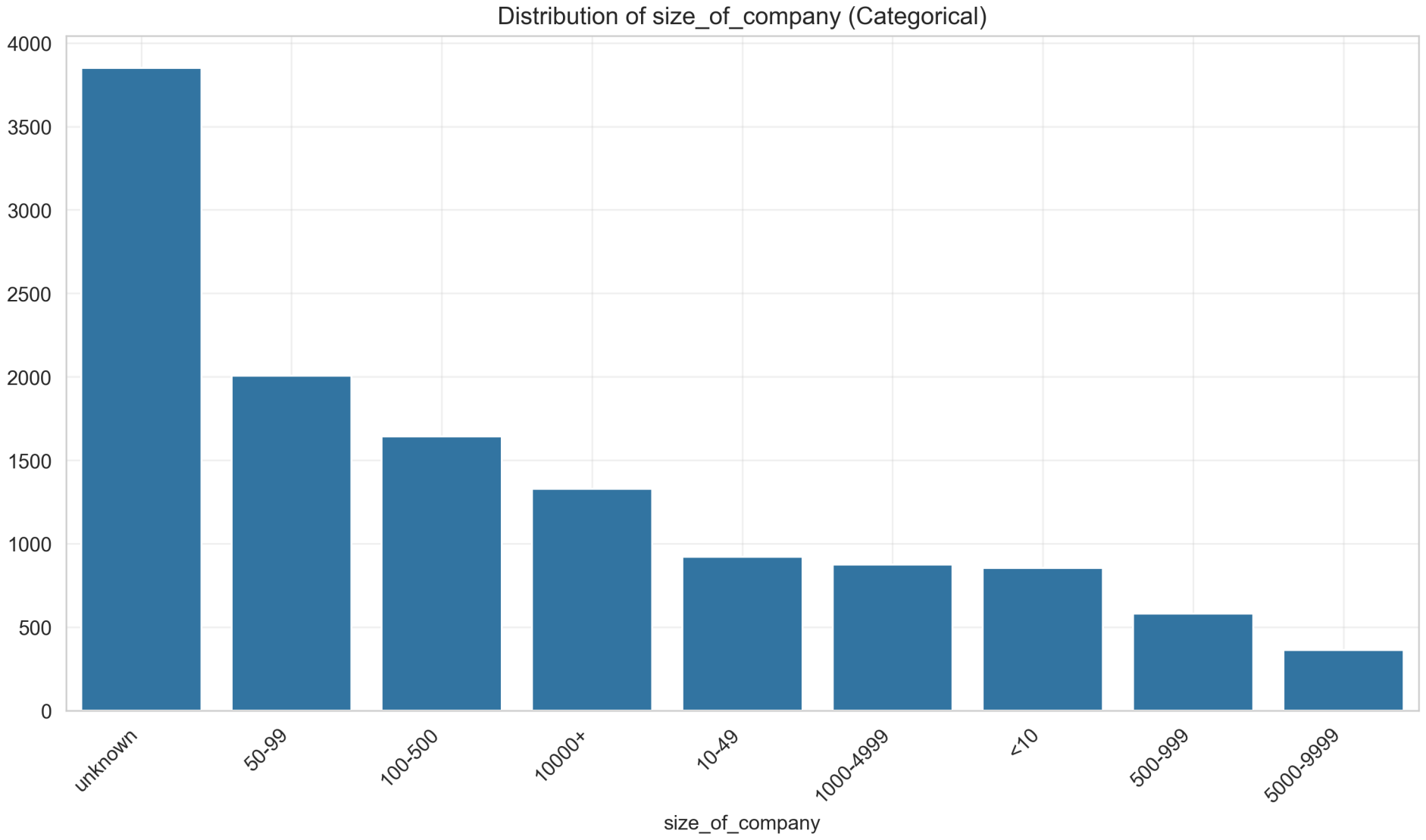

Distribution of gender (Categorical)
8000
7000
6000
5000
4000
3000
2000
1000
0
Male
Unknown
Female
Other
gender

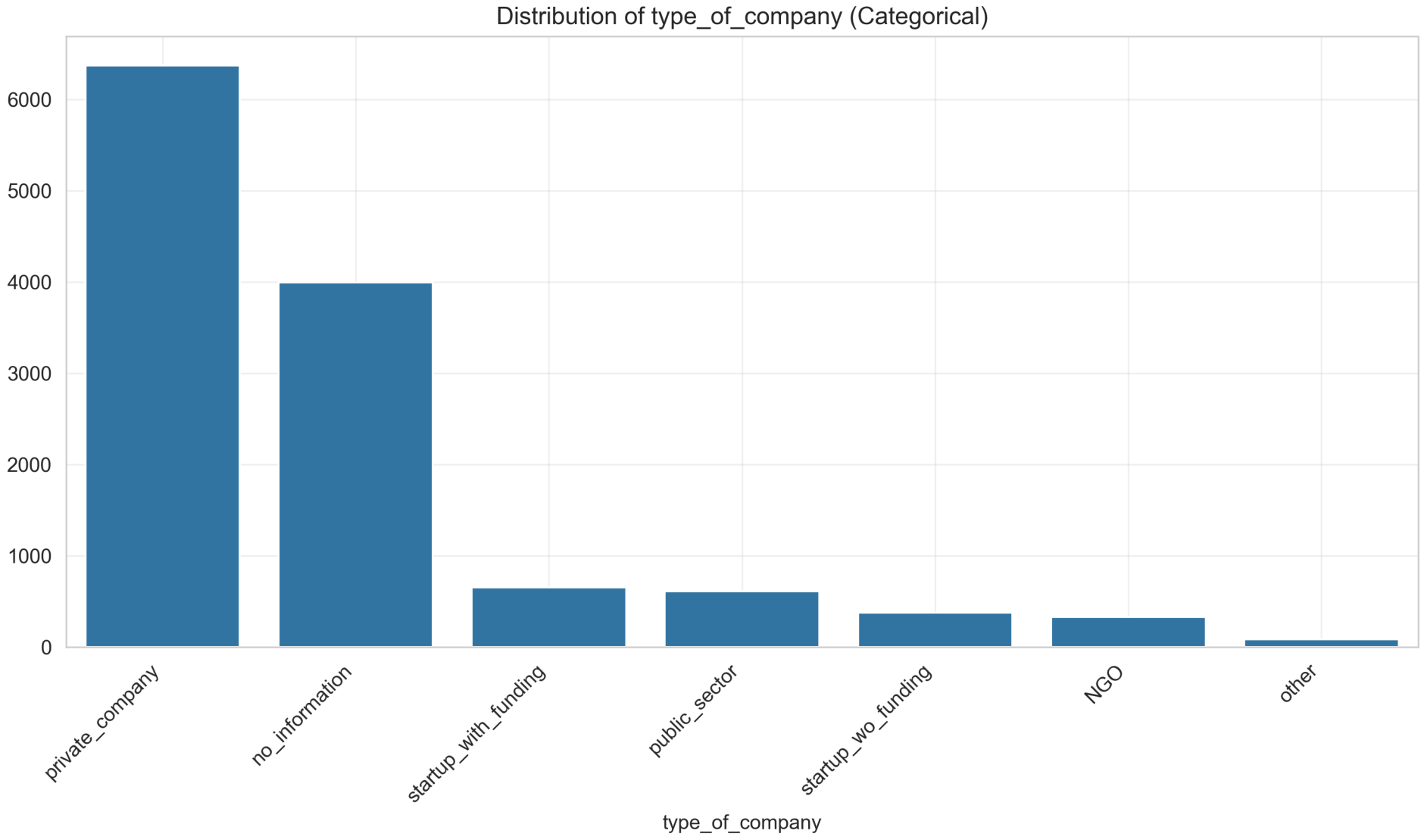

Distribution of hours_of_training (Numeric)
Count
hours_of_training
0
200
400
600
800
1000
0
50
100
150
200
250
300
350

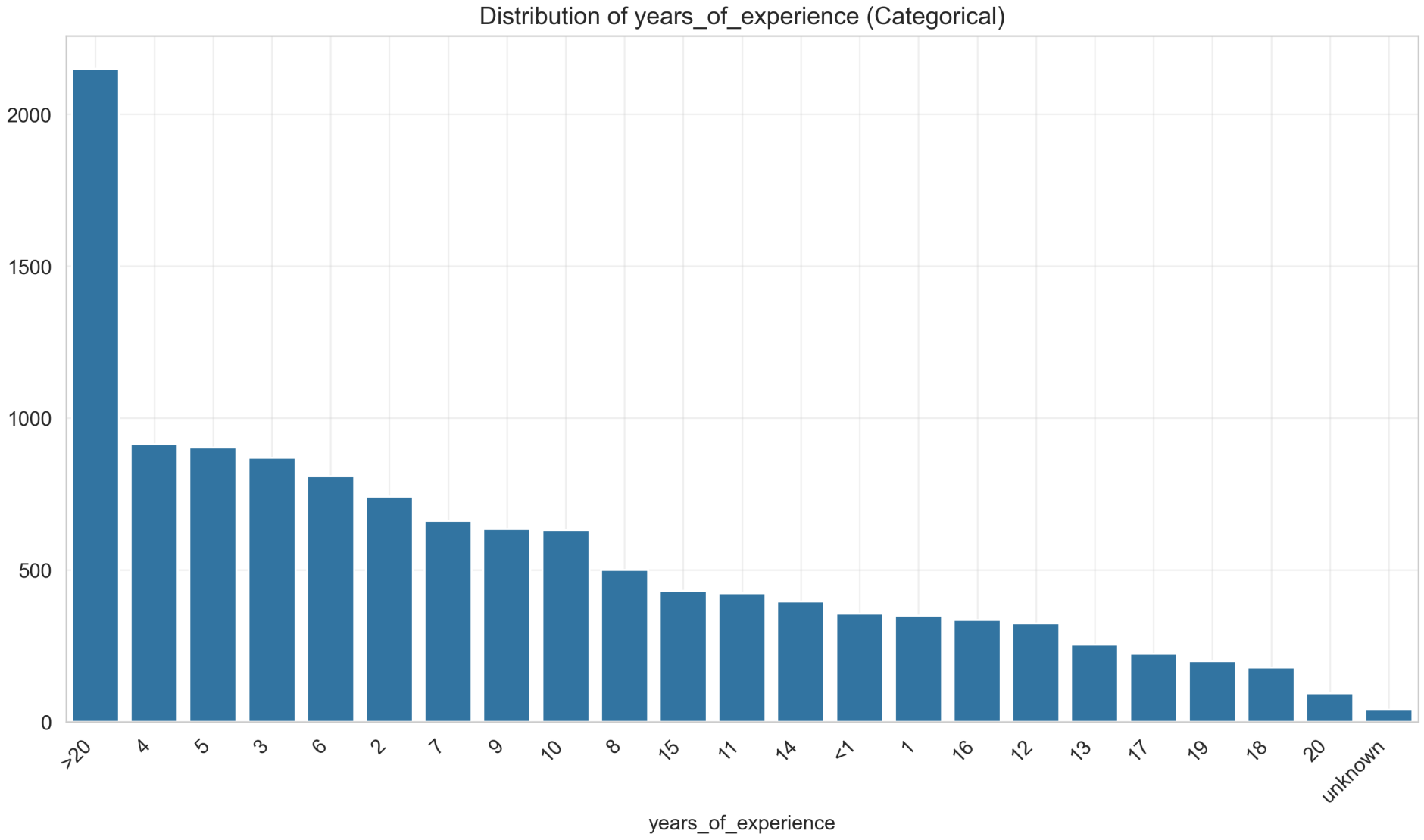

Distribution of is_certified (Numeric)
Count
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
is_certified

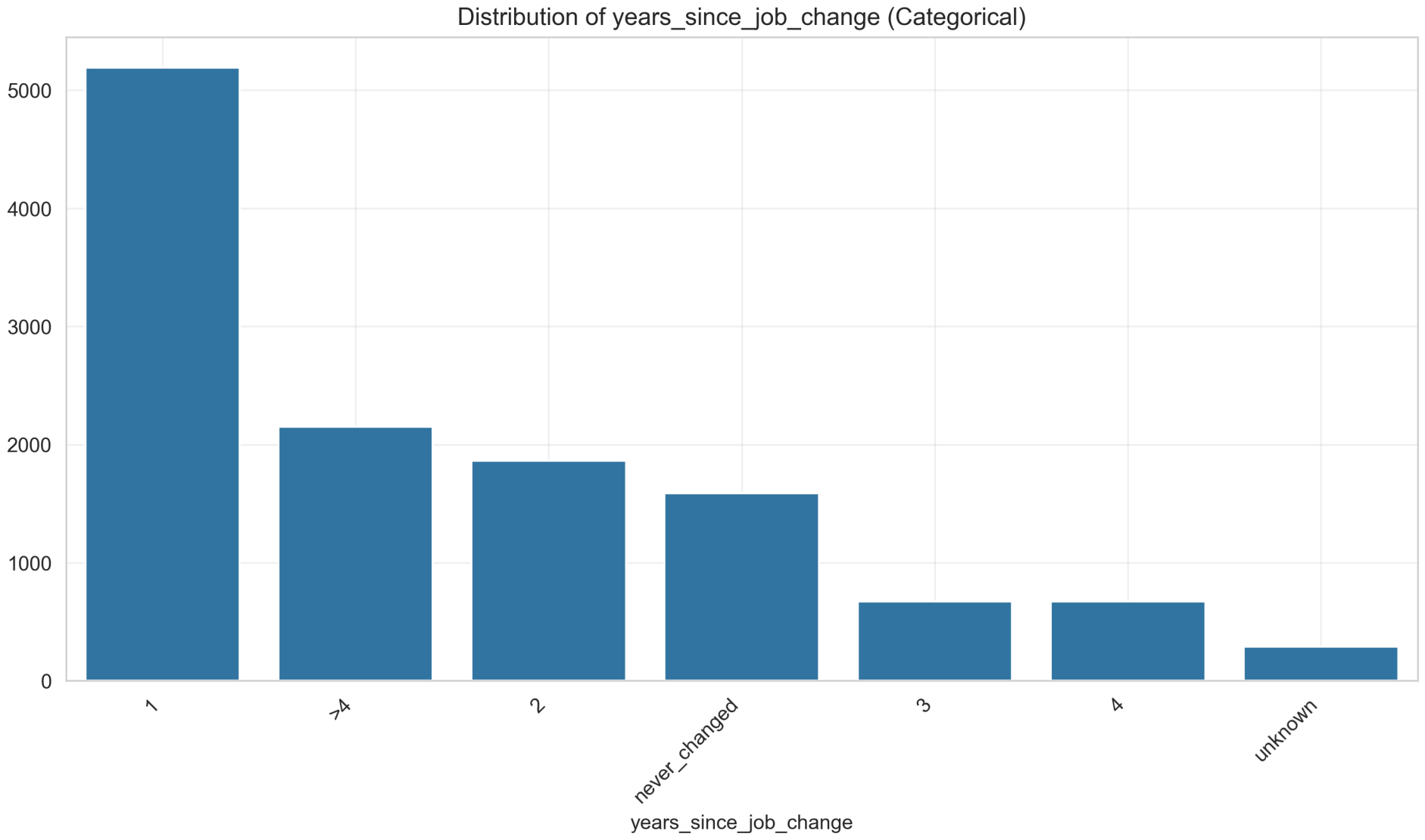

Distribution of is_studying (Categorical)
8000
6000
4000
2000
0
no
full_time
part_time
no_information
is_studying


Distribution of relative_wage (Numeric)
Count
4000
3000
2000
1000
0
80
100
120
140
160
relative_wage

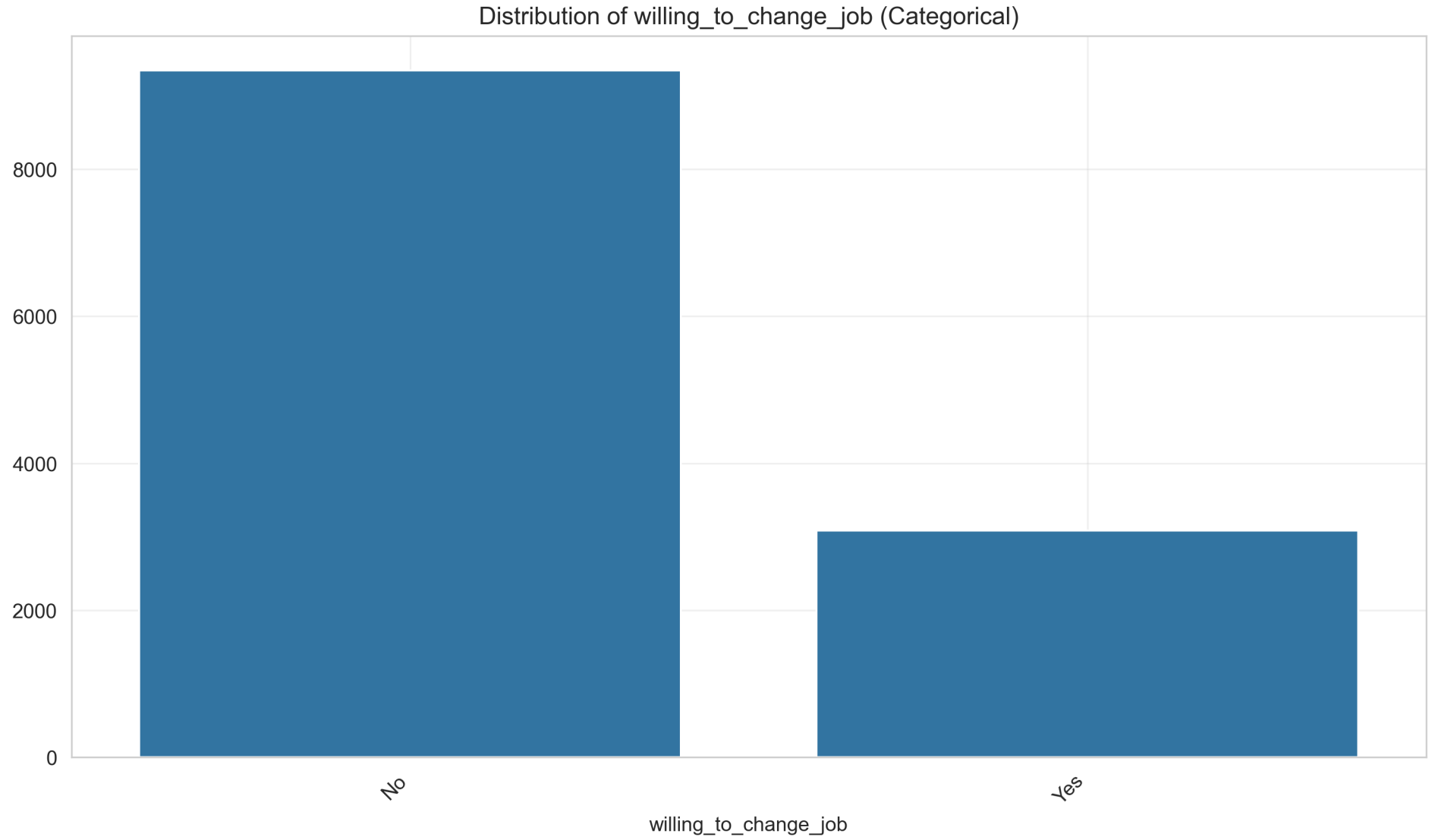

Distribution of size_of_company (Categorical)
4000
3500
3000
2500
2000
1500
1000
500
0
unknown
50-99
100-500
10000+
10-49
1000-4999
<10
500-999
5000-9999
size_of_company


Distribution of type_of_company (Categorical)
6000
5000
4000
3000
2000
1000
0
private_company
no_information
startup_with_funding
public_sector
startup_wo_funding
NGO
other
type_of_company


Distribution of years_of_experience (Categorical)
2000
1500
1000
500
0
>20
4
5
3
6
2
7
9
10
8
15
11
14
<1
1
16
12
13
17
19
18
20
unknown
years_of_experience


Distribution of years_since_job_change (Categorical)
5000
4000
3000
2000
1000
0
1
>4
2
never_changed
3
4
unknown
years_since_job_change

- Target histogram:



Data Preprocessing Pipeline

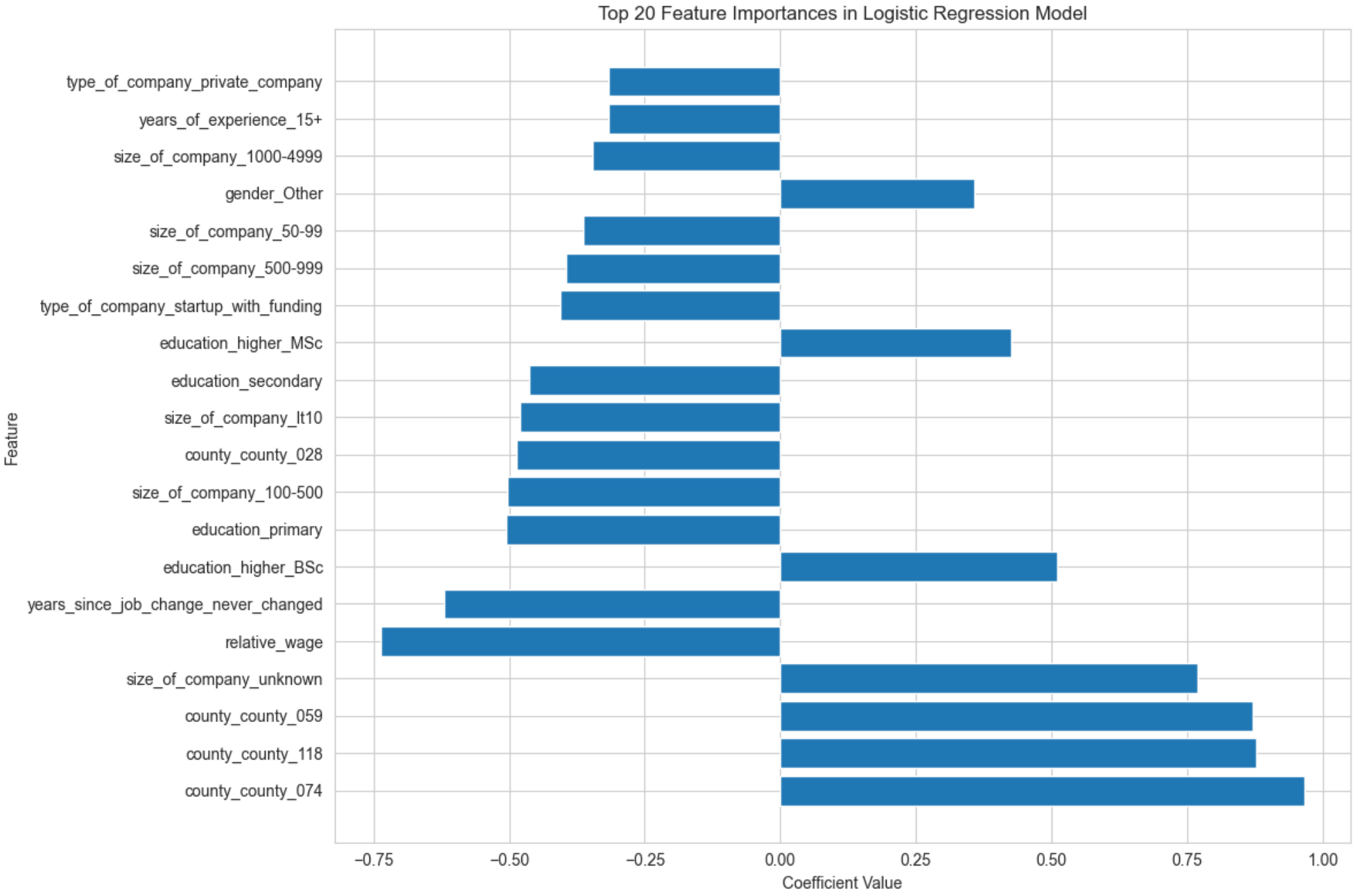
1. **Categorical Variables** (One-hot encoding)
    - gender
    - education
    - field_of_studies
    - is_studying
    - county
    - years_since_job_change
    - years_of_experience
    - size_of_company
    - type_of_company
2. **Numerical Variables** (Standardization)
    - age
    - relative_wage
    - hours_of_training

3. **Special Handling**
    - County grouping: Top 10 most frequent counties kept, rest grouped as "Other"
    - Experience binning: Converted to categorical ranges
        - 0-3 years
        - 3-7 years
        - 7-15 years
        - 15+ years
    - Special values handling:
      - '>20' years converted to 21
      - '<1' year converted to 0

Model Selection

After testing multiple algorithms including Random Forest, XGBoost, Gradient Boosting, and KNN, **Logistic Regression** was chosen.

Model	CV Score	CV Std	Test Score	Training Time (s)
XGBoost	0.7942	0.0095	0.6757	0.798
Logistic Regression	**0.7602**	**0.0103**	**0.7554**	**0.197**
Gradient Boosting	0.7021	0.0023	0.6908	7.985
Random Forest	0.6748	0.0115	0.6711	8.580
KNN	0.6617	0.0100	0.6639	0.593


Top 20 Feature Importances in Logistic Regression Model
type_of_company_private_company
years_of_experience_15+
size_of_company_1000-4999
gender_Other
size_of_company_50-99
size_of_company_500-999
type_of_company_startup_with_funding
education_higher_MSc
education_secondary
size_of_company_lt10
county_county_028
size_of_company_100-500
education_primary
education_higher_BSc
years_since_job_change_never_changed
relative_wage
size_of_company_unknown
county_county_059
county_county_118
county_county_074
Feature
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Coefficient Value