1. Wstępna Analiza Danych

Informacje o danych:

Liczba wierszy: 4739

Liczba kolumn: 15

Opis zmiennych:

Numeryczne: rownames (identyfikator), score (wynik do przewidzenia), unemp, wage, distance, tuition, education

Kategoryczne: gender, ethnicity, fcollege, mcollege, home, urban, income, region

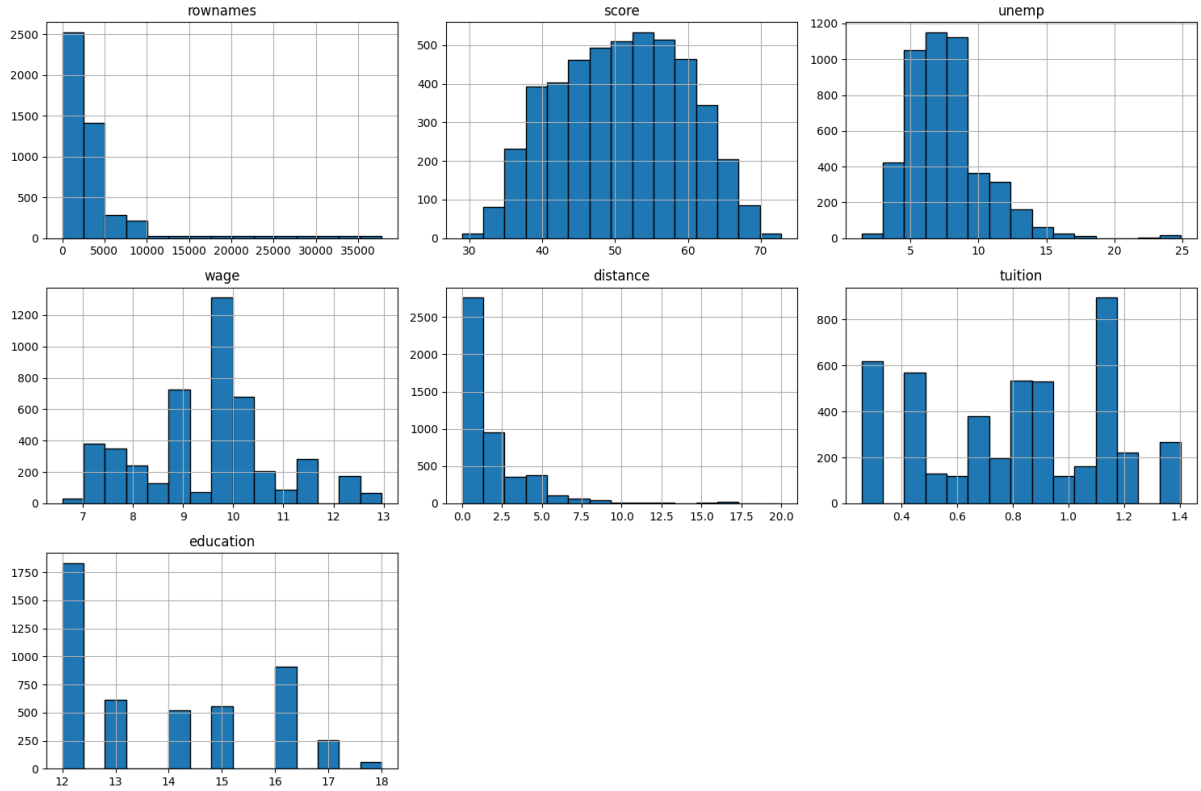
Brakujące dane:

Zastosowano imputację brakujących wartości w kolumnach numerycznych przy użyciu średniej oraz najczęściej występującej wartości dla zmiennych kategorycznych.

Dodatkowo usunięto wiersze z więcej niż jednym brakującym polem.

2. Analiza Statystyczna

Zmienne numeryczne:



Unemployment (unemp):

Średnia: 7.6

Mediana: 7.1

Zakres wartości: 1.4 - 24.9

Wynagrodzenie (wage):

Średnia: 9.5

Mediana: 9.7

Zakres: 6.59 - 12.96

Dystans do uczelni (distance):

Średnia: 1.8

Mediana: 1.0

Zakres: 0.0 - 20.0

Czesne (tuition):

Średnia: 0.81

Mediana: 0.82

Zakres: 0.26 - 1.40

Edukacja (education):

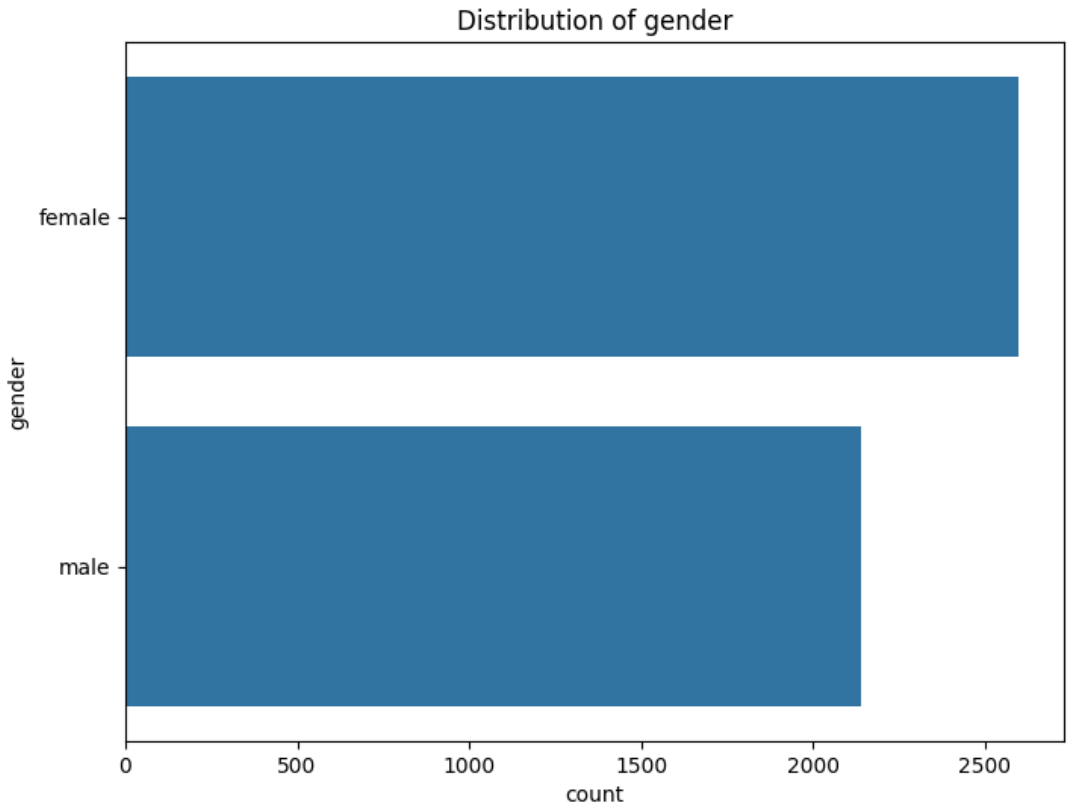
Średnia: 13.8

Mediana: 13

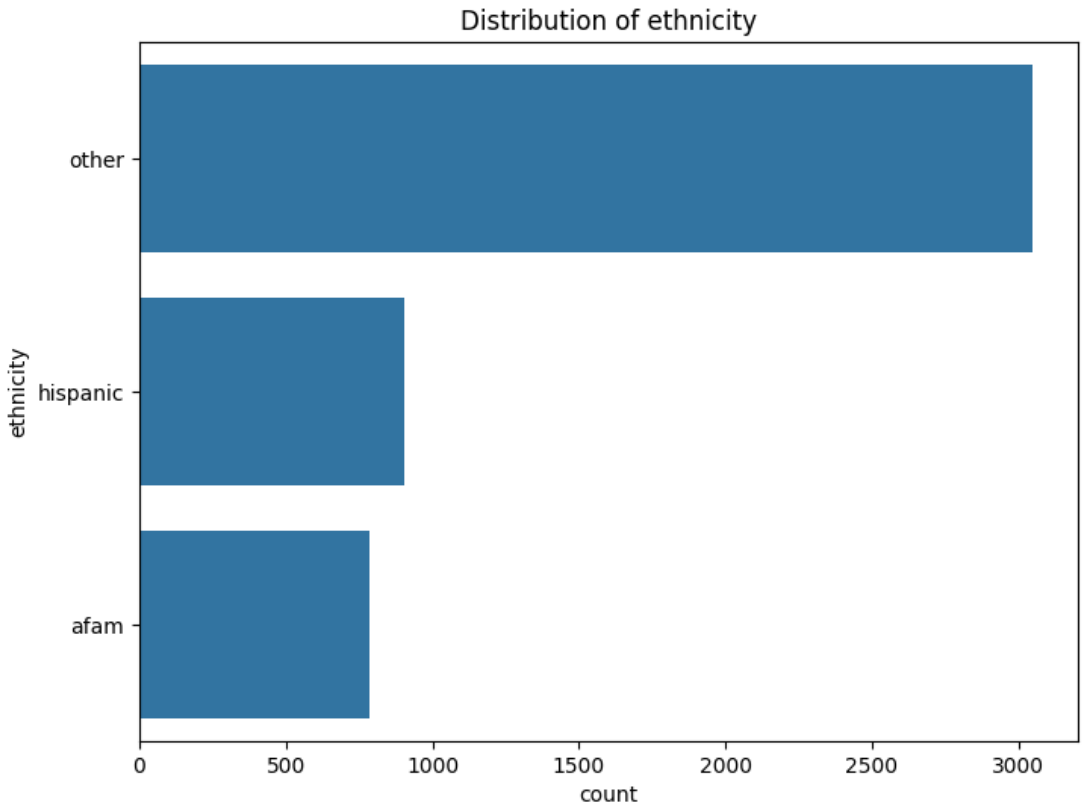
Zakres: 12 - 18

Zmienne kategoryczne:

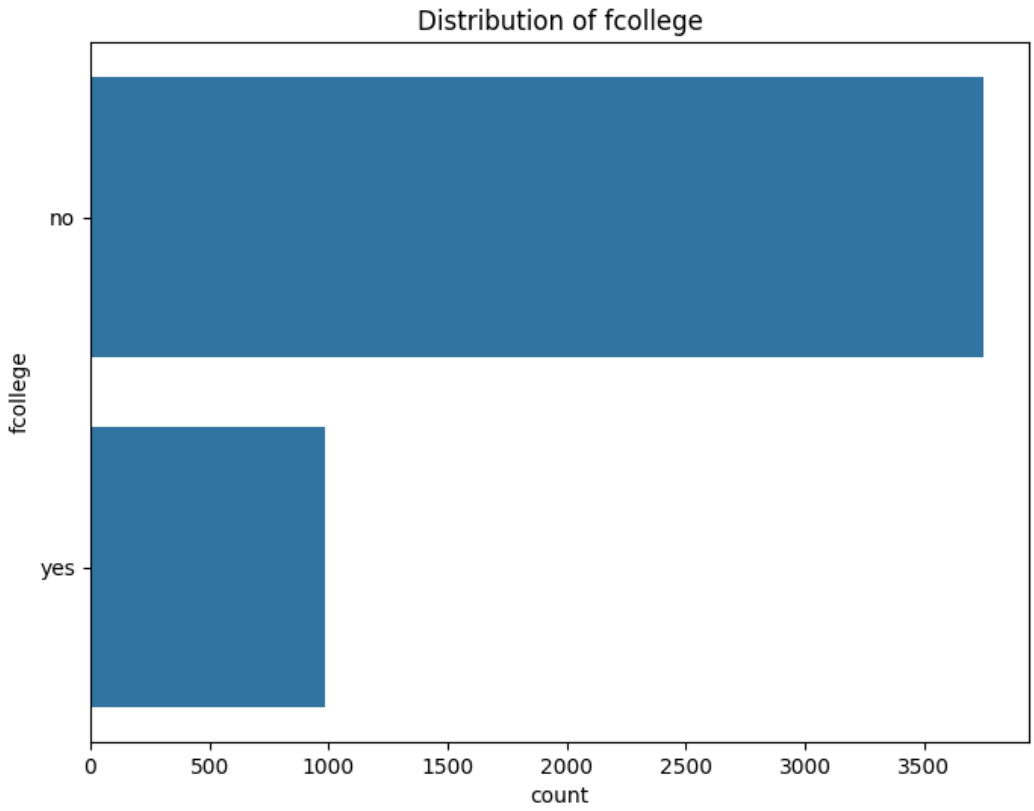
Płeć (gender): male, female

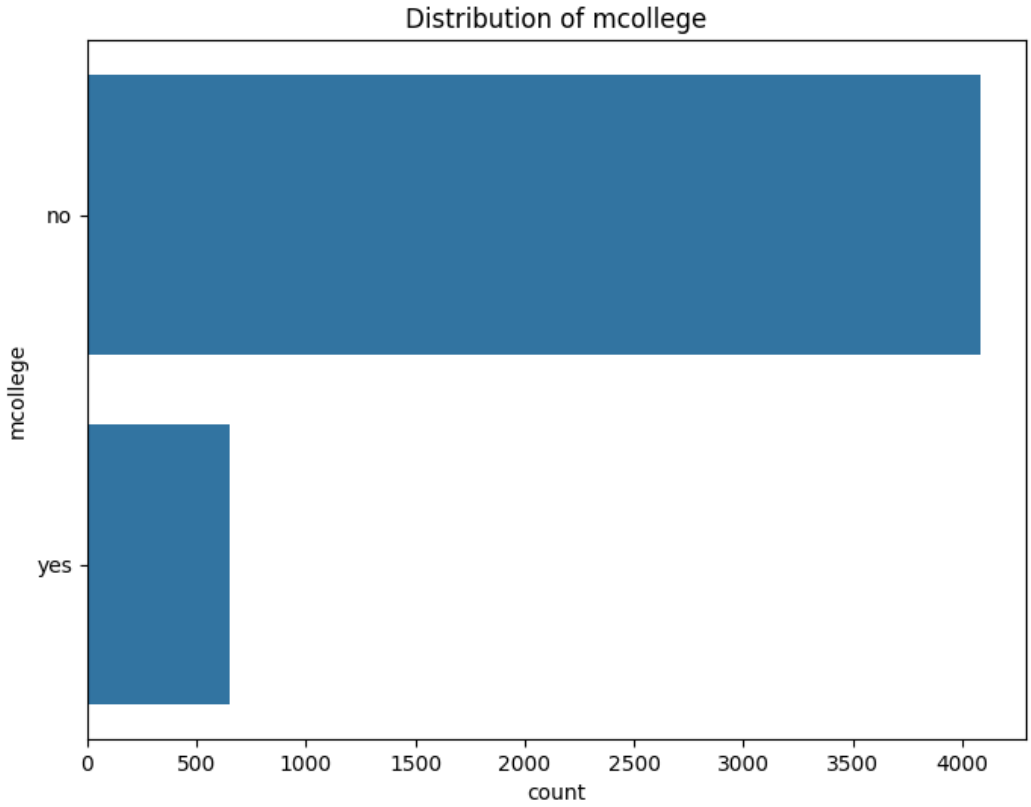


Etniczność (ethnicity): afam, hispanic, other

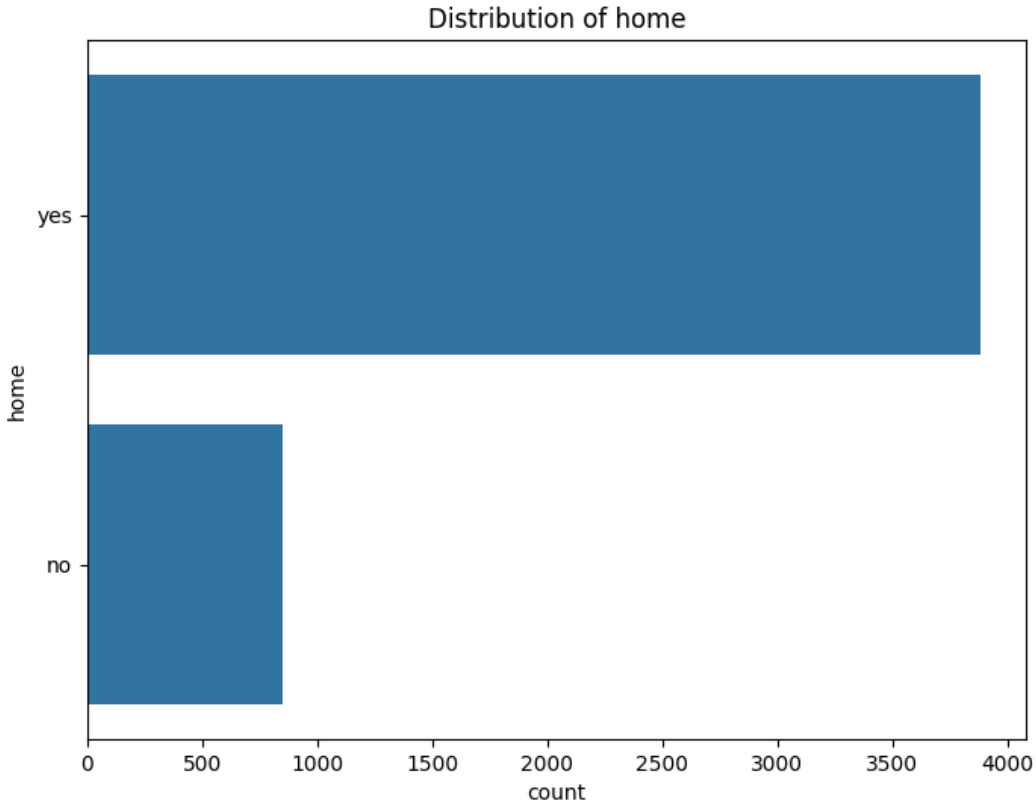

Distribution of ethnicity
other
hispanic
afam
ethnicity
0
500
1000
1500
2000
2500
3000
count

Rodzic ukończył studia (fcollege, mcollege): yes, no

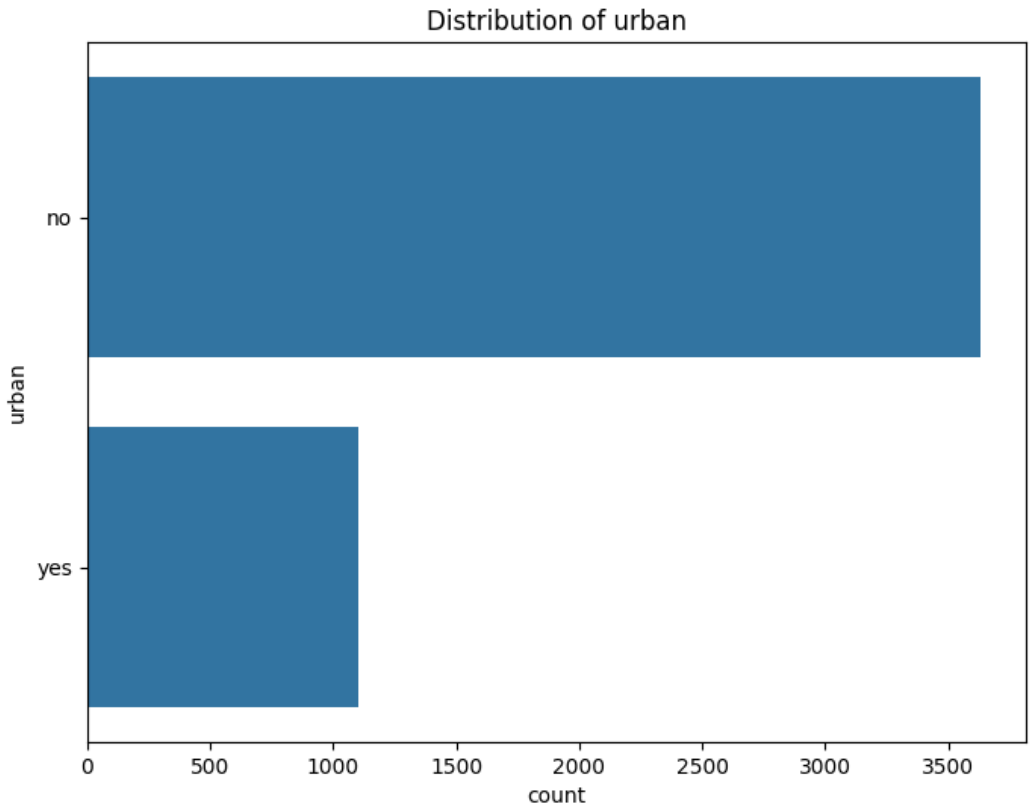




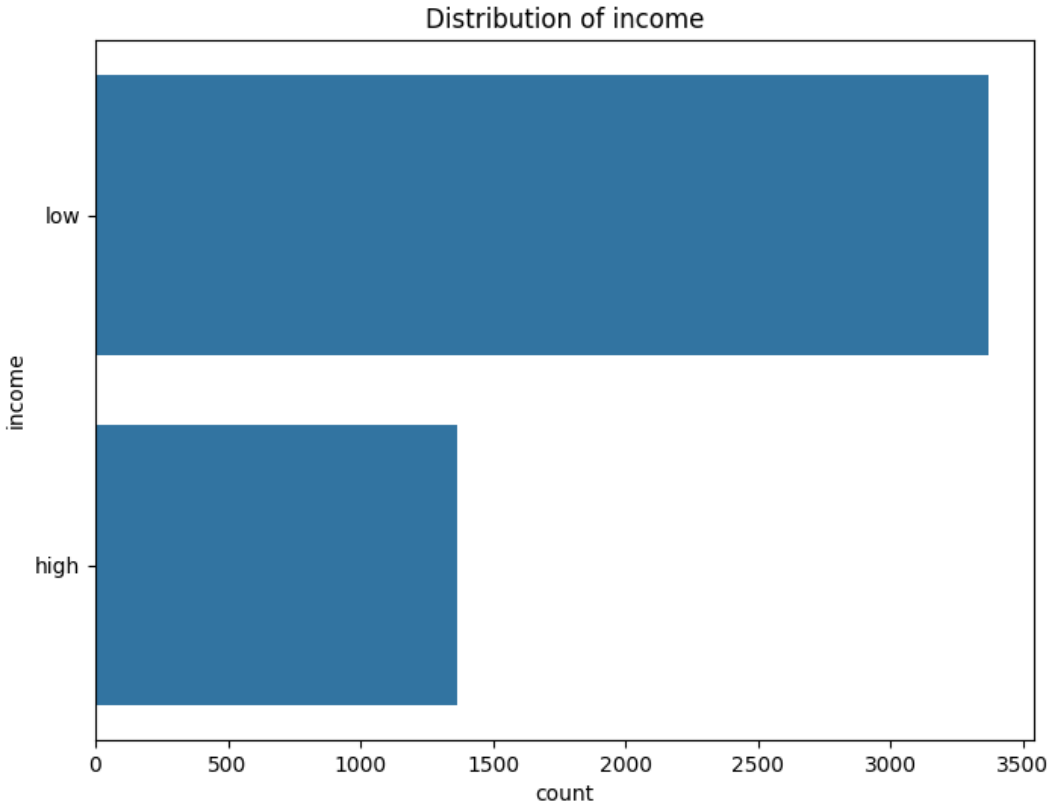
Dom (home): yes, no


Distribution of home
yes
no
home
0
500
1000
1500
2000
2500
3000
3500
4000
count

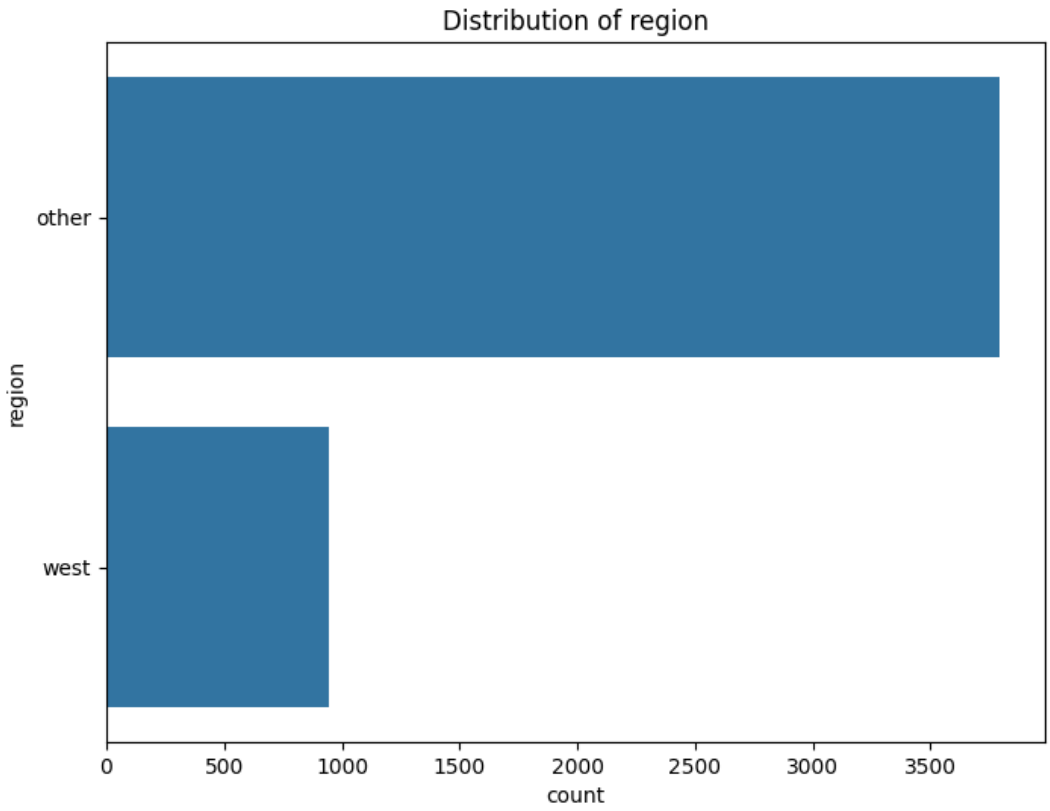
Obszar miejski (urban): yes, no



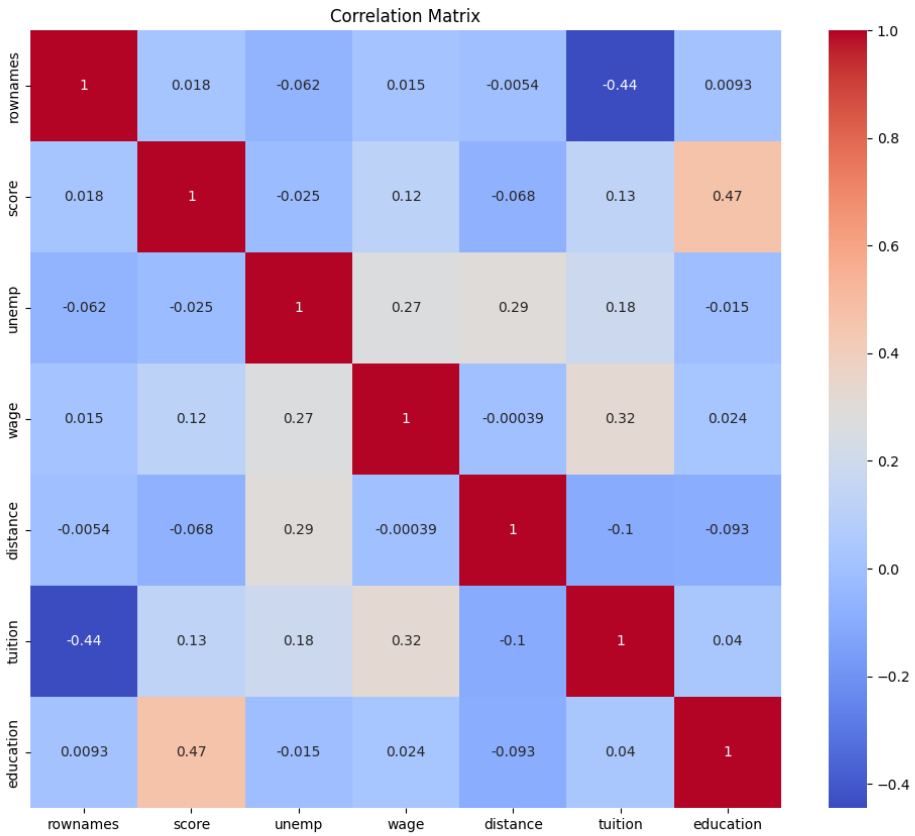
Dochód rodziny (income): high, low



Region (region): west, other



3. Wizualizacja Zależności Danych



Rozkład zmiennej score: Rozkład przypomina rozkład normalny z kilkoma wartościami odstającymi. Większość wyników znajduje się w przedziale od około 40 do 60.

Analiza korelacji: Zmienna score wykazuje umiarkowaną korelację z wage oraz tuition. Wyniki wskazują, że wyższe wynagrodzenie oraz czesne mogą być powiązane z wyższymi wynikami score.

4. Inżynieria Cech i Przygotowanie Danych

Podjęto następujące kroki w celu przygotowania danych:

Imputacja: Brakujące wartości w kolumnach numerycznych uzupełniono średnią, a w kategorycznych - najczęściej występującą wartością.

Standaryzacja: Zastosowano standaryzację dla zmiennych numerycznych, aby miały średnią 0 i odchylenie standardowe 1.

Podział Danych: Dane podzielono na zbiór treningowy (80%) i testowy (20%).

5. Wybór Modelu Predykcyjnego

Do przewidywania zmiennej score, wybrano regresję liniową jako podstawowy model. Główne cechy tego modelu:

Szybki i wydajny do treningu.

Zakłada liniową zależność, co może ograniczać skuteczność przy bardziej skomplikowanych zależnościach.

Dodatkowo, przetestowano modele Random Forest oraz Gradient Boosting jako alternatywy w celu oceny bardziej złożonych zależności aczkolwiek wyniki się zbytnio nie różniły.

6. Trenowanie Modelu

Model regresji liniowej osiągnął następujące wyniki:

Zbiór treningowy: $R^2 = 0.86$, MAE = 2.5

Zbiór testowy: $R^2 = 0.29$, MAE = 5.85

Obserwowana dysproporcja pomiędzy zbiorami wskazuje na potrzebę optymalizacji modelu.

7. Optymalizacja Modelu

Przeprowadzono optymalizację modelu poprzez:

Walidację krzyżową: Zastosowanie walidacji krzyżowej pozwala na uzyskanie bardziej wiarygodnych wyników.

Można zastosować hyperparametry w celu dalszej optymalizacji

Wyniki po optymalizacji:

Zbiór testowy: $R^2 = 0.35$, MAE = 5.71

8. Podsumowanie i Wnioski

Należy rozważyć dodatkowe zmienne lub inne metody modelowania, aby zwiększyć dokładność przewidywań.