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

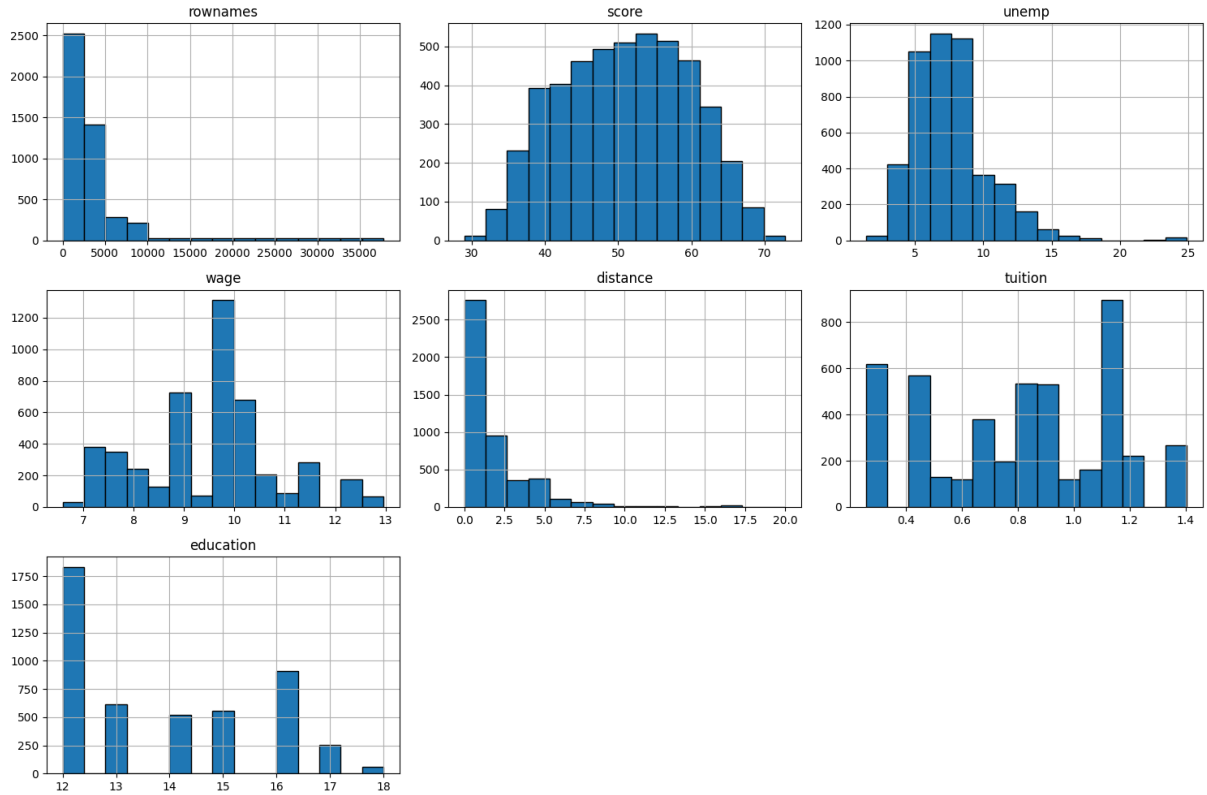
dane:

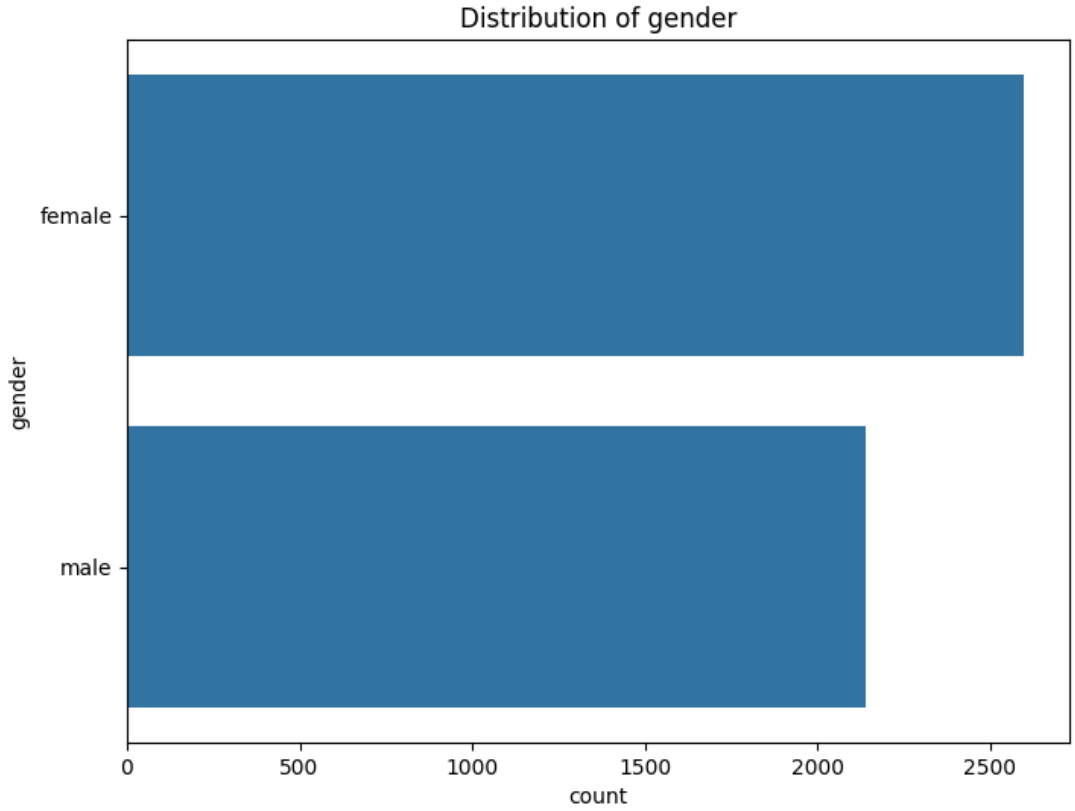
Zastosowano imputację brakujących wartości w kolumnach numerycznych przy użyciu średniej oraz najczęściej występującej wartości dla zmiennych kategorycznych.

Dodatkowo usunięto wiersze z więcej niż jednym brakującym polem.

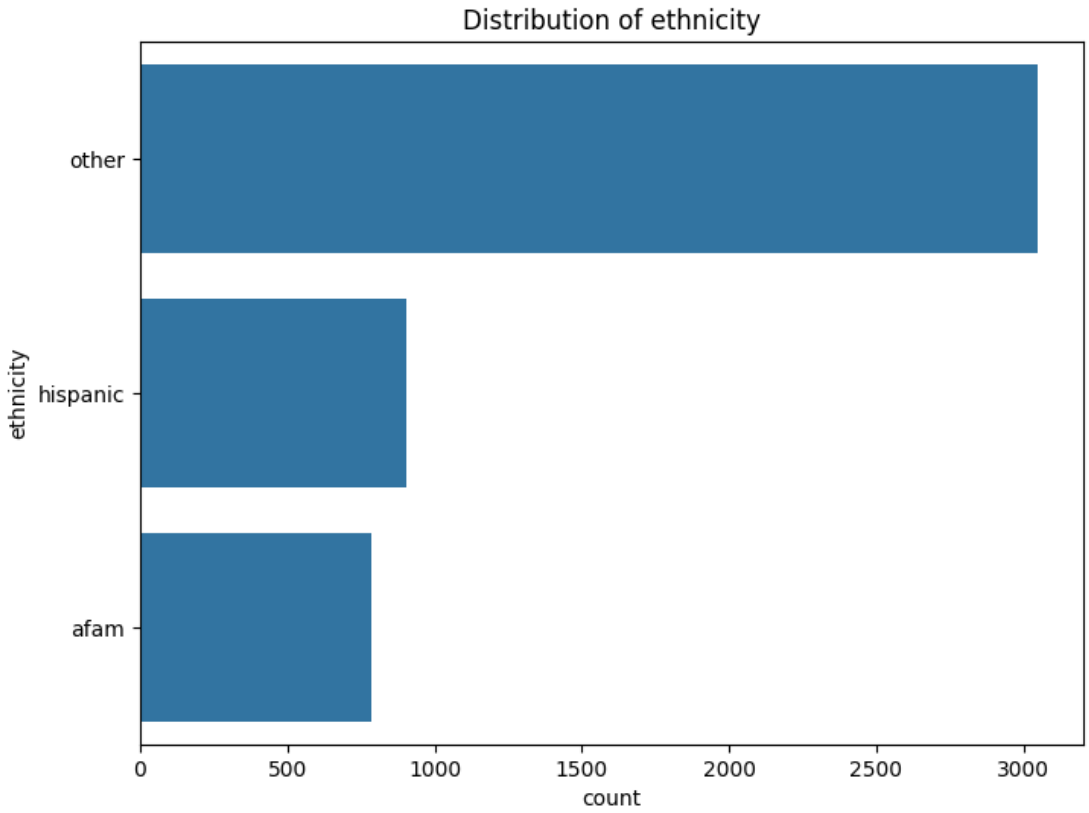
2. Analiza Statystyczna

Zmienne numeryczne:

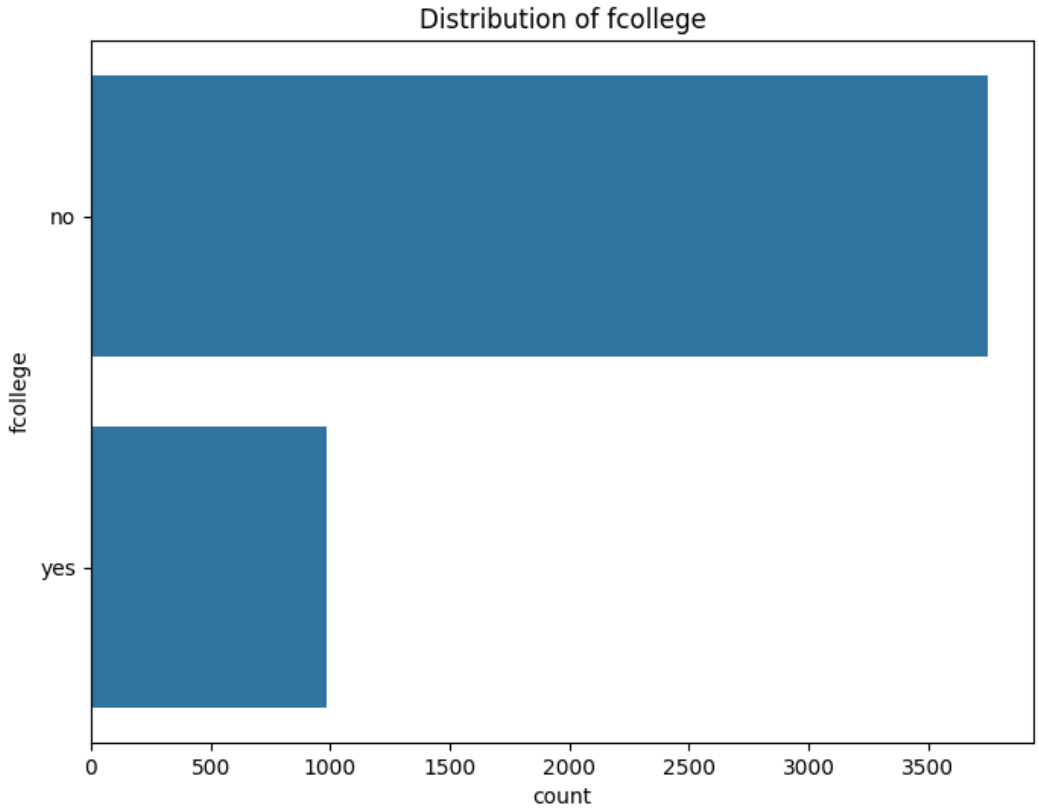


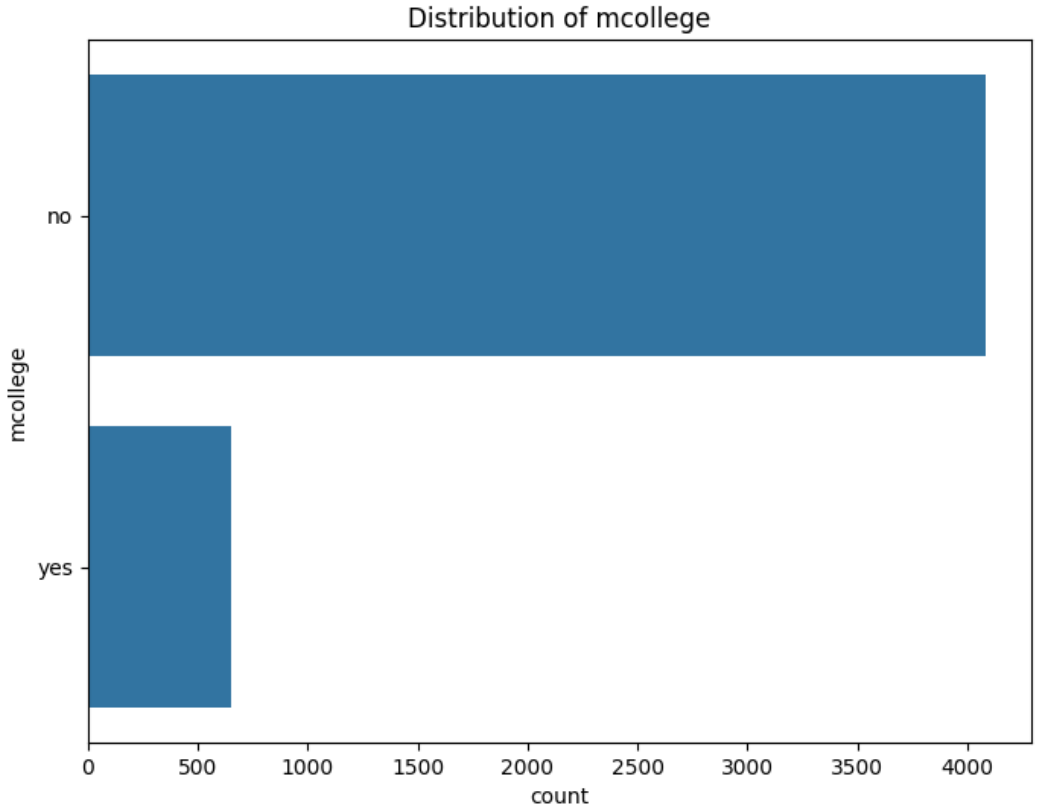


Etniczność (ethnicity): afam, hispanic, other

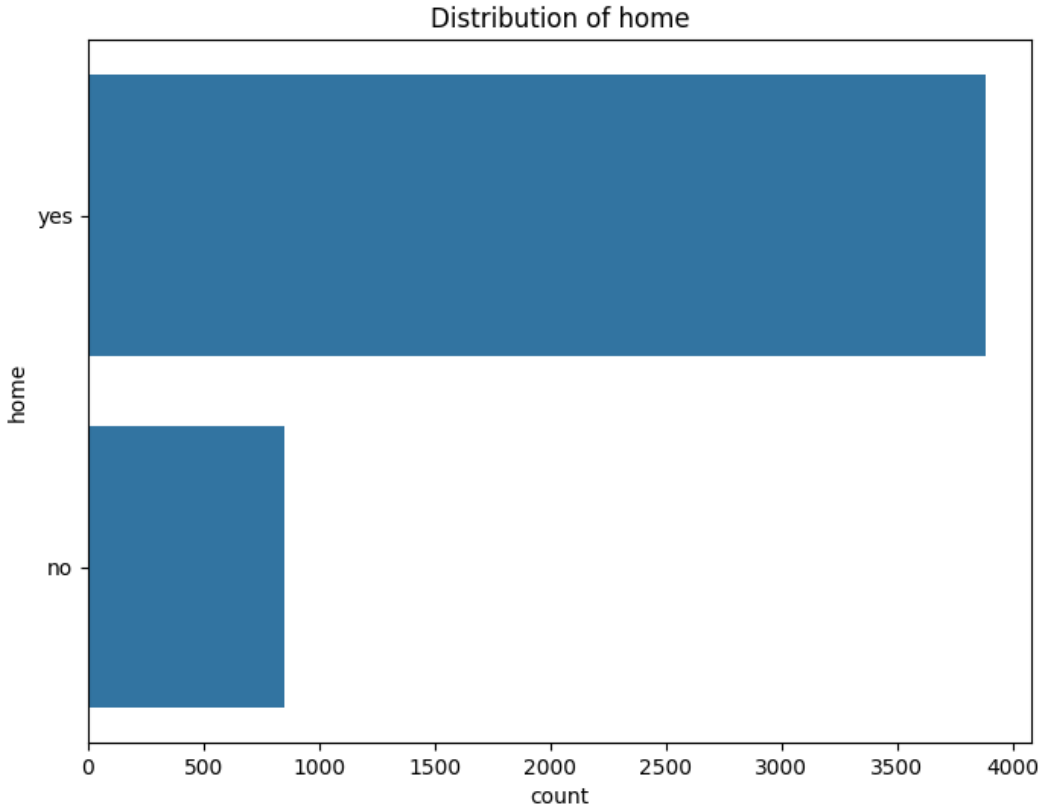

Distribution of ethnicity
other
hispanic
afam
ethnicity
0
500
1000
1500
2000
2500
3000
count

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

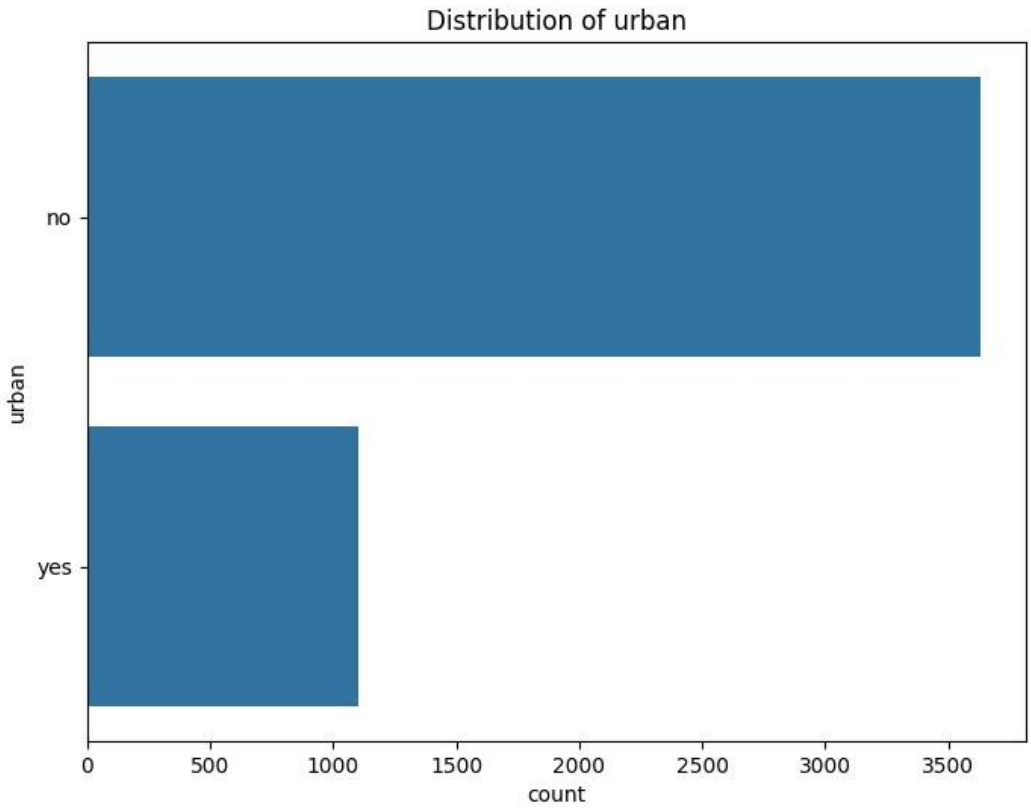




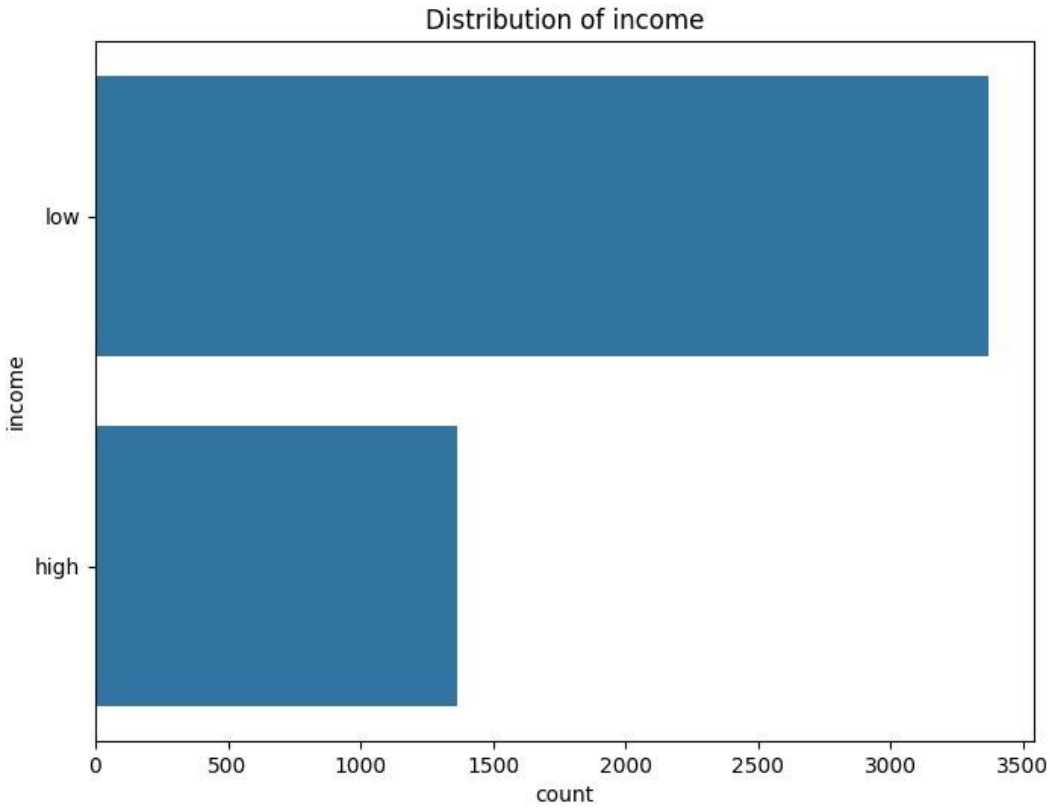
Dom (home): yes, no


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

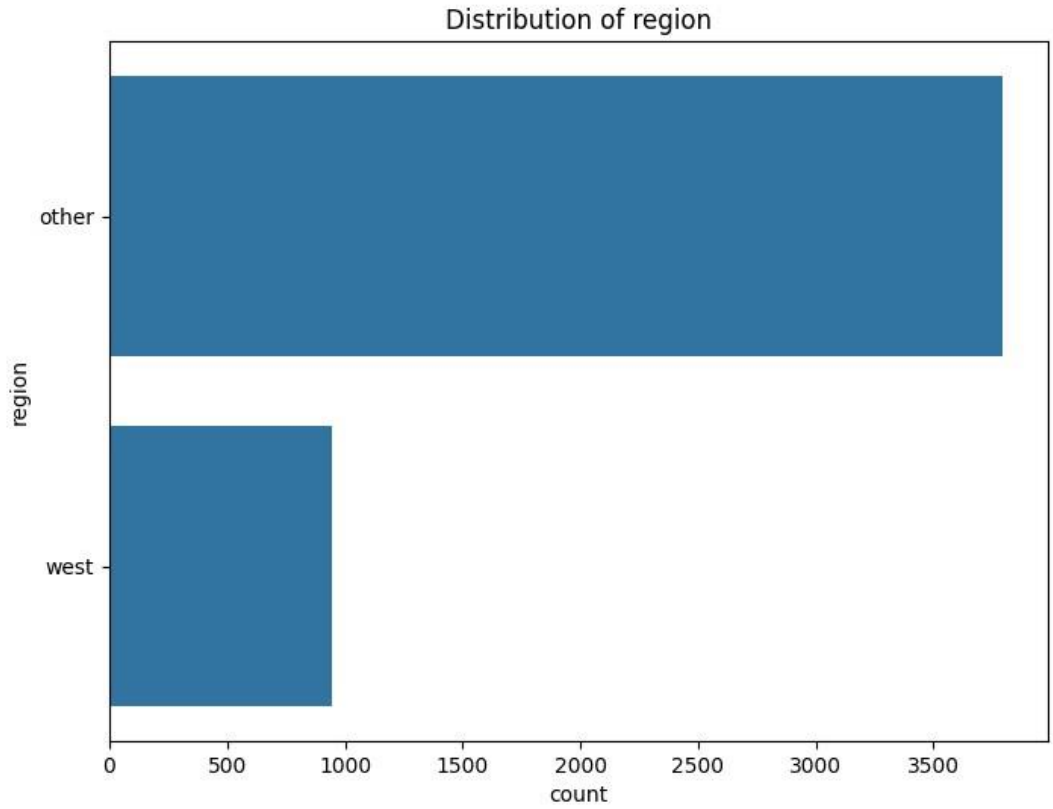
Obszar miejski (urban): yes, no



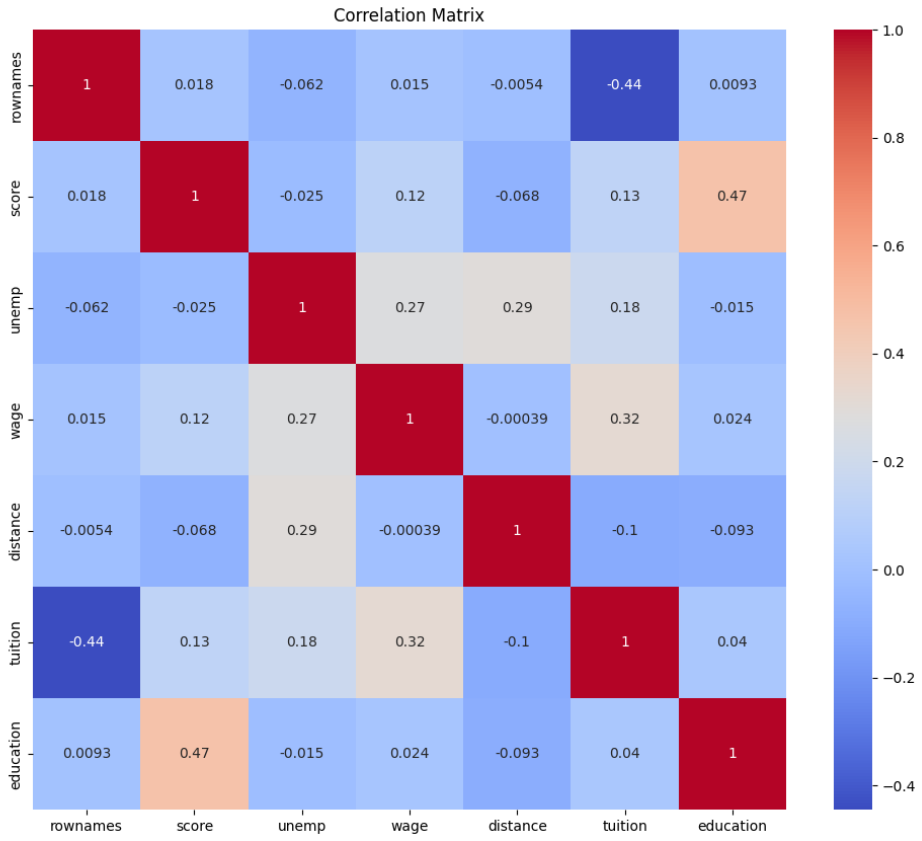
Dochód rodziny (income): high, low



Region (region): west, other



3. Wizualizacja Zależności Danych



Rozkład zmiennej score

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

Analiza korelacji

Zmienna score wykazuje umiarkowaną korelację z wage i tuition. Wyniki sugerują, że wyższe wynagrodzenie oraz czesne mogą być powiązane z wyższymi wartościami score.

Inżynieria Cech i Przygotowanie Danych

Podjęto następujące kroki w celu przygotowania danych:

- Imputacja: Brakujące wartości w kolumnach numerycznych zostały uzupełnione średnią, natomiast w kolumnach kategorycznych - najczęściej występującą wartością.

- Standaryzacja: Przeprowadzono standaryzację zmiennych numerycznych, aby miały średnią 0 i odchylenie standardowe 1.

- Podział Danych: Dane zostały podzielone na zbiór treningowy (80%) i testowy (20%).

Wybór Modelu Predykcyjnego

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

- Szybkość i wydajność podczas treningu.

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

Dodatkowo, przetestowano modele Random Forest oraz Gradient Boosting, aby ocenić bardziej złożone zależności, aczkolwiek wyniki nie różniły się znacząco.

Trenowanie Modelu

Obserwowana dysproporcja pomiędzy zbiorami wskazuje na konieczność optymalizacji modelu.

Optymalizacja Modelu

Model został zoptymalizowany poprzez:

- Walidację krzyżową: Zastosowanie walidacji krzyżowej pozwoliło uzyskać bardziej wiarygodne wyniki.

- Optymalizację hyperparametrów: Dalsza optymalizacja przy użyciu hyperparametrów.

Wyniki po optymalizacji:

- Zbiór testowy: R^2 = 0.3545, MAE = 5.7321

Podsumowanie i Wnioski

Aby zwiększyć dokładność przewidywań, warto rozważyć uwzględnienie dodatkowych zmiennych lub zastosowanie innych metod modelowania.