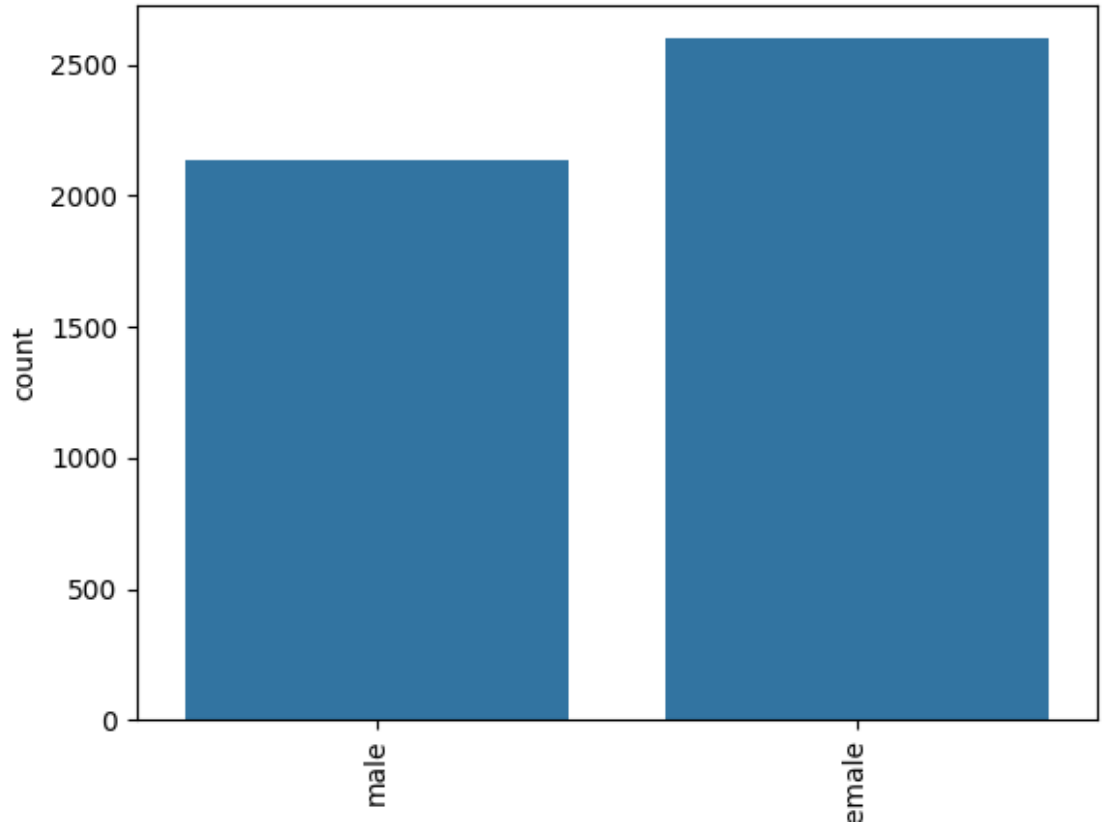
Dokumentacja LAB 3 – s24667

1.Eksploracja i wstępna analiza danych

CollegeDistance zawiera 4739 wierszy, 7 kolumn numerycznych i 8 kolumn kategorycznych. W całym zbiorze danych nie występują żadne braki. W kolumnie ‘distance’ pojawiają się wartości 0, aczkolwiek można to zinterpretować, że te osoby zakwaterowane są na terenie uczelni.

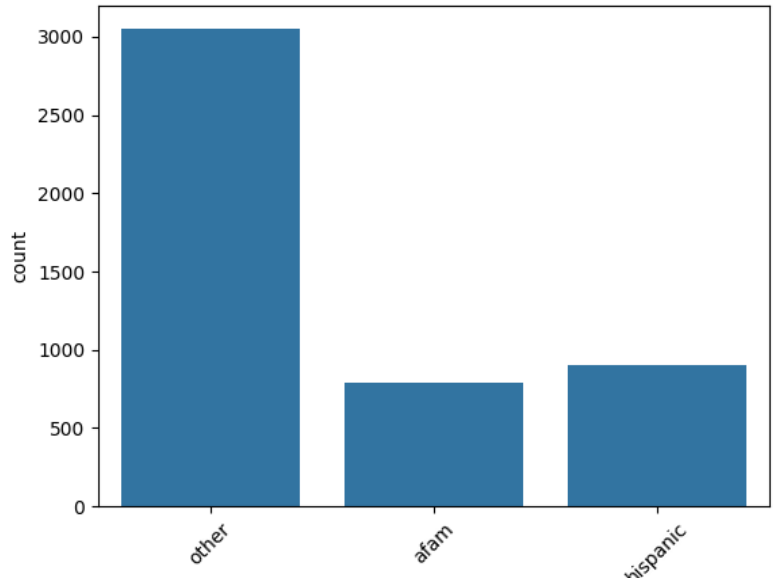
Opis poszczególnych cech:

-gender: płeć



Większa część uczniów w tym zbiorze danych jest płci żeńskiej

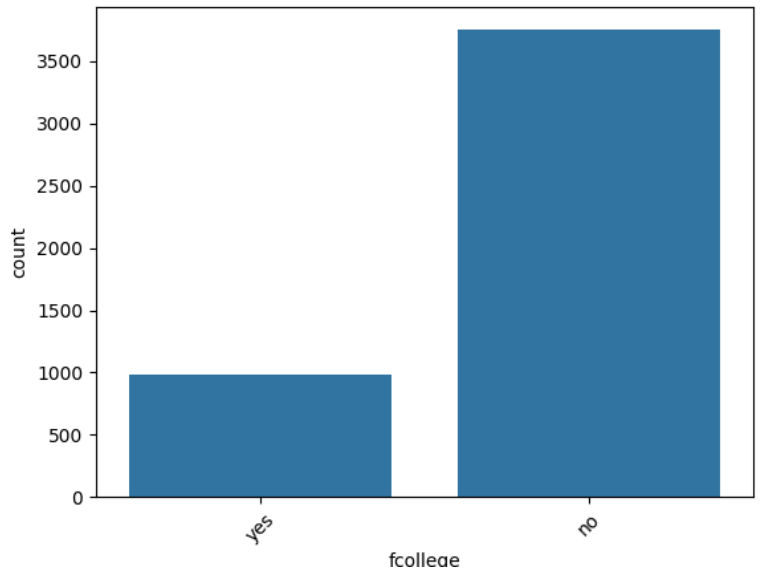
-ethnicity: pochodzenie etniczne



Większa część uczniów z tego zbioru danych nie jest pochodzenia afroamerykańskiego ani latynoskiego.

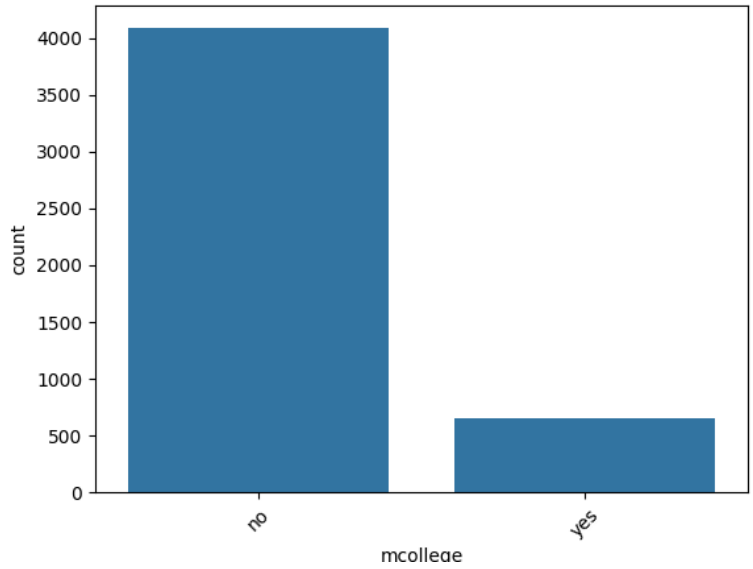
-score: wynik testu łączonego rocznego.

-fcollege: Czy ojciec osoby ukończył studia?



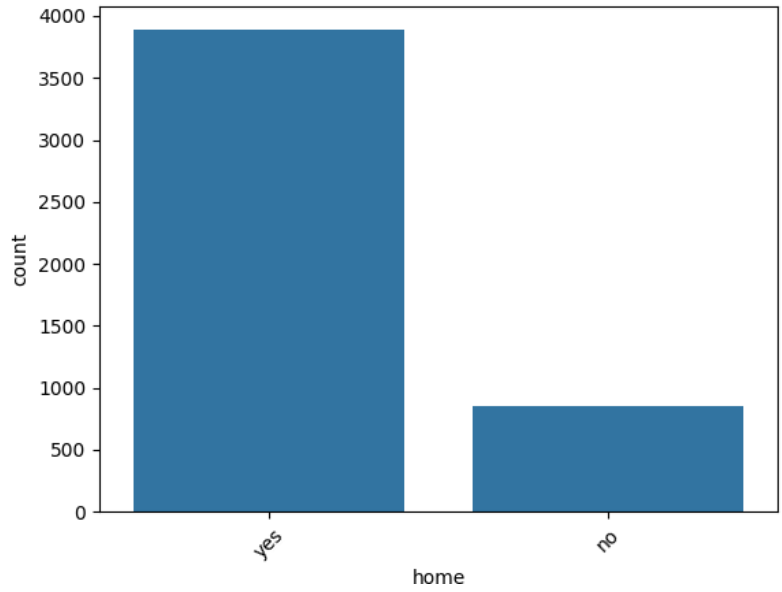
Około 80% matek uczniów ze zbioru danych nie ukończyło studiów

-mcollege: Czy matka osoby ukończyła studia?



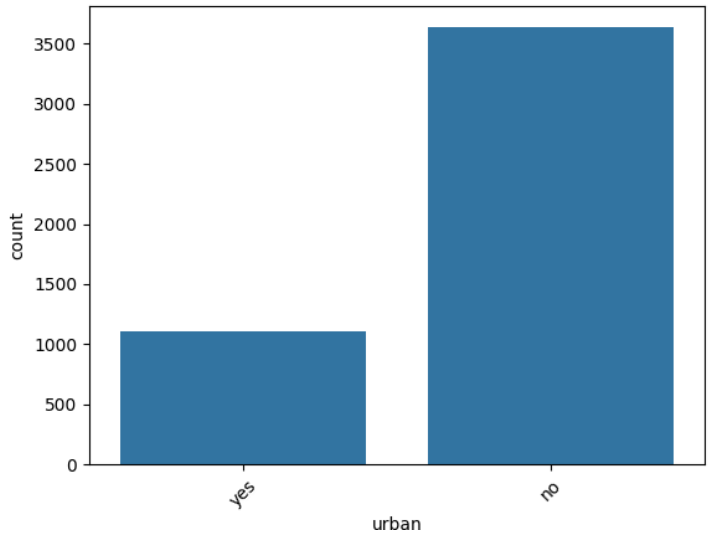
Około 85% ojców uczniów ze zbioru danych nie ukończyło studiów

-home: Czy rodzina osoby posiada własne mieszkanie?



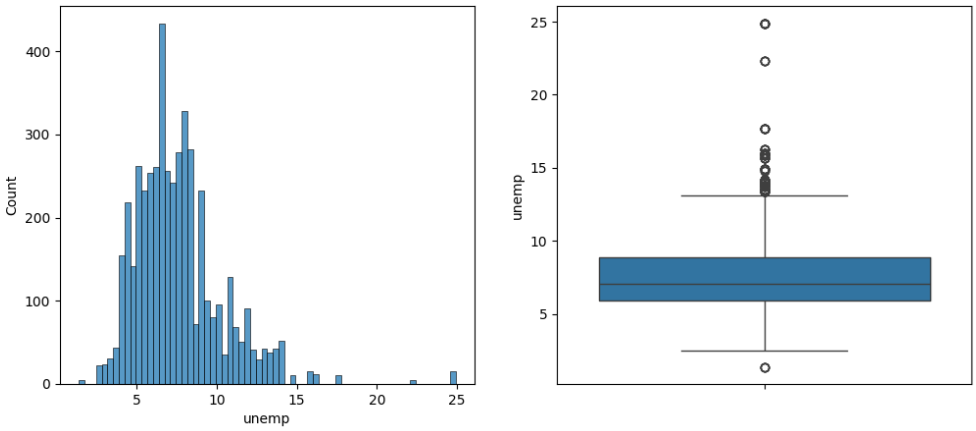
Około 85% rodzin uczniów ze zbioru danych posiada własne mieszkanie

-urban: Czy szkoła znajduje się w obszarze miejskim



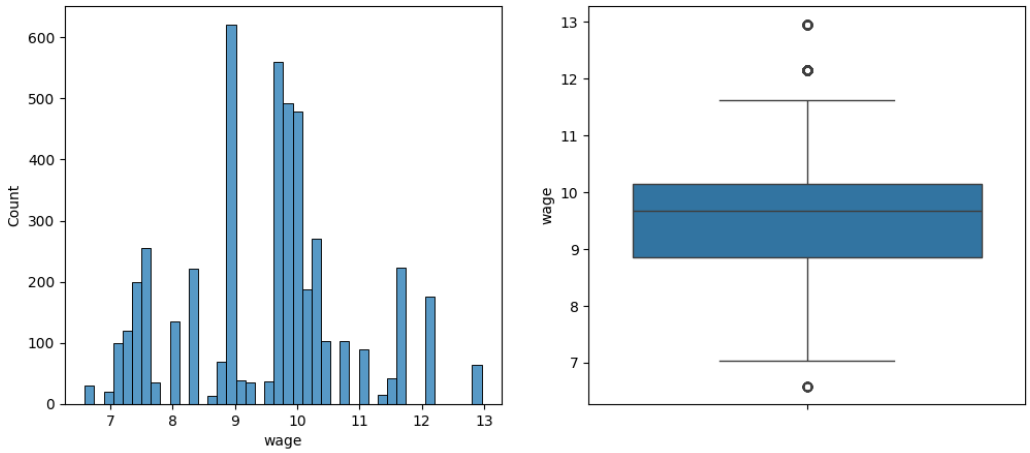
Około 75% szkół do których uczęszczają uczniowie ze zbioru danych znajduje się na obszarze miejskim

-unemp: Stopa bezrobocia w powiecie w 1980



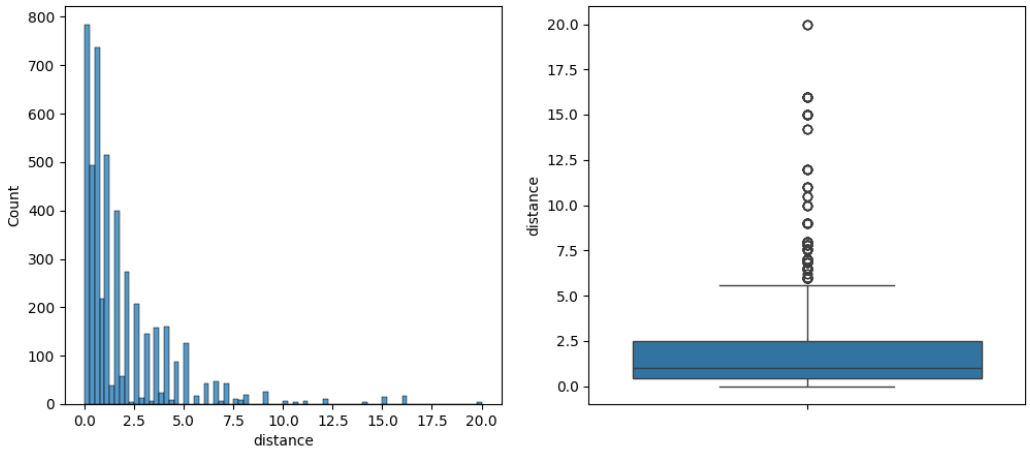
Powiat ucznia o stopie procentowej powyżej 13 odstaje od normy, w większości przypadków wynosi między 6-9

-wage: państwowa stawka godzinowa w przemyśle w 1980



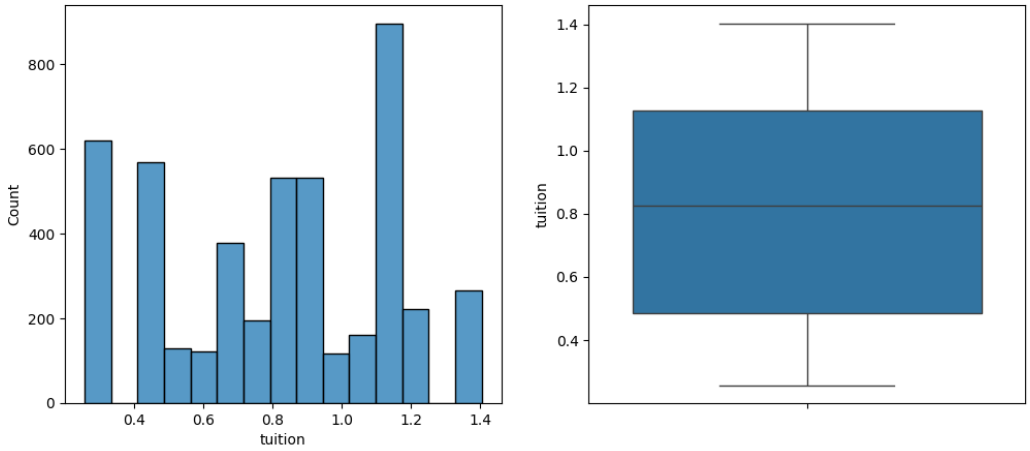
Średnia wartość jest w zakresie od 9 do 10 USD, stawka powyżej 12 i poniżej 7 odstaje od normy.

-distance: odległość od czteroletniej uczelni (w 10 milach)



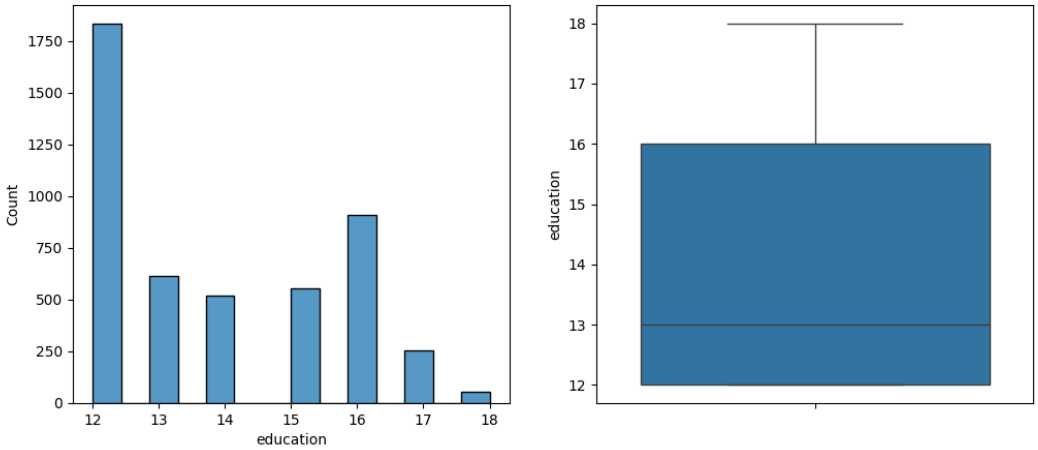
W większości przypadków, uczeń miał między 0 a 25 mil do najbliższej czteroletniej uczelni. Uczniowie z odległością powyżej 50 odstają od normy.

-tuition: średnie czesne na czteroletniej uczelni państwowej (w 1000 USD)



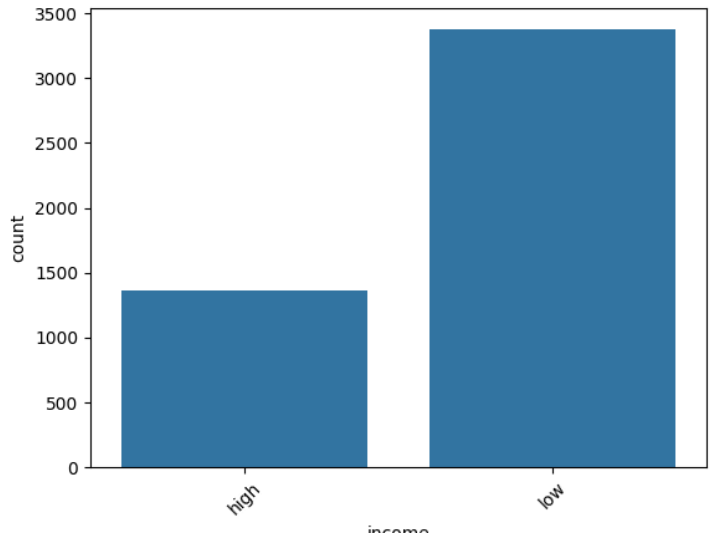
W przypadku większości uczelni, czesne wynosiły między 3000 – 11000 USD

-education: liczba lat edukacji



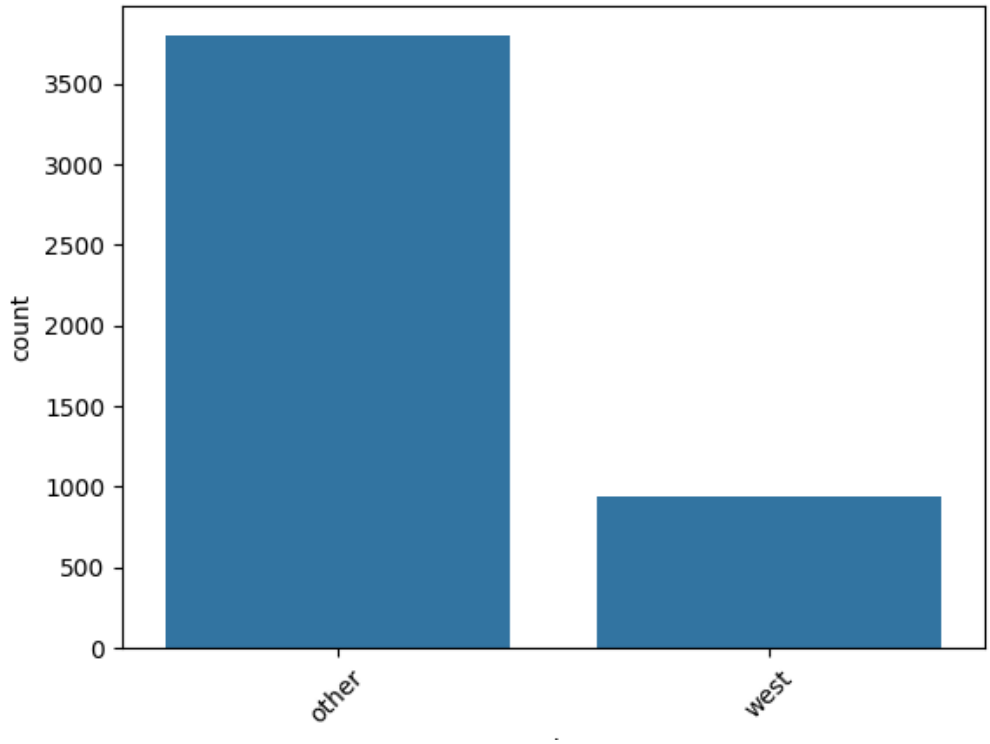
W większości przypadków uczniów ze zbioru danych, liczba lat edukacji wynosi między 12 a 16

-income Czy rodzina zarabia powyżej 25000USD rocznie?

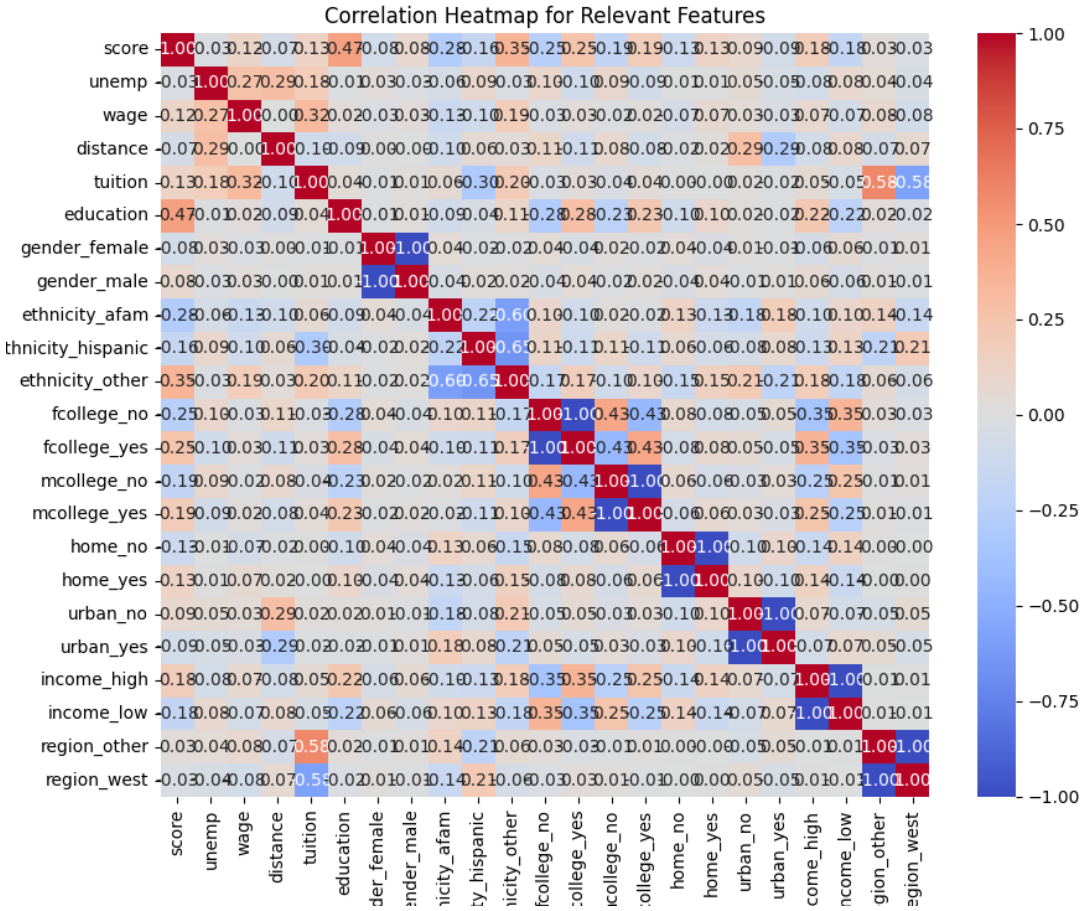


Około 70% rodzin uczniów ze zbioru danych zarabia poniżej 25000USD rocznie.

-region Region (West, other)



Około 20% uczniów ze zbioru danych pochodzi z zachodniej części Ameryki.



Największą korelację z cechą score posiada: education, ethnicity_other oraz fcollege_no

2.Inżynieria cech i przygotowanie danych

Wykonano następujące działania w celu przygotowania danych:

- Kategoryzacja: wykorzystano OneHotEncoder w celu przekształcenia wartości kategorycznych w wartości binarne

- Imputacja: w przypadku wystąpienia brakujących wartości następuje ich uzupełnienie:

    -dla wartości numerycznych, uzupełniane są średnią wartości w kolumnie

    -dla wartości kategorycznych, uzupełniane są najczęściej pojawiającą się kategorią

- Standaryzacja: dokonano standaryzacji danych numerycznych przy pomocy StandardScaler

- Podział danych na zbiór treningowy i testowy

3.Wybór i trenowanie modelu

RANDOM FOREST - Wybrany został model lasu losowego, dobrze radzi sobie z danymi o wysokiej złożoności oraz nieliniowością.

4.Ocena i optymalizacja modelu

Uzyskane wyniki dla zbioru testowego:
MAE = 5.8995341170264295 R^2
= 0.2542072511350574

W celu polepszenia uzyskanych wyników, została wykonana optymalizacja modelu.
Przeprowadzono tunowanie hiperparametrów i walidację krzyżową przy pomocy GridSearchCV

Uzyskane wyniki po optymalizacji modelu:
MAE = 5.6972939478006
R^2 = 0.329420189183869