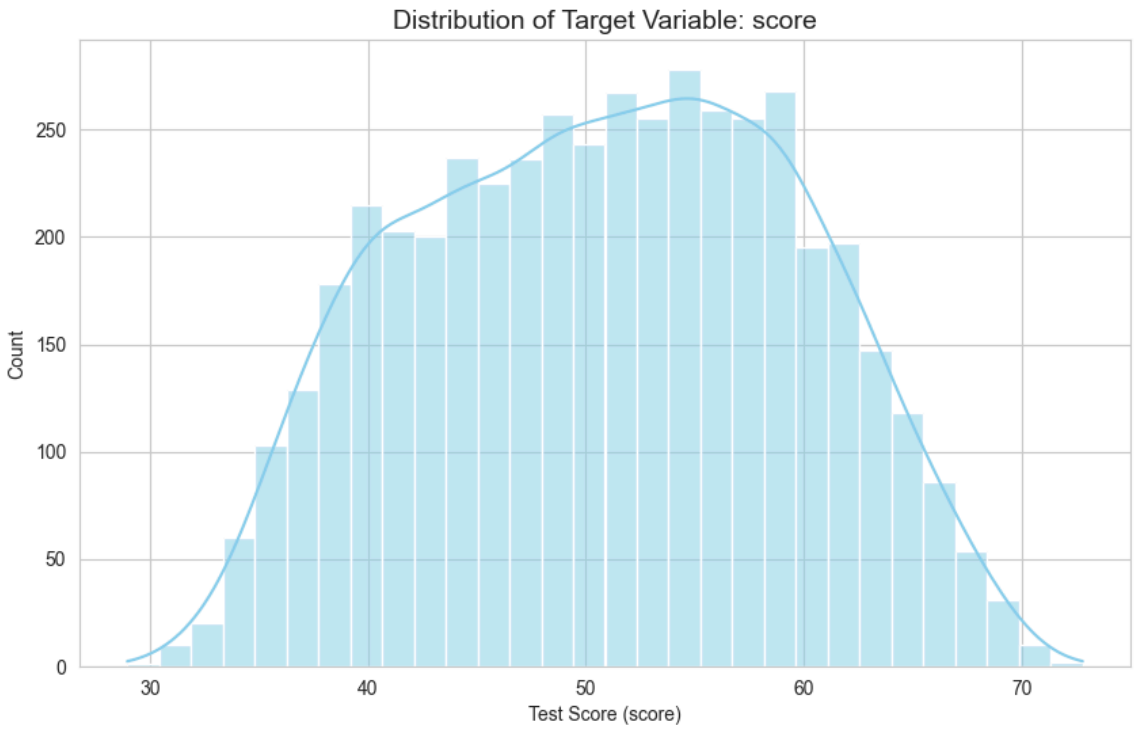
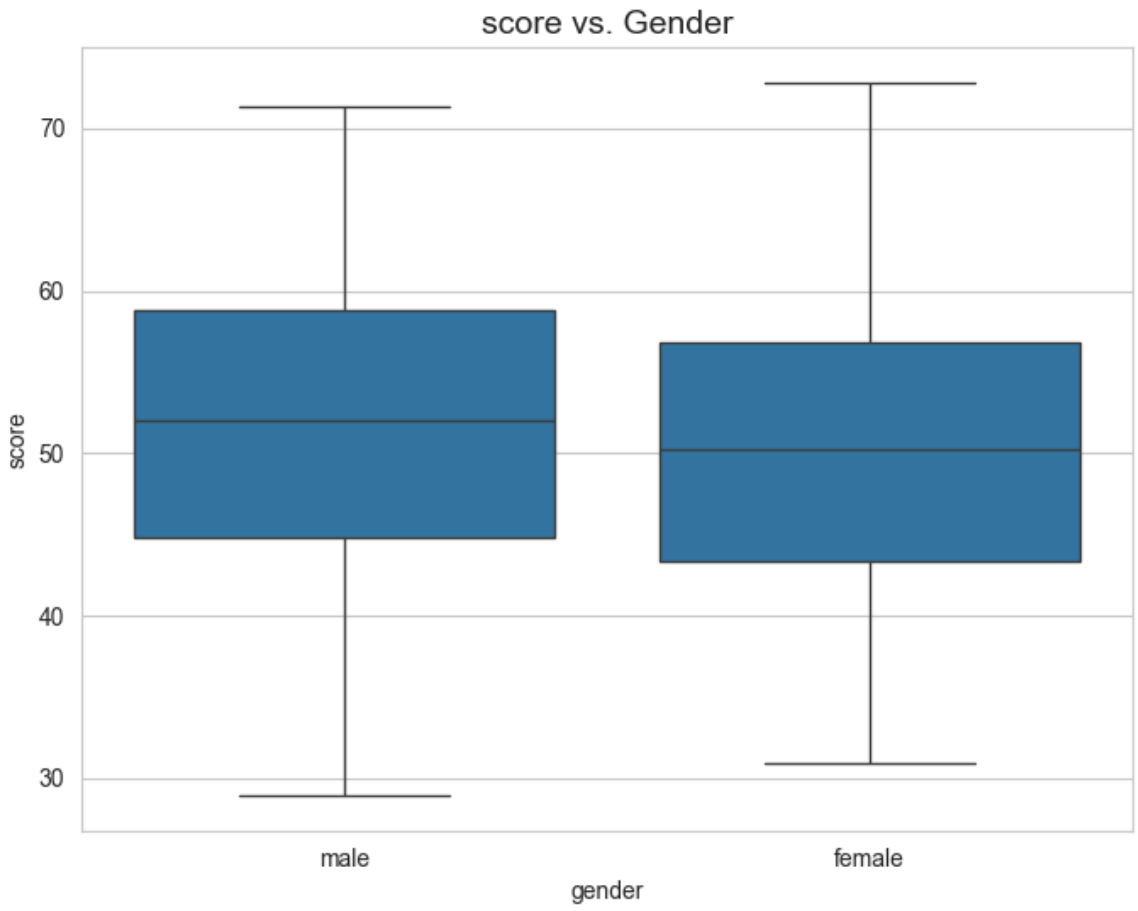
1. Wstępna Analiza Danych (EDA)

Wczytanie i braki: Załadowano zbiór `CollegeDistance`. **Nie zidentyfikowano żadnych brakujących wartości (NaN)**, więc dane były od razu czyste.

Charakterystyka: Najsilniejsza liniowa zależność występuje między `score` a `education`.




Correlation Matrix of Numerical Features
score
unemp
wage
distance
tuition
education
1.00 -0.03 0.12 -0.07 0.13 0.47
-0.03 1.00 0.27 0.29 0.18 -0.01
0.12 0.27 1.00 -0.00 0.32 0.02
-0.07 0.29 -0.00 1.00 -0.10 -0.09
0.13 0.18 0.32 -0.10 1.00 0.04
0.47 -0.01 0.02 -0.09 0.04 1.00
score
unemp
wage
distance
tuition
education
1.0
0.8
0.6
0.4
0.2
0.0



---

2. Przygotowanie Danych i Inżynieria Cech

Standaryzacja: Cechy numeryczne przeskalowano za pomocą `StandardScaler`.

Kodowanie: Cechy kategoryczne ('gender', 'ethnicity', 'fcollege', 'mcollege', 'home', 'urban', 'region') zakodowano metodą One-Hot Encoding.

Podział: Dane podzielono na zbiory **Treningowy (80%)** i **Testowy (20%)**.

---

3. Wybór i Wyniki Modeli

Wytrenowano trzy modele regresji (Regresja Liniowa, Random Forest, Gradient Boosting) w celu wyboru najlepszego algorytmu.

Wyniki R2 (Test Set):

```
- Regresja Liniowa: 0.3523

- Random Forest: 0.2924

- Gradient Boosting: 0.3659
```

Wybór: Model **Gradient Boosting** osiągnął najwyższy wynik R2 i został wybrany jako model końcowy, ponieważ najlepiej dopasowuje się do nieliniowych wzorców w danych.

---

4. Ocena Końcowa i Wnioski

Najlepszy Model: Gradient Boosting Regressor (R2=0.3659).

Końcowe Metryki (Test Set):

```
- R2: 0.3659

- MAE: 5.7015

- MSE: 48.0831
```

Wnioski: Uzyskany R2 jest umiarkowany. Sugeruje to, że brakuje kluczowych czynników predykcyjnych poza dostarczonymi danymi (np. czynników psychologicznych, indywidualnej motywacji).