1. 동일한 test set에 대한 4가지 알고리즘의 RMSE를 표로 나타내고, RMSE 순위 에 대한 이유를 서술

	Model	RMSE
0	Linear Regression	10.68500
1	Regression Tree	10.95076
2	Visual tree	10.56225
3	Hybrid tree regression	10.68500

순위 1위: Visual Regression tree

공동 2위: Linear Regression, Hybrid tree regression

4위: Regression Tree

이유: RMSE가 낮을수록 성능이 좋으므로 RMSE가 낮은 순으로 순위를 매겼다

2. 모델 해석

- (1) Linear Regression 해석

선형회귀방정식

(Height)=89.7592+(Gender)×19.9140+(Weight)×0.8220

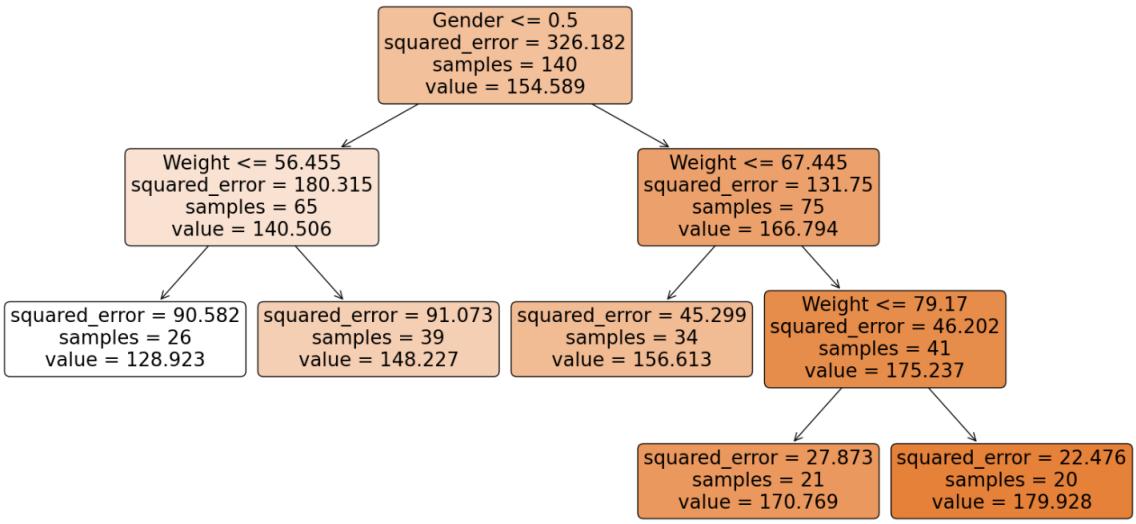
- Gender가 Male(Gender=1)일 때 키가 19.9140cm만큼 오를 것이다라고 해석이 가능함.
- Weight가 한 단위 증가하면 키가 0.8220cm만큼 오를

것이라고 해석이 가능함

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Height   R-squared:                       0.818
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     307.5
Date:                Wed, 10 Apr 2024   Prob (F-statistic):           2.22e-51
Time:                        13:12:18   Log-Likelihood:                -484.59
No. Observations:                 140   AIC:                             975.2
Df Residuals:                     137   BIC:                             984.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         89.7592      3.565     25.178      0.000      82.710      96.809
Gender        19.9140      1.389     14.335      0.000      17.167      22.661
Weight         0.8220      0.056     14.788      0.000       0.712       0.932
==============================================================================
Omnibus:                        3.666   Durbin-Watson:                   1.900
Prob(Omnibus):                  0.160   Jarque-Bera (JB):                3.785
Skew:                           0.166   Prob(JB):                        0.151
Kurtosis:                       3.734   Cond. No.                         363.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

- (2) Regression Tree 해석



- 의사결정트리 그림으로 Gender가 0.5이하이고 Weight가 56.455이하이면 키가 128.923cm라고 해석이 가능하

다.

- (3) Visual Regression Tree 해석

```python
from sklearn.linear_model import LinearRegression

# 왼쪽 리프 노드 데이터셋에 대한 선형 회귀 모델 생성 및 적합
lr_model_left = LinearRegression()
lr_model_left.fit(left_leaf_X_train[['Weight']], left_leaf_y_train)

# 오른쪽 리프 노드 데이터셋에 대한 선형 회귀 모델 생성 및 적합
lr_model_right = LinearRegression()
lr_model_right.fit(right_leaf_X_train[['Weight']], right_leaf_y_train)

# 왼쪽 리프 노드의 선형 회귀 계수 및 절편
left_leaf_coeff = lr_model_left.coef_[0]
left_leaf_intercept = lr_model_left.intercept_

# 오른쪽 리프 노드의 선형 회귀 계수 및 절편
right_leaf_coeff = lr_model_right.coef_[0]
right_leaf_intercept = lr_model_right.intercept_

print("왼쪽 리프 노드 선형 회귀 계수 및 절편:", left_leaf_coeff, left_leaf_intercept)
print("오른쪽 리프 노드 선형 회귀 계수 및 절편:", right_leaf_coeff, right_leaf_intercept)
```

```
왼쪽 리프 노드 선형 회귀 계수 및 절편: 0.7381624087037346 94.93438428051107
오른쪽 리프 노드 선형 회귀 계수 및 절편: 0.9075239687629273 103.72940873121418
```

- 위 그림은 Female(Left)과 Male(Right)로 나눈 다음에 각각 회귀 방정식을 구한 것이다.

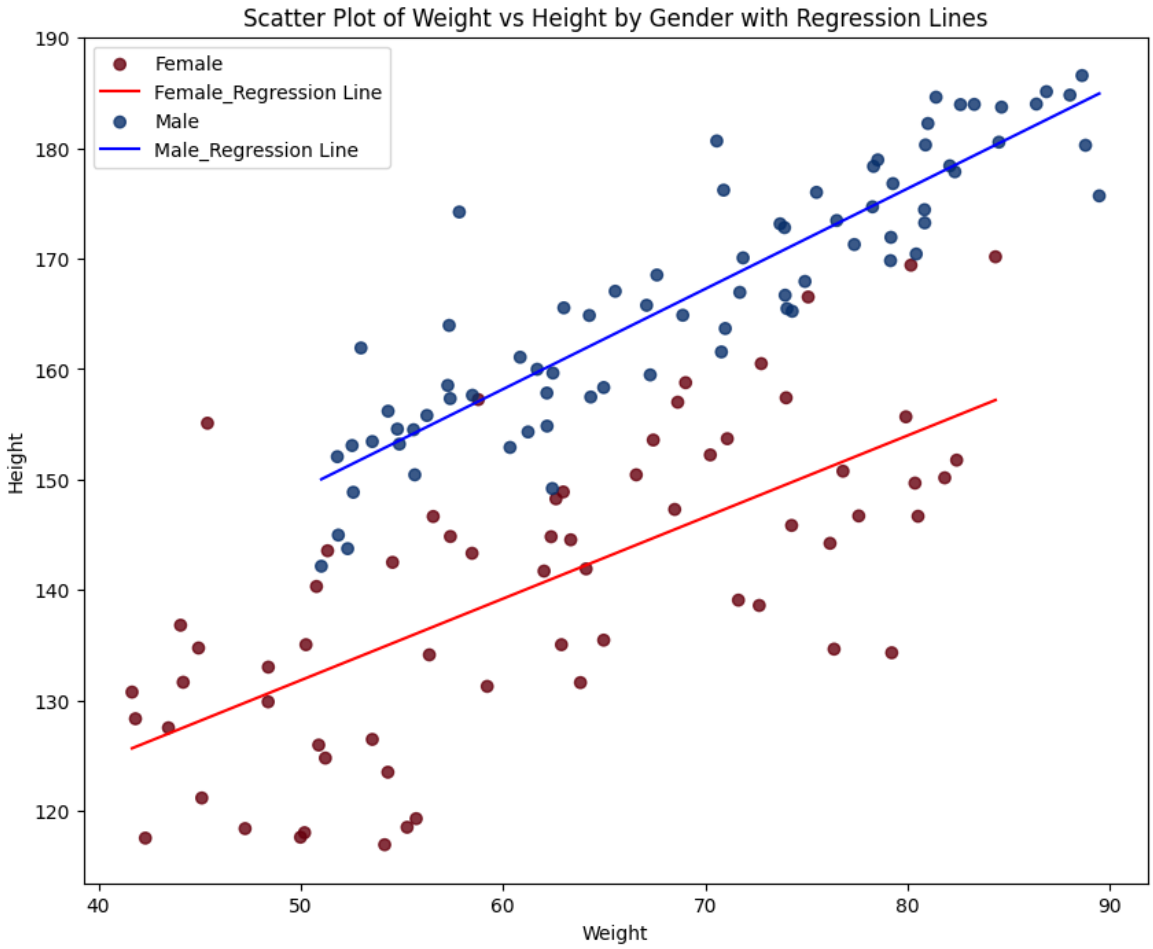
  Female: (Height)=0.738×(Weight)+94.934

  해석: Weight가 한 단위 증가하면 0.738cm만큼 키가 클 것이다라고 해석이 가능함

  Male: (Height)=0.9075×(Weight)+103.729

  해석: Weight가 한 단위 증가하면 0.9075cm만큼 키가 클 것이다라고 해석이 가능함

- 아래 그림은 Female/male 별로 Weight에 따른 Height를 시각화한것이다.



해석: 두 범주(Female, Male) 모두 Weight과 Height이 양의 상관관계가 있음을 그래프를 통해 확인할 수 있다.

- (4) Hybrid decision tree and regression 해석

Hybrid decision tree and regression 은 기존의 Linear regression를 보완한 알고리즘이다. 잔차에 대한 회귀식은 다음과 같다.

(Residual)=(Gender) × (-6.374)+(Weight) × (0.822)인데

해석을 하자면 성별이 남성(Gender=1)일 때가 잔차가 더 작음을 알 수 있다. 따라서 예측값의 편차가 여성보다는 남성일때가 더 작다는 것을 확인할 수 있다.

```
from sklearn.linear_model import LinearRegression

# 선형 회귀 모델 생성 및 적합

lr_model2 = LinearRegression()
lr_model2.fit(pd.DataFrame(new_train_X),pd.DataFrame(residuals_train))
```

```
▾ LinearRegression
LinearRegression()
```

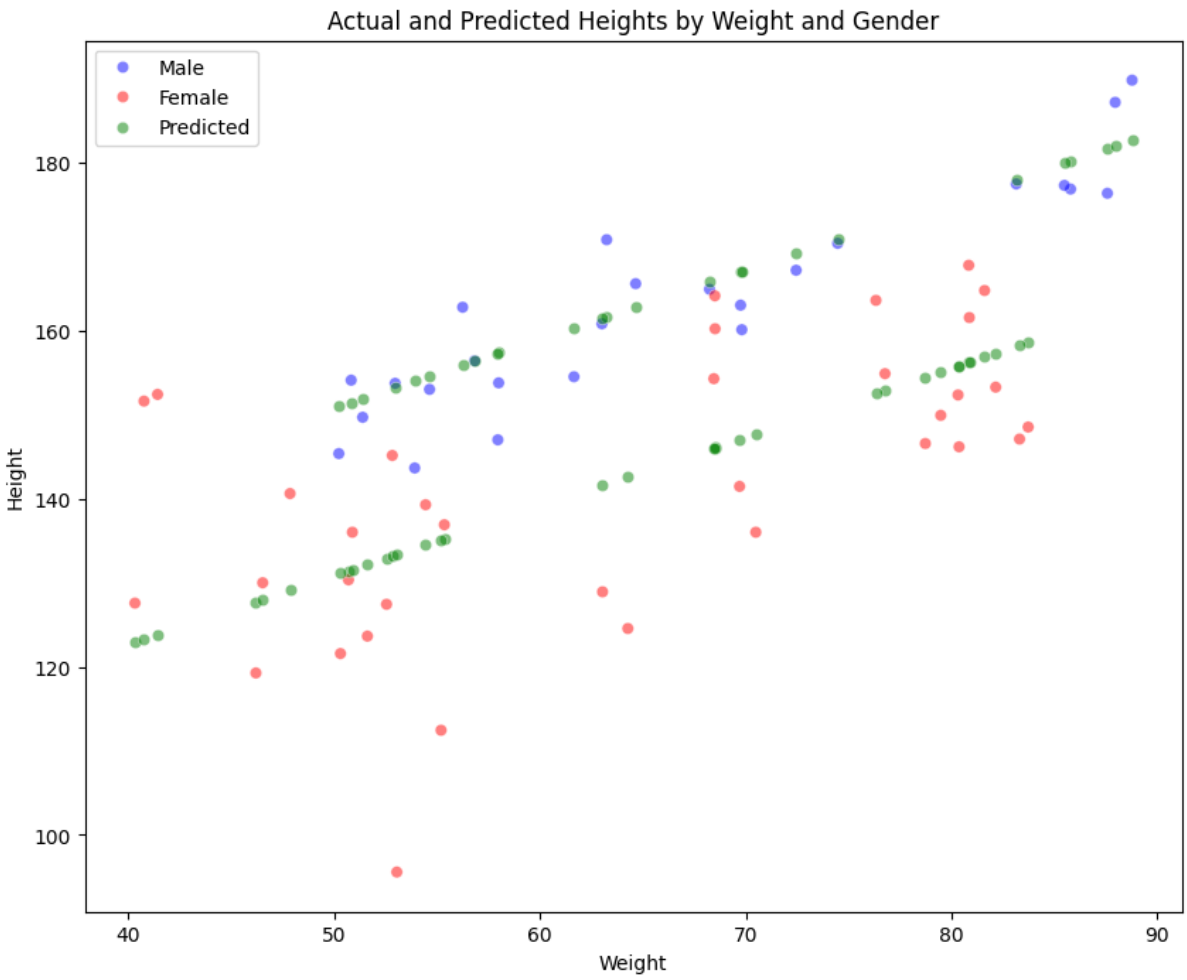
```
# 잔차 회귀식

coef = lr_model2.coef_[0]
intercept = lr_model2.intercept_
print(coef)
print(intercept)
```

```
[-6.37401705  0.82198931]
[-50.74645851]
```

- regression tree의 예측 값과 (2)에서 추정한 회귀선의 예측 값을 더하면 됩니다.
- $\hat{y} = \hat{y}_{tree} + \hat{y}_{reg}$

위 식에서 $\hat{y}_{reg}$는(Gender) × (-6.374)+(Weight) × (0.822)이므로 성별이 남성일 때 기존의 결정트리의 예측값보다 -6.347 만큼 작아진다고 해석할 수 있다.



- 위 그래프를 통해서는 예측값들이 Weight과 Height이 양의 상관관계임을 확인할 수 있다.

3. 모델 해석 비교

Linear Regression을 통해서는 쉽게 선형 회귀방정식을 구해서 해석을 할 수 있었고 Regression tree는 의사결정나무그림을 통해 규칙기반으로 해석을 할 수 있었다. Visual Regression Tree는 남/녀로 나누고 각각의 선형회귀방정식을 구해서 남/녀별로 분리해서 해석할 수 있었다. Hybrid tree Regression은 잔차를 대상으로 한 Linear Regression도 구할 수 있어서 남 녀 중에 어느 범주가 흩어짐 정도가 높은지를 파악할 수 있었다.

이 중에서 가장 좋은 모델은 Hybrid tree regression이라고 생각이된다. 왜냐하면 decision tree와 잔차에 대한 Linear regression tree가 합쳐진 알고리즘으로 decision tree를 통해서 도식화를 통해 나온 Height의 예측값에다가 잔차에 대한 회귀식의 예측값을 더해서 rule 기반으로 해석이 가능하다. 또한 어느 성별이 잔차가 높은지도 해석할 수 있어서 가장 좋은 모델이라고 생각이 든다.