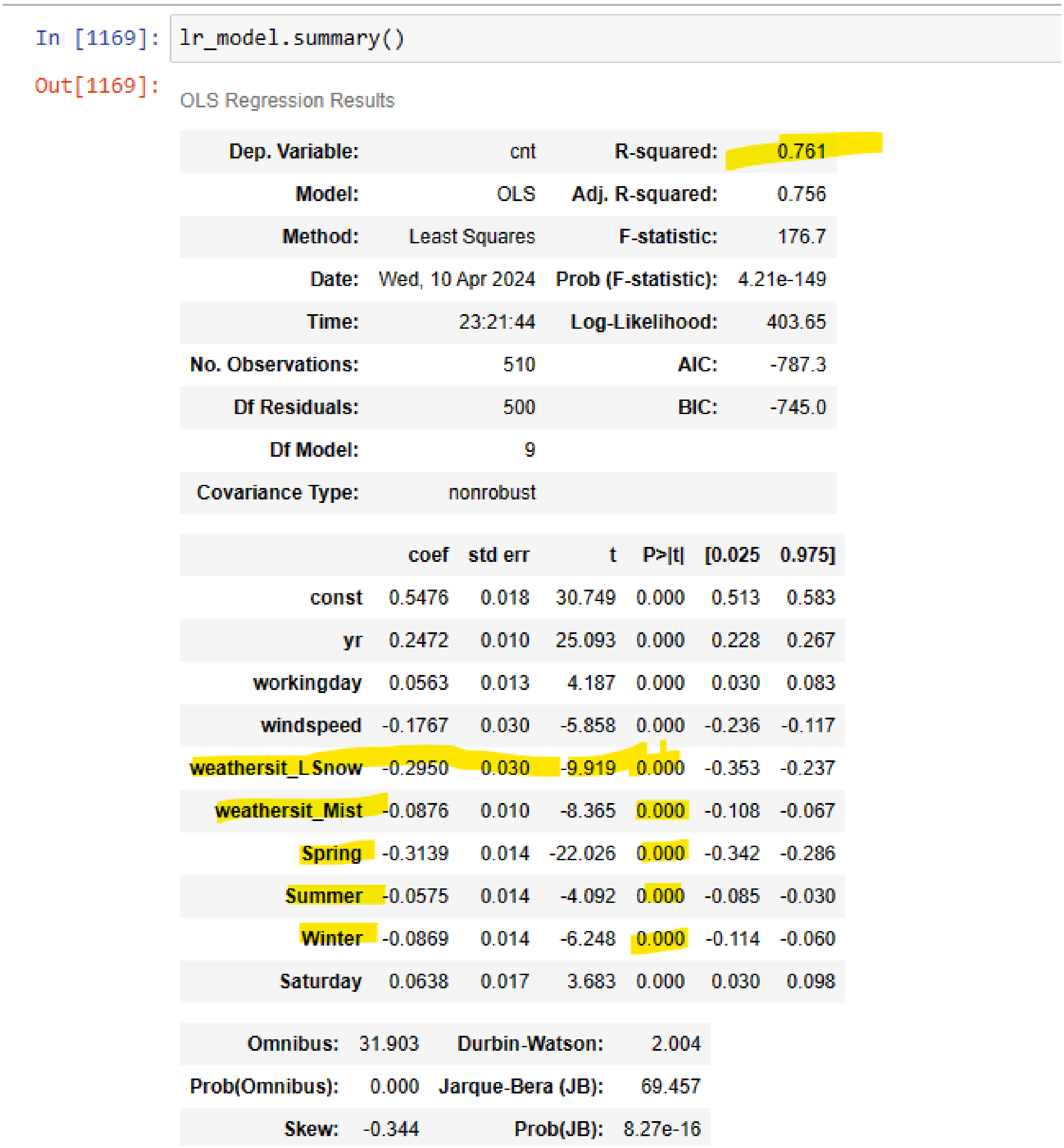
# Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Categorical dummy variables has significant impact on the R square values

Like : Season and Weather, the P value is almost 0 which mean significance for R-Squared or preduction



```
In [1169]: lr_model.summary()
```

Out[1169]:

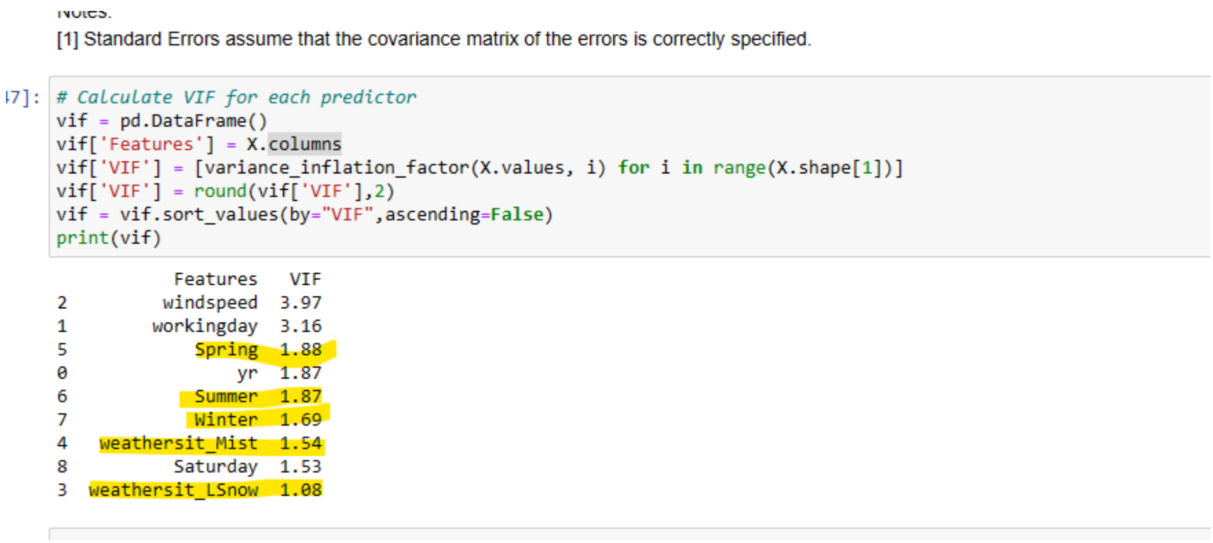
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | cnt | R-squared: | 0.761 |
| Model: | OLS | Adj. R-squared: | 0.756 |
| Method: | Least Squares | F-statistic: | 176.7 |
| Date: | Wed, 10 Apr 2024 | Prob (F-statistic): | 4.21e-149 |
| Time: | 23:21:44 | Log-Likelihood: | 403.65 |
| No. Observations: | 510 | AIC: | -787.3 |
| Df Residuals: | 500 | BIC: | -745.0 |
| Df Model: | 9 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.5476 | 0.018 | 30.749 | 0.000 | 0.513 | 0.583 |
| yr | 0.2472 | 0.010 | 25.093 | 0.000 | 0.228 | 0.267 |
| workingday | 0.0563 | 0.013 | 4.187 | 0.000 | 0.030 | 0.083 |
| windspeed | -0.1767 | 0.030 | -5.858 | 0.000 | -0.236 | -0.117 |
| weathersit_LSnow | -0.2950 | 0.030 | -9.919 | 0.000 | -0.353 | -0.237 |
| weathersit_Mist | -0.0876 | 0.010 | -8.365 | 0.000 | -0.108 | -0.067 |
| Spring | -0.3139 | 0.014 | -22.026 | 0.000 | -0.342 | -0.286 |
| Summer | -0.0575 | 0.014 | -4.092 | 0.000 | -0.085 | -0.030 |
| Winter | -0.0869 | 0.014 | -6.248 | 0.000 | -0.114 | -0.060 |
| Saturday | 0.0638 | 0.017 | 3.683 | 0.000 | 0.030 | 0.098 |

| | | | |
|---|---|---|---|
| Omnibus: | 31.903 | Durbin-Watson: | 2.004 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 69.457 |
| Skew: | -0.344 | Prob(JB): | 8.27e-16 |

VIF value of categorical variables is < 5 means high significance in prediction



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

```
47]: # Calculate VIF for each predictor
     vif = pd.DataFrame()
     vif['Features'] = X.columns
     vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
     vif['VIF'] = round(vif['VIF'],2)
     vif = vif.sort_values(by="VIF",ascending=False)
     print(vif)
```

```
            Features   VIF
2          windspeed  3.97
1         workingday  3.16
5             Spring  1.88
0                 yr  1.87
6             Summer  1.87
7             Winter  1.69
4    weathersit_Mist  1.54
8           Saturday  1.53
3   weathersit_LSnow  1.08
```

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

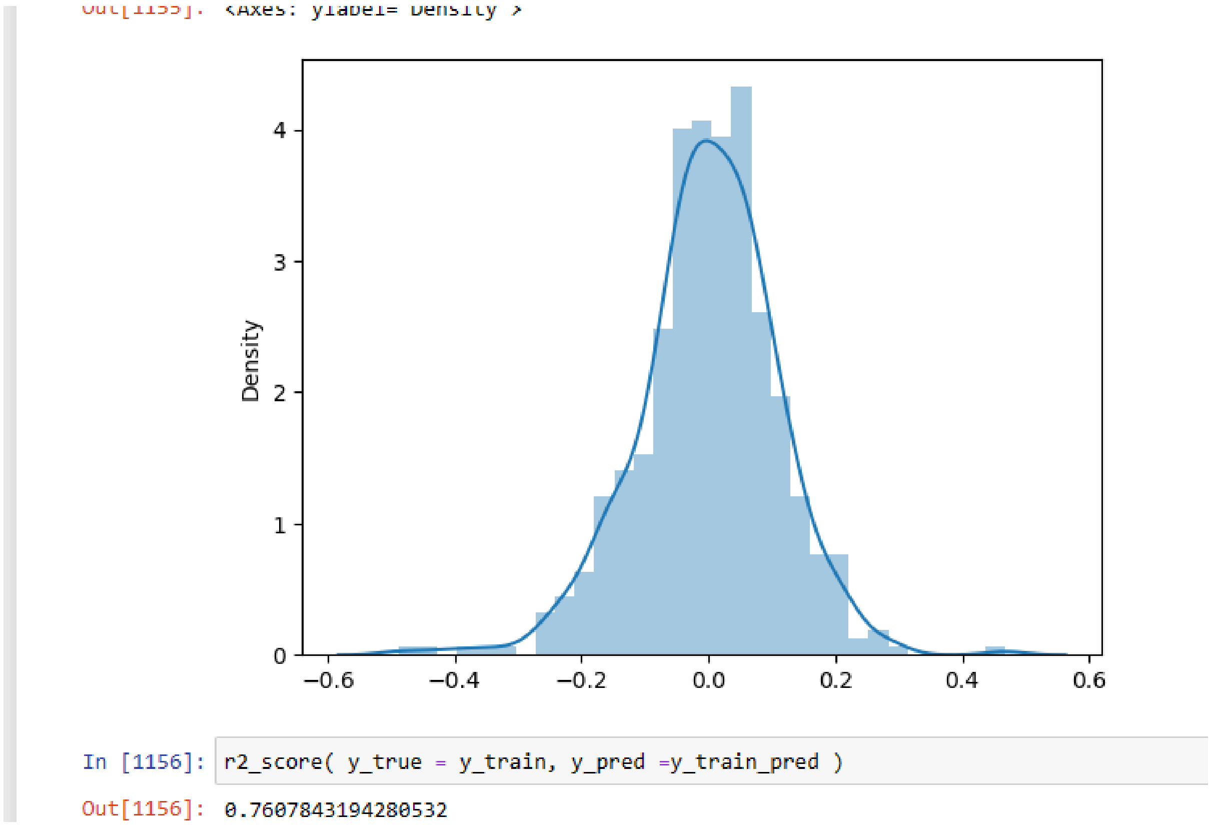
   Drop_first is required in dummy variable as lets say out of 3 dummy variable 2 variables are enough to predict the third variable values, so it is redundant variable and reduces the correlations among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

4.

5. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   ➔ Using scater plot draw a liner line to validate the assumption
   ➔ Checking using distplot of residuals

   res = y_train - y_train_pred
   sns.distplot(res)

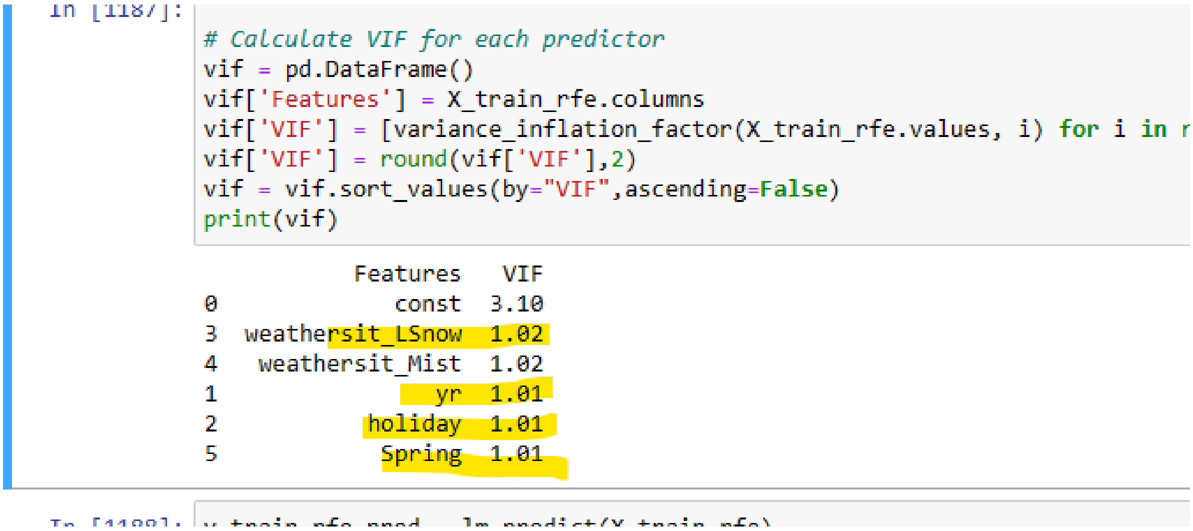


```
Out[1155]: <Axes: ylabel='Density'>
```

```
In [1156]: r2_score( y_true = y_train, y_pred =y_train_pred )

Out[1156]: 0.7607843194280532
```

6. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- ➔ Holiday,
- ➔ Ligh snow, (- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
- ➔ Spring



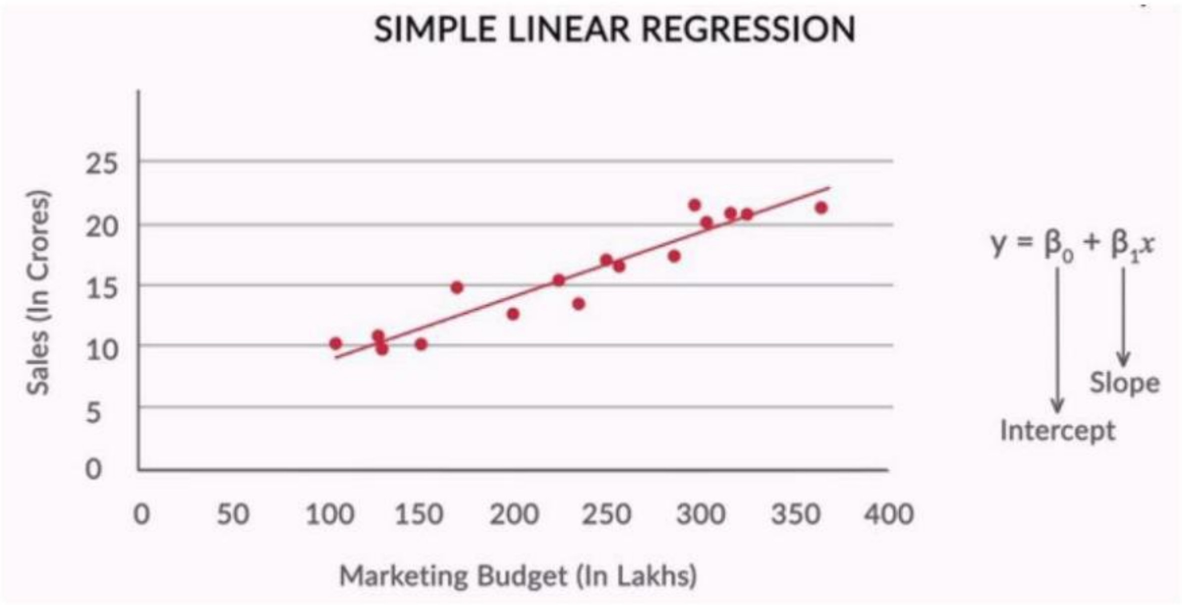
```
In [1187]:
# Calculate VIF for each predictor
vif = pd.DataFrame()
vif['Features'] = X_train_rfe.columns
vif['VIF'] = [variance_inflation_factor(X_train_rfe.values, i) for i in r
vif['VIF'] = round(vif['VIF'],2)
vif = vif.sort_values(by="VIF",ascending=False)
print(vif)

            Features   VIF
0              const  3.10
3  weathersit_LSnow  1.02
4    weathersit_Mist  1.02
1                 yr  1.01
2            holiday  1.01
5             Spring  1.01

In [1188]: y_train_rfe_pred = lm.predict(X_train_rfe)
```

# General Subjective Questions

## 1. Explain the linear regression algorithm in detail. (4 marks)

→ It's a machine learning algorithm where in Manchine learns from the data,
Moslty Supervised learning, meaning use data and regression model like continuous variables as input and output based on cofficients like linear equation.
Classification : Categorization of input data as spam emails etc.

Simple linear regression : Output based on one dependent variable with coefficient and one independent variable
Y=B0+B1X



Multiple Liner regression : Output based on multiple dependent variable with coefficient and one independent variable

Y = B0 + B1x1 + B2X2 …….BnXn

## 2. Explain the Anscombe's quartet in detail. (3 marks)

→ Basically state four data sets
→ like Linear relationship , straight linear line
→ Not linear relationship even though straight line
→ One data point or variable effect the correlation with other variable but vice versa is not true
→ variable is outlier and very less impact on correlation

## 3. What is Pearson's R? (3 marks)

Correlation co-effcient and how 2 variables are corelated

It ranges from -1 to 1

+ve correlation means both the variables are +vely correlated and if one increases it causes other to increase

-ve correlation means both the variables are -vely correlated , means if one increases other decreases

## 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

When we have many independent variables and were on different scales like one is from 10 to 100, other is 1000 to 10000, then ise scaling, so that easy to interpret

Stadardizing scaling is using mean is zero and standard deviation is one

X=x-mean(x)/sd(x)

Min max scaling (normalized scaling) is that it lies between 1 and 0

X = x-min(x)/max(x)-min(x)

## 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

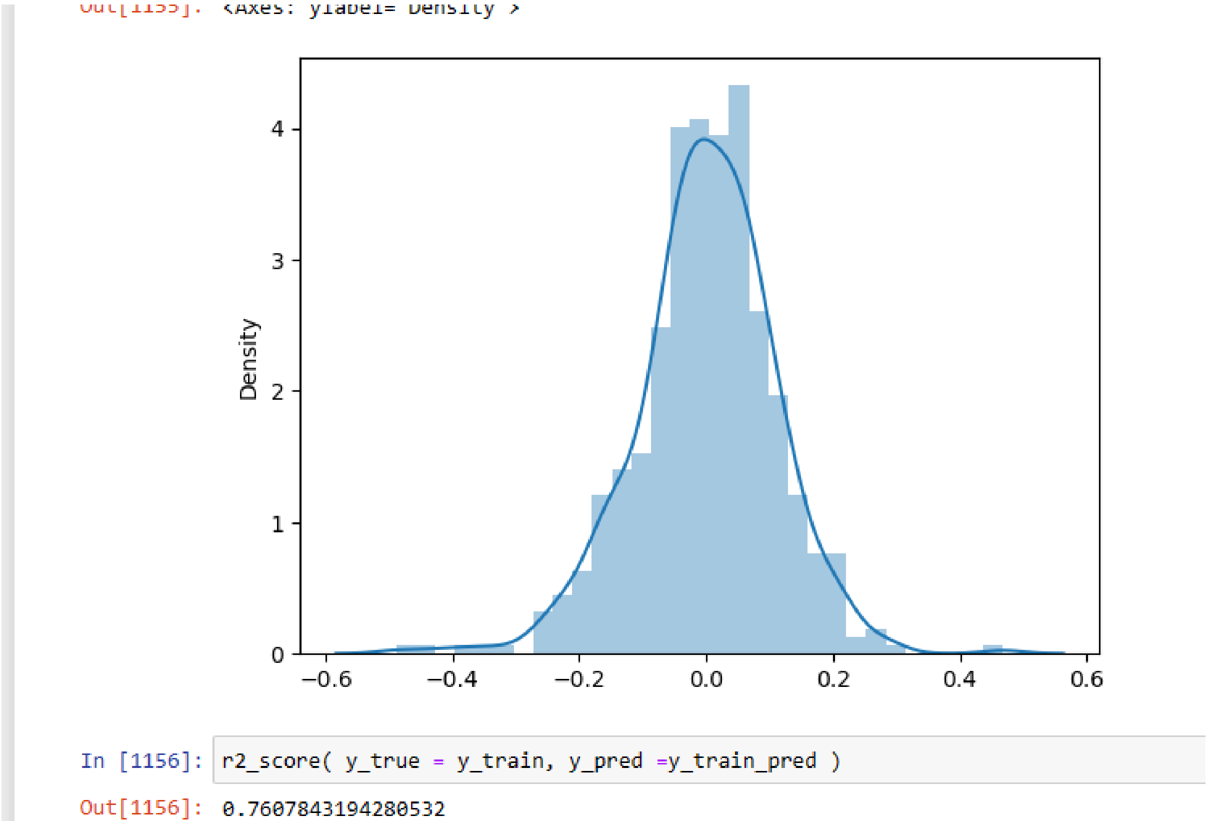
VIF = 1/1-R2

Inase R2=1 VIF becomes infinite , R2=1 means perfect correlation so to fix it we have drop one of the variable

## 7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

This is to check residual is near 0 or not

If its devition from the 0 that means its not normally distributed so need to calculate again



```
Out[1155]: <Axes: ylabel='Density'>
```

```
In [1156]: r2_score( y_true = y_train, y_pred =y_train_pred )
Out[1156]: 0.7607843194280532
```