**Name - Neelam Mahapatro**

**Batch - C46**

**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?

**Ans** – The boxplot results for different categorical variables obtained during the analysis are as below.

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

1. Season , year, month and weathersit categorical variables have significant impacts on the bike hire counts.

Example – Year 2019 has much bike hire counts then 2018 year. Spring season has much less bike hire counts than any other season

1. However, variables like weekday doesnot have any impact on the output variables.
2. Why is it important to use drop\_first=True during dummy variable creation?

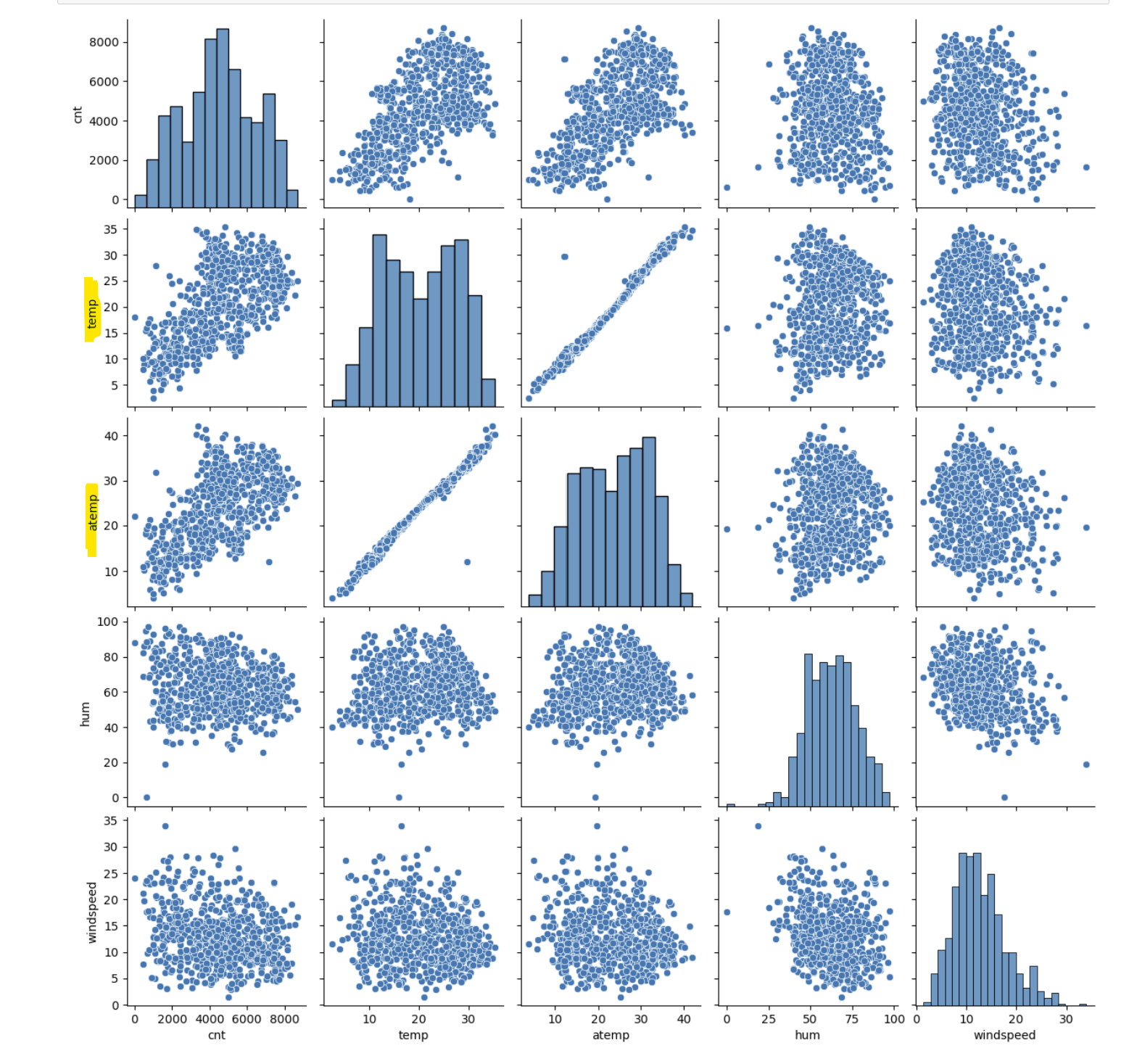
**Ans -** If a category variable has N levels then during dummy variable creation it creates N number of variables. But with respect to multi-collinearity, one variable will be redundant i.e when other variables are zero, this will be one, When any one of other variable it 1, this is 0.

**Hence,** it is essential to use drop\_first to True during creation of dummy variables and p-value will also not get affected by multicollinearity.

Example – Year variable – As per the data set whether it is 2018 or 2019 based on the based on 0 and 1 value respectively. If we create dummy variable on this category, it will create 2 variables ( is\_2018 and is\_2019). Here, when is\_2018 becomes 1, is\_2019 automatically becomes 0. Hence, we can get rid of is\_2018, only one variable (is\_2019) is sufficient and required.

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

**Ans** – By pair-plot among numerical variables, temp and atemp has the highest correlation with the bike rent counts.



1. How did you validate the assumptions of Linear Regression after building the model on the training set?

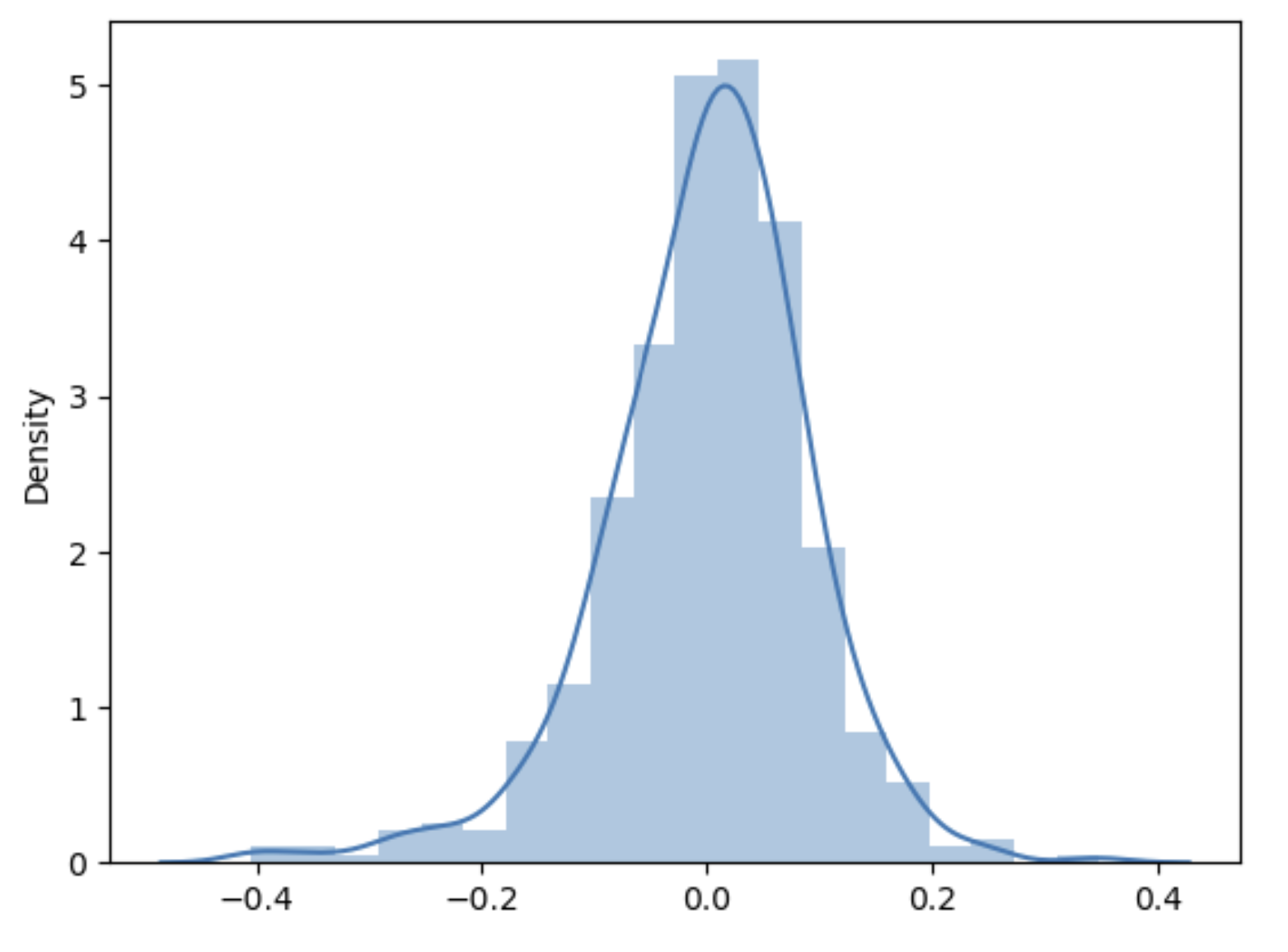
**Ans** –

* 1. By plotting Error Terms –

**-** Error Terms has the normal distribution

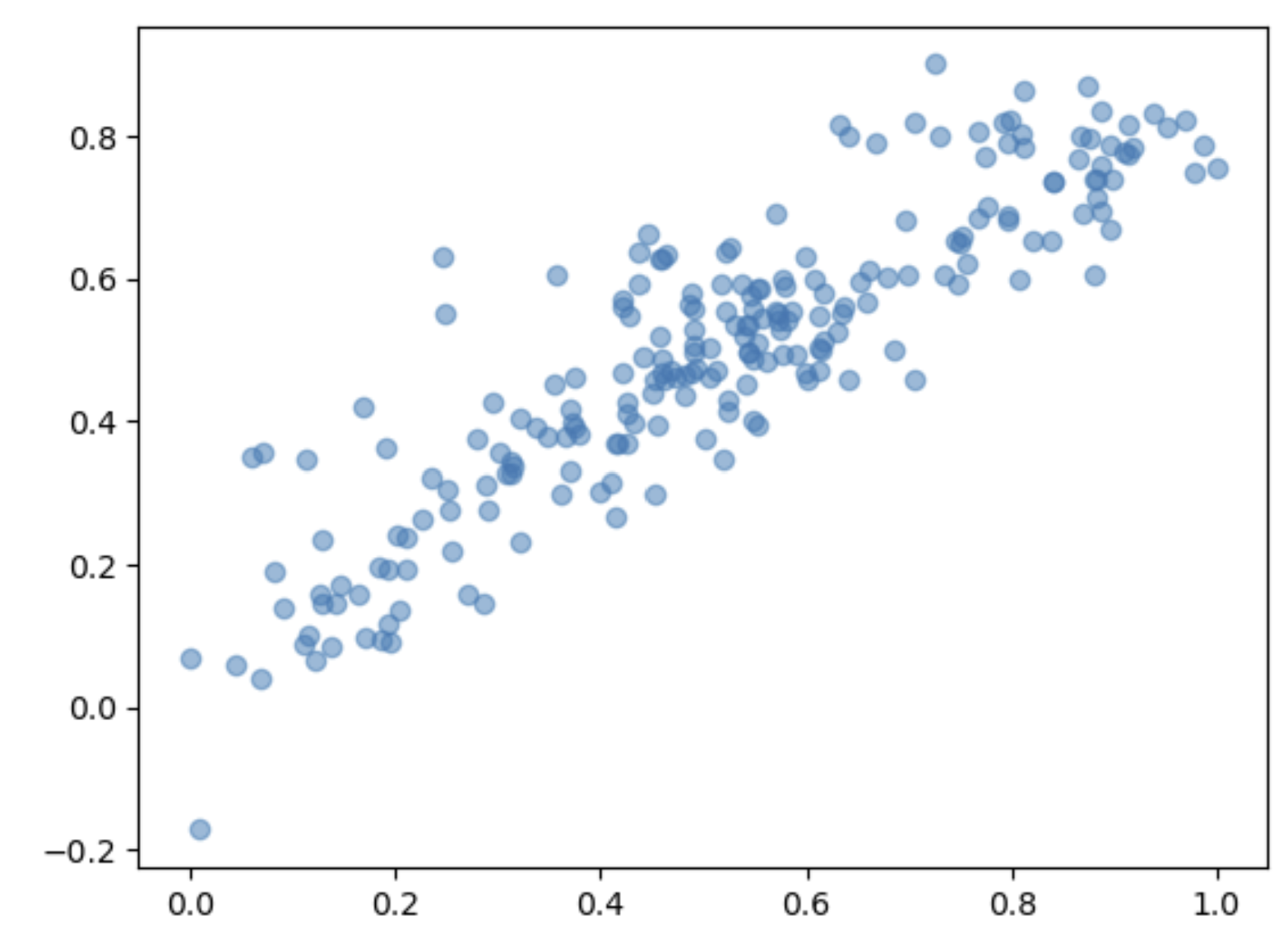
- Error Terms has the mean of 0

- It has a constant variance



* 1. Plot between y\_test\_pred and y\_test

- This has a linear pattern and y\_test\_pred nearly equal to y \_test which is expected



* 1. R2 and Adjusted R2 score for Train and Test data set

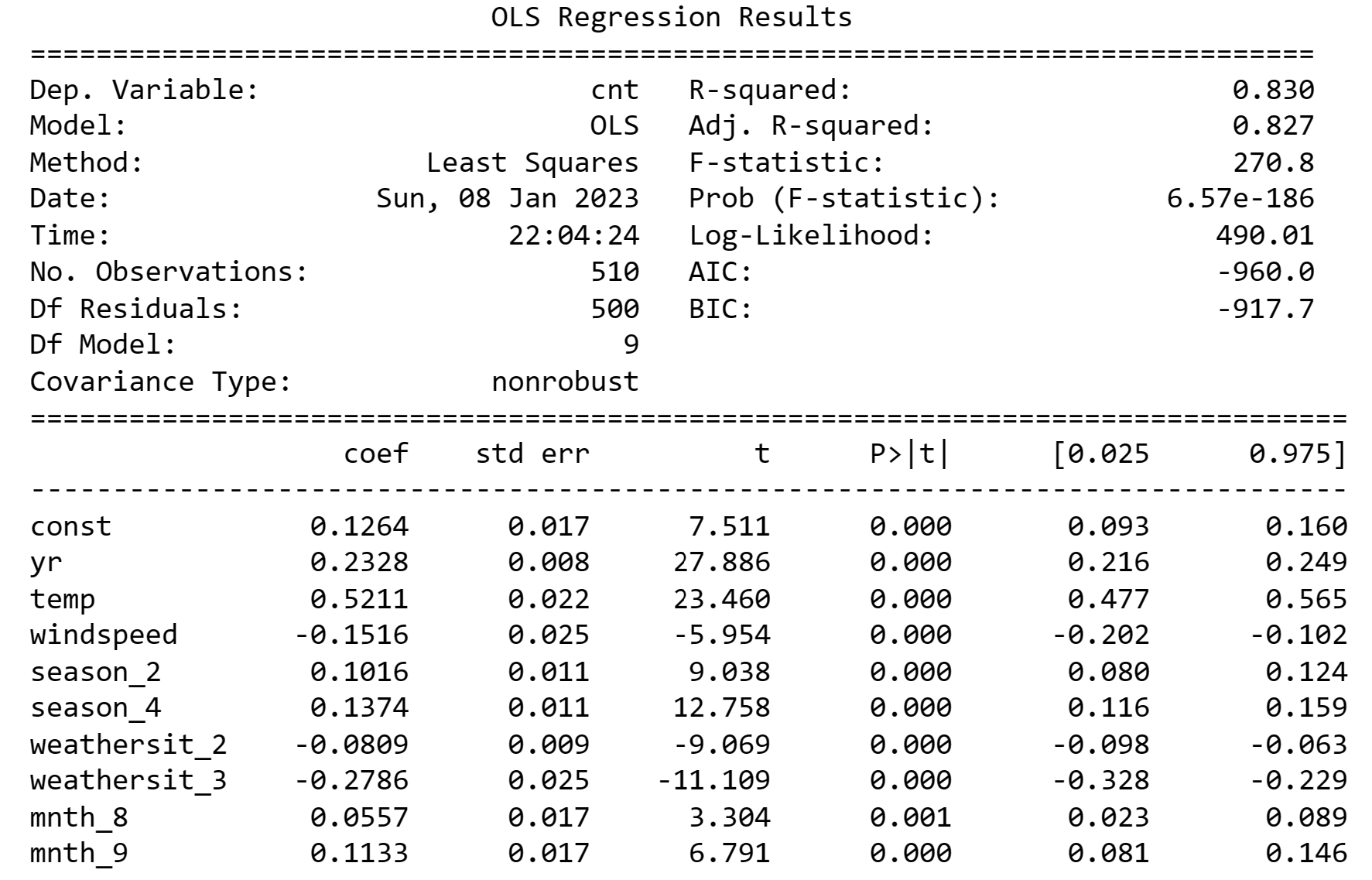
- R2 score is higher and hence the model is good

- R2 score of Train set is higher then that of Test set

Text

Description automatically generated

Summary of the Linear Regression –



1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

**Ans** – Based on the Final model, the top 3 features contributing significantly for demand of shared bikes are –

1. Year : 0.2328
2. Temperate : 0.5211
3. Weathersit\_3 : -0.2786

The summary of the model is also attached above.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

2. Explain the Anscombe’s quartet in detail.

3. What is Pearson’s R?

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.