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?

A: The categorical variables from the dataset are ‘Season’, ‘WeatherSit’, ‘Holiday’ and ‘Year’ as these can be clearly spitted into categories

The following can be inferred from the output of OLS regression test about the above categorical variables

Season (We have dropped the variable spring)

1. Rentals of summer is 11.65 percent higher compared to that of spring
2. Rentals of fall is 7.46 percent higher compared to that of spring
3. Rentals of winter is 16.30 percent higher compared to that of spring

This means Rentals are comparatively higher during Winter

Weather (We have dropped the variable clear)

1. Rentals during misty weather is 5.37 lower compared to clear weather
2. Rentals during light rain is 24.09 percent lower compared to clear weather

This means Rentals are comparatively lower during rainy or misty weather. People prefer cycling during clear weather conditions

Holidays

There is an 8.65 percent drop in bike rentals during holidays compared to working days

Year

There is an increase of 23.09 percent in bike rentals in 2019 compared to that of 2018 which indicates an increase in demand over the year

2. Why is it important to use drop\_first=True during dummy variable creation?

A: This helps in preventing multicollinearity and hence helps in creating independent variables

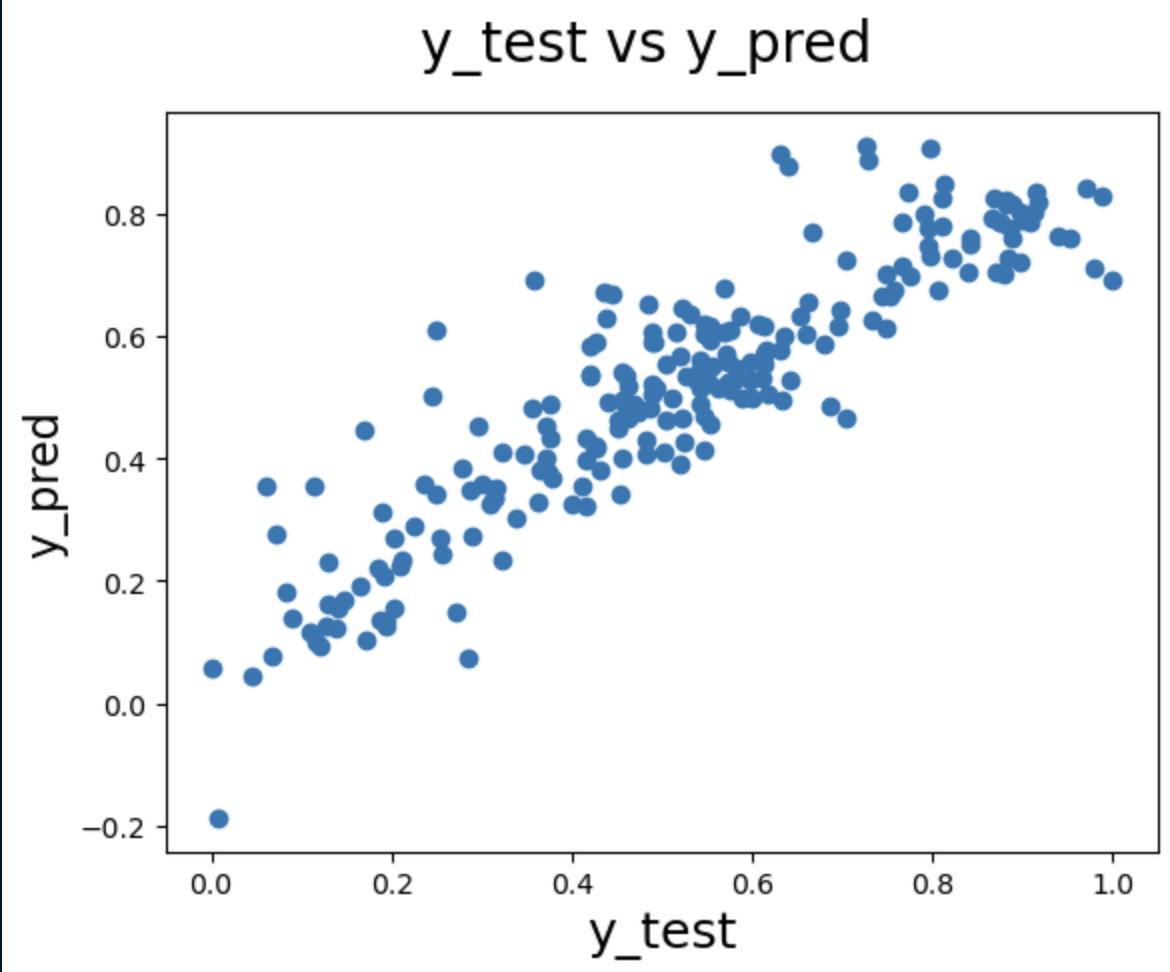
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

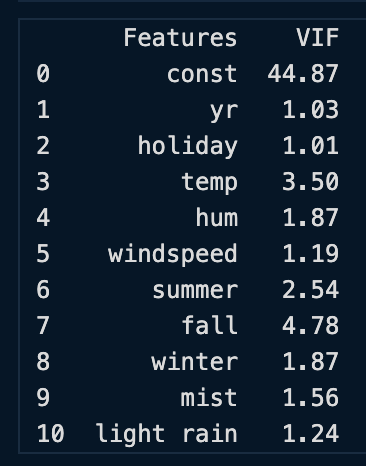
A: The variables with the highest correlation with the target variable is temp and atemp with temp at 64% and atemp at 65% which is obtained from heatmap and pairplot

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

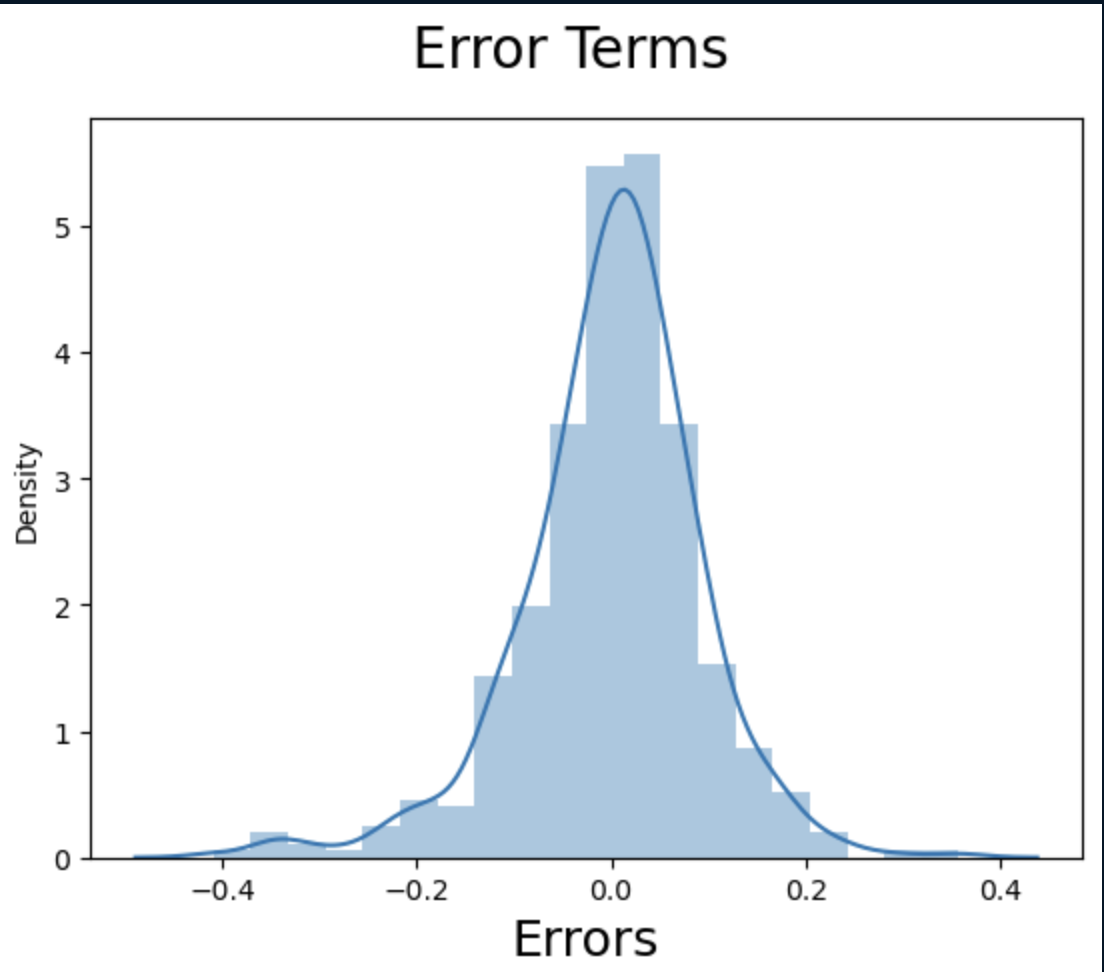
A: The R2 value of the output is 0.8022 which indicates a strong fitment of variables.

The plot between y\_test and y\_pred is also shows a strong linear relationship between the predicted and actual values



The VIF values of all the variables except the target variable are under 5 as given below  


It is also seen from the notebook that the errors are normally distributed



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

A: Temperature with a coefficient of .4961

Year with a coefficient of .2307 and

Light Rain with a coefficient of -.2409

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.