LINEAR REGRESSION

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) Bike rentals have high correlation with “temp”, ”Season: Summer & Winter”, “ Month: Sep”, “Weather: Lightrain_snow”, “holiday”, “windspeed”.

Year 2019 has increased bike rentals than year 2018.

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

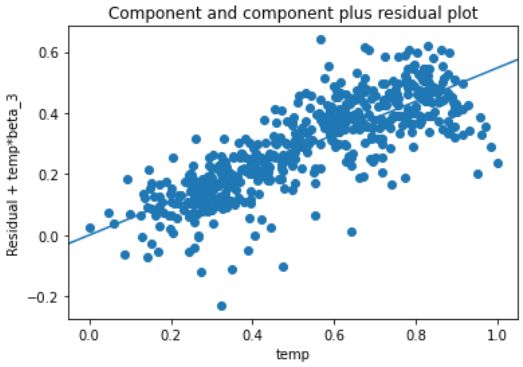
A) It is used to remove the unnecessary column.

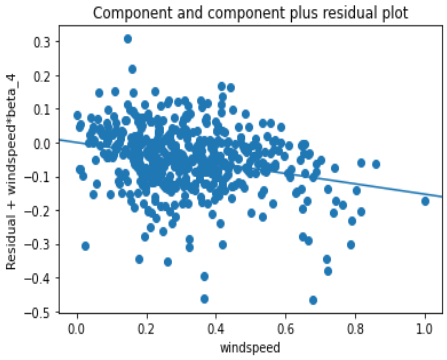
3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A) Temperature

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

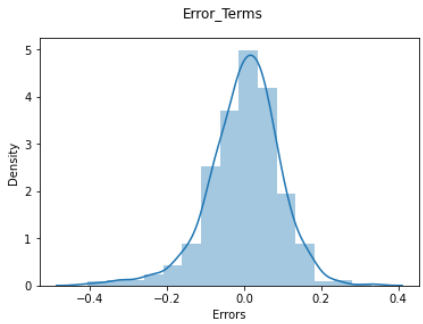
A) i)The relationship between model and predictor variables is linear.



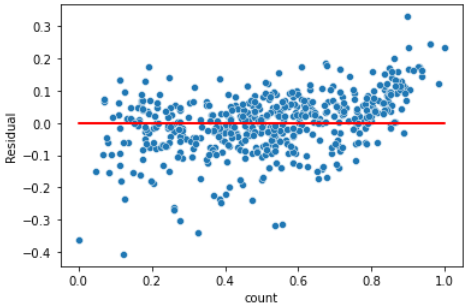


ii) The Durbin_Watson value for the final model LRm7 is 2.097, which indicates that there is almost no auto-correlation.

iii) The histogram that shows that the error terms have a normal distribution.



iv) The homoscedasticity is well preserved as there is no visible pattern of residual values.



v) All the VIF values of the model are below 5 which indicates that there is very less multicollinearity among the predictor variables.

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

A) The top 3 predictor variables according to the Final_Model that influences bike booking are:

1) Temperature(temp)

A coefficient value of '0.548008' indicated that a temperature has significant impact on bike rentals.

2) Year(yr)

A coefficient value of '0.232861' indicates that a year-wise the rentals are increasing.

3) Lightrain_snow

A coefficient value of '-0.282869' indicates that lightsnow and rain refrains people from bike rentals.

These three variables should be prioritized to maximize bike rentals.

The bike availability and promotions should be increased during summer to increase bike rentals.

General Subjective Questions

1) Explain the linear regression algorithm in detail

A) Linear Regression is a machine learning algorithm which comes under supervised learning. It is useful in finding the best linear relationship between independent and dependent variables. i.e finding the best fitting straight line through the data.
It is done using the Sum of squared residuals method.
Uses of Linear relationship:
Prediction of targets and price
Risk management

2) Explain the Anscombe's quartet in detail.

A) As the name suggests, Anscombe's quartet has four datasets which are nearly similar to each other's statistical description (like variance, mean)but have different distributions and plots.

3) What is Pearson's R?

A) Pearson's coefficient which is known as correlation coefficient used mostly in linear regression analysis.

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

A) Scaling is the method used to normalize the range of independent variables between 0 and 1 in general.
Normalization is MinMaxScaling method consists of rescaling the variables in the range of 0-1
Standardisation makes the values with zero mean and unit variance.

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

A) Whenever there is perfect correlation between the variables the VIF value is infinite. The variable can be explained exactly by the linear combination of other variable.

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

A) Q-Q plot is a quantile-quantile plot which is used to assess if a set of data came from theoretical distribution of variables such as normal, or uniform distribution.

It helps in knowing whether the two datasets came from the same population.

Its use to show how close two distributions are and is useful in determining the normality of a linear distribution.