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?

#From the box plot we can see the pattern that there is a high demand in Fall and Summer Season

#From the box plot we can see that 2019 has high demand

#From the box plot we can see that Aug,Sep,Oct has a high demand

#From the box Plot we can see that Sat/Wed/THurs day has a demand compare to other days

#From the box plot we can see that weather 'Clear' has a demand.

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

-it reduces the correlations created among dummy variables.

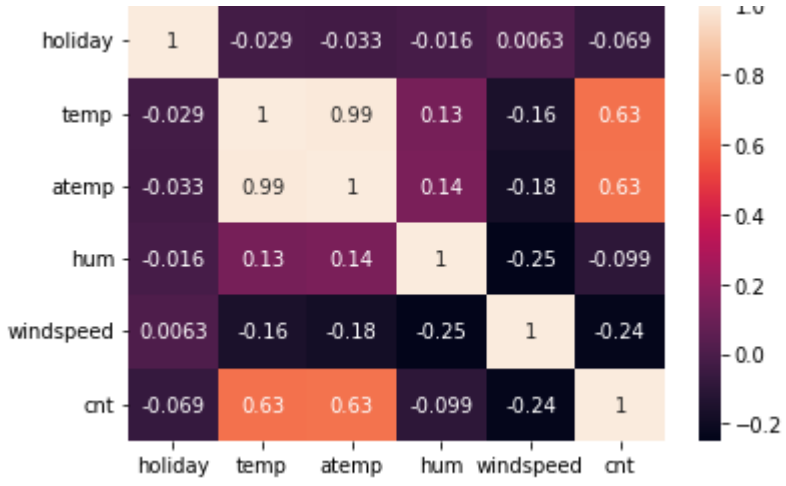
-reducing the extra column created during dummy variable creation.

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

Cnt-0.63

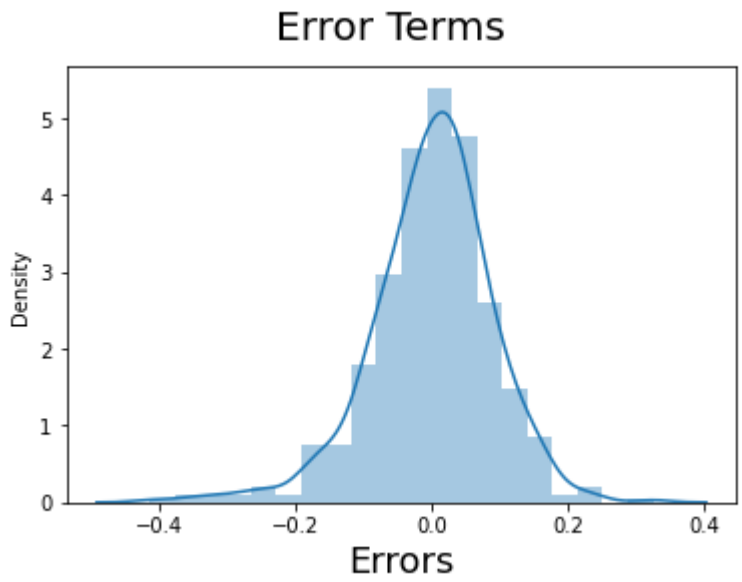
Atemp-0.99

We can exclude atemp as it's not a deciding factor. The same pattern it identified on Pair plot



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

Created the distplot to identify the residual.



it is proof that Error Distribution Is Normally Distributed Across 0.

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

- temp - 0.5471(Positive Correlation)

    a unit increase in temp variable increases the bike hire numbers by 0.5471

- yr_2019- 0.2327 (Positive Correlation)

    a unit increase in yr_2019 variable increases the bike hire numbers by 0.2327units

- season_winter -0.1323(Positive Correlation)

    a unit increase in season_winter variable increases the bike hire numbers by 0.1323 units

- weathersit_light (-0.2892) `(Negative Correlation)`

    a unit increase in weathersit_light variable decrease the bike hire numbers by 0.2892 units

General Subjective Questions

1.Explain the linear regression algorithm in details?

-It's a linear relationship between input variable and output variable.

-In a simple words we can all it as dependent variable is continuous in nature.

-It's completely based on the Supervised learning.

-Mathematically the linear expression would be denoted by following formula, y=mx+c . Here m is the slope and c is y-intercept line.

-Below are the types,

    01. Simple Linear Regression – One predictor Variable.

02.Multi Linear Regression – Two or More predictor variables.

As part of the linear regression we calculate the Residual sum Square with that we calculate the howmuch the target value varies from the linear regression line.

Also, We calculate Total sum of square and it gives the details that howmuch the data point move around the mean.

R-Square is an another method wherein we use to calculate that how close the data are to the regression line.

2.Explain the Anscombe's quartet in detail.

- a group of four data sets which are nearly identical in simple descriptive statistics

- provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

-Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

- It's a bivariate correlation.

- It is a statistic that measures the linear correlation between two variables.

- It has a numerical value that lies between -1.0 and +1.0.

- It cannot capture nonlinear relationships between two variables

- It cannot differentiate between dependent and independent variables.

- It's the covariance of the two variables divided by the product of their standard deviations.

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

- The Pre-Processing step which is applied to independent variables to normalize the data within a particular range.

- Let's consider that the data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account

- The scaling affects the coefficients

<u>Normalisation:</u>

Minimum and maximum value of features are used for scaling

It brings all of the data in the range of 0 and 1

<u>Standardisation:</u>

Mean and standard deviation is used for scaling

It is not bounded to a certain range.

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

-It's perfect correlation

- We need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

- The corresponding variable may be expressed exactly by a linear combination of other variables.

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

-plots of two quantiles against each other

-It is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

-Q–Q plot will approximately lie on a line if it's linear regression.