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?

   Ans:

   1. COUNT OF BIKE RENTALS INCREASED AND BECAME POPULAR IN YEAR 2019 THAN 2018
      (from 'Year' variable)
   2. COUNT OF BIKE RENTALS IS MORE DURING CLEAR WEATHER
      (from 'Weathersit' variable)
   3. FALL AND SUMMER ARE MORE FAVOURABLE FOR BIKE RENTALS THAN SPRING
      (from 'Season' variable)

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

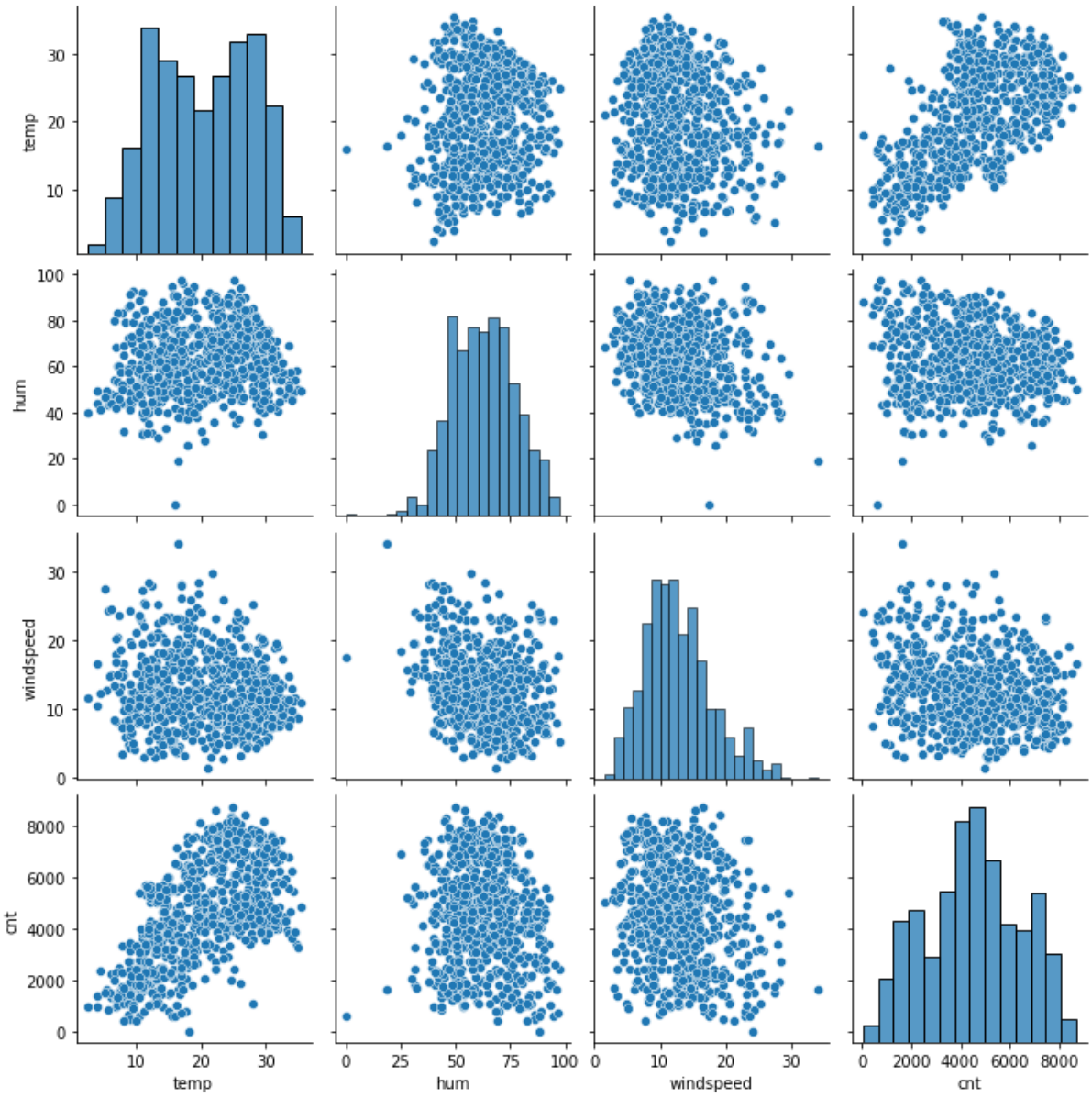
   Ans:

   1. TO AVOID MULTICOLLINEARITY (IF WE DON'T DROP ,DUMMY VARIABLES WILL BE CORRELATED) AND AFFECTS THE MODEL ADVERSELY
   2. TO AVOID REDUNDANT FEATURES

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

Ans:

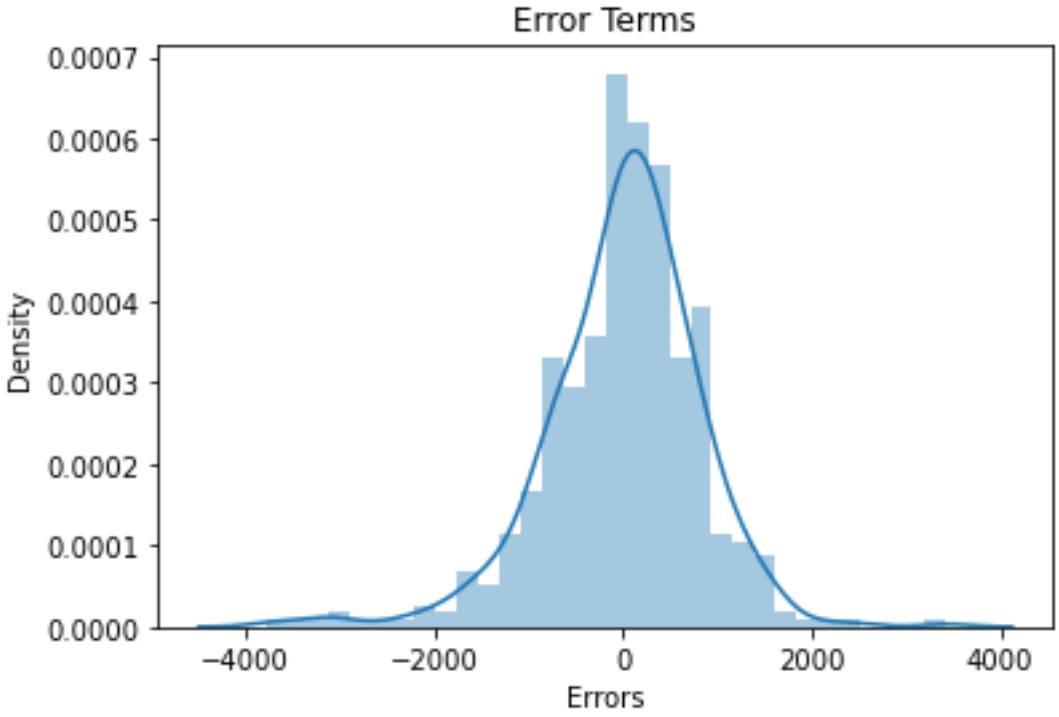
COUNT (TARGET VARIABLE)HAS SIGNIFICANTLY HIGH CORRELATION WITH TEMPERATURE (TEMP)



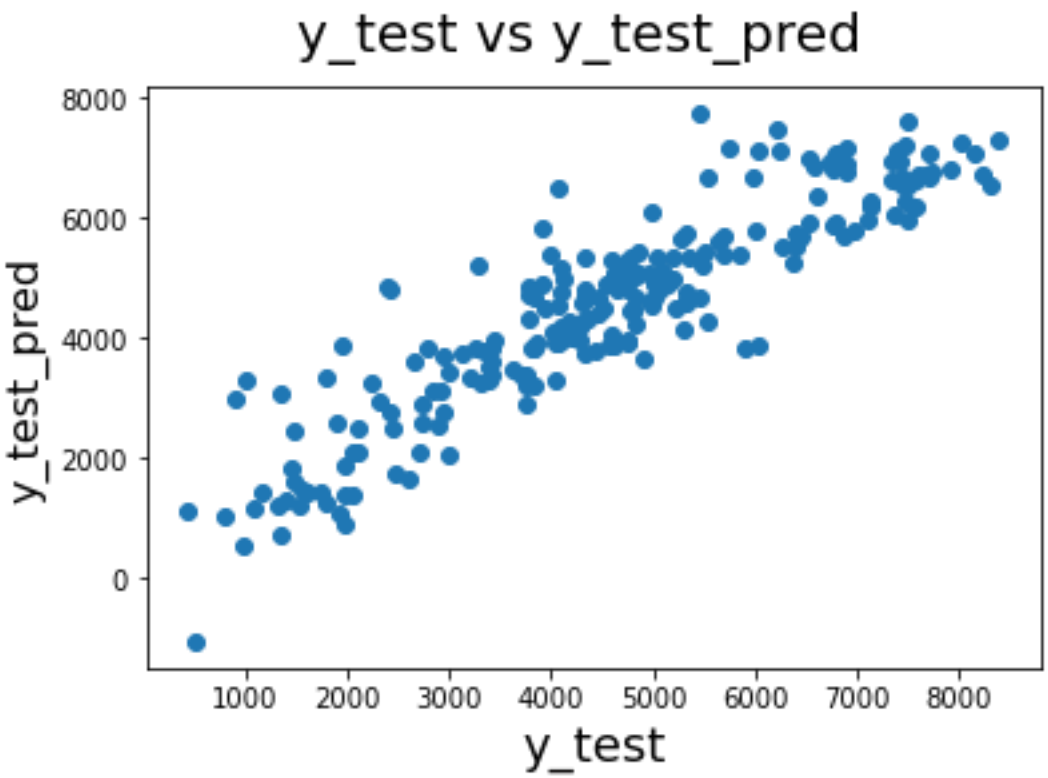
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Ans:

   RESIDUAL ERRORS FOLLOW NORMAL DISTRIBUTION



   MAINTAINS LINEAR RELATION
   BETWEEN DEPENDANT VARIABLE (TEST AND PREDICTED)



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

   Ans:

   As per our final model Top 3 features are

   1. Temperatue: 0.4171 indicates a unit increase in temp variable increases the bike hire numbers by 0.4171
   2. Weather Situation: -0.2617 indicates a unit increase weathersit(weathersit_bad) decreases the bike hike numbers by 0.2617
   3. Yr: 0.2150 indicates a unit increase in yr increases the bike hike numbers by 0.2150

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans:

   Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

   In Regression, we plot a graph between the variables which best fit the given data points. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).To calculate best-fit line linear regression uses a traditional slope-intercept form.

2. Explain the Anscombe's quartet in detail.

   Ans:

   Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

   Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be

plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Simple understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson's R?

Ans:

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

A correlation can be calculated between two numerical values (e.g., age and salary) or between two category values (e.g., type of product and profession). However, a company may also want to calculate correlations between variables of different types. One method to calculate the correlation of a numerical variable with a categorical one is to convert the numerical variable into categories. For example, age would be categorized into ranges (or buckets) such as: 18 to 30, 31 to 40, and so on.

As well as the correlation, the covariance of two variables is often calculated. In contrast with the correlation value, which must be between − 1 and 1, the covariance may assume any numerical value. The covariance indicates the grade of synchronization of the variance (or volatility) of the two variables.

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

   Scaling: *It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.*
   Why Scaling Performed:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that ***scaling just affects the coefficients*** *and none of the other parameters like* ***t-statistic, F-statistic, p-values, R-squared****, etc.*

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1.* ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

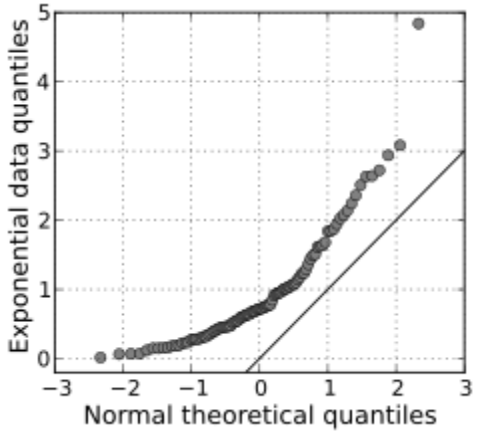
   The ***variance inflation factor*** *(VIF)* quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot 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.