Assignment Based Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer: From analyzing the categorical variables from the dataset, I am able to infer the following:

a. In the fall season, most of the bike were rented in both the years
b. Bikes rented in 2019 were more than 2018 irrespective of the seasons.
c. No bikes were rented in heavy snow weather.
d. The maximum bikes were rented in clear weather.
e. From June to October there was spike in rents.
f. From June to October there were more bikes rented than the other months for both working and non-working days.
g. For working days there were more rents than the non-working days throughout the year.
h. There were no rents in month of March, June and august during holidays.
i. During holidays, the rents were more in July.
j. During non-holidays, the most number of rents were made in September, October.

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

Answer: Dummy variables are numerical representations of the categorical variables. The key idea behind dummy variable is to create new indicator variables for a categorical variable having N levels with N-1 indicator variable. Suppose we have a categorical variable as a weather situation which have 3 values 'clear', 'mist' and 'light snow'. The combination of 0 and 1 represent different weather situation.

Weather situation	Clear	Mist	Light Snow
Clear	1	0	0
Mist	0	1	0
Light snow	0	0	1

If we drop the first column 'Clear', the combination 10 will represent 'Mist' and the combination 01 will represent 'Light snow'. So we don't need to create another column for 'Clear' which can be represented by '00'. So for N levels, we create N-1 indicator variable.

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

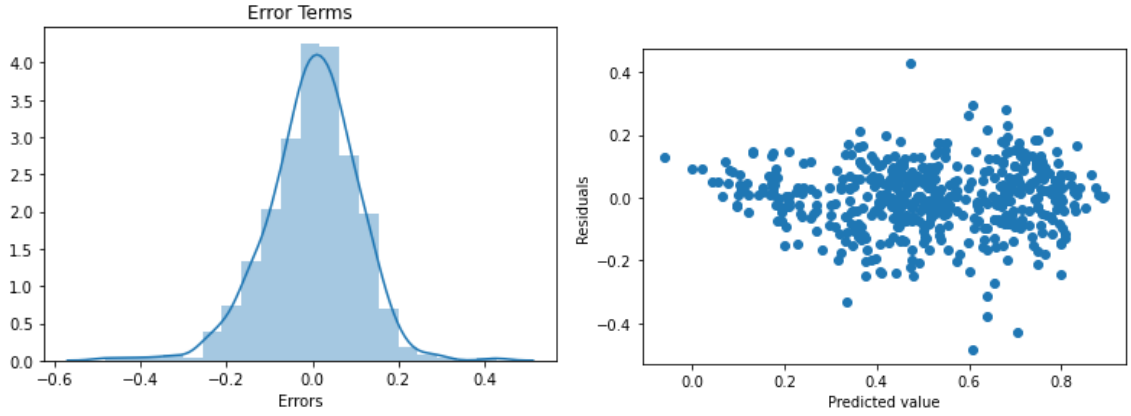
Answer: From the pair plot, we can clearly see that the numerical variable 'temp' ha the highest correlation with the target variable 'cnt'.

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

Answer: The assumptions that we make after building the LR model on the training set are-

a. Error terms are normally distributed with mean as zero. For that we plotted a histogram which represents the residuals.
b. Error terms are independent of each other.

c. Error terms have constant variance.



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

Answer: As per the final model, the top 3 predictors are:

a. yr: A coefficient value of 0.248 indicates that a unit increase in yr variable, increases the 'cnt' number by 0.248 units.
b. Light snow: A coefficient of 0.303 indicates that a unit increases in light snow variable, decreases the 'cnt' by 0.303 units.
c. Spring: A coefficient of 0.258 indicates that a unit increase in spring variable, decreases the 'cnt' by 0.258 units.

General Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a ML algorithm that finds the best fit linear relationship on any given data, between independent and dependent variable. More specifically, that output variable can be calculated from a linear combination of the input variables(x).

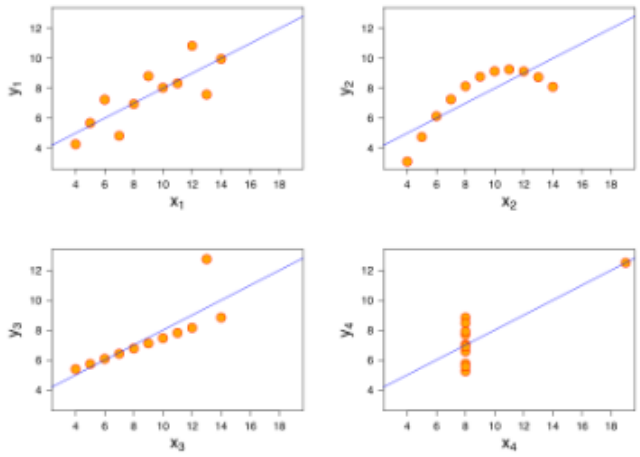
The equation for a linear regression model would be:

$y = B_0 + B_1 X_1 + B_2 X_2 + \ldots + B_n X_n$

The above equation B_0 is the y-intercept and B_i is the coefficient of a feature variable X_i. The significance of the model is that it can easily be interpreted and understand the marginal changes and their consequences. If the value of X_i increases by 1 unit keeping all the other variables constant, then total increase in the value of y is B_i. Mathematically, the intercept term B_0 is the response when all the predictor terms are set to zero or not considered.

2. Explain the Anscombe's quartet in detail.

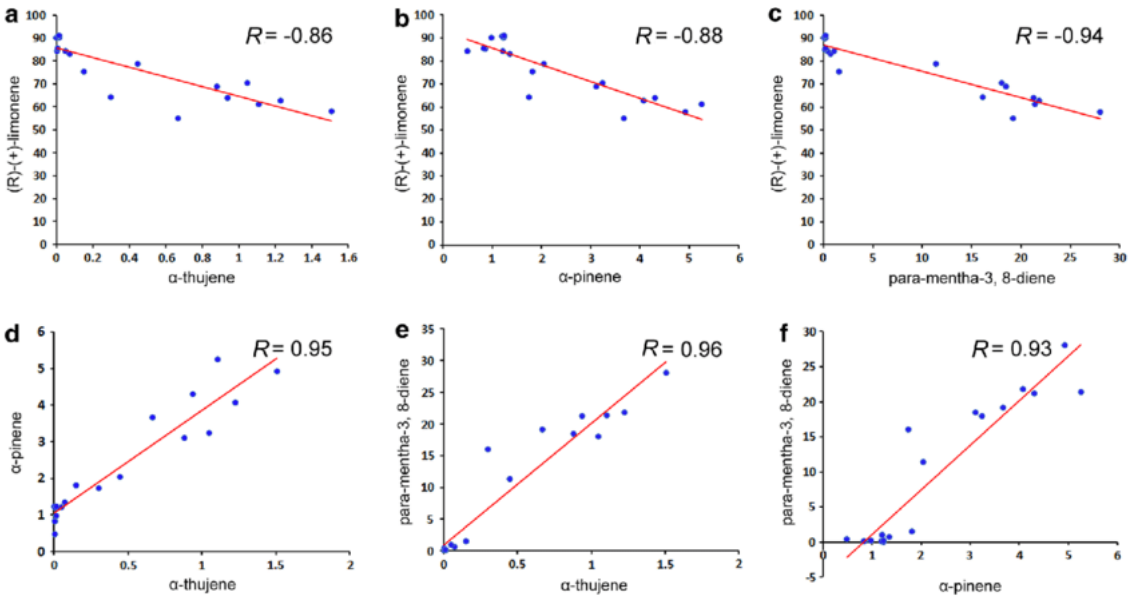
Answer: The shortcomings of linear relations can be explained by the Anscombe's quartet.



As we can see in all the above 4 graphs, the linear regression are exactly same. But there are some peculiarities in the datasets which have fooled the linear regression line. The 1st one seems to be doing a descent job, the 2nd one clearly shows that the linear regression can only model linear relationships and is incapable of handling any other kind of data. The 3rd and 4th one shows that the linear regression model is sensitive to outliers. If there would be no outliers than we would have gotten a great line fitted through the data. Therefore, Anscombe's quartet emphasizes the importance in visualization in data analysis.

3. What is Pearson's R?

Answer: Correlation between datasets is a measure of how well they are related to each other. The most common measure of correlation in stats is the Pearson Correlation. It is the test statistics that measure the statistical relationship, or association between two continuous variables. It is known as the best method of measuring the association between variables of interest as it is based on the method of covariance. It tells about the magnitude of correlation as well as direction of relationship.



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

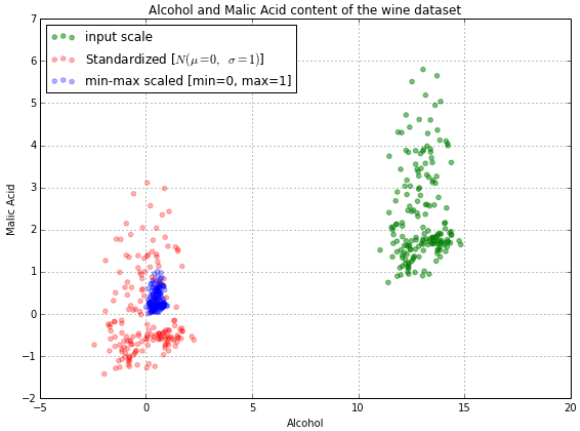
Answer: Scaling is a technique to standardize the values of a feature so that it can be easy to make inferences about the features. Suppose if we have two features on two different scales such as salary and age. Both the features have different range. Age may vary from 20 to 60. But the salary would vary from 20,000 to 20,00,000. So, it would be difficult to infer from the coefficients about impact of each feature on the target variable. Therefore, scaling is required. Scaling just affects the coefficients and none of the other parameters like t-statistic, p-value, R^2 etc.

Normalization scaling – It brings all the data in the range of 0 and 1.

$$x = x\text{-}\min(x) \,/\, \max(x)\text{-}\min(x)$$

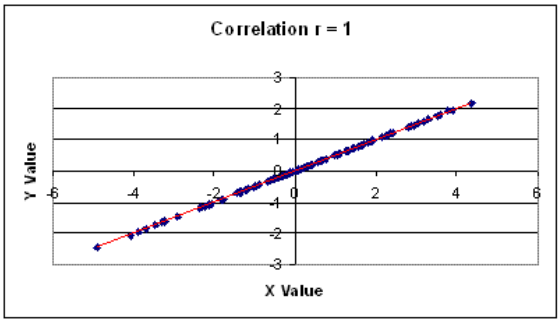
Standardization scaling – It brings all the data into a standard normal distribution which has mean zero and standard deviation of 1.

$$x = x\text{-}\text{mean}(x) \,/\, \text{sd}(x)$$



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

Answer: Since VIF = 1 / (1-R^2), For a VIF to be infinite, R^2 needs to be equal to 1. This happens when the independent variable is perfectly correlated to the other variables. The below graph shows such a case.



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

Answer: Q-Q (Quantile - Quantile) plot is a graphical tool. This helps in a scenario of linear regression when we have training and test data set separately and the we can confirm using Q-Q plot that the both the data sets are from same populations with the same distributions. Q-Q plot is a scatter plot created by plotting two sets of quantitles against one another. If both the sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.