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?

   The variables Yr and light_snow is having high effect on the dependent variable since their coefficient is high. In one of the years, the shared bike demand was high since the coefficient is positive and in the light_snow atmosphere, the shared bike demand is negative since the coefficient is negative.

   The other categorical variables are not having high coefficient implying that the other variables are not having high impact on the demand of shared bikes.

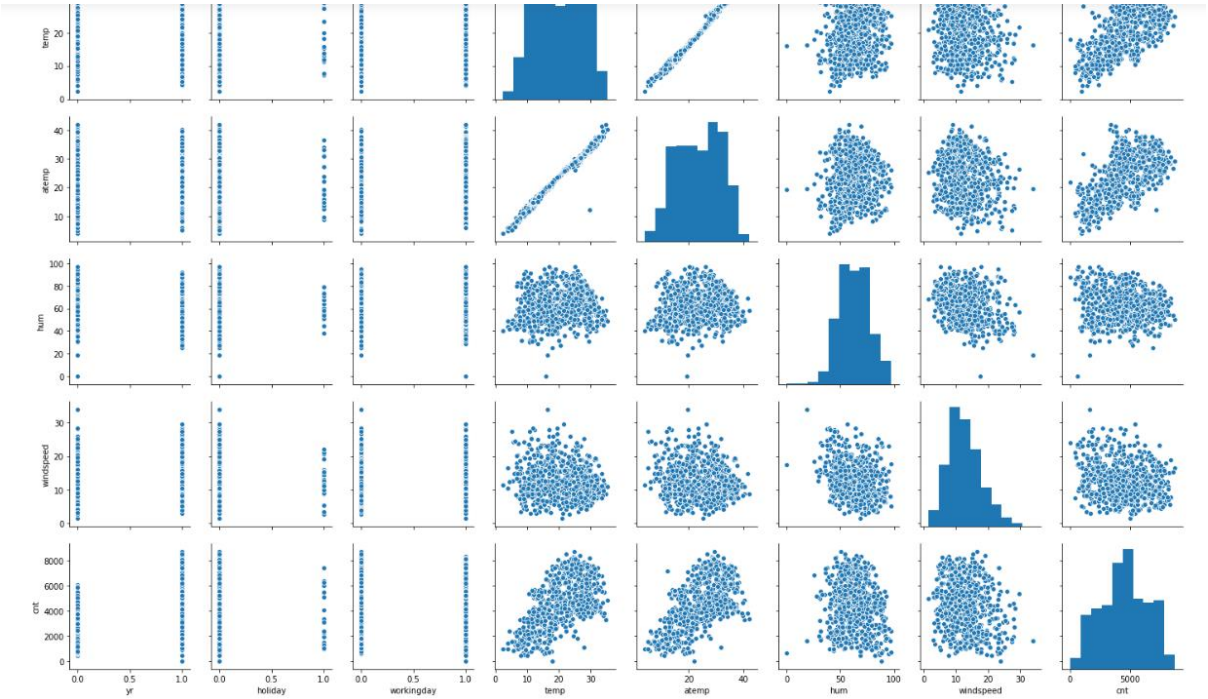
2. Why is it important to use drop_first=True during dummy variable creation?

   Using drop_first=True is more common in statistics and often referred to as "dummy encoding" while using drop_first=False gives you "one hot-encoding" which is more common in ML.

   If drop_first = True is used, It makes sure that out of k distinct values of a categorical variable, there are k-1 dummy variables created.

   If drop_first = True is not used. It would create k dummy variables out of k distinct values of categorical variables.

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

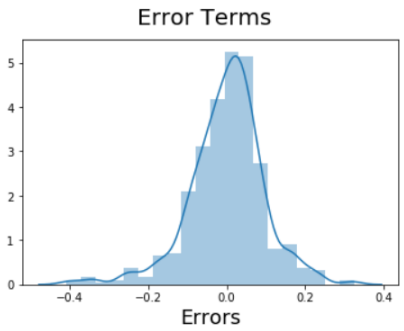


   Temp and atemp are having high correlation with the target variable.

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

   I have validated the model with the three assumptions of multiple linear regression.

   Error terms are normally distributed over the mean 0:



   Overfitting is not present in the model because the r-square of test and the train sets are almost equal. If there is a high r-square for training set and low r-square for test set, then it would have been an issue of over fitting.

   Multicollinearity is also not present because the VIF values of each feature in the model is very less.

   F-statistic is also high, implying that the relationship of independent variables with the target is not a coincidence but there is a significant relationship.

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

   Temp, yr and light_snow weather situation are contributing more to the demand of shared bikes.

   Temp and yr are having positive correlation. Raise in temperature is causing an increase in the demand of shared bikes. In an a particular year, there is a high demand of shared bikes.

   Light_snow kind of weather situation is causing lesser demand of shared bikes.

General subjective questions:

1. Explain Linear Regression Algorithm in detail

   Linear regression is a linear approach of finding the exact relationship between independent variables and dependent variable. If there is only one independent variable present in the data, then it is called simple linear regression. If there are more than one independent variable present in the data, then it is called multiple linear regression.

   Considering a simple linear equation $y = mx + c$

   Where y is the target/dependent variable and x is the independent variable.

M is the slope of the line that is fit in the given data points and determines the rate of change of y when the x changes.
C is the intercept which determines the value of y when x is 0.

When there are multiple independent variables, the linear regression equation is given by

$$yi=\beta 0+\beta 1xi1+\beta 2xi2+...+\beta pxip+\epsilon$$

Linear regression is a process of determining the coefficients of variables and the intercept by minimising the difference in the predicted values of 'y' and actual values of 'y'

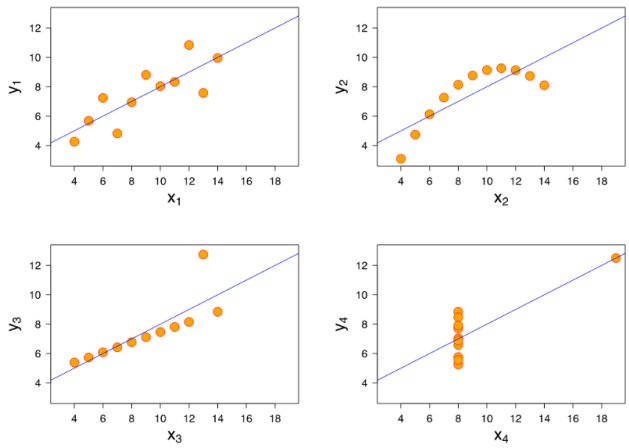
The equation formed with the determined values of x coefficients is called the linear regression function and it helps us predict the values of y.

2. Explain the Anscombe's quartet in detail:

Upon calculating the summary statistics of the data, we will not be able to determine the exact characteristics of the distribution of the data.

The summary statistics such as average, variance and correlation coefficient statistics are misleading if the distribution of the data is not seen.

For example, let us consider the below data distributions,



If we compute the statistics of the data, The average of x, y values, variance of x, y values, correlation coefficients are all same for the given data sets. But when plotted, the distribution of the data is different.

If we assume that the data is following a linear pattern and if we try to fit the linear regression model, it will cause low accuracy in the predictions.

3. What is Person's R?
Pearson's R or the Pearson's correlation coefficient is the numeric value that describes the relationship between two continuous variables.

The first step of calculating Pearson's correlation coefficient is using a scatter plot for the plotting of two variables. If the scatter of the data points is less, the correlation coefficient will be high. Else, it will be low.

Positive correlation coefficient indicates that both the values increase or decrease together and negative correlation coefficient indicates that raise of one value causes the fall of other value and vice versa.

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

   Scaling is a technique used to get the values of a feature into a given range. It is performed to handle highly varying magnitudes or values or units. If scaling is not done, it is difficult to interpret the coefficients of each variable as their ranges would differ.

   Normalization / Min-Max scaling: It rescales the data values to any point between 0 and 1.

   x' = (x-min)/(max-min)

   Standardization: It re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. (Basically a Z-value)

   x' = (x-mean)/standard deviation

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

   VIF is the variance inflation factor. It is used to know about the multicollinearity in the data. Multicollinearity is a variable depending on another variable. High value of VIF indicates high collinearity and low value indicates low collinearity.

   VIF is given by 1/(1-R2)

   In VIF, each feature is a regression against all other features. If R2 is more, it means that the present feature is more correlated with the other features. When R2 reaches 1, VIF becomes infinity. So, if there is a perfect correlation, the VIF value becomes infinity. An infinite VIF value indicates that 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.

   Using Q-Q plots, we can determine if two data sets form a same distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. We can also understand if the two datasets have common location and scale, similar distributional shapes and similar tail behaviour.

   The resultant graph should for a near- straight line if the two data sets come from a same distribution.

   We use Q-Q plots in linear regression model to check if the points lie approximately on the line and If they don't, that would mean that residuals are not gaussian and inturn errors are not gaussian. This implies that, for small sample sizes, our estimator is not gaussian and the standard confidence levels and significance tests are invalid.