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 categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.
- The numerical variables are 'cnt','temp','atemp','hum', and 'windspeed' and the categorical variables are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.

- “season” –
  - Based on the data available, the most favourable seasons for biking are summer and fall.
  - We can expect a lot of bookings during these seasons
  - Spring has comparatively lower bookings among all the seasons.

- “workingday” –
  - Working day represents weekday and weekend/holiday information.
  - Cannot conclude from the data if this is lower or higher for weekdays or weekends.

- “weathersit” –
  - Most preferred weather situation is clear.
  - People do not prefer to use bikes in light snow or rainy conditions.
  - There is no data available for heavy rain/snow days.

- “weekday” –
  - Cannot conclude from the data if this is lower or higher for each of the weekdays.

- “yr” –
  - Data is available for 2018 and 2019.

- “mnth” –
  - The bike rental ratio is higher for April, June, July, August, September and October months.

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

Since we are creating dummy variables for each of the categorical columns so it becomes easier to analyse the treds, this process also adds one redundant variable which will have high colinearity and hence we have to drop the redundant columns with the drop_first= True command.

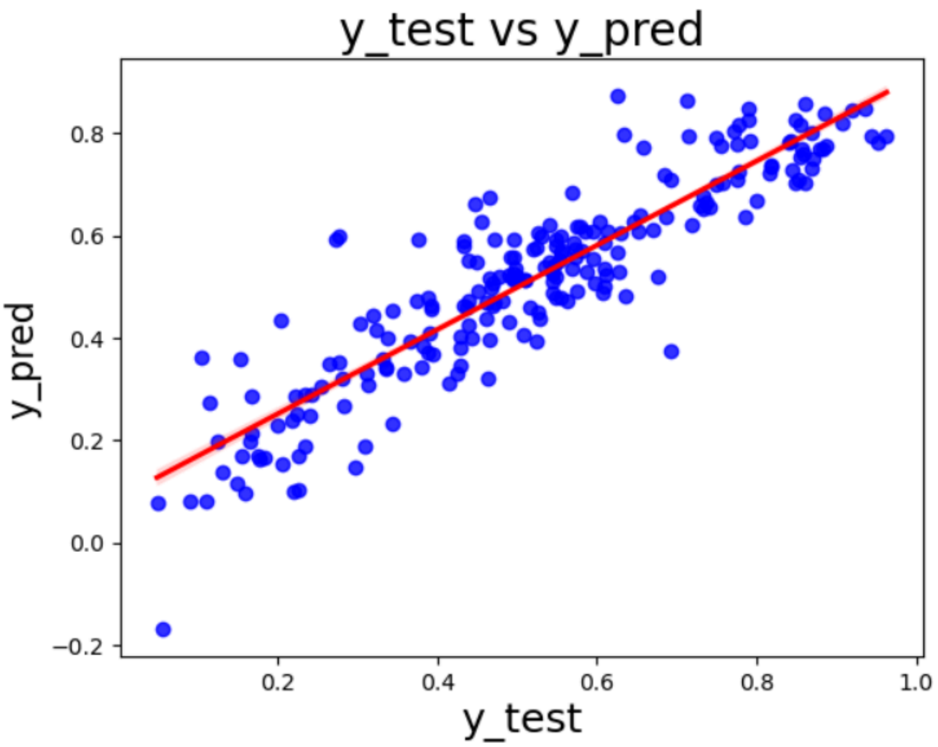
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- "temp" and "atemp" are the variables which have the highest correlation with target variable.
- "atemp" is however dropped as it is a derived value from the "temp" variable and has high colinearity with "atemp"

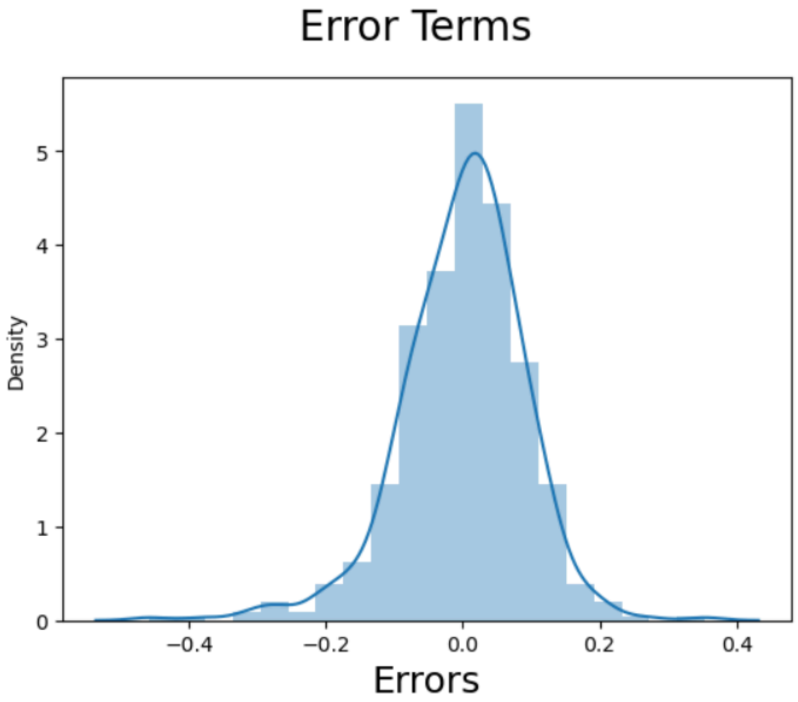
.

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

1. Linear relationship is observed independent and dependent variables – We can see a clear linear relationship in the graph shown below.



2. Error terms are normally distributed: We can see a normal distribution of error terms with mean 0 and with the help of a histogram as shown below.



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

The top 3 variables are:

'temp':

- Temperature is the Most Significant Feature which affects the Business positively.
- Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively**.**

'year':

We have seen a considerable increase within one year and can expect the same with the next.

'mnth_sep':

Winter season is playing the crucial role in the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- The process of determining the optimal linear relationship between the independent and dependent variables is known as linear regression.
- The algorithm maps the relationship between independent variables and dependent variables using the best-fitting line.
- There are 2 types of linear regression algorithms
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \in$ is the line equation for MLR.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, ..., \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The target variable's probability can be predicted by using the cost functions to determine the optimal values for the B0, B1, B2,... BN. In order to find the best-fitted line for predicting the dependent variable, the minimization strategy is employed to lower the cost functions.
- Errors occur when mapping the real numbers to the line, which we discover while searching for the best-fit line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
  - $RSS = = \sum_{i=1}^{n} \left(y_i - y_{pred}\right)^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe's quartet in detail.

- Variance and standard deviation are two statistics that are typically thought to be sufficient to explain some data fluctuation without actually examining every data point. The data's broad trends and other features are well-described by the statistics..
- Anscombe's Quartet indicates that even when two or more data sets have many comparable statistical features, they may not plot identically.

3. What is Pearson’s R?

The strength of the relationship and correlation between the various variables are measured by the Pearson's R, also referred to as Pearson's correlation coefficients. Values between -1 and 1 are returned by the Pearon's R. The coefficients can be interpreted as follows:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma x^{*}y)-(\Sigma x)^{*}(\Sigma y)}{\sqrt{\left[n\Sigma x^{2}-(\Sigma x)^{2}\right]^{*}\left[n\Sigma y^{2}-(\Sigma y)^{2}\right]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx^2 *= the sum of squared x scores*

Σy^2 *= the sum of squared y scores*

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

- The regression model's data preparation stage is scaling. These many datatypes are normalized to a specific data range via the scaling method.
- The majority of the time, the features in the gathered data set vary greatly in terms of range, magnitude, and unit. Erroneous modelling results from algorithms that solely consider magnitude rather than units when scaling is neglected. Scaling is required to resolve this problem by bringing all of the variables' magnitudes to the same level. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x-(x)}{(x) -(x)}$$

Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

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

$$VIF = \frac{1}{1-R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

The quantile-quantile graphs are called Q-Q plots. A graphical tool is utilised to evaluate whether the two data sets have a common distribution. There are three possible types of theoretical distributions: normal, exponential, and uniform. In linear regression, the Q-Q plots help determine if the train and test data sets come from populations with similar distributions. Here's another way to verify that the data sets' normal distribution is a straight line with the patterns described below.

- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.