## Assignment-based Subjective

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

Ans : Following are our categorical variables :

- **'Season'** : The seasons like ‘winter’, ‘spring’ have a positive effect on the dependent variable.

  *As per analysis : a) If all variables remain constant, we can expect the demand of bike to increase by 0.003 in spring months*
  *b) If all variables remain constant, we can expect the demand of bike to increase by 0.127 in winter months*

- **'Mnth'** : Few months have positive effect on the variable namely ‘Sept’,while months like ‘Jul’ has negative effect on the dependent variable

  *As per analysis: a) If all variables remain constant, we can expect the demand of bikes to increase by 0.055 in Sept month.*
  *b)If all variables remain constant, we can expect the demand for bikes to decrease by 0.101 in July month.*

- **Weathersit** : The situations like ‘Mist and Cloudy’ have a positive effect on demand for bikes, could be because such weather situations call for a bike ride while situations like heavy rain or snow have a negative effect on demand for the bikes.

  *As per analysis : a) If all variables remain constant, we can expect the demand for bikes to decrease by 0.277 in weather situation of Light Rain or Snow*
  *b)If all variables remain constant, we can expect the demand for bikes to decrease by 0.059 in weather situation of Mist and cloudy*

- '**Yr**': In coming years, the demand for bikes will have positive effect

  *As per analysis : a) If all variables remain constant, we can expect the demand for bikes to increase by 0.242 every year*

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

**Ans** : **drop_first =True,** drops the first column during the dummy variable creation process. Its importance revolves around decreasing the correlation amongst dummy variables by eliminating the extra columns created during the dummy creation process.

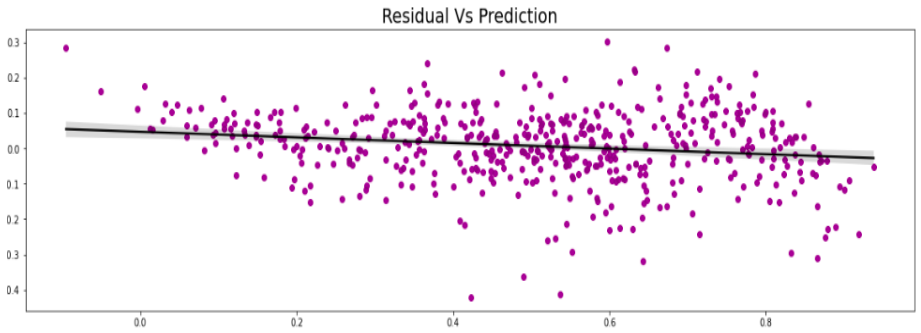
*For ex : I*n the given dataset, we have a month variable with 12 columns. After creating dummy variables for it, the same dataset will represent 11 columns. So let's suppose "Jan" has been dropped, but it still can be represented when all remaining 11 columns are showing '0'.

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

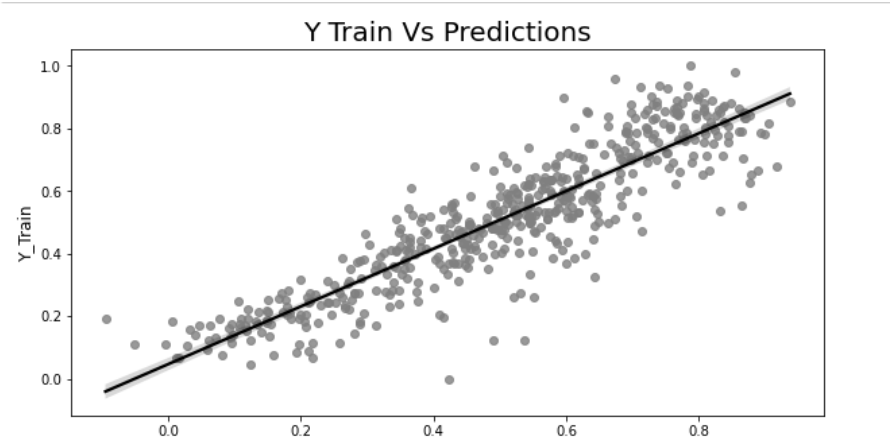
**Ans :** The highest correlation is observed with 'registered' variable.

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

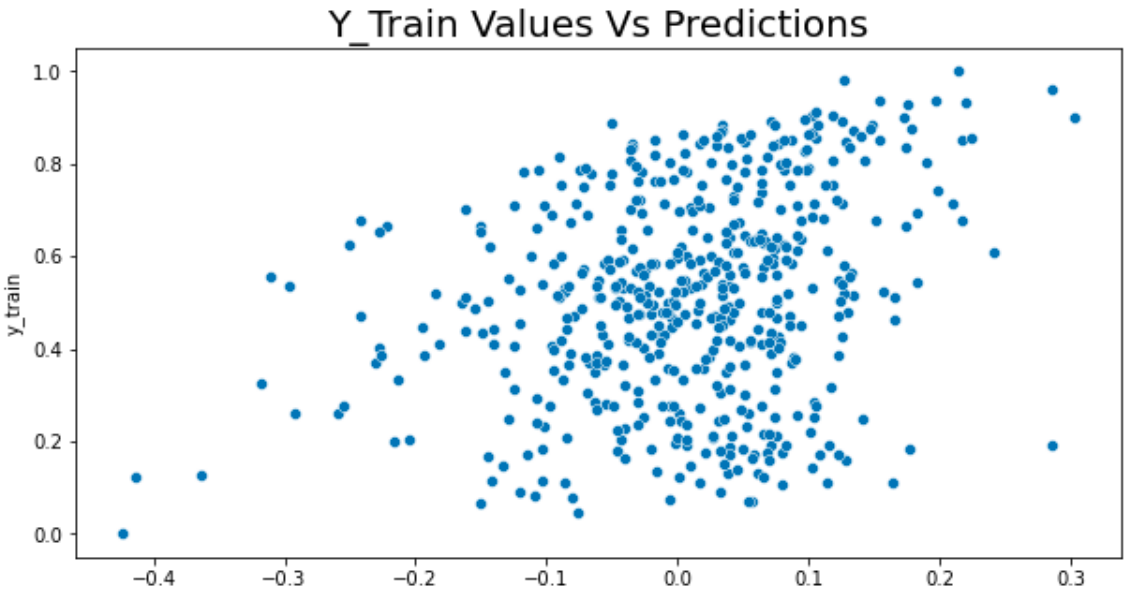
Ans : For checking Linearity, we plot a graph for y_train and X_train while performing residual analysis, which is shown below



For Homoscedasticity, we plot a graph between y_train and y_train_predicted which is shown below



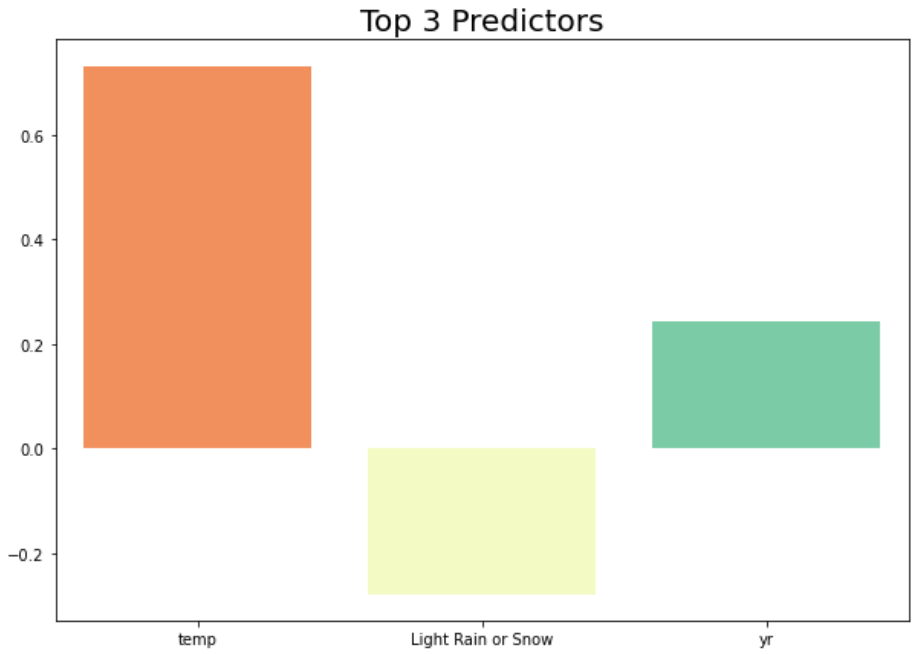

For Independence, we plot a scatter plot between y_train and residuals, which is shown below



**Question 4 :** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans : The top 3 features are temp, light Rain or Snow, yr.

| | |
|---|---|
| temp | 0.731187 |
| Light Rain or Snow | -0.276852 |
| yr | 0.241892 |

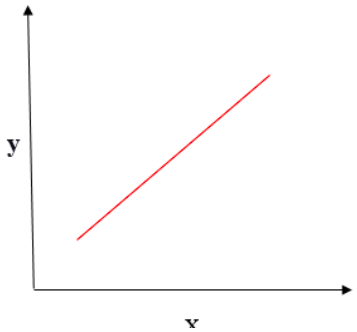


## General Subjective Questions

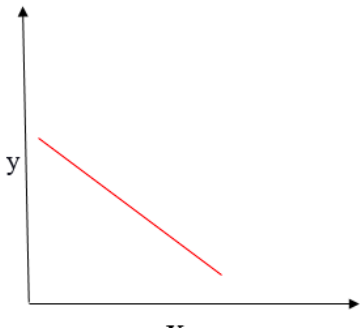
**1.** Explain the linear regression algorithm in detail.

**Ans :** It is a machine learning algorithm which performs regression related tasks on supervised learning. It helps in predicting the relationship between the target variable and independent variables, and how the change in independent variables affect the target variable.

It can be positive or negative.




***Positive LR***




***Negative LR***

Linear regression basically finds our linear relationship between the input variable and the output variable and helps in finding the best fit line for the chosen model

Simple linear regression is depicted by equation

$$Y = mx + c$$

where, m= slope

c= intercept

Multiple linear regression is depicted by equation

$$Y = b1x1 + b2x2 \ldots bnxn$$

So basically, we plot a graph between variables to choose the best fit line which has the least error/loss.

**2.** Explain the Anscombe's quartet in detail. (3 marks)

Ans Anscombe's Quartet is a group of four data sets which seems almost identical in nature but have a distinct distribution of data which confuses the regression model.

Let's talk about the 4 data sets

1. This data set fit the linear regression model aptly
2. This data doesn't fit the model since data taken is non-linear in nature
3. It involves outliers which are beyond the control of linear regression model
4. It depicts an example where one high-leverage point caters to producing high correlation

**Purpose** : Helps in visualising all the data variables while developing and implementing the machine learning algorithm

**3.** What is Pearson's R?

**Ans** : Pearson correlation coefficient is the measure of strength of a linear association. The values varies between -1 to +1 depicting:

+1 : highly positively correlated
-1 : negatively correlated

Following assumptions are made before checking the Pearson's R:

- The relationship between X and the mean of Y is linear.
- The variance of residual is the same for any value of X.
- Observations are independent of each other.
- For any fixed value of X, Y is normally distributed.

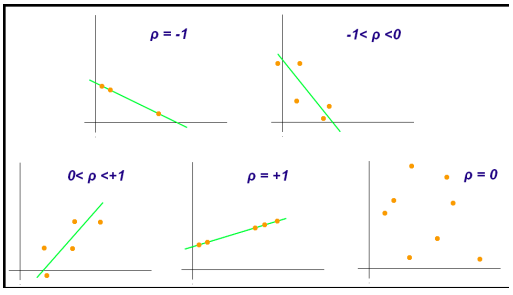



Fig. *Graph for different values of p*

**4.** What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

**Ans** : **Scaling** is a technique to normalise the range of independent variables of the data set. It's part of preprocessing step in linear regression

We perform scaling to bring the distributed values to a same magnitude, helping to perform better regression analysis. If we don't perform scaling, the machine learning will take more time to understand and perform the analysis

The difference between the normalised scaling and standardised scaling os that ,
Normalisation rescale the variable with the distribution varying between 0 and 1 while Standardisation uses mean value and the distribution varies between 0 mean value while variance is kept at 1.

The Normalisation uses minimum  and maximum value of the variable while standardisation uses standard deviation.

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

**Ans** : VIF basically provides the measure of variance and how much it increases due to collinearity. It's used to deal with multicollinearity

VIf = infinite, exists in the case of perfect correlation that means the given variable can be precisely expressed by the linear combination of other existing variables. In this case, R square is close to 1

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

**Ans.**  A Quantile-Quantile plot is a plot of 2 quantiles against one another.  A graphical way that helps to analyse if the given data set has a distribution such as  uniform, normal, exponential.

Use : It can be used on sample data as well. It helps in finding out outliers

Importance : It basically helps in determining if the two data sets are from a population with a common distribution. It provides graphical presentation of distribution of properties like location scale, skewness etc.