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?**

   EDA on Categorical Variable:

   a. SEASON :
      i. Fall, Summer and Winter have a higher average rental bike per day than overall average rental
      ii. Fall and Summer have a higher average rental bike per day than Winter and Spring

   b. YR :
      i. 2019 year have a higher average rental bike per day than overall average rental
      ii. 2019 year have a higher average rental bike per day than 2018 year

   c. MNTH :
      i. 6,9,8,7,5 and 10 months have a higher average rental bike per day than overall average rental
      ii. 6,9,8,7,5 and 10 have a higher average rental bike per day than 4,11,3,12,2,1

   d. HOLIDAY :
      i. Non-Holidays have a higher average rental bike per day than overall average rental
      ii. Non-Holidays have a higher average rental bike per day than Holidays

   e. WEEKDAY :
      i. Weekday 2,3,4,5,6 have a higher average rental bike per day than overall average rental
      ii. Weekday 2,3,4,5,6 have a higher average rental bike per day than Weekday 1,2

   f. WORKINGDAY :
      i. Working day have a higher average rental bike per day than overall average rental
      ii. Working day have a higher average rental bike per day than Non-Working Day

   g. WEATHERSIT :
      i. Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than overall average rental
      ii. Clear, Few clouds, Partly cloudy, Partly cloudy Weather have a higher average rental bike per day than Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

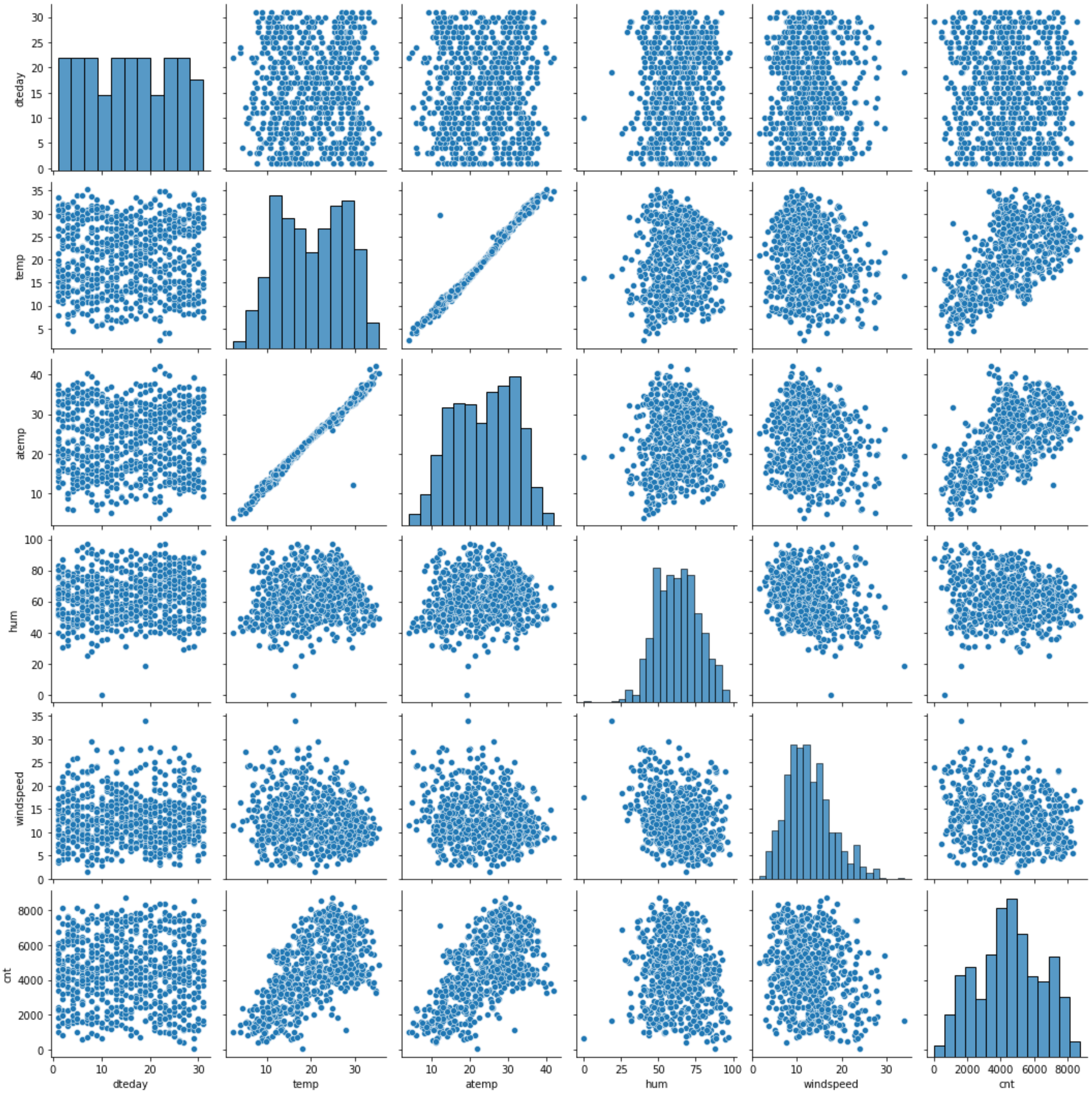
2. **Why is it important to use drop_first = True during dummy variable creation?**

   It is important to use drop first as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   By dropping one of the one-hot encoded columns from each categorical feature, we make sure that there are no reference columns and the remaining columns become linearly independent.
   If there are n categorical values then n-1 dummy columns have to be created.
   In the BoomingBikes Sharing Assignment the following are the categorical columns for which we have to create the dummy variables.

   a. SEASON
   b. MNTH
   c. WEEKDAY
   d. WEATHERSIT

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

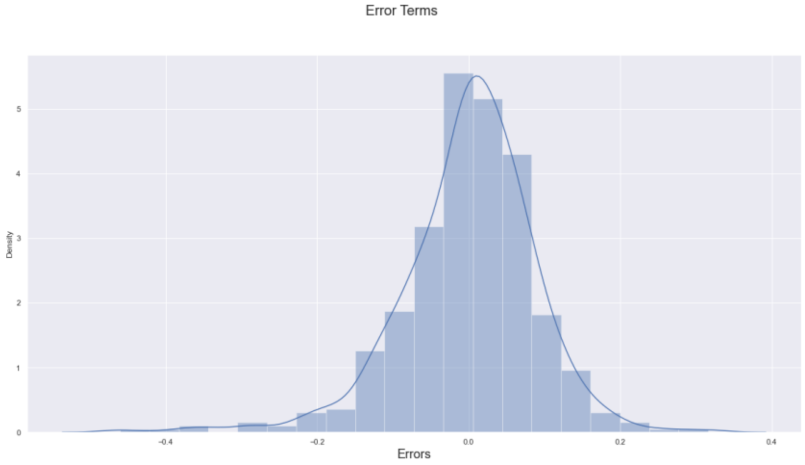
Pair plot for the numerical variables:



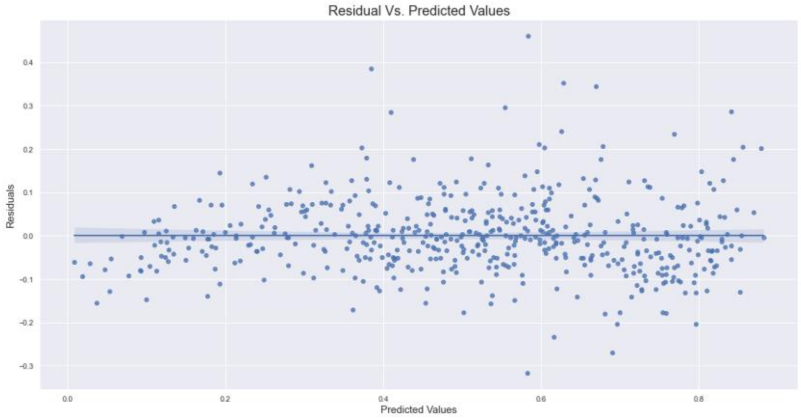
Looking at the above pair plot we can see the “temp” and “atemp” have the linear correlation with the target variable “cnt”. Looking at the couple of points on the edges of the scatter plot we can conclude that the “temp” has the highest correlation with the target variable.

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

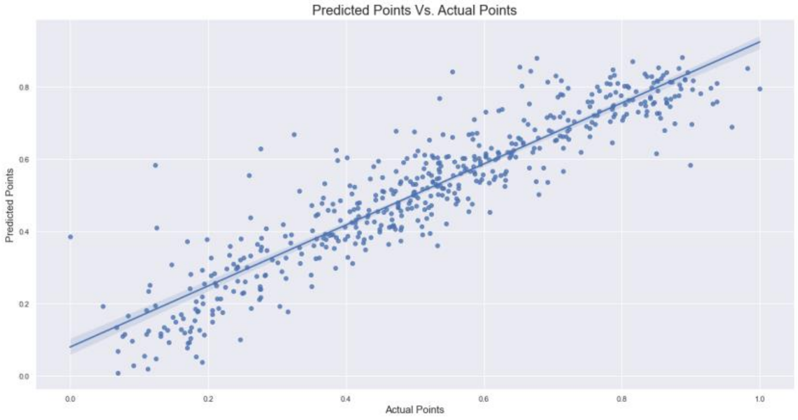
a. Assumption of Normally Distributed Error Terms
Validated by plotting a distribution of the residuals/error.



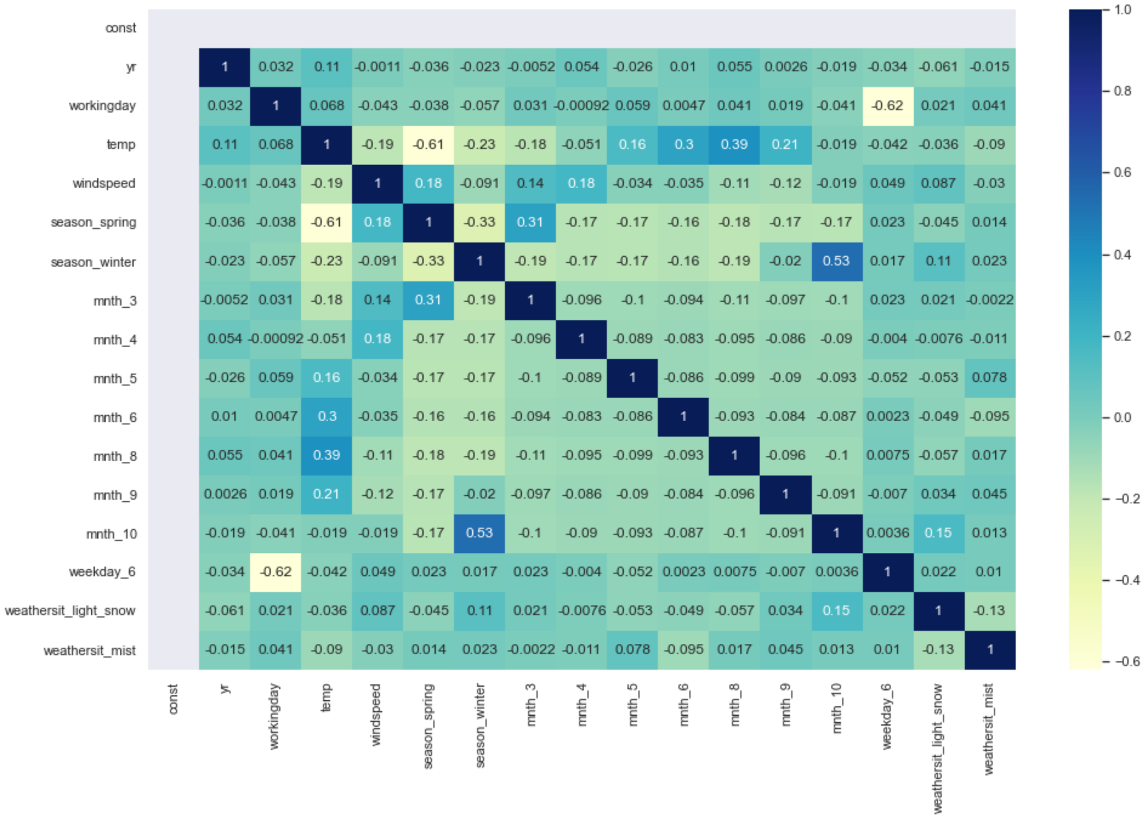
b. Assumption of Error Terms Being Independent
Validated by plotting a regression plot of the residual’s vs predicted values(0-1).



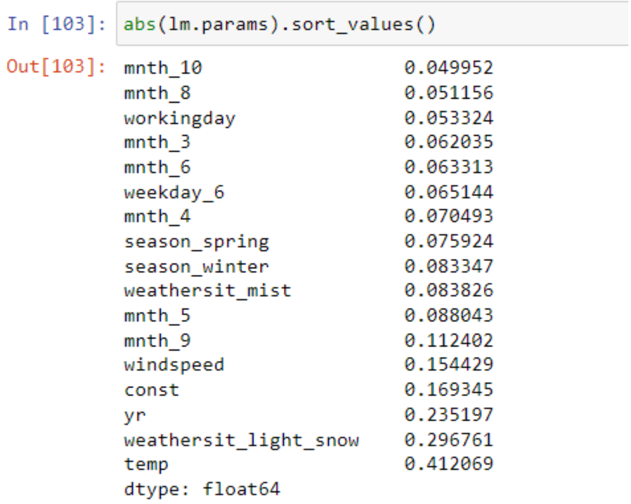
c. Assumption of Homoscedasticity (Constant Variance)
Validated by plotting a regression plot of the Actual values vs predicted values(0-1) for the target variable.



d. Assumption of No or little multi-correlation
Validated by plotting a heat map for the correlation matrix and Variance inflation factor (VIF) of the independent variables



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



```
In [103]: abs(lm.params).sort_values()

Out[103]: mnth_10                  0.049952
          mnth_8                   0.051156
          workingday               0.053324
          mnth_3                   0.062035
          mnth_6                   0.063313
          weekday_6                0.065144
          mnth_4                   0.070493
          season_spring            0.075924
          season_winter            0.083347
          weathersit_mist          0.083826
          mnth_5                   0.088043
          mnth_9                   0.112402
          windspeed                0.154429
          const                    0.169345
          yr                       0.235197
          weathersit_light_snow    0.296761
          temp                     0.412069
          dtype: float64
```

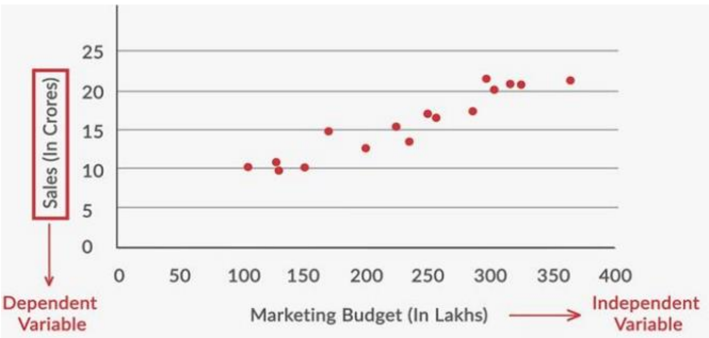
Based on the absolute coefficient values for each of the variables the top 3 features contributing towards demand are:

a. temp
b. weathersit_light_snow
c. yr

General Subjective Questions

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

Linear : The relationship between the input and output variables is a straight line.



Regression : output variable is continuous e.g., sales and demand.

There are two types of linear regression models:

a. Simple linear regression
b. Multiple linear regression

Linear regression can be represented as:

Linear Regression: Single Variable

$$\underbrace{\boxed{\hat{y}}}_{\text{Predicted output}} = \underbrace{\beta_0 + \beta_1}_{\text{Coefficients}} \underbrace{\boxed{x}}_{\text{Input}} + \underbrace{\boxed{\epsilon}}_{\text{Error}}$$

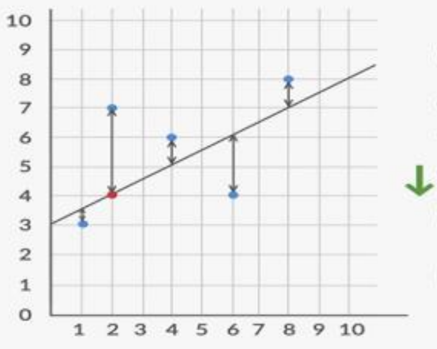
Linear Regression: Multiple Variables

$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}\boxed{x_1} + \ldots + \underbrace{\beta_p}\boxed{x_p} + \boxed{\epsilon}$$

Assumption made in linear regression model :

a. **Linearity**: The relationship between X and the mean of Y is linear.
b. **Homoscedasticity**: The variance of residual is the same for any value of X.
c. **Independence**: Observations are independent of each other.
d. **Normality**: For any fixed value of X, Y is normally distributed.

We can find the best-fit line by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot**.**



$$Y = \underset{\underset{\text{Intercept}}{\downarrow}}{\beta_o} + \underset{\underset{\text{Slope}}{\downarrow}}{\beta_1} X$$

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

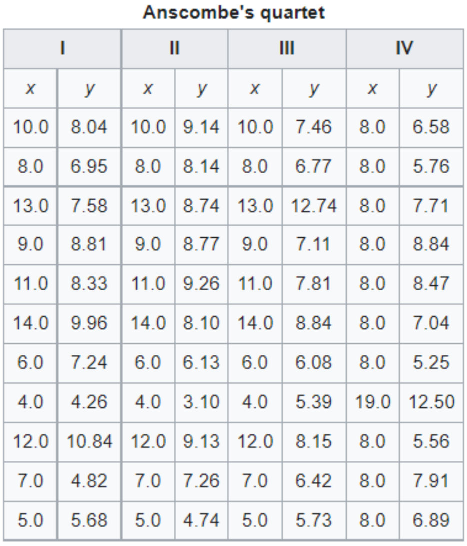
$$\downarrow e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$\text{RSS} = (Y_1-\beta_o-\beta_1X_1)^2 + (Y_o-\beta_o-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_o-\beta_1X_n)^2$$

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1X_i)^2$$

2. **Explain the Anscombe's quartet in detail.**

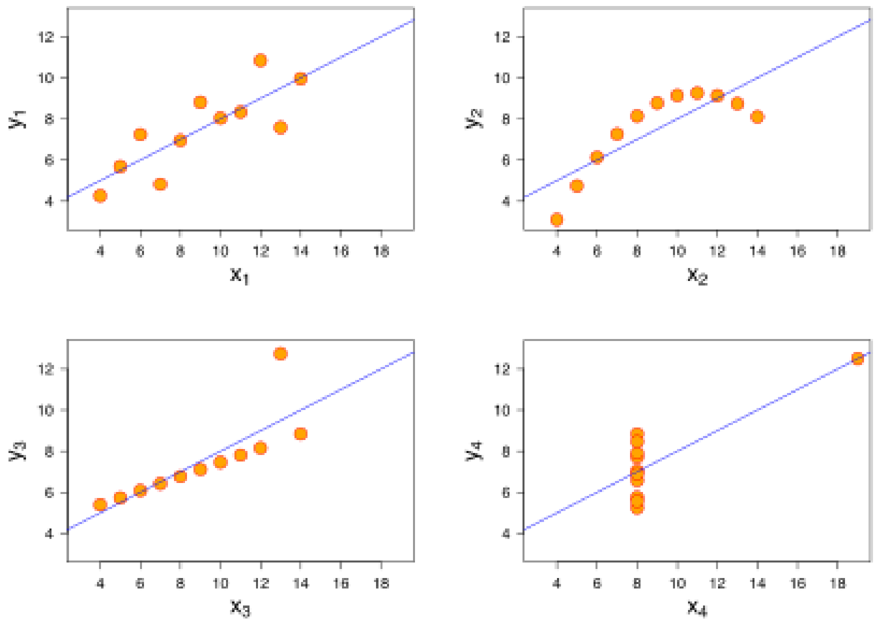
Anscombe's quartet has 4 dataset that have 11 datapoint (x , y).



Anscombe's quartet

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

It demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.



a. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

b. The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

c. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

d. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. **What is Pearson’s R?**

It’s the measure of linear correlation between two variables in a dataset**.**

Pearson’s R value lies between -ve 1 to +ve 1

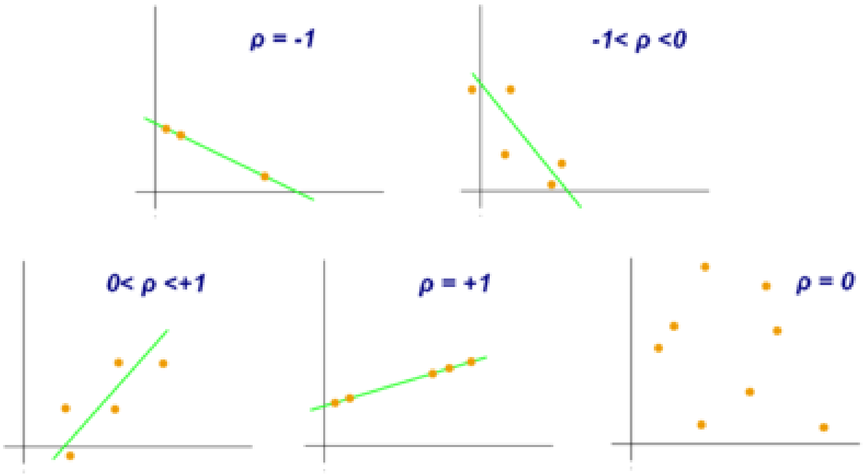
Where,

-1 would represent perfect negative correlation

0 would represent no correlation

+1 would represent perfect positive correlation

It is the covariance of 2 variables by product of their standard deviation.



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

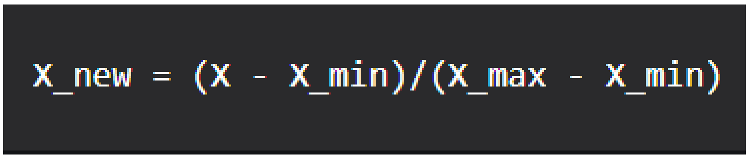
Scaling is one of the important steps in creating an efficient linear regression model. It is done during data preparation step. It is applied on the numerical variables to normalize the data in a specific range. Scaling helps in speeding up the calculations in an algorithm.

Scaling is performed to normalize the variables. It brings all the variables in a specified range i.e., same level of magnitude. As mentioned, it is also used to speed up the calculation of minimizing the cost function.

There are main two types of scaling are :

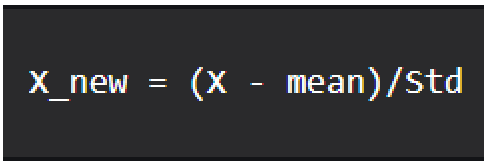
- Normalized Scaling : Scales the data between 0 and 1. Normalization is useful when the data is needed in the bounded intervals.

  E.g., we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.



```
X_new = (X - X_min)/(X_max - X_min)
```

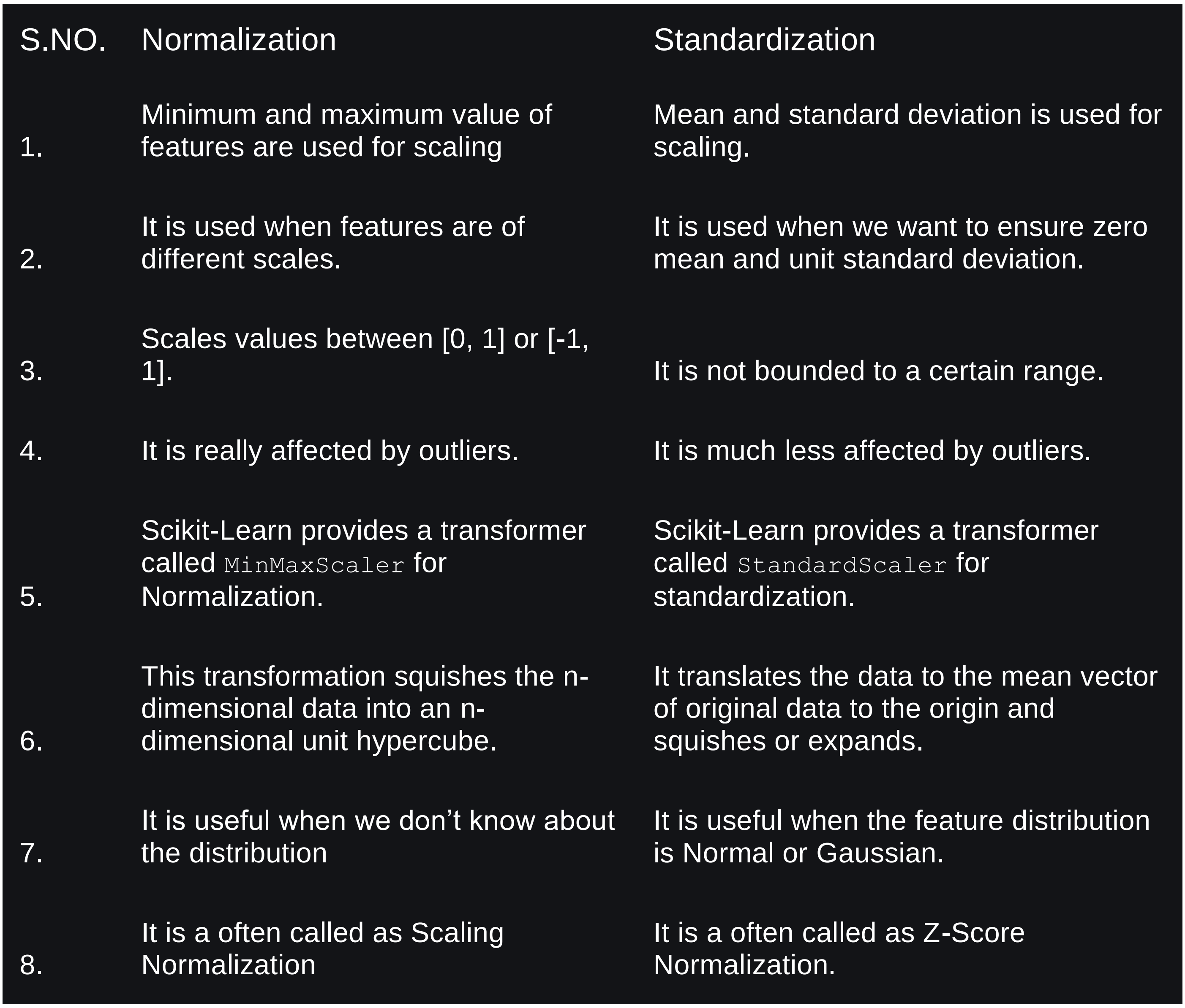
- Standardisation Scaling : Scales the data into standard normal distribution where the mean of the dataset is zero.



```
X_new = (X - mean)/Std
```

The difference:

- In normalization, you're changing the range of your data, while in Standardisation, you're changing the shape of the distribution of your data.
- In Normalization, information about the outliers is lost, while in Standardisation, information about the outliers is not lost.



S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Source : https://www.geeksforgeeks.org/normalization-vs-standardization/

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

VIF – Variance Inflation Factor is the measure of multi-collinearity between the independent variables.

VIF can be calculated using the below formula:

$$VIF_1 = \frac{1}{1 - R^2}$$

VIF values ranges from 1 to infinity

- VIF = 1 :
  R^2 = 0
  indicates the total absence of collinearity between these variable and other variables in the model.
- VIF = 5 :
  5 = 1/1-R^2
  1- R^2 = 0.2
  R^2 = 0.8
  It means that the other variables contribute to 80% of the variance for the give variable.
- VIF = Infinity :
  R^2 = 1
  indicates the given can be perfectly predicted by other variables in the model

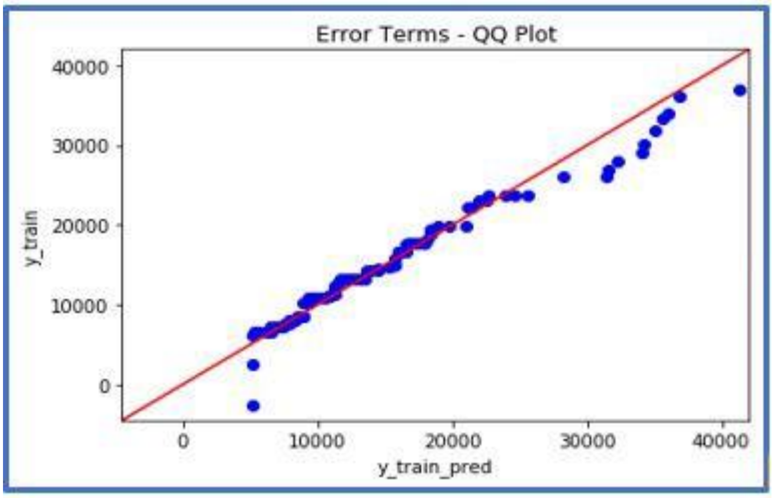
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Quantile-Quantile (Q-Q) plot is a graphical technique to check if a sample follows a particular distribution. It can also be used to check if two different samples follow the same distribution. It is a scatter plot between the quantiles of the two distributions that need to be compared. A Q-Q plot approximately results in the straight-line y=x if the two distributions are identical.

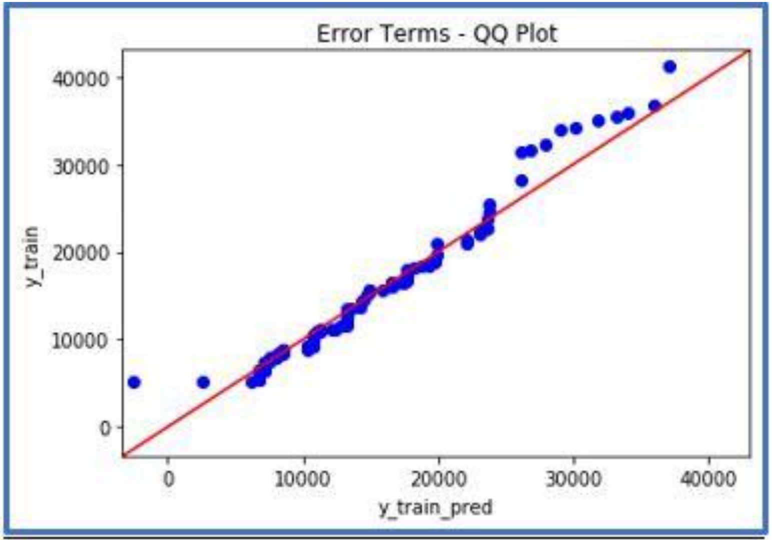
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X -values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis