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?

- **Seasonal Analysis:** Fall has the highest average rentals, followed closely by summer.
- Year-wise Rentals: 2019 sees a notable increase with a median rise of approximately 2000 rentals compared to 2018.
- **Monthly Trend:** September tops the monthly rental count, with surrounding months showing substantial demand. The trend aligns with seasonal patterns, indicating a correlation between rentals and seasons.
- **Temperature:** has the top impact on the bike rentals
- **Holiday vs. Working Days:** Holidays generally result in lower rental counts compared to working days. Holidays exhibit greater variability in rental demand.
- **Weekday Analysis:** Overall, no significant difference in rentals across weekdays is observed. Thursdays and Sundays stand out with higher variability in rental counts compared to other weekdays.

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

To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data.

Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.
If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known **as multicollinearity** and it, in turn, leads **to Dummy Variable Trap.**

Mathematically it can be explained by considering a regression model which is used to find population rise in 3 different states as below where X3 represents the 3 different states name.

$Y = \beta 0 + \beta 1 (X1) + \beta 2 (X2) + \beta 3(X3) + \varepsilon$ ---(i)

As X3 is a categorical variable that contains 3 different state names, we can assign 3 dummy variables D1 as [100], D2 as [010], and D3 as [001] for each state in our equation.

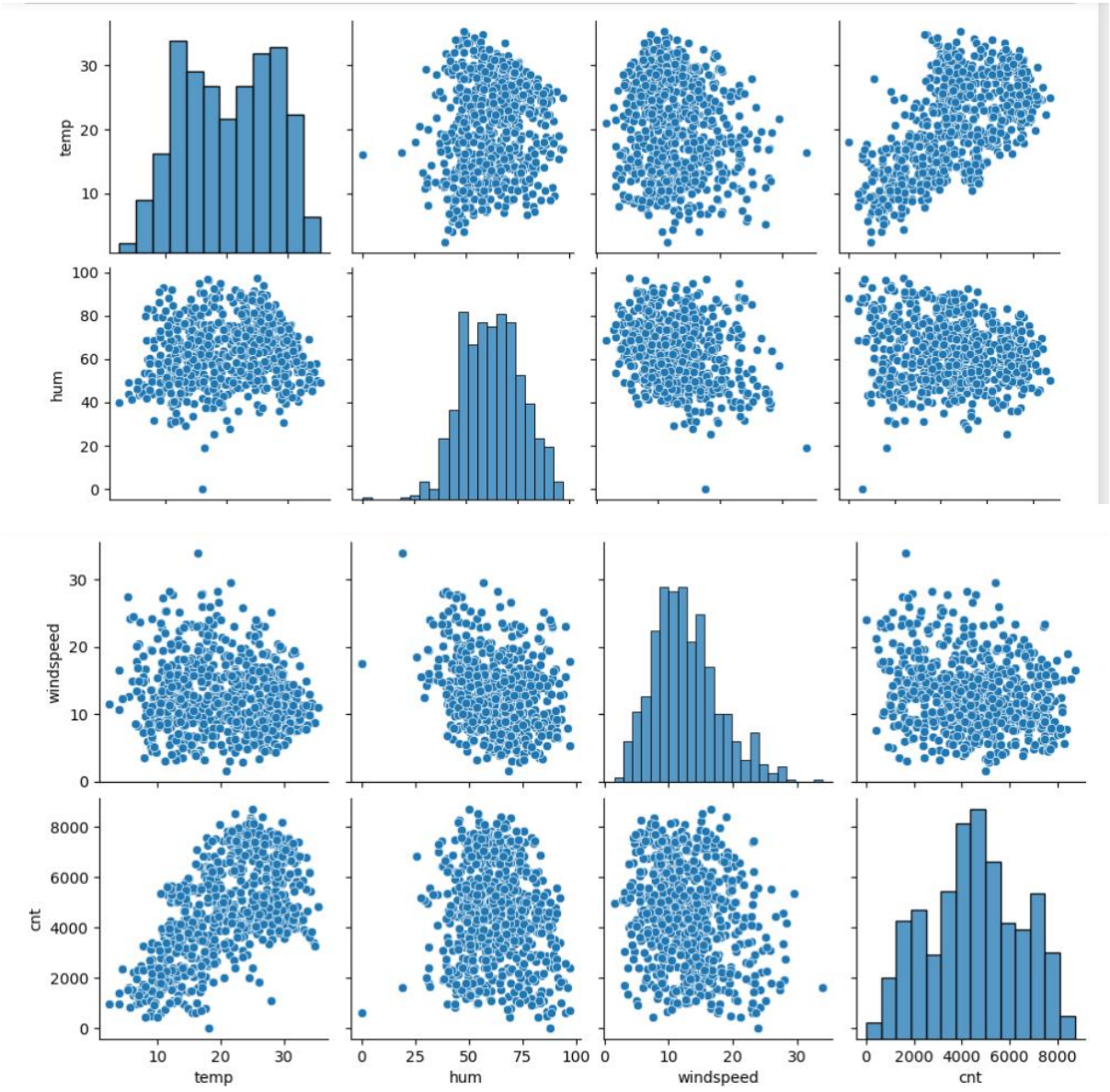
D1 + D2 + D3 = 1
D3 = 1- (D1 + D2) ---(ii)

The last equation indicates D3 is perfectly explained by the other two dummy variables D1 and D2.

In a broader sense, we can conclude that if there are n dummy variables, n-1 dummy variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity.

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

temp have high correlation with target variable of 0.63 which is the highest among all numerical variables.

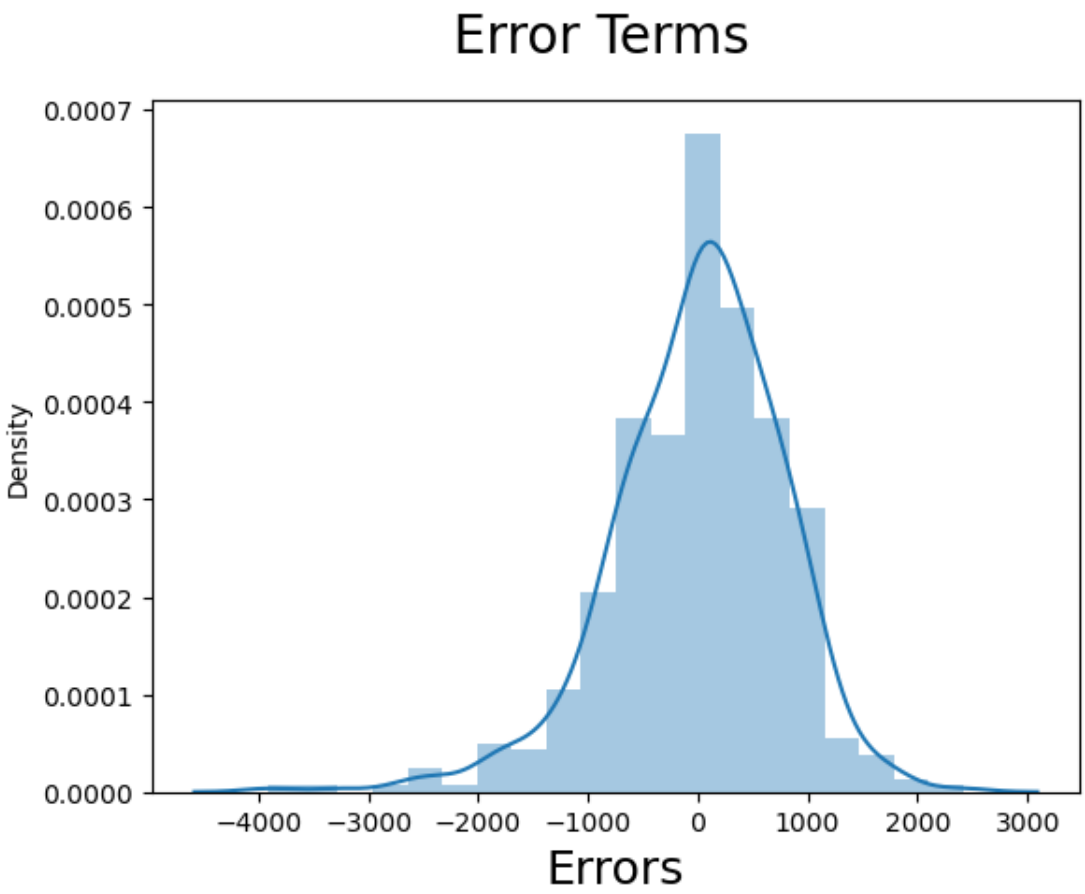


4.How did you validate the assumptions of Linear Regression after building the model on the training set?

We validated this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.

Checked the following assumptions a s well:

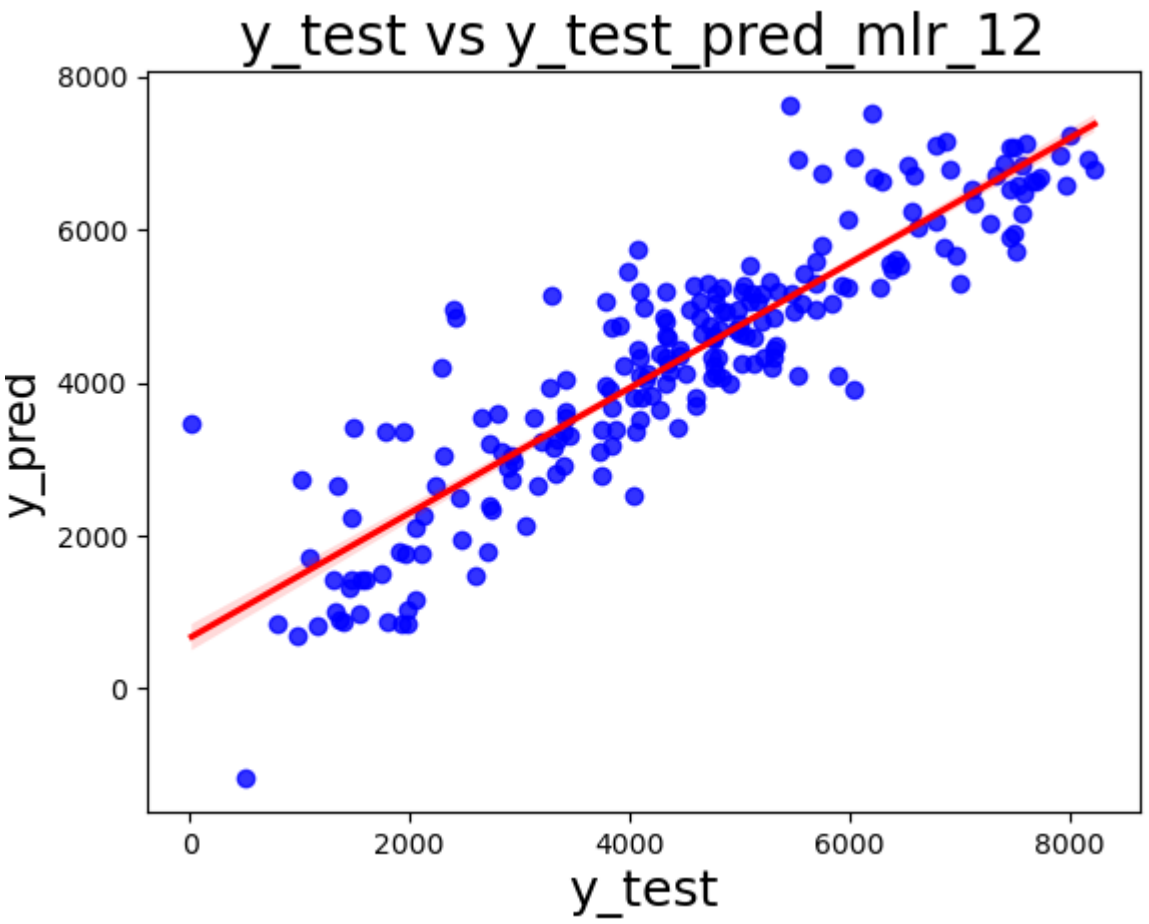
•Error Terms do not follow any pattern.



- **Multicollinearity check using VIF(s).**

	Features	vif
1	temp	4.71
2	windspeed	4.02
4	Winter	2.42
0	yr	2.06
7	November	1.72
3	Spring	1.69
9	Cloudy_mist	1.52
5	December	1.44
6	July	1.37
8	September	1.23
10	Light_Rain_Thunder	1.07

• **Linearity Check.**



•**Ensured the overfitting by looking the R2 value and Adjusted R2.**

```
#Calculate the r square for test

r_squared = r2_score(y_test, y_test_pred_mlr_12)
r_squared
```

```
0.7912052419702502
```

Below is the summary of the final model based on which the predictions were made :

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.842
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	241.6
Date:	Sun, 28 Jan 2024	Prob (F-statistic):	1.28e-191
Time:	21:52:25	Log-Likelihood:	-4120.7
No. Observations:	510	AIC:	8265.
Df Residuals:	498	BIC:	8316.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2349.7216	207.312	11.334	0.000	1942.408	2757.035
yr	2030.2929	70.757	28.694	0.000	1891.274	2169.312
temp	3695.8589	268.081	13.786	0.000	3169.150	4222.568
windspeed	-1041.5648	214.738	-4.850	0.000	-1463.469	-619.660
Spring	-1082.0269	131.315	-8.240	0.000	-1340.026	-824.028
Winter	581.5041	121.090	4.802	0.000	343.593	819.415
December	-372.3484	139.609	-2.667	0.008	-646.643	-98.054
July	-616.1436	148.379	-4.153	0.000	-907.669	-324.618
November	-504.6424	162.342	-3.109	0.002	-823.602	-185.683
September	421.8712	132.161	3.192	0.002	162.209	681.533
Cloudy_mist	-766.6253	75.466	-10.159	0.000	-914.897	-618.354
Light_Rain_Thunder	-2319.4878	219.220	-10.581	0.000	-2750.198	-1888.778

Omnibus:	69.093	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.043
Skew:	-0.766	Prob(JB):	3.89e-31
Kurtosis:	5.061	Cond. No.	14.3

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

- Year(yr)
- Temperature(tmp)
- Windspeed

General Subjective Questions

1.Explain the linear regression algorithm in detail?

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

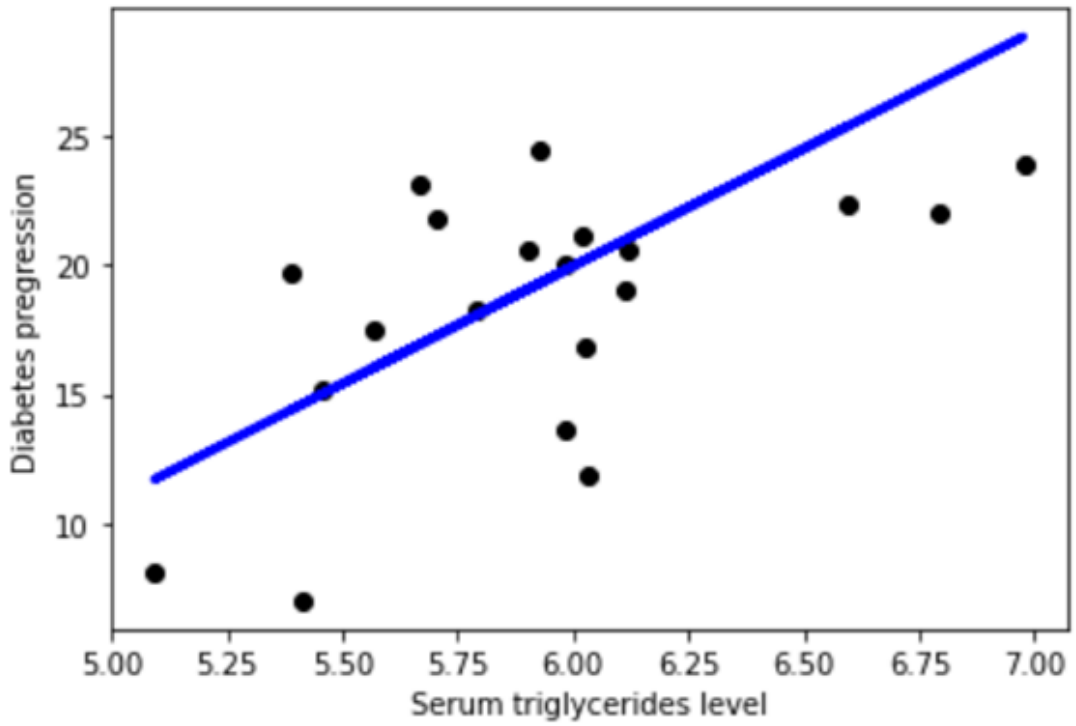
The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots$$

In the example above, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

In the following image, a linear regression model is described by the regression line

y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes pregression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example demonstrates a linear regression model with two variables. Although it is not possible to visualize models with more than three variables, practically, a model can have any number of variables.



A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

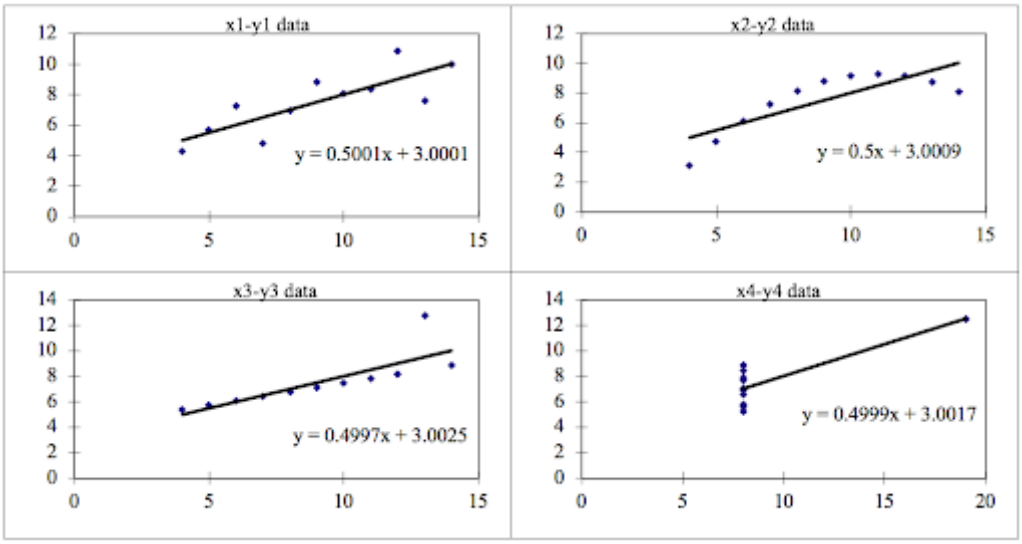
2. Explain the Anscombe's quartet in detail?

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

When these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

Anscombe's quartet four Dataset:

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

When r = 1 positve strong correlation

When r = -1 negative strong corelation

When r = 0 no correlation

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

There are two types of scaling :

1. **Normalization/Min-Max Scaling:** It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. **Standardization Scaling:** Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$X' = \frac{X - \text{Mean}}{\text{Standard deviation}}$$

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

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, 'i' refers to the ith variable. If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

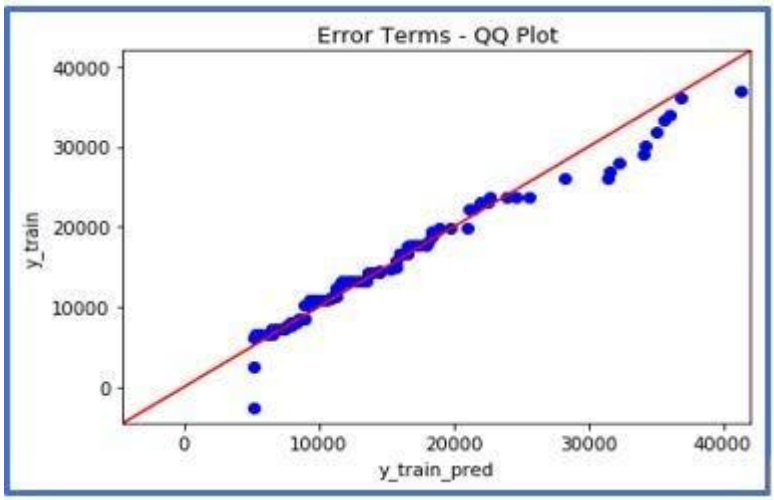
It is used to check following scenarios:

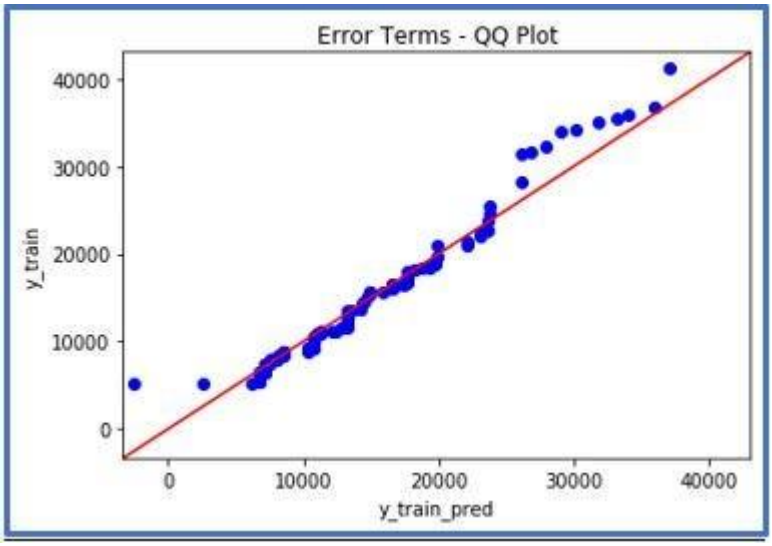
i. come from populations with a common distribution

ii. ii. have common location and scale

iii. iii. have similar distributional shapes

iv. iv. have similar tail behavior





Python:

statsmodels.api provide **qqplot** and **qqplot_2samples** to plot Q-Q graph for single and two different data sets respectively.