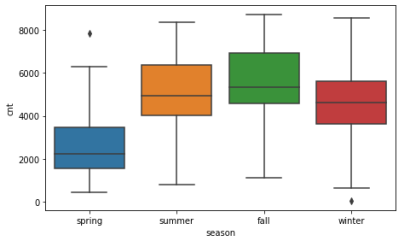
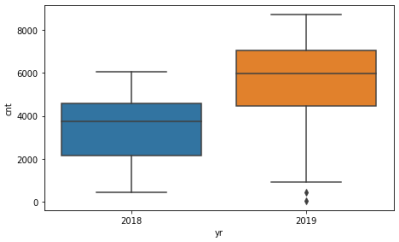
Assignment-based Subjective Questions

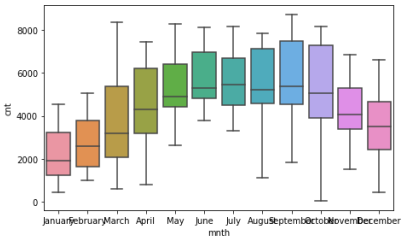
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

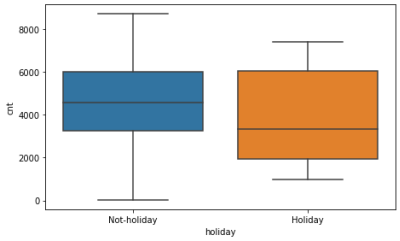
Ans: Based on our analysis of the categorical variables on dataset we can infer the below:

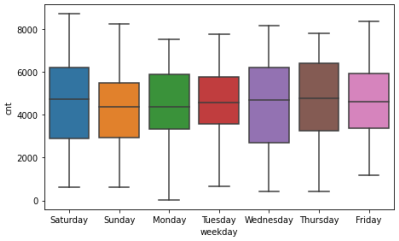
- Frequency of Bike Rental is maximum in Fall season whereas its minimum in Spring season
- Count of Bike Rental was increased in 2019 as compared to 2018.
- August, September and October are the months where the bike rentals are highest whereas in January and February it's the least.
- Count of Bike Rental is more on Working day as compared to non-working day.
- Frequency of Bike Rental when Weather is Clear, Few clouds or Partly cloudy whereas its low when its Light Snow, Light Rain.
- Count of Bike Rental is low on Sunday's as compared to other days of the week.

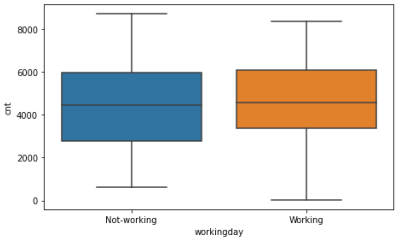


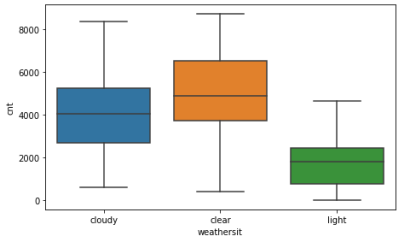












2. Why is it important to use *drop_first=True* during dummy variable creation?

Ans: When our dataset has categorical data and during preparing our dataset to be used for Machine Learning we need to convert our dataset into binary vector representation. For this we use the pandas get_dummies or One-Hot encoding.

For each unique value in the categorical column a new column is created having values as 0 and 1.
So, this 'N' columns can be represented using N-1 columns which reduces complexity and correlations among the dataset

Season	Summer	Spring	Fall	Winter
Summer	1	0	0	0
Spring	0	1	0	0
Fall	0	0	1	0
Winter	0	0	0	1

Here above if we drop one column still, we can represent the data without missing any information.

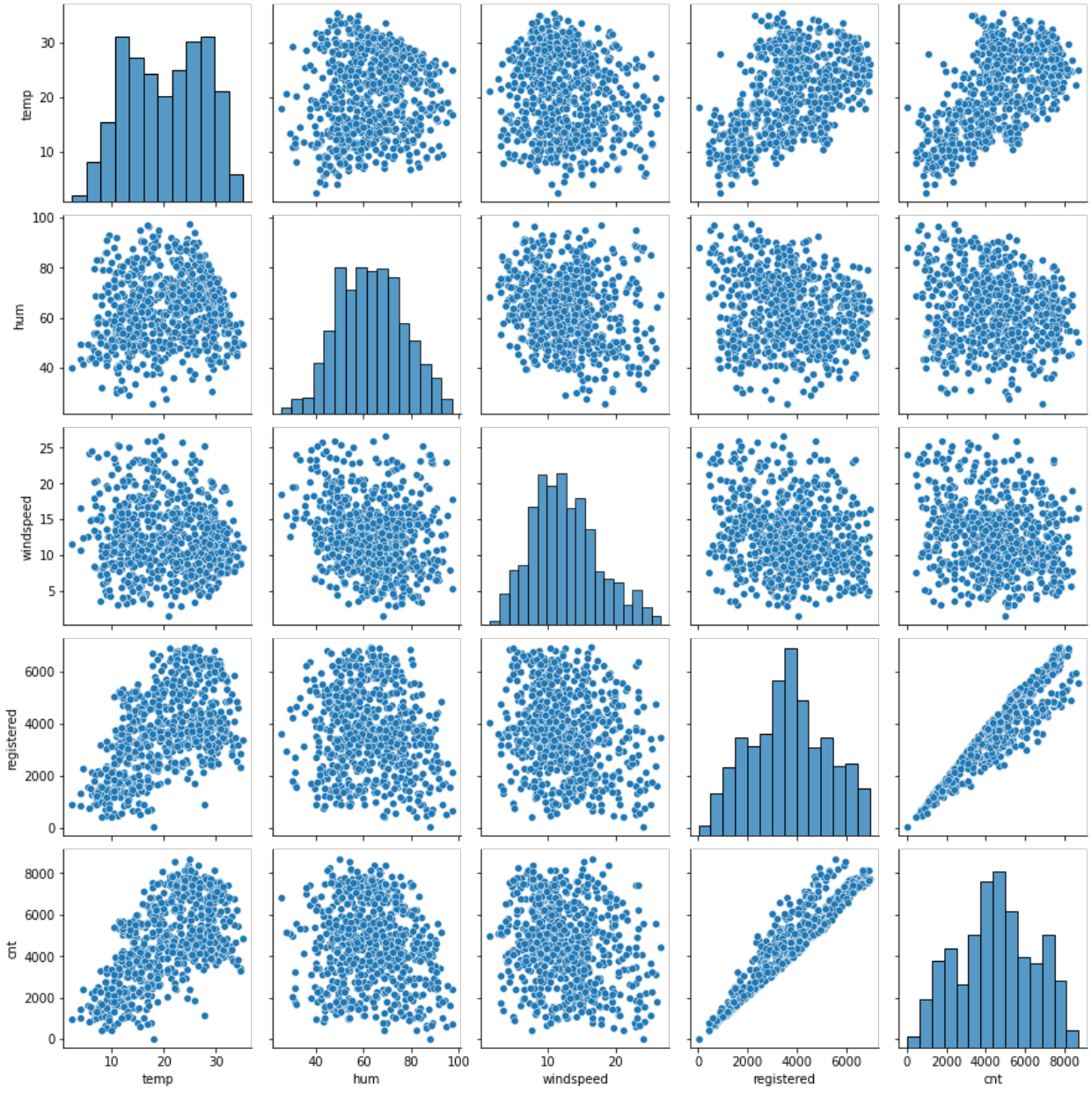
Season	Spring	Fall	Winter
Summer	0	0	0
Spring	1	0	0
Fall	0	1	0
Winter	0	0	1

So, we can say when the data is 000 -> its summer season.

Hence **drop_first=Tru**e is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans: Looking at the pair-plot for the numerical variables we can infer **registered** and **temp** variables have highest correlation with target variable



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

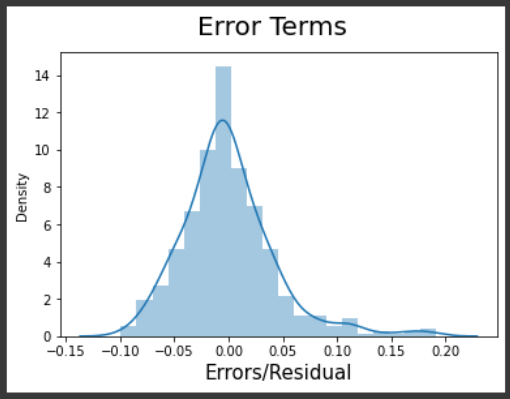
Ans: Following are Assumptions of Linear Regression

- There is linear relation between X & Y
- Error Terms are normally distributed with Mean 0
- Error Terms are independent of each other
- Error terms have constant Variance.

We validated the assumption of our Linear regression model by plotting a distribution graph or the Residual/Error term.

An **error term / residual i**s a value which represents differs of actual data and the predicted data from the model.

```
[ ]  1  # residual/error terms
     2  res = y_train - y_train_pred
     3
     4  # plotting the distribution of error terms
     5  fig = plt.figure()
     6  sns.distplot(res, bins = 20)
     7  plt.xlabel('Errors/Residual',fontsize=15)
     8  fig.suptitle('Error Terms',fontsize=20)
     9  plt.show()
```



From the plot we could infer that error terms are normally distributed with Mean at 0.

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

Ans: Following is the equation of our fitted line is

cnt = 0.0271 * *weekday* - 0.0963 * *workingday* + 0.149 * *temp* - 0.0470 * *hum* - 0.0222 * *windspeed* + 0.9394 * *registered*

From this we can conclude that following features affect the demand of Bike Rental:

- Increases with increase of Registered users
- Increases when there is Weekdays
- Increases with Temprature
- Decreases with high humidity
- Decreases with high windspeed
- Decreases when Working day is a non-working day or no working day.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.966
Model:	OLS	Adj. R-squared:	0.965
Method:	Least Squares	F-statistic:	1986.
Date:	Mon, 12 Sep 2022	Prob (F-statistic):	0.00
Time:	19:41:54	Log-Likelihood:	865.04
No. Observations:	503	AIC:	-1714.
Df Residuals:	495	BIC:	-1680.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0425	0.012	3.408	0.001	0.018	0.067
weekday	0.0271	0.008	3.305	0.001	0.011	0.043
workingday	-0.0963	0.006	-15.628	0.000	-0.108	-0.084
temp	0.1497	0.011	13.601	0.000	0.128	0.171
hum	-0.0470	0.011	-4.377	0.000	-0.068	-0.026
windspeed	-0.0222	0.010	-2.234	0.026	-0.042	-0.003
registered	0.9394	0.012	80.425	0.000	0.916	0.962
summer	0.0290	0.005	6.323	0.000	0.020	0.038

Omnibus:	109.936	Durbin-Watson:	1.865
Prob(Omnibus):	0.000	Jarque-Bera (JB):	265.832
Skew:	1.109	Prob(JB):	1.88e-58
Kurtosis:	5.787	Cond. No.	14.7

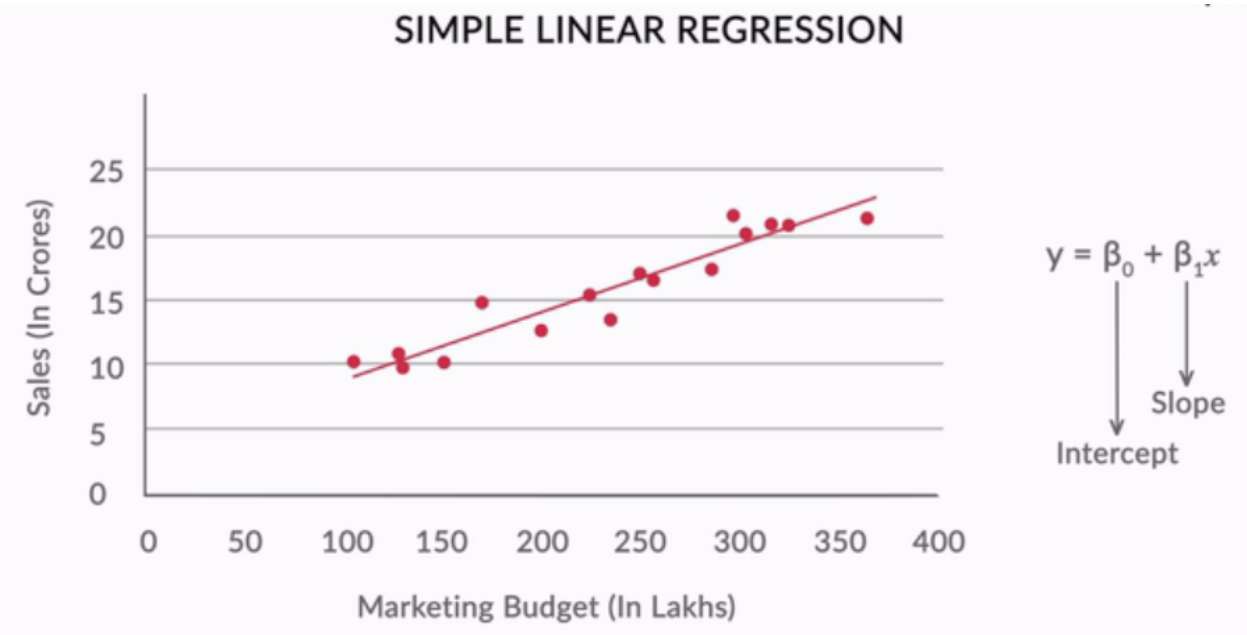
General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans: Regression is a type of Machine Learning model where output variable to be predicted is a continuous variable. Regression and classification fall under **supervised learning methods –** in which you have previous years data with labels and you use that to build the model.

Linear regression is a form of machine learning where we train a model to predict the behaviour of your data based on some dependent variables. In the case of linear regression, it predicts linear i.e., two variables on the x-axis and y-axis should be linearly correlated.

The straight line is plotted on the plot of these two points.



The standard equation of the regression line is given by the following expression:
$Y = \beta_o + \beta_1.X$

Here, x and y are two variables on the regression line.
β_1 = Slope of the line
β_o = y-intercept of the line
X = Independent variable from dataset
Y = Dependent variable from dataset

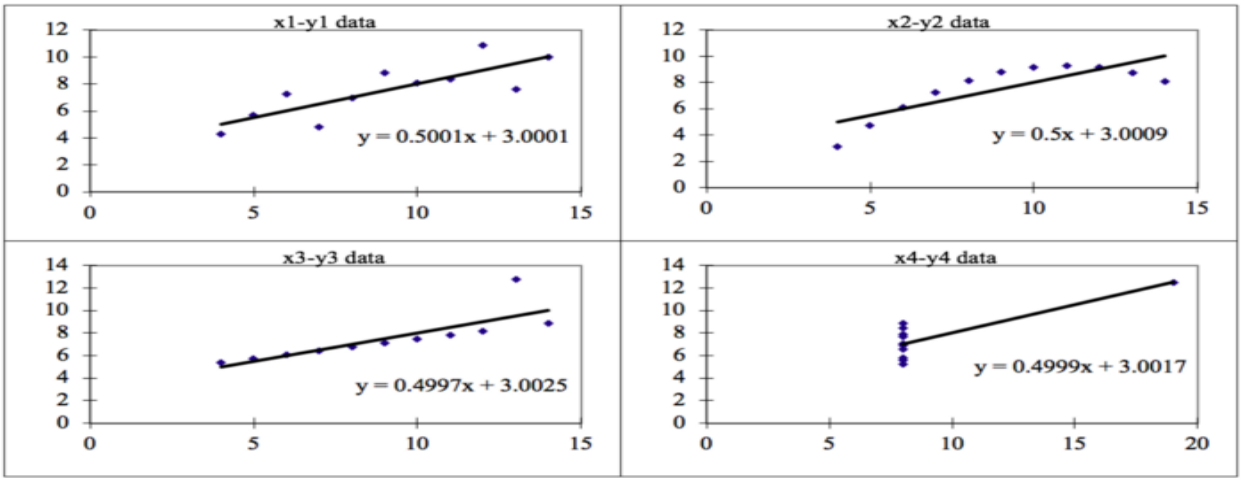
2. Explain the Anscombe's quartet in detail?

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot:



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data as its non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression

Conclusion:

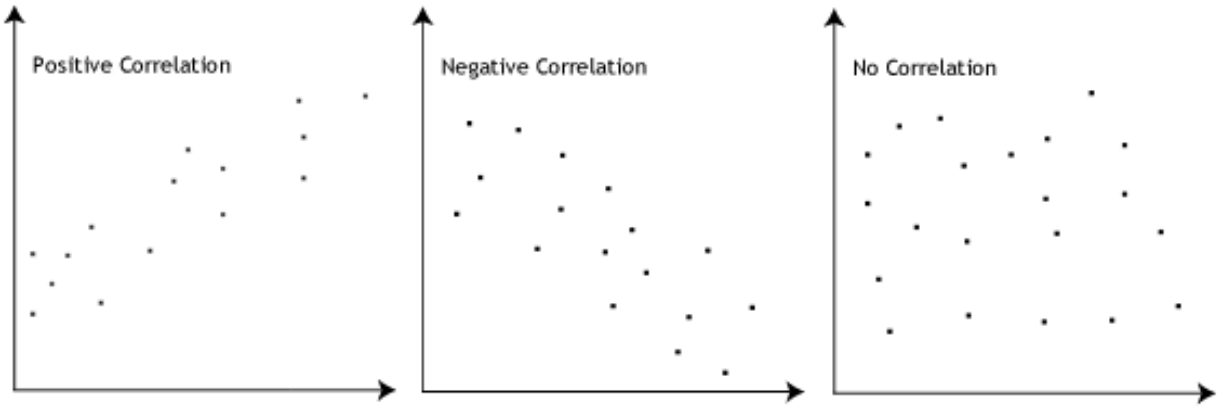
All the important features in the dataset must be visualised before implementing any machine learning algorithm on them which help to make a good fit model.

3. What is Pearson's R?

Ans: Pearson correlation coefficient (PCC), also referred to as Pearson's r or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations i.e. its normalised measurement of the covariance, so that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Here,

- r = correlation coefficient
- xi=values of the x-variable in a sample
- x=mean of the values of the x-variable
- yi=values of the y-variable in a sample
- y=mean of the values of the y-variable

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

Ans: It is a part of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Mostly data set contains numeric features which is highly varying in values, units and range. It also helps in speeding up the calculations in an algorithm.

If scaling is not done then algorithm only takes size in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of size.

Scaling affects the coefficients and not other parameters like F-statistic, p-values or R-squared.

Normalization means rescales the values into a range of [0 - min,1-max]. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

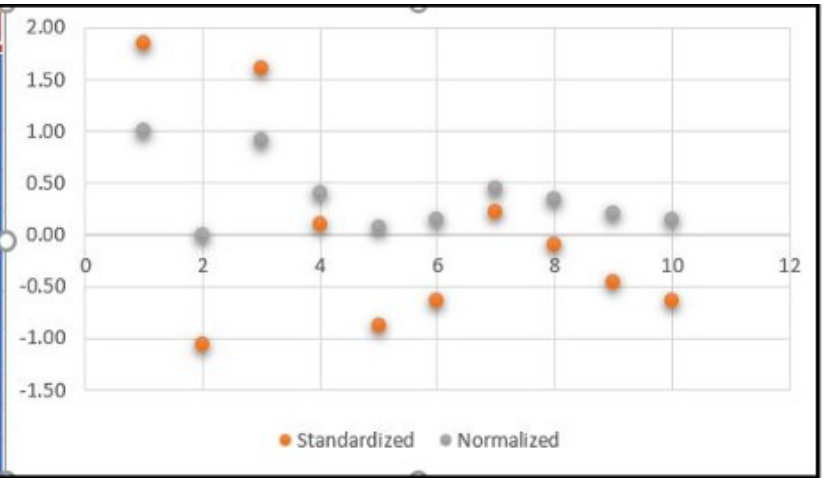
Standardization means rescales data to have a mean of [0-mean and standard deviation of klearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

S.NO.	Normalisation	Standardisation
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.
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.

Below example of Standardized and Normalized scaling

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

Ans: Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car etc.

Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable.

Multicollinearity can be detected via various methods. Most common – **VIF (Variable Inflation Factors)**.

" VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. "

VIF is calculated as below:

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

If there is perfect correlation, then VIF = infinity. This shows that there is a perfect correlation between two independent variables.

In perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

In order to resolve this, we need to drop one of the variables from the dataset which is leading this perfect multicollinearity.

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.

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. E.g., the median is a quantile where 50% of the data fall below that point and 50% lie above it.
The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help assess if a dataset 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.

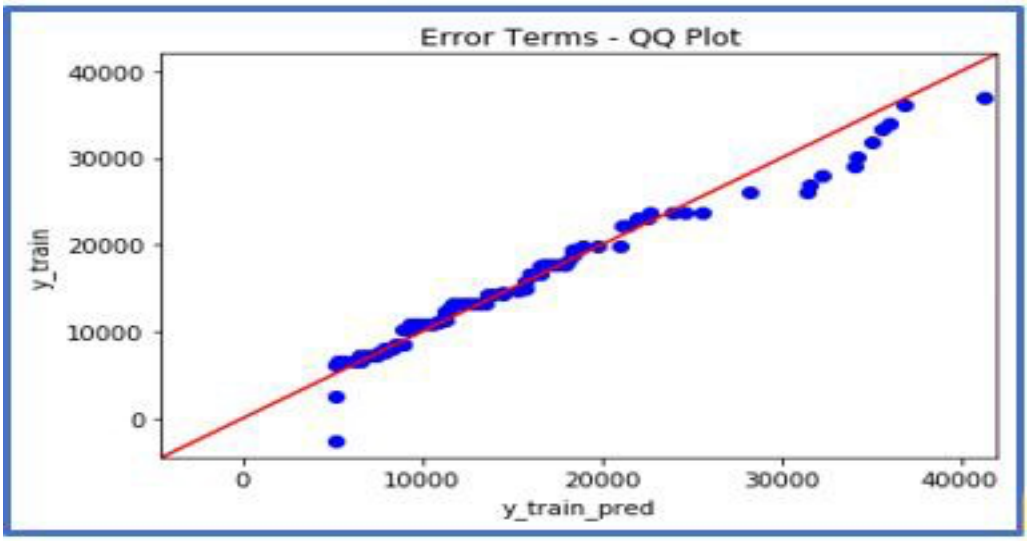
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

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.

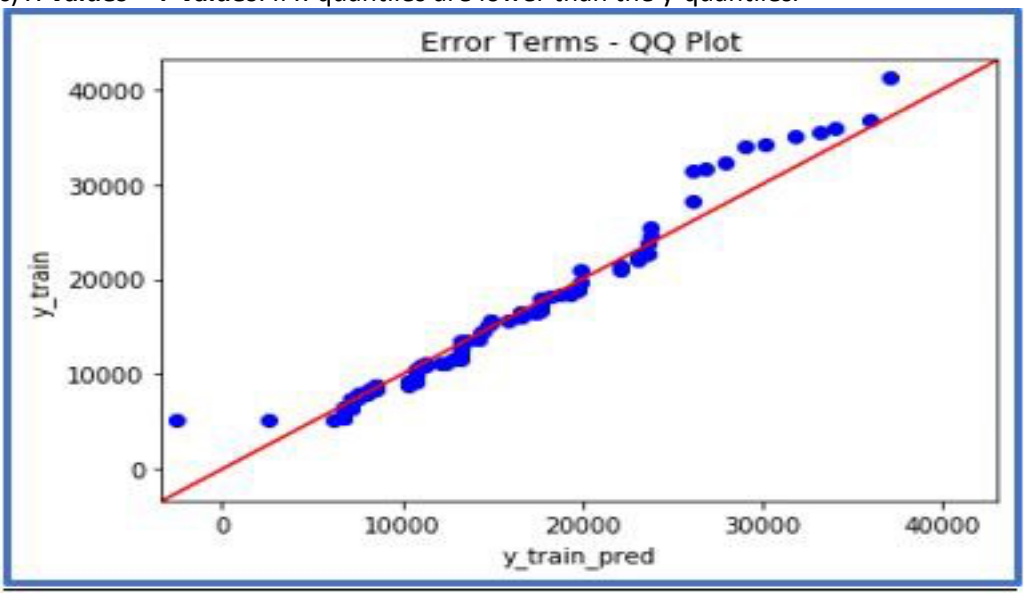
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

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

Python:
statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two data sets.