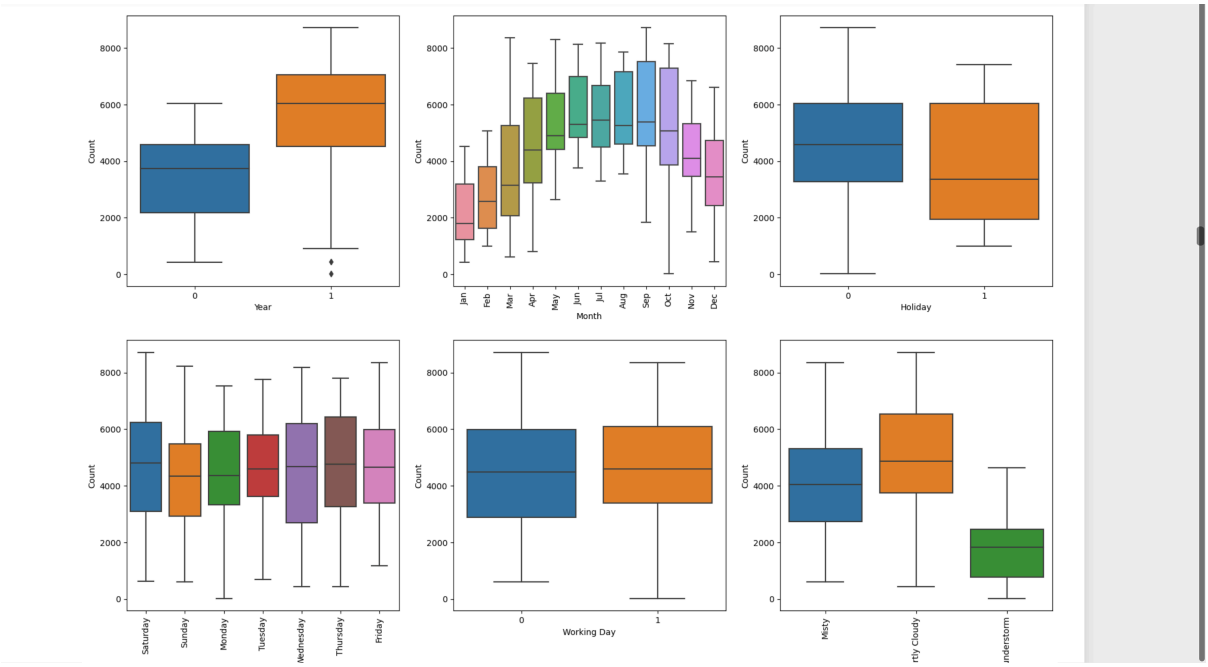
Assignment-based Subjective Questions

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

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2889	0.027	10.570	0.000	0.235	0.343
yr	0.2272	0.008	28.426	0.000	0.212	0.243
workingday	0.0510	0.010	4.864	0.000	0.030	0.072
temp	0.4272	0.028	15.001	0.000	0.371	0.483
hum	−0.1683	0.021	−7.880	0.000	−0.210	−0.126
windspeed	−0.1162	0.020	−5.783	0.000	−0.156	−0.077
spring	−0.1082	0.015	−7.032	0.000	−0.138	−0.078
winter	0.0612	0.013	4.894	0.000	0.037	0.086
Dec	−0.0443	0.015	−2.896	0.004	−0.074	−0.014
Jan	−0.0456	0.017	−2.624	0.009	−0.080	−0.011
Sep	0.0542	0.015	3.577	0.000	0.024	0.084
Saturday	0.0664	0.015	4.577	0.000	0.038	0.095
Light Rain/Thunderstorm	−0.1825	0.026	−7.024	0.000	−0.234	−0.131



- Month: Months August to September has highest demand
- Year: The demand for bikes has increased in the year 2019
- Holiday: The demand is in range of (2000-6000) in holidays compared to non-holidays whose demand is in the range of (3500+-6000)
- WeekDay: There is High demand on Thursdays (2500+-6000+) on Thursdays and the higher range is almost in the same range during almost all days except Sundays.
- WorkingDay: There is no major difference in demand between Working day and non-working day.
- Weathersit: The most favourable weather is when it's clear/partly cloudy whereas the least favourable situation is when it has Light Rains/Thunderstorms
- Season: The demand is high during fall season, whereas the demand is the least during spring season.

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

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

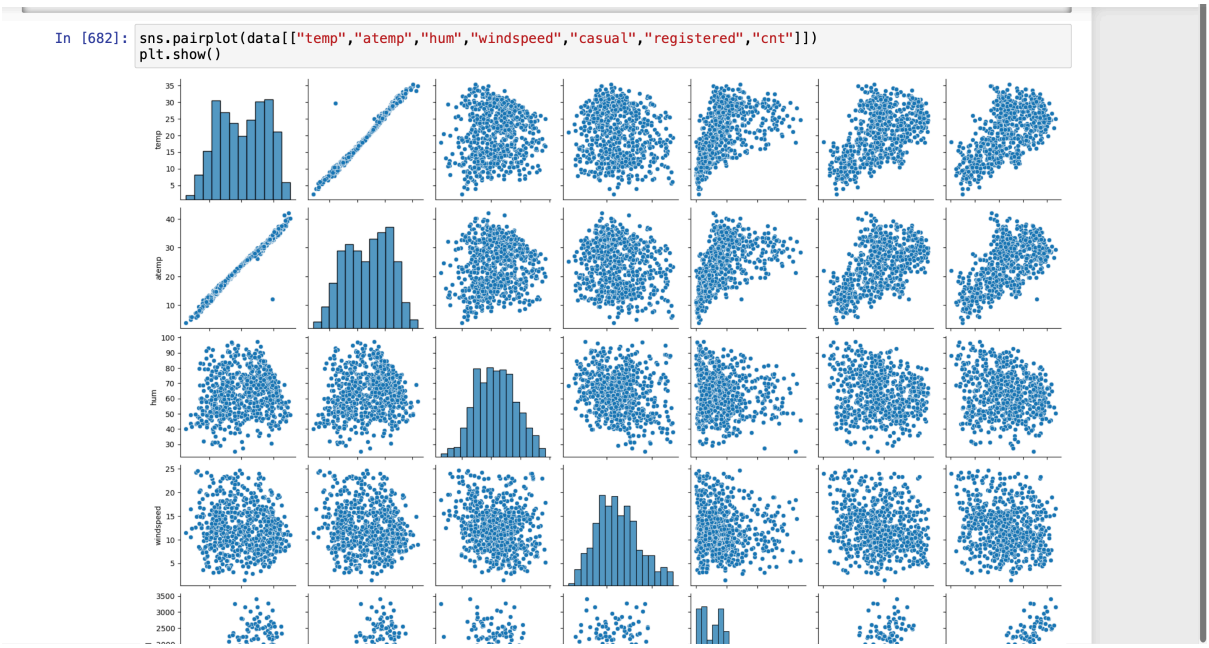
For ex: If there's a variable called Working status and it has three options Full-day, Half-day and Leave, without dropping a column the dummy variables looks like :

Full-day	Half-Day	Leave
1	0	0
0	1	0
0	0	1

If we use drop_first=True, it drops a redundant column and Leave can be represented as 0-0

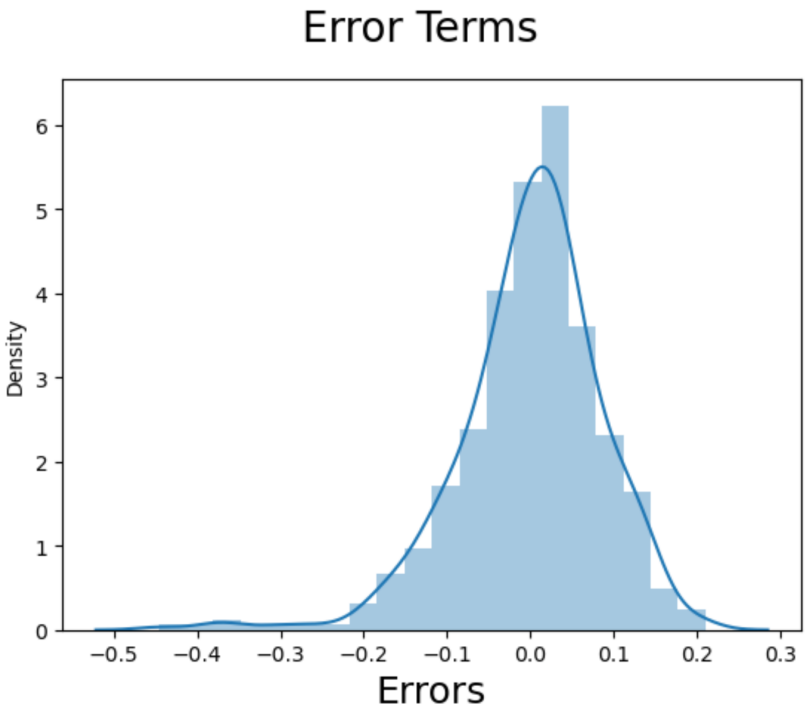
Full-day	Half-Day
1	0
0	1
0	0

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   By looking at the pair plot, temp has highest correlation.
   atemp and temp are highly correlated to each other, hence ignoring atemp.

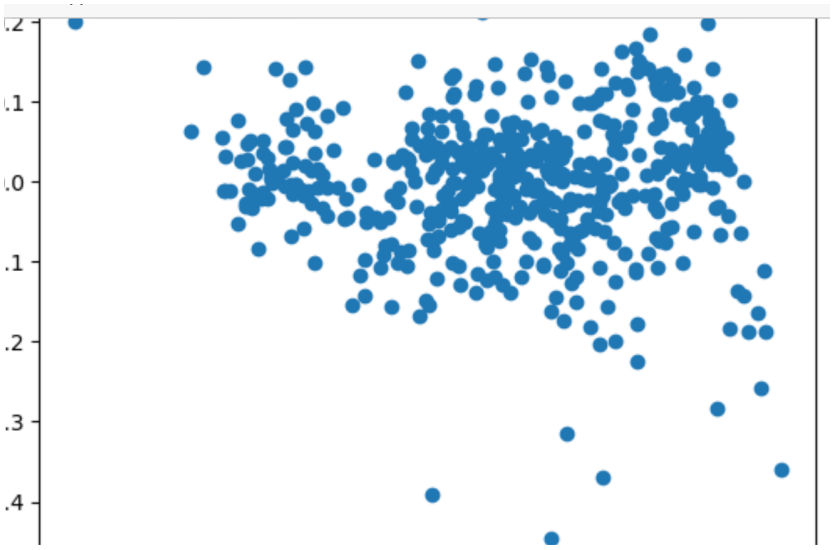


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

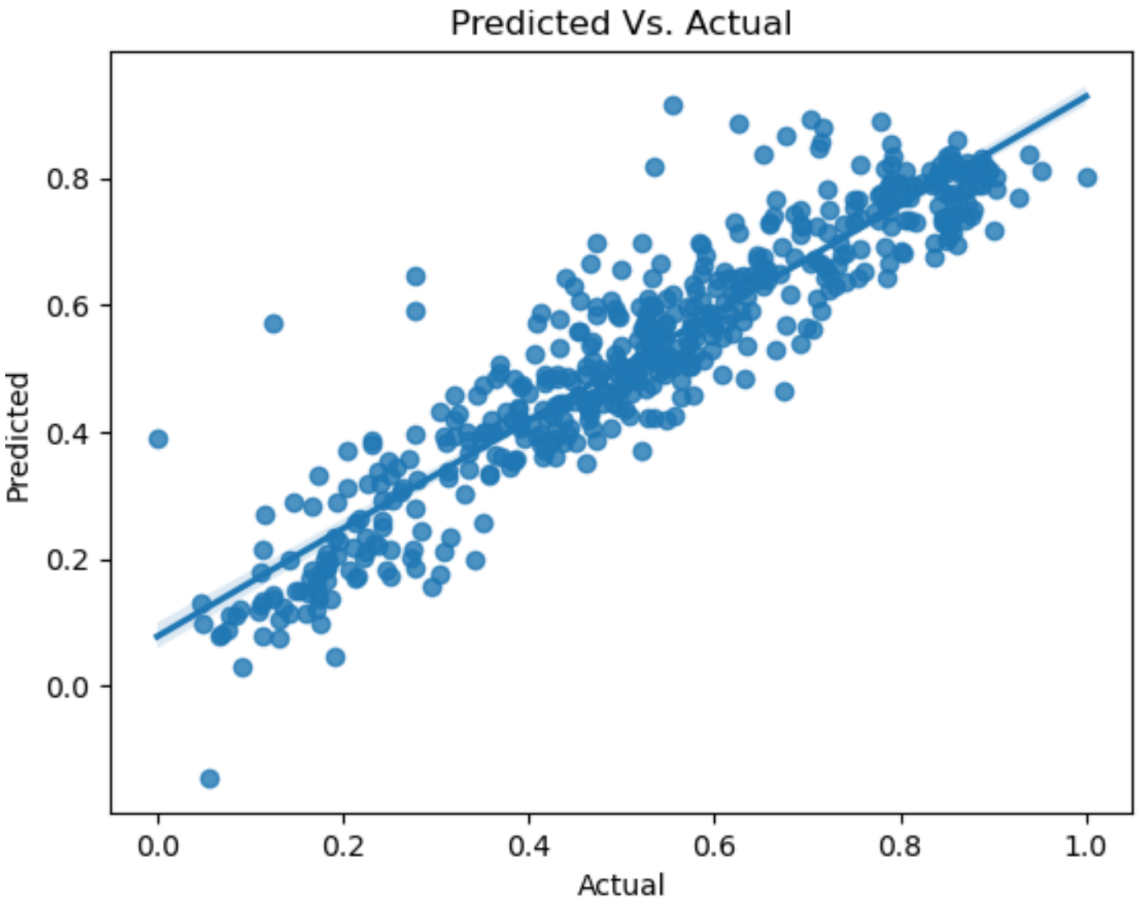
- **Residual Analysis:**
  The Error terms are normally distributed with mean at Zero.



- Error terms are independent of each other since we don't see any specific pattern.



- There is linear relationship between dependant and independent variables, Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.



All feature variables are having P value >0.05
R-squared: 0.850
Adj. R-squared: 0.847

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

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2889	0.027	10.570	0.000	0.235	0.343
yr	0.2272	0.008	28.426	0.000	0.212	0.243
workingday	0.0510	0.010	4.864	0.000	0.030	0.072
temp	0.4272	0.028	15.001	0.000	0.371	0.483
hum	-0.1683	0.021	-7.880	0.000	-0.210	-0.126
windspeed	-0.1162	0.020	-5.783	0.000	-0.156	-0.077
spring	-0.1082	0.015	-7.032	0.000	-0.138	-0.078
winter	0.0612	0.013	4.894	0.000	0.037	0.086
Dec	-0.0443	0.015	-2.896	0.004	-0.074	-0.014
Jan	-0.0456	0.017	-2.624	0.009	-0.080	-0.011
Sep	0.0542	0.015	3.577	0.000	0.024	0.084
Saturday	0.0664	0.015	4.577	0.000	0.038	0.095
Light Rain/Thunderstorm	-0.1825	0.026	-7.024	0.000	-0.234	-0.131

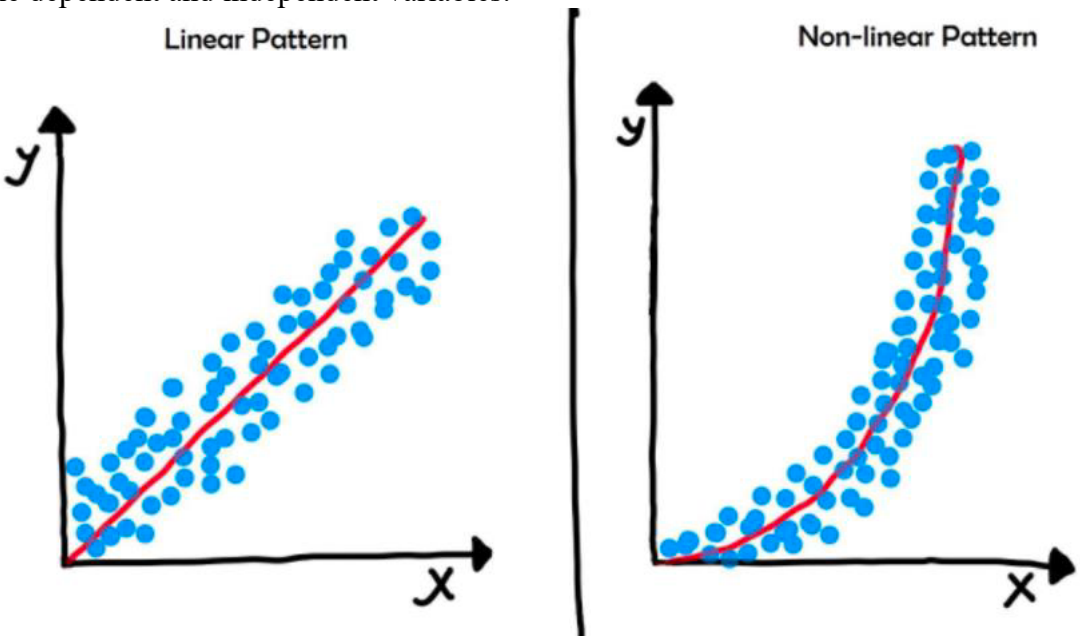
Top 3 features significantly contributing towards demand of shared bikes are:

1. Temp (coef 0.4272)
2. Yr (coef 0.2272)
3. Day – Saturday

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression finds the best linear relationship between the independent and dependent variables. It is a method of finding the best straight-line fitting to the given data. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.
The assumptions of linear regression are:
a. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables.



b. Assumptions about the residuals:
1), Normality Distribution : It is assumed that error terms are normally distributed.
2) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
3) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.
4) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

c. Assumptions about the estimators:
1) The independent variables are measured without error.
2) The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

There are two types of Linear Regression Algorithm

- Simple Linear Regression
  - Single independent variable is used to predict the value of a dependant variable.
  - $Y = \beta_0 + \beta_1 X$
- Multiple Linear Regression
  - More than one independent variable is used to predict the value of a dependant variable.
  - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$
  - β_0 is intercept
  - $\beta_1, \beta_2, \ldots, \beta_p$ are slopes

The main goal while using linear regression is to find best fit line i.e error between predicted and actual values should be minimal. This is found by minimising the Residual Sum of Squares: $\sum(y_i - \beta_0 - \beta_1 x_i)^2$
The Strength of linear regression depends on:

- $R^2 = (1\text{-}RSS/TSS)$
- Residual Standard Error

2. Explain the Anscombe's quartet in detail.

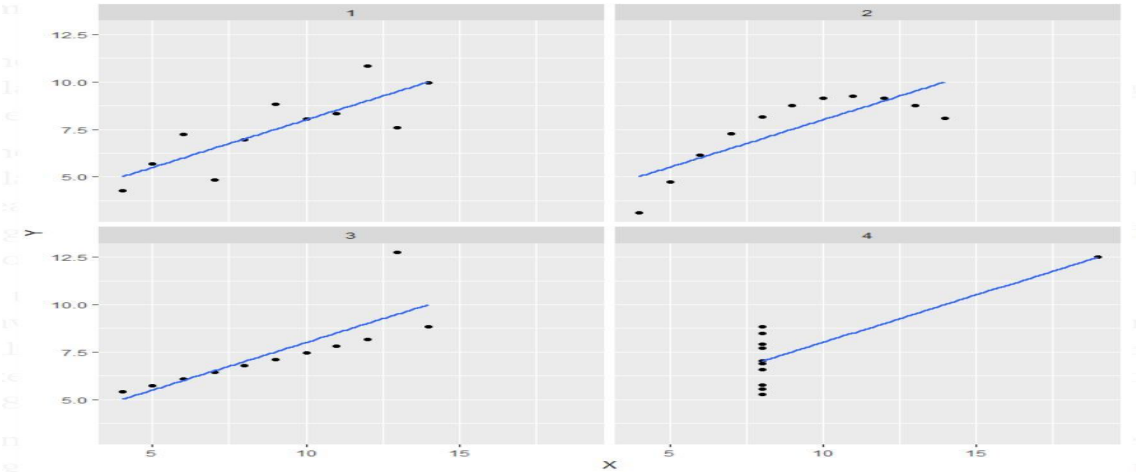
It's a group of 4 data sets that are nearly identical when used simple descriptive statistics, but they have different distributions and they also appear different when the data sets are put in a graph. Which tells us to visualize the data before applying any algorithm to build models so that various anomalies present in the data such as outliers, diversity of data etc.. can be identified.

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

The above data sets are the 4 data-sets of 11 points.
When we use descriptive statistics, we can see that all the values of mean, standard deviation are similar.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



But when data is visualized, we can see that

- In the first plot the data has some sort of linear relationship
- In the second plot there is a non-linear relationship
- In the third plot there is an outlier(far away from the line)
- In the fourth plot shows one high leverage point(far right) is enough to produce a high correlation coefficient

3. What is Pearson's R?

Pearson's R also known as Pearson's correlation Coefficient is a measure of linear correlation between two sets of data. Pearson's R returns values between -1 and 1. Where

- Coefficient value 1 shows strong relationship
- Coefficient value -1 shows inverse relationship
- Coefficient value 0 shows no relationship

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

n: sample size

$x_i y_i$: Individual sample points indexed with i

$\overline{x} : \frac{1}{n}\sum_{i=1}^{n} x^i$ and similarly, for $\overline{y}$

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

Scaling is a method used to normalize the range of independent variables or features of data. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization:
Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as: x= x-min(x)/max(x)-min(x)

Standardization:
Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point.

Standardization has a disadvantage over normalization since it loses outliers.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

Few advantages:

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:

If two data sets —

1. come from populations with a common distribution

2. have common location and scale

3. have similar distributional shapes

4. have similar tail behavior

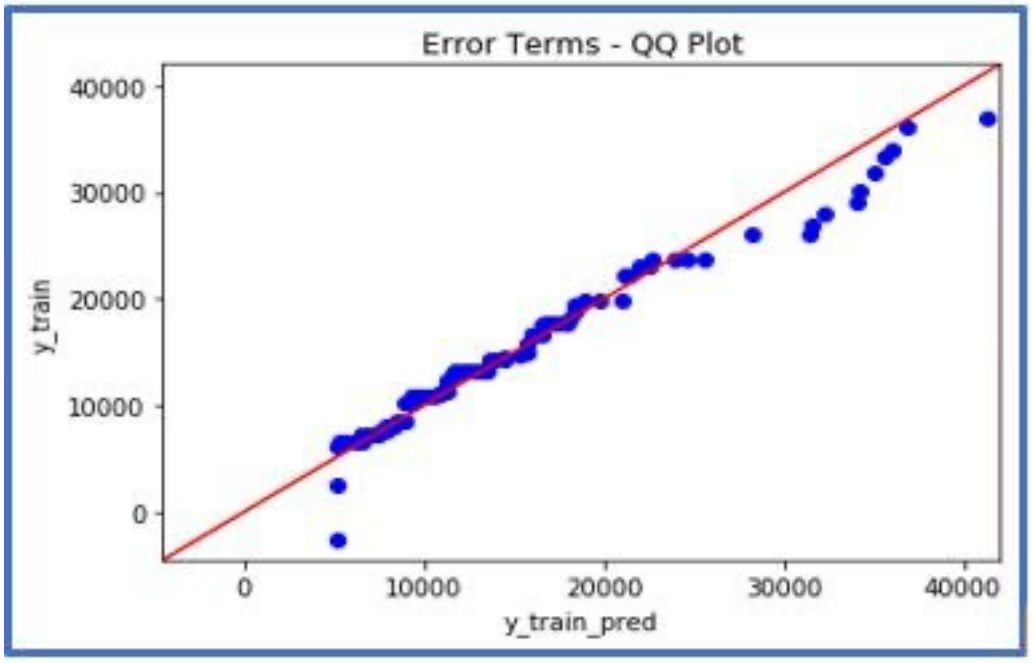
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

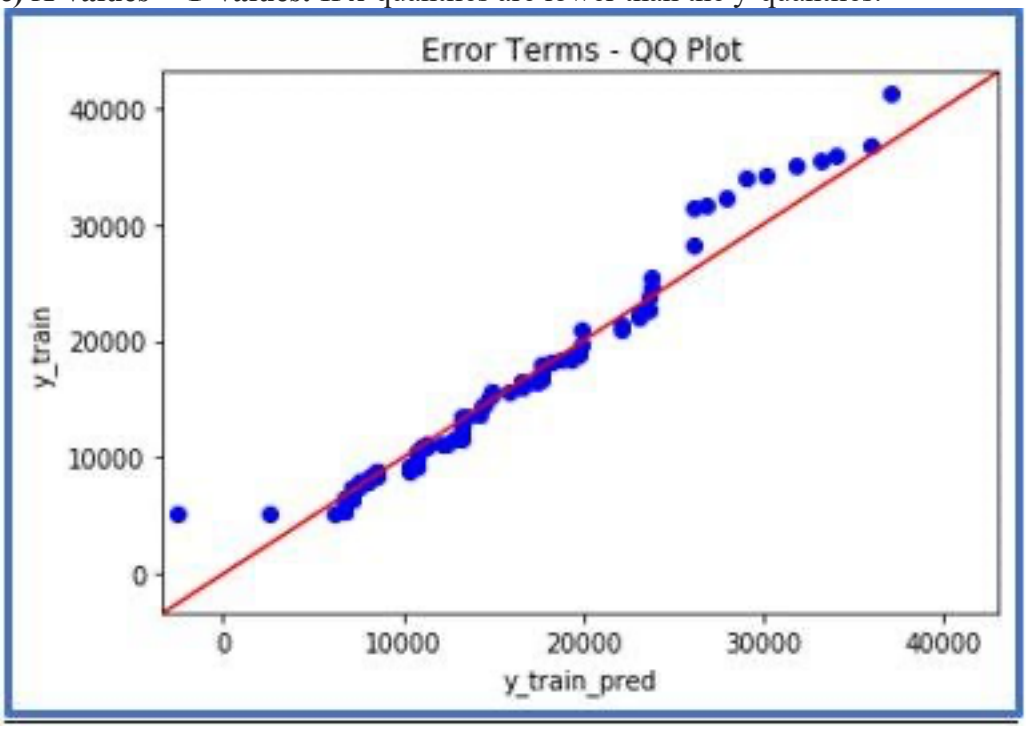
Below are the possible interpretations for two data sets.

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