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? (3 marks)
   Answer:
   The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2059	0.028	7.423	0.000	0.151	0.260
yr	0.2341	0.008	27.985	0.000	0.218	0.251
workingday	0.0219	0.009	2.460	0.014	0.004	0.039
temp	0.4495	0.032	13.968	0.000	0.386	0.513
windspeed	-0.1487	0.026	-5.775	0.000	-0.199	-0.098
spring	-0.0928	0.018	-5.047	0.000	-0.129	-0.057
winter	0.0771	0.015	5.149	0.000	0.048	0.107
weathersit_2	-0.0799	0.009	-8.988	0.000	-0.097	-0.062
weathersit_3	-0.2878	0.025	-11.495	0.000	-0.337	-0.239
month_3	0.0517	0.015	3.440	0.001	0.022	0.081
month_4	0.0430	0.019	2.264	0.024	0.006	0.080
month_5	0.0542	0.017	3.258	0.001	0.022	0.087
month_9	0.0800	0.016	5.001	0.000	0.049	0.111

   spring, winter falls under season category and have been dummy encoded. weathersit_2 and weathersit_3 falls under weathersit category and have been dummy encoded. Similarly, month variables fall under mnth category and have been dummy encoded. We can infer from above image that these variables are statistically significant and explain the variance in model very well.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 marks)
   Answer:
   Reason 1:
   To avoid dummy variable trap. Dummy variable trap might lead to multicollinearity issue among dummy variables. This may lead to violation of assumptions of linear

regression. So, if we have k levels where k >=3, we only use k-1 levels while dummy variable encoding. The dropped level gets handled by intercept as a base case.

Reason 2:

Suppose there is category of colours like Red, Green and Blue. This category belongs to nominal categorical variable. There is no order or relation among Red, Green, Blue. We can't simply label encode them as 1, 2, 3 because this will confuse our model which might lead to order-based bias like Red < Green < Blue. To avoid this, we dummy encode cases like this. Also, model can't understand string/text data therefore it is necessary to convert them into numbers.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer:

   So, before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.945. ***But we are not using casual and registered in our pre-processed training data for model training.*** *casual + registered = cnt. This might leak out the crucial information and model might get overfit.*

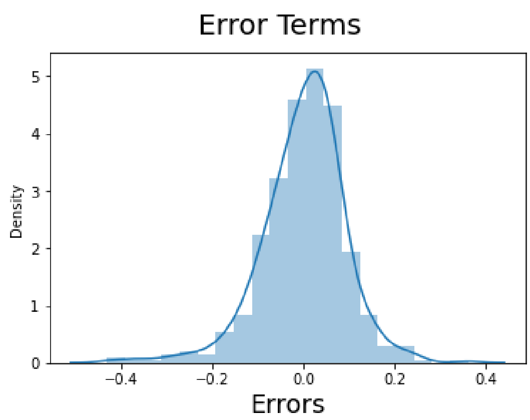
   So, excluding these two variables ***atemp*** *is having highest correlation with target variable* ***cnt*** *which is followed by* ***temp****.*
   As per the correlation heatmap, correlation coefficient between ***atemp*** *and* ***cnt*** *is* ***0.631****. And correlation coefficient between* ***temp*** *and* ***cnt*** *is* ***0.627****.*

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer:

   To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

   - ***Residual Analysis:***
     We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like:



The residuals are following the normally distribution with a mean 0. All good!

- ***Linear relationship between predictor variables and target variable:***
  *This is happening because all the predictor variables are statistically significant (p-values are less than **0.05**). Also, R-Squared value on training set is **0.832** and adjusted R-Squared value on training set is **0.828**. This means that variance in data is being explained by all these predictor variables.*

- ***Error terms are independent of each other:***
  Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

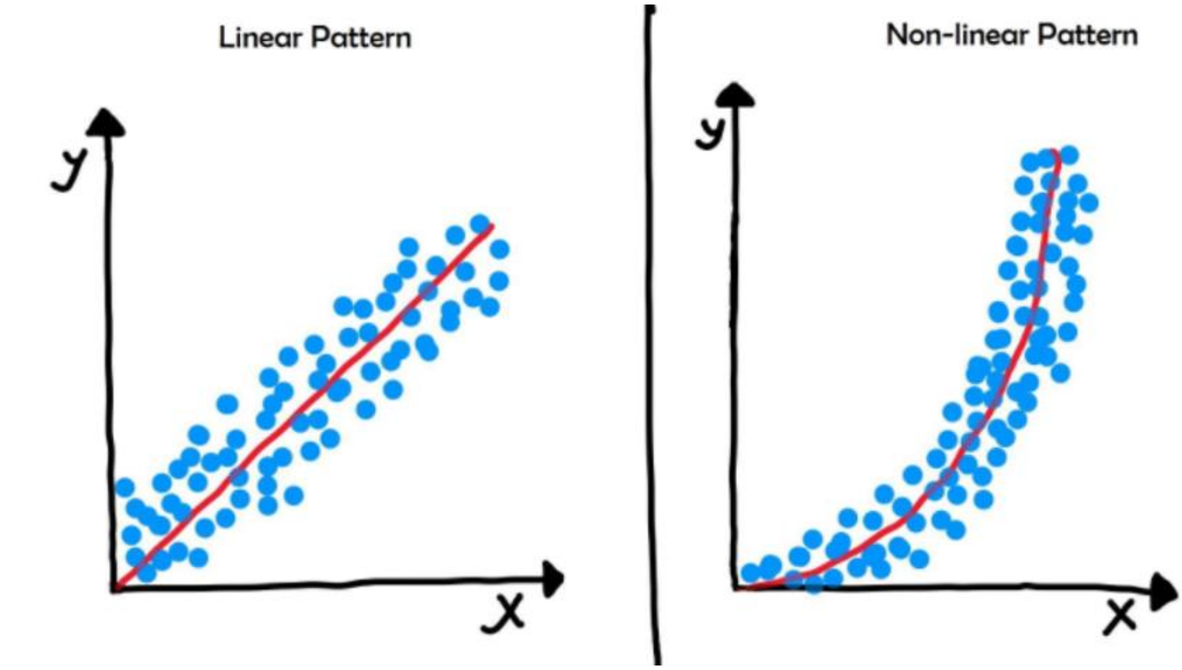
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   Top 3 features significantly contributing towards demand of shared bikes are:
   1) ***temp** (coef: **0.4495**)*
   2) ***yr** (coef: **0.2341**)*
   3) ***month_9** (coef: **0.0800**)*

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Answer:
   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.

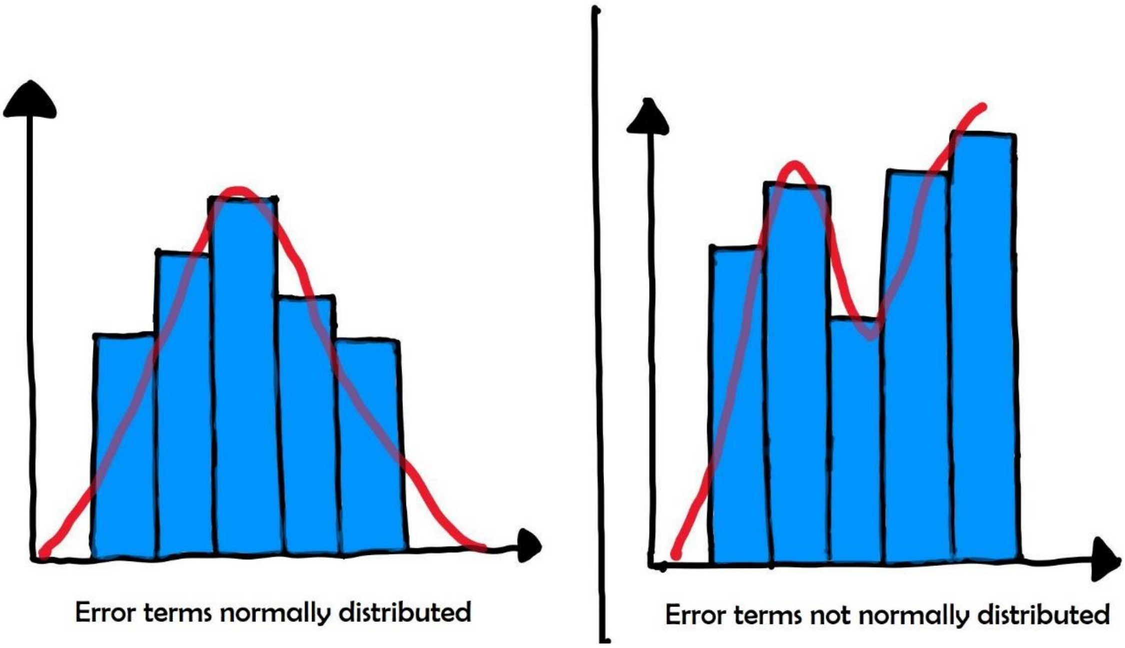
   The assumptions of linear regression are:

   a. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables.

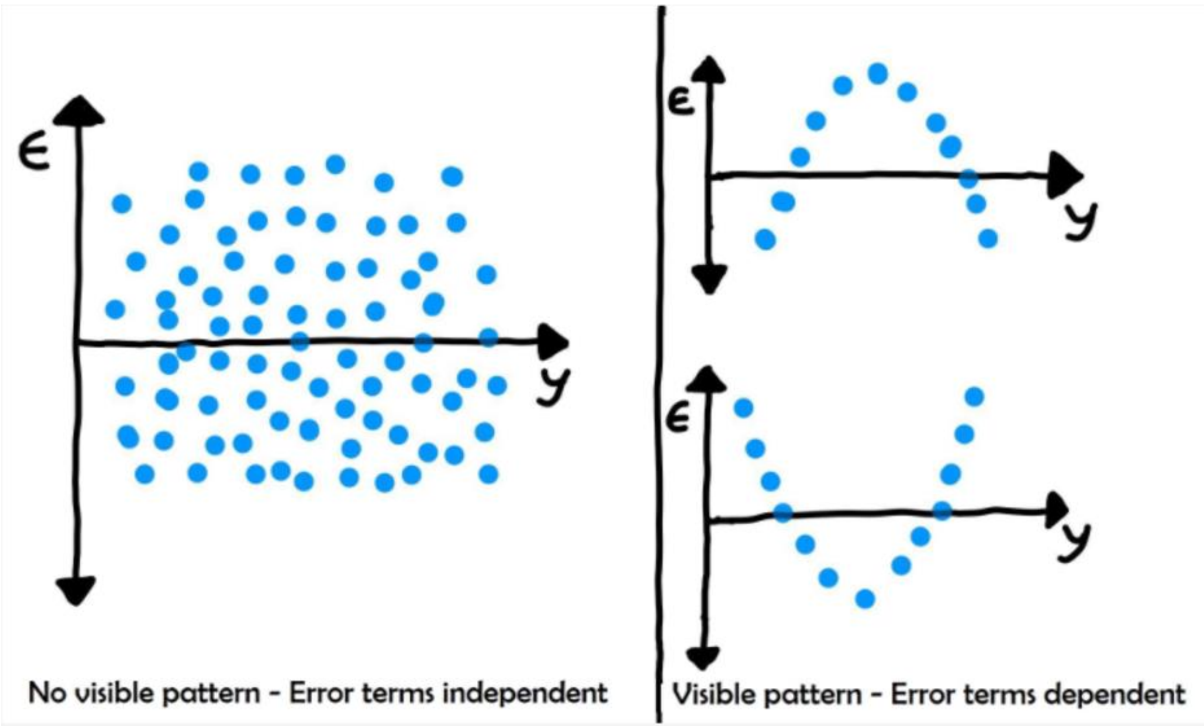


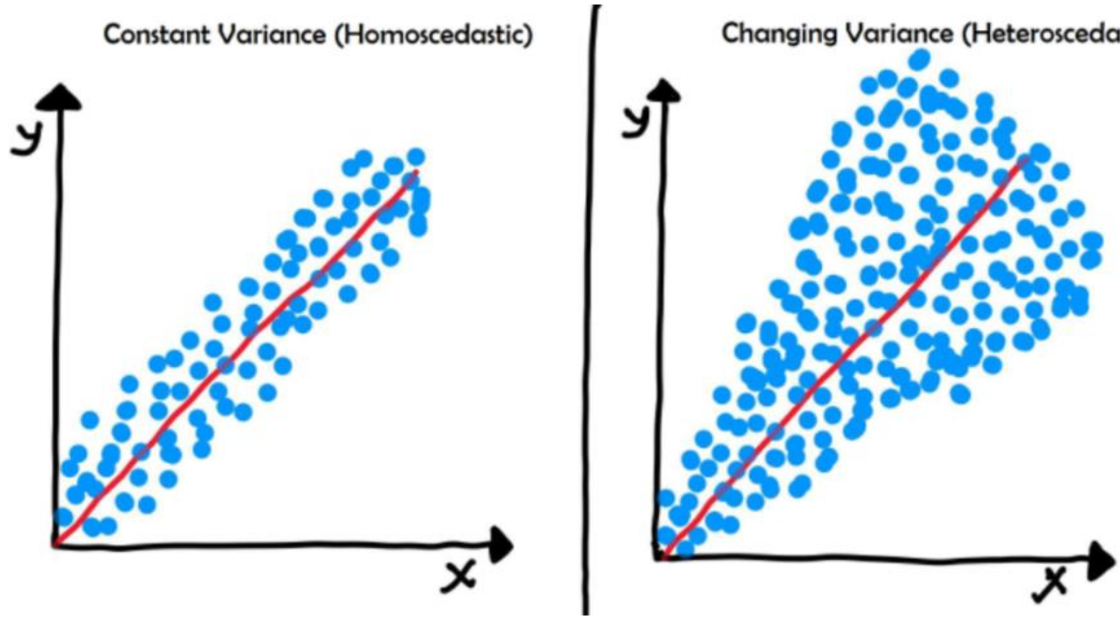
b. Assumptions about the residuals:

1) *Normality assumption: It is assumed that the error terms, ε(i), 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.*

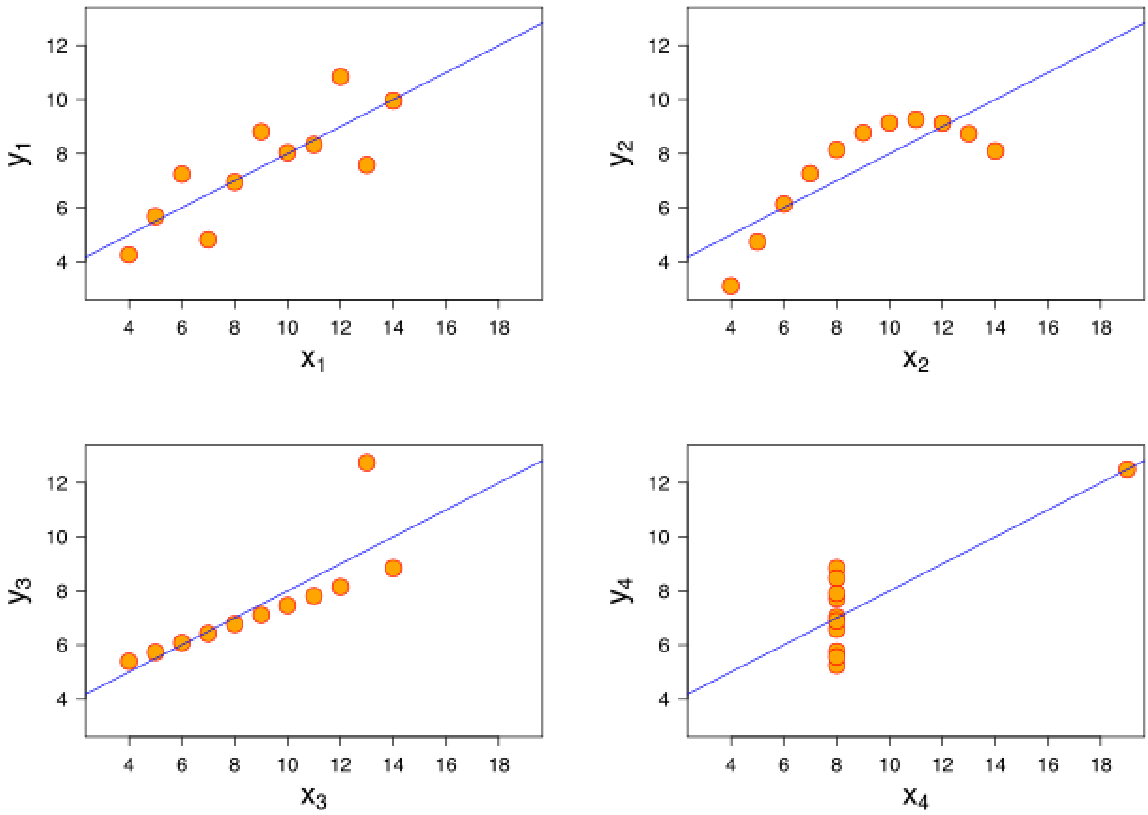




2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

1) *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.*
2) *The second graph (top right) is not distributed normally; 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.*
3) *In the third graph (bottom left), the distribution 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.*
4) *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.*

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

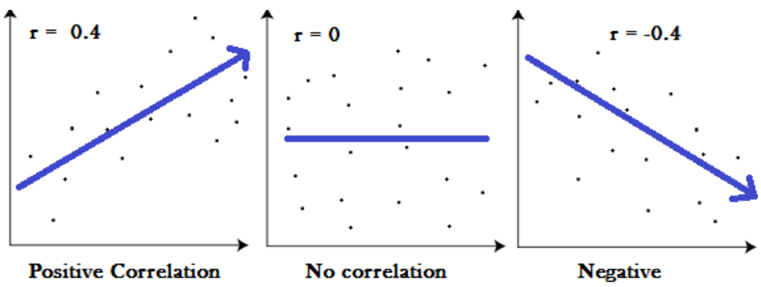
Anscombe's quartet

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

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R or correlation coefficient is a measure of 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 normalised 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 relationship or correlation.



1) *A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.*
2) *A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.*
3) *Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.*

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:
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:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

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 by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   Answer:

   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. (3 marks)

   Answer:

   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. For example, 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.

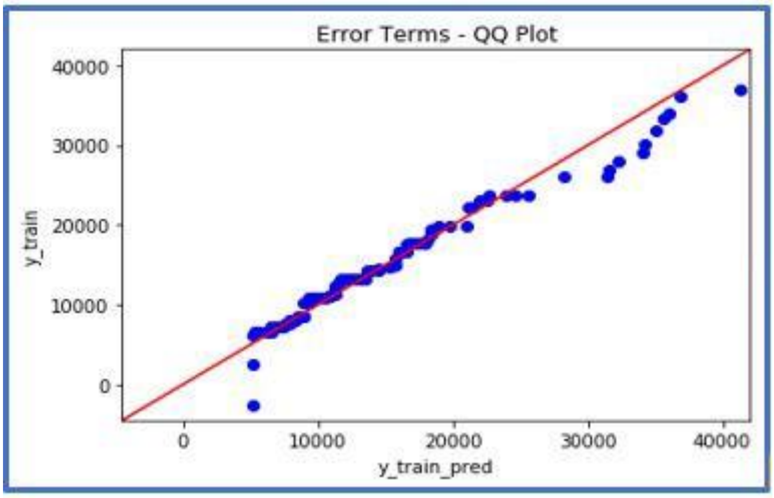
   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Interpretation:

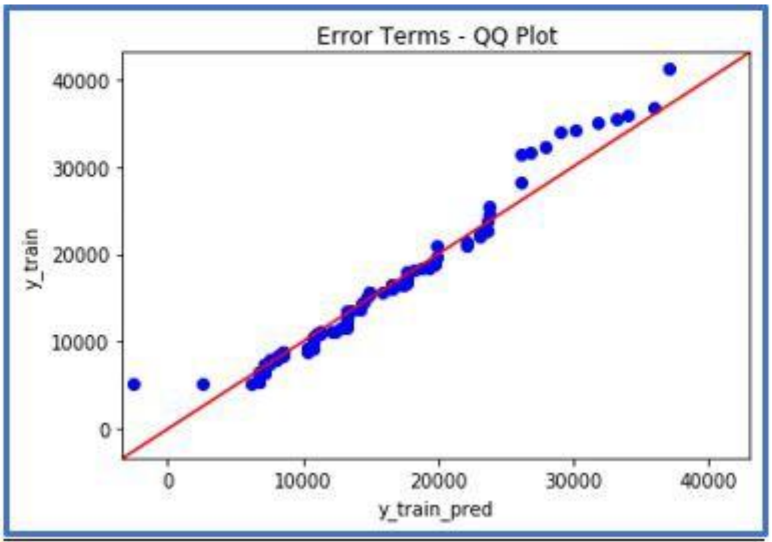
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

1) *Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*
2) *Y-values < X-values: If y-quantiles are lower than the x-quantiles.*



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



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

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