Subjective Questions Solutions

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 : Final Model Equation is :

cnt = - 0.4458 + 1.0456 * yr - 0.3696 * holiday + 0.4732 * temp - 0.1139 * windspeed - 0.3278 * spring

+ 0.2011 * summer + 0.3577 * winter - 1.2234 * LightSnowAndRain - 0.3519 * MistAndCloudy + 0.2911 * September

From , The equation we can see categorical values like year , spring ,LightSnowAndRain , MistAndCloudy and winter and others plays an important role in deciding the target variable cnt . Their Coefficients are also pretty good.

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

Ans : **drop_first=True is** important because it helps in reducing the extra column created from dummies and so helps in reducing the corelations created among dummy variables .

In the dataset given , we have created dummy variable for season having values as spring , summer , fall and winter . In that we have called

seasons = pd.get_dummies(df['season'], drop_first = True)

By this it dropped fall column as from other column value we can know whether it’s a fall season or not i.e when all spring =0 , summer =0 and winter = 0 then it means fall = 1 otherwise 0.

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

with the target variable?

Ans : After the final pre-processing of data (i.e when instant , casual and registered columns got deleted) , temp and atemp have the highest corelation with the target variable .

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

training set?

Ans : Did the residual analysis and verified whether error terms or residuals are independent and normally distributed.

Then calculated the r2 score of test data and found it approx. range of training r2 score.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Ans : : Final Model Equation is :

cnt = - 0.4458 + 1.0456 * yr - 0.3696 * holiday + 0.4732 * temp - 0.1139 * windspeed - 0.3278 * spring + 0.2011 * summer + 0.3577 * winter - 1.2234 * LightSnowAndRain - 0.3519 * MistAndCloudy + 0.2911 * September

Most Important featur**es are :**

1. Yr
2. LightSnowAndRain
3. Spring

This we decided based on RFE results if features selected = 3 only

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.675
Model:	OLS	**Adj. R-squared:**	0.673
Method:	Least Squares	**F-statistic:**	350.4
Date:	Wed, 15 Nov 2023	**Prob (F-statistic):**	4.61e-123
Time:	13:11:34	**Log-Likelihood:**	-437.03
No. Observations:	510	**AIC:**	882.1
Df Residuals:	506	**BIC:**	899.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2249	0.040	-5.649	0.000	-0.303	-0.147
yr	1.1177	0.051	21.993	0.000	1.018	1.218
spring	-1.2547	0.059	-21.201	0.000	-1.371	-1.138
LightSnowAndRain	-1.2821	0.150	-8.523	0.000	-1.578	-0.987

Omnibus:	48.416	**Durbin-Watson:**	1.892
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	76.648
Skew:	-0.642	**Prob(JB):**	2.27e-17
Kurtosis:	4.399	**Cond. No.**	6.99

However, based on the model important features are :

1. LightSnowAndRain
2. Yr
3. temp

General Subjective Questions :

1. Explain the linear regression algorithm in detail.

Ans : Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

Linear Regression is of two types: Simple and Multiple.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$Y = b0 + b1*x$

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

$Y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 \ldots\ldots\ldots + bn* xn$

Mathematical Approach for Linear regression :

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

Assumptions of Linear Regression

a. Linearity: dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

b. Normality: The X and Y variables should be normally distributed. Histograms, KDE plots, Q-Q plots can be used to check the Normality assumption.

c. Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot.

d. Independence/No Multicollinearity: The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score.

e. The error terms should be normally distributed.

Evaluation Metrics for Regression Analysis

1. R squared or Coefficient of Determination: The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model.
   Adjusted R2 is always lower than R2.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

3. Mean Squared Error (MSE): Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

4. Root Mean Squared Error (RMSE): It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician

Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets

is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

Each graph plot shows the different behavior irrespective of statistical analysis.

Apply the statistical formula on the above

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets

Average Value of x = 9

Average Value of y = 7.50

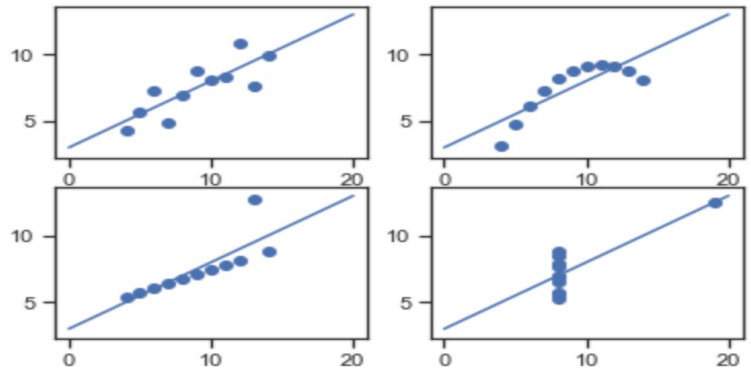
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

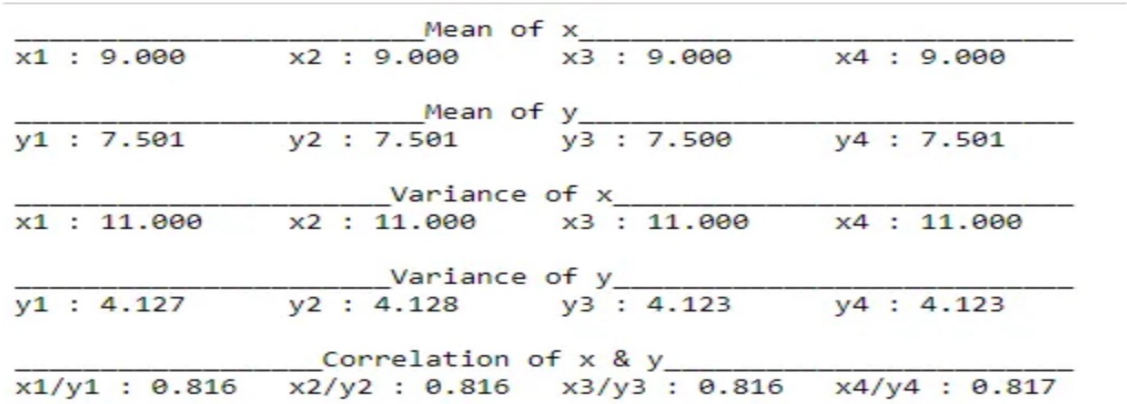
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour .



Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.



Output of Python code

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

"Visualization gives you answers to questions you didn't know you had." — Ben Schneiderman

3. What is Pearson's R?

The Pearson correlation coefficient is a descriptive statistic meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

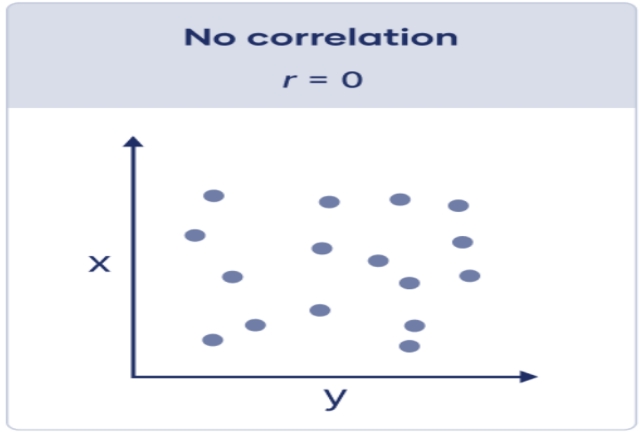
The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

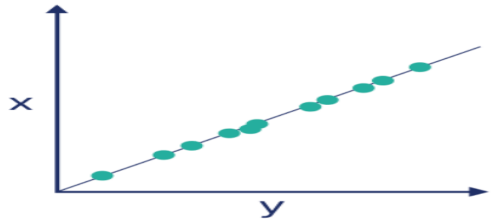
The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

When r is 1 or –1, all the points fall exactly on the line of best fit:

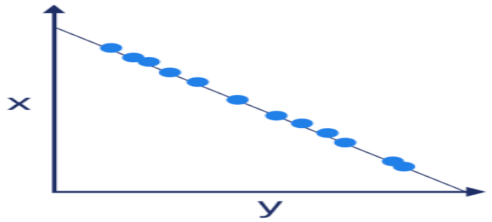
When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



Perfect positive correlation
$r = 1$

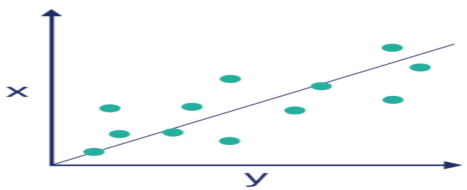


Perfect negative correlation
$r = -1$

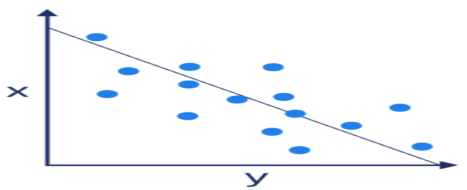


When r is between 0 and .3 or between 0 and −.3, the points are far from the line of best fit:

Weak positive correlation
$.3 > r > 0$

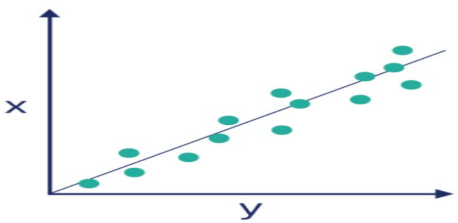


Weak negative correlation
$0 > r > -.3$

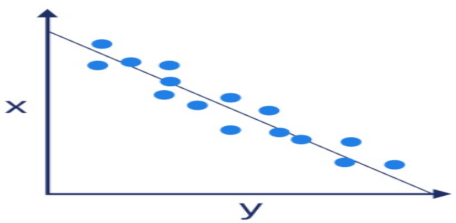


When r is greater than .5 or less than −.5, the points are close to the line of best fit:

Strong positive correlation
$r > .5$



Strong negative correlation
$r < -.5$



The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

•Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

•The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

•The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

•The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

•Formula for calculating the Pearson correlation coefficient (r):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

and standardized scaling?

Ans : Feature Scaling is one of the important pre-processing that is required for standardizing/normalization of the input data. When the range of values are very distinct in each column, we need to scale them to the common level. The values are brought to common level and then we can apply further machine learning algorithm to the input data.

Suppose we have different features, in which one of the feature might have data represented in Kilometre, another column might have data represented in Metre and the last column might have data representation in centimetre. Before applying the algorithm to the data, we need to first bring them to the common scale which might be “Metre”, “Kilometre” or “Centimetre” to have effective analysis and prediction.

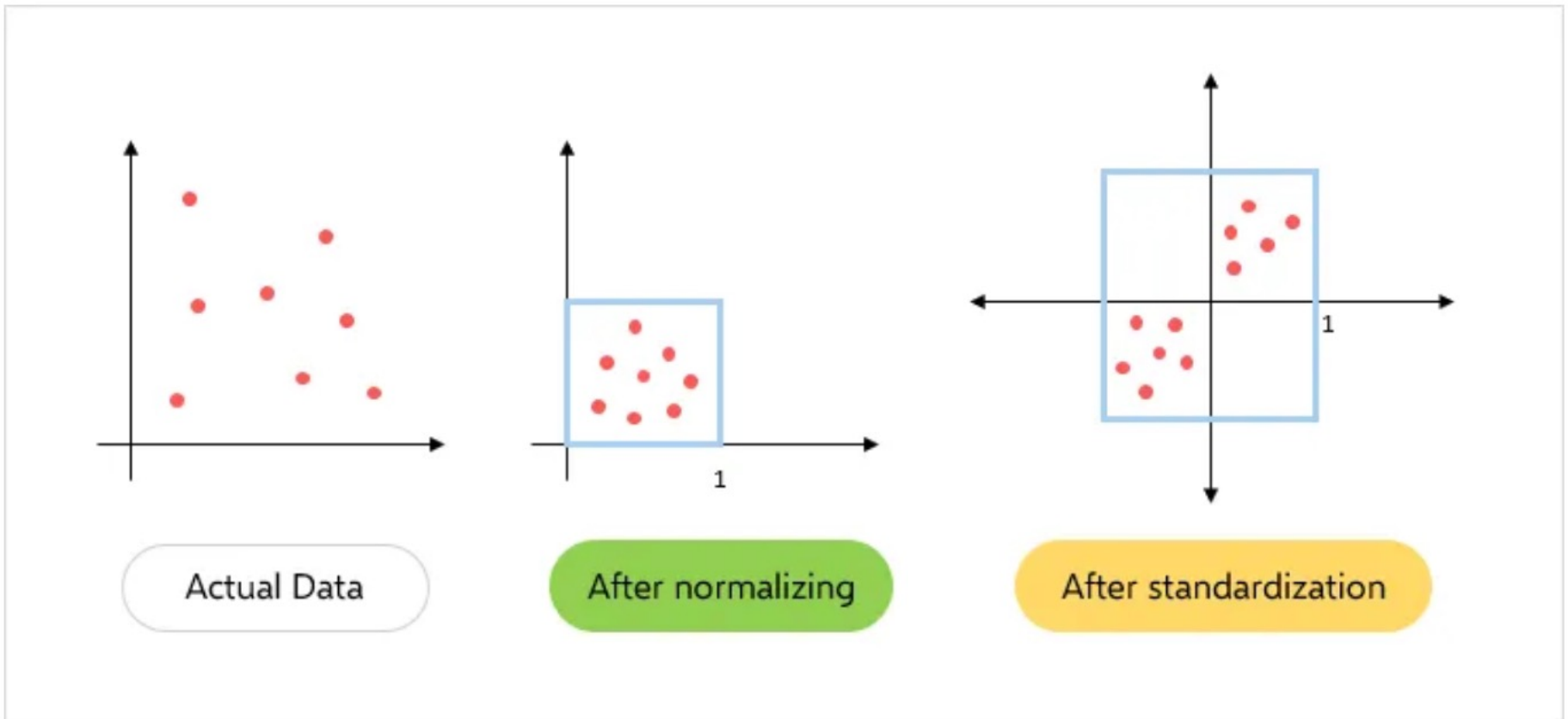
- Standarization
- Standarization (or Z-score normalization) rescaling of the features so that they have the properties of a standard normal distribution with μ=0 and σ=1, where μ is the mean (average) and σ is the standard deviation from the mean.
- Standardizing the features so that they are centered(μ) around 0 with a standard deviation(σ) of 1 is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms.

Python impl :
from sklearn.preprocessing import StandardScalerscale StandardScaler().fit(data)
scaled_data = scale.transform(data)

- Normalization
- Normalization is a technique often applied as a part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the range of value or losing information
- An alternative approach to Z-score normalization (or standardization) is the so-called Min-Max scaling (Normalization).
- The data is scaled to a fixed range usually 0 to 1.
- Python Impl :
- from sklearn.preprocessing import MinMaxScaler
  norm = MinMaxScaler().fit(data)
  transformed_data = norm.transform(data)
- A Min-Max scaling is typically done via the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Difference between Standardization and normalization



Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans :

- If there is perfect corelation then VIF is infinity .
- VIF = 1/(1-R2)

So, When R2 is 1 then VIF is infinity

This shows a perfect correlation between two independent variables . To solve this , we need to drop one variable from the dataset which is causing this multicollinearity . An Infinite VIF indicates that the corresponding variable may be expressed exactly be a linear combinations of other variables .

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans :

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

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 —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

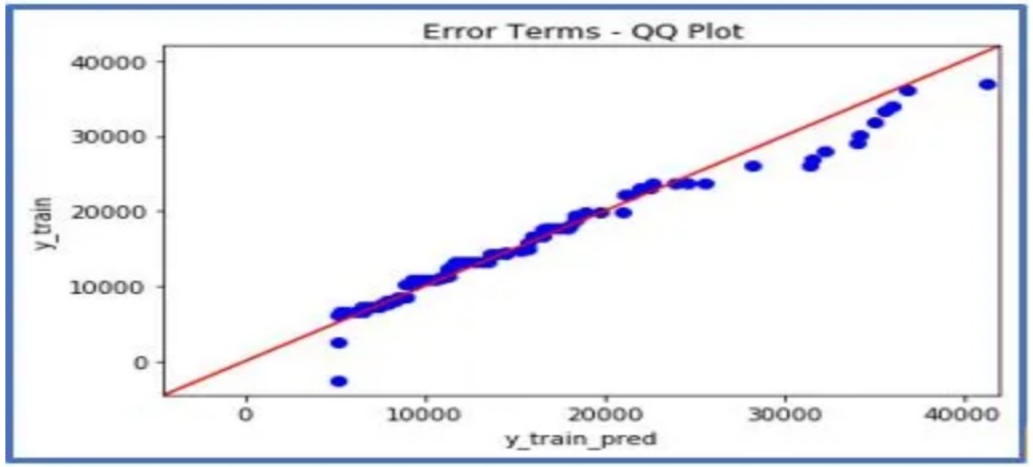
Interpretation:

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

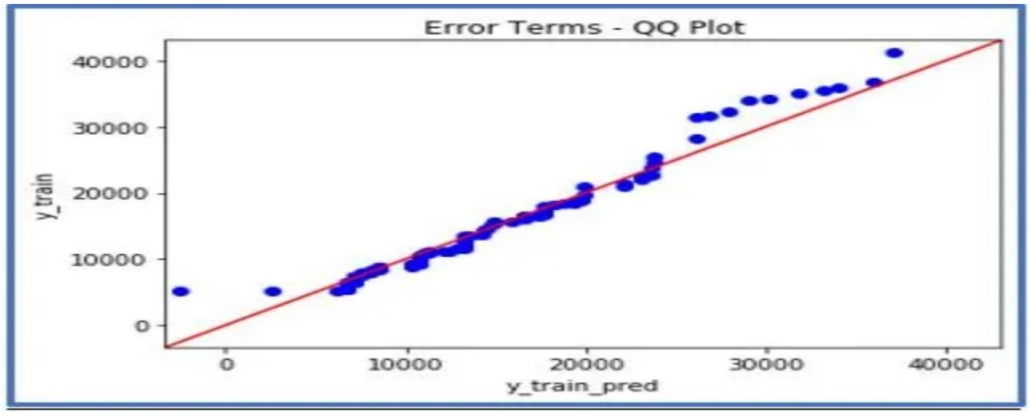
Below are the possible interpretations for two data sets.

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



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

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