Linear Regression Assignment

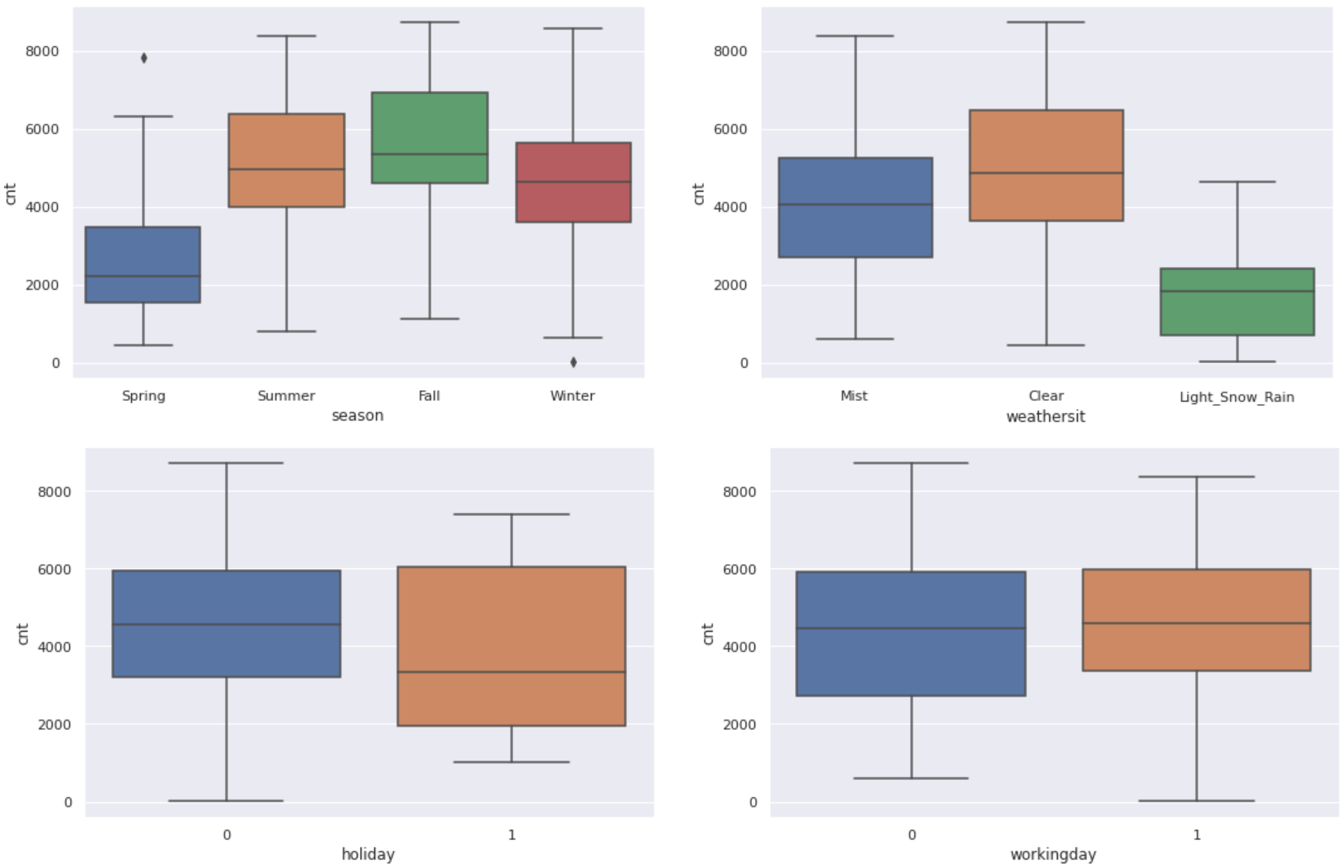
Assignment Based Questions

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

Answer:

After Performing EDA on the dataset, we can infer the following (Charts Attached below):

1. The demand of bikes is less in the season of spring compared to the other seasons.
2. More people tend to rent a bike when the weather seems to be clear.
3. The demand of bike will fall drastically if the weather is not pleasant (light or heavy rains and snow).
4. The company should look out for holidays and keep a good amount of stock ready for renting out the bikes as the demand will tend to go up on holidays.
5. When we look at months combined 2018 and 2019 data, we see that the demand goes up in the month of June and starts falling from the month of October.
6. If we look at the time series data, we see that the overall demand has increased in 2019 compared to 2018.



2. Why is it important to use drop_first=True during dummy variable creation?
Answer:
suppose that we have a column called Season and In Season we have 4 different categories let’s name those 3 categories as ‘Summer’, ‘Winter’, ‘Fall’, ‘Spring’. Now to pass these values to the model we have to create a dummy variable. Instead of creating 4 dummy variables we can achieve the desired results in 3 dummy variables. Let’s say if we set drop_first = True, it will not create dummy variable of the variable Summer. And we will have dummy variables of Winter, Fall and Spring. Now let’s say we have a record for which the value of Season was set to Winter, in the dummy we created the value for Winter would be set to ‘1’ and for Fall and Spring value would be set to ‘0’.
So the codes will we as follows:

Winter:	100
Fall:	010
Spring:	001
Summer:	000

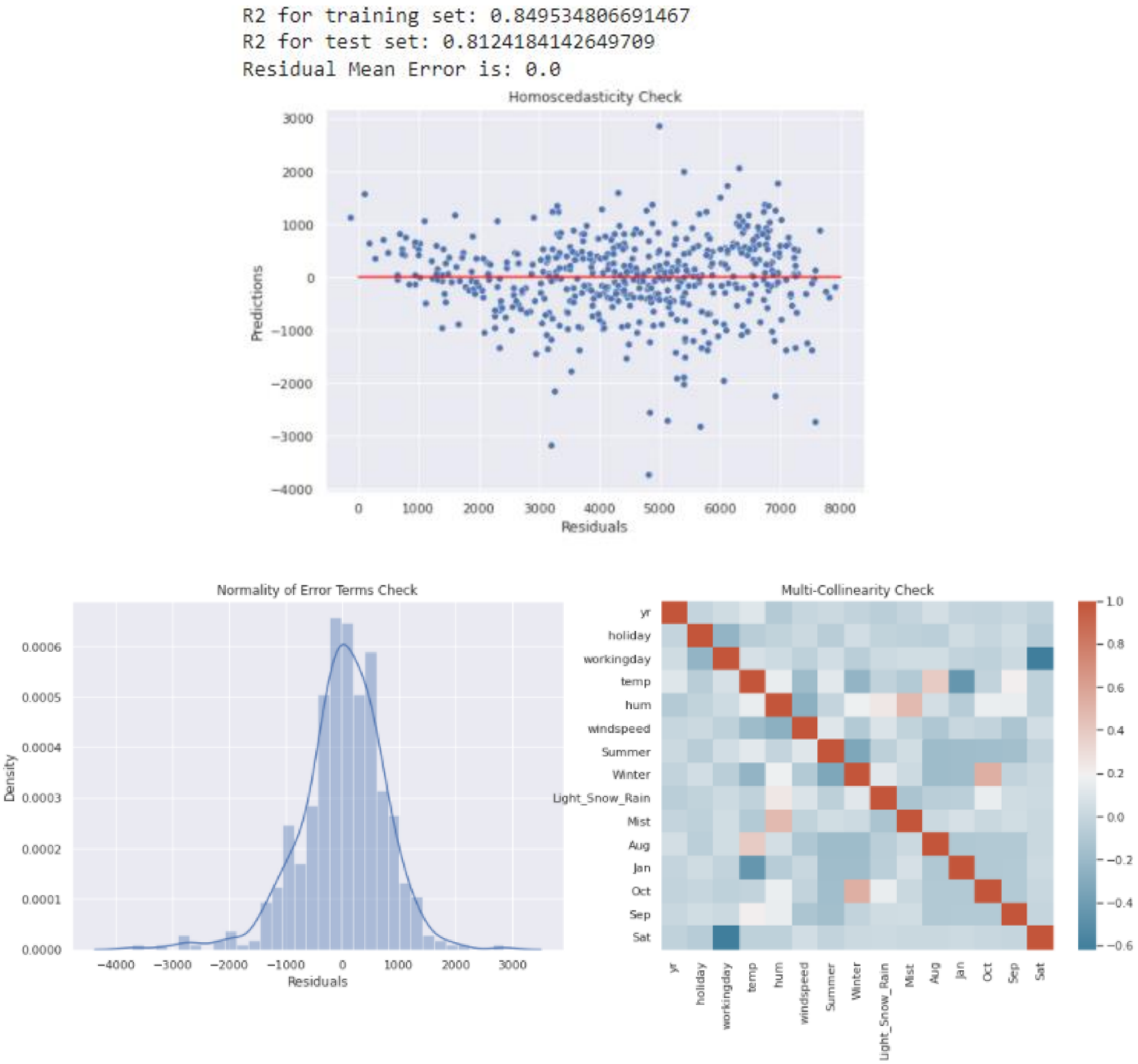
as we see above we do not need to create a dummy variable for summer separately, if all other dummy variables is set to 0, will indirectly mean that the value is set to Summer.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer:
Looking at the pair-plot we can say that temp and atemp variables has the highest correlation of 0.63 with the target variable.

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

1. To Validate the Assumptions of Linear Regression we checked for the Mean of Residual Error, which is obtained by subtracting the predicted values from the actual values of the test set.
   The mean of residual should be equal to 0 to pass this check.
2. We then look at the homoscedasticity – which means to check if the variance of the residuals has to be constant. We checked this by plotting a scatter plot of the predictions and residuals and saw how the data was spread from ‘0’, does the variance appear to be constant or not.
   The variance has to be constant to pass this check.
3. We then check if our residuals or the error terms are distributed normally or not.
   It is important for the residuals to form a normal curve to pass the Linear Regression assumption.
4. We also need to check if the multi collinearity exists or not, if it does we have to drop the variables which cause multi collinearity for holding the assumption of least or none multi collinearity to be true.
   Multi Collinearity has to be eliminated for this assumption to hold true

Attaching a Snapshot below of the charts I used to validate the assumptions



```
R2 for training set: 0.849534806691467
R2 for test set: 0.8124184142649709
Residual Mean Error is: 0.0
```

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

Answer:

Based on the Final model following are our top 3 features contributing significantly towards explaining the demand:

1. Temp (coefficient: 4379.1712)
2. Weathersit: Light Snow, Light Rain (coefficient: -2186.2627)
3. Year (coefficient: 2005.1589)

General Subjective Questions

1. **Explain the Linear Regression algorithm in detail.**

Answer:

Linear Regression is a supervised algorithm in which the model finds the best linear line between the independent and dependent variable. We have 2 types of linear regression, 1. Simple Linear Regression and 2. Multiple Linear Regression. Although Simple Linear Regression is only for academic purpose, we cannot find any real-world application of it as the problems are lot complex and simple linear regression assumes that we have only one independent variable. On the other hand, multiple linear regression has more than one independent variable to find the relationship with the target or dependent variable.

Equation of Linear Regression:

a. Simple Linear Regression:
y=b0 + b1x

b. Multiple Linear Regression:
y=b0+b1x1+b2x2+…bnxn

We use linear regression model to find the best fit linear line and the best values of intercept and coefficients minimizing our residuals. Error/Residuals is the difference between the actual values of y and the values we predict from our model.

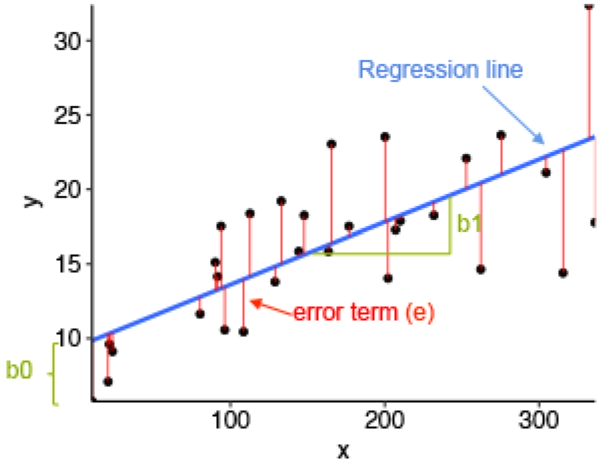


Fig: 2.1

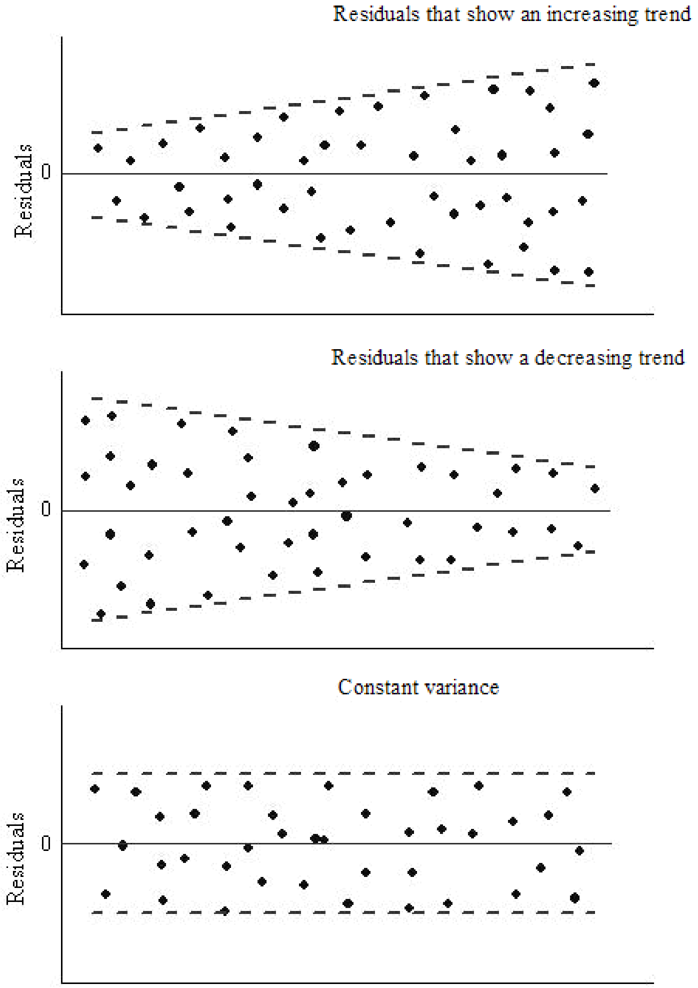
- X is our independent variable, and y is our target variable which we will be predicting using our linear regression model.
- The Scattered black dots are the actual data points.
- From the formula b0 (intercept) is at 10 and b1 is the slope of our line.
- The blue line is the best fit line predicted by the model.
- The vertical red lines from our best fit line to actual data points is termed as error or residual. The sum of all the errors is known as the sum of residuals.

Linear Regression Assumptions:

1. Linearity: there should be linearity between the target and the independent variables. We can check this assumption by plotting a scatter plot between both variables.

2. Normality: The dependent and independent variables should be normally distributed

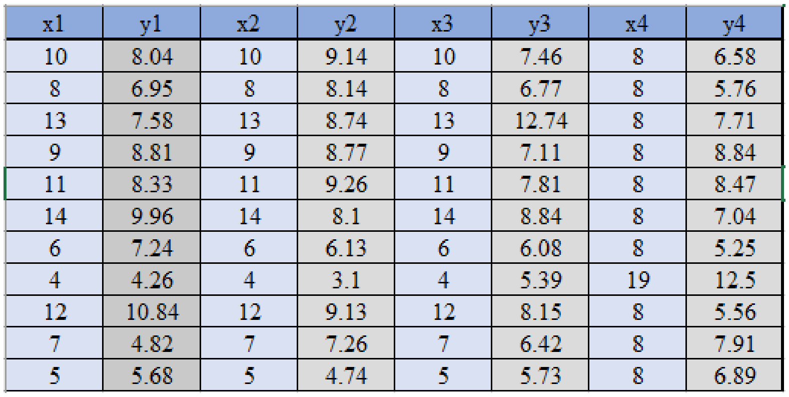
3. Homoscedasticity: The variance of the residuals should be constant. We can check this assumption by plotting a scatter plot of residuals.



5. The Residuals should be normally distributed: Histograms could be used to check the distribution of error terms.

4. No Multi-collinearity: The independent variables should have least or no multi collinearity.

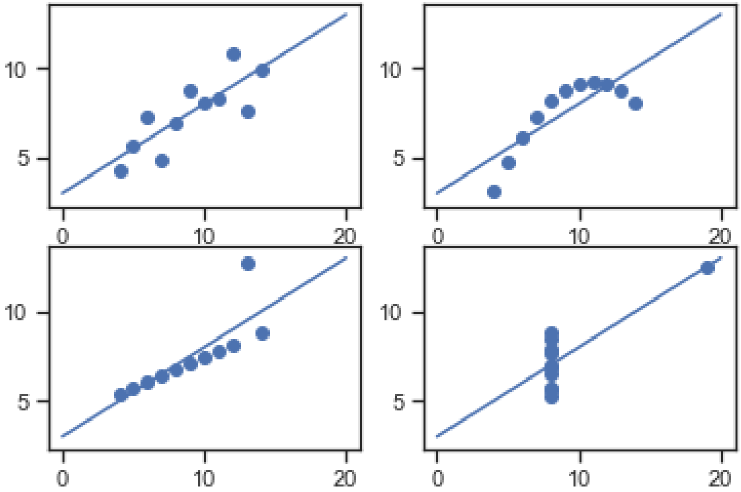
2. **Explain the Anscombe’s quartet in detail.**
Answer:
Anscombe’s Quartet compromises of 4 datasets that have identical simple statistical properties, yet give different results when we plot the data on graph. Each dataset consists of 11 samples. They were introduced in 1973 by the statistician Francis Anscombe to demonstrate the importance and graphing data before analyzing and the effect of outliers on statistical properties. The thing to analyze about these data sets is that all share the same descriptive statistics but their graphical representation come out to be different from each other. Each graph shows different behavior irrespective of similar statistical analysis.
Example:



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: 9.0
Average value of y: 7.50
Variance of x: 11.0
Variance of y: 4.12
Correlation coefficient: 0.82
Linear regression equation: y=0.5x+3
Above are the descriptive statistics for all for 4 data sets used in the example, lets plot these now.



As we see the graphical representations is coming out to be different for all the datasets, having same statistics.

3. **What is Pearson’s R?**
Answer:
Pearson’s R also called Pearson’s correlation. Correlation means to find out the relation/association between the two variables and correlation coefficients are used to find out how strong this relationship is between the two input variables of the dataset. Pearson’s coefficient of correlation is one of the popular and widely used tool to get the correlation of different variables.
Let’s take a basic example of predicting house prices, logically when the area of the house increases so does the price, this is positive correlation.
Pearson’s R measures how strong is the linear relationship between 2 continuous variables using the below formula:

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Value of correlation ranges from -1 to 1, if it is ‘0’ it specifies that there is no correlation between the two variables. A value greater than ‘0’ indicates a positive relationship between the variables and a value less than ‘0’ indicates negative relationship.
Assumptions for a Pearson Correlation:

- Data should be random.
- Variables should be continuous in nature.
- Variance of the data should be constant.

Extreme outliers affect the person correlation coefficient.

4. **What is scaling? Why is Scaling performed? What is the difference between normalized scaling and standardized scaling?**
Answer:
When we have a dataset with multiple features varying degrees of magnitude, range, and units. This is a big issue when it comes to training the data as few machine learning algorithms are highly sensitive to these features. For example, one feature is entirely in kgs while the other is in gms, another one is cm, m and so on.

if we pass these features to our algorithm we will not know if our model is performing well or not or we won’t be able to evaluate our model.
Scaling is the technique where we bring all our features to the same scale, ideally the scale ranges from 0 to 1.

Normalized scaling vs standard scaling:
Normalization is a scaling technique in which values are rescaled so that they end up ranging between 0 and 1. Another name for normalization is min max scaling. Below is the formula for normal scaling:

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

here, Xmax and Xmin are the maximum and minimum values of the feature X.

Standardization is another scaling technique where we minus the actual value from the mean of the variable and divide it by the standard deviation. Below is the formula for normal scaling.

$$X' = \frac{X - \mu}{\sigma}$$

Mu is the mean of the feature values and sigma is the standard deviation of the feature values. Note that in this case the values are not restricted to a particular range.

When to use what?
For example, you are working on financial data and do not wish to ignore the outliers as it could be some significant information there you can use standard scaler and if you do not mind you outliers being rounded off to 1 or 0 we can use normalization here.

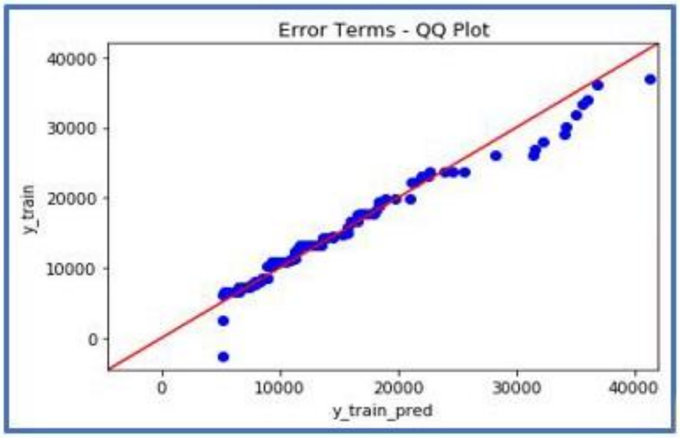
5. **You might have observed that sometimes the value of VIF is infinite. Why does that happen?**
Answer:
VIF (variance inflation factor) is a tool which helps us identify if there exists multi collinearity in the data set and if yes how strongly the variables are related. it will provide us with a VIF score looking at which we can handle multi collinearity scenarios.
If all the independent variables are independent of each other, then VIF equals 1. if there is a perfect correlation, then VIF equals infinity. A large value of VIF indicates that the variable is correlated with one or more variables and this needs to be handled.
A general rule of thumb is that if VIF > 10 then there is multi collinearity and it is a serious issue.
Note that, in some cases we might choose to live with high VIF values if it does not affect our results but to hold the assumption of no multi collinearity we need to take care of variables with high VIF values as it will help us achieve a more general model.
A VIF values < 5 is considered to be good, again it depends on the business problem you are tackling. There is no hard and fast rule for the thresholds.

6. **What is Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
Answer:
Q-Q or Quantile-Quantile plot is a plot of quantiles of first data set and is compared with the quantiles of the 2nd dataset.
In Linear regression when we receive test and train data separately, we can create Q-Q plots and check if the distribution of the data is same or not.

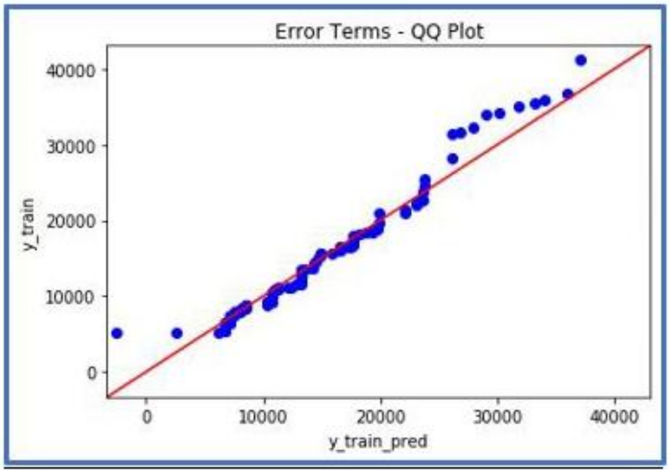
- It can be used with small datasets.
- Many aspects like shifts in location, scale, symmetry and the presence of outliers can all be known using this graphical representation.
- It is used to check the following scenarios:
  - If the two datasets Come from populations with a common distribution.
  - If the two datasets Have common location and scale.
  - If the two datasets Have similar distributional shapes.
  - If the two datasets Have similar tail behavior.

Interpretation:

1. Similar distribution: if all points lies on or near the line we say it is a similar distribution.
2. Y-values < X-values: if y-quantiles are lower than the x-quantiles, below line



3. X-values < Y-values: if x-quantiles are lower than the y-quantiles, above line



4. Different distributions: if all point of quantiles lies away from the line of 45 degrees.