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 – Demand of rental bike [cnt] which is dependent variable does very much depend on categorical columns such as seasons where we can see that:

season	sum	mean	count	sum_perc	count_perc
Fall	1061129	5644.30	188	32.24	25.75
Summer	918589	4992.33	184	27.91	25.21
Winter	841613	4728.16	178	25.57	24.38
Spring	469514	2608.41	180	14.27	24.66

Rental bike counts during fall was highest when compared all other seasons similarly for other categorical columns like Months , Weather, Weekday, Holiday, Working day and year also shows trend of rental bike count.

Below was the observation for categorical column:

- People tend to take rental bikes mostly during clear weather and avoid mostly when there is lightning/rainstorm.
- Falls are preferred for rental bikes.
- Demand for rental bikes are higher during May to September
- Weekdays don't much help in analysis we still need to analyze more to understand if it can be a good predictor or not.
- Holidays surely impact the rental bike count , there should be no holiday.
- It should be working day, which helps in increasing rental count.
- 2019 was much better than 2018 for rental bike count.

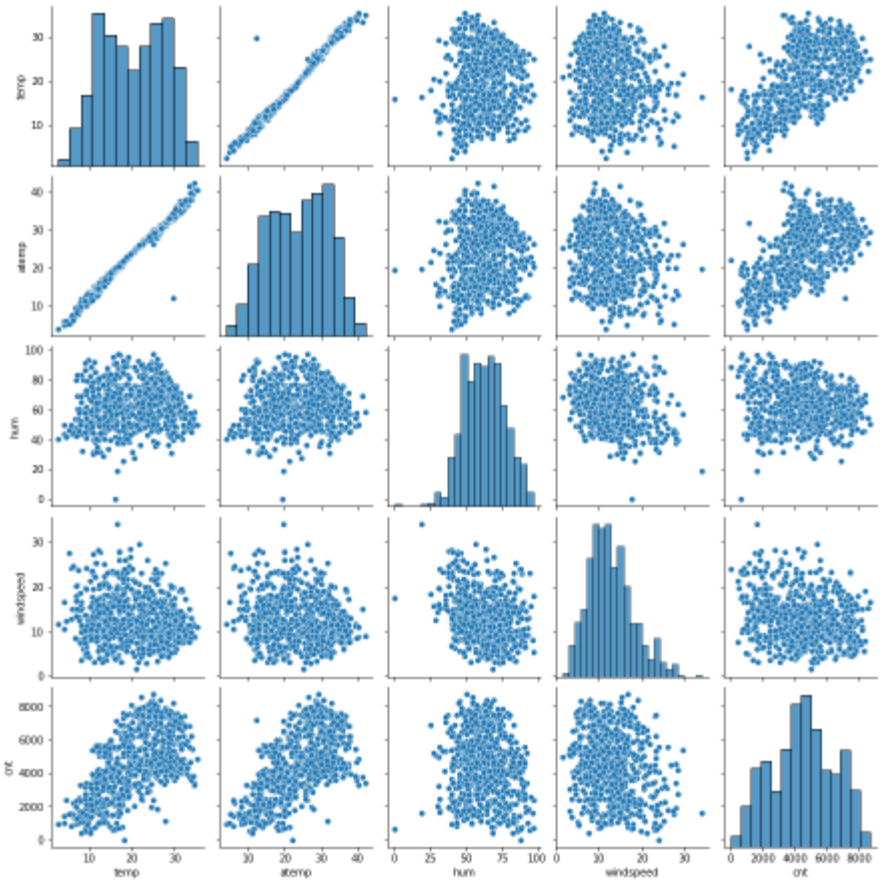
Clearly from above points we can say that categorical variable does affect rental bike counts.

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

Ans- We need to use drop_first=True as it is important. It helps in reducing the extra column created during dummy variable creation. Hence, it reduces the correlations created among dummy variables. For example, we would have got 4 different columns for seasons if we would have not used drop_first=True, but if since we used it during analysis, one of the column – Fall got reduced from dataset which indicate that if all the columns have value = 0 it means that directly the value is Fall.

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

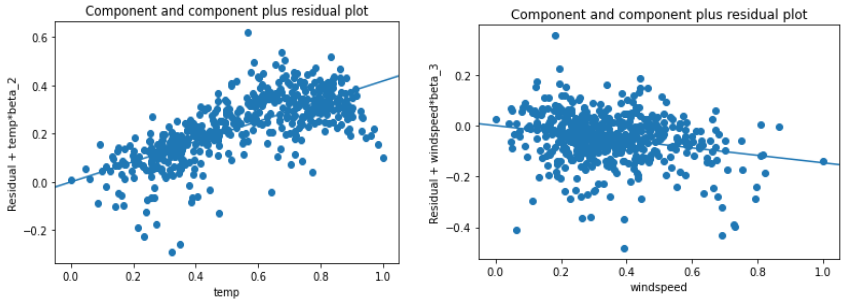
Ans- We can clearly see after looking the pair plot below that temp and atemp show extreme/highest correlation with target variable.



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

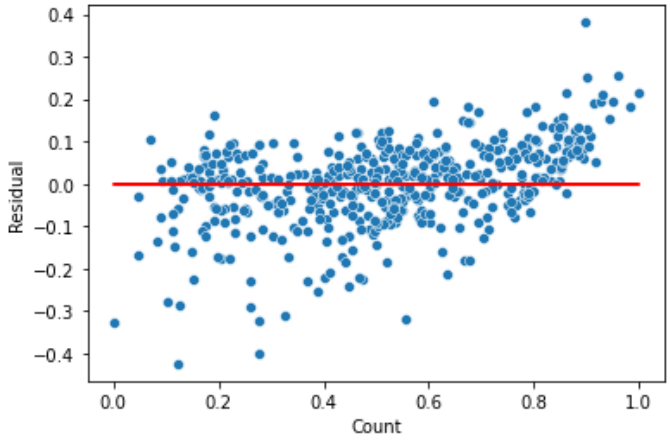
Ans- We had to validate the model on training dataset whether we still stick to Liner Regression model. We checked the model on below points:

- We try to form a scatter graph between model and predictor and we did observe a linear relationship.

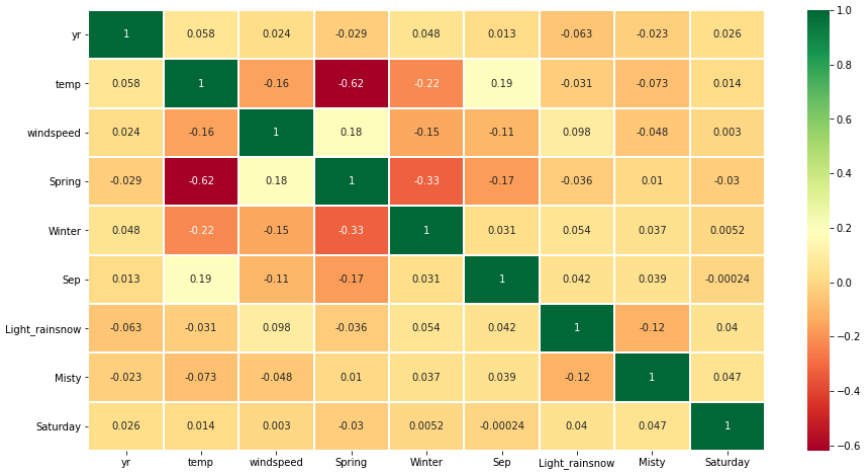


- **Homoscedasticity:**

  Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.

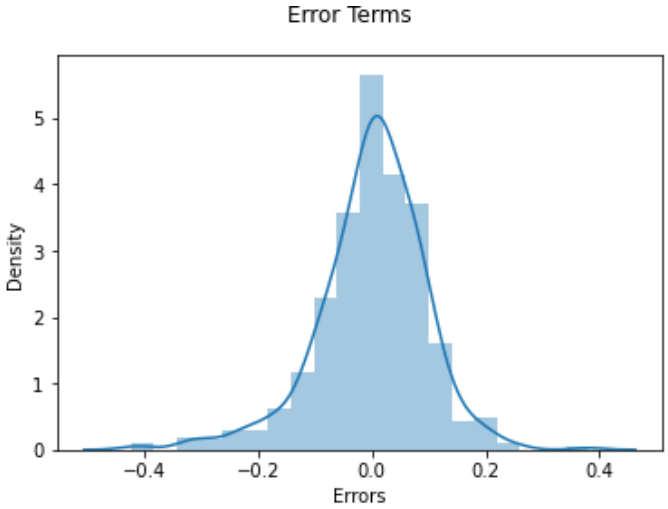


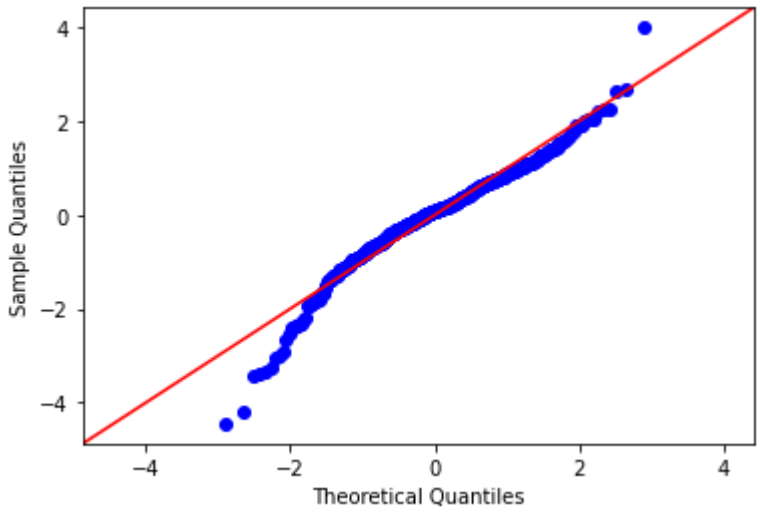
- **Absence of Multicolinearity:**
  Multicolinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case).



- **Normality of error:**

  If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.





Only when all the conditions are satisfied we can say that there is no violation on the linear regression assumption.

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

Ans- As per the final model, the top 3 predictor variables that influence bike booking are:

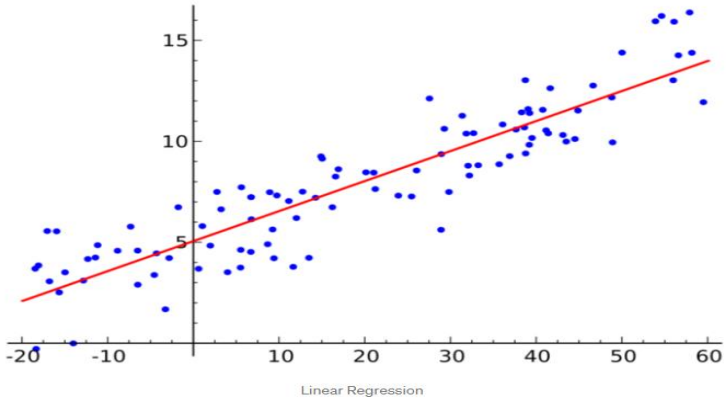
- **Temperature (Temp):** A coefficient value of '0.417791' indicated that a temperature has significant impact on bike rentals
- **Light Rain & Snow (weathersit =3):** A coefficient value of '-0.309695' indicated that the light snow and rain deters people from renting out bikes
- **Year (yr):** A coefficient value of '0.231845' indicated that a year wise the rental numbers are increasing

It is recommended to give utmost importance to these three variables while planning to achieve maximum bike rental booking. As high temperature and good weather positively impacts bike rentals, it is recommended that bike availability and promotions to be increased during summer months to further increase bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans- Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y).
When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, often refers to the method as multiple linear regressions.
Different techniques can be used to prepare or train the linear regression equation from data. It is common to therefore refer to a model prepared this way as Least Squares Regression.
Below graph shows linear algorithm:



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modeled based on the linear equation shown below:

$Y = mx + C$
The motive of the linear regression algorithm is to find the best values for m and C.
Types of Linear Regression -
Linear regression can be further divided into two types of the algorithm:

- Simple Linear Regression:
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- Multiple Linear regression:
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Assumptions of Linear Regression -
Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- Linear relationship between the features and target:
  Linear regression assumes the linear relationship between the dependent and independent variables.
- Small or no multicollinearity between the features:
  Multicolinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. It is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- Homoscedasticity Assumption:
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- Normal distribution of error terms:
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.
- No autocorrelations:
  the linear regression model assumes any autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model.
  Autocorrelation usually occurs if there is a dependency between residual errors.

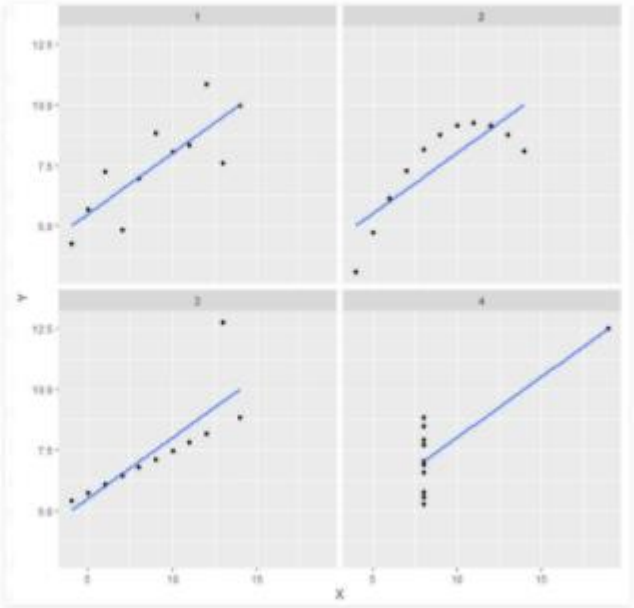
2. Explain the Anscombe's quartet in detail.

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.
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Output:



Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

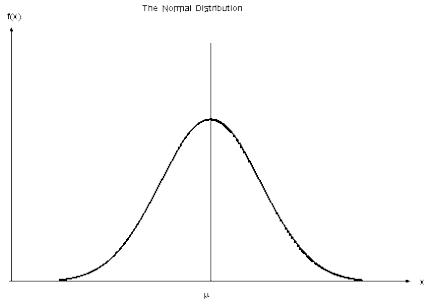
Application:
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

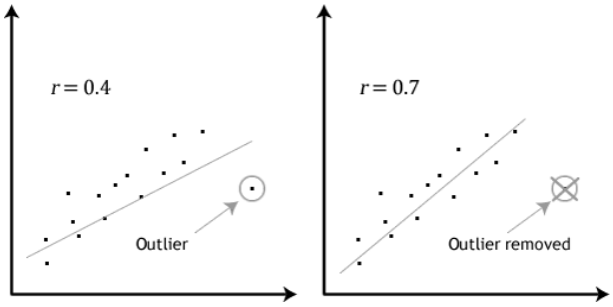
Ans- Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression.

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

Assumptions
1. For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.



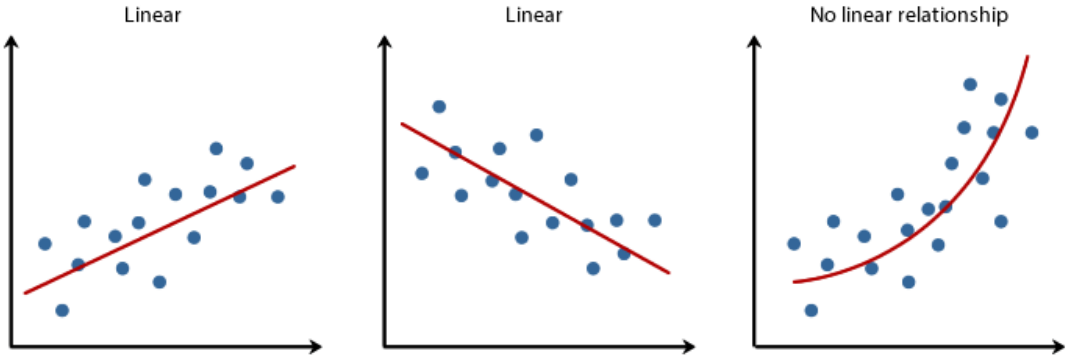
Normal Distribution
2. There should be no significant outliers. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.



Outliers

3. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

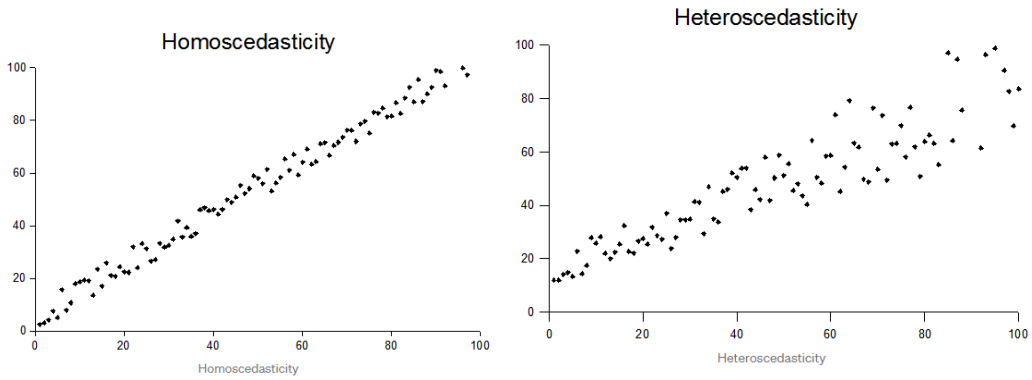
4. The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric .



Linear and non-Linear Relationships

5. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example if you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks.

6. Homoscedasticity: Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is Homoscedasticity. As a bonus — the opposite of Homoscedasticity is heteroscedascity (the violation of homoscedasticity) which is present when the size of the error term differs across values of an independent variable.



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

Ans- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude 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 magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans -VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$

$[VIF]\_1=1/(1-R\_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X\_2=C+ \alpha\_1 X\_1+\alpha\_3 X\_3+\cdots$

$VIF\_2=1/(1-R\_2^2)$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regressions. Here are the various options:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

6. 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.
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.

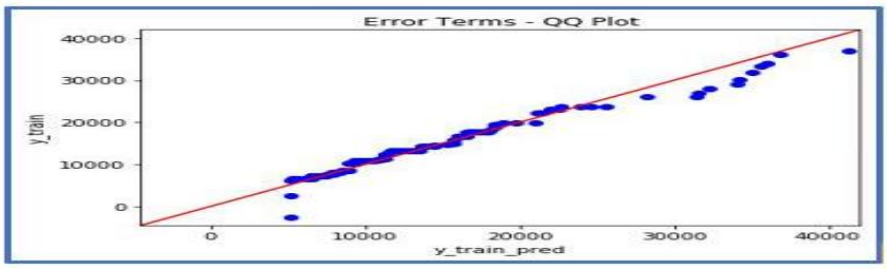
It is used to check following scenarios:
If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

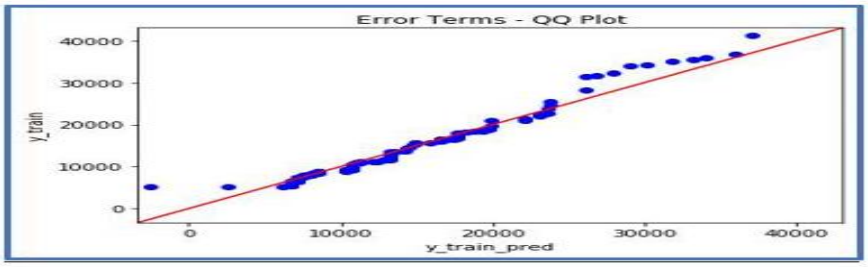
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

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.



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