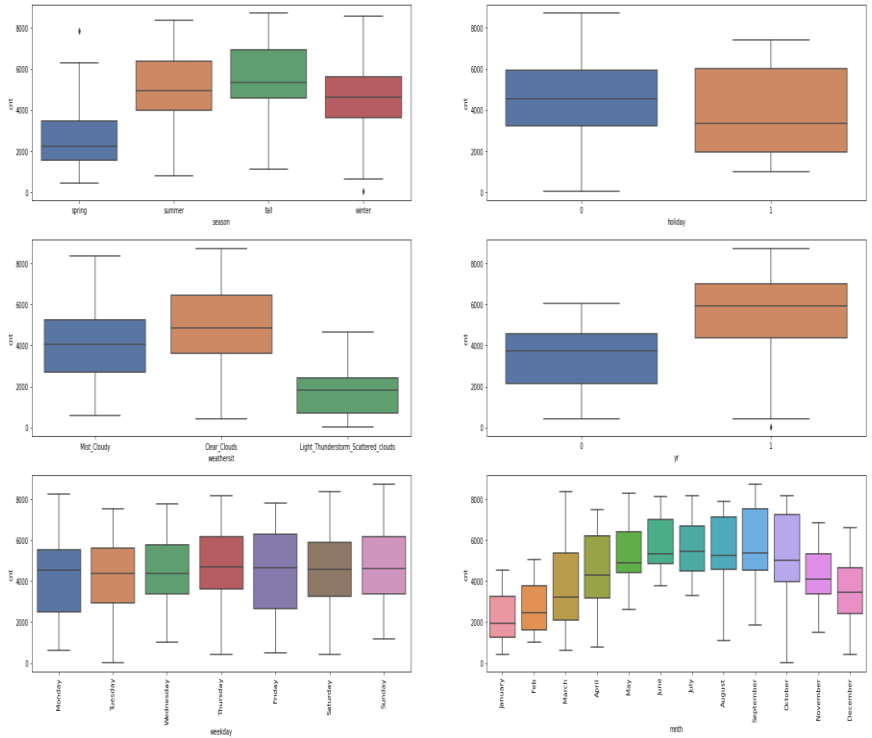
Assignment-based Subjective Questions

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



season: The demads of bikes is less in spring season compared to other seasons

holidays: Demands of bikes is less during holidays

weathersit: Demands are high when the weather is clear with few clouds. Demands are very less during Thunderstrom\Light Rain\Light Snow compared to Mist cloudy.We do not see any demands of bikes during Heavy rains + Ice Pallets + Thunderstorm + Mist, Snow + Fog

yr: Demands of bikes are high in the year 2019

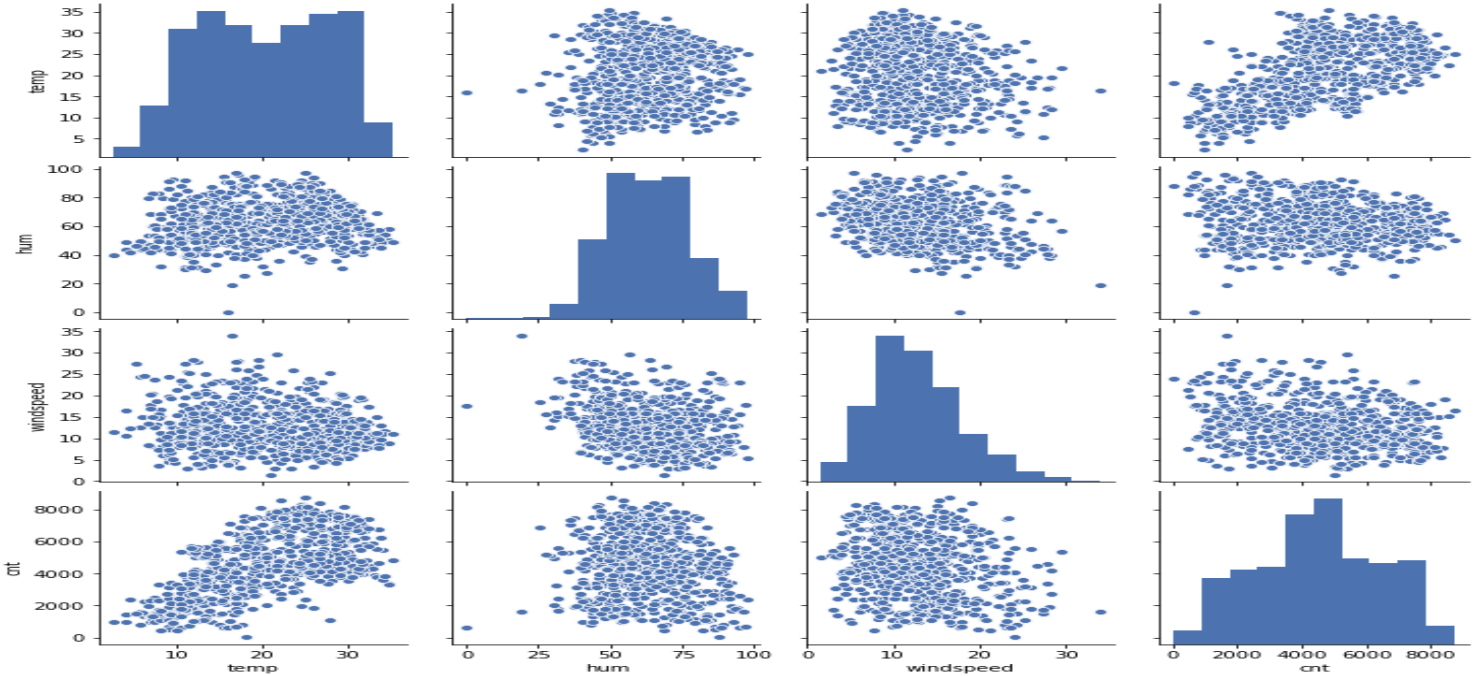
weekday: Demands of bikes is almost similar during weekday

mnth: January month has the lowerst bike demands and months March, June, July, September, October has the highest bike demands

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

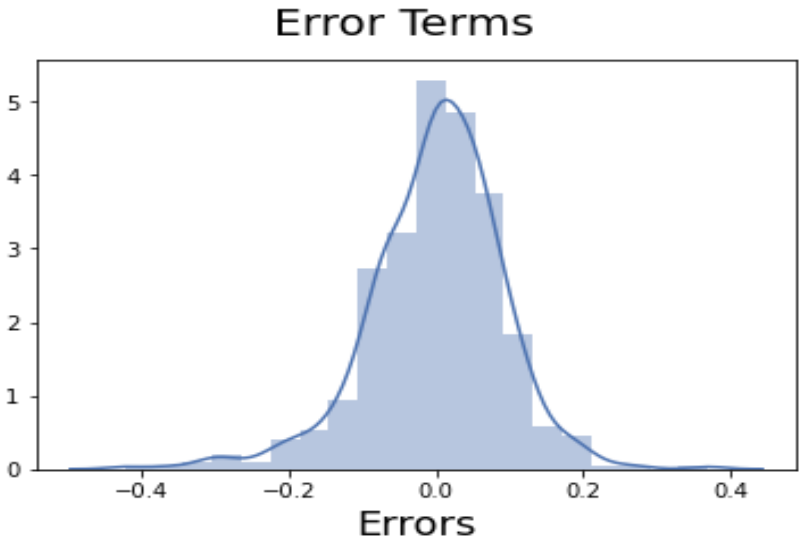
drop_first=**True** is **important to use**, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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



From the graph we can infer that column temp has the highest correlation with the target variable cnt

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



	Features	VIF
2	temp	4.60
3	windspeed	4.00
0	yr	2.06
4	spring	1.65
7	Mist_Cloudy	1.51
5	winter	1.40
8	July	1.35
9	September	1.20
6	Light_Thunderstorm_Scattered_clouds	1.08
1	holiday	1.04

Validate the assumptions of Linear Regression

- The distribution plot show normal distribution with mean 0
- no multicollinearity between predicted variables and the VIF is less than 5

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

The equation of our best fitted line is:

cnt=0.252+(0.234×yr)-(0.099×holiday)+(0.451×temp)−(0.140×windspeed)-(0.111×spring)+(0.047×winter)-(0.286×Light_Thunderstorm_Scattered_clouds)−(0.081×Mist_Cloudy)−(0.073×July)+(0.058*September)

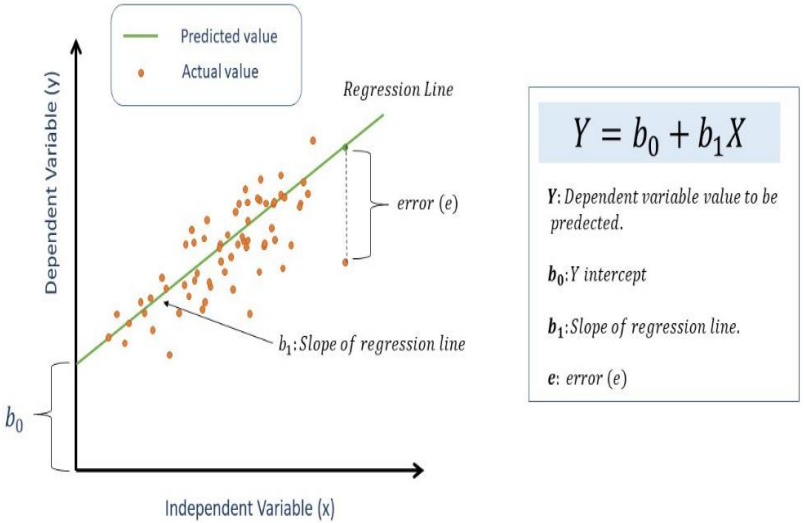
From R-Sqaured and adj R-Sqaured value of both train and test dataset we could conclude that the above variables can well explain almost 80% of bike demand.
Coeffiencients of the variables explains the factors effecting the bike demand

Top 3 features contributing significantly towards explaining the demand of the shared bikes1:

1. Temperature - (0.451) : a unit increase in the temp variable, increases the bike hire demands by 0.451 units
2. year - (0.234) : a unit increase in the year variable, increases the bike hire demands by 0.234 units
3. Light_Thunderstorm_Scattered_clouds - (-0.286) : a unit increase in the Light_Thunderstorm_Scattered_clouds variable, decreases the bike hire demands by 0.286 units

General Subjective Questions

Question 1:
Explain the linear regression algorithm in detail



In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

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 Residual Sum of Squares Method. Linear regression models can be classified into two types depending upon the number of independent variables:

Simple linear regression: This is used when the number of independent variables is 1.

Multiple linear regression: This is used when the number of independent variables is more than 1.

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$

The assumptions of linear regression are:

Assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.

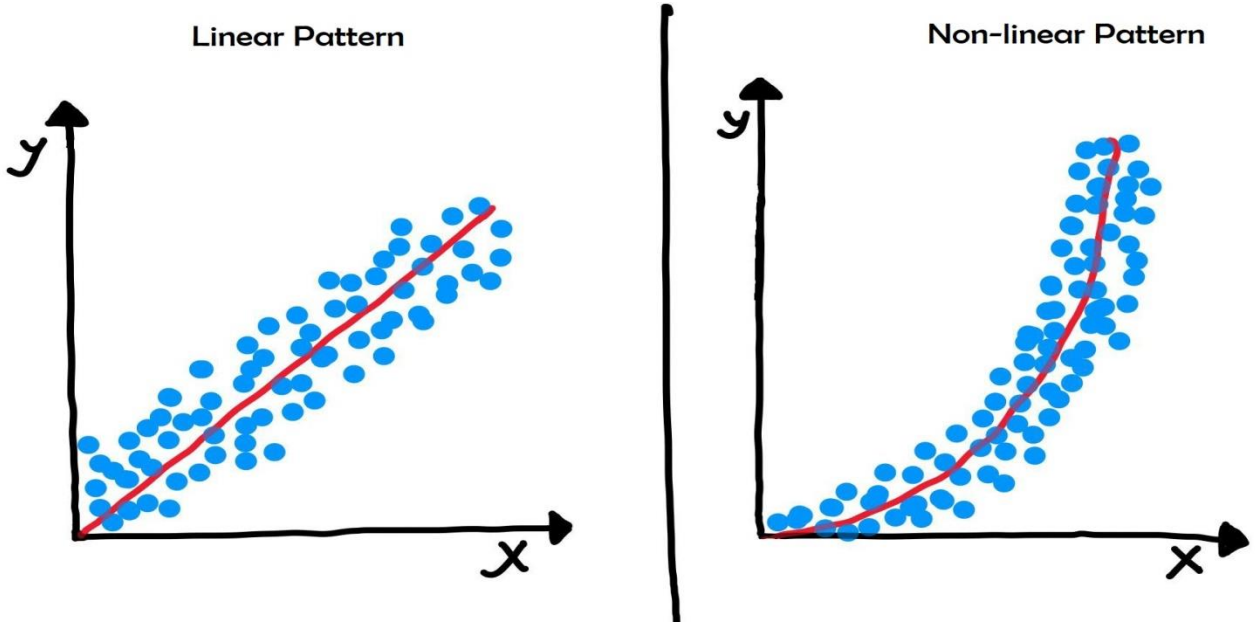
Assumptions about the residuals:

- **Normality assumption**: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
- **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.
- **Constant variance assumption**: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
- **Independent error assumption**: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

Let's understand the importance of each assumption one by one:

1. There is a *linear relationship* between X and Y:

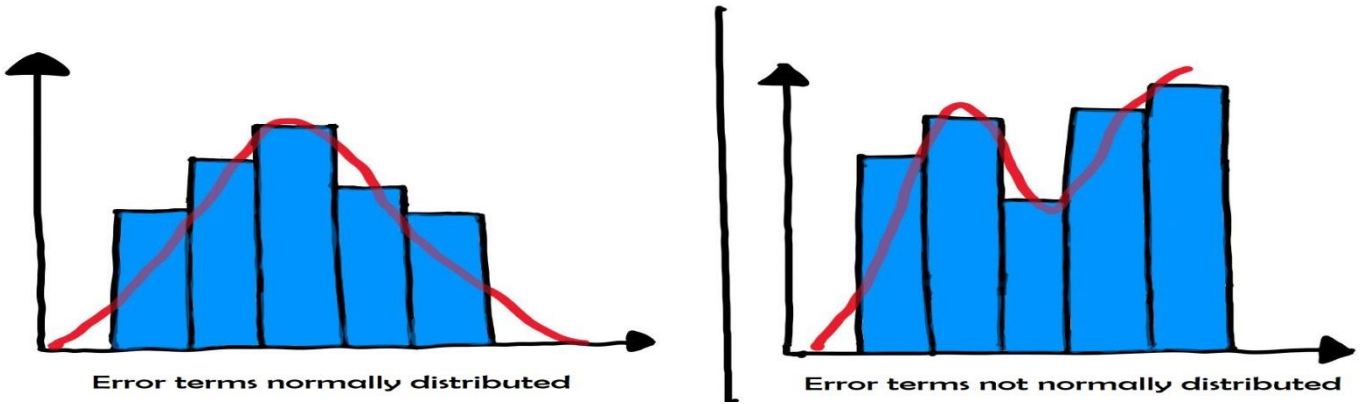
X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.



2. Error terms are *normally distributed* with mean zero(not X, Y):

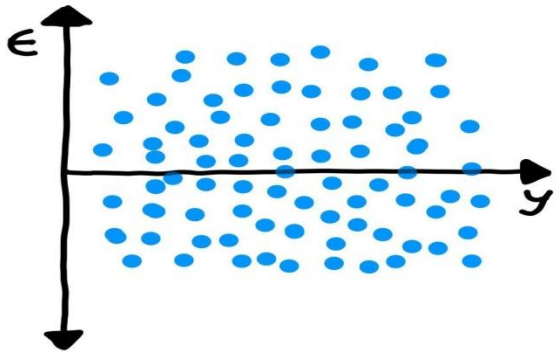
There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.

But if you are willing to make some inferences on the model that you have built , you need to have a notion of the distribution of the error terms. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable. (You'll see this in a later segment)
The assumption of normality is made, as it has been observed that the error terms generally follow a **normal distribution with mean equal to zero** in most cases.

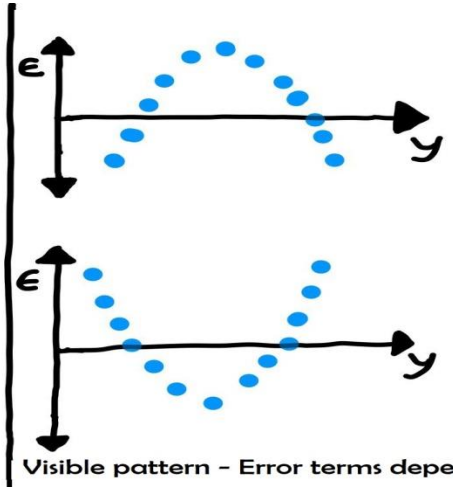


3. Error terms are *independent* of each other:

The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).



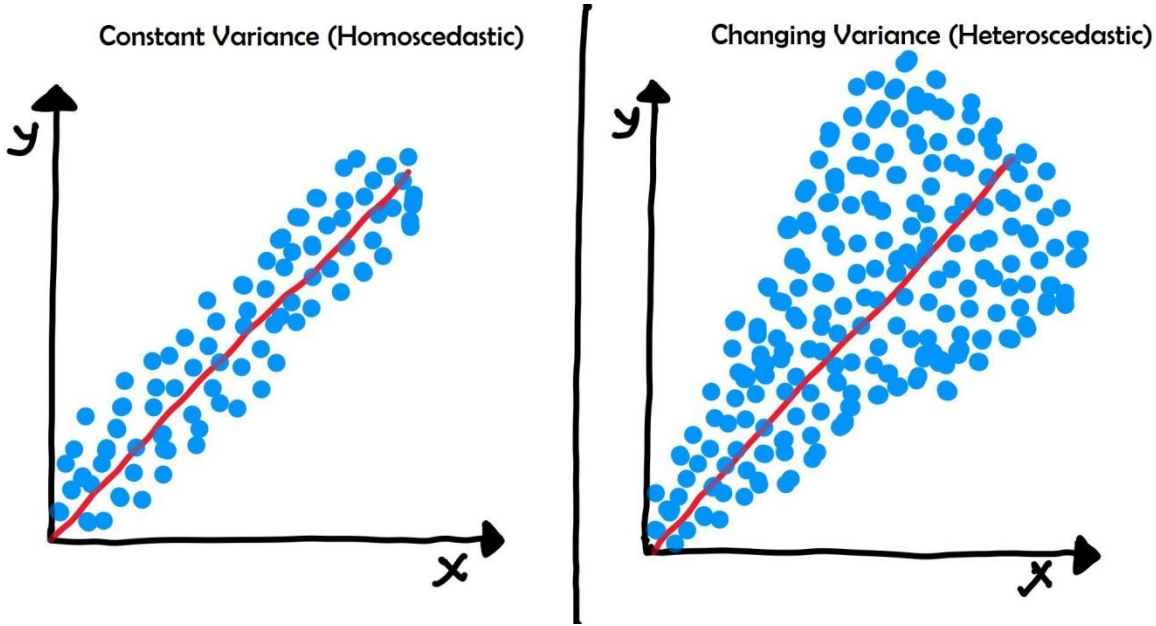
No visible pattern - Error terms independent



Visible pattern - Error terms dependent

4. Error terms have *constant variance* (homoscedasticity):

The variance should not increase (or decrease) as the error values change.
Also, the variance should not follow any pattern as the error terms change.

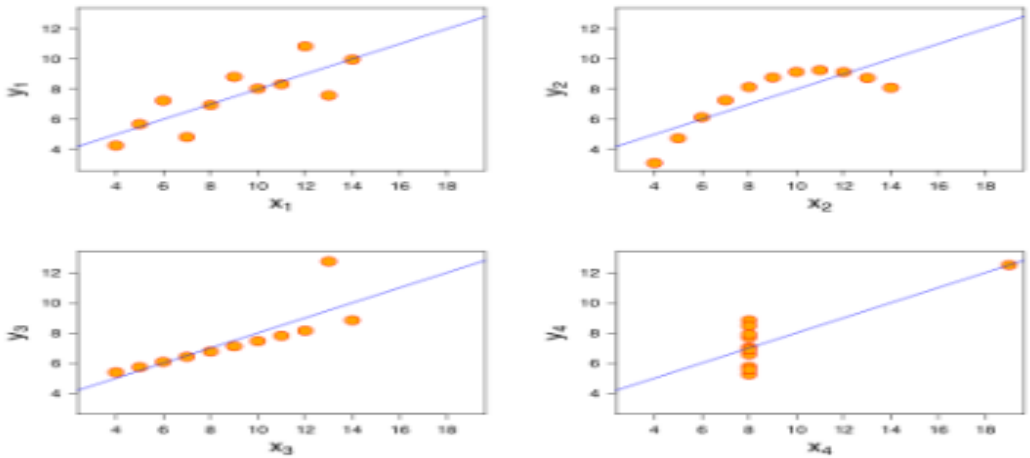


Question 2:
Explain the Anscombe’s quartet in detail.

You should never just run a regression without having a good look at your data because simple linear regression has quite a few shortcomings:

- It is sensitive to outliers.
- It models linear relationships only.
- A few assumptions are required to make the inference.

These phenomena can be best explained by the Anscombe's Quartet, shown below:



Overall moral: first- and second-order summary statistics don't say everything you might want to know about your data, so remember to plot it.

The top right panel shows that, even though we are taught that correlation measures a linear association between two variables, we can have high correlations (0.816 in this case) even when the relationship is nonlinear.

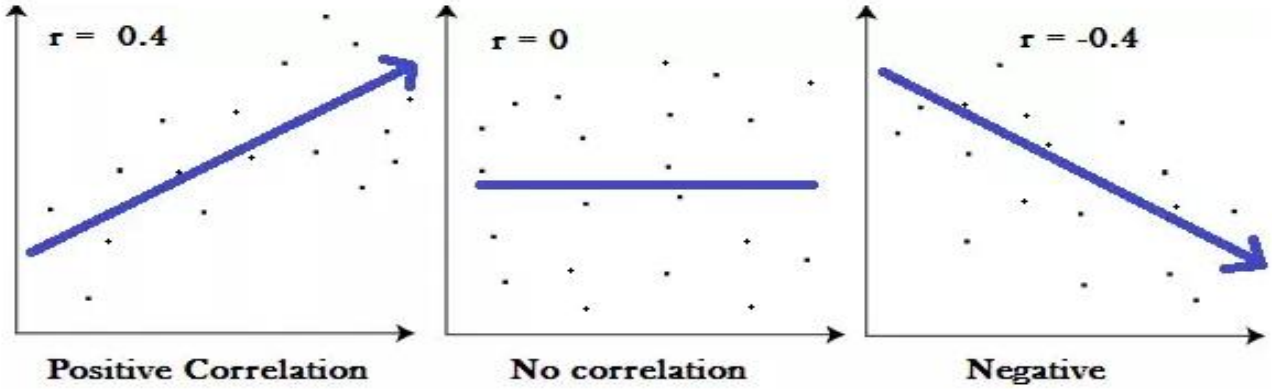
The bottom two panels show that these summary statistics are sensitive to outliers. A generalization of the bottom right panel arises often in real life: you might have, say, two noisy clouds corresponding to two groups. You observe a correlation induced by the grouping, but after controlling for the group, the correlation disappears.

Had the outlier not been present, we could have got a great line fitted through the data points. So, we should never run a regression without having a good look at our data.

Question 3:
What is Pearson's R?

Pearson's Correlation Coefficient (r), defined as the (sample) covariance of the variables divided by the product of their (sample) standard deviations, measures the strength of a linear relationship between two quantitative variables. The results will be between -1 and 1.
If two variables are correlated, they can possibly have any relationship and not just a linear one.

But the important point to note here is that there are two correlation coefficients that are widely used in regression. One is Pearson's R correlation coefficient, which is the correlation coefficient that you learnt about in the linear regression model. This correlation coefficient is designed for linear relationships, and it might not be a good measure for a non-linear relationship between the variables.

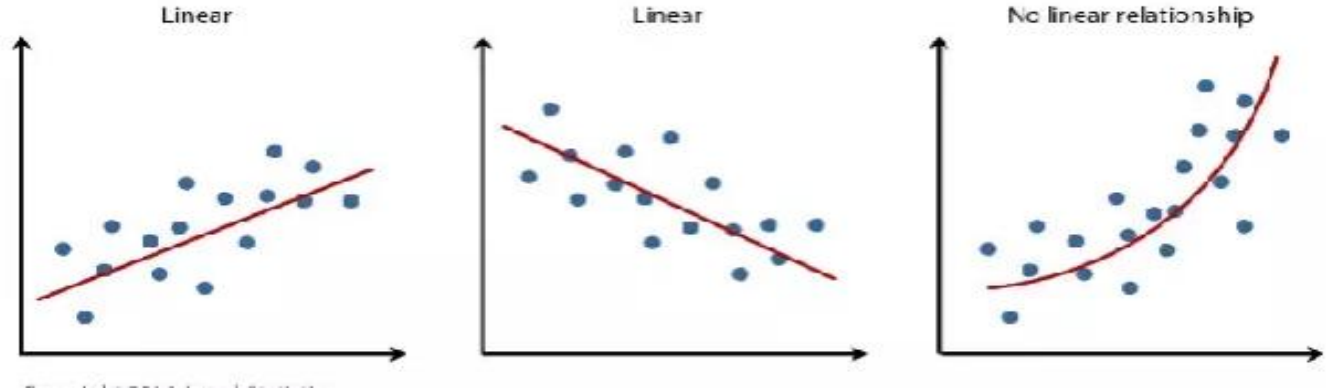


Usually, you get a number somewhere in between those values. The closer the value of r gets to zero, the greater the variation the data points are around the straight line of best fit. Positive values indicate direct relationships (as one variable increases, the other increases as well). Negative values indicate inverse relationships

Some things to consider:

• Correlation does not imply causality. That is, high correlations does not mean one variable causes variations or have an effect over the other.

• Small values do not mean that there is no relationship between variables, only that the linear relationship is low.



A t-test is used to establish if the correlation coefficient is significantly different from zero, and, hence that there is evidence of a linear association between the two variables.

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Techniques to perform Feature Scaling
Consider the two most important ones:
1. Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

2. Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Difference between normalized scaling and standardized scaling

Sl no	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Variance inflation factor (VIF) is used to check the presence of multicollinearity in a data set. It is calculated as:

$$VIF_i = \frac{1}{1-R_i^2}$$

Here, VIFi is the value of VIF for the ith variable, Ri2 is the R2 value of the model when that variable is regressed against all the other independent variables.

If the value of VIF is high for a variable, it implies that the R2 value of the corresponding model is high, i.e., other independent variables are able to explain that variable. In simple terms, the variable is linearly dependent on some other variables.

The common heuristic we follow for the VIF values is:
> 10: VIF value is definitely high, and the variable should be eliminated.
> 5: Can be okay, but it is worth inspecting.
< 5: Good VIF value. No need to eliminate this variable.

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.

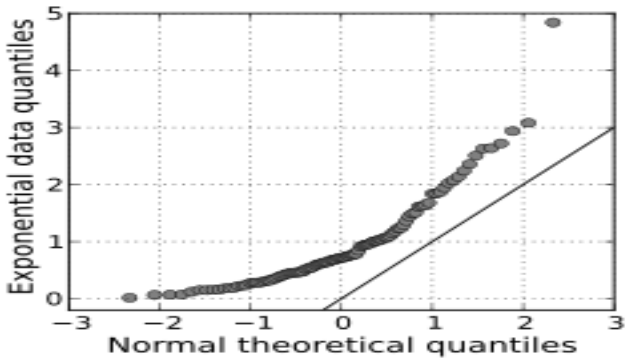
The user has to select the variables to be included by ticking off the corresponding check boxes. An **infinite VIF** value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).

Question 6:
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



A Q Q plot showing the 45 degree reference

USE of Q-Q plot

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

importance of a Q-Q plot in linear regression

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.