Assignment-based Subjective Questions:

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

Answer 1: As per the multiple univariate and bivariate analysis, we can infer few interesting points regarding the demand of the bikes. Here, we are predominantly focusing on the 'cnt' variable, most of the bivariate analysis is done by considering it as one of the attribute and remaining ones as other attribute. Below are the few points which are inferred from the analysis on categorial variables:

1. Demand is maximum for **fall** season as compared to any other season and it increased drastically in 2019.
2. For both the years (2018, 2019), the demand is high between **May** and **September**. The **lowest** demand is in **January** then it starts increasing gradually.
3. It is clear from the plots that the demand is high in **clear** weather while **lowest** in **light or snow or rain**.
4. People prefer more bikes on **holidays** as compared to **non-holidays**.
5. There is not much difference in the demand based on **working** or **non-working** day but has been increased 2019.
6. There is no massive difference in the demand based on the **weekday**, but it is lowest at the start of the week.
7. **temp** and **atemp** are highly correlated with each other. These attributes are clearly showing the linear relation with **cnt**.
8. The requirement of bikes is increasing drastically with the change in **year**.

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

Answer 2: Below are the 2 main reasons of dropping the first dummy variable:

- To avoid multicollinearity: If we don't drop it then dummy variables will be correlated which affects the model adversely.
- To avoid the redundant variable.

We know that the **number of variables should be less for better interpretation of the model**. drop_first = True is important to use because it helps in reducing the extra column created during dummy variable creation. When we create the dummy variables from categorial variable G, with k unique values, it creates k dummy variables to represent all the values of G. We can represent the same information with k-1 dummy variables. Which means we always get 1 redundant dummy variable which can be deleted for better interpretation.

Let's take the below scenario to understand why it is important to drop first column.

A column City contains 3 unique values Delhi, Hyderabad, Bangalore, and we need to convert it to dummy variable. Initially the column will look like below:

City
Bangalore
Hyderabad
Delhi
Hyderabad
Delhi

After creating the dummy variables for this column, it will look like below:

Bangalore	Hyderabad	Delhi
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1

So, the interpretation will be like:

100 = Bangalore
010 = Hyderabad
001 = Delhi

Now, we can provide the same information using 2 columns instead of 3 and the representation will be like:

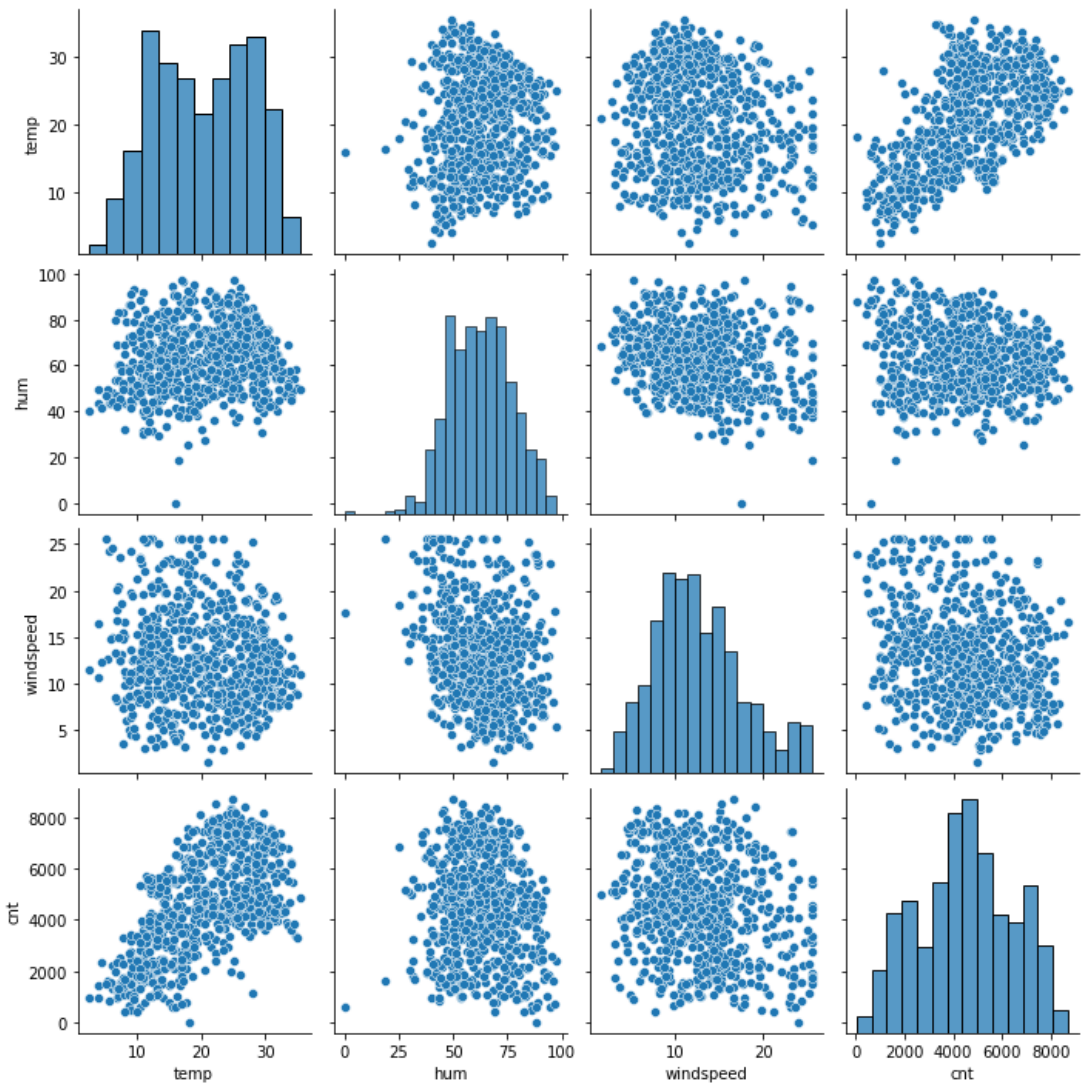
00 = Bangalore
10 = Hyderabad
01 = Delhi

Hyderabad	Delhi
0	0
1	0
0	1
1	0
0	1

So, we can see that the same level of information can be provided using only 2 variables and the extra dummy variable was a redundant variable which can be deleted in all the cases.

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

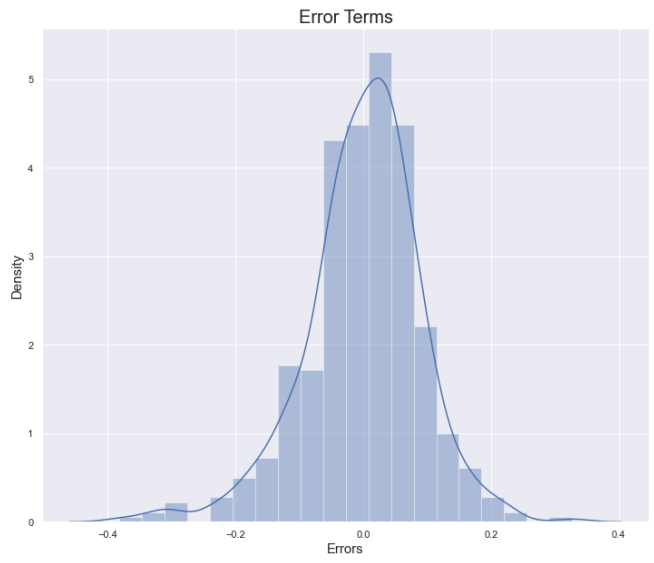
Answer 3: By looking at the pair-plot, it is clearly visible that 'temp' has the highest correlation with the target variable (cnt).

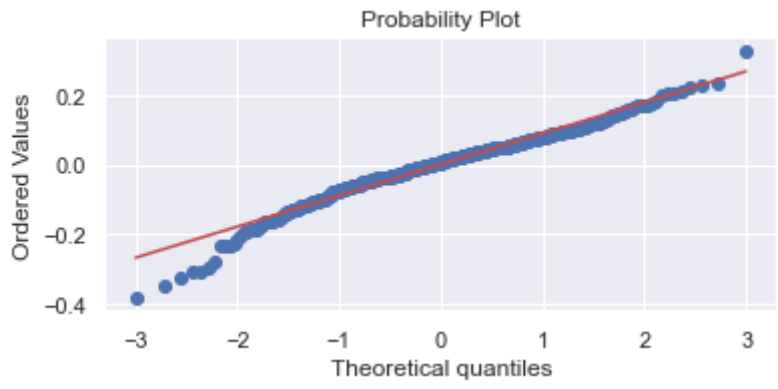


Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

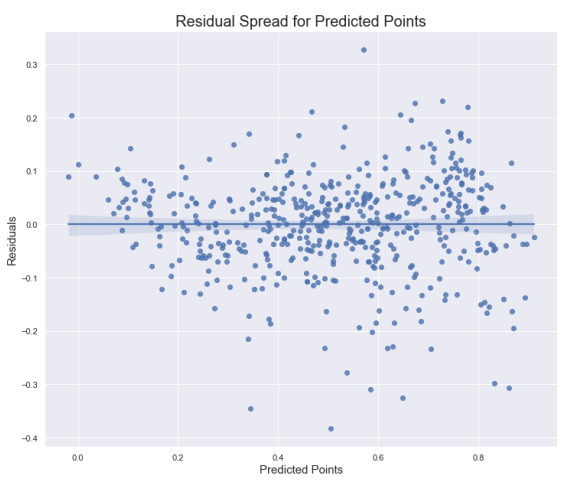
Answer 4: While building the model, validation of assumption is very important step which has been carried out in my analysis. Below are the assumptions which are validated:

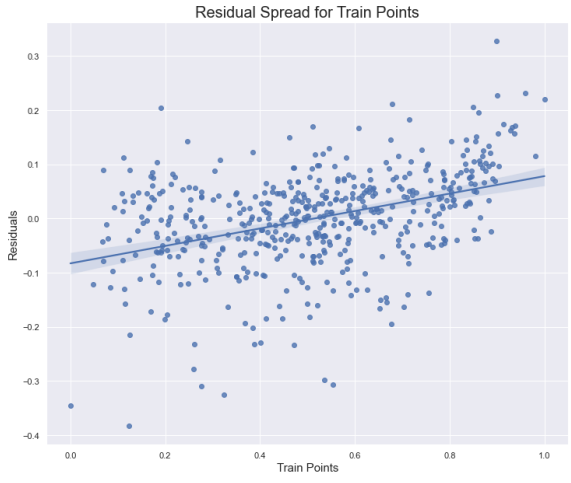
1. **Error terms are normally distributed:** I have calculated the residual (y_train - y_train_pred) and plotted the histogram to see the distribution.





2. **Homoscedasticity:** There is no pattern found in the plot between error and y points.

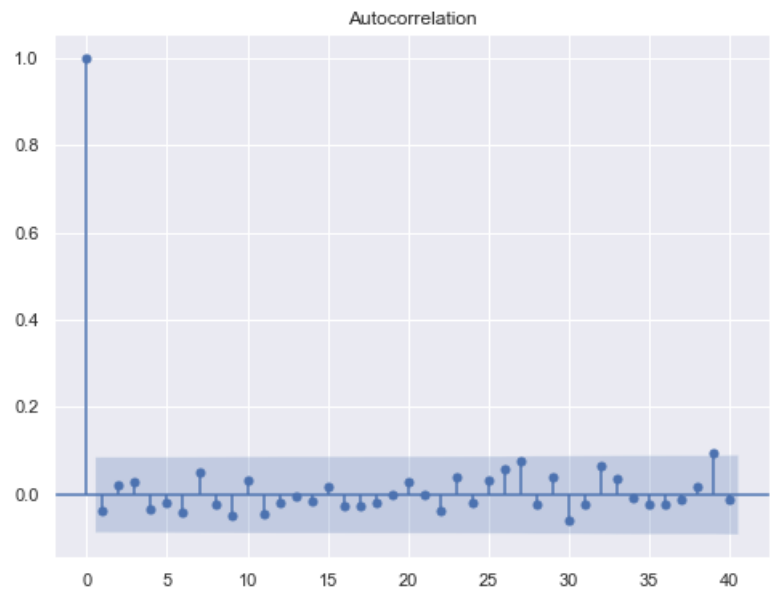




3. **Multicollinearity:** At each step, I have calculated the VIF (Variance Inflation Factor) to make sure the attributes are multicollinear. At the end, VIF is less than 5 for all the Features.

	Features	VIF
0	const	45.06
4	hum	1.86
3	temp	1.60
11	MistCloud	1.55
8	July	1.43
6	summer	1.33
7	winter	1.29
10	LightSnowAndRain	1.24
9	September	1.19
5	windspeed	1.18
1	yr	1.03
2	holiday	1.02

4. **Autocorrelation:** Plotted the graph of residuals to check the correlation of error terms. There is no autocorrelation between error terms.



5. **Linear Relationship:** We have seen there is some linear relationship between dependent and independent variables.

As described above, all the assumptions are validation after building the model.

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

Answer 5: Top contributing features are calculated with the help of coefficient. Features with higher coefficient values are highly contributing to the demand. As per the model, below are the top 3 features which are contributing significantly to the demand:

a) **Temperature** (0.5976)
b) **Weather Situation**: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2319)
c) **Year** (0.228)

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer 1: In machine learning, linear regression algorithm is used to find the relation between dependent and independent attributes. It is a supervised learning model i.e., learning a function that maps an input to an output based on example input-output pairs a method. Linear regression works on numerical attributes which means that input and output both are in numeric form. Regression model performs a regression task where a target value is getting predicted based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. If we have linear relation between input and output, then change in input affects the output.

The mathematical formulation of linear regression is given below:

Y = mX + c

where Y = dependent variable,
X = independent variable,
m = slope between X and Y,
c is the constant also known as Y intercept

The pictorial linear of linear regression is given below:

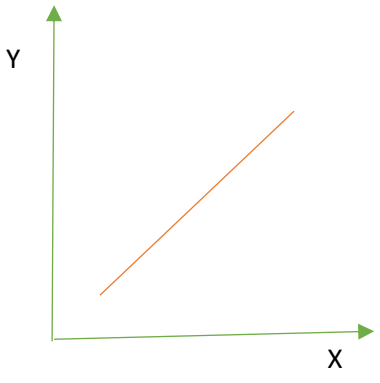


Fig 1: Positive Correlation

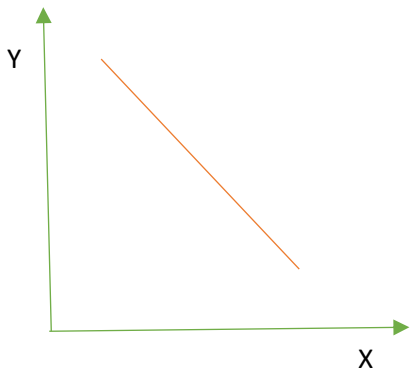


Fig 2: Negative Correlation

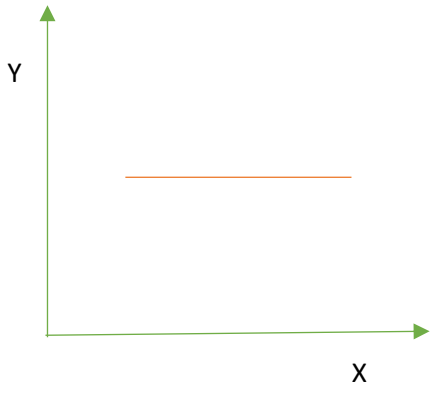


Fig 3: No Correlation

Positive Correlation: If increase in X causes increase in Y then the relation is called positive correlation. (Fig 1)

Negative Correlation: If increase in X causes decrease in Y then the relation is called negative correlation. (Fig 2)

No Correlation: If increase in X doesn't change the value of Y then the relation is calles no correlation. (Fig 3)

Types of Linear Regression: There are 2 types of linear regression:

1. **Simple Linear Regression:** When we use a single independent variable to get the output then it is called simple linear regression.

   Y = β0 + β1X

2. **Multiple Linear Regression:** When we use a more than one independent variable to get the output then it is called multiple linear regression.

   Y = β0 + β1X1 + β2X2 + β3X3 + ……+ βnXn where n = number of independent variables

Assumptions of Linear Regression: Below are the assumptions of the linear regression:

1. **Linear Relationship:** The relationship between the independent and dependent variables should be linear. This can be tested using scatter plots.
2. **Normal Distribution:** All the error terms should be normally distributed, means it should be centered to 0. This can be tested by plotting a histogram.
3. **Multicollinearity:** There shouldn't be multicollinearity in the data. Multicollinearity happens when the independent variables are highly correlated with each other. Multicollinearity can be tested by calculating the VIF.
4. **No Autocorrelation:** There shouldn't be autocorrelation in the data. Autocorrelation means single column data values are related to each other. In other words, f(x+1) is dependent on value of f(x). Autocorrelation can be tested with scatter plots.
5. **Homoscedasticity:** Homoscedasticity is followed. In other words residuals are equal across regression line. There shouldn't be any visible patterns in the error terms. Homoscedasticity can also be tested using scatter plot.

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

Answer 2: In most of the cases we get the dataset in the spreadsheet where we can apply some formula or keys to get the statistical summary of the data like mean, median, percentage, etc. At high level this statistical data provides some insights but, it is completely different in the real world. In that scenario we can't rely on the statistical data all the time. To explain this issue, Anscombe's quartet was introduced.

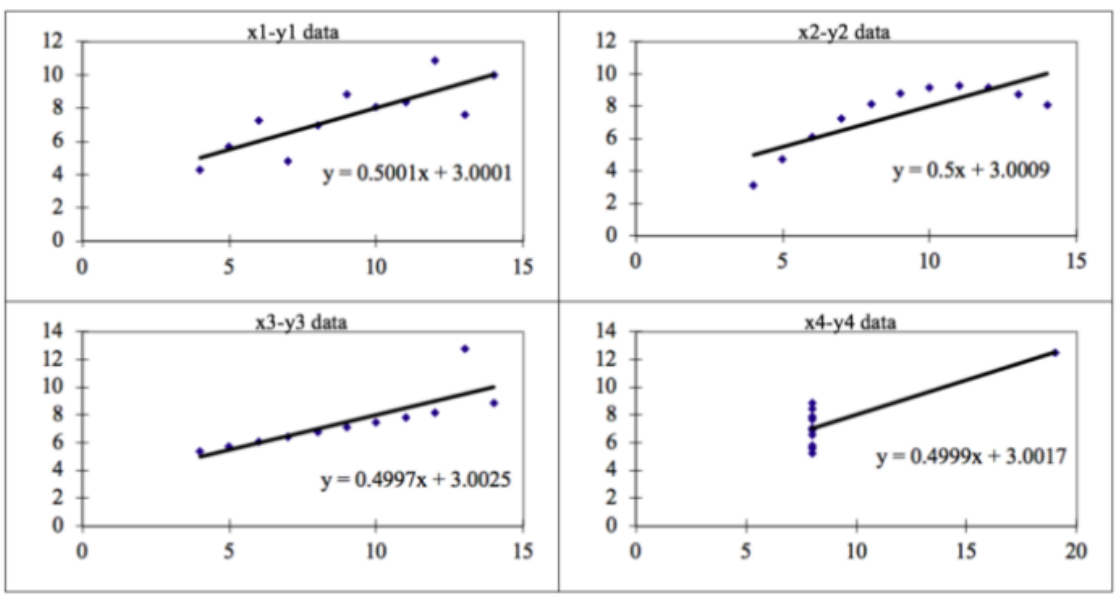
Anscombe's Quartet was constructed in 1973 by Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. It can be defined as a group of four data sets each containing exactly eleven (x, y) pairs which are nearly identical in simple **descriptive statistics**, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Below is the dataset to describe it in detail:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The statistical information for all these four datasets is approximately similar and can be computed as follows:

- The average value of x and y is 9 and 7.50 respectively for each dataset
- The standard deviation of x and y is 3.16 and 1.94 respectively for each dataset.
- The correlation between x and y is 0.82 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



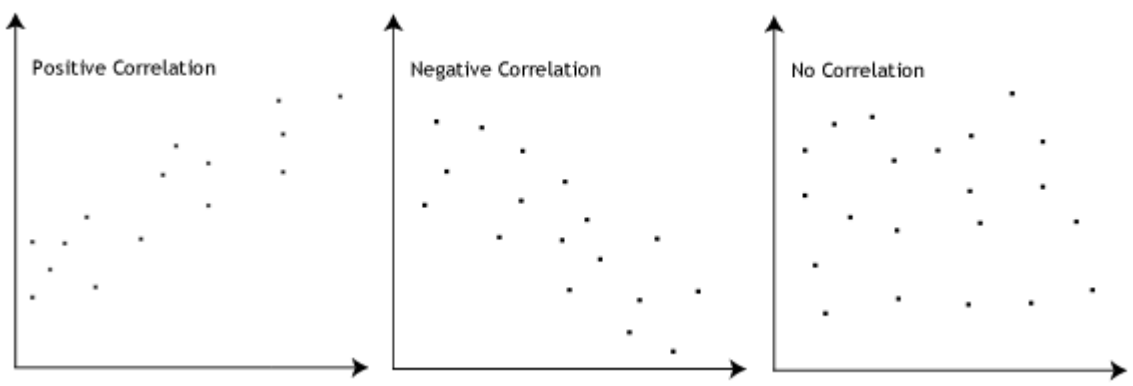
The four datasets can be described as:

- Dataset 1: This fits the linear regression model well.
- Dataset 2: This could not fit linear regression model because the data is non-linear.
- Dataset 3: This distribution is linear, but the outliers can't be handled by linear regression model.
- Dataset 4: It shows the greater number of outliers involved in the dataset which cannot be handled by linear regression model.

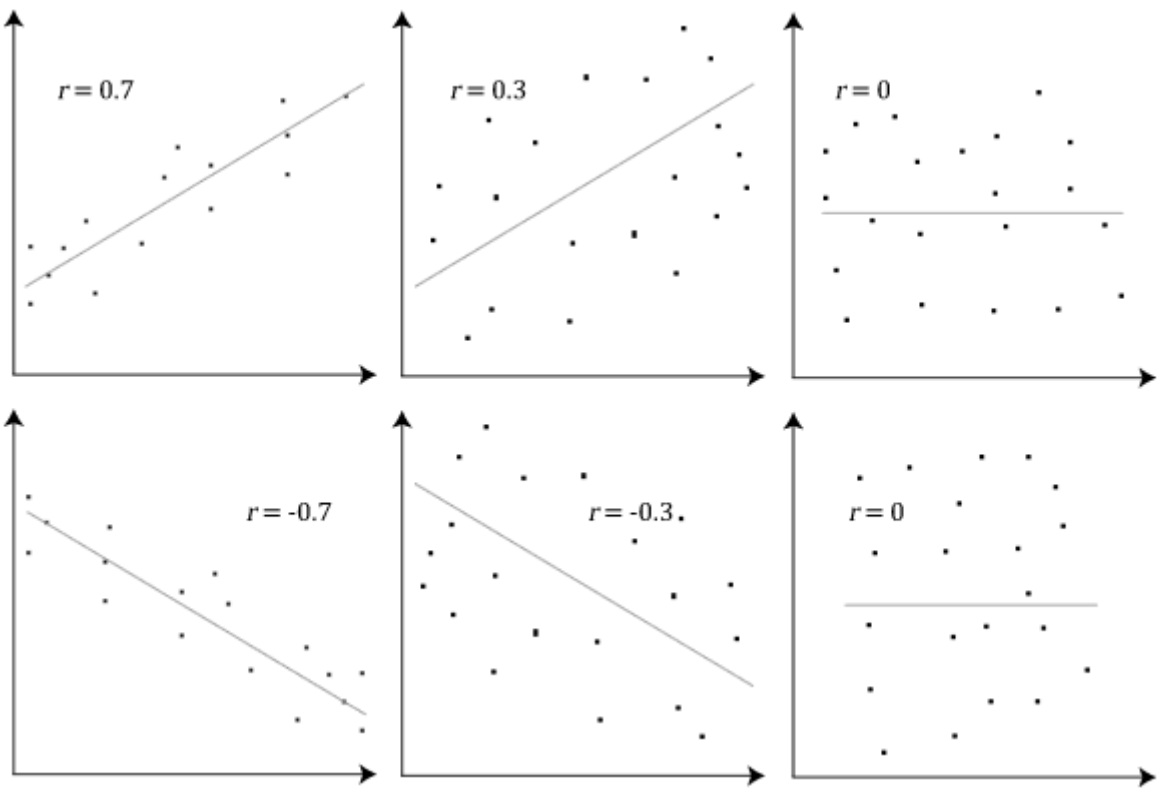
This quartet emphasizes the importance of visualization in data analysis. Looking at the graphs and data reveals a lot of the structure and a clear picture of the dataset.

Question 3: What is Pearson's R?

Answer 3: The Pearson's R or Pearson Correlation Coefficient is a measure of the strength of a linear correlation between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).



The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



To calculate the Pearson product-moment correlation, one must first determine the covariance of the two variables in question. Next, one must calculate each variable's standard deviation. The correlation coefficient is determined by dividing the covariance by the product of the two variables' standard deviations. Below is the formula to calculate it:

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Based on the r value, we can define the association. Below are the guidelines which can be used to whether an association is strong or not:

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

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

Answer 4: Scaling is a technique to standardize the independent featured available in the dataset. In most of the cases we get a greater number of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. It also impacts the system performance if we have datasets with different scales.

This step is performed during the data pre-processing to handle highly varying magnitudes or units. If the feature scaling is not done, then a machine learning algorithm tends to weigh greater values as higher and consider smaller values as lower values regardless of the unit of the values. This also affects in interpreting the model. One can interpret that higher coefficient variable is contributing more to the model which is not correct. So, we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

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-square, etc.

We can scale the features using two very popular method:

1. **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one. Below is the formula used:

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

2. **MinMax Scaling/Normalization:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

Difference between normalization and standardization:

S.NO.	Normalization	Standardization
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	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.
6	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7	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?

Answer 5: If we get this situation where VIF is infinite then it means a **perfect correlation**. Or in other words, we can say that both the features are identical. In such scenarios, we need to drop one feature before proceeding with the model building. VIF is calculated based on the R square(R2) value.

$$VIF = \frac{1}{(1 - R^2)}$$

In case of perfect correlation, $R^2 = 1$

$$VIF = \frac{1}{(1 - 1)}$$

$$VIF = \frac{1}{0}$$

$$VIF = \infty$$

We can see from the above calculation, for the perfect correlation, VIF will be infinite.

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

Answer 6: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. It 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.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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 of Q-Q plot:

1. It can be used with sample sizes
2. 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.

Q-Q plot is getting used on two datasets to check the below points:

1. If both come from populations with a common distribution.
2. If both have common location and scale.
3. If both have similar distributional shapes
4. If both have similar skewness or different.
5. If both have similar tail behavior.