

Linear Regression- Bike Case Study Assignment



Shashi Khurana

10/1/2023

Assignment-based Subjective Questions

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

Best Fit Equation is as follows. (Categorical values in bold for reference)

$$\text{Cnt} = 0.2343 \times \text{yr} + 0.48 \times \text{temp} + 0.0615 \times \text{summer} + 0.0972 \times \text{winter} + 0.0557 \times \text{working day} - 0.1501 \times \text{Windspeed} - 0.0817 \times \text{cloudy} - 0.2902 \times \text{Rainy} - 0.0551 \times \text{Spring} + 0.1405$$

Based on the analysis following was inferred.

- Count of total rental bikes (count) in year 2019 is 23.43% higher than that in 2018.
- Unit increase in temperature increases count by 48%
- Count is 9.7% higher in Winter season.
- Count is 29.02 % lower when there is light snow or rain.
- Count is 8.1 % lower when the weather is misty and cloudy.
- Unit increase in windspeed decreases count by 15%
- Count is 5.5% more on working day.
- Count is 5.5% lower in Spring season.

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

Using this it will drop the first category of each categorical variable, create k-1 dummy variables for each categorical variable and perform dummy encoding. It is important to use, as it helps in reducing the extra column created during dummy variable creation. Therefore, reduces the correlations created among dummy variables.

In our assignment we end up Creating Dummy Variables for following categorical variables

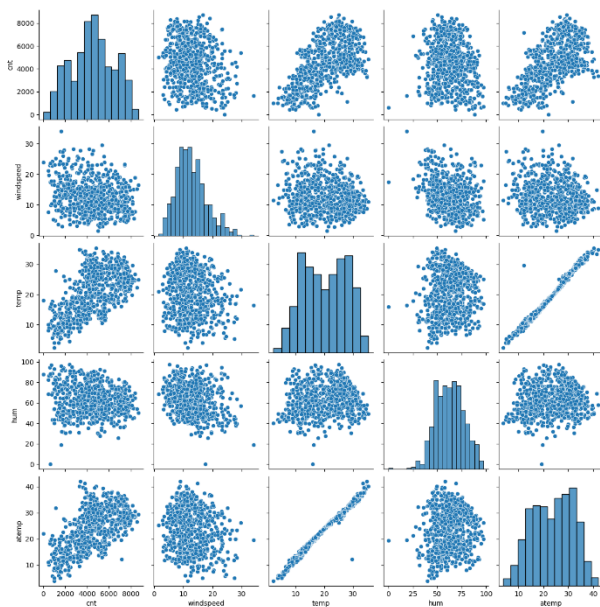
```
season_dummy = pd.get_dummies(bike['season'], drop_first = True)
```

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

```
weathersit_dummy = pd.get_dummies(bike['weathersit'], drop_first = True)
```

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

Highest correlation is between temp and target variable. (One thing to note here is temperature and feel like temperature has similar correlation with Target variable) both temperature and atemp are highly correlated that is evident from the pair plot.

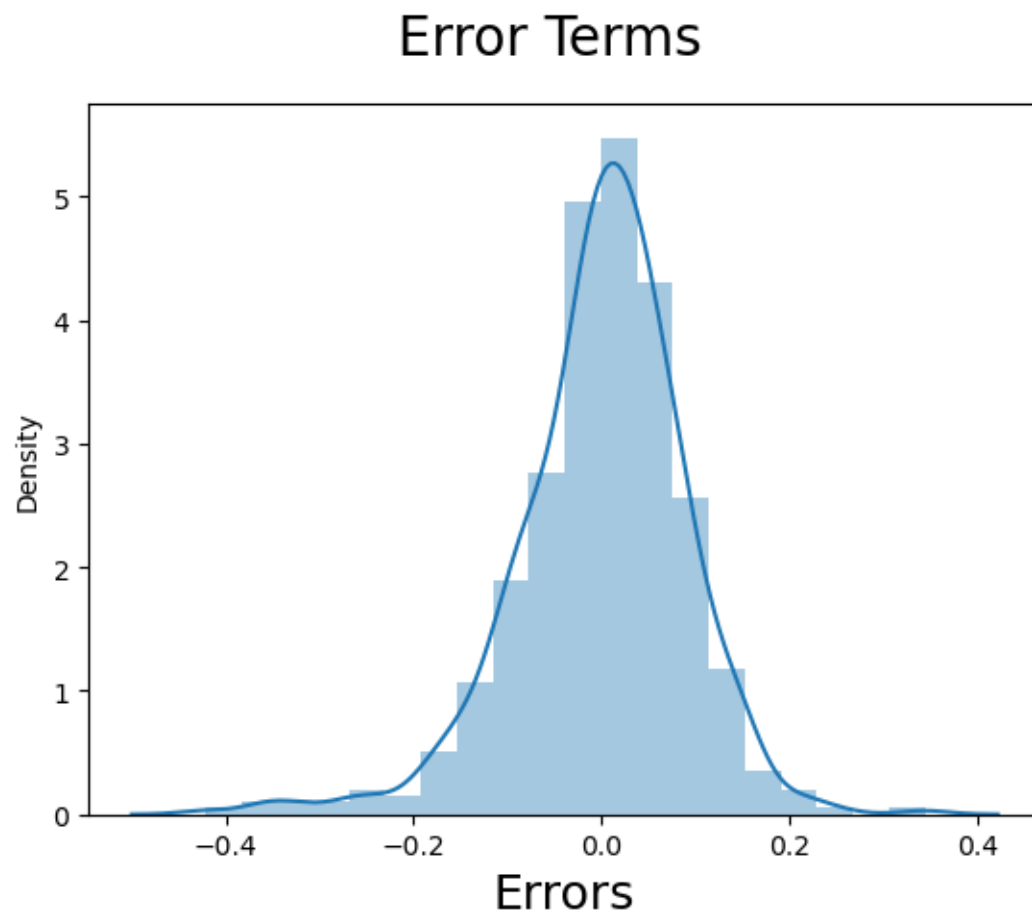


Q. How did you validate the assumptions of Linear Regression after building the model on the training set?

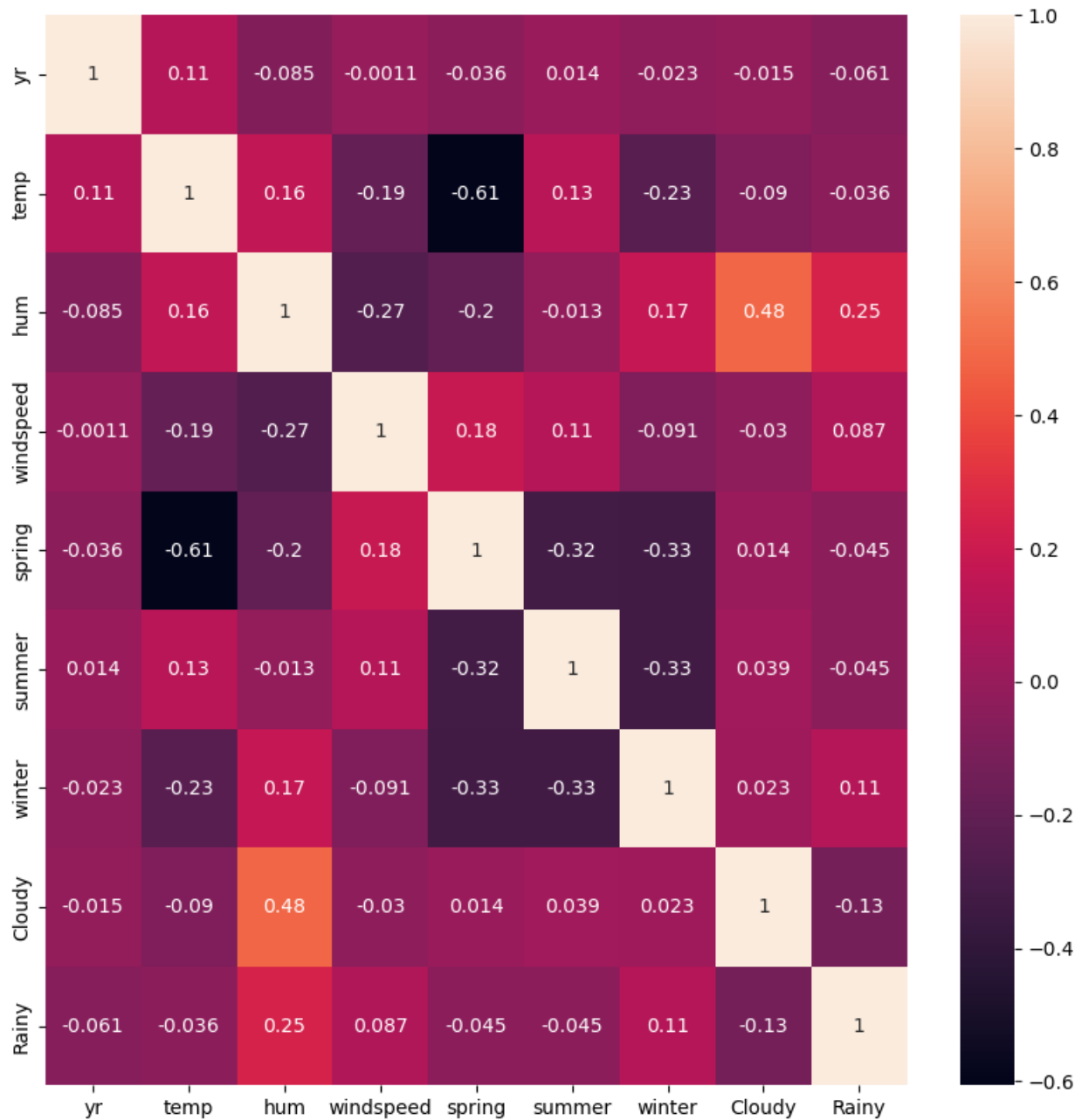
1. Error terms are normally distributed. - Residual analysis

Following is the histogram plot for the error terms. we should check whether the residuals/errors are normally distributed (after having fitted the model)

if they are not the inference may be misleading.

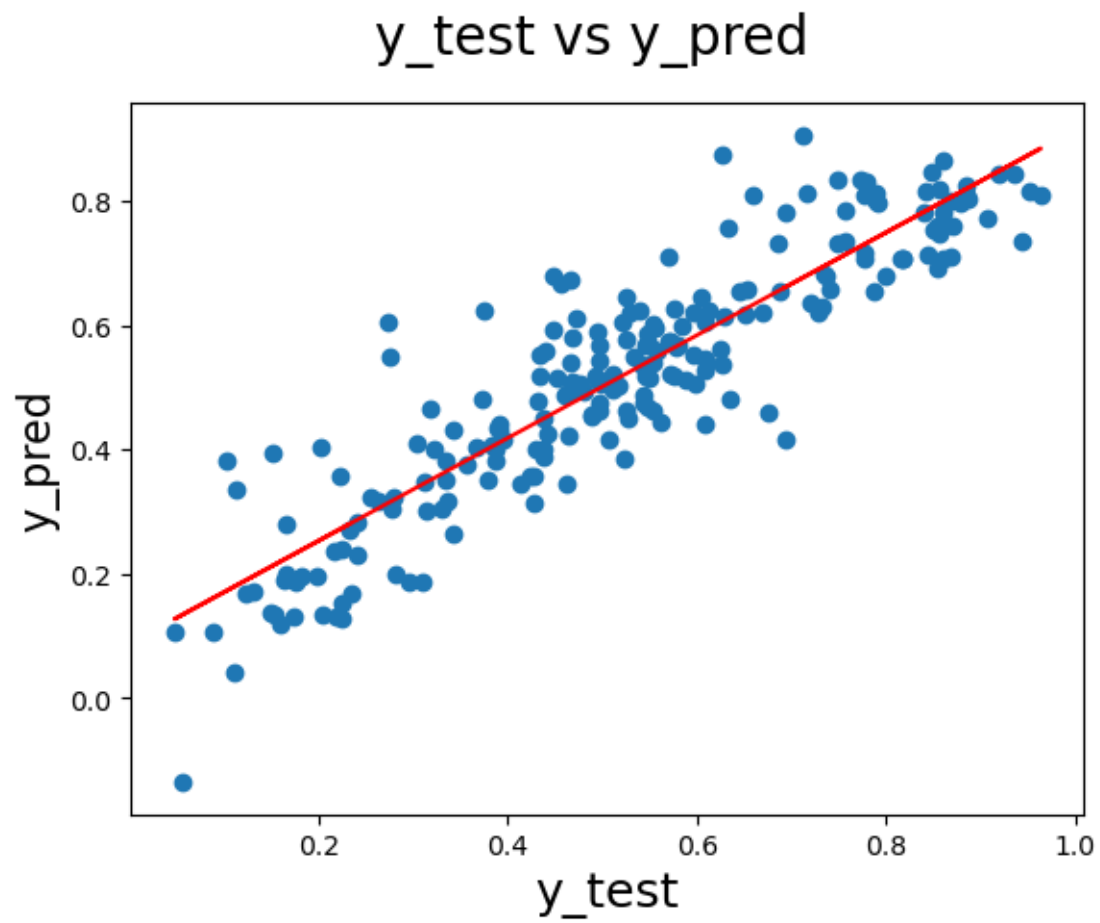


2. Minimizing correlation between independent variables.





3. Linear relationship exists between X and Y.



Following scatter plot is developed using test data that predicts the y value based on the test dataset and is almost linear.



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

Following features are contributing significantly towards explaining the demand of shared bikes. (Positive and negative)

Positive Contributors (Features) for Bike demand	Temperature 
	Year 

Negative Contributors (Features) for Bike demand	Light Snow or rain(Weather Situation – Rainy) 
	Windspeed 

Best Fit Equation is as follows.

$$\text{Cnt} = 0.2343 \times \text{yr} + 0.48 \times \text{temp} + 0.0615 \times \text{summer} + 0.0972 \times \text{winter} + 0.0557 \times \text{working day} - 0.1501 \times \text{Windspeed} - 0.0817 \times \text{cloudy} - 0.2902 \times \text{Rainy} - 0.0551 \times \text{Spring} + 0.1405$$

Summary / inference for top 4 features effecting demand either way.

- Unit increase in temperature increases count by 48%
- Count of total rental bikes (count) in year 2019 is 23.43% higher than that in 2018.
- Count is 29 % lower when there is light snow or rain.
- Unit increase in windspeed decreases count by 15%

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear Regression

In the simplest words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: **Simple** and **Multiple**.

Simple Linear Regression is where only one independent variable is present and the model must 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 b_0 is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_0 + b_1x$$

Equation of Multiple Linear Regression, where b_0 is the intercept, $b_1, b_2, b_3, b_4, \dots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4, \dots, x_n$ and y is the dependent variable.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots + b_nx_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.

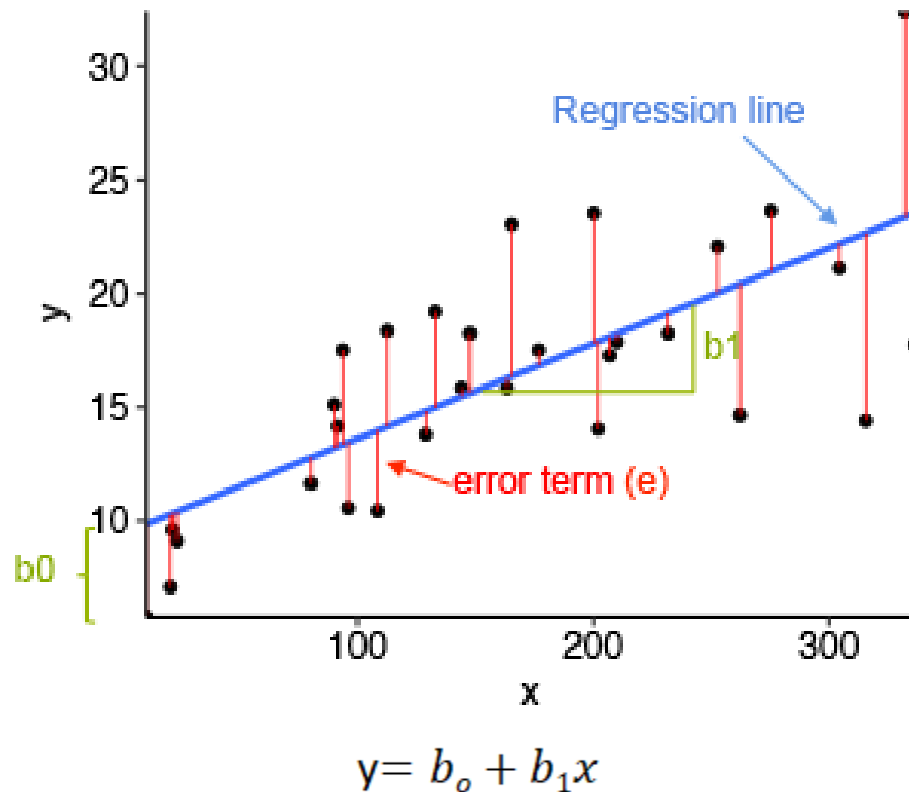


Image Source: Statistical tools for high-throughput data analysis

In the above diagram,

- x is our independent variable which is plotted on the x -axis and y is the dependent variable which is plotted on the y -axis.
- Black dots are the data points (actual values)
- b_0 is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:

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

i.e

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

Rsq, AdjRsq, MSE, RMSE, MAE – 5 evaluation metrics

Assumptions of Linear Regression –

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

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. If the assumption is violated, then the points will form a funnel shape otherwise they will be constant.

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. If the VIF score is greater than 5 then the variables are highly correlated.

The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.

No Autocorrelation: The error terms ($y_{act} - y_{pred}$) should be independent of each other. Autocorrelation can be tested using the **Durbin Watson test**. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

They were constructed in 1973 by the Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

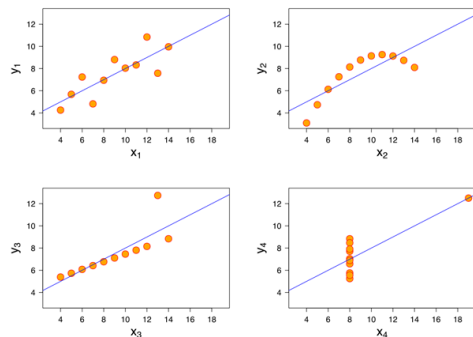
This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well, but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

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

Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

Q3. What is Pearson's R?

Pearson's R was developed by **Karl Pearson** and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula:

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

Where:

N	=	number of pairs of scores
$\sum xy$	=	sum of the products of paired scores
$\sum x$	=	sum of x scores
$\sum y$	=	sum of y scores
$\sum x^2$	=	sum of squared x scores
$\sum y^2$	=	sum of squared y scores

Example:

- Statistically significant relationship between age and height.
- Relationship between temperature and ice cream sales.
- Relationship among job satisfaction, productivity, and income.

- Which two variables have the strongest co-relation between age, height, weight, size of family and family income.

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Formula of Standardized scaling:

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

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

The value of VIF is calculated by the below formula:

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

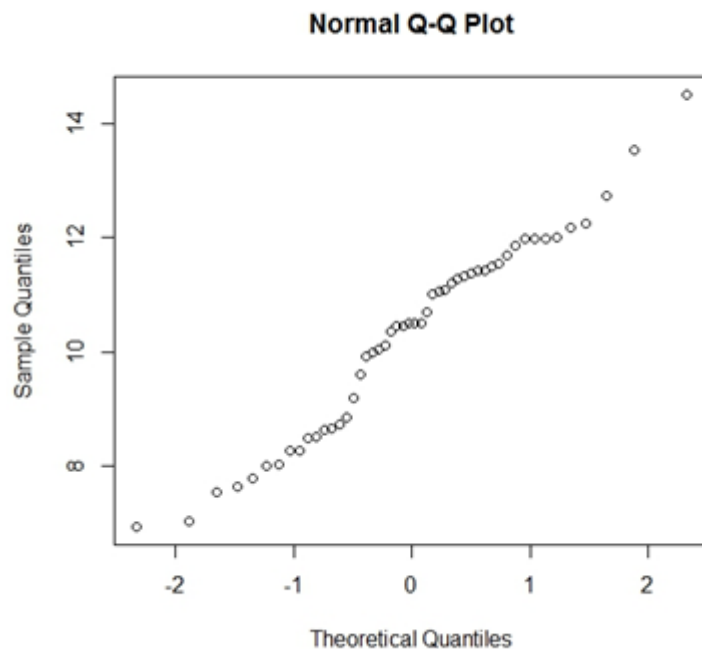
Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

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

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

- I. The sample sizes do not need to be equal.
- II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- III. The q-q plot can provide more insight into the nature of the difference than analytical methods.