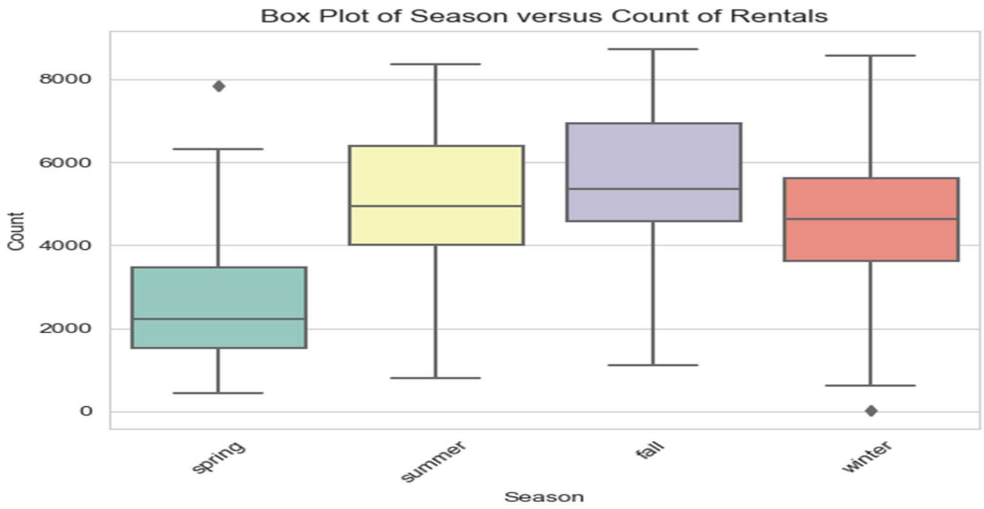
Assignment-based Subjective Questions

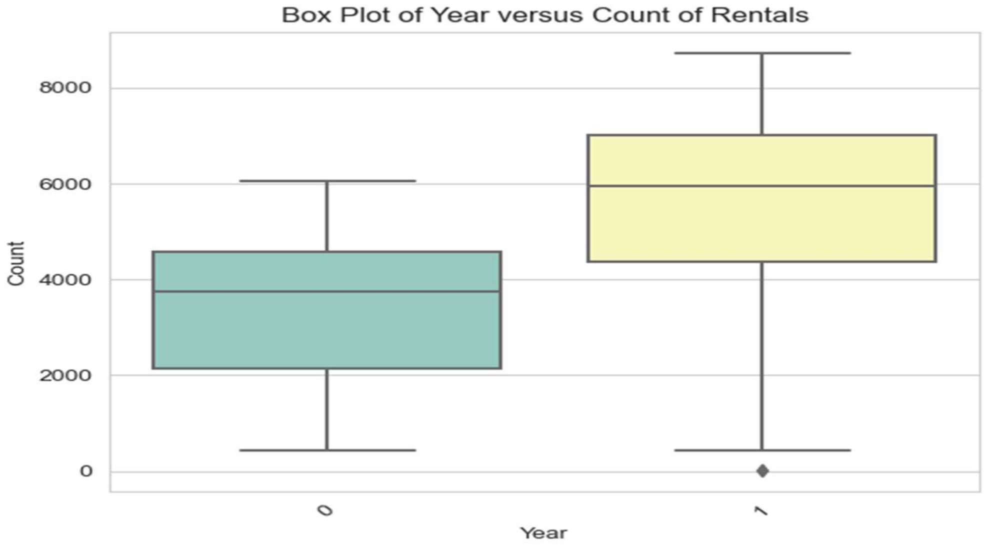
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   The categorical variables are ‘season’, ‘year’, ‘month’, ‘holiday’, ‘weekday’, ‘workingday’ and ‘weathersit’. These categorical variables have a major effect on the dependent variable ‘count’.

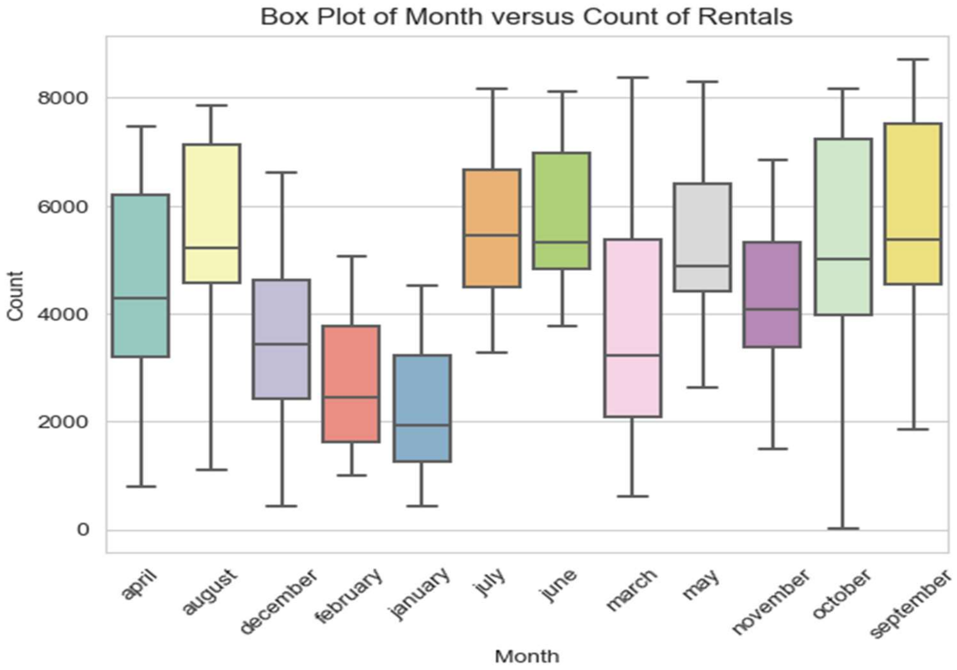
   - **Season:** Rentals peak in the Fall, followed closely by the Winter and Summer season.



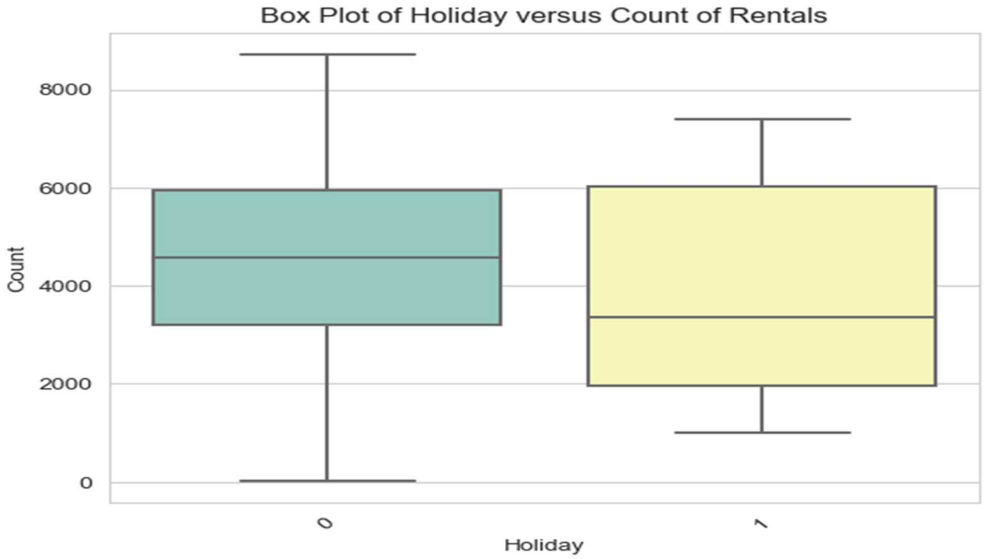
   - **Year:** There is a notable increase in the number of rentals in 2019 compared to 2018.



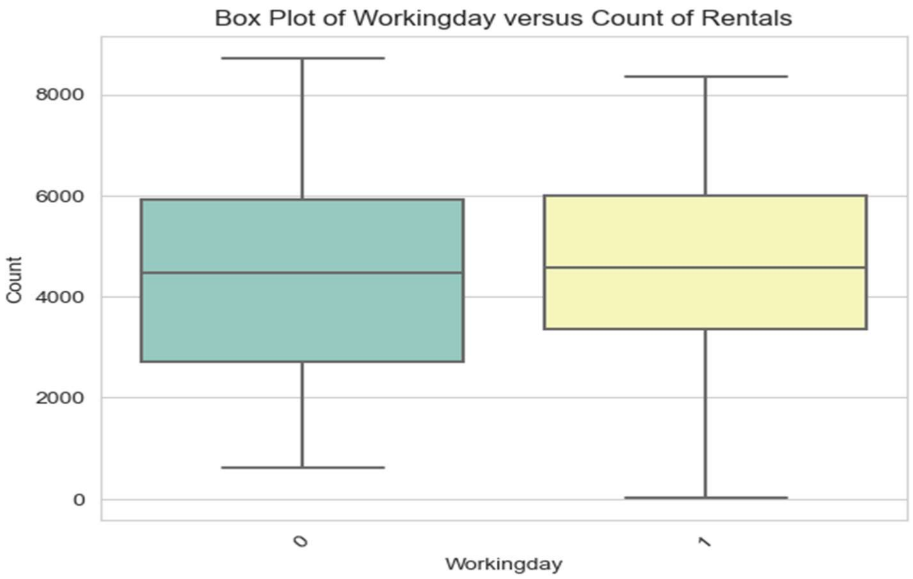
- **Month:** September sees the highest number of rentals.



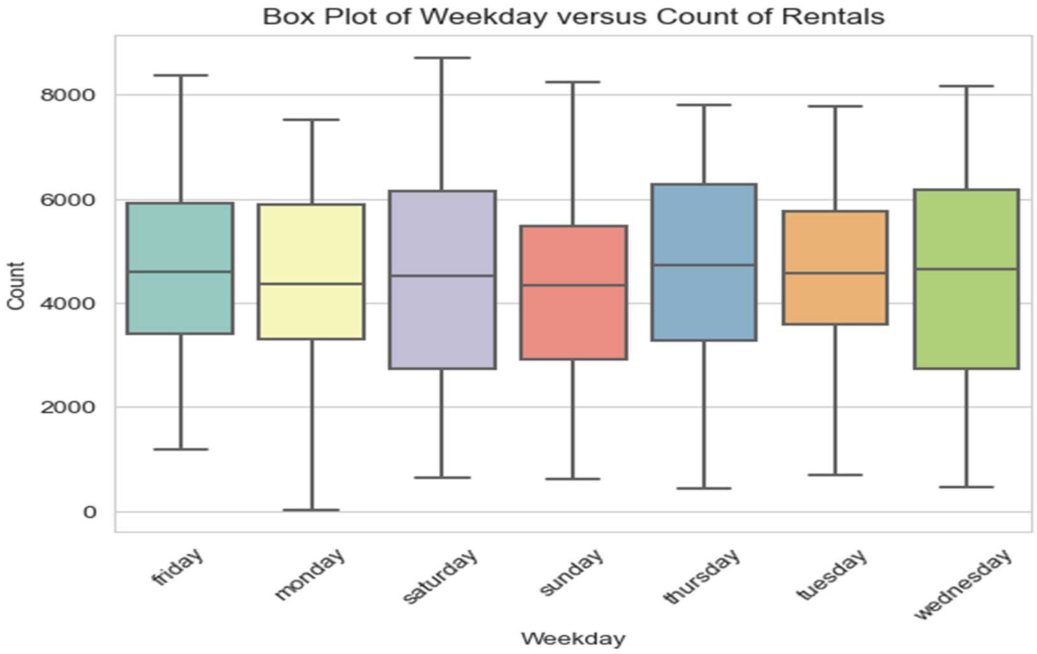
- **Holiday**: Rentals decrease on holidays.



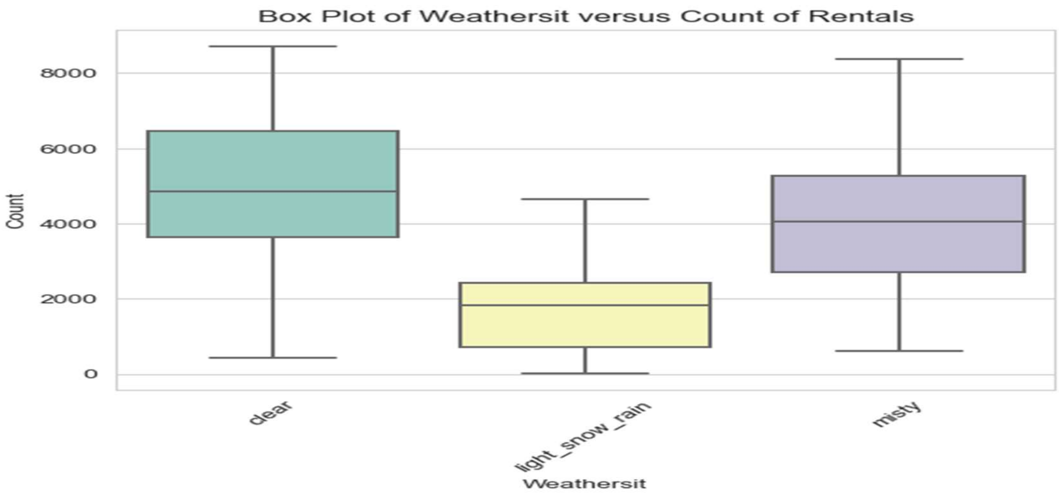
- **Working Day**: Rentals increase on working days.



- **Weekday:** There is no significant variation in the number of rentals across different weekdays.



- **Weather Situation:** Rentals are highest during clear weather conditions.



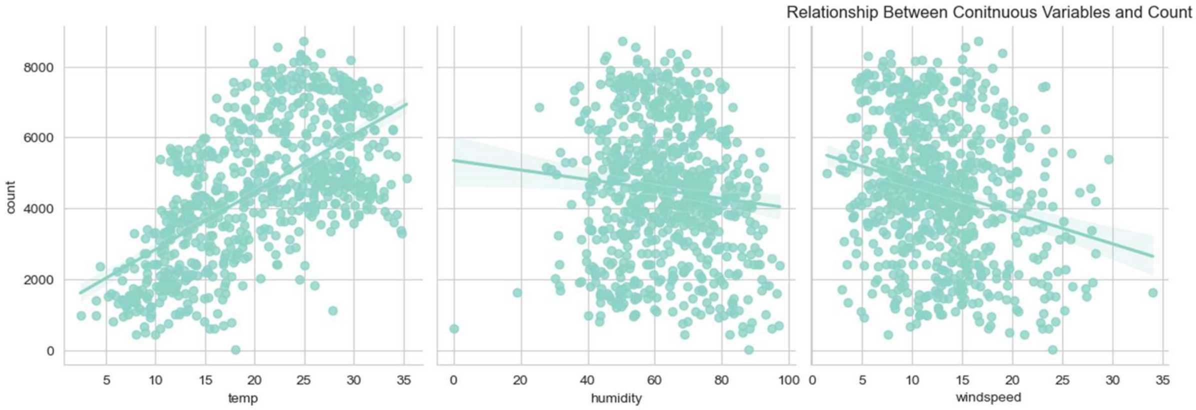
2. **Why is it important to use drop_first=True during dummy variable creation?**

   It is important to use drop_first=True during dummy variable creation to avoid multicollinearity.
   By setting drop_first=True when creating dummy variables, you drop one of the 'n' dummy variables. This dropped variable serves as the reference category or baseline, against which the other categories are compared.
   This approach:
   - **Prevents Multicollinearity:** Dropping one dummy variable avoids the problem of perfect multicollinearity, as the remaining dummy variables will not be perfectly collinear with each other. This helps ensure that the model's estimates are stable and interpretable.
   - **Simplifies Interpretation:** The coefficient of each remaining dummy variable represents the effect of that category relative to the reference category. This makes it easier to interpret the influence of each category on the dependent variable.

   Example:
   Consider a categorical variable Colour with three levels: Red, Blue, and Green. If you create dummy variables for each colour, you'll end up with three columns: Red, Blue and Green. If you include all three in your model, you'll face multicollinearity because:
   The sum of Red, Blue, and Green will always be 1.
   By setting drop_first=True, you might drop Green, leaving you with:
   Red
   Blue

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

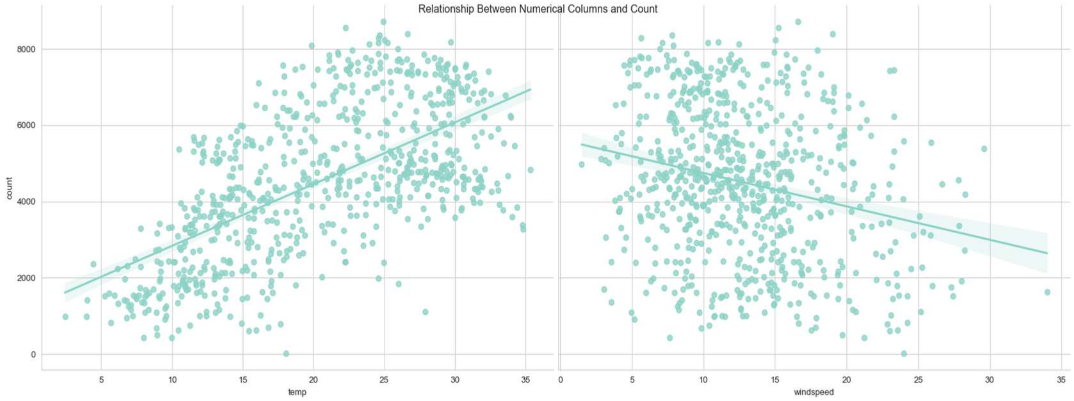
   By looking at the pairplot 'temp' have the highest correlation with the target variable 'count'. **temp** is having the correlation value of 0.63 with the target variable 'count'.



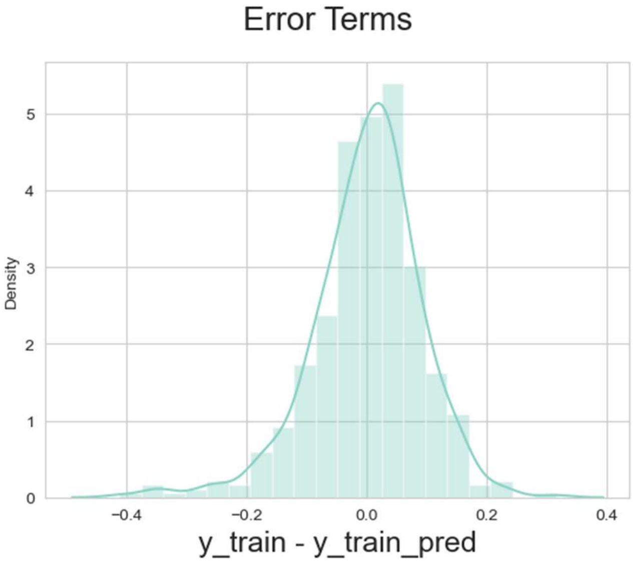
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

The assumption of Linear Regression Model has been validated based on

- **There is a linear relationship between X and Y.**
  By plotting a regression plot between the predictors and target variable 'count'.
  The below graph shows there is linear relationship between numerical columns temp and windspeed with the target variable.

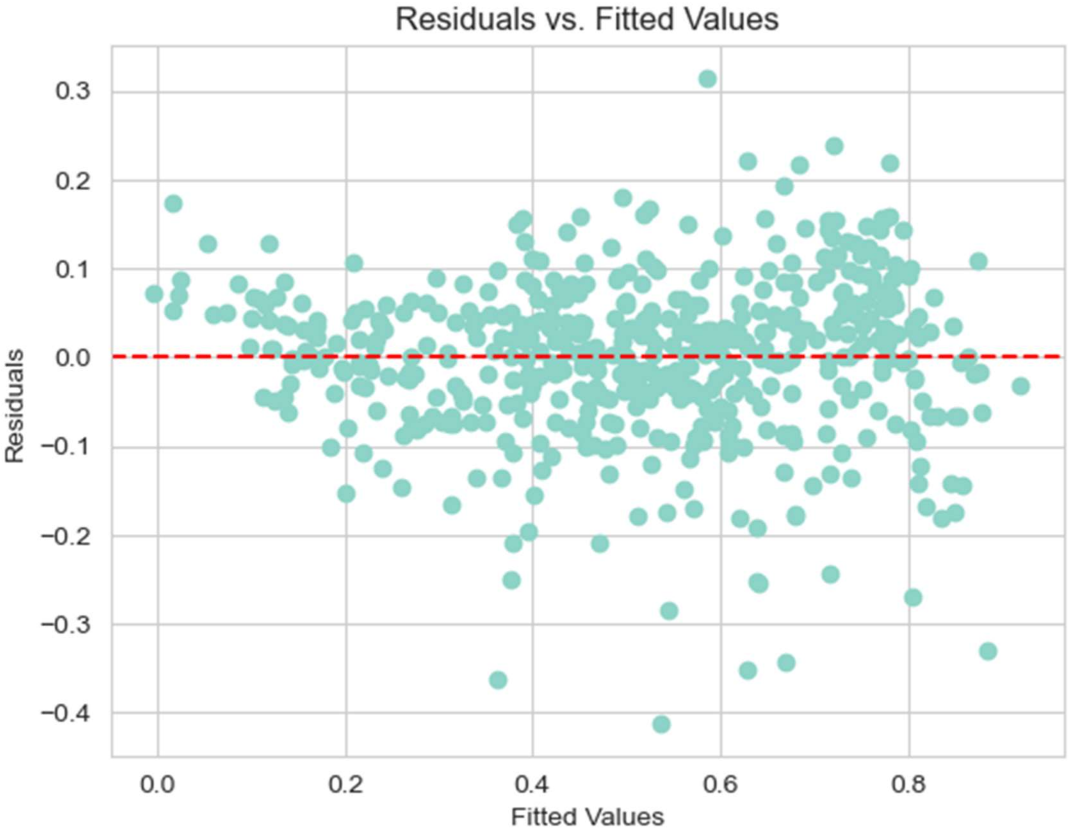


- **Error terms are normally distributed (not X, Y).**
  By plotting the histogram of the error terms we can validate that the error terms are normally distributed.



- **Error terms are independent of each other**
  Durbin-Watson value of final model lr_5 is 2.019 which signifies there is no autocorrelation.

- **Error terms have constant variance (homoscedasticity)**
  By plotting a scatter plot of Residuals vs. Fitted Values (Predicted Values) we can validate this assumption.



- **There is No Multicollinearity between the predictor variables.**
  The VIF calculation indicates that multicollinearity is not an issue among the predictor variables, as all VIF values are within the acceptable range of below 5.

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

Top 3 features that has significant impact towards explaining the demand of the shared bikes the top three predictor variables influencing bike bookings are:

- **Temperature:** With a coefficient of 0.4917, a unit increase in temperature is associated with an increase of 0.4917 units in bike bookings.
- **Year:** With a coefficient of 0.2339, a unit increase in the year variable is associated with an increase of 0.2339 units in bike bookings.
- **Weather Situation:** Light_snow_rain with a coefficient of -0.2847, a unit increase in the light_snow_rain variable is associated with a decrease of 0.2847 units in bike bookings. Misty with a coefficient of -0.0802, a unit increase in the misty variable is associated with a decrease of 0.0802 units in bike bookings.

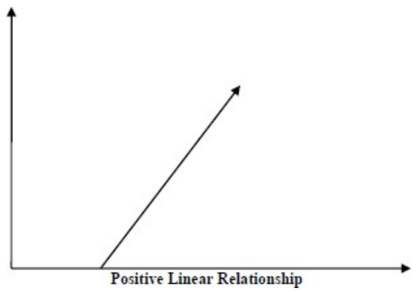
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

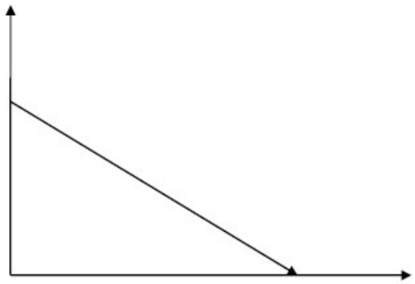
Furthermore, the linear relationship can be positive or negative in nature as explained below:

- **Positive Linear Relationship**: A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

- **Negative Linear relationship:** A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Linear regression is of the following two types −

- **Simple Linear Regression:**

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.
Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y= b_o + b_1x$$

- **Multiple Linear Regression:**

Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
Equation of Multiple Linear Regression, where b_o is the intercept, $b_1,b_2,b_3,b_4…,b_n$ are coefficients or slopes of the independent variables $x_1,x_2,x_3,x_4…,x_n$ and y is the dependent variable.

$$y= b_o + b_1x_1 + b_2x_2 + b_3x_3 \dots . +b_nx_n$$

Assumptions -

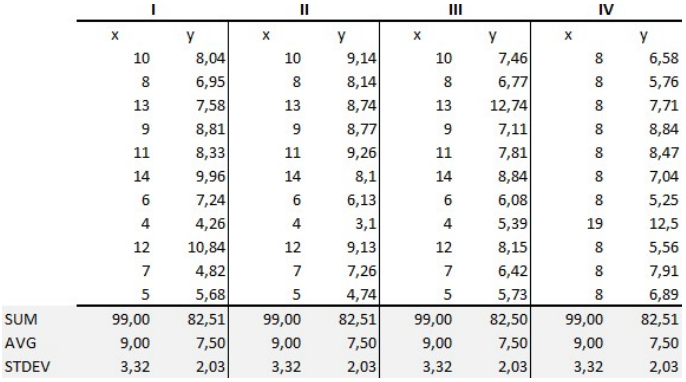
The following are some assumptions about dataset that is made by Linear Regression model −

- **Multi-collinearity** - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- **Auto-correlation** - Another assumption of Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

- **Relationship between variables** - Linear regression model assumes that the relationship between response and feature variables must be linear.
- **Normality of error terms** - Error terms should be normally distributed.
- **Homoscedasticity** - There should be no visible pattern in residual values.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

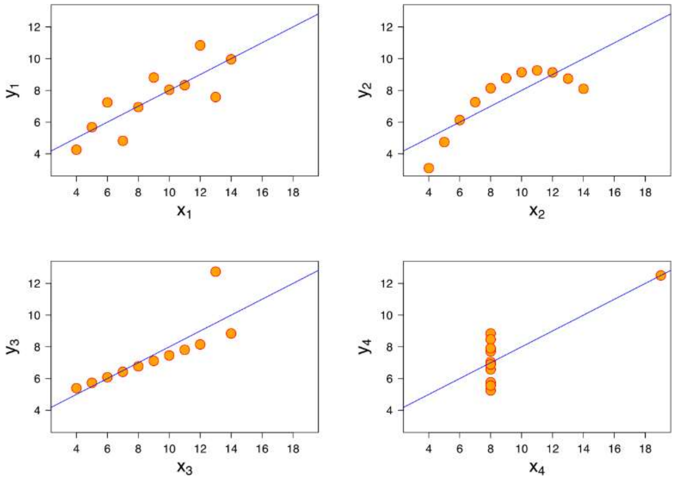


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 appears 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 anoutlier.
- 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.

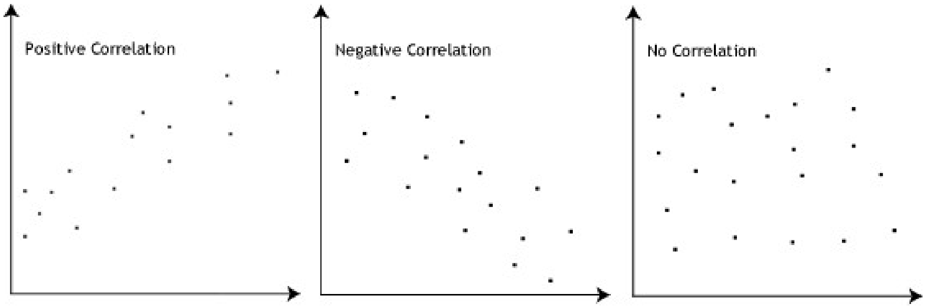
3. What is Pearson’s R?

Pearson's r measures how strongly two variables are related in a linear way.

- **Positive Correlation (+r):** If both variables increase or decrease together, the correlation coefficient is positive.
- **Negative Correlation (-r):** If one variable increases while the other decreases, the correlation coefficient is negative.
- **No Correlation (0):** If there's no clear pattern or relationship between the variables, the correlation coefficient is zero.

Pearson's r ranges from +1 to -1:

- **+1:** Perfect positive relationship (both variables move in the same direction).
- **-1:** Perfect negative relationship (one variable goes up as the other goes down).
- **0:** No relationship between the variables.



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

Scaling is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables.

Scaling is performed for the following reason:

- It ensures numerical stability and efficiency which can leads to better the model performance
- In algorithms like gradient descent, scaling can help speed up the convergence.

- Prevents features with larger scales from dominating the model's behaviour, ensuring all features contribute equally.
- Helps algorithms that are sensitive to the initial values of the parameters.

Most of the time, datasets contain features with values that vary widely in size, units, and range. If we don't scale the data, algorithms might focus only on the size of the values, ignoring their units, which leads to incorrect modelling. To fix this, we need to scale the data so all features have similar magnitudes.

It's important to note that scaling only affects the coefficients of the model. It doesn't change other parameters like t-statistics, F-statistics, p-values, or R-squared.

Difference between normalized scaling and standardized scaling

Normalized Scaling	**Standardized Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

If the value of VIF is infinite, it indicates that there is perfect multicollinearity in the dataset. It occurs when one predictor variable is an exact linear combination of one or more other predictor variables.

Perfect Correlation: When a predictor variable is perfectly correlated with other predictors, the variance of its residuals (the error term when regressing this variable on the other predictors) is zero.

Division by Zero: VIF is calculated as:

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

where R-square is the coefficient of determination from regressing the predictor variable on all other predictor variables.
If R^2=1 (indicating perfect correlation), the denominator becomes zero, making the VIF value infinite.

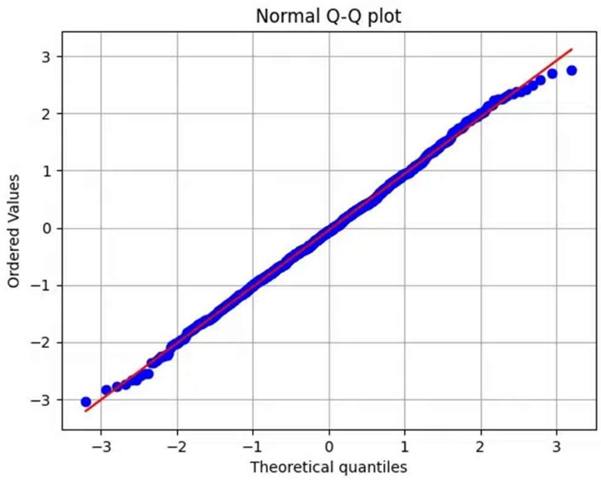
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A quantile-quantile (Q-Q) plot is a type of graph that helps us see if a dataset fits a certain distribution, like a normal distribution, or if two sets of data come from the same group. It's like a visual test to check if our data matches what we expect. Q-Q plots are often used in statistics and data analysis to make sure our assumptions about the data are correct and to spot any differences from the expected pattern.

Steps to draw Q-Q plot:

- **Collect the Data:** Gather your numerical dataset that you want to analyse.
- **Sort the Data:** Arrange the data in order from smallest to largest.
- **Choose a Theoretical Distribution:** Decide which theoretical distribution you want to compare your data to, like the normal distribution.
- **Calculate Theoretical Quantiles:** Find the quantiles for your chosen theoretical distribution. For example, if using a normal distribution, use its inverse cumulative distribution function to get the expected quantiles.
- **Plotting:**
  - Put the sorted data values on the x-axis.
  - Put the theoretical quantiles on the y-axis.
  - Each point on the plot shows an observed data value (x) and its expected value (y) from the theoretical distribution.
  - Draw a line or connect the points to see how well the data fits the distribution.

If the points on the plot fall roughly along a straight line, it means your data fits the expected distribution well. If the points deviate from the straight line, it means your data does not fit the expected distribution and you might need to investigate further.



The data points in the above Q-Q plot roughly form a straight line, it means that the dataset fits well with the assumed theoretical distribution, which in this case, is the normal distribution.

Advantages of Q-Q Plot:

1. Q-Q plots can compare datasets of different sizes without needing the same sample size.

2. They can compare datasets with different units or scales.
3. They provide a clear visual comparison of your data against a theoretical distribution.
4. They easily show deviations from the expected distribution, helping to identify data issues.
5. They help assess if data follows a distribution, identify outliers, and understand data patterns.