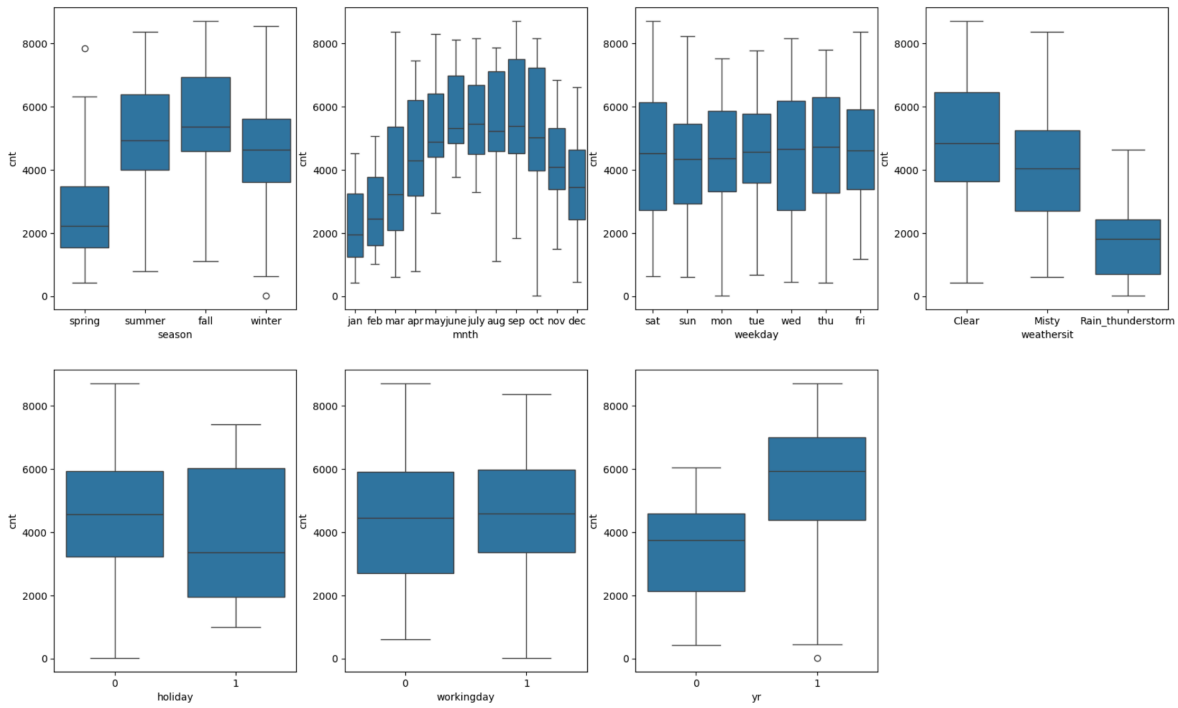
Assignment-based Subjective Questions

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

   Answer: Based on box plot which were plotted for categorical columns; following is the analysis

   a. Fall season seems to have attracted more booking.
   b. Most of the bookings has been done during the month of june, july, aug, sep and oct.
   c. No significant impact because of weekday
   d. Clear weather attracted more booking which seems obvious.
   e. Bookings are less on holidays as people won't travel to workplaces and hence less bookings
   f. Bookings in 2019 were much higher than 2018. This may be because people are getting aware of this.



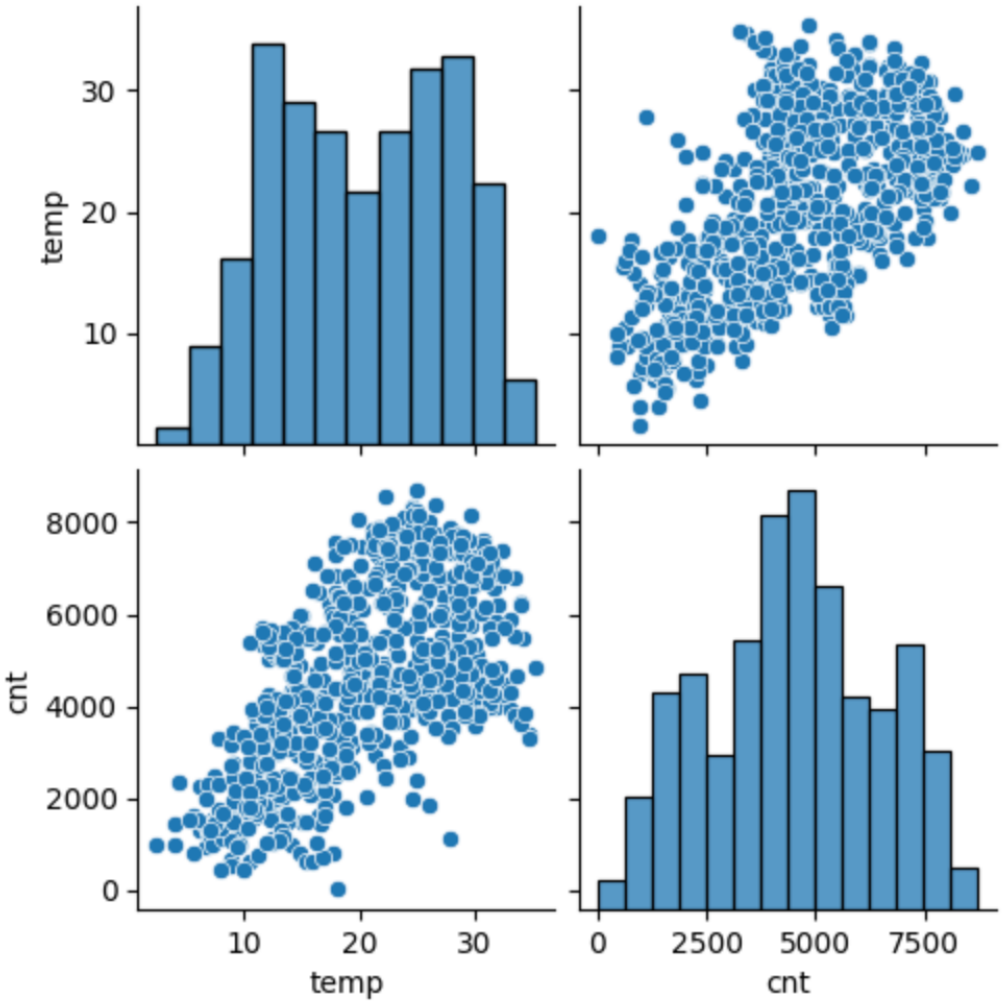
2. Why is it important to use **drop_first=True** during dummy variable creation?

Answer : drop_first= True is important to use, as it helps in reducing the extra column created during dummy variable creation.

For eg we have 4 types of values in a categorical column and we create dummy variable of that column . So we will have 4 combinations T F F F, F T F F, F F T F, F F F T so if 1 variable is F F F hence its true. Hence we can reduce 1 exrta column.

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

   Answer : temp variable has the highest correlation with the target variable

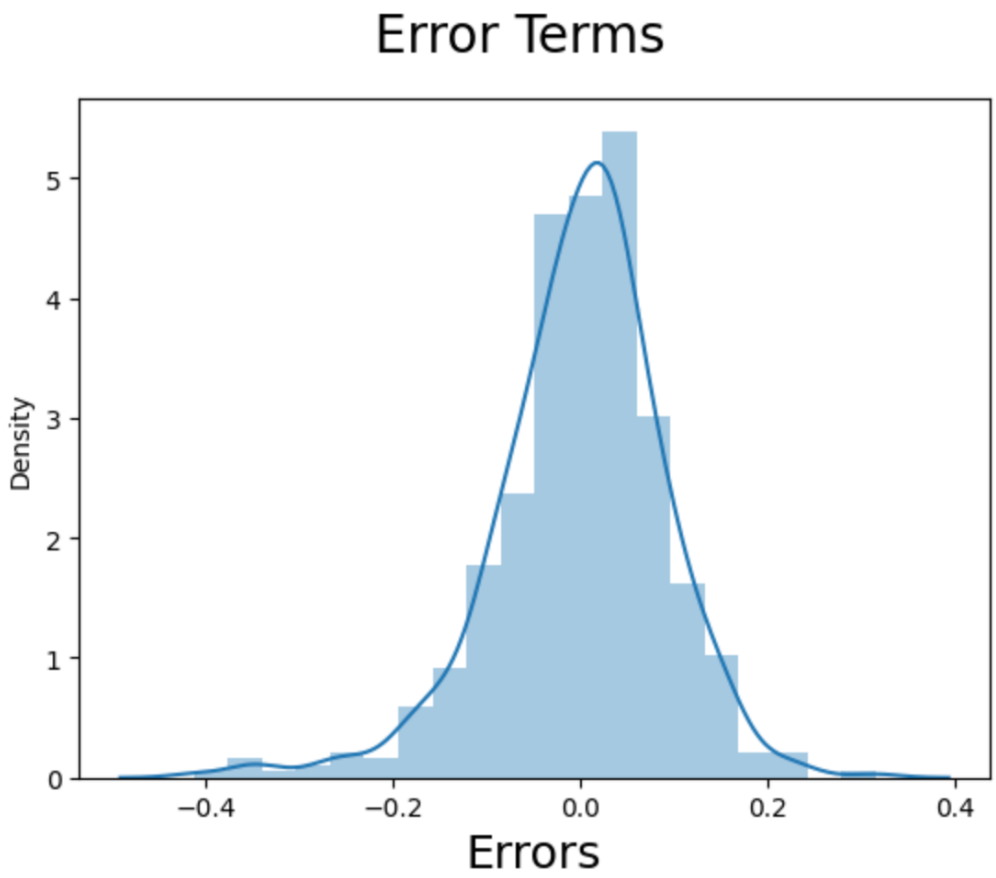


4. How did you validate the assumptions of Linear Regression after building the model on the

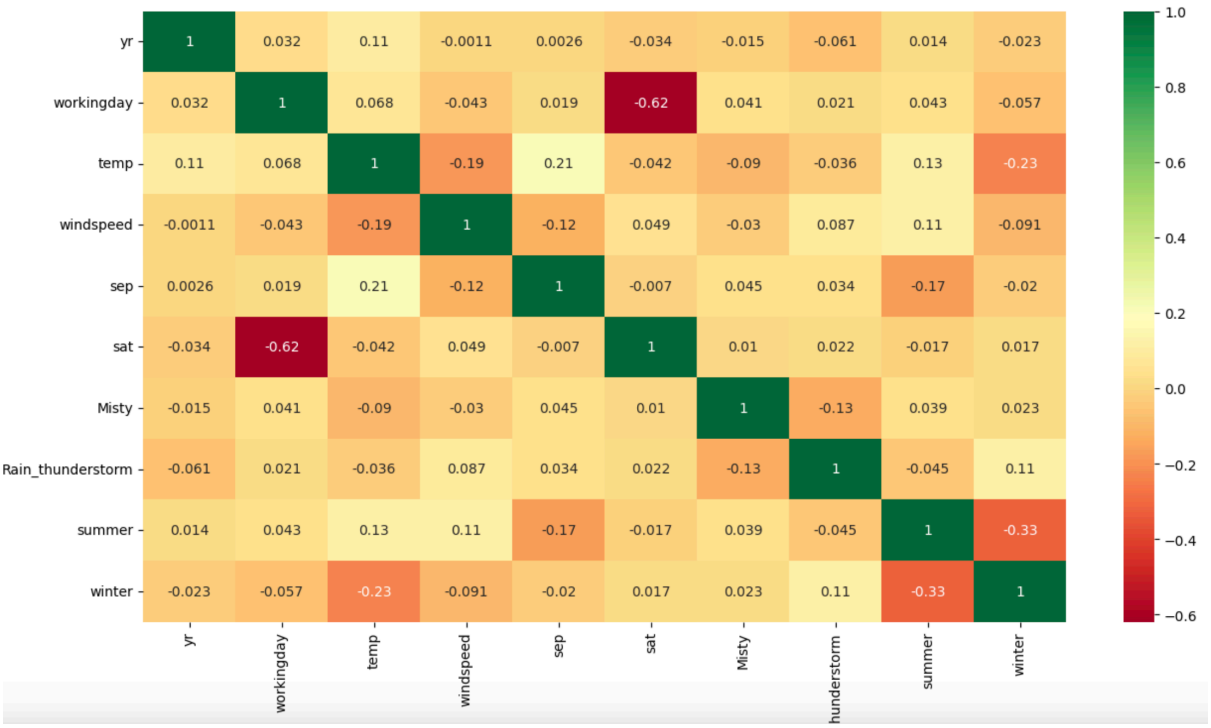
   training set?

   Answer: I have validated the assumption of linear regression by checking

   a. Normality of error terms



   b. Multicollinearity check



c. Homoscedasticity

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

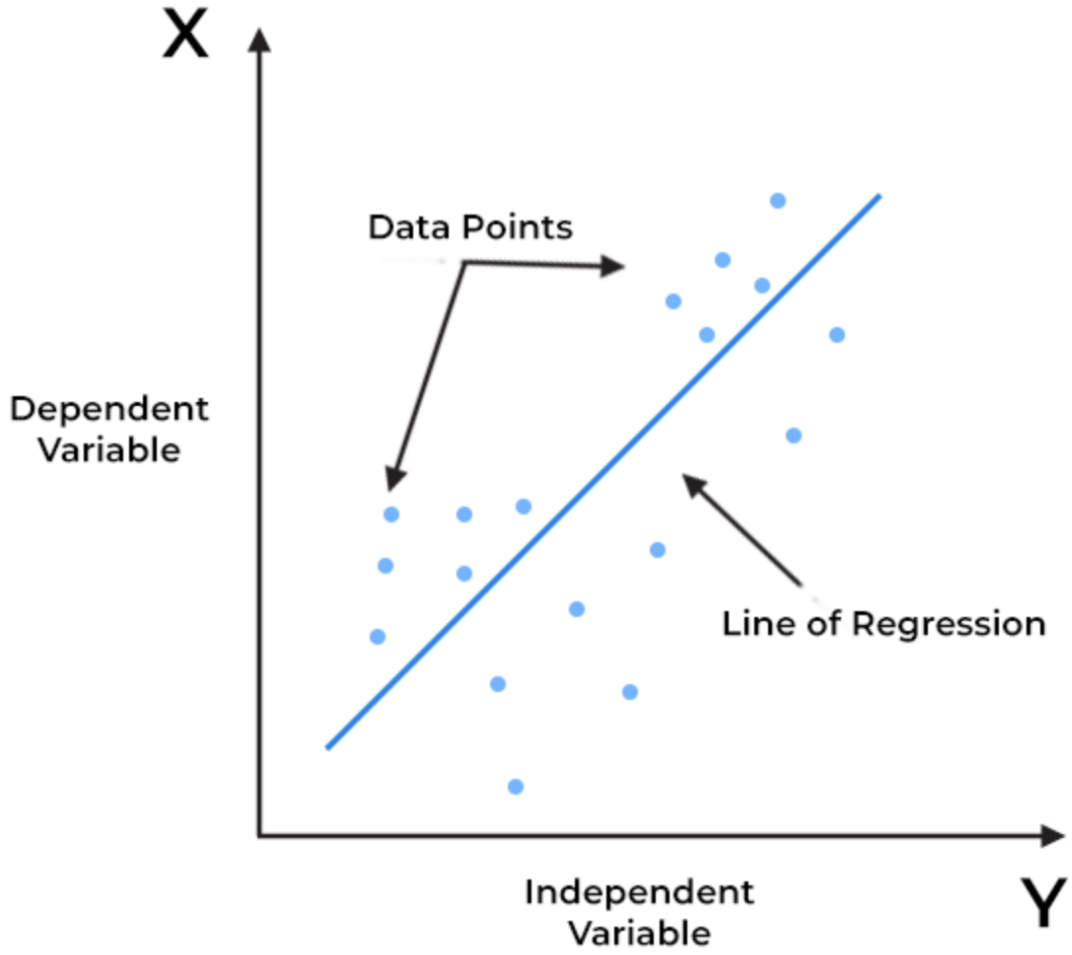
Answer: Top 3 features are as below

a. Temp
b. Working day
c. Winter

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer : Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis. The independent variable is the predictor variable that remains unchanged due to the change in other variables. However, the dependent variable changes with changes in the independent variable. Linear regression is a supervised learning algorithm A sloped straight line represents the linear regression model.



In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Mathematically the relationship can be represented as

Y = mx+c

Where Y is dependent variable ; x I independent variable

M = slope of regression line which represents effect of X on Y

C is the constant

Linear relationship can be positive or negative.

a. Positive: if dependent and independent variables increase simultaneously
b. Negative: If independent variable increases ; dependent decreases

Key assumptions about the data which is made by Linear Regression Model are as below

a. Linear relationship

The first important assumption of linear regression is that the dependent and independent variables should be linearly related.

b. Normal distribution of residuals

The second assumption relates to the normal distribution of residuals or error terms, i.e., if residuals are non-normally distributed, the model-based estimation may become too wide or narrow.

c. Multicollinearity

The third assumption relates to multicollinearity, where several independent variables in a model are highly correlated. More correlated variables make it difficult to determine which variable contributes to predicting the target variable. Also, standard errors inevitably increase

due to correlated variables.The goal, therefore, is to have minimal or lesser multicollinearity.

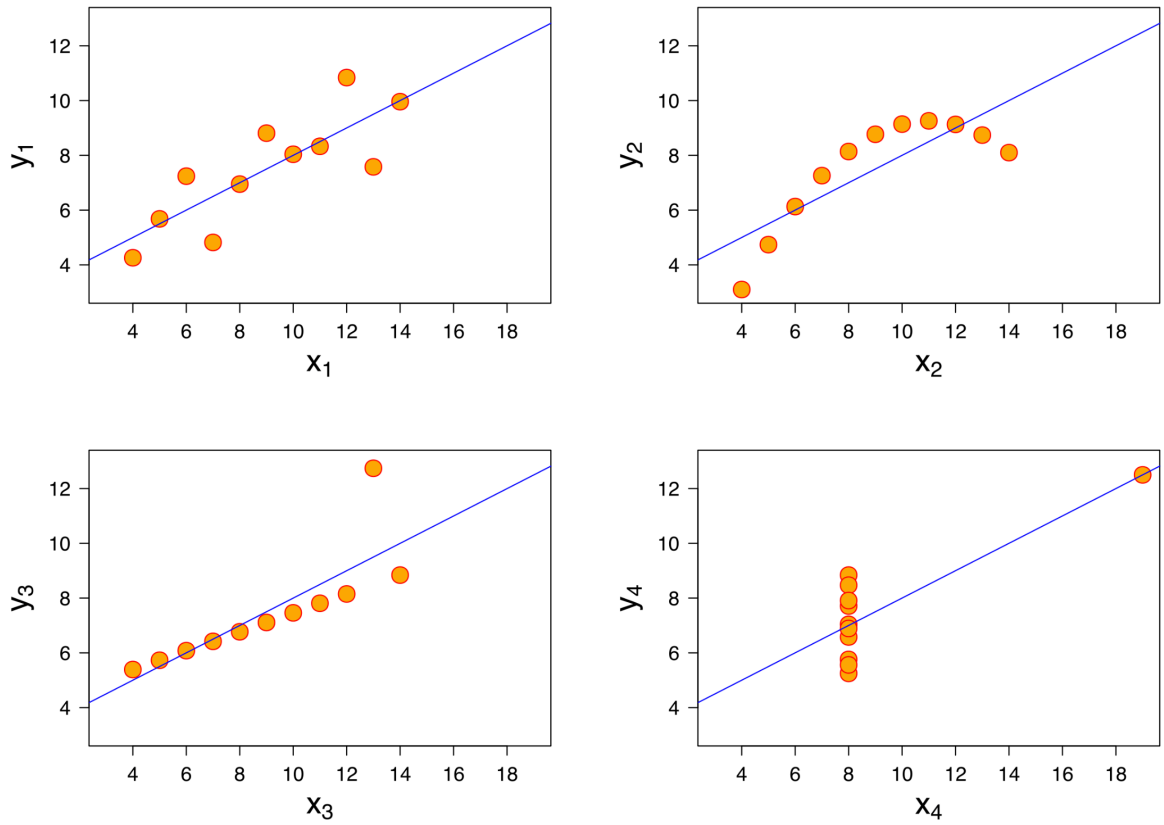
d. Homoscedasticity

Another assumption of linear regression analysis is referred to as homoscedasticity. Homoscedasticity relates to cases where the residuals (error terms) between the independent and dependent variables remain the same for all independent variable values.
In sother words, the residuals or error terms must have ‘constant variance.’

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

Answer: Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
It shows us that we should not blindly trust summary statistics or standard methods of analysis. It tells us to look closely at our data, question our assumptions, and use a variety of analytical tools to get a full picture



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables
-
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.
-
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
-
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Question 3. What is Pearson's R?

Answer : The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
0 to 1 means positive correlation. When one variable changes, the other variable changes in the **same direction**.

0 mean No correlation. There is **no relationship** between the variables.

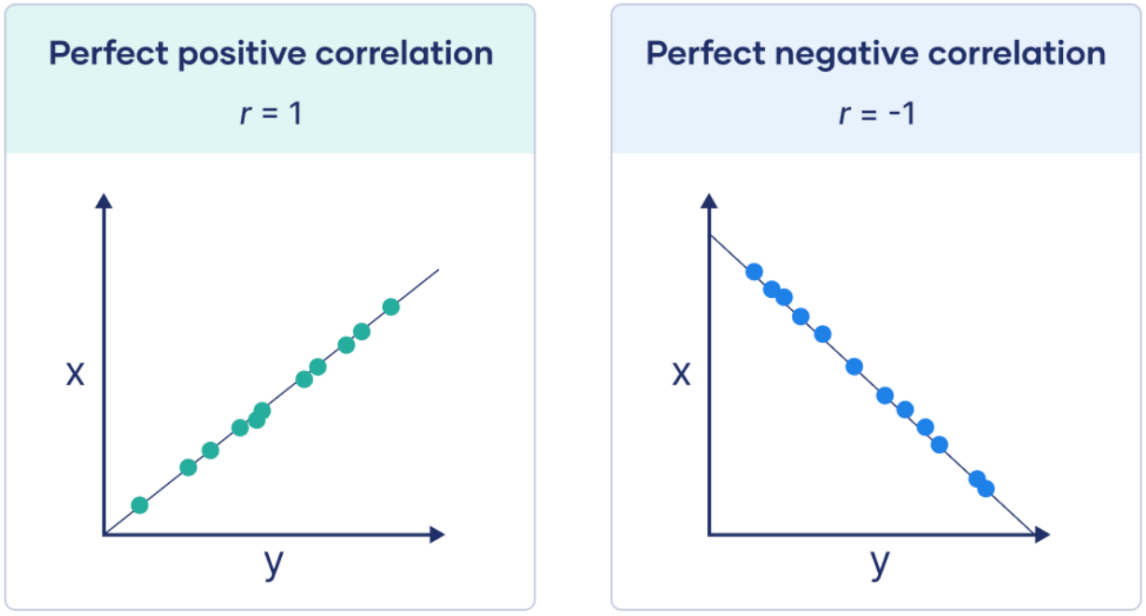
Between 0 and -1 means Negative correlation. When one variable changes, the other variable changes in the **opposite direction**.

When to use the Pearson correlation coefficient

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

- **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

- The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.
- When *r* is 1 or −1, all the points fall exactly on the line of best fit:



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

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. 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.

Why Scaling is Performed:

- Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of input features. For example, distance-based algorithms like k-nearest neighbors or gradient descent

optimization in linear models can be influenced by the magnitude of features.

- Convergence Speed: In iterative optimization algorithms, such as gradient descent, normalizing features can help the algorithm converge faster.
- Regularization: Regularization techniques, which penalize large coefficients, can be affected by the scale of features. Scaling helps in preventing one feature from dominating due to its larger magnitude.
- Interpretability: Scaling makes it easier to interpret the coefficients of features in linear models.

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. It is simple, preserves the shape of the original distribution, and is suitable when the data distribution is not Gaussian.

Xnormalised = (X- Xmin) / (Xmax -Xmin)

Standardization Scaling:

Standardized scaling transforms the values of a feature to have a mean of 0 and a standard deviation of 1.

sklearn.preprocessing.scale helps to implement standardization in python.

It is robust to outliers, makes the data comparable across different features, and is suitable for algorithms that assume Gaussian-distributed features.

Xstandardized = (X - mean(X))/std(X)

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

Answer: If there Is a perfect correlation ; then VIF = infinity . A large value of VIF indicates there is a correlation between variables

When the value of VIF is infinite it shows the perfect correlation between 2 independent variables. In case of perfect correlation, we get R2 = 1 which means 1/(1-R2) = infinity.

To solve this problem; we should drop one of the variables which is causing this multi collinearity.

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

Answer: 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.

Use of Q-Q Plot in Linear Regression:

Normality Check:

Q-Q plots are commonly used to assess whether the residuals (the differences between observed and predicted values) in a linear regression model follow a normal distribution.

In linear regression, the assumption of normally distributed residuals is essential for valid statistical inference, including hypothesis testing and confidence interval estimation.

Comparison with Theoretical Distribution:

The Q-Q plot compares the quantiles of the observed residuals against the quantiles of a theoretical normal distribution.

If the points on the Q-Q plot roughly follow a straight line, it suggests that the residuals are approximately normally distributed.

Importance of Q-Q Plot in Linear Regression:

Assumption Checking:

- Linear regression assumes that the residuals are normally distributed with constant variance. Deviations from this assumption can affect the accuracy of hypothesis tests and confidence intervals.
- Q-Q plots provide a visual tool to assess the normality assumption of the residuals.

Identification of Outliers:

- Q-Q plots can help identify potential outliers in the residuals. Outliers might appear as points deviating significantly from the expected straight line on the Q-Q plot.

Decision Making:

- When deciding whether to rely on parametric statistical tests or make certain inferences about the coefficients, it's crucial to consider the normality of residuals. Q-Q plots help in this decision-making process.

Model Improvement:

- If the Q-Q plot suggests non-normality, it may indicate areas for model improvement. This could involve transforming variables or exploring alternative modeling approaches.