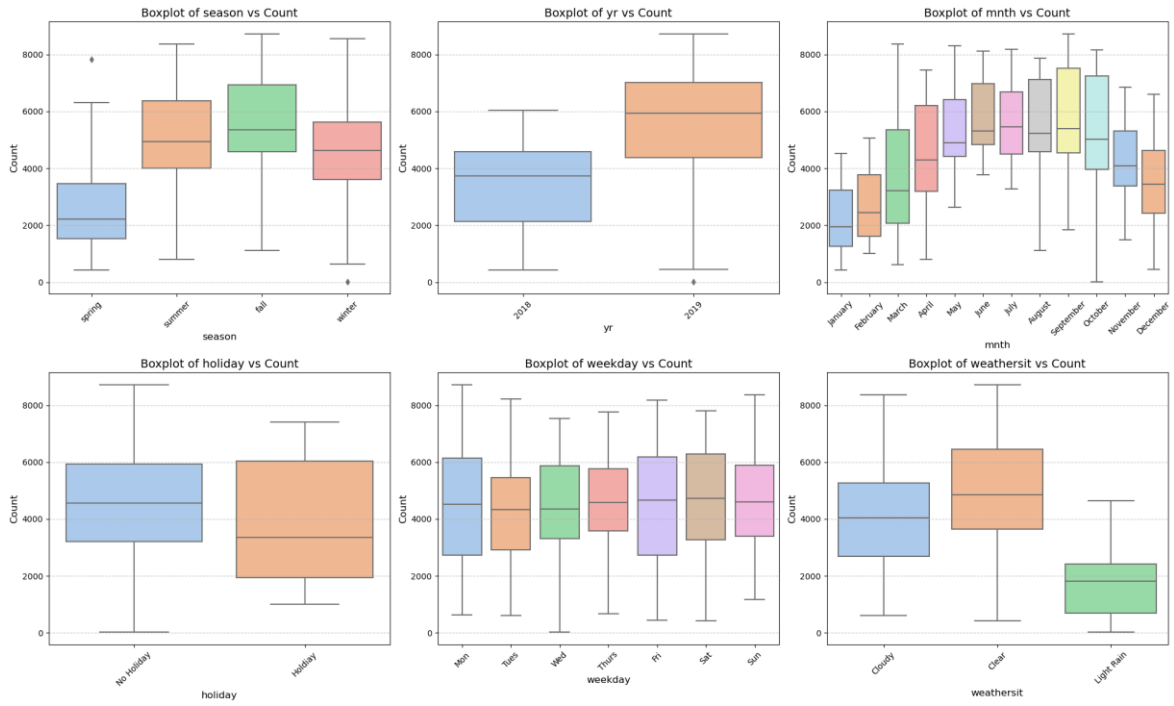
Assignment-based Subjective Questions:

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

Below are Inferencing about the dependent variable and categorical variables:

- **Seasons and Months**: We see that people tend to rent more bikes in certain seasons, like summer and fall, when the weather is warmer and more pleasant. In contrast, during spring and winter, when it's colder, the number of bike rentals decreases. This pattern is consistent when we look at individual months. For instance, from January to June, the number of bike rentals steadily increases, reaching its peak around August and September, before dropping again during the colder months.
- **Holidays vs. Regular Days**: On holidays, like Christmas or New Year's Day, fewer people rent bikes compared to regular days when there's no holiday. This could be because people are busy with holiday activities or spending time with family and friends, rather than biking.
- **Weather Conditions**: The type of weather also affects bike rentals. On days with clear skies and good weather, more people tend to rent bikes for outdoor activities or commuting. However, on rainy days or when the weather is cloudy, fewer people are inclined to rent bikes, likely because they prefer to stay indoors or use other modes of transportation.

Below, you'll find subplots visualizing the relationships between the categorical variables (excluding 'workingday') and 'cnt':



In summary, these factors—seasons, holidays, and weather conditions—play a significant role in determining bike rental patterns. Understanding these relationships helps us predict bike rental demand more accurately and plan accordingly.

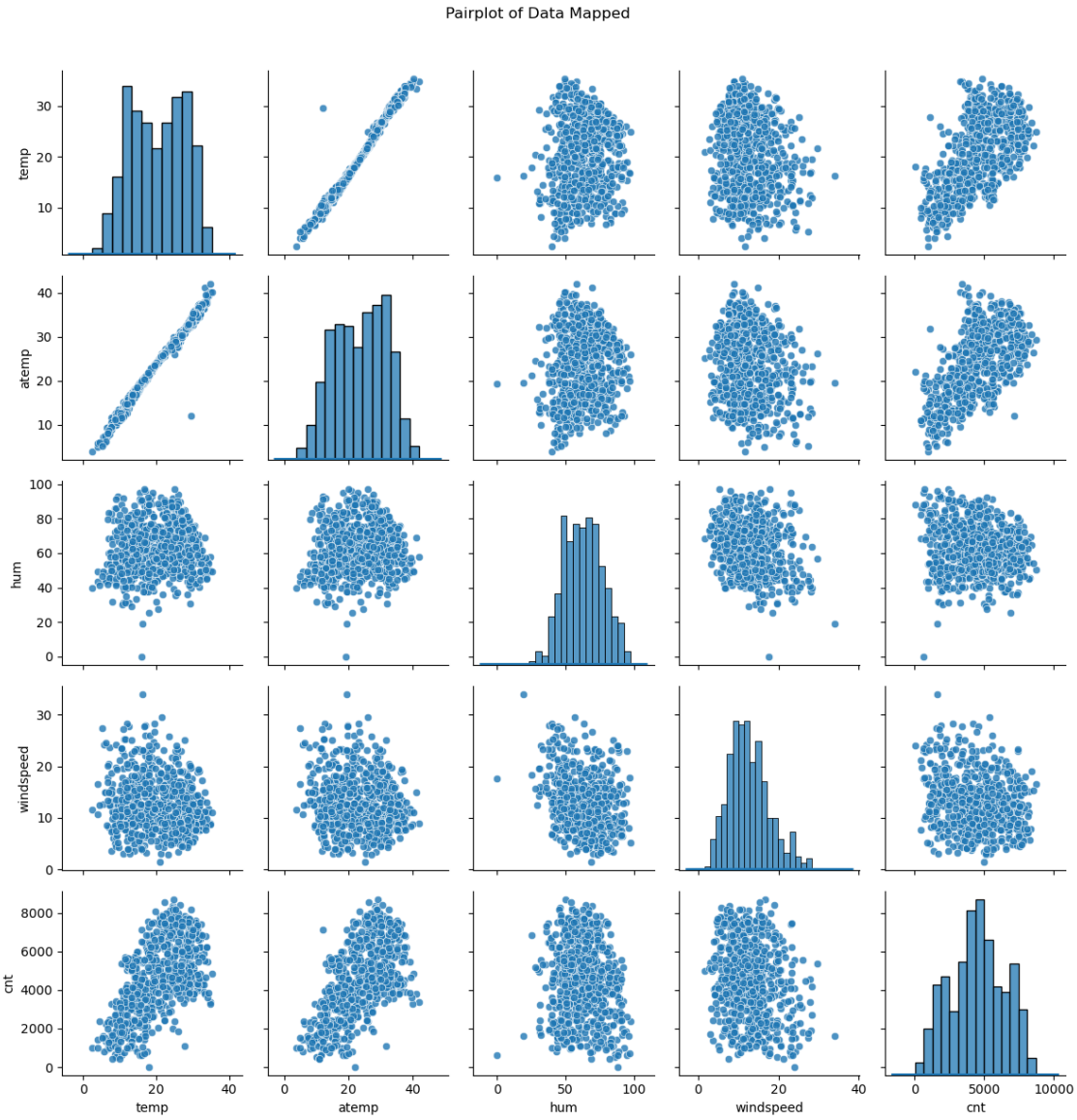
2. Why is it important to use drop_first=True during dummy variable creation?

When creating dummy variables for categorical variables with n levels, you typically create $n-1$ dummy variables. These $n-1$ variables are sufficient to represent all levels of the categorical variable without introducing redundancy. Including all n dummy variables would result in perfect multicollinearity because one level could be perfectly predicted from the others. Dropping one dummy variable helps avoid multicollinearity issues in the model.

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

Upon examining the pair-plot of numerical variables, such as 'feeling temperature' and 'temperature', we notice a strong correlation with the target variable 'cnt'. This observation aligns with our previous findings regarding the categorical variables and bike rentals. Specifically, seasons characterized by pleasant temperatures, like summer and fall, experience a surge in bike rentals. This trend is further supported by the month-wise analysis.

Below, you'll find pairplot visualizing the relationships between the numerical variables :

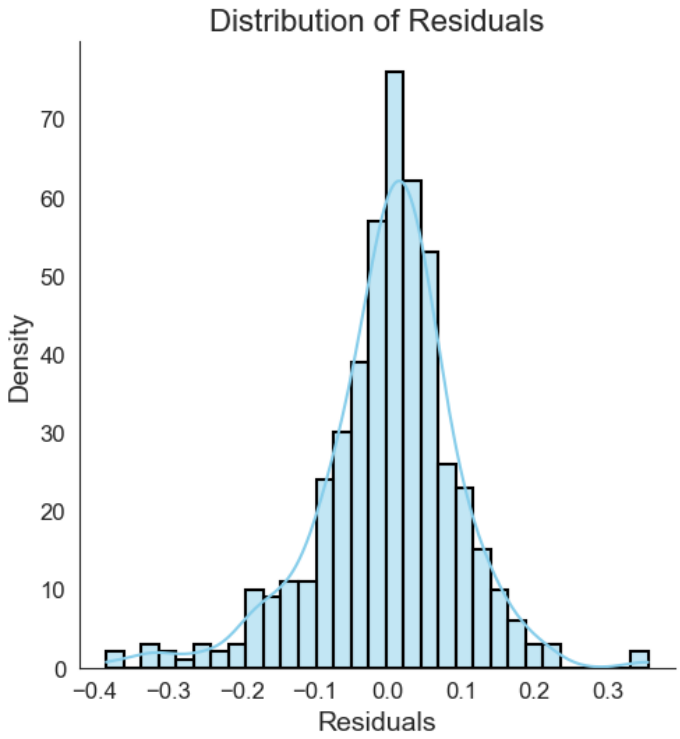


In summary, the pair-plot analysis of numerical variables reveals a notable correlation between 'feeling temperature', 'temperature', and the target variable 'cnt'. These findings corroborate our earlier observations regarding the impact of seasonal variations on bike rentals, particularly during periods with pleasant temperatures. The visual exploration of numerical variables provides additional support for the relationship between weather conditions and bike rental demand.

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

To validate the assumptions of linear regression, including linearity, normality, multicollinearity, and homoscedasticity, we conducted residual analysis on the final model. Residual analysis helps ensure that the dataset meets these assumptions. We examined the distribution plot of the residuals to check for normality and centeredness around zero.

Below is the distribution plot of the residual, it shows normal distribution and it is centred at 0:



To address multicollinearity, we employed the variance inflation factor (VIF). Variables with VIF values exceeding 5 were removed to mitigate multicollinearity in the dataset.

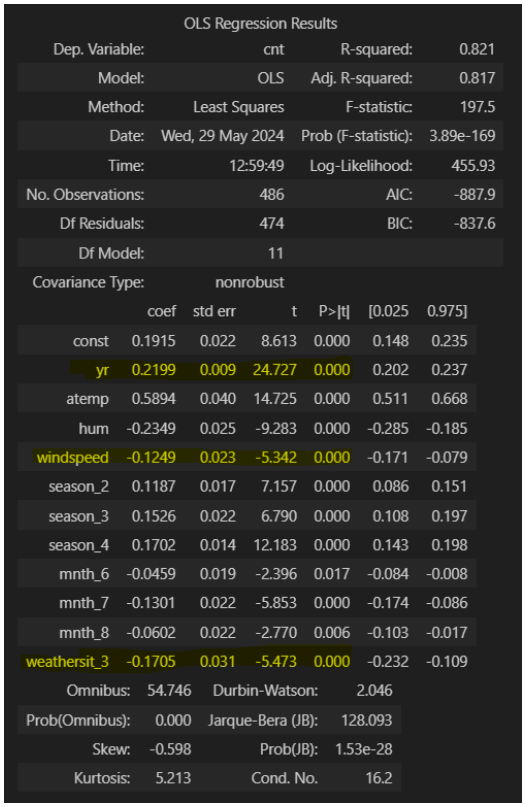
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

In the final model, the top three features that emerge as statistically significant are:

- 'yr'
- 'windspeed'
- 'weathersit_3'

These variables are deemed the most significant based on their p-values, all of which are 0, indicating strong statistical significance. Additionally, their VIF (Variance Inflation Factor) values support their significance. The variables are listed in increasing order of their VIF values.

To validate these findings, screenshots of the model summary and VIF values are provided for verification:

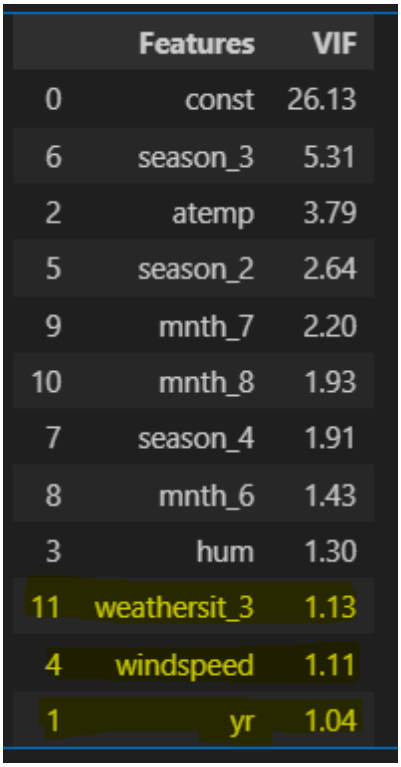


OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	197.5
Date:	Wed, 29 May 2024	Prob (F-statistic):	3.89e-169
Time:	12:59:49	Log-Likelihood:	455.93
No. Observations:	486	AIC:	-887.9
Df Residuals:	474	BIC:	-837.6
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1915	0.022	8.613	0.000	0.148	0.235
yr	0.2199	0.009	24.727	0.000	0.202	0.237
atemp	0.5894	0.040	14.725	0.000	0.511	0.668
hum	-0.2349	0.025	-9.283	0.000	-0.285	-0.185
windspeed	-0.1249	0.023	-5.342	0.000	-0.171	-0.079
season_2	0.1187	0.017	7.157	0.000	0.086	0.151
season_3	0.1526	0.022	6.790	0.000	0.108	0.197
season_4	0.1702	0.014	12.183	0.000	0.143	0.198
mnth_6	-0.0459	0.019	-2.396	0.017	-0.084	-0.008
mnth_7	-0.1301	0.022	-5.853	0.000	-0.174	-0.086
mnth_8	-0.0602	0.022	-2.770	0.006	-0.103	-0.017
weathersit_3	-0.1705	0.031	-5.473	0.000	-0.232	-0.109

Omnibus:	54.746	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	128.093
Skew:	-0.598	Prob(JB):	1.53e-28
Kurtosis:	5.213	Cond. No.	16.2



	Features	VIF
0	const	26.13
6	season_3	5.31
2	atemp	3.79
5	season_2	2.64
9	mnth_7	2.20
10	mnth_8	1.93
7	season_4	1.91
8	mnth_6	1.43
3	hum	1.30
11	weathersit_3	1.13
4	windspeed	1.11
1	yr	1.04

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a statistical technique which help us understand the linear relationship between the independent and the dependent variables. The main of this algorithm is to determine the linear equation that best predicts dependent variables using independent variables.

Two types of Linear regression:

- Simple Linear Regression: Involves a single independent variable and a dependent variable.
- Multiple Linear Regression: Involves multiple independent variables and a dependent variable.

The general equation of the linear regression model is:

$$\mathbf{Y = B0 + b1X1 + b2X2 + b3X3 + ... + bnXn + E}$$

Where:

- Y is the dependent variable.
- X1, X2, X3, . . ., Xn are the independent variables.
- B0 is the intercept.
- b1, b2, b3, . . . , bn are the coefficients.
- E is the error term (residual).

The coefficients are estimated using the least squares method, which minimizes the sum of the squared differences between observed and predicted values. To find the sum of the squared differences for a particular instance we use the below formula:

$$RSS=\sum(yi-yi\hat{})2$$

To apply Linear Regression, the variables must satisfy certain assumptions:

- The relationship between the independent and dependent variables is linear.
- Observations are independent of each other.
- Constant variance of the residuals.
- Residuals are normally distributed.
- Independent variables are not highly correlated with each other.

Advantages

- Simple to understand and implement.
- Provides interpretable results.
- Efficient for small to medium-sized datasets.

Disadvantages

- Assumes a linear relationship, which may not always be accurate.
- Sensitive to outliers.
- Assumes homoscedasticity and independence of residuals.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties but very different distributions and appear quite different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and the effect of outliers and the limitations of summary statistics.

Properties of Anscombe's Quartet

All four datasets in Anscombe's quartet have the following nearly identical properties:

- **Mean of x values**: 9
- **Variance of x values**: 11
- **Mean of y values**: 7.50
- **Variance of y values**: 4.12
- **Correlation between x and y**: 0.816
- **Linear regression line**: y=3+0.5xy=3+0.5x
- **Residual sum of squares (RSS)**: 13.75

Despite these similar statistical properties, the datasets have very different distributions and reveal different patterns when plotted. Here is a detailed look at each of the four datasets and their plots:

Dataset 1

- This dataset forms a linear relationship between xx and yy.
- When plotted, it shows a classic linear trend that fits well with the regression line y=3+0.5xy=3+0.5x.

Dataset 2

- This dataset also shows a linear relationship but with one outlier.
- The outlier significantly influences the correlation and regression line.
- When plotted, it is clear that the outlier distorts the perceived relationship.

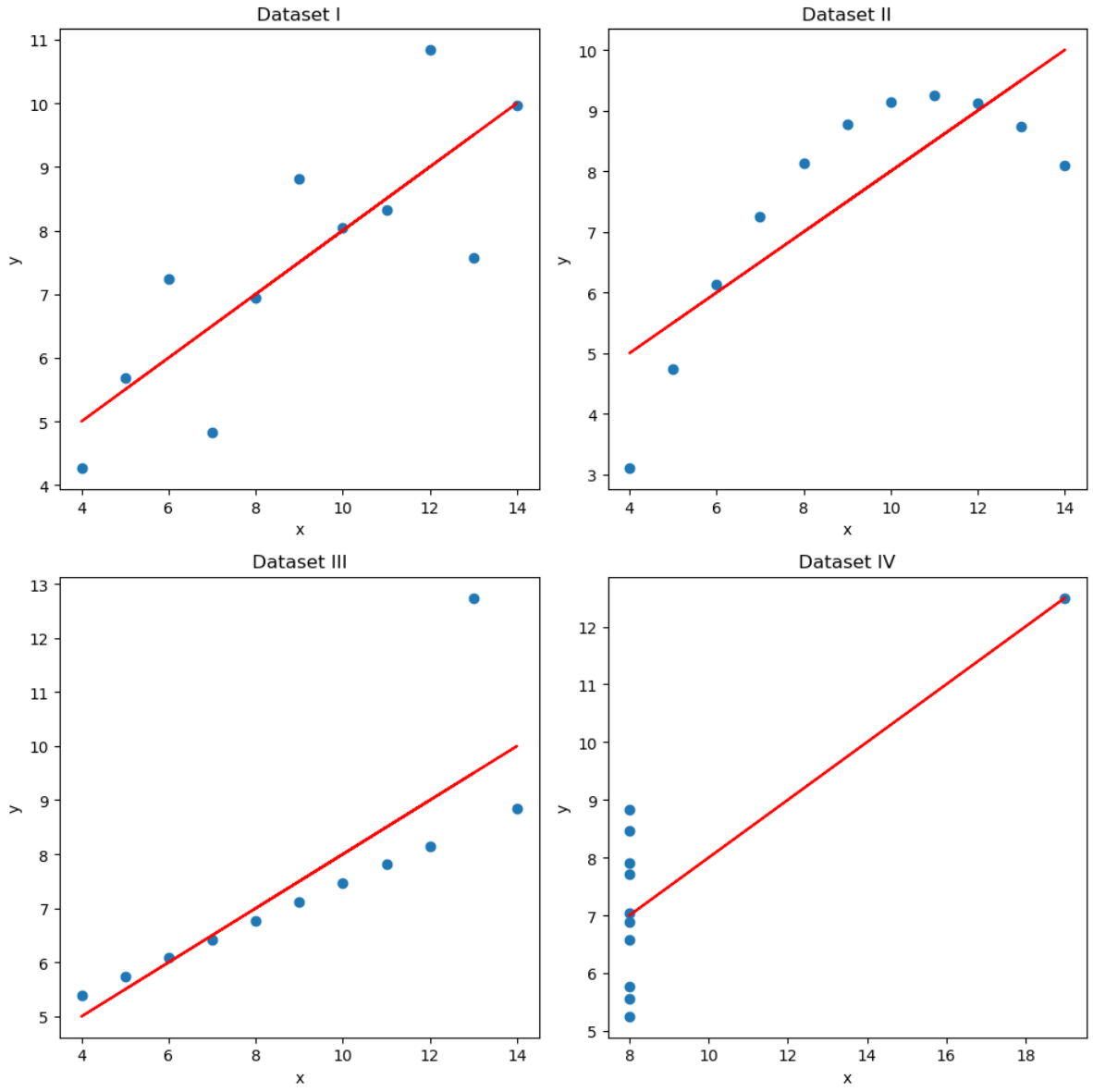
Dataset 3

- This dataset forms a perfect linear relationship except for one influential point.
- All points but one lie on a straight line, but the outlier greatly affects the regression.
- The plot reveals that the linear regression line does not represent the majority of the data points well.

Dataset 4

- This dataset has a vertical relationship between xx and yy due to one influential point.
- Most of the points have the same xx value except for one outlier.
- The plot shows that the regression line is misleading because of the influential outlier.

Below is the plot for Anscombe's Quartet using Seaborn dataset library:



Importance and Lessons from Anscombe's Quartet

Anscombe's quartet demonstrates several critical points:

- **Graphing Data**: Always visualize your data before performing statistical analyses. Graphs can reveal patterns, trends, and outliers that summary statistics might not.
- **Effect of Outliers**: Outliers can significantly affect statistical measures and can lead to misleading conclusions if not identified and handled properly.
- **Limits of Summary Statistics**: Similar summary statistics (mean, variance, correlation) do not imply similar data distributions. Different datasets can have identical statistical properties yet differ in shape, spread, and structure.

In summary, Anscombe's quartet underscores the necessity of comprehensive data exploration, emphasizing visualization to uncover the true characteristics and relationships within the data.

3. What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient or simply Pearson's correlation, is a measure of the linear correlation between two variables. It quantifies the degree to which a linear relationship exists between two continuous variables.

The value of Pearson's R ranges from -1 to 1, where:

- +1 indicates a perfect positive linear correlation.
- -1 indicates a perfect negative linear correlation.
- 0 indicates no linear correlation.

Formula for Pearson's R is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Where:

- r is the Pearson correlation coefficient.
- xi and yi are the individual sample points indexed with i.
- x^- and y^- are the means of the x and y values, respectively.
- n is the number of sample points.

Interpretation

- **r = 1**: Perfect positive correlation. As xx increases, yy increases in a perfectly linear manner.
- **r = -1**: Perfect negative correlation. As xx increases, yy decreases in a perfectly linear manner.
- **r = 0**: No linear correlation. There is no linear relationship between xx and yy.
- **0 < r < 1**: Positive correlation. As xx increases, yy tends to increase.
- **-1 < r < 0**: Negative correlation. As xx increases, yy tends to decrease.

Use Cases for Pearson's R is widely used in various fields, including:

- **Statistics**: To measure the strength and direction of the linear relationship between two variables.
- **Finance**: To determine the correlation between asset returns, which helps in portfolio diversification.
- **Research**: To assess the linear relationship between variables in scientific studies.
- **Machine Learning**: To select features that have strong linear relationships with the target variable.

While Pearson's R is a powerful tool, it has several limitations:

- **Linearity**: Pearson's R only measures linear relationships. It does not capture non-linear relationships between variables.
- **Sensitivity to Outliers**: Outliers can significantly affect the value of Pearson's R, leading to misleading interpretations.
- **Assumes Normality**: Pearson's R assumes that the variables are normally distributed. If the data is heavily skewed, the correlation coefficient may not be accurate.

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

Scaling is a data preprocessing technique used to adjust the range and distribution of data features so they can be effectively analyzed and modeled. Scaling ensures that different features contribute equally to the analysis and prevents any single feature from dominating the results due to its larger magnitude.

Scaling is performed for several reasons:

- **Improving Algorithm Performance**: Many machine learning algorithms, especially those based on distance metrics (e.g., k-nearest neighbors, support vector machines), and gradient descent optimization (e.g., linear regression, logistic regression, neural networks), perform better when features are on a similar scale.
- **Ensuring Faster Convergence**: In optimization algorithms like gradient descent, scaling helps achieve faster convergence by preventing large feature values from causing large gradient updates, which can lead to oscillations and slow convergence.
- **Maintaining Numerical Stability**: Scaling helps maintain numerical stability in algorithms by preventing the model from producing large numerical values, which can lead to computational issues.
- **Enhancing Interpretability**: Scaled features can improve the interpretability of the model coefficients, as the coefficients become more comparable when the features are on a similar scale.

There are two common types of scaling: normalization and standardization.

Normalization (or min-max scaling) transforms the features to a fixed range, usually [0, 1]. Each feature value is scaled according to the following formula:

$$x' = x - \min(x) / (\max(x) - \min(x))$$

Where:

- x' is the normalized value.
- x is the original value.
- $\min(x)$ is the minimum value of the feature.
- $\max(x)$ is the maximum value of the feature.

Use Case: Normalization is useful when the distribution of data is not Gaussian (normal distribution) and when you know the boundaries of your data, as it scales the data to a specific range.

Standardization (or z-score normalization) transforms the features to have a mean of 0 and a standard deviation of 1. Each feature value is scaled according to the following formula:

$$x'=(x-\mu)/\sigma$$

Where:

- x' is the standardized value.
- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Use Case: Standardization is useful when the data follows a Gaussian distribution and you want to center the data around the mean with a unit variance. This is particularly helpful for algorithms that assume the data is centered around zero.

Key Differences

Aspect	Normalization	Standardization
Formula	$x'=x-\min(x)/(\max(x)-\min(x))$	$x'=(x-\mu)/\sigma$
Range of Values	Typically [0, 1] (can be changed)	Mean of 0 and standard deviation of 1
Use Case	When data is not Gaussian; for bounded features	When data is Gaussian; for centering
Effect on Outliers	Sensitive to outliers	Less sensitive to outliers
Data Distribution	Does not change the shape of the distribution	Centers data around mean; unit variance

In summary, scaling is a crucial preprocessing step in machine learning that ensures features are on a comparable scale, leading to improved performance and stability of algorithms. Normalization is used for bounded features and non-Gaussian distributions, while standardization is used for Gaussian distributions and centering data around zero.

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

VIF can become infinite when there is perfect multicollinearity. This means that the predictor variable Xi is perfectly linearly dependent on one or more of the other predictor variables. In mathematical terms:

Perfect multicollinearity implies $Ri2$=1.

Substituting $Ri2$=1 into the VIF formula gives:

$$\mathbf{VIF(}Xi\mathbf{) = 1/(1-1) = 1/0 = \infty}$$

This situation occurs when there is an exact linear relationship between one predictor and the others. As a result, the regression model cannot distinguish between the collinear predictors, leading to an infinite VIF.

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 graphical tool used to compare the distribution of a dataset to a theoretical distribution, often the normal distribution. It is a scatter plot that shows the quantiles of the sample data on the y-axis against the quantiles of the theoretical distribution on the x-axis.

Use and Importance of a Q-Q Plot in Linear Regression

In linear regression, a Q-Q plot is primarily used for checking the assumption of normality of the residuals (errors). The key assumptions of linear regression include:

- **Linearity**: The relationship between the predictors and the response variable is linear.
- **Homoscedasticity**: The residuals have constant variance at all levels of the predictor variables.
- **Independence**: The residuals are independent.
- **Normality**: The residuals are normally distributed.