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?

In the dataset, there are several categorical variables, including Season, Weather Situation, Holiday, Month, Working Day, and Weekday. We visualized these variables using boxplots and observed their influence on the dependent variable as follows:

1. Season: The boxplot revealed that the spring season had the lowest count of bike rentals, while the fall season had the highest count. Summer and winter fell in between these two extremes.

2. Weather Situation: During heavy rain or snow, bike rentals were non-existent, indicating extremely unfavorable weather conditions. The highest rental count was observed when the weather was forecasted as 'Clear or Partly Cloudy.'

3. Holiday: Rentals were generally lower during holidays, likely due to a change in user behavior during holiday periods.

4. Month: September had the highest number of rentals, while January had the fewest. This observation aligns with the Weather Situation variable, as January typically experiences cold and snowy weather.

5. Weekday: Weekday had minimal influence on the dependent variable, suggesting that the day of the week didn't significantly impact bike rentals.

6. Working Day: Working days showed a substantial increase in bike rentals compared to non-working days, indicating that people are more likely to rent bikes for commuting on workdays.

Additionally, we noticed that the median bike rentals have been increasing year by year, with 2019 having a higher median than 2018. This could be attributed to the growing popularity of bike rentals and an increasing awareness of environmental concerns among people.

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

Using **drop_first=True** during dummy variable creation is important to avoid multicollinearity and enhance the interpretability of regression models, especially in scenarios where categorical variables are being encoded using one-hot encoding. Here's why it's important:

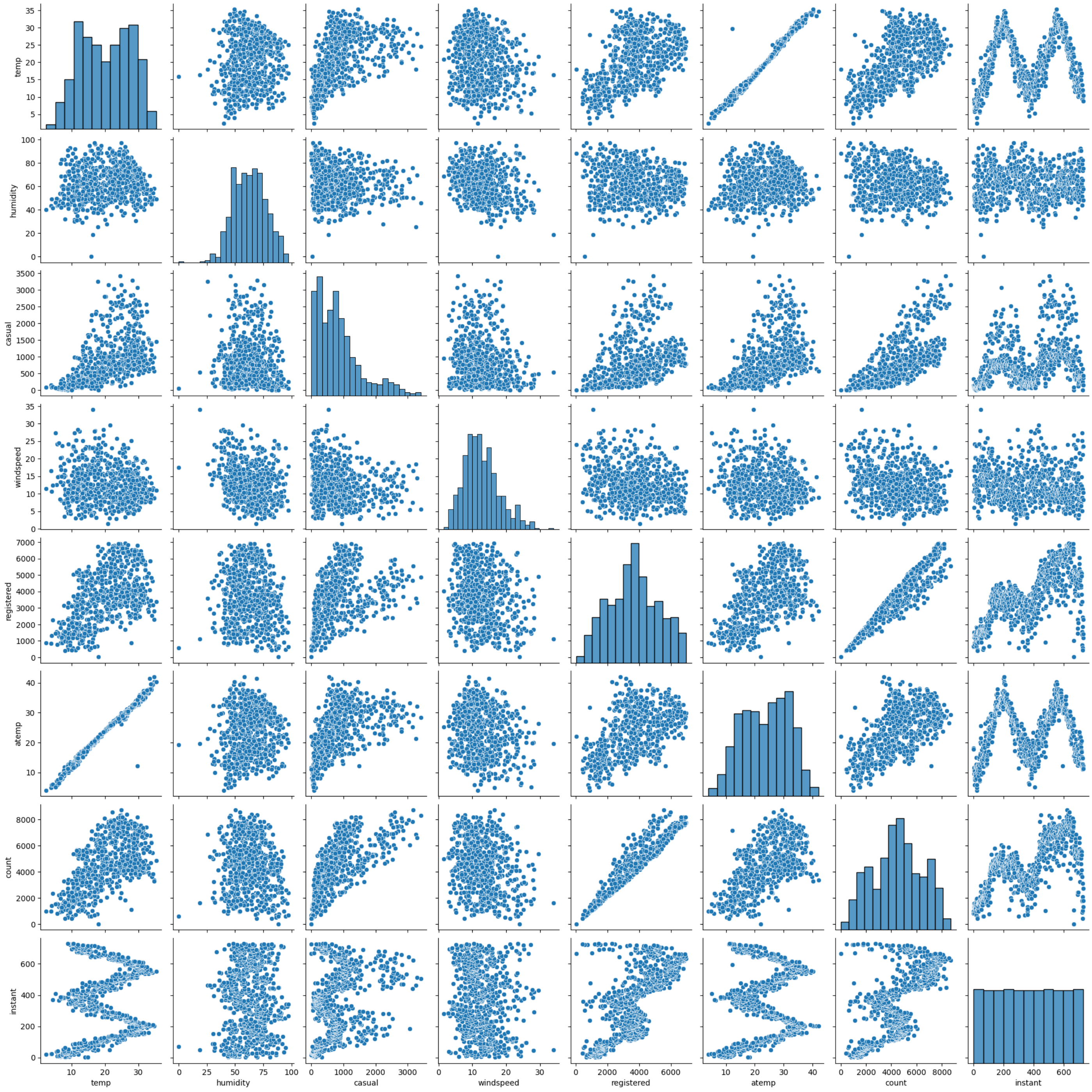
1. Multicollinearity: When you create dummy variables for categorical data, one-hot encoding typically results in one category serving as the reference category (represented as all zeros in the dummy variables). If you include all dummy variables in the regression model, you introduce multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated, which can lead to unstable parameter estimates and difficulty in interpreting the model.

2. Redundancy: Including all dummy variables in the model makes one of them redundant. For example, if you have three categories (A, B, and C) with dummy variables D1 and D2, the information about category C is already captured when D1 and D2 are known (D3 = 1 - D1 - D2). This redundancy can lead to issues in regression analysis.

3. Interpretability: Including all dummy variables without dropping one can make the interpretation of the model less intuitive. By setting `drop_first=True`, you make one category the reference category, and the coefficients of the remaining dummy variables represent the difference between those categories and the reference category. This makes it easier to interpret the impact of each category on the dependent variable.

In summary, using **drop_first=True** during dummy variable creation is essential to prevent multicollinearity, reduce redundancy, and enhance the interpretability of regression models. It simplifies the interpretation of coefficients and ensures that the model is both stable and meaningful in its predictions.

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



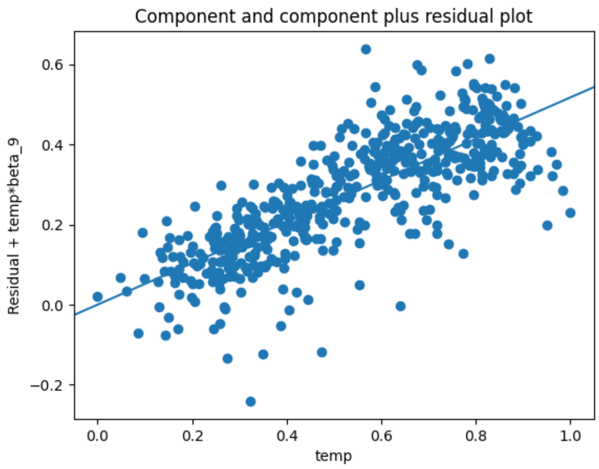
As observed in the pairplot, the "**count**" (the target variable) exhibits the strongest correlations with "**temp**," "**atemp**," "**casual**," and "**registered**."

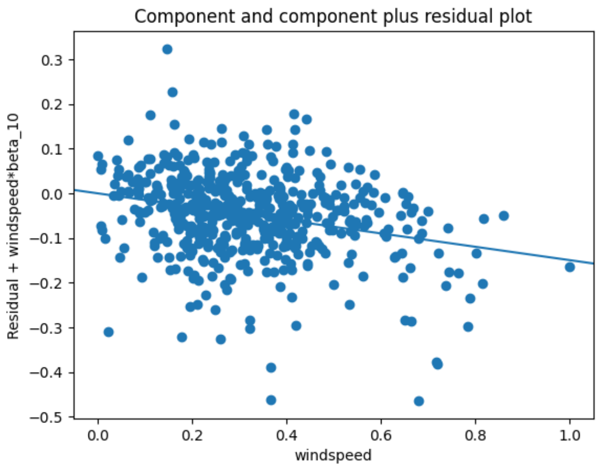
The corresponding correlation coefficients are as follows:

- "**temp**" = 0.63
- "**atemp**" = 0.63
- "**casual**" = 0.67
- "**registered**" = 0.95

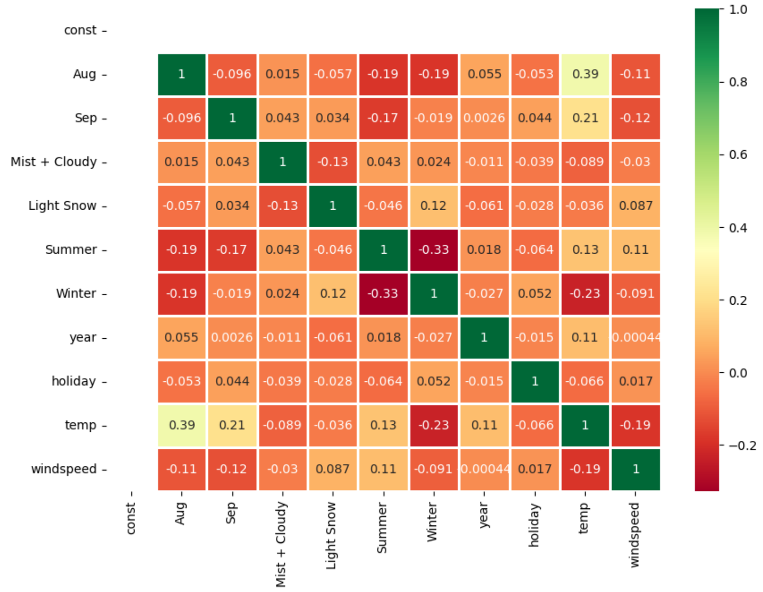
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

To evaluate whether this assumption holds, a simple approach is to construct a scatter plot of x against y. If the data points form a straight line on the graph, it suggests a linear relationship between the dependent and independent variables, confirming the assumption's validity.



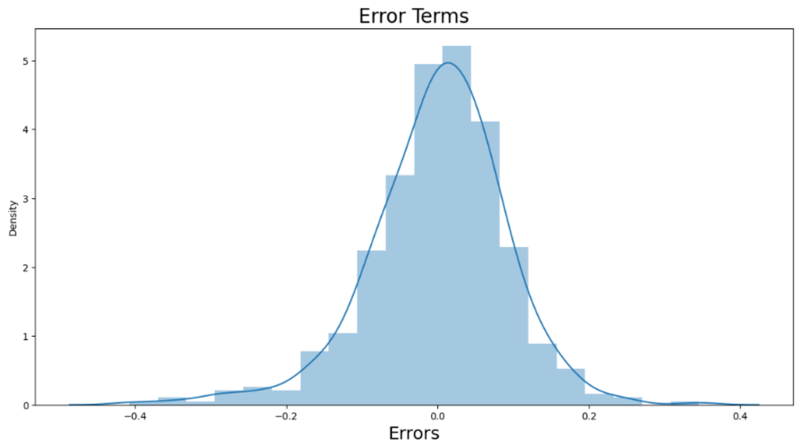


Also, the heat map to check multi-collinearity validated it.



	const	Aug	Sep	Mist + Cloudy	Light Snow	Summer	Winter	year	holiday	temp	windspeed
const											
Aug		1	-0.096	0.015	-0.057	-0.19	-0.19	0.055	-0.053	0.39	-0.11
Sep		-0.096	1	0.043	0.034	-0.17	-0.019	0.0026	0.044	0.21	-0.12
Mist + Cloudy		0.015	0.043	1	-0.13	0.043	0.024	-0.011	-0.039	-0.089	-0.03
Light Snow		-0.057	0.034	-0.13	1	-0.046	0.12	-0.061	-0.028	-0.036	0.087
Summer		-0.19	-0.17	0.043	-0.046	1	-0.33	0.018	-0.064	0.13	0.11
Winter		-0.19	-0.019	0.024	0.12	-0.33	1	-0.027	0.052	-0.23	-0.091
year		0.055	0.0026	-0.011	-0.061	0.018	-0.027	1	-0.015	0.11	0.00044
holiday		-0.053	0.044	-0.039	-0.028	-0.064	0.052	-0.015	1	-0.066	0.017
temp		0.39	0.21	-0.089	-0.036	0.13	-0.23	0.11	-0.066	1	-0.19
windspeed		-0.11	-0.12	-0.03	0.087	0.11	-0.091	0.00044	0.017	-0.19	1

Furthermore, when examining the distribution of residuals, it was found to follow a normal distribution with a mean value of 0.



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

The top three predictor variables that significantly impact bike bookings, as determined by our final model, are as follows:

1. Temperature (temp): This variable has a coefficient of 0.5173, meaning that for each unit increase in temperature, bike rentals increase by 0.5173 units.

2. Weather Situation (weathersit_3): With a coefficient of -0.2819, a two-unit increase in the lightsnow variable results in a reduction of 0.2828 units in bike rentals compared to cloudy. Lightsnow corresponds to conditions like Light Snow, Light Rain + Thunderstorm, and Scattered Clouds, Light Rain + Scattered Clouds.

3. Year (year): Year has a coefficient of 0.2326, indicating that a one-unit increase in the year variable leads to an increase of 0.2326 units in bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression, a supervised machine learning algorithm, specializes in regression tasks. It seeks to establish a relationship between a dependent variable (denoted as “y”) and an independent variable (referred to as “x”) by fitting a straight line to the data points. This linear relationship between “y” and “x” underpins the name “linear regression.”

In this context:
- The independent variable “x” is also known as the predictor variable, serving as the input for making predictions.
- The dependent variable “y” is commonly referred to as the output variable, representing the variable we aim to predict.

Mathematically, this relationship is expressed as:
y = mx + c

Here, the elements signify:
- “y” is the dependent variable we seek to predict.
- “x” stands for the independent variable used for predictions.
- “m” denotes the slope of the regression line, signifying the impact of “x” on “y.”
- “c” represents a constant known as the Y-intercept. When “x” equals zero, “y” takes on the value of “c.”

This linear relationship can manifest in two primary forms:
1. **Positive Linear Relationship:**
 - A positive linear relationship occurs when both the independent and dependent variables increase simultaneously.

2. **Negative Linear Relationship**:
 - Conversely, a negative linear relationship emerges when an increase in the independent variable corresponds to a decrease in the dependent variable, and vice versa.

Linear regression comes in two main variations:
1. **Simple Linear Regression:**
 - This method explores the relationship between a single independent variable and the dependent variable.

2. **Multiple Linear Regression**:
 - Multiple linear regression extends the concept to incorporate multiple independent variables, allowing for more complex modeling of relationships and takes the following form:
$\mathbf{Y= B_0+B_1x_1+B_2x_2+B_3x_3\ldots B_nx_n}$

Where :

B_1 = coefficient for X1 variable
B_2= coefficient for X2 variable
B_3= coefficient for X3 variable and so on...
B_0 is the intercept (constant term)

Several key assumptions underlie the linear regression model's performance:

1. **Multi-collinearity**:
   - The model assumes little or no multi-collinearity in the dataset. Multi-collinearity occurs when independent variables exhibit dependencies among themselves, which can hinder model accuracy.

2. **Auto-correlation**:
   - Another assumption is that there is minimal or no auto-correlation in the data. Auto-correlation arises when residual errors exhibit dependencies, potentially leading to inaccurate model predictions.

3. **Relationship Between Variables**:
   - Linear regression assumes that the relationship between response and feature variables must be linear in nature. In other words, the effect of the independent variables on the dependent variable is linear.

4. **Normality of Error Terms**:
   - Error terms, representing the differences between predicted and actual values, should adhere to a normal distribution. Deviations from normality can impact model performance.

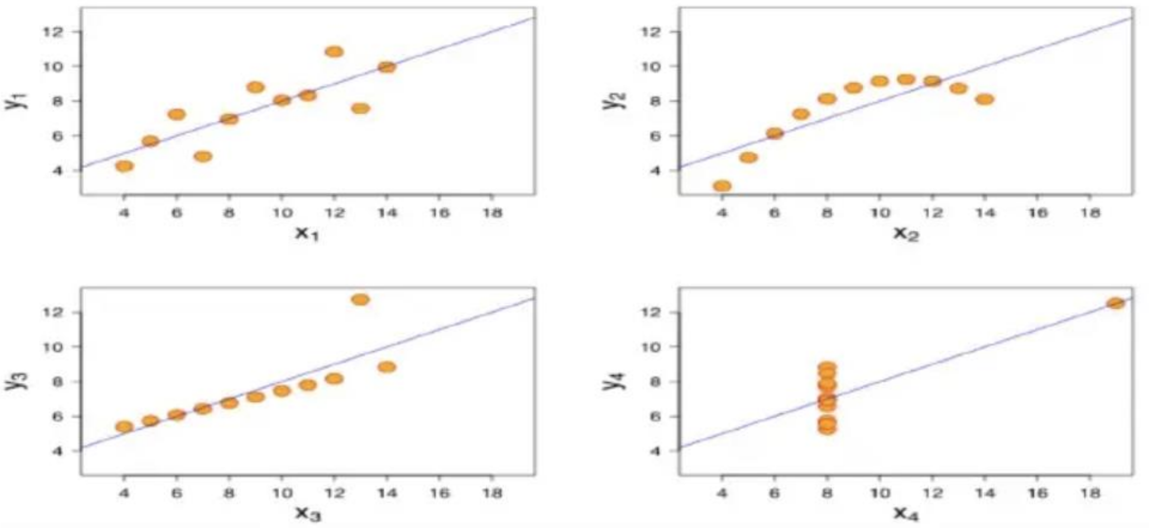
5. **Homoscedasticity**:
   - The model assumes that there is no discernible pattern in residual values, ensuring consistent variability across different levels of the independent variable. Violations of this assumption can indicate heteroscedasticity, affecting the reliability of the model's predictions.

In summary, linear regression is a fundamental and interpretable machine learning algorithm that models the relationship between variables through a linear approach. It comes in various forms and makes specific assumptions about the dataset, all of which are essential for the model's accuracy and effectiveness.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. This quartet was created by the British statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in understanding and interpreting data. Anscombe's goal was to show that a reliance solely on summary statistics, such as means, variances, and correlation coefficients, can be misleading, and that graphical exploration of data is a crucial part of data analysis.

The quartet consists of four datasets, each containing 11 data points. Here is a brief description of each dataset in Anscombe's quartet:



1. **Dataset I**:
   - This dataset represents a simple linear relationship with some random noise. When graphed, it appears as a scatterplot with a linear trend.

2. **Dataset II**:
   - Unlike the first two datasets, Dataset III does not follow a linear relationship. It consists of an apparent quadratic relationship. This dataset highlights the importance of considering nonlinear relationships.

3. **Dataset III**:
   - Similar to Dataset I, this dataset also represents a linear relationship, but it contains an outlier that significantly impacts the line of best fit. The presence of this outlier affects the overall pattern.

4. **Dataset IV**:
   - Dataset IV is designed to look like there's no clear relationship between the variables, with one data point being an influential outlier. It emphasizes the effect of outliers on correlation and regression analysis.

The key takeaway from Anscombe's quartet is that summary statistics alone can be insufficient for understanding complex datasets. Graphical representation is essential for revealing underlying patterns, outliers, and relationships that may be overlooked when focusing solely on numerical summaries. Data visualization techniques, such as scatterplots, can provide valuable insights into data structure, trends, and potential issues that may impact statistical analysis and model interpretation.

In summary, Anscombe's quartet serves as a compelling reminder of the importance of data visualization in the data analysis process and the potential limitations of relying solely on summary statistics. It emphasizes that visually exploring and understanding data is an integral part of meaningful data analysis.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as “r” or “Pearson's r,” is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics to assess the degree of correlation between two variables. Pearson's r ranges from -1 to 1 and provides insights into how closely the two variables are related.

Key characteristics of Pearson's correlation coefficient include:

1. Range:

- Pearson's r varies between -1 and 1.
- An r value of 1 indicates a perfect positive linear relationship, where one variable increases as the other increases.
- An r value of -1 signifies a perfect negative linear relationship, where one variable decreases as the other increases.
- An r value of 0 indicates no linear relationship between the variables.

2. Interpretation:

- An r value close to 1 or -1 suggests a strong linear relationship. The sign (positive or negative) indicates the direction of the relationship.
- An r value close to 0 implies a weak or no linear relationship.

3. Strength:

- The closer the absolute value of r is to 1, the stronger the linear relationship.
- Values around 0.7 or -0.7 are generally considered moderately strong, while values above 0.9 or below -0.9 are considered very strong.

4. Assumptions:

- Pearson's correlation coefficient assumes that the relationship between the two variables is linear.
- It also assumes that the variables are continuous and follow a bivariate normal distribution.

5. Calculation:

- The formula for Pearson's r is:

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

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

Pearson's r is commonly used in various fields, including statistics, social sciences, economics, and natural sciences. It helps researchers and analysts understand the degree and direction of association between two variables. However, it's important to remember that Pearson's correlation coefficient only measures linear relationships and may not capture more complex or nonlinear associations between variables. In cases where the relationship is not linear, alternative correlation measures, or methods like Spearman's rank correlation, may be more appropriate.

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

Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the values of variables to a specific range or distribution. It is performed to bring variables to a common scale or format, making them more suitable for analysis, modeling, and comparison. Scaling is particularly useful when dealing with features or variables that have different units, ranges, or magnitudes.

Here are some key reasons why scaling is performed:
1. Equalization of Scales: Variables often have different units and ranges. Scaling makes it possible to compare and analyze these variables on a common scale, which is important for many statistical and machine learning techniques.

2. Algorithm Sensitivity: Many machine learning algorithms, such as k-nearest neighbors and support vector machines, are sensitive to the scale of the input variables. Scaling can improve the performance and convergence of these algorithms.

3. Interpretability: Scaling can make it easier to interpret the coefficients or feature importance in linear models or decision trees. Without scaling, it can be challenging to compare the importance of variables with different units and magnitudes.

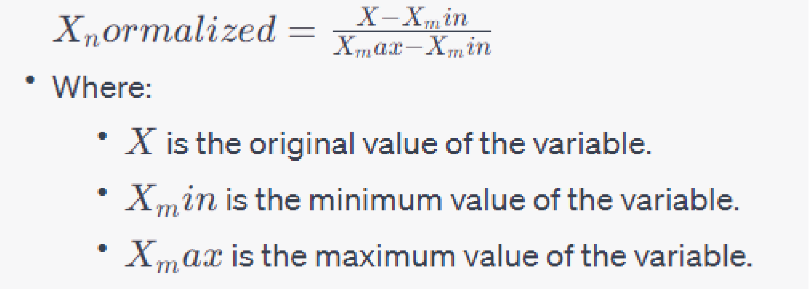
4. Reduced Computational Complexity: Some optimization algorithms converge faster and are more computationally efficient when variables are on a similar scale.

5. Visualization: Scaling variables can make it easier to visualize and compare data, especially in situations where the variable ranges are vastly different.

There are two common types of scaling techniques: normalized scaling and standardized scaling. Here's how they differ:

1. Normalized Scaling (Min-Max Scaling):

- Normalized scaling transforms the variable values to a specific range, typically between 0 and 1.
- The formula for normalized scaling is:



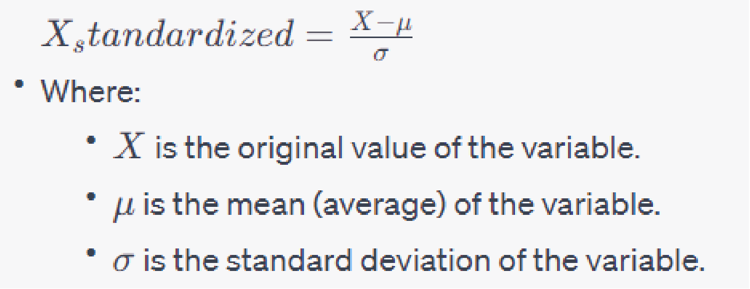
$$X_normalized = \frac{X - X_min}{X_max - X_min}$$

- Where:
  - X is the original value of the variable.
  - X_min is the minimum value of the variable.
  - X_max is the maximum value of the variable.

- Normalized values are bounded between 0 and 1, with 0 representing the minimum value in the dataset and 1 representing the maximum.

2. Standardized Scaling (Z-Score Scaling):

- Standardized scaling transforms the variable values to have a mean of 0 and a standard deviation of 1.
- The formula for standardized scaling is:



$$X_standardized = \frac{X - \mu}{\sigma}$$

- Where:
  - X is the original value of the variable.
  - μ is the mean (average) of the variable.
  - σ is the standard deviation of the variable.

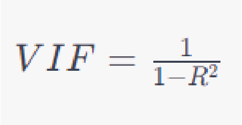
- Standardized values have a mean of 0 and a standard deviation of 1, which makes them centered around the mean with a consistent scale.

In summary, scaling is a crucial data preprocessing step to ensure that variables are on a common scale, making them more suitable for analysis and modeling. Normalized scaling scales variables to a specific range (0 to 1), while standardized scaling transforms variables to have a mean of 0 and a standard deviation of 1. The choice of scaling method depends on the specific requirements of the analysis or modeling task and the characteristics of the data.

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

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple linear regression model. Multicollinearity occurs when two or more independent variables in the regression model are highly correlated, which can lead to problems in the model's interpretation and stability. VIF is a numerical value that quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.

The formula for calculating the VIF for a particular independent variable is as follows:



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

Where:
- R^2 is the coefficient of determination from a regression model in which the variable in question is regressed against all the other independent variables in the same model.

Now, VIF can be influenced by the presence of multicollinearity in the dataset. When VIF becomes infinite, it usually indicates an extreme case of multicollinearity. This happens for the following reasons:

1. Perfect Multicollinearity: Infinite VIF values occur when there is perfect multicollinearity, meaning one independent variable in the model is a linear combination of the other independent variables. In other words, one variable can be exactly predicted from a combination of the other variables in the model. This results in a situation where the coefficient of determination (R^2) in the VIF formula becomes equal to 1.

2. Overparameterization: Infinite VIF can also occur if the model is overparameterized. Overparameterization happens when there are more independent variables in the model than there are observations in the dataset. In such cases, the regression model cannot be estimated properly, leading to infinite VIF values.

In practice, when you encounter infinite VIF values, it's a clear indication that there is a fundamental issue with your regression model. It's important to identify and address the underlying problem of multicollinearity, which may involve removing one or more variables, transforming variables, or rethinking the model specification.

Handling multicollinearity effectively is crucial for building stable and interpretable regression models.

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 assess how well a dataset conforms to a specific theoretical distribution, typically the normal distribution. It compares the quantiles (percentiles) of the observed data with those of the expected distribution.

In linear regression, the Q-Q plot is essential for:

1. Assumption Checking: It helps verify the normality assumption, which states that the residuals (the differences between observed and predicted values) should follow a normal distribution with a mean of 0.

2. Identifying Departures from Normality: Departures from the expected distribution are visible in the Q-Q plot as deviations from a straight line. This is crucial for understanding how well the normality assumption holds.

3. Model Validity: A reliable Q-Q plot is part of model diagnostics, ensuring that the linear regression model meets the necessary assumptions for accurate parameter estimation and hypothesis testing.

In brief, Q-Q plots are valuable tools in linear regression for assessing the normality of residuals and overall model reliability. They help data analysts and statisticians verify key assumptions and make more informed model-related decisions.