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?

Based on the coefficients and their corresponding signs (positive or negative) for the categorical variables in the model, we can make several inferences about their effect on the dependent variable

	coef
const	0.3173
yr	0.2358
temp	0.4371
hum	-0.1535
windspeed	-0.1131
season_spring	-0.1570
season_winter	0.0610
mnth_mar	0.0640
mnth_sep	0.0752
weathersit_2	-0.0552
weathersit_3	-0.2256

Season:

- **Spring:** demand is significantly lower in the spring, as indicated by the negative coefficient (-0.1570).
- **Winter:** Demand is slightly lower in the winter compared to the reference category, as suggested by the positive coefficient (0.0610). However, the effect is less pronounced than in the spring.

Month:

- **March (Coefficient = 0.0640) , September (Coefficient = 0.0752):** Demand is slightly higher in March and September

Weather Situation:

- **weathersit_2: [Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist]**

  This category has a negative coefficient (-0.0552), indicating a slight decrease in demand.

- **weathersit_3: [Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds]**

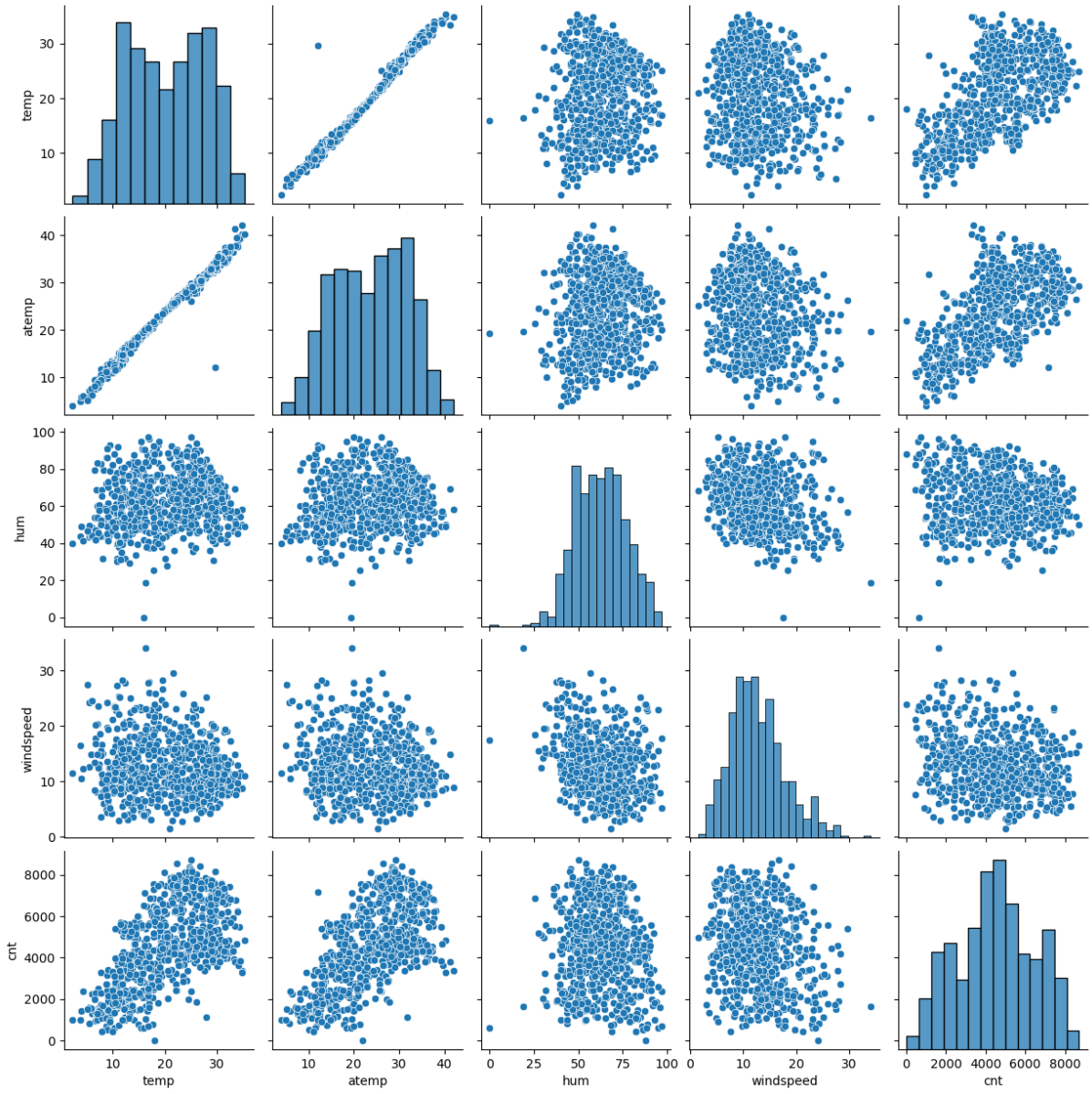
  This category has a significant negative coefficient (-0.2256), suggesting a substantial decrease in demand.

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

Based on the pair plot, **temp or atemp** has highest correlation with **cnt** target variable**.**

Looking at the scatter plot between **temp or atemp** and **cnt** , shows strong positive linear relationship , as the datapoints fall along the straight line indicating strong positive relationship

Also, we can see that temp and atemp are showing perfect linear relationship, we will use any one of the predictors in our modeling.



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

To avoid multicollinearity, we use **drop_first=True** during dummy variable creation.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This can lead to unstable coefficients, inflated standard errors, and difficulty in interpreting the model's results.

When creating dummy variables for categorical variables with k categories, k-1 dummy variables are typically sufficient. This is because the last category can be inferred from the absence of the other k-1 categories.

> *For example, if a categorical variable has three categories (A, B, C), only two dummy variables are needed: A and B. If A and B are both 0, then the category must be C.*

By setting **drop_first=True**, we explicitly drop the first dummy column, preventing perfect multicollinearity between the dummy variables. This ensures that the model can accurately estimate the effects of the categorical variable without redundant information.

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

Linear regression models rely on several key assumptions.

Validating these assumptions is essential to ensure the reliability and accuracy of the model's results.

1. **Linearity: The relationship between the dependent variable and the independent variables is linear**

   How to Validate - After fitting the model, plot the residuals (errors) on the y-axis and the fitted values (predicted values) on the x-axis. The residuals should be randomly scattered around the horizontal line (zero), without any clear pattern. If you observe a pattern, such as a curve or trend, it indicates a violation of the linearity assumption.

2. **Error terms are normally distributed**

   How to Validate –

   - Plot a histogram of the residuals to visually inspect the distribution. A bell-shaped curve suggests normality.
   - Q-Q Plot (Quantile-Quantile Plot): Plot the quantiles of the residuals against the quantiles of a normal distribution. If the residuals are normally distributed, the points should lie along the 45-degree line. Significant deviations from this line indicate a violation of the normality assumption.

3. **Error terms are independent of each other**

   How to Validate - For time series data, plot the residuals over time. The residuals should not show any patterns or trends over time; they should be randomly scattered.

4. **Error terms have constant variance (homoscedasticity)**

How to Validate - Create a plot of residuals against predicted values. If the spread of residuals is constant, homoscedasticity is met

5. **No Multicollinearity**
   How to Validate
   - By Looking at pairwise correlations between the independent variables can sometimes be useful to detect multicollinearity
   - Variance Inflation Factor: By calculating VIF for each independent feature. VIF values above 5 or 10 suggest high multicollinearity and indicate that some variables may need to be dropped or combined.

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 contributing to demand of shared bikes, as per the final model,

1. **temp (Temperature)**: Coefficient = 0.4371
   Temperature has the highest positive coefficient, indicating that it is the most significant feature in explaining the demand for shared bikes. As temperature increases, the demand for shared bikes increases.

2. **yr (Year)**: Coefficient = 0.2358
   The year is the second most significant feature. A positive coefficient suggests that bike demand has increased over time (perhaps from one year to the next).

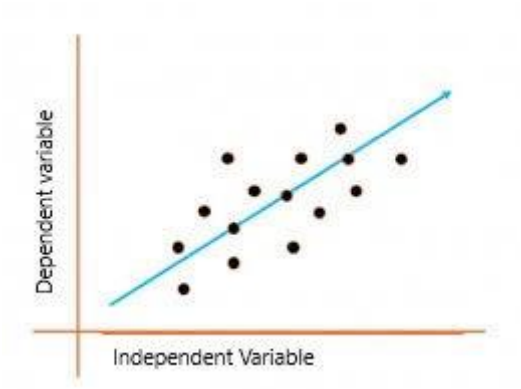
3. **weathersit_3 (Weather Situation 3)**: Coefficient = -0.2256
   [Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds]

   This feature has a strong negative coefficient, indicating that this weather condition (likely poor weather) significantly reduces the demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (also called the target or outcome) and one or more independent variables (features). It assumes a linear relationship between the variables, meaning that the dependent variable can be expressed as a linear combination of the independent variables.



Linear regression is a quiet and the simplest statistical regression technique used for predictive analysis in machine learning. It shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent (output) variable i.e. Y-axis, called linear regression.

Types of Linear Regression

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for a simple linear regression model with one independent variable is:

$y = \beta_o + \beta_1 x + \varepsilon$

where:

- y is the dependent variable
- x is the independent variable
- β_o is the intercept (the value of y when x is 0)
- β_1 is the slope (the change in y for a unit change in x)
- ε is the error term, representing the random variation that cannot be explained by the model

Multiple Linear Regression

This involves more than one independent variable and one dependent variable.

For multiple linear regression with more than one independent variable, the equation becomes:

$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$

where:

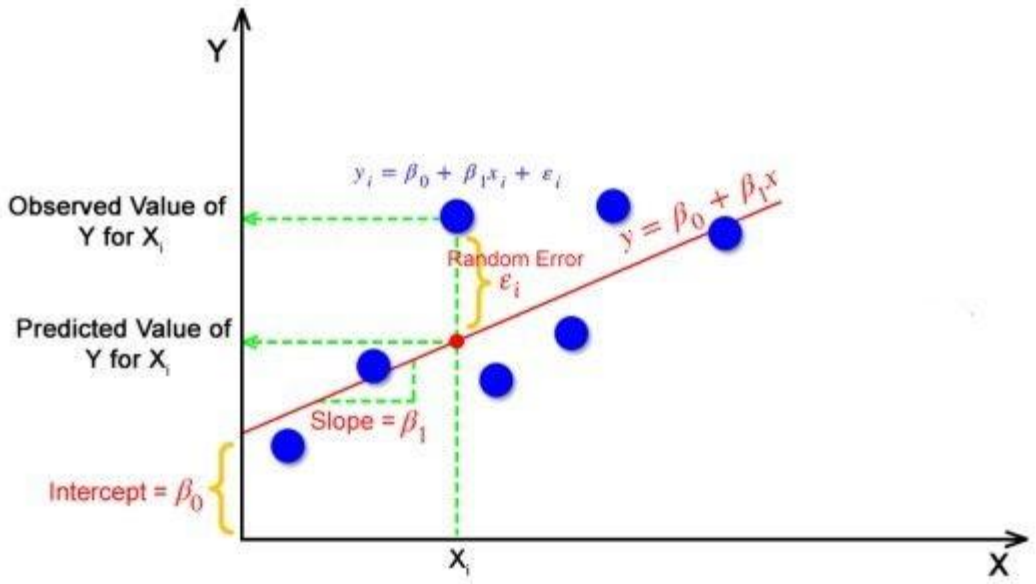
- $x_1, x_2, ..., x_p$ are the independent variables

Goal of Linear Regression:

But how does the regression find out which is the best-fit line?

The goal of the linear regression algorithm is to get the **best values for B 0 and B 1** to find the best-fit line. The best-fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

This is typically done using the **least squares method**, which minimizes the sum of the squared residuals (the differences between the actual values and the predicted values).



<u>Key Concepts and Steps in Multiple Linear Regression</u>

1. **Model Fitting (Ordinary Least Squares - OLS):**
    - Just like in simple linear regression, the OLS method is used to estimate the coefficients by minimizing the sum of the squared residuals (the differences between observed and predicted values).
    - The goal of OLS in this case is to estimate the parameters : β0,β1,…,βn by minimizing the sum of the squared residuals

$$\text{Minimize} \quad \sum_{i=1}^{m} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}))^2$$

2. **Assumptions of Multiple Linear Regression:**
    - Linearity: The relationship between the dependent variable and the independent variables is linear.
    - Independence: The observations are independent of each other.
    - Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.

- Normality: The residuals (errors) are normally distributed.
- No Multicollinearity: The independent variables are not highly correlated with each other.
- No Overfitting: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the overfitting of a model. This usually leads to high training accuracy and very low test accuracy.
- Feature Selection: With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

3. Multicollinearity

- Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a multiple regression model are highly correlated, making it difficult to assess the individual effects of each variable on the dependent variable.
- Detecting Multicollinearity includes two techniques:
- Correlation Matrix: Examining the correlation matrix among the independent variables is a common way to detect multicollinearity. High correlations (close to 1 or -1) indicate potential multicollinearity.
- VIF (Variance Inflation Factor): VIF is a measure that quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A high VIF (typically above 10) suggests multicollinearity.

4. **Evaluation Metrics for Linear Regression**

The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

Root Mean Squared Error: The Root Mean Squared Error is the square root of the variance of the residuals. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values. Mathematically it can be represented as,

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n} \left(y_i^{Actual} - y_i^{Predicted}\right)^2 / n}$$

To make this estimate unbiased, one has to divide the sum of the squared residuals by the degrees of freedom rather than the total number of data points in the model. This term is then called the Residual Standard Error (RSE). Mathematically it can be represented as,

$$RSE = \sqrt{\frac{RSS}{df}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 \Big/ (n-2)}$$

Coefficient of Determination or R-Squared (R2)

R-squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1. Overall, the higher the value of R-squared, the better the model fits the data.

Mathematically it can be represented as,

$$R^2 = 1 - (RSS/TSS)$$

- **Residual sum of Squares (RSS)** is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.

$$RSS = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

- **Total Sum of Squares (TSS)** is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

$$TSS = \sum (y_i - \bar{y}_i)^2$$

where y hat is the mean of the sample data points.

R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables (i.e. it is not a normalized measure), it can change with the change in the unit of the variables.

Why Linear Regression is Important?

Linear regression is important for a few reasons:

- ***Simplicity and interpretability***: It's a relatively easy concept to understand and apply. The resulting simple linear regression model is a straightforward equation that shows how one variable affects another. This makes it easier to explain and trust the results compared to more complex models.
- ***Prediction***: Linear regression allows you to predict future values based on existing data. For instance, you can use it to predict sales based on marketing spend or house prices based on square footage.
- ***Foundation for other techniques***: It serves as a building block for many other data science and machine learning methods. Even complex algorithms often rely on linear regression as a starting point or for comparison purposes.

- ***Widespread applicability***: Linear regression can be used in various fields, from finance and economics to science and social sciences. It's a versatile tool for uncovering relationships between variables in many real-world scenarios.

In essence, linear regression provides a solid foundation for understanding data and making predictions. It's a cornerstone technique that paves the way for more advanced data analysis methods.

2. Explain the Anscombe's quartet in detail?

Anscombe's quartet comprises a set of **four datasets, that have nearly identical simple descriptive statistics (mean, variance, correlation coefficient, and linear regression parameters**), yet they exhibit very different distributions and patterns when graphed. It was created by the statistician **Francis Anscombe** in 1973 to illustrate the importance of visualizing data before relying on statistical summaries alone.

The Four Datasets:

- **Dataset I:**

This dataset shows a roughly linear relationship between x and y, which is what you would expect from the summary statistics. When plotted, it forms a straight line with some random noise.

- **Dataset II:**

This dataset has a clear non-linear relationship. All the x-values are the same except for one outlier, which skews the summary statistics. When plotted, it shows a parabolic curve.

- **Dataset III:**

This dataset has a linear relationship between x and y, but it contains an outlier that strongly influences the regression line. When plotted, the points are mostly in a vertical line with one outlier far to the right.

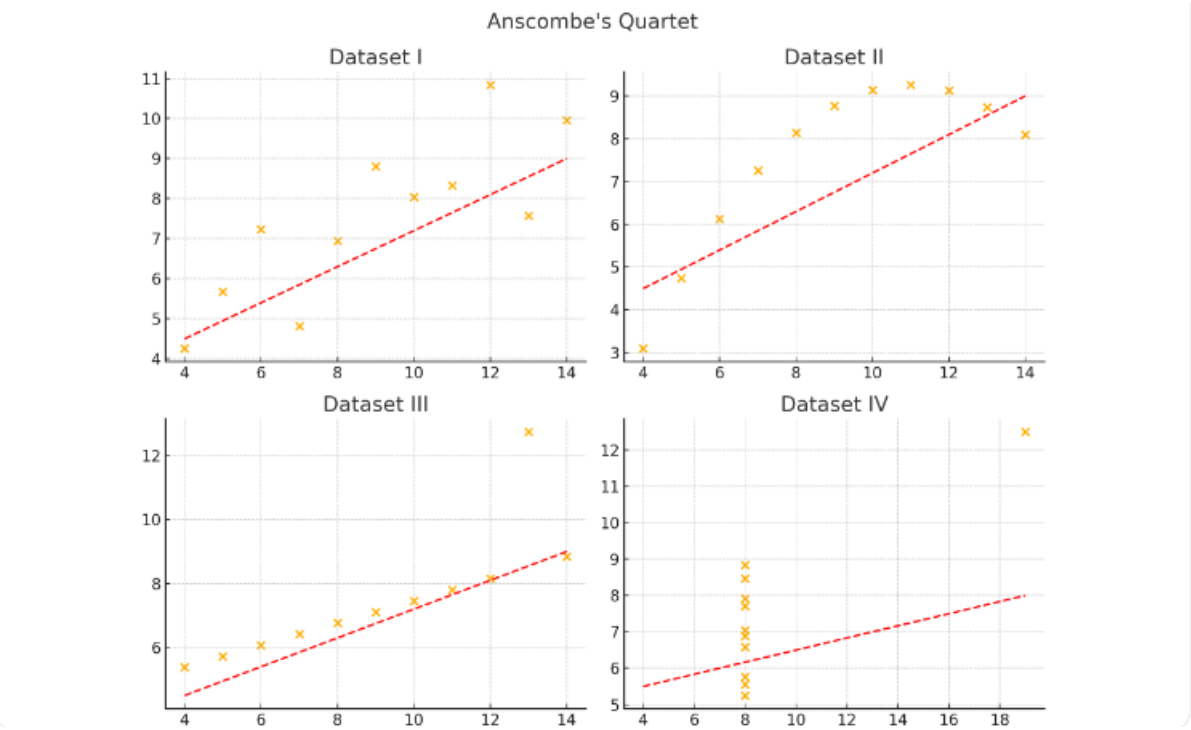
- **Dataset IV:**

This dataset consists of mostly identical y-values, except for one outlier. The correlation and regression statistics are similar to the others, but the plot shows a horizontal line of points with one outlier far above.

Key Points:

- ***Visual exploration is crucial***: While statistical measures can provide valuable insights, they alone may not capture the full picture. Visualizing the data can reveal patterns, outliers, and non-linear relationships that might be missed by numerical summaries.
- ***Outliers can have a significant impact***: A single outlier can dramatically change the regression line and the overall interpretation of the relationship between the variables.
- ***Non-linear relationships***: Statistical measures based on linear models might not be appropriate for non-linear relationships.

- ***Correlation does not imply causation***: Even with a high correlation coefficient, it's essential to consider other factors and potential confounding variables before drawing causal conclusions.



Implications:

- **Data visualization:** Always visualize your data to identify potential patterns, outliers, and non-linear relationships.
- **Critical thinking:** Don't rely solely on statistical measures without considering the underlying data and its context.
- **Multiple perspectives:** Explore different visualizations and statistical analyses to gain a comprehensive understanding of the data.

3. What is Pearson's R?

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.

- r = 1: Perfect positive correlation, meaning the variables increase or decrease together perfectly.
- r = -1: Perfect negative correlation, meaning one variable increases as the other decreases perfectly.
- r = 0: No correlation between the variables.

Interpretation:

- **Strength**: The absolute value of r indicates the strength of the relationship:
    - |r| close to 1: Strong correlation
    - |r| close to 0: Weak correlation
- **Direction**: The sign of r indicates the direction of the relationship:
    - Positive r: The variables increase or decrease together
    - Negative r: One variable increases as the other decreases

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.

Use Cases:

- **Identifying relationships:** Pearson's correlation is used to determine if there's a linear relationship between two variables.
- **Measuring strength of relationships:** It quantifies the strength of the relationship.
- **Understanding direction:** It indicates whether the relationship is positive or negative.

Formula

$$r = \Sigma((x - \bar{x})(y - \bar{y})) / \sqrt{(\Sigma(x - \bar{x})^2 * \Sigma(y - \bar{y})^2)}$$

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

Scaling is a technique used in data preprocessing to transform numerical data to a common range or scale. This is often necessary because different variables in a dataset may have vastly different units or magnitudes, which can lead to issues during modeling.

Why is scaling performed?

- **Improving model performance:** Many machine learning algorithms assume that features are centered around zero and have comparable scales. If features are on vastly different scales, some features may dominate others in the model, leading to suboptimal performance.
- **Ensuring fair comparison:** When comparing variables on different scales, scaling ensures that all variables contribute equally to the model.
- **Interpreting coefficients:** In some cases, scaling can make it easier to interpret the coefficients of a model.

Types of Scaling:

1. **Normalized scaling (Min-Max Scaling)**

    1. Normalized scaling, also known as Min-Max Scaling, transforms the data to a specific range, usually between 0 and 1.
    2. The formula for min-max scaling is:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

    Where:

    - X is the original feature value.
    - Xmin is the minimum value of the feature.
    - Xmax is the maximum value of the feature.

    After applying min-max scaling, all feature values will be within the range [0, 1].

    When to Use Normalized Scaling:

    1. When the data does not contain outliers, as min-max scaling is sensitive to outliers.
    2. When we want to scale the data to a fixed range, such as [0, 1], which can be useful in algorithms that require features to be within a specific range, like neural networks and image processing tasks.

2. **Standardized scaling: (z-score scaling)**

    1. **Standardized scaling**, also known as Z-score Normalization, transforms the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Where:

- X is the original feature value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

After applying standardization, the feature values will be centered around zero with a standard deviation of one.

When to Use Standardized Scaling:

- When the data contains outliers, as standardization is less sensitive to outliers compared to min-max scaling.
- When the data follows a Gaussian (normal) distribution, as standardization works well with normally distributed data.
- When the algorithm assumes that data is centred around zero, such as in Principal Component Analysis (PCA), linear regression, and logistic regression.

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 detect multicollinearity in a regression model. Specifically, VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors in the model.

A VIF of 1 indicates no multicollinearity, while a VIF greater than 1 suggests increasing levels of multicollinearity.

When VIF becomes infinite, it essentially means that one of the predictor variables can be perfectly predicted by a linear combination of the other predictors

Understanding VIF

The VIF for a predictor Xi is calculated as:

$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

Where Ri^2 is the coefficient of determination (R-squared) obtained by regressing Xi on all the other predictors.

Why Does VIF Become Infinite?

The value of VIF becomes infinite when $Ri^2=1$. This occurs under the following condition:

Perfect Multicollinearity: Perfect multicollinearity arises when one predictor variable is a perfect linear combination of one or more other predictors. In this case, the predictor Xi can be expressed exactly as a linear combination of the other predictors. As a result, when you try to regress Xi on the other predictors, the regression fit will be perfect, and $Ri^2=1$.

Mathematically, this means that the denominator in the VIF formula, $1- Ri^2$, becomes zero. Since division by zero is undefined, the VIF value is considered infinite (or extremely large in practical computation).

Implications of Infinite VIF

1. **Unstable Coefficient Estimates**
2. **Model Interpretation Issues**
3. **Possible Solutions**: To address the issue of infinite VIF, you can:
    - **Remove one of the collinear variables**: If two or more predictors are perfectly correlated, consider removing one of them from the model.
    - **Combine the variables**: In some cases, you might create a composite variable by combining collinear variables.
    - **Use regularization techniques**: Ridge regression and Lasso regression are regularization methods that can handle multicollinearity by adding a penalty to the size of the coefficients.

Example:

In the current bike sharing case study, we can see that 'holiday' and 'workingday' has VIF values as infinity which indicates perfect multicollinearity.

It means 'holiday' and 'workingday' are perfectly correlated with other variables in the dataset.

	Feature	VIF
0	const	0.000000
1	yr	1.059334
2	holiday	inf
3	workingday	inf

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically a normal distribution. The plot helps to assess whether the data follows the theoretical distribution or not.

In a Q-Q plot:

- The x-axis represents the theoretical quantiles (i.e., the quantiles of the theoretical distribution you are comparing against, usually a normal distribution).
- The y-axis represents the empirical quantiles (i.e., the quantiles of the observed data).

If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight diagonal line (45-degree line).

How a Q-Q Plot is Constructed

1. Sort the Data: The observed data points are sorted in ascending order.
2. Determine the Theoretical Quantiles: The corresponding quantiles of the theoretical distribution are calculated.
3. Plot the Points: Each data point is plotted against the corresponding theoretical quantile. If the data follows the theoretical distribution, the points will form a straight line.

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

In linear regression, one of the key assumptions is that the residuals (errors) of the model are normally distributed. This assumption is critical because:

- Normality of Residuals: The normality of residuals is an assumption underlying various statistical tests (e.g., t-tests) used in regression analysis. If residuals are not normally distributed, the results of these tests (such as p-values) may not be valid.
- Validity of Confidence Intervals: The accuracy of confidence intervals and prediction intervals also depends on the assumption that residuals are normally distributed.

How Q-Q Plots Are Used in Linear Regression:

1. Assessing Normality of Residuals: In linear regression, after fitting the model, a Q-Q plot is used to assess whether the residuals follow a normal distribution. If the residuals follow a normal distribution, they should fall along the 45-degree reference line in the Q-Q plot.
    - Straight Line: If the points form a straight line, this indicates that the residuals are normally distributed, and the assumption of normality is satisfied.
    - S-shaped Curve: If the points deviate from the line in an "S" shape, this indicates that the residuals are skewed, either positively or negatively.
    - Heavy Tails: If the points at the ends of the plot deviate from the line, this suggests that the distribution has heavier or lighter tails than a normal distribution.
2. Identifying Outliers: Extreme deviations from the line in the Q-Q plot can indicate outliers in the residuals. These outliers may have a disproportionate influence on the model, leading to biased estimates and predictions.
3. Model Diagnostics: A Q-Q plot is a useful diagnostic tool to evaluate whether the linear regression model is appropriate for the data. If the residuals are not normally distributed, it might suggest that the model needs to be adjusted, for example, by transforming the dependent variable or using a different modeling approach.

Importance of Q-Q Plot in Linear Regression

1. Assumption Checking: It helps in checking the normality assumption of residuals, which is crucial for hypothesis testing and confidence interval estimation.
2. Model Validation: The Q-Q plot provides a visual validation of the model assumptions, helping to ensure that the model is reliable and the inferences drawn are valid.
3. Identifying Issues: It helps identify potential issues such as non-normality, skewness, heavy tails, and outliers, which can affect the model's performance and interpretation.
4. Decision-Making: Based on the Q-Q plot, we can decide whether to proceed with the current model or consider alternative approaches, such as transforming variables or using a different model.