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?**
2.

To check the effect of the categorical variables on the target, bar plots were created which revealed the following:

1. Season

The `season` variable significantly affects the demand for bikes. The bar plot shows that bike rentals are highest during the summer and fall seasons, while the demand drops during winter and spring. This pattern indicates that warmer weather conditions in summer and fall encourage more bike usage.

2. Year (`yr`)

The `yr` variable, which distinguishes between the years 2018 and 2019, shows a noticeable increase in bike rentals in 2019 compared to 2018. This suggests that the popularity and adoption of bike-sharing services have increased over time.

3. Month (`mnth`)

The `mnth` variable affects bike demand, with higher rentals observed during warmer months like June, July, and August. Conversely, colder months such as January and February see lower bike rentals. This trend aligns with the seasonal variation in bike demand.

4. Holiday

The `holiday` variable shows that bike rentals are slightly lower on holidays compared to regular days. This may be because people use bike-sharing services more for commuting to work rather than leisure on holidays.

5. Weekday

The `weekday` variable indicates that bike rentals are fairly consistent throughout the week, with a slight dip on weekends. This suggests that bike-sharing services are used more frequently for weekday commutes.

6. Weather Situation (`weathersit`)

The `weathersit` variable significantly impacts bike rentals. Clear weather conditions see the highest bike rentals, while adverse weather conditions like heavy rain or snow lead to a significant drop in bike usage. This is expected as people are less likely to use bikes in poor weather conditions.

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

When creating dummy variables for categorical features, it's important to use `drop_first=True` to avoid the dummy variable trap. Here’s a detailed explanation of why this is necessary:

The Dummy Variable Trap

The dummy variable trap refers to the situation where there is multicollinearity in the model due to the inclusion of all dummy variables for a categorical feature. Multicollinearity occurs when two or more predictor variables are highly correlated, meaning that one variable can be predicted from the others. This can lead to several issues in regression models, such as:

- Unreliable Coefficient Estimates: The regression coefficients become unstable and can change significantly with small changes in the model or data.

- Difficulty in Interpretation: It becomes hard to determine the individual effect of each predictor variable on the target variable.

- Inflated Standard Errors: Standard errors of the coefficients increase, leading to wider confidence intervals and less statistically significant results.

Avoiding the Dummy Variable Trap

To avoid the dummy variable trap, we use `drop_first=True` when creating dummy variables. This approach drops the first category of each categorical variable, which serves as a reference category. By doing this, we prevent multicollinearity and ensure that the model is correctly specified.

Example

Consider a categorical variable `season` with four categories: spring, summer, fall, and winter. Without `drop_first=True`, creating dummy variables would result i

- `season_spring`

- `season_summer`

- `season_fall`

- `season_winter`

Including all four dummy variables in the regression model would lead to multicollinearity because knowing the values of three of them would automatically determine the value of the fourth.

With `drop_first=True`, one of the categories (e.g., `season_spring`) is dropped, and the remaining dummy variables represent comparisons against this reference category

- `season_summer`

- `season_fall`

- `season_winter`

Now, the model will use `season_spring` as the baseline, and the coefficients of the remaining dummy variables will indicate the effect of being in those seasons relative to spring,

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

1. Pair-Plots Visualiz

The pair-plots allow us to visually inspect the relationships between the numerical variables and the target variable (`cnt`). We can see the scatter plots and the distribution of each variable.

2. Correlation Matrix:

The correlation matrix quantifies the linear relationships between each pair of variables.

```
               cnt      temp     atemp       hum  windspeed
cnt       1.000000  0.627494  0.631066  0.126963  -0.234545
temp      0.627494  1.000000  0.991701  0.126963  -0.157944
atemp     0.631066  0.991701  1.000000  0.139988  -0.183643
hum       0.126963  0.126963  0.139988  1.000000  -0.248489
windspeed -0.234545 -0.157944 -0.183643 -0.248489  1
```

3. Variable with Highest Correlation:

From the correlation matrix, we can see that `atemp` (feeling temperature) has the highest correlation with `cnt` (total bike rentals), with a correlation coefficient of approximately 0.631.

The numerical variable with the highest correlation with the target variable (`cnt`) is `atemp` (feeling temperature).

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

1. Linearity:
   - The plot of residuals vs. fitted values should show no clear pattern. The residuals should be randomly scattered around zero, indicating a linear relationship.
2. Normality:

    - The Q-Q plot compares the distribution of residuals to a normal distribution. If the residuals follow the reference line closely, they are approximately normally distributed.
3. Homoscedasticity:
    - In the residuals vs. fitted values plot, the spread of residuals should be constant across all levels of fitted values. No funnel shape or pattern should be visible.
4. Independence:
    - This assumption is checked based on the data collection process. Ensure that the data points are independent of each other.
5. Multicollinearity:
    - Calculate VIF for each predictor. A VIF value greater than 10 indicates high multicollinearity. The VIF values printed in the output should be below this threshold.

By performing these validation steps, we ensure that the linear regression model's assumptions are met, resulting in a more reliable and interpretable model. The residual analysis, Q-Q plot, and VIF calculations provide insights into the model's adequacy and highlight any potential issues with the assumptions

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

yr_2019	2001.438674
season_winter	821.498930
atemp	801.162869

General Subjective Questions

1. **Explain Linear Regression in detail ?**

Linear Regression Algorithm: A Detailed Explanation

Key Concepts of Linear Regression

1. Types of Linear Regression:
   - Simple Linear Regression : Involves a single independent variable.
   - Multiple Linear Regression -: Involves two or more independent variables.

2. Model Representation :
   - The linear regression model can be represented as:
     \[
     y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon
     \]
     where:
     - \(y \) is the dependent variable.
     - \(\beta_0 \) is the intercept.
     - \(\beta_1, \beta_2, \cdots, \beta_n \) are the coefficients for the independent variables \(x_1, x_2, \cdots, x_n \).
     - \(\epsilon \) is the error term (residual).

3. Assumptions of Linear Regression :
   - Linearity : The relationship between the dependent and independent variables is linear.
   - Independence : Observations are independent of each other.
   - Homoscedasticity : Constant variance of the error terms.

- Normality : The error terms are normally distributed.
- No multicollinearity : Independent variables are not highly correlated with each other.

Steps in Linear Regression

1. Data Collection :
   - Gather data that includes the dependent variable and one or more independent variables.

2. Data Preprocessing :
   - Handle missing values.
   - Convert categorical variables to numerical using techniques like one-hot encoding.
   - Scale the data if necessary.

3. Splitting the Data :
   - Split the data into training and testing sets to evaluate the model's performance.

4. Fitting the Model :
   - Use the training data to fit the linear regression model. This involves finding the best-fit line that minimizes the sum of squared residuals (the difference between observed and predicted values).

5. Evaluating the Model :
   - Use metrics like R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to evaluate the model's performance on the test data.

6. Making Predictions :
   - Use the model to make predictions on new data.

Mathematical Details

1. Cost Function :

   - The cost function used in linear regression is the Mean Squared Error (MSE):

     \[
     J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\beta(x^{(i)}) - y^{(i)})^2
     \]

     where \(m \) is the number of training examples, \(h_\beta(x^{(i)}) \) is the predicted value, and \(y^{(i)} \) is the actual value.

2. Gradient Descent :

   - An optimization algorithm used to minimize the cost function by iteratively updating the coefficients:

     \[
     \beta_j := \beta_j - \alpha \frac{\partial J(\beta)}{\partial \beta_j}
     \]

     where \(\alpha \) is the learning rate.

3. Normal Equation :

   - An analytical method to directly compute the coefficients:

     \[
     \beta = (X^T X)^{-1} X^T y
     \]

     where \(X \) is the matrix of input features, and \(y \) is the vector of output values.

Assumptions Validation

1. Linearity:

- Check if the relationship between the independent variables and the dependent variable is linear.
- Plot residuals vs. fitted values (predicted values) to see if the residuals are randomly distributed around zero.

2. Normality:
   - Check if the residuals are normally distributed.
   - Use a Q-Q plot (quantile-quantile plot) to visually inspect if the residuals follow a normal distribution.

3. Homoscedasticity:
   - Check if the residuals have constant variance.
   - Plot residuals vs. fitted values to see if the spread of residuals is constant across all levels of fitted values.

4. Independence:
   - Ensure that the residuals are independent of each other.
   - This assumption is usually checked based on the data collection process rather than residual plots.

5. Multicollinearity:
   - Ensure that the independent variables are not highly correlated.
   - Use Variance Inflation Factor (VIF) to check for multicollinearity.

Model Evaluation Metrics

1. R-squared (R^2) :
   - Represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. An R^2 value of 1 indicates a perfect fit, while an R^2 value of 0 indicates that the model does not explain any of the variance in the target variable.

2. Mean Absolute Error (MAE) :
 - Measures the average magnitude of the errors in a set of predictions, without considering their direction.

3. Mean Squared Error (MSE) :
 - Measures the average squared difference between the observed actual outcomes and the predictions.

4. Root Mean Squared Error (RMSE) :
 - The square root of the mean squared error, providing a measure of the average magnitude of the prediction errors.

2. Explain the Anscombe's quartet in detail.

Introduction

Anscombe's quartet, named after the statistician Francis Anscombe, is a collection of four datasets that have nearly identical simple descriptive statistics but exhibit very different distributions and graphical properties. It was constructed to demonstrate the importance of visualizing data before performing statistical analysis.

Key Characteristics

1. Identical Statistical Properties :
 - Each dataset in Anscombe's quartet has the same mean and variance for both \(x \) and \(y \) variables.
 - They share the same linear regression line.
 - They have the same correlation coefficient between \(x \) and \(y \).

2. Different Distributions :
 - Despite the identical statistical properties, each dataset has a distinct distribution when graphed.

The Four Datasets

Each of the four datasets (I, II, III, IV) consists of eleven \((x, y) \) points, and they all share the following statistical properties:

- Mean of \(x \) : 9
- Mean of \(y \) : 7.5
- Variance of \(x \) : 11
- Variance of \(y \) : 4.12
- Correlation between \(x \) and \(y \) : 0.816
- Linear regression line : \(y = 3 + 0.5x \)

Detailed Descriptions

1. Dataset I :
   - This dataset forms a standard linear relationship between \(x \) and \(y \).
   - When plotted, it appears as a typical scatter plot with points closely aligned along the regression line.

2. Dataset II :
   - This dataset consists of a clear non-linear relationship.
   - When plotted, the data points form a parabolic shape.
   - A linear regression model does not adequately describe this dataset.

3. Dataset III :
   - This dataset contains an outlier in the \(x \) values.
   - The majority of points lie along a horizontal line, but one outlier significantly affects the regression line.
   - Without the outlier, the relationship between \(x \) and \(y \) would be almost non-existent.

4. Dataset IV :
- This dataset has an outlier in the \(y \) values.
- Most data points have the same \(x \) value and lie along the regression line, but one outlier skews the correlation.

Importance of Anscombe's Quartet

1. Visualization :
- Anscombe's quartet emphasizes the necessity of visualizing data before analyzing it. Graphical representations can reveal patterns, relationships, and anomalies that are not apparent from summary statistics alone.

2. Misleading Statistics :
- The quartet demonstrates how relying solely on statistical properties can be misleading. Identical means, variances, and correlations can mask fundamentally different data distributions.

3. Robust Analysis :
- It highlights the importance of using robust statistical methods and diagnostics to understand the underlying data structure.

Visual Representation

To fully appreciate Anscombe's quartet, one should visualize the datasets. Here are the typical scatter plots for each dataset:

- Dataset I : A linear scatter plot with points closely clustered around the regression line.
- Dataset II : A parabolic scatter plot showing a clear non-linear relationship.

- Dataset III : A scatter plot with a horizontal line of points and one influential outlier in the \(x \) direction.
- Dataset IV : A scatter plot with points along the regression line and one significant outlier in the \(y \) direction.

---

3. What is Pearsons R ?

Introduction

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which two variables move in relation to each other. Named after Karl Pearson, it is widely used in statistics to determine the strength and direction of a linear relationship between two continuous variables.

Key Characteristics

1. Range :
   - The value of Pearson’s R ranges between -1 and 1.
   - \(R = 1 \) indicates a perfect positive linear relationship.
   - \(R = -1 \) indicates a perfect negative linear relationship.
   - \(R = 0 \) indicates no linear relationship.

2. Sign :
   - A positive value indicates that as one variable increases, the other variable also increases.
   - A negative value indicates that as one variable increases, the other variable decreases.

3. Magnitude :
   - The closer the value of R to 1 or -1, the stronger the linear relationship between the two variables.
   - The closer the value of R to 0, the weaker the linear relationship between the two variables.

Calculation

Pearson’s R is calculated using the following formula:

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

where:
- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

Interpretation

1. Perfect Positive Correlation (R = 1) :
   - Every increase in one variable is associated with a proportionate increase in the other variable.
   - Points lie exactly on a straight line with a positive slope.

2. Perfect Negative Correlation (R = -1) :
   - Every increase in one variable is associated with a proportionate decrease in the other variable.
   - Points lie exactly on a straight line with a negative slope.

3. No Correlation (R = 0) :

- There is no linear relationship between the two variables.
- Points are scattered with no discernible pattern.

4. Moderate Correlation ($0 < |R| < 1$) :
   - Indicates the degree to which one variable tends to increase or decrease as the other variable increases.
   - Points lie closer to the line of best fit as $|R|$ approaches 1.

ent of each other.

Applications

1. Exploratory Data Analysis :
   - Pearson's R is often used in the initial stages of data analysis to identify potential relationships between variables.

2. Hypothesis Testing :
   - It is used in hypothesis testing to determine if there is a significant linear relationship between two variables.

3. Regression Analysis :
   - Pearson's R helps in assessing the goodness of fit for regression models by quantifying the strength of the linear relationship between the dependent and independent variables.

Example Interpretation

Suppose we have two variables, \(X \) and \(Y \):

- R = 0.85 : Strong positive linear relationship.
- R = -0.75 : Strong negative linear relationship.
- R = 0.30 : Weak positive linear relationship.
- R = -0.10 : Weak negative linear relationship.

- R = 0 : No linear relationship.

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

Scaling is the process of transforming the features of your data so that they are on a similar scale. This is particularly important in machine learning algorithms that rely on the distance between data points, such as gradient descent-based methods, support vector machines, and k-nearest neighbors. Without scaling, features with larger ranges can dominate the model and lead to suboptimal performance.

Why is Scaling Performed?

1. Improving Model Performance :
   - Algorithms that involve distance calculations or gradient descent optimization (e.g., linear regression, logistic regression, neural networks) can perform better and converge faster with scaled features.

2. Ensuring Equal Importance :
   - Features with larger magnitudes can dominate the learning process if not scaled. Scaling ensures that all features contribute equally to the model.

3. Enhancing Numerical Stability :
   - Scaling can help prevent numerical instability in algorithms that perform complex calculations by keeping the values within a similar range.

4. Improving Interpretability :
   - Scaling can make it easier to interpret the importance of different features, especially when dealing with coefficients in regression models.

Types of Scaling: Normalized Scaling vs. Standardized Scaling

# Normalized Scaling (Min-Max Scaling)

1. Definition :
   - Normalization scales the features to a fixed range, usually between 0 and 1.

2. Formula :
   \[
   x' = \frac{x - \min(x)}{\max(x) - \min(x)}
   \]
   - \(x \): Original value
   - \(x' \): Scaled value
   - \(\min(x) \): Minimum value in the feature
   - \(\max(x) \): Maximum value in the feature

3. When to Use :
   - When you want to maintain the relationships in the original data, such as for algorithms like k-nearest neighbors and neural networks.
   - When the data does not contain outliers, as normalization can be sensitive to them.

4. Example :
   - A feature with values ranging from 10 to 100 would be scaled to a range of 0 to 1.

# Standardized Scaling (Z-score Standardization)

1. Definition :
   - Standardization scales the features to have a mean of 0 and a standard deviation of 1.

2. Formula :
\[
x' = \frac{x - \mu}{\sigma}
\]
- \(x \): Original value
- \(x' \): Scaled value
- \(\mu \): Mean of the feature
- \(\sigma \): Standard deviation of the feature

3. When to Use :
- When the algorithm assumes that the data is normally distributed, such as linear regression, logistic regression, and linear discriminant analysis.
- When dealing with data that has outliers, as standardization is less sensitive to outliers compared to normalization.

4. Example :
- A feature with values centered around 50 with a standard deviation of 10 would be scaled so that the transformed values have a mean of 0 and a standard deviation of 1.

Differences Between Normalized and Standardized Scaling

1. Range :
- Normalization : Transforms data to a fixed range, typically 0 to 1.
- Standardization : Transforms data to have a mean of 0 and a standard deviation of 1.

2. Sensitivity to Outliers :
- Normalization : Sensitive to outliers, as the minimum and maximum values can significantly affect the scaling.
- Standardization : Less sensitive to outliers, as it focuses on the mean and standard deviation.

3. Use Cases :

- Normalization : Preferred when the distribution of the data is not Gaussian or when you want to preserve the relationship between the original data values.
- Standardization : Preferred when the data is normally distributed or for algorithms that assume normally distributed data.

–

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

Introduction to VIF

Variance Inflation Factor (VIF) is a measure of multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to the presence of collinear (correlated) independent variables. VIF is calculated for each independent variable as follows:

\[
VIF_i = \frac{1}{1 - R_i^2}
\]

where \(R_i^2 \) is the coefficient of determination of the regression model that predicts the \(i \)-th independent variable using all other independent variables.

Why VIF Becomes Infinite

VIF becomes infinite (or very large) when there is perfect multicollinearity, meaning that one independent variable is an exact linear combination of one or more other independent variables. This can occur due to several reasons:

1. Perfect Multicollinearity :
   - Perfect multicollinearity arises when one independent variable can be perfectly predicted from the others. In such cases, R_i^2 becomes 1, and the denominator of the VIF formula becomes zero, leading to an infinite VIF value.

2. Duplicate Variables :
   - Including duplicate or nearly duplicate variables in the dataset can cause perfect multicollinearity. For example, if two columns in the dataset contain identical values, their VIF will be infinite.

3. Dummy Variable Trap :
   - The dummy variable trap occurs when all categories of a categorical variable are included as separate dummy variables without dropping one as a reference category. This leads to perfect multicollinearity because the dummy variables sum to one, making one variable a perfect linear combination of the others.

4. Linear Dependence :
   - If one variable is a linear combination of others (e.g., $x_3 = 2x_1 + 3x_2$), it will result in perfect multicollinearity, causing infinite VIF values.

Example Scenarios

1. Duplicate Variables :
   - Suppose you have a dataset with variables x_1 and x_2, and x_2 is an exact duplicate of x_1. The regression model for x_2 using x_1 will have $R_i^2 = 1$, leading to an infinite VIF for x_1 and x_2.

2. Dummy Variable Trap :
   - If a categorical variable "Color" has three categories: Red, Green, and Blue, and we create three dummy variables without dropping one, the resulting VIFs will be infinite. This is because one dummy variable can be perfectly predicted using the other two (e.g., \(Red = 1 - Green - Blue \)).

How to Address Infinite VIF

1. Remove Perfectly Collinear Variables :
   - Identify and remove one of the perfectly collinear variables to eliminate perfect multicollinearity.

2. Drop One Dummy Variable :
   - In the case of dummy variables, drop one category to avoid the dummy variable trap. This is typically done by using the `drop_first=True` parameter in pandas' `get_dummies()` function.

3. Examine the Data :
   - Carefully examine the data for any linear dependencies or duplicate columns and address them appropriately.

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a specified distribution, typically the normal distribution. The Q-Q plot compares the quantiles of the sample data with the quantiles of a theoretical distribution. If the data follows the specified distribution, the points on the Q-Q plot will approximately lie on a straight line.

Components of a Q-Q Plot

1. Sample Quantiles :
   - Quantiles calculated from the observed data.

2. Theoretical Quantiles :
   - Quantiles calculated from the specified theoretical distribution, often the normal distribution.

3. Reference Line :
   - A 45-degree line (y = x) that represents perfect agreement between the sample and theoretical quantiles. If the data follows the theoretical distribution, the points should lie close to this line.

Steps to Create a Q-Q Plot

1. Sort the Data :
   - Sort the observed data in ascending order.

2. Calculate Sample Quantiles :
   - Compute the quantiles of the observed data.

3. Calculate Theoretical Quantiles :
   - Compute the quantiles from the theoretical distribution (e.g., normal distribution) corresponding to the same probabilities as the sample quantiles.

4. Plot the Quantiles :
   - Plot the sample quantiles on the y-axis against the theoretical quantiles on the x-axis.

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

In the context of linear regression, a Q-Q plot is primarily used to check the assumption that the residuals (errors) of the model are normally distributed. This is important because many statistical tests and confidence intervals for linear regression coefficients rely on the normality assumption.

1. Assumption Checking :
   - Normality of Residuals : Linear regression assumes that the residuals are normally distributed. A Q-Q plot helps to visually assess this assumption. If the residuals are normally distributed, the points will lie approximately along the reference line.
   - Identifying Deviations : Deviations from the reference line in a Q-Q plot indicate departures from normality. For example, systematic deviations such as an S-shaped pattern might suggest heavy tails, skewness, or other non-normal characteristics.

2. Model Diagnostics :
   - Detecting Outliers : Points that fall far from the reference line, especially at the ends of the plot, may indicate outliers in the data.
   - Assessing Fit : A Q-Q plot can be used to evaluate the goodness-of-fit of a linear regression model. If the residuals deviate significantly from normality, it might suggest that the model is not adequately capturing the underlying relationship in the data.

Interpreting a Q-Q Plot

1. Linear Pattern :
   - If the points form a straight line, the data is consistent with the specified distribution. For a normal Q-Q plot, this indicates normality.

2. S-shaped Curve :
   - An S-shaped curve suggests heavy tails in the data. The data has more extreme values than expected under the normal distribution.

3. Inverted S-shaped Curve :

- An inverted S-shaped curve suggests light tails in the data. The data has fewer extreme values than expected under the normal distribution.

4. Curvature :
   - Curvature away from the line indicates skewness. Points bending upwards suggest positive skewness, while points bending downwards suggest negative skewness.

5. Outliers :
   - Points that lie far from the line, especially at the extremes, indicate potential outliers.

Example Interpretation

Suppose you have a Q-Q plot for the residuals of a linear regression model:

- Straight Line : Indicates that the residuals are normally distributed, satisfying the normality assumption of linear regression.
- Upward Curve : Indicates that the residuals have heavy tails (positive kurtosis).
- Downward Curve : Indicates that the residuals have light tails (negative kurtosis).
- Asymmetrical Deviations : Suggest skewness in the residuals.