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

Analysis of Categorical Variables: From the examination of categorical columns using box plots and bar plots, several insights can be drawn:

- The fall season appears to attract more bookings, with a noticeable increase in booking counts from 2018 to 2019 across all seasons.
- Peak booking months are May, June, July, August, September, and October, showing an increasing trend from the beginning of the year until mid-year, followed by a decrease towards the year-end.
- Clear weather conditions tend to result in higher booking numbers, as expected.
- Thursday, Friday, Saturday, and Sunday exhibit higher booking counts compared to the start of the week.
- Non-holidays see fewer bookings, aligning with the expectation that people may prefer spending time at home with family during holidays.
- Booking frequencies appear similar on working and non-working days.
- The year 2019 witnessed a significant increase in bookings compared to the previous year, indicating positive progress in business.

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

Importance of drop_first=True in Dummy Variable Creation: The utilization of **drop_first=True** is crucial during dummy variable creation as it eliminates the redundant column generated, thus reducing correlations among dummy variables. By excluding the first level, this parameter helps prevent multicollinearity issues among the created dummy variables.

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

Highest Correlation with Target Variable in Pair-Plot: The 'temp' variable demonstrates the highest correlation with the target variable in the pair-plot among numerical variables.

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

Validation of Linear Regression Assumptions: To validate the assumptions of the Linear Regression model, the following five aspects were considered:

- **Normality of Error Terms:** Ensuring that error terms are normally distributed.
- **Multicollinearity Check:** Verifying the absence of significant multicollinearity among variables.
- **Linear Relationship Validation:** Confirming the presence of linearity among variables.
- **Homoscedasticity:** Ensuring no discernible pattern in residual values.
- **Independence of Residuals:** Confirming the absence of autocorrelation.

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 Bike Demand in Final Model: The three most influential features contributing significantly to explaining the demand for shared bikes in the final model are:

- Temperature
- Weather
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also known as the dependent variable) based on one or more predictor variables (independent variables). The algorithm models the relationship between the independent variables and the dependent variable as a linear equation. In simple terms, it tries to fit a straight line through the data points to make predictions.

Here's a detailed explanation of the Linear Regression algorithm:

1. Mathematical Representation:

Linear Regression assumes a linear relationship between the independent variables (X) and the dependent variable (Y). The general form of a linear regression equation for a single variable is:

$Y = \beta 0 + \beta 1.X + \epsilon$

- Y is the dependent variable.
- X is the independent variable.
- $\beta 0$ is the y-intercept (constant term).
- $\beta 1$ is the slope of the line.
- ϵ represents the error term, which accounts for unobserved factors affecting Y.

2. Objective:

The primary objective of Linear Regression is to find the values of $\beta 0$ and $\beta 1$ that minimize the sum of squared differences between the predicted (Ypred) and actual (Yactual) values. This process is known as the method of least squares.

3. Training the Model:

- Given a dataset with input features X and corresponding target values Y, the algorithm learns the optimal values of $\beta 0$ and $\beta 1$ during the training phase.
- The optimization process involves adjusting the parameters to minimize the cost function, which quantifies the difference between predicted and actual values.

.

4. Gradient Descent (Optional):

- Optimization techniques like Gradient Descent can be employed to find the minimum of the cost function iteratively.
- The algorithm adjusts the parameters in the opposite direction of the gradient until convergence.

5. Assumptions of Linear Regression:

- **Linearity:** The relationship between variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** Residuals have constant variance.
- **Normality of Residuals:** Residuals are normally distributed.
- **No Multicollinearity:** Predictor variables are not highly correlated.

6. Predictions:

- Once trained, the model can be used to make predictions on new, unseen data.
- Predictions are made by plugging the new input features into the learned linear equation.

7. Evaluation:

- Model performance is assessed using metrics like R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), etc.
- These metrics help quantify how well the model generalizes to new data.

8. Extensions:

- Linear Regression can be extended to Multiple Linear Regression when there are multiple independent variables.
- Regularization techniques like Ridge and Lasso can be applied to prevent overfitting.

Linear Regression is a straightforward yet powerful algorithm, widely used in various fields for tasks such as prediction, forecasting, and understanding relationships between variables.

2. Explain the Anscombe’s quartet in detail

nscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. Despite having vastly different appearances when graphically plotted, these datasets share almost identical statistical properties. The quartet was designed to illustrate the importance of visualizing data rather than relying solely on summary statistics. Here's a detailed explanation:

1. **Datasets Overview:**
    - Anscombe's Quartet consists of four datasets, each containing 11 data points.
    - Each dataset has two variables, labeled as X and Y.
2. **Statistical Properties:**
    - Despite the visual differences in the scatterplots, all four datasets have the same:
        - Mean of X, mean of Y.
        - Variance of X, variance of Y.
        - Correlation coefficient between X and Y.
        - Linear regression line.

3. **Visual Differences:**
    - When the datasets are graphically represented, they demonstrate how different patterns can arise even when summary statistics are nearly identical.
    - For example, one dataset might exhibit a linear relationship, another a quadratic relationship, and so on.
4. **Implications:**
    - The quartet emphasizes the limitations of relying solely on numerical summaries like means, variances, and correlation coefficients to understand the underlying structure of the data.
    - It encourages the use of graphical tools for data exploration and visualization to gain a more comprehensive understanding of the data distribution and patterns.
5. **Educational Significance:**
    - Anscombe's Quartet is often used in statistics education to highlight the importance of visualizing data.
    - It serves as a cautionary example, reminding analysts and researchers that datasets with similar summary statistics may exhibit different patterns when visualized, underlining the need for a holistic approach to data analysis.

In summary, Anscombe's Quartet serves as a powerful illustration of the concept that relying solely on summary statistics may not capture the full complexity of a dataset. Visualization plays a crucial role in understanding the nuances and patterns within the data.

3. What is Pearson's R? (3 marks)

Pearson's R:

1. **Definition:**
    - Pearson's correlation coefficient, denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.
    - It assesses how much one variable changes in relation to another.
2. **Range and Interpretation:**
    - r ranges from -1 to 1.
        - $r=1$ indicates a perfect positive linear correlation.
        - $r=-1$ indicates a perfect negative linear correlation.
        - $r=0$ indicates no linear correlation.
    - The sign indicates the direction of the correlation: positive or negative.
3. **Calculation:**
    - Pearson's r is computed using the covariance of the two variables (X and Y) divided by the product of their standard deviations.
4. **Interpretation:**
    - A positive r indicates a positive linear relationship: as one variable increases, the other tends to increase.

- A negative *r* indicates a negative linear relationship: as one variable increases, the other tends to decrease.
- The magnitude of r reflects the strength of the correlation, with r close to 1 or -1 indicating a stronger linear relationship.

5. **Use in Statistics:**
    - Pearson's r is widely used in various fields, including economics, biology, psychology, and many others, to quantify and understand the association between two variables.
    - It is a crucial tool in regression analysis, helping assess the relationship between the independent and dependent variables.

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

1.
    - Scaling is the process of transforming numerical variables to a standard range or distribution, making them comparable and preventing one variable from dominating others in a dataset.
2. **Purpose of Scaling:**
    - **Comparable Units:** Variables might have different units or scales. Scaling ensures that all variables contribute equally to the analysis.
    - **Algorithm Convergence:** Many machine learning algorithms, such as gradient descent-based optimization, converge faster when variables are on similar scales.
    - **Interpretability:** Scaling aids in making models more interpretable by ensuring that coefficients represent the variable's impact in a comparable manner.
3. **Difference Between Normalized Scaling and Standardized Scaling:**
    - **Normalized Scaling (Min-Max Scaling):**
        - **Range:** Scales values between 0 and 1.
        - **Characteristics:**
            - Preserves the relative differences in the original data.
            - Sensitive to outliers, as extreme values can disproportionately impact the scaling.
    - **Standardized Scaling (Z-score Scaling):**
        - **Range:** Centers the data around zero with a standard deviation of 1.
        - **Formula:** $\text{standardized} = \frac{X - \mu}{\sigma}$ where μ is the mean of the variable, and σ is its standard deviation.
        - **Characteristics:**
            - Less sensitive to outliers, as the scale is determined by the standard deviation.
            - Does not preserve the original data distribution but is robust against extreme values.
    - **Which to Choose:**

- Normalized scaling is suitable when preserving the original distribution and relative differences is crucial.
- Standardized scaling is preferred when the emphasis is on reducing the impact of outliers and achieving a standardized distribution.

Scaling is essential for ensuring fair and effective comparisons among variables in a dataset. The choice between normalized and standardized scaling depends on the specific requirements of the analysis and the characteristics of the data.

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

The phenomenon of obtaining infinite values for the Variance Inflation Factor (VIF) typically occurs when there is perfect multicollinearity among the predictor variables in a regression model. Perfect multicollinearity arises when one or more independent variables can be precisely predicted from a linear combination of other variables in the model.
The VIF is a measure that quantifies how much the variance of an estimated regression coefficient is inflated due to multicollinearity. Specifically, for a given predictor variable, the VIF is calculated as:
VIFi=1−Ri21 **where:** Ri2.
=Unadjusted coefficient of determination forregressing the ith independent variable on therem aining ones
The presence of perfect multicollinearity is a severe issue for regression analysis, as it implies redundancy among predictor variables. In such cases, the regression coefficients become indeterminate, making it impossible to uniquely estimate the contribution of each variable to the dependent variable. Perfect multicollinearity can be caused by various factors, such as duplicate variables, linearly dependent variables, or inclusion of derived variables that are linear combinations of others.
To address issues related to infinite VIF values, it is essential to identify and remedy multicollinearity in the dataset. This can involve removing redundant variables, transforming variables, or using regularization techniques like Ridge regression to penalize large coefficients and mitigate multicollinearity. Overall, dealing with multicollinearity is crucial for maintaining the stability and reliability of regression models.

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

1. Definition of Q-Q plot
    - A Q-Q plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. The term "Q-Q" stands for quantile-quantile.
2. **Construction:**
    - A Q-Q plot is created by plotting the quantiles of the observed data against the quantiles of the expected distribution. If the points in the plot fall approximately along a straight line, it suggests that the data follows the assumed distribution.
3. **Use in Linear Regression:**

- **Assumption of Normality:**
    - In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed.
    - The Q-Q plot is often used to visually inspect whether the residuals conform to a normal distribution.
- **Interpretation:**
    - If the points in the Q-Q plot align closely with the diagonal line, it suggests that the residuals are approximately normally distributed.
    - Deviations from the line indicate departures from normality, helping identify potential issues with the assumption.

4. **Importance in Linear Regression:**
    - **Model Validity:**
        - Validating the normality assumption of residuals is crucial for the accuracy and reliability of linear regression models.
        - A Q-Q plot provides a visual tool for researchers to check whether the residuals meet the normality requirement.
    - **Identifying Skewness or Outliers:**
        - Q-Q plots are effective in identifying skewness or heavy tails in the distribution of residuals.
        - Outliers or systematic deviations from the line can be indicative of issues that may impact the robustness of the regression model.
    - **Diagnostic Tool:**
        - Q-Q plots are an integral part of diagnostic checks in linear regression analysis.
        - They complement statistical tests for normality and provide a more intuitive understanding of the distributional properties of residuals.
    - **Model Improvement:**
        - If the Q-Q plot suggests departures from normality, researchers may consider transformations or other adjustments to improve the model's adherence to assumptions.