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

a. The fall season has shown the highest number of bookings, suggesting a strong preference for biking during this season, possibly due to favorable weather conditions.
b. There has been a significant increase in bookings from 2018 to 2019 across all seasons, indicating overall growth and increased acceptance of the bike-sharing service.
c. May through October (late spring to early fall) see the highest booking rates, aligning with better weather and possibly more leisure time during summer vacations.
d. Booking trends start strong at the beginning of the year, peak mid-year, and then gradually decline towards the year-end, reflecting seasonal weather impacts and possibly varying user needs throughout the year.
e. Clear weather conditions have predictably led to more bookings, underscoring the weather dependency of bike-sharing services.
f. Thursday, Friday, Saturday, and Sunday are the busiest days in terms of bookings, suggesting that the service is popular both for weekend recreational activities and end-of-week commuting.
g. Bookings are lower on holidays, which might indicate that during holidays, potential users prefer to stay home or use other modes of transport for leisure outings.
h. The data shows negligible difference in booking numbers between working and non-working days, suggesting that the service caters effectively to both routine commuters and casual users alike.

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

Using drop_first=True during dummy variable creation in data preprocessing is important to avoid the issue known as **multicollinearity**, particularly in models that require matrix inversion like linear regression

Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. In the context of dummy variables, including a dummy for every category means one variable can be perfectly predicted from the others.

When you create dummy variables for a categorical feature with n categories, you actually only need n-1 dummy variables to capture all the information.

Dropping one category (usually the first alphabetically) means simplifying the model without losing any information. This simplification helps in interpreting the model results because the base category (the dropped one) serves as a reference against which other categories are compared.

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

‘temp’ variable has the highest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    a. **Normality of Residuals**
        i. The residuals of the model are normally distributed.
    b. **Independence of Residuals**
        i. No correlation in the residuals.
    c. **Multicollinearity**
        i. Predictors are not too highly correlated.
    d. **Linearity**
        i. The relationship between the predictors and the target variable is linear.
    e. R-squared, Adjusted R-squared: Indicative of the model fit.
    f. F-statistic and its p-value: Testing the overall significance of the model.

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

1. **Explain the linear regression algorithm in detail.**

    Linear regression is a way to predict a numeric outcome (like sales, prices, scores) based on one or more influencing factors (like advertising budget, house size, or hours studied).

Key Components:

Variables: You have one outcome variable you're trying to predict (like price of a house). This is often called the dependent variable. You also have one or more predictor variables (like the size of the house or the number of rooms) that you think will influence the outcome. These are called independent variables.

Relationship: The main idea in linear regression is that the dependent variable is a linear combination of the independent variables, plus some error. Basically, you can think of it as a line (or a plane in higher dimensions) that best fits your data points.

Fit the Line: The algorithm calculates the best line that minimizes the differences (errors) between the predicted values and actual values. This line is a prediction equation.

Prediction Equation: This equation tells you what value of Y you can expect for a given value of X. For example, in a simple housing price model, the equation might look like y = mx + c Here, the slope tells you how much the price increases per square foot of the house.

Use the Model: Once the model is built, you can use it to predict the outcome for any new data. For example, if you know the square footage of a new house, you can plug it into your equation to predict its price.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous example created by the statistician Francis Anscombe in 1973 to demonstrate a crucial point about the importance of data visualization and the potential pitfalls of relying solely on statistical properties for data analysis. The quartet consists of four different datasets, each with eleven x-y pairs. Remarkably, these datasets have nearly identical statistical properties but look very different when graphed.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistic that measures the strength and direction of a linear relationship between two variables. Here's how it works in simple terms:

- **Strength of Relationship**: It tells you how closely two variables move together. If one variable increases, does the other increase as well, decrease, or stay the same?
- **Direction of Relationship**: It indicates whether the relationship is positive (both variables increase or decrease together) or negative (one variable increases while the other decreases).
- **Range**: Pearson's R can range from -1 to 1.
    - **+1**: A perfect positive relationship (as one variable increases, the other also increases).
    - **-1**: A perfect negative relationship (as one variable increases, the other decreases).
    - **0**: No linear relationship (the variables do not move together in a specific linear pattern).

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

Scaling is a method of transforming data to a common scale, often used in preprocessing before applying machine learning algorithms. This technique is crucial for many models that are sensitive to the magnitude of features, such as support vector machines, k-nearest neighbors, and many algorithms that use distance calculations or gradient descent in their computations.

Why is Scaling Performed?

1. **Uniformity**: It brings all the features to a similar scale, thus preventing any single feature from dominating the model due to its range.
2. **Improved Performance**: Many algorithms converge faster and perform better when features are on a similar scale because gradient descent converges faster when all the gradients are on a similar scale.
3. **Distance-based Algorithms**: For methods that rely on the calculation of distances (like k-NN and k-means), scaling ensures that the distance metric gives equal importance to all features.
4. **Prevent Bias**: Without scaling, models might become biased towards features having a wider range.

Normalized Scaling vs. Standardized Scaling

1. Normalized Scaling (Min-Max Scaling)

- **Definition**: This type of scaling adjusts the data to a fixed range, usually 0 to 1, or -1 to 1.
- **When to Use**: Min-Max scaling is useful when you need a bounded range and the data distribution is not Gaussian. It is also useful when scaling does not distort differences in the ranges of the features.

2. Standardized Scaling (Z-score Normalization)

- **Definition**: This scaling technique adjusts the data such that the mean of the transformed features is 0 and the standard deviation is 1. It assumes that the data follows a Gaussian distribution.
- **When to Use**: Standardization is most useful when data follows a normal distribution. This scaling does not bound values to a specific range, which may be necessary for some algorithms that assume data is centred around zero and with a standard deviation of one.

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 the level of multicollinearity in regression analyses. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This correlation can cause difficulties in accurately estimating the relationship between predictors and the outcome variable.

Why VIF Can Be Infinite

The value of VIF becomes infinite (or exceedingly high) in situations of perfect multicollinearity, where one independent variable is an exact linear combination of other independent variables.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a certain theoretical distribution, usually the normal distribution. The plot displays the quantiles of the dataset against the quantiles of the theoretical distribution. If the data adhere to the assumed distribution, the points in the Q-Q plot will fall approximately along a straight line.

Importance of Q-Q Plots in Linear Regression

- **Model Validity**: Ensuring that the residuals are normally distributed helps validate the use of linear regression. Many statistical tests used for hypothesis testing in regression analysis (like t-tests for coefficients) assume normality of residuals.
- **Improving Model Accuracy and Interpretability**: By identifying and correcting issues like non-normality and outliers, you can improve both the accuracy and interpretability of your regression model.
- **Diagnostic Tool**: The Q-Q plot serves as an essential diagnostic tool in the model checking phase, helping to ensure that the conclusions drawn from the model are well-founded.