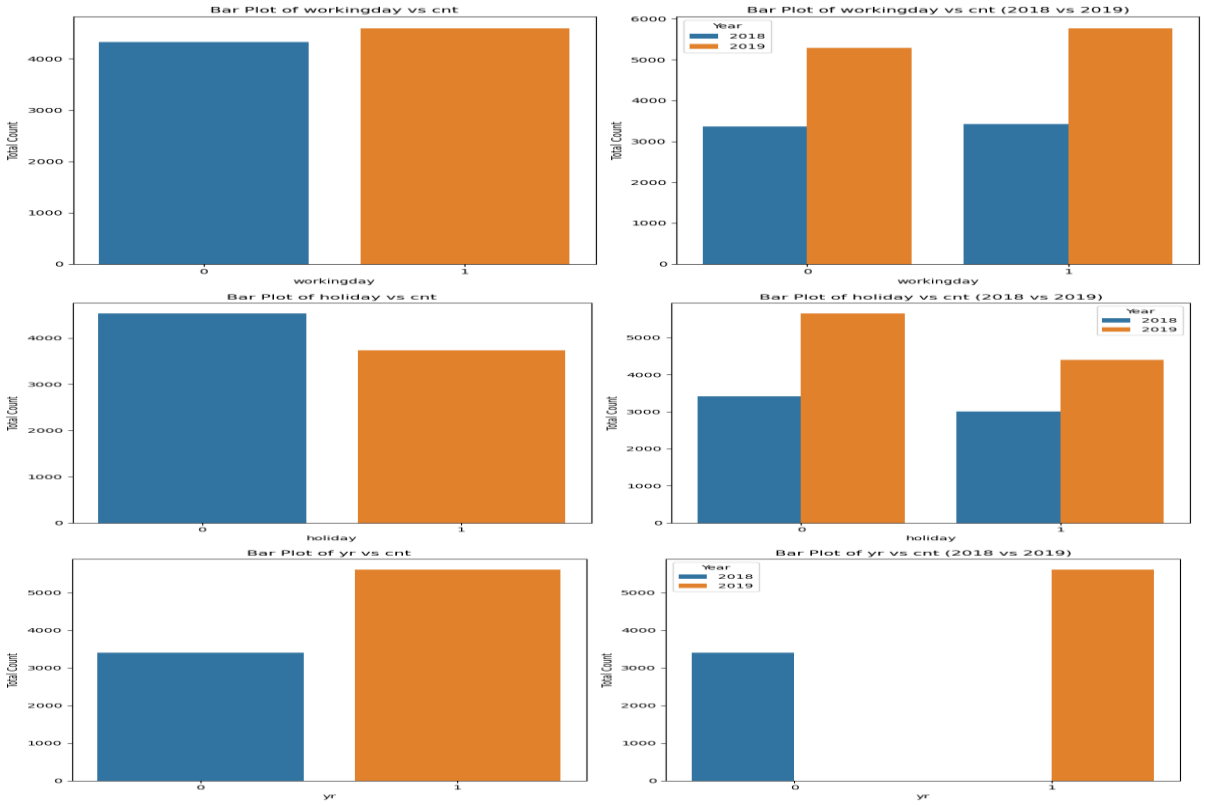
Assignment-based Subjective Questions

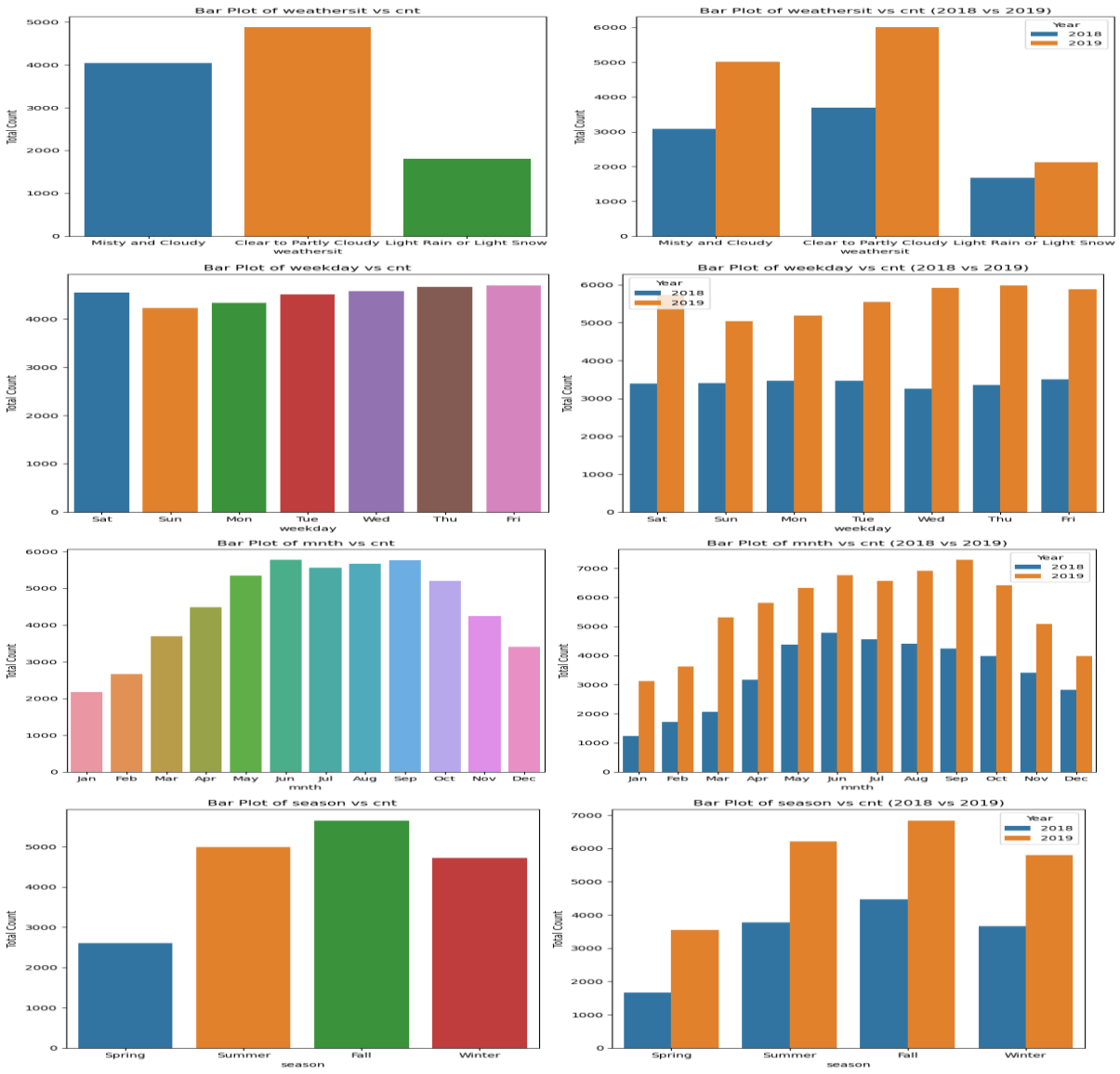
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   Answer:

   From the analysis of Categorical variables' impact on the dependent variable we have observed these trends:

   a. **Seasonal Trends**: The fall season experiences the highest number of bike rentals. Additionally, bike rentals in 2019 saw an overall increase across all seasons compared to 2018.
   b. **Monthly Trends**: Bike registrations are higher between May and October. In 2019, there was a rise in bike registrations for each month in comparison to 2018.
   c. **Weather Conditions**: Favorable weather encourages more bike rentals. In 2019, bike registrations increased across all weather conditions compared to 2018.
   d. **Holiday**: Fewer people rent bikes on holidays.
   e. **Daily Trends**: The number of bikes rented each day of the week shows minimal variation, indicating consistent demand throughout the week.
   f. **Working vs. Non-Working Days**: Bike rentals are higher on working days, likely due to increased use for commuting purposes.





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

Answer:

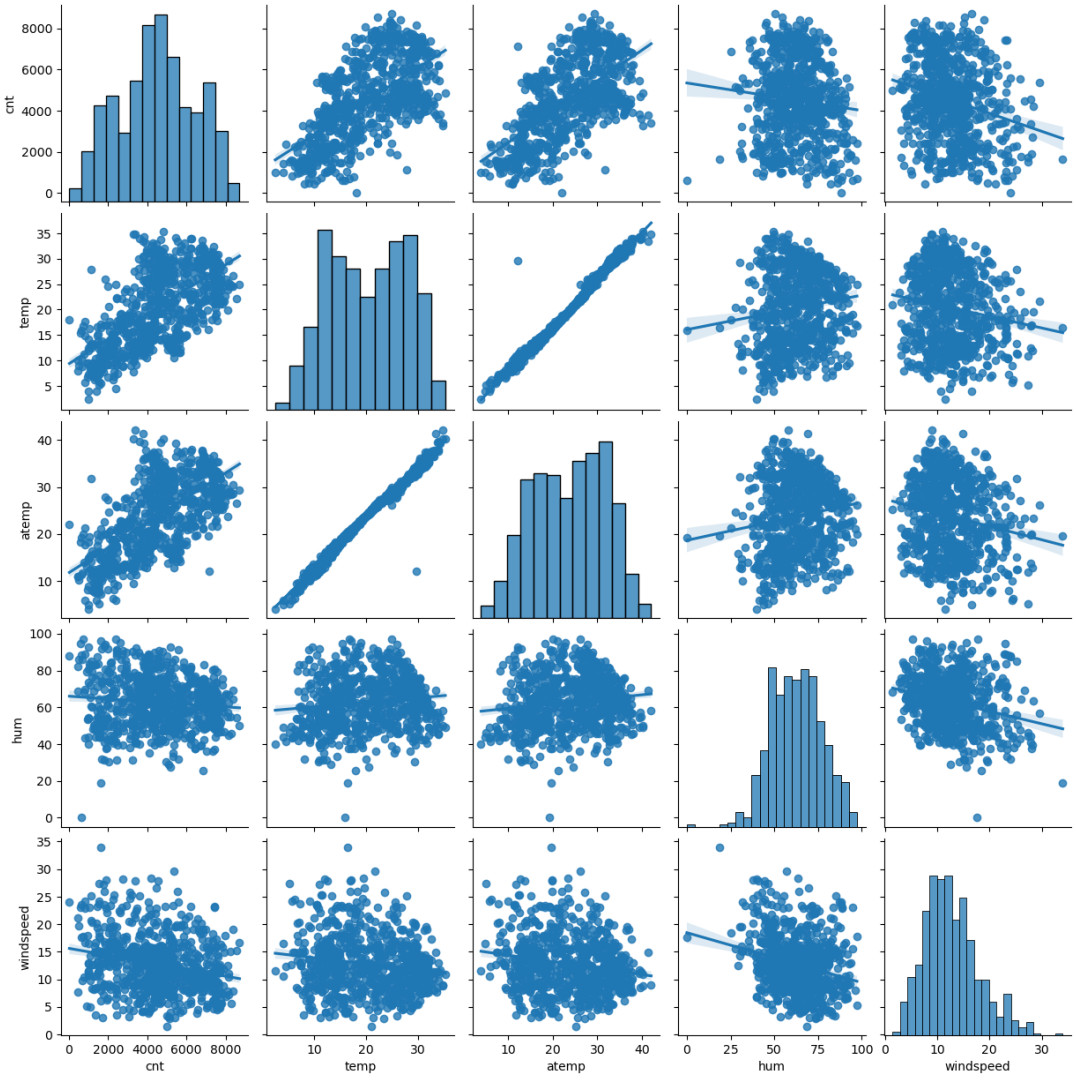
Using **drop_first=True** when creating dummy variables is important because it prevents **multicollinearity** (when variables overlap and mess up the model). By dropping one category, you set it as a **reference point**, so the model compares the others to it. This makes the model **simpler**, reduces unnecessary variables, and makes it easier to understand the results. In short we can say that it stops your model from being confused by redundant info.

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

Answer:

The variable temp (temperature) has a positive relationship with the target variable cnt, similar to atemp. The scatter plot shows that cnt and temp have a positive linear relationship, but it's a bit weaker compared to atemp. This makes sense because atemp better reflects how the temperature actually feels, which might influence the activities or demand shown by cnt.

In short, both temp and atemp are positively related to cnt, but atemp seems to have a slightly stronger correlation based on the pair-plot.

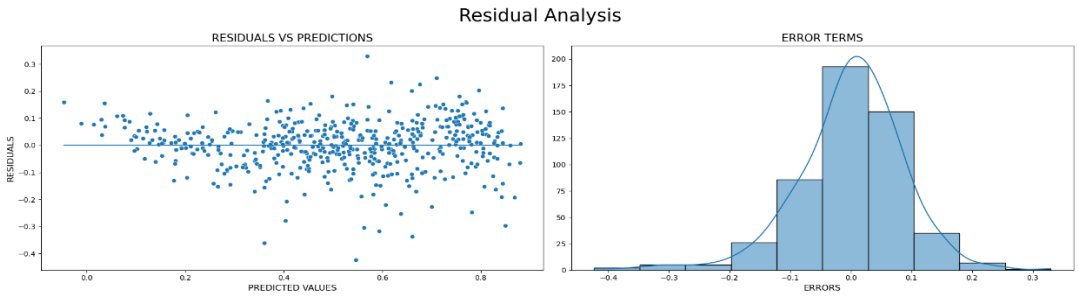


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

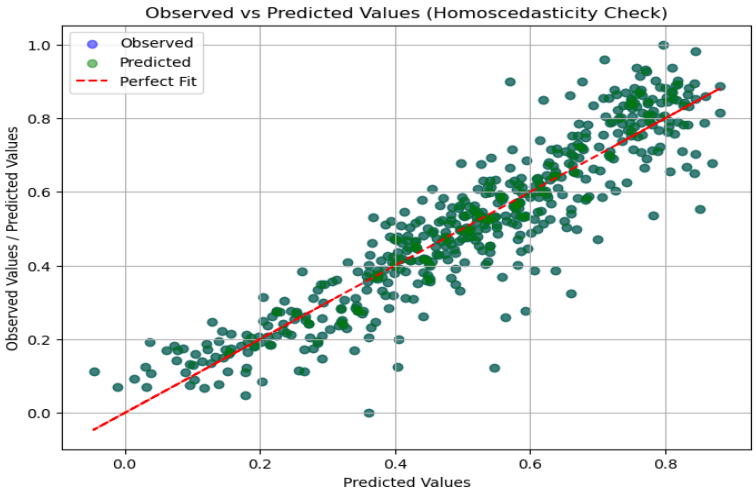
Answer:

In order to validate the assumptions of Linear Regression after building the model on the training set I did the Residual Analysis and created these plots of residual errors:

- **Normality of Residuals**
  - The average of the residuals is very close to 0.
  - The errors are normally distributed around a mean of 0.



- **Independence of Errors**
  - The scatter plot of the residuals from the regression model shows them randomly scattered around zero with no clear pattern, indicating that the error terms are independent of each other.
- **Homoscedasticity**
  - Since the data points are fairly evenly scattered around the red "perfect fit" line without any visible patterns (like a funnel shape), it suggests that the assumption of homoscedasticity is likely satisfied, meaning the model's residuals are evenly distributed.



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

Answer:

Here are the Top 3 features based on the final model contributing to explain the demand of the shared bikes:

- **temp (0.473)**: For each unit increase in temperature, the predicted bike rental count increases by 0.473 units, indicating that higher temperatures lead to more bike rentals.
- **weathersit_Light Rain or Light Snow (-0.292)**: When the weather changes to light rain or light snow, the predicted bike rental count decreases by 0.292 units, meaning that adverse weather conditions reduce bike rentals.
- **yr (0.234)**: Each passing year is associated with an increase of 0.234 units in the predicted bike rental count, suggesting that bike rentals have generally increased over time.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

A **linear regression model** is a statistical technique used to predict the value of one variable (called the **dependent variable** or **target variable**) based on one or more other variables (called **independent variables** or **features**). The model does this by fitting a straight line through the data points that best represents the relationship between the dependent and independent variables.

Steps Involved in Linear Regression:

a. **Identify the dependent and independent variables**:
   - The **dependent variable (Y)** is the outcome you want to predict. In the bike-sharing example, it could be the number of bike rentals (**cnt**).
   - The **independent variables (X)** are the factors that influence the dependent variable. These could include **temperature**, **weather condition**, and **year** in the bike-sharing context.

b. **Fit the linear regression model**: The goal is to find a line that best fits the data. The model calculates the line by minimizing the differences between the actual data points and the predicted values (called residuals).

c. **Find the coefficients**: The **coefficients** (β) are the key outputs of the model. Each coefficient tells us how much the dependent variable is expected to change when the independent variable increases by one unit, keeping other factors constant.

d. **Model equation**: The linear regression model can be represented mathematically as:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

- **Y** is the predicted value of the dependent variable (e.g., bike rentals).
- $\mathbf{X_1, X_2, ..., X_n}$ are the independent variables (e.g., temperature, weather condition, year).
- $\boldsymbol{\beta_0}$ is the **intercept** (the predicted value when all independent variables are zero).
- $\boldsymbol{\beta_1, \beta_2, ..., \beta_n}$ are the **coefficients** for each independent variable.
- **ε** is the error term or residual (the difference between the observed and predicted values).

e. **Interpret the coefficients**:
   - A **positive coefficient** (β) means that as the corresponding feature increases, the predicted target variable increases.
   - A **negative coefficient** means that as the corresponding feature increases, the predicted target variable decreases.

Assumptions of simple linear regression are:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Summary:

In linear regression, you start by defining your dependent variable (what you want to predict) and independent variables (the factors that influence the prediction). Then, the model fits a straight line to the data, calculates the coefficients for each independent variable, and provides an equation that allows you to predict future outcomes. In the Upgrad's bike-sharing assignment, factors like temperature, weather conditions, and time are used to predict bike rental numbers, with each coefficient explaining the effect of that factor on bike rentals.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is a set of four different datasets that share nearly identical statistical properties, such as mean, variance, correlation, and linear regression line, yet reveal different patterns when plotted. This showcases the importance of visualizing data rather than relying solely on summary statistics.

Important points about Anscombe's Quartet:

1. **Similar Summary Statistics**: Each dataset has:
    - The same mean for both the x and y variables.
    - The same variance for both variables.
    - The same correlation coefficient (around 0.816).
    - Nearly identical linear regression equations.
2. **Distinct Patterns**: Despite their similarities in statistics, each dataset shows different relationships:
    - **Dataset 1**: Shows a clear linear relationship, making the regression line a good fit.
    - **Dataset 2**: Exhibits a nonlinear relationship, indicating that the data does not fit a straight line well.
    - **Dataset 3**: Contains an outlier that skews the regression line, affecting the overall analysis.
    - **Dataset 4**: Has most x values the same, except for one outlier, resulting in a vertical distribution and a misleading regression line.

Learnings from Anscombe's Quartet:

- **Visualizing Data**: Always create plots to explore data patterns. Visuals can highlight trends, outliers, or anomalies that numbers alone might obscure.
- **Limitations of Summary Statistics**: While helpful, summary statistics can overlook critical details about the data.
- **Awareness of Outliers and Nonlinear Patterns**: Outliers can significantly influence regression results, and nonlinear trends may lead to incorrect conclusions if only linear models are used.

In summary, Anscombe's quartet emphasizes the need to analyze data from multiple perspectives. Combining visualizations with statistical analysis provides a clearer understanding of the relationships in the data, ensuring more accurate interpretations.

3. What is Pearson's R?

Answer:

Pearson's R, or the Pearson correlation coefficient, is a way to measure how strongly two continuous variables are related to each other in a straight-line fashion. Its value can range from -1 to +1:

- A value of **+1** means there is a perfect positive relationship; as one variable goes up, the other variable also goes up in a straight line.
- A value of **-1** means there is a perfect negative relationship; as one variable goes up, the other goes down in a straight line.
- A value of **0** means there's no linear relationship between the two variables.

How It's Calculated:

To calculate Pearson's R, you look at how each variable varies from its average. The formula is:

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

Where:

- Xi and Yi are the individual data points for the two variables,
- X(bar) and Y(bar) are the averages of those variables.

Key Points:

- Pearson's R only measures straight-line relationships and might miss more complex ones.
- It can be affected by outliers, or extreme values, which can change the result significantly.
- It is commonly used in various areas like finance, social studies, and science to understand how two things are connected.

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

Answers

Scaling is the process of adjusting the range of feature values in a dataset to ensure that no single feature disproportionately influences the model.

Why is Scaling Performed?

Scaling is done to ensure that features with different units or ranges contribute equally to the model, improving algorithm performance. For example, in a dataset with height (in cm) and weight (in kg), scaling helps treat both features equally.

Difference Between Normalized Scaling and Standardized Scaling:

- **Normalized Scaling** (Min-Max Scaling): Rescales values to a specific range, usually [0, 1]. The formula is: X′= (x –xmin)/ (xmax – xmin)

  For example, a temperature of 30°C in a range of 20°C to 40°C would be scaled to 0.5.

- **Standardized Scaling:** Centers values around a mean of 0 with a standard deviation of 1. The formula is : X′= (x – xmean)/ xstd

For example, a score of 80 with a mean of 70 and a standard deviation of 10 would be standardized to 1.

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

Answer:

The Variance Inflation Factor (VIF) is used to assess how much the variance of a regression coefficient increases due to multicollinearity among predictor variables. When you encounter a **VIF value of infinity**, it usually means:

1. **Perfect Multicollinearity**: This occurs when one predictor is a perfect linear combination of another. For instance, if you have variables X1X_1X1 and X2X_2X2 where X2=2X1X_2 = 2X_1X2=2X1, they are perfectly correlated, leading to an infinite VIF for one of the variables.
2. **Redundant Variables**: Including duplicate features in your dataset can also cause infinite VIF values, as they do not provide any additional information beyond what's already represented.
3. **Data Issues**: Sometimes, missing values or inconsistencies in the data can result in calculation errors, which can also lead to infinite VIF.

An infinite VIF indicates that the model struggles to isolate the impact of correlated variables, making coefficient estimates unreliable. In such cases, it's important to address the multicollinearity by removing or combining the correlated variables.

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

Answer:

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, usually the normal distribution. It plots the quantiles of your data against the quantiles of the expected distribution.

Use of a Q-Q Plot:

The main purpose of a Q-Q plot is to check if your data follows a specific distribution. If the points on the plot form a straight diagonal line, your data likely matches the theoretical distribution well.

Importance in Linear Regression:

1. **Normality Check**: Linear regression assumes that the residuals (the differences between observed and predicted values) are normally

distributed. A Q-Q plot helps confirm this assumption by showing how closely the residuals align with a normal distribution.

2. **Model Diagnostics**: If the points deviate from the diagonal line, it may indicate problems with the model, suggesting that adjustments might be needed.
3. **Outlier Detection**: The plot can also help spot outliers—points that stray far from the line can indicate unusual observations that require further investigation.

In short, Q-Q plots are essential for ensuring that the assumptions of linear regression are met, leading to more reliable and accurate models.