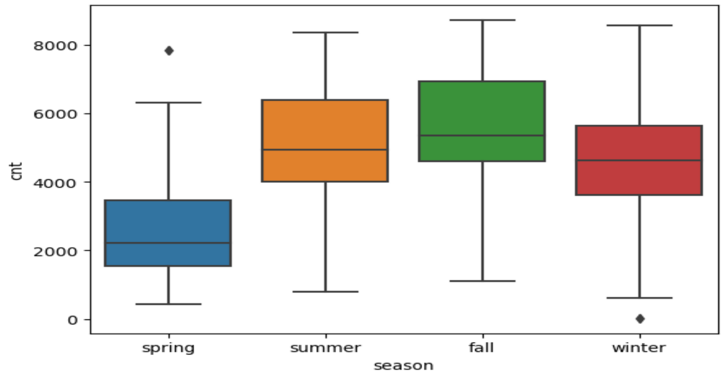
Linear Regression Assignment- Subjective Questions and Answers

Assignment-based Subjective Questions & Answers

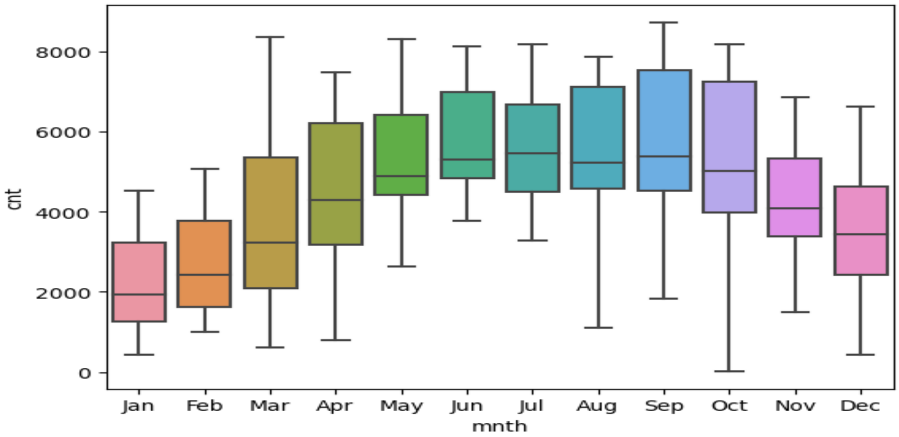
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1. From analysing the categorical variables in dataset and reviewing the box plots in the image, here are the following key insights into their effect on the dependent variable(cnt) (total bike rentals):

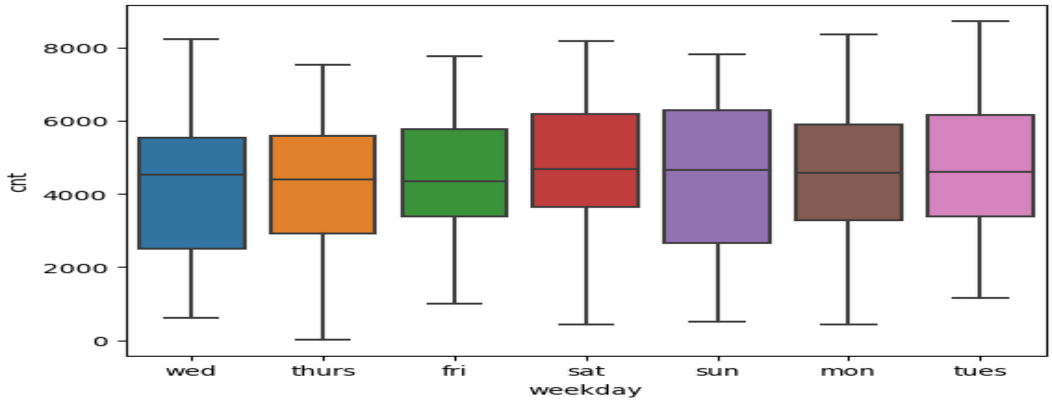
1. Season's Impact on Bike Demand:
   - Different seasons show varying median rental counts.
   - Fall appears to have the highest rentals, while spring has the lowest rentals.
   - This suggest that weather conditions influence bike-sharing demand significantly.



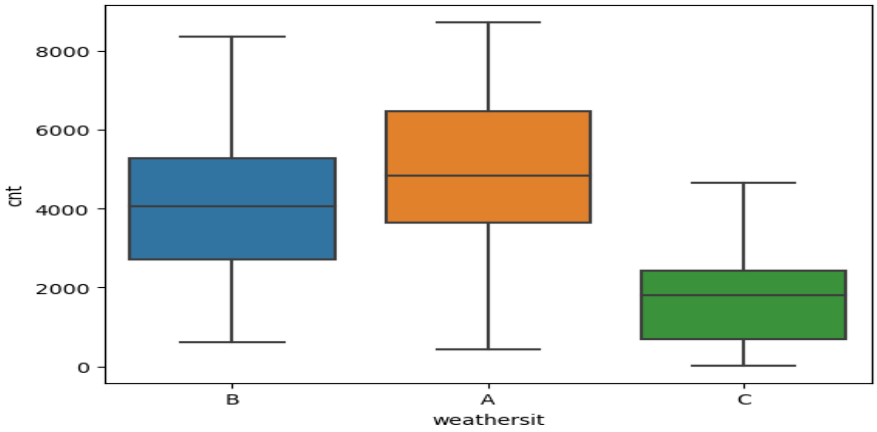
2. Month's Effect on Demand:
   - Months like May, June, July and September show higher rentals.
   - Winter months (Dec, Jan, Feb) show lower rental counts, possibly due to cold weather reducing outdoor activities.



3. Weekday vs. Weekend Effect on Demand:
    - Some weekdays (like Monday, Tuesday, Wednesday) have lower rentals compared to weekends (Saturday, Sunday).



4. Weather Effect on Demand:
    - Clear weather (weathersit_A) has the highest median rentals, while bad weather (weathersit_C) shows significantly lower demand.
    - This confirms that extreme weather (rain, fog, snow) reduces bike usage.

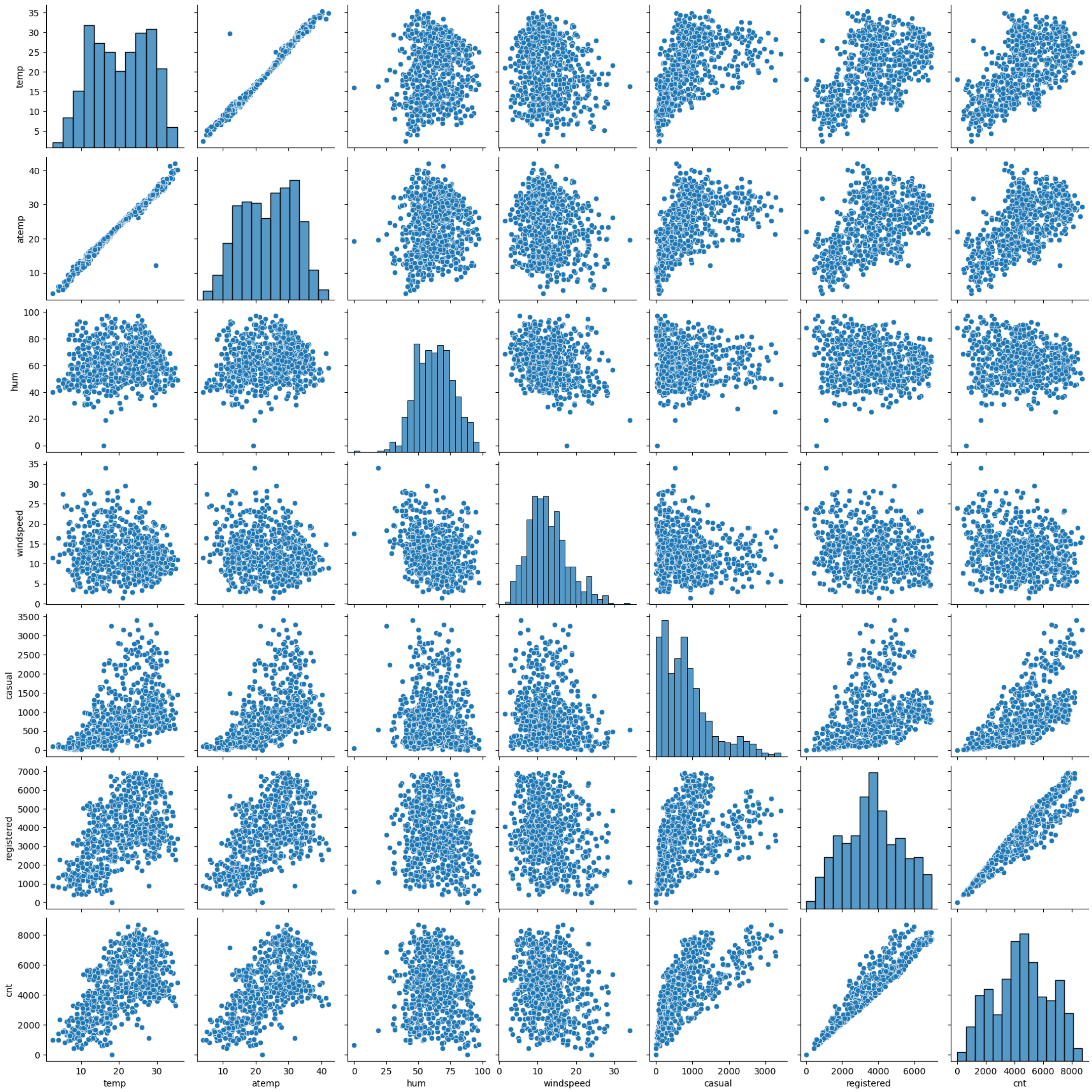


Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans2. Using drop_first=True during dummy variable creation in Panda's get_dummies() function is important because it helps prevent multicollinearity, which occurs when one dummy variable is completely predictable from the others.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans3. Looking at the pair-plot among the numerical variables, "registered" has the highest correlation with the target variable "cnt"(total bike rentals). The scatter plot between "registered" and "cnt" shows a strong positive linear relationship.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans4. Following are the key assumptions need to be validated:

a) Linearity: The dependent variable(cnt) should have linear relationship with independent variables.
☑ check: Residual vs Fitted plot

b) No Multicollinearity: Independent variables should not be highly correlated with each other. ☑ Check: Variance Inflation Factor (VIF)

c) Normality of Residuals: The residuals should follow a normal distribution.
☑ Check: Histogram

d) Homoscedasticity (Constant Variance of Residuals): The spread of residuals should be consistent across predicted values.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans5. Based on the OLS Regression Results, the top three features that significantly contribute to explaining bike rental demand (cnt) are:

a) Temperature (temp)
- Coefficient: 0.6009 → Highest positive impact
- t-value: 27.083 → Very strong significance
- p-value: 0.000 → Highly significant
- Interpretation: Warmer temperatures lead to higher bike rentals, as people prefer outdoor activities in favourable weather.

b) Year(yr)
- Coefficient: 0.2328 → Strong positive impact
- t-value: 24.467 → Highly significant
- p-value: 0.000 → Strong significance
- Interpretation: Bike-sharing demand increased significantly in 2019 compared to 2018, possibly due to market growth or higher adoption rates.

c) Weather Situation (weathersit_C)
- Coefficient: -0.2504 → Negative impact
- t-value: -8.841 → Very significant
- p-value: 0.000 → Strong effect
- Interpretation: Bad weather conditions (rain, snow, fog) significantly decrease bike rentals.

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.777
Model:                            OLS   Adj. R-squared:                  0.775
Method:                 Least Squares   F-statistic:                     292.4
Date:                Fri, 18 Apr 2025   Prob (F-statistic):         1.93e-160
Time:                        13:44:43   Log-Likelihood:                 421.78
No. Observations:                 510   AIC:                            -829.6
Df Residuals:                     503   BIC:                            -799.9
Df Model:                           6
Covariance Type:            nonrobust
```

```
==============================================================================
                 coef    std err        t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1008      0.018      5.481      0.000       0.065       0.137
yr             0.2328      0.010     24.467      0.000       0.214       0.251
holiday       -0.0877      0.030     -2.920      0.004      -0.147      -0.029
temp           0.6009      0.022     27.083      0.000       0.557       0.644
windspeed     -0.1377      0.029     -4.790      0.000      -0.194      -0.081
season_winter  0.1025      0.011      8.941      0.000       0.080       0.125
weathersit_C  -0.2504      0.028     -8.841      0.000      -0.306      -0.195
==============================================================================
Omnibus:                       38.009   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.470
Skew:                          -0.513   Prob(JB):                     2.72e-14
Kurtosis:                       4.374   Cond. No.                         9.58
==============================================================================
```

General Subjective Questions & Answers

Q1. Explain the linear regression algorithm in detail.

Ans1. Linear regression is a fundamental machine learning algorithm used for predicting a dependent variable(Y) based on one or more independent variables (X). It assumes a linear relationship between the variables.

1. Mathematical Equation of Linear Regression:

   Linear Regression models the relationship using the equation: [Y= \beta_0 + \beta_1X_1 + \beta_2X_2 + ……. + \beta_nX_n + \epsilon] where:

   - (Y) → Dependent variable (target/prediction)
   - (X_i) → Independent variables (features)
   - (\beta_0) → Intercept (constant term)
   - (\beta_i)→ Coefficients (weights for each feature)
   - (\epsilon)→ Error term (difference between actual and predicted values)

2. Types of Linear Regression:

   A. Simple Linear Regression (SLR):

   - One independent variable (X) and one dependent variable (Y).

- Equation:
  [Y= \beta_0 + \beta_1X + \epsilon]
- Example: Predicting house price based on square footage.

B. Multiple Linear Regression (MLR)

- Multiple independent variables (X1, X2, ……, Xn) affect Y.
- Equation:

  [Y = \beta_0 + \beta_1X_1 + \beta_2X_2 +….+ \beta_nX_n + \epsilon]

- Example: Predicting house price using square footage, number of bedrooms, and location.

3. How the Algorithm Works

Step 1: Compute Cost Function (Mean Squared Error – MSE)

The model aims to minimize the error using MSE: [MSE= \frac{1}{n} \sum_{i=1}^{n} (Y_i-\hat{Y_i}^2] Where:

- (Y_i) = actual value
- (\hat{Y_i} = predicted value

Step 2: Optimize Weights Using Gradient Descent

- Gradient Descent updates weights (βi) iteratively: [\beta_i = \beta_i - \alpha \frac{\partial MSE}{\partial \beta_i}]
- Here, α is the learning rate controlling step size.

Step 3: Model Evaluation

After training, evaluate using:

- R-squared ((R^2)) → Measures how well the model explains variance in Y.
- Adjusted (R^2) → Accounts for feature redundancy.
- Root Mean Squared Error (RSME) → Measures error magnitude.

4. Assumptions of Linear Regression:

To ensure accurate predictions, the model assumes:

☑ Linearity: Y must have a linear relation with X.

☑ No Multicollinearity: Independent variables must not be highly correlated.

☑ Normal Distribution of Residuals: Errors should follow a bell curve.

☑ Homoscedasticity: Variance of residuals must be constant across predictions.

☑ Independence of Residuals: No autocorrelation in errors.

5. Limitations of Linear Regression
    - Sensitive of Outliers → Extreme data points distort predictions.
    - Assumes Constant Variance → Violations lead to unreliable models
    - Works Only for Linear Data → Fails for complex relationships.

Q2. Explain the Anscombe's quartet in detail.

Ans2. Anscombe's Quartet: Understanding the Importance of Data Visualization

Anscombe's Quartet is a set of four different datasets that have nearly identical summary statistics (mean, variance, correlation, regression line) but very different distributions when plotted visually. It was introduced by Francis Anscombe in 1973 to highlight the importance of data visualization in statistical analysis.

1. The Four Datasets
   Each dataset consists of 11 (x,y) pairs, and they all share these key statistical properties:

- Mean of X ≈ 9.0
- Mean of Y ≈ 7.5
- Variance of X ≈ 10.0
- Variance of Y ≈ 4.12
- Correlation (X, Y) ≈ 0.816
- Same regression line:
  [y = 3 + 0.5x]

A Closer look at Each Dataset

a. First Dataset: Linear relationship, well-fitted by regression.
b. Second Dataset: Quadratic (non-linear) relationship --- regression line is misleading.
c. Third Dataset: Vertical clustering with an outlier, distorting correlation.
d. Fourth Dataset: Nearly constant X values – outlier drives regression result.

2. Key Lesson: Why Visualization Matters
   If you relied only on summary statistics, all four datasets would appear identical. But when plotted:
    - Dataset 2 & 3 show non-linearity or clustering, invalidating a simple regression model.
    - Dataset 4 has a single outlier drastically affecting results.

- Dataset 1 behaves as expected for linear regression.

This emphasizes that numerical summaries alone can hide important patterns, making visualization essential in statistical analysis.

3. Practical Applications
   - Outlier Detection: Identifies anomalies affecting statistical models.
   - Model Validation: Ensures a linear model is appropriate.
   - Data Storytelling: Helps interpret trends before making decisions.
   - Scientific Research & AI: Used to validate datasets in machine learning.

4. Summary
   Anscombe's Quartet proves that data visualization is just as important as numerical analysis. Before trusting summary statistics, always plot the data to check for hidden Patterns.

Q3. What is Pearson's R?

Ans3. Pearson's R (Pearson Correlation Coefficient)

Pearson's R is a statistical measure that quantifies the linear relationship between two variables. It tells us how strongly and in what direction two variables are related.

1. Formula for Pearson's R

Pearson's correlation coefficient (r) is calculated as:

[r= \frac{\sum(X_i - \bar{X})(Y_i -\bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\times\sqrt{\sum(Y_i-\bar{Y})^2}}]

Where:

- (X) and (Y) = The two variables
- (\bar{X}) and (\bar{Y}) = Mean of each variable
- Numerator= Covariance between (X) and (Y)
- Denominator = Product of standard deviations of (X) and (Y)

2. Interpretation of Pearson's R

Pearson's R ranges from -1 to +1, indicating:

- +1: Perfect positive correlation (both variables increase together)
- 0: No correlation (random, no linear relationship)
- -1: Perfect negative correlation (one variable increases, the other decreases)

3. When to use Pearson's R?

☑Checking linear relationships

☑Feature selection in machine learning

☑Understanding dependencies between variables

4. Limitation:

Pearson's R only captures linear relationships. If the data is curved or nonlinear, other correlation methods like Spearman's correlation may be better.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4. Scaling is the process of transforming numerical features in a dataset so they have a consistent range or distribution. It ensures that different variables are comparable, especially when they have vastly different units or magnitudes.

Scaling is performed for several reasons:

- Improves Model Performance → Many machine learning algorithms(e.g., Linear regression, logistic regression, KNN, SVM) are sensitive to varying feature scales. Scaling ensures that no feature dominates the training process.
- Speeds Up Convergence→ Gradient-based optimization methods (e.g., gradient descent) work more efficiently when features are scaled, leading to faster convergence.
- Enhances Interpretability→ Standardized or normalized values make it easier to compare features across different scales.
- Required for Certain Models→ Models using Euclidean distance(KNN, K-means clustering, PCA) require feature scaling for accurate distance calculations.

Difference Between Normalized Scaling & Standardized Scaling

Scaling can be done using two popular methods:

a. Normalization (Min-Max Scaling)
    - Formula:

    [X' = \frac{X-X_{\text{min}}}{X_{\text{max}} – X_{\text{min}}}]

    - Effect: Rescales values to range [0,1] (or sometimes [-1,1])
    - Use Case: Best for bounded values or datasets without extreme outliers.
b. Standardization (Z-Score Scaling)
    - Formula: [X' = \frac{X - \mu}{\sigma}]

- Effect: Centers data around mean=0 and scales to unit variance.
- Use Case: Best for datasets with outliers or assumptions of normality.

Key Takeaway:

Use Normalization when features have bounded values and no extreme outliers.

Use Standardization when features follow a normal distribution or have outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans5. An infinite Variance Factor (VIF) typically occurs due to perfect multicollinearity, meaning one variable is a perfect linear combination of others. Here's why it happens:

1. Perfect Multicollinearity (Exact Linear Dependency)
   - If one feature is an exact multiple or sum of other features, its VIF becomes infinite.
   - Example: If you include both temp and atemp, which are highly correlated, one can be nearly predicted by the other.

2. Dummy Variable Trap
   - Occurs when categorical variables are fully represented using dummy encoding without dropping one category (drop_first=True)

3. Identical Columns or Highly Correlated Features
   - If two variables store the same information, one can be perfectly predicted from the other.
   - Example: Including both "casual" and "registered" when their sum equals "cnt", which leads to a singular matrix and makes VIF explode.

4. Small Sample Size Relative to Features
   - If there are more features than observations, some variables might be highly dependent, causing infinite VIF.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans6. A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, typically the normal distribution.

It helps assess whether residuals from a regression model follow a normal distribution, which is a key assumption in linear regression.

Use of a Q-Q plot work:

1. Sort residuals (or any dataset values) in ascending order.
2. Calculate theoretical quantiles from a normal distribution.
3. Plot residual quantiles against theoretical quantiles.
4. Interpret results:
    - Straight diagonal line: Residuals follow a normal distribution.
    - Curved pattern: indicates skewness (data is not normally distributed).
    - S-shaped plot: Indicates heavy tails (outliers affecting distribution).

Importance of Q-Q Plot in Linear Regression

Linear regression assumes residuals are normally distributed.

Violations can affect statistical tests, confidence intervals, and model interpretability.

Use in regression validation:

- Helps detect non-normality in residuals.
- Identifies skewness or heavy tails, leading to improved model adjustments.
- Supports decision-making for transformations (e.g.,log, square root) or robust regression.