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)**

   From the univariate analysis:

   1) Bike demand is lower at the beginning of the year, increases to a peak in mid-year, and then decreases toward the end of the year
   2) Bike demand is lower in winter compared to other seasons.
   3) There is no significant variation in demand between weekdays and weekends.

   From Bivariate Analysis:

   4) It appears that temperature and "atemp" have a similar relationship with bike count, so we will drop "atemp" from the analysis.

   From Multivariate Analysis:

   5) The cnt shows a positive correlation with temperature and a negative correlation with wind speed. Other variables can be disregarded due to their very low correlation values.

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

   Using drop_first=True during dummy variable creation (one-hot encoding) helps prevent the "dummy variable trap," which occurs when dummy variables are highly correlated (multicollinearity). By dropping one dummy variable from each categorical feature, we eliminate this perfect multicollinearity, ensuring that the model's coefficients remain interpretable and stable. This approach reduces the number of features without sacrificing information, as the dropped category can be inferred from the others.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   'temp' variable has highest correlation with the target variable

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

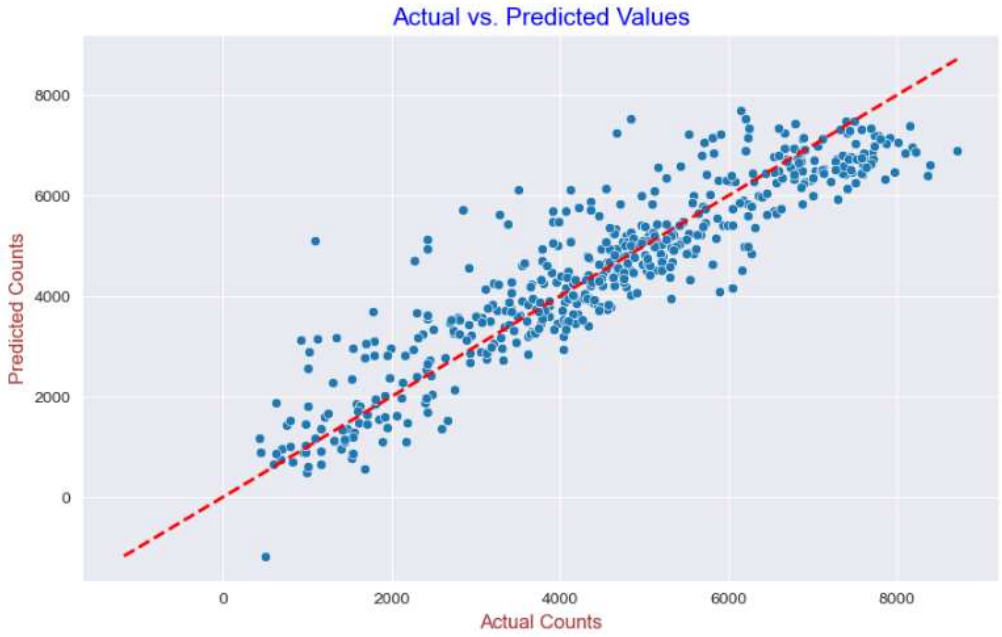
   1) R2_score_test set is 0.826, the model can explain 82.6% of the variance in bike rental demand based on the independent variables used.

   ```
   R-squared on Test Set: 0.826120546994689
   ```

   2) Linearity

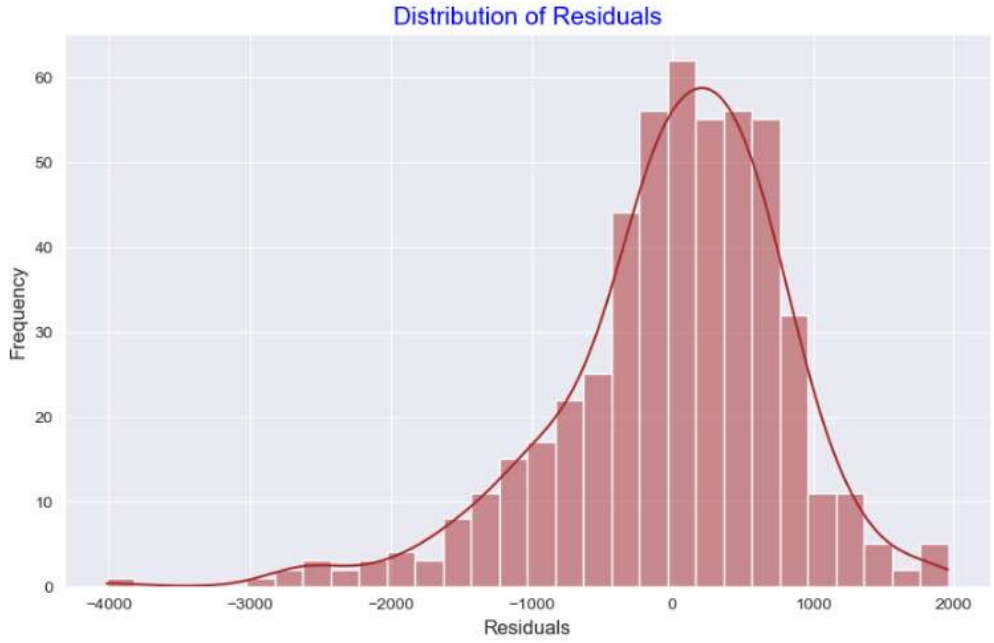
      Assumption: The relationship between the independent variables and the dependent variable is linear.

      Validation: I visually examined scatter plots of predicted versus actual values, looking for linear patterns. A linear pattern suggests that the assumption of linearity holds, while non-linear patterns may indicate that a linear model is not suitable.

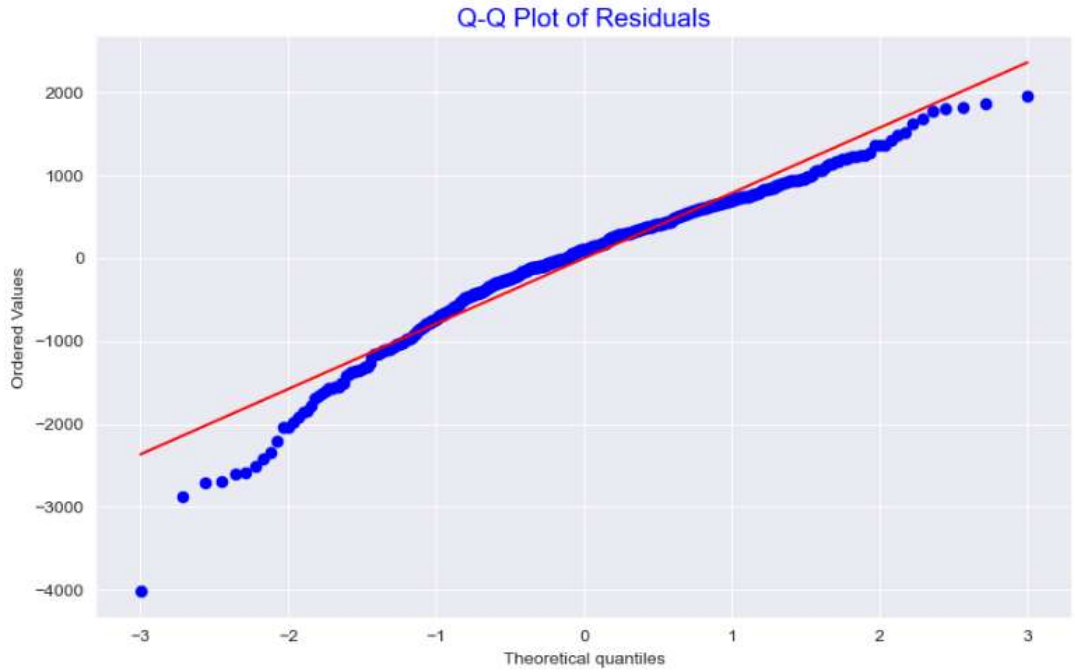


3) Normality of Residuals

Assumption: The residuals (differences between observed and predicted values) of the model are normally distributed



Validation: I plotted a histogram of the residuals, and a bell-shaped curve centered around zero indicates that they are approximately normally distributed. Additionally, I used a Q-Q plot to compare the residuals' distribution to a normal distribution. Points that closely align with the diagonal line in the Q-Q plot suggest that the residuals are normally distributed.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

1) temp (Temperature):

Coefficient(coef): 3892.2865
P-value: Significantly low ($p < 0.001$)
Interpretation: Temperature is the most significant predictor of bike demand. Its positive coefficient indicates that higher temperatures are associated with increased bike rentals, highlighting it as the top contributing factor.

2) Yr_1(Year: 2019):

Coefficient(coef): 1917.1324
P-value: Significantly low ($p < 0.001$)
Interpretation: The year 2019 (encoded as yr_1) significantly influences bike rental demand, indicating a marked increase from 2018 to 2019. This suggests a growing popularity or expansion of the bike-sharing service during that period.

3) weathersit_Light_Snow_Rain (Weather Situation: Light Snow or Rain):

Coefficient(coef): -1434.9070
P-value: Significantly low ($p < 0.001$)
Interpretation: This feature has a significant negative impact on bike demand, indicating that light snow or rain greatly reduces the number of bike rentals. Thus, it is a critical factor affecting demand, though in a negative manner.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The objective is to find the best-fitting line (or hyperplane) that minimizes the difference between predicted and actual values.

Mathematical Formulation

The general equation of a linear regression model can be expressed as:

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+ \ldots . + \beta_nx_n + \epsilon$$

where:

- Y Dependent variable (target).
- X is the dependent variable.
- β_0 Intercept (constant term).
- $\beta_1. B_{2,- - }\beta_n$ are the coefficients for the independent variables $x_1, x_2 - - - x_n$
- ϵ is the error term (the difference between the predicted and actual values).

Types of Linear Regression

- Simple Linear Regression: Involves one independent variable.
- Multiple Linear Regression: Involves two or more independent variables.

Evaluation Metrics

After training the model, performance can be evaluated using metrics like:

- **R-squared**: Indicates how well the independent variables explain the variability of the dependent variable.
- **Mean Absolute Error (MAE)**: The average absolute difference between predicted and actual values.
- **Mean Squared Error (MSE)**: The average of the squared differences.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's Quartet is a set of four datasets that are designed to illustrate the importance of graphing data before analyzing it. Despite having nearly identical statistical properties (such as means, variances, and regression lines), the datasets exhibit very different distributions and relationships when visualized. Developed by statistician Francis Anscombe in 1973, the quartet highlights how relying solely on statistical measures can be misleading. It emphasizes the necessity of visualizing data to gain a true understanding of underlying patterns.

Key Properties:

The four datasets in Anscombe's quartet have the same or very similar values for the following properties:

- Mean of both x and y variables.
- Variance of x and y variables.
- Correlation between x and y variables.
- Linear regression line (y = mx + c) that best fits the data, including the slope (m) and y-intercept (c).
- Coefficient of determination (R^2), which measures the proportion of the variance in the dependent variable that is predictable from the independent variable. Despite these similarities in statistical summaries, the datasets have very different distributions and appear distinct when graphed. Each set illustrates a different case or problem in regression analysis.

Te quartet consists of four datasets, each containing 11 pairs of (x,y) values:

- **Dataset I:** Shows a strong linear relationship.
- **Dataset II:** Appears linear but with a few outliers that can affect the regression line.
- **Dataset III:** Contains a perfect quadratic relationship.
- **Dataset IV:** Shows a linear relationship with one outlier that significantly skews the regression line.

Key Insights

- **Importance of Visualization:** Anscombe's Quartet demonstrates that statistical summaries alone can be misleading. Visual inspection is crucial to understand data behavior.
- **Sensitivity to Outliers:** The influence of outliers can vary greatly among datasets, affecting the interpretation of statistical results.
- **Model Selection:** Different models (e.g., linear vs. polynomial regression) may be more appropriate based on the visual characteristics of the data.

Conclusion

Anscombe's Quartet serves as a compelling reminder of the need for thorough exploratory data analysis (EDA). By illustrating how different datasets can yield the same statistical results while presenting vastly different patterns, it underscores the critical role of visualization in statistical analysis and model interpretation.

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

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that evaluates the strength and direction of the linear relationship between two continuous variables. Here are the key aspects of Pearson's R:

Definition

Pearson's R quantifies how closely two variables move in relation to each other. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation (as one variable increases, the other also increases).
- -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases).
- 0 indicates no linear correlation (the variables do not show any linear relationship).

The formula for calculating Pearson's R is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- n = number of data points
- x and y = individual sample points from the two datasets

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling refers to the process of transforming the features of a dataset so that they have a similar range or distribution. This is essential in machine learning and statistical analysis, where the scale of data can significantly impact model performance and convergence.

Why is Scaling Performed?

Improves Model Performance: Many algorithms, particularly those that rely on distance calculations (e.g., k-NN, SVM), perform better when features are on a similar scale.

Speeds Up Convergence: Gradient-based optimization methods (like gradient descent) converge faster when features are scaled, as it prevents certain features from dominating the loss function.

Enhances Interpretability: Scaling can make it easier to interpret model coefficients, especially in linear models.

Difference Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling:
   - **Definition**: Rescales the features to a fixed range, usually [0, 1]. It is calculated using the formula:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

   - **Usage**: Best for when the data has a known minimum and maximum and when you want to maintain the relative proportions.
   - **Effect**: Useful when you want all features to contribute equally to distance metrics.
2. Standardized Scaling:
   - **Definition**: Centers the data around the mean with a unit standard deviation. It is calculated using the formula:

$$X' = \frac{X - \mu}{\sigma}$$

     where μ is the mean and σ is the standard deviation.
   - **Usage**: Best when the data follows a Gaussian (normal) distribution, or when you want to handle outliers by giving them less influence.
   - **Effect**: It results in a distribution with a mean of 0 and a standard deviation of 1.

Summary

In summary, scaling is crucial for improving model performance and interpretability. Normalized scaling adjusts data to a specific range, while standardized scaling centers it around the mean with a unit variance. Each method serves different purposes based on the nature of the data and the requirements of the analysis.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. It quantifies the extent of multicollinearity in regression models.

Why Does VIF Become Infinite?

Perfect Multicollinearity:
VIF can become infinite when one predictor variable is a perfect linear combination of one or more other predictor variables. This means that the predictor's information is entirely redundant, leading to perfect multicollinearity.

- For example, if you have variables X_1 and X_2 where $X_2 = 2 \times X_1$, the correlation between them is 1, resulting in infinite VIF for either variable.

2. **Mathematical Definition**:
   - VIF for a variable X_i is calculated as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the R-squared value obtained by regressing X_i against all other predictors. If $R_i^2 = 1$ (indicating perfect correlation), the denominator becomes zero, resulting in infinite VIF.

3. Implications:
Infinite VIF indicates a serious problem in the regression model, suggesting that the model cannot uniquely estimate the coefficients for the involved predictors. This can lead to unreliable and unstable estimates.

Summary
In summary, VIF becomes infinite due to perfect multicollinearity, where one predictor can be perfectly explained by others, leading to a breakdown in the regression analysis. It's essential to identify and address this issue to ensure a valid regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A **Q-Q (Quantile-Quantile)** plot is a graphical tool used to assess if a dataset follows a specified theoretical distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

How a Q-Q Plot Works

Quantiles: The data is divided into equally sized intervals, and the quantiles are calculated for both the sample data and the theoretical distribution.

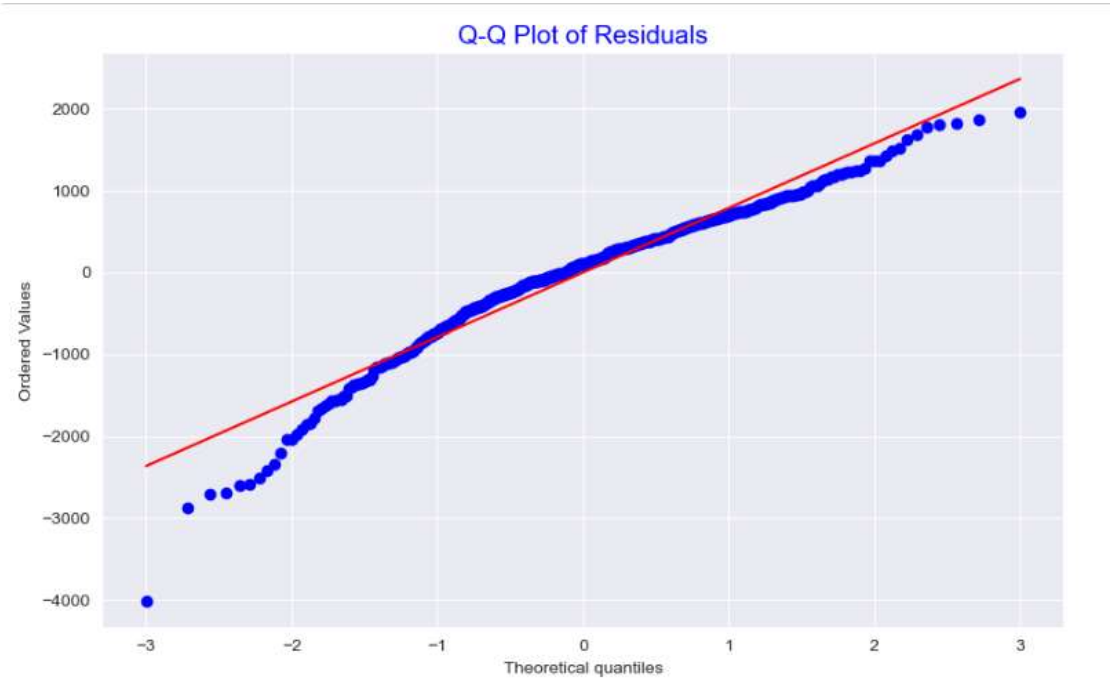
Plotting: Each point on the plot corresponds to a quantile of the sample data (y-axis) versus a quantile of the theoretical distribution (x-axis).

Interpretation:

- If the points lie approximately along a straight line (usually the 45-degree line), the data is likely normally distributed.
- Deviations from this line indicate departures from normality.

Use and Importance of Q-Q Plots in Linear Regression

I have plotted Q-Q plot in the current assignment:



Assessing Normality of Residuals:
In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps visually evaluate this assumption by comparing the distribution of the residuals to a normal distribution.

Identifying Non-Normality:
If the Q-Q plot shows significant deviations from the straight line (e.g., heavy tails, skewness), it indicates that the residuals may not be normally distributed. This non-normality can affect the validity of statistical tests and confidence intervals derived from the regression model.

Model Diagnostics:
Q-Q plots serve as a diagnostic tool to identify issues in the regression model. If the residuals are not normally distributed, it may suggest the need for transformations of the response variable or the inclusion of additional predictors.

Improving Model Fit:
By analyzing the Q-Q plot, practitioners can make informed decisions about potential adjustments to the model, such as using polynomial terms or other non-linear transformations to better capture the underlying data structure.

Conclusion
In summary, a Q-Q plot is a vital tool in linear regression for assessing the normality of residuals, which is crucial for the validity of the regression analysis. By identifying departures from normality, it aids in diagnosing model fit and making necessary adjustments to improve the robustness of the analysis.