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?**
    - A clear sky weather contributes significantly to the rise in demand as the other weathers shows a dip in it.
    - There is a dip in the demand if it's a holiday.

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

    The reason for dropping one of the dummy variables is to avoid redundancy. Since the sum of all dummy variables for a given observation will always be 1. When we drop one dummy variable, the remaining k−1 dummy variables serve as relative indicators. The dropped dummy variable acts as the reference category, against which the effects of the other categories are measured. This makes it easier to interpret the model coefficients, as each coefficient will show the effect of that category relative to the reference category.

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

    'cnt' is highly (positively) correlated with 'casual' and 'registered' and further it is high with 'atemp' and 'temp'(Which are same).

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

    We can perform residual analysis and plot the graph between count and the difference of Y variable and the predicted Y variable, if it shows the plot to be normally distributed with the mean value of the residual error as 0 , then the model on the training set is said to be following the linear regression.

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

    1. The demand of the bikes gets reduced when the weather is cloudy or if there is a snowfall, high humidity and windspeed also contributes to the same.
    2. We have definitely seen the rise in demand of bikes when the temperature gets higher or if its summer season.
    3. Comparatively, to other months, September shows rise is demand.

General Subjective Questions

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

   Linear regression is a widely used algorithm in statistics and machine learning for predicting a continuous target variable based on one or more predictor variables. The primary goal of linear regression is to model the relationship between the predictors (features) and the target variable (outcome) by fitting a linear equation to the observed data.

   Key Concepts of Linear Regression

   - Linear regression assumes a linear relationship between the dependent variable y and the independent variables X. This relationship is represented by a linear equation.
   - For simple linear regression (one predictor), the model is given by: y=β0+β1x+e
     - y is the dependent variable (target).
     - X is the independent variable (feature).
     - β0 is the intercept.
     - β1 is the slope of the line (coefficient of x).
     - e is the error term (residual), capturing the difference between the predicted and actual values.
   - For multiple linear regression (more than one predictor), the model extends to: y=β0+β1x1+β2x2+⋯+βnxn+e
   - The objective of linear regression is to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between the observed values and the predicted values.
   - Linear regression typically uses the Mean Squared Error (MSE) as the loss function, which measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model.

   $$n1\sum i = 1n(yi - y\hat{}i)2$$

2. Fitting the Model:
   - The model parameters (coefficients) are estimated by minimizing the MSE using optimization techniques. The most common method is the **Ordinary Least Squares (OLS)** method, which analytically solves for the coefficients that minimize the sum of squared residuals.
3. Assumptions:
   - **Linearity**: The relationship between predictors and the target is linear.
   - **Independence**: Observations are independent of each other.
   - **Homoscedasticity**: The residuals (errors) have constant variance.
   - **Normality**: The residuals are normally distributed.

   Steps to Perform Linear Regression

1. Data Preparation:
   - **Collect Data**: Gather data that includes both the features and the target variable.
   - **Preprocess Data**: Handle missing values, encode categorical variables, and scale features if needed.
2. Split Data:
   - **Training and Testing**: Split the dataset into training and testing sets to evaluate the model's performance.
3. Train Model:

- **Fit the Model**: Use the training data to fit the linear regression model. This involves estimating the coefficients that minimize the MSE.

4. Evaluate Model:
   - **Predict**: Use the model to make predictions on the test set.
   - **Assess Performance**: Evaluate the model's performance using metrics such as R-squared, Adjusted R-square, MSE, and Mean Absolute Error (MAE).
5. Interpret Results:
   - **Coefficients**: Analyse the coefficients to understand the impact of each predictor on the target variable.
   - **Model Diagnostics**: Check residual plots, perform hypothesis tests, and validate assumptions.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four distinct datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression lines. However, when these datasets are visualized, they reveal vastly different distributions and relationships between the variables. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how different datasets with similar statistical properties can have different structures.

Key Points of Anscombe's Quartet:

1. Identical Statistical Properties:
- All four datasets have the same mean for the x and y values.
- They have the same variance for x and y.
- Each dataset has the same correlation coefficient between x and y.
- The linear regression line (y = mx + c) is nearly the same for all four datasets.

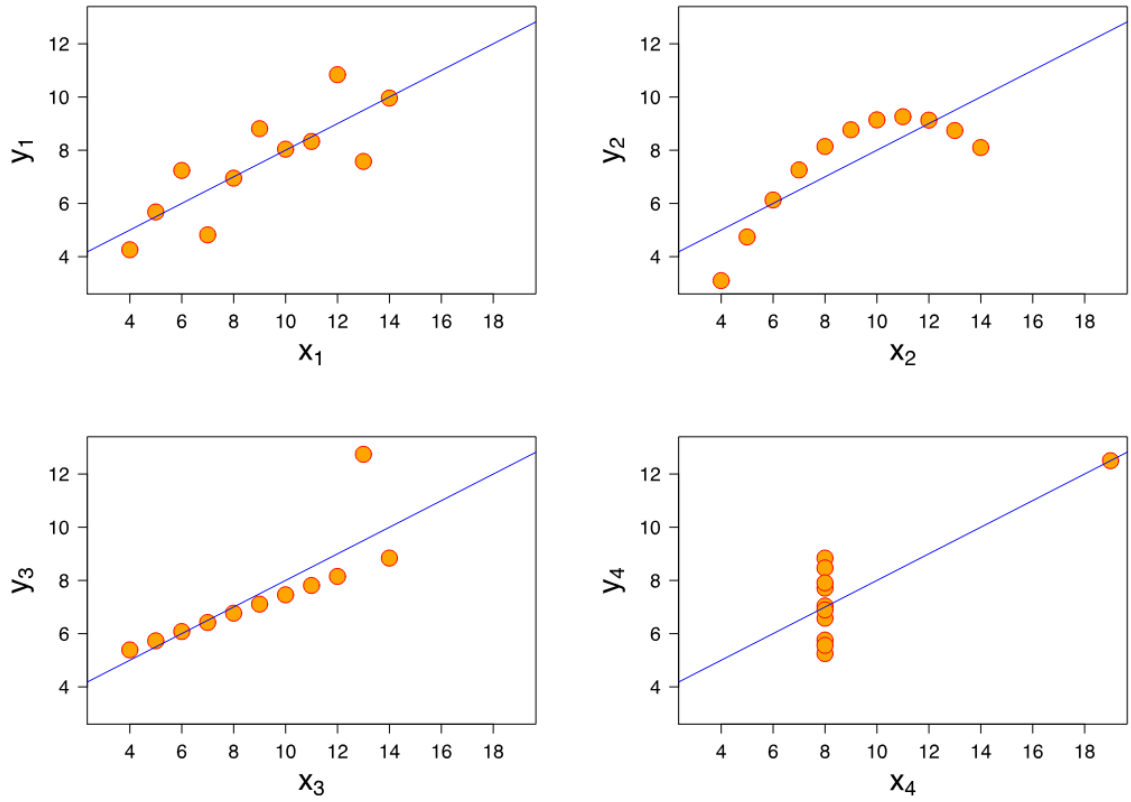
2. Visual Differences:

Dataset 1: Shows a typical linear relationship between x and y, which would be expected based on the regression analysis.

Dataset 2: The data is more curvilinear, indicating a non-linear relationship despite the linear regression line.

Dataset 3: Contains an outlier, which heavily influences the regression line, leading to misleading interpretations if only the statistics are considered.

Dataset 4: All x-values are the same except for one, creating a vertical line. The single differing point (an outlier) forces the regression line to fit in a misleading way.



Importance of Anscombe's Quartet:
Visual Exploration: The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.
Misleading Conclusions: If one relies solely on statistical summaries without visual inspection, one might draw incorrect or oversimplified conclusions about the data.
Teaching Tool: Anscombe's Quartet is widely used in statistics education to teach the importance of graphical analysis and to caution against the over-reliance on summary statistics.

Practical Takeaway:
Even though two datasets might have identical statistical properties, their underlying distributions and relationships can be entirely different. Always visualize your data to understand its true nature before relying solely on statistical summaries or models.

3. What is Pearson's R?

Pearson's R in Linear Regression is a measure of the linear correlation between two variables, typically the independent variable (predictor) and the dependent variable (outcome). It quantifies the strength and direction of the linear relationship between these two variables.

- Pearson's R, also known as the Pearson correlation coefficient, is a statistic that ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship between the variables.

- The Pearson's R is calculated as the ratio of the covariance of the two variables to the product of their standard deviations. Mathematically, it is expressed as:

- [R = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}]
  where:
- (\text{Cov}(X, Y)) is the covariance of the variables (X) (independent) and (Y) (dependent),
- (\sigma_X) and (\sigma_Y) are the standard deviations of (X) and (Y), respectively.

Importance in Linear Regression

In the context of linear regression, Pearson's R helps to understand how well the independent variable predicts the dependent variable. A high absolute value of R indicates that the independent variable is a good predictor of the dependent variable, which supports the linear regression model. However, it's important to remember that correlation does not imply causation; a strong correlation doesn't necessarily mean that one variable causes changes in the other.

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

Scaling refers to the process of adjusting the range of features in your dataset so that they can be compared on a common scale. This is particularly important in algorithms where the distance between data points or the magnitudes of features matter, such as in gradient descent optimization, support vector machines, or k-nearest neighbors.

Scaling is performed to:

- Improve Model Performance: Algorithms like gradient descent converge faster when features are scaled because the optimization process is more efficient when the features are within the same range.
- Ensure Equal Contribution Scaling ensures that all features contribute equally to the model. Without scaling, features with larger ranges could disproportionately influence the model's predictions.
- Enhance Interpretability It makes it easier to interpret the results of the model, particularly in distance-based algorithms.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling

Definition Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. It adjusts the values to be within a specific range without affecting the relative differences between data points.

Formula:
$Xnorm=Xmax-XminX-Xmin$

Use Case Normalization is useful when you need to bound the values within a certain range, for instance, in algorithms that need bounded inputs like neural networks.

Standardized Scaling

Definition Standardization scales the data based on the mean and standard deviation, resulting in features with a mean of 0 and a standard deviation of 1.

Formula:

$$Xstd = \sigma X - \mu$$

where \(\mu\) is the mean and \(\sigma\) is the standard deviation of the feature.

In summary, normalization constrains data within a specific range, making it useful for bounded inputs, while standardization adjusts data based on statistical properties, making it more suitable when the data's distribution is important.

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 multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among the independent variables. A VIF value becomes infinite when perfect multicollinearity is present, meaning that one independent variable is an exact linear combination of one or more other independent variables.

Cause of Infinite VIF

An infinite VIF occurs when the correlation between one independent variable and a combination of the other independent variables is perfect (correlation coefficient of 1 or -1). In this case, the model cannot distinguish between the perfectly correlated variables, and the variance of the affected variable's coefficient is infinitely inflated. Mathematically, this happens when the determinant of the matrix (X'X) used to compute VIF is zero, leading to a division by zero.

Implication and Solution

Implication: An infinite VIF indicates severe multicollinearity, which can destabilize the regression coefficients, making them highly sensitive to changes in the model. This undermines the reliability of the model and can lead to incorrect interpretations.

Solution: To address infinite VIF, you may need to:

Remove one of the perfectly correlated variables from the model. Combine the correlated variables into a single feature through techniques like Principal Component Analysis (PCA).

Rethink the model design to avoid including redundant predictors.

In summary, an infinite VIF signifies that one variable is a perfect linear combination of others, leading to perfect multicollinearity, which must be addressed to ensure the stability and reliability of the regression model.

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 specified theoretical distribution, such as the normal distribution. In the context of linear regression, a Q-Q plot is particularly useful for evaluating the assumption that the residuals (errors) of the model are normally distributed.

- A Q-Q plot compares the quantiles of the observed data with the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points in the Q-Q plot will lie approximately along a straight line.
- **Axes**:
    - **X-axis**: Quantiles of the theoretical distribution (e.g., standard normal distribution).
    - **Y-axis**: Quantiles of the observed data.
- : To visually inspect how well the observed data conforms to the theoretical distribution.

Creating and Interpreting a Q-Q Plot

1. Theoretical Distribution: Determine the theoretical distribution to compare against. Commonly, the normal distribution is used in linear regression.
2. Plot Construction:
    - **Calculate Quantiles**: Compute quantiles for both the observed data and the theoretical distribution.
    - **Plot Points**: Plot the quantiles of the observed data against the quantiles of the theoretical distribution.
3. Interpretation:
    - **Straight Line**: If the points lie approximately along a straight line (usually the 45-degree line), the data likely follows the theoretical distribution.
    - **Deviations**: Significant deviations from the straight line indicate departures from the theoretical distribution. For example, if the points form a curve, it may suggest skewness or kurtosis in the data.

Use and Importance in Linear Regression

In linear regression, the residuals (the differences between observed and predicted values) should ideally follow a normal distribution for several reasons:

1. Assumption Checking:
    - **Normality of Residuals**: One of the key assumptions of linear regression is that residuals are normally distributed. This assumption is important for accurate hypothesis testing and confidence interval estimation.

    - **Q-Q Plot for Residuals**: By creating a Q-Q plot of the residuals, you can visually check if they follow a normal distribution.
2. Model Validity:
    - **Model Diagnostics**: If residuals are not normally distributed, it might indicate issues with the model. For instance, it might suggest that a linear model is inappropriate, or there might be outliers or other issues affecting the residuals.
    - **Improving Model**: Identifying deviations from normality can help in diagnosing model fit issues and guiding adjustments to improve model performance.
3. Hypothesis Testing:
    - **Statistical Tests**: Many statistical tests and confidence intervals derived from linear regression rely on the assumption of normally distributed residuals. Ensuring this assumption is met helps validate the reliability of these tests.