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?**

- Bike Rentals are more during the Fall season and then in summer.
- Bike Rentals are more in the year 2019 compared to 2018.
- Bike Rentals are more in partly cloudy weather.
- Bike Rentals are more on Wednesday and Thursday.

2. **Why is it important to use drop first=True during dummy variable creation?**
   It is important to use drop first=True to prevent multicollinearity and improve Model interpretability.

- **Multicollinearity**: if we have n unique categories in a categorical feature, you would typically create n dummy variables. However, including all n dummy variables can lead to multicollinearity, as one of the dummy variables can be predicted perfectly from the others due to their sum being constant.
- **Interpretability**: When interpreting the coefficients of a regression model, having one less dummy variable for each categorical feature makes it easier to understand the effect of each category relative to a reference category. The coefficient associated with each remaining dummy variable represents the change in the response variable compared to the reference category.

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

   Temperature (atemp) and (temp) is highly correlated with Target variable.

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

- **Normal Distribution of Errors:** Residuals should be normally distributed, centred at mean value zero.
- **No Multicollinearity:** Calculate correlation and VIF to get the level of multicollinearity.
- **Constant Variance:** The residuals should have constant variance across all levels of the independent variables. A plot of residuals against predicted values should ideally show an even spread of points around zero.
- **No Outliers or High Influence Points:** Outliers data points can heavily influence regression coefficients and predictions. Identify outliers using methods like residual plots.

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

- Temperature has highest coefficient of .506. That means if temp increases by one unit then increase in bike rental demand by 50%.

- During Summer season, Bike Rentals are more in demand.
- Sprint Season is negatively correlated with Bike Rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- **Linear Regression** is a fundamental supervised machine learning algorithm used for predicting a continuous numeric output based on one or more input features. It models the relationship between the input features and the target variable as a linear equation.
- **Problem Statement**: Linear Regression is used when you have a dataset with pairs of input features (X) and corresponding target values (y). The goal is to find a linear equation that best fits the data.
- **Hypothesis Function**: In simple linear regression (one independent variable), the hypothesis function is represented as:
- $y = mx + b$
- **Cost Function (Loss Function)**: The cost function measures the difference between the predicted values and the actual values for all data points. The most common cost function used in linear regression is the Mean Squared Error (MSE):
- $MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$
- **Training the Model**: During the training phase, the algorithm adjusts the coefficients using Gradient Descent to find the optimal values that minimize the cost function.
- **Prediction**: Once the model is trained and the coefficients are learned, you can use the learned equation to make predictions on new, unseen data points.
- **Evaluation**: The model's performance is evaluated using various metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (coefficient of determination), which provide insights into how well the model fits the data.
- **Assumptions**: Linear Regression makes certain assumptions, including linearity between variables, independence of residuals, constant variance (homoscedasticity), and normally distributed residuals.

2. Explain the Anscombe's quartet in detail.

The four datasets in Anscombe's Quartet consist of 11 data points each, and they share the following characteristics:

- **Mean of x**: All datasets have a mean x-value of 9.
- **Mean of y**: All datasets have a mean y-value of approximately 7.5.
- **Regression Line**: A linear regression line fitted to each dataset has an equation of **y = 3 + 0.5x** for all four datasets.
- **Correlation Coefficient**: The correlation coefficient (Pearson's correlation) between x and y for each dataset is approximately 0.816.

Despite these common statistical properties, the datasets differ significantly when visualized:

Dataset I:

Scatter plot: Points follow a clear linear relationship.

Regression line fits well.

Dataset II:

Scatter plot: Points follow a linear relationship with one outlier.

The regression line is heavily influenced by the outlier.

Dataset III:

Scatter plot: Points form a curved pattern, indicating a non-linear relationship.

A linear regression line is not appropriate for this dataset.

Dataset IV:

Scatter plot: Points are mostly concentrated around two distinct y-values for each x-value.

The dataset can be split into two subsets, each with its own linear relationship.

The key takeaway from Anscombe's Quartet is that relying solely on summary statistics like means, variances, and correlation coefficients can be misleading. Visualization is crucial to understanding the underlying patterns in the data and making appropriate decisions. The quartet emphasizes the importance of graphing data before drawing conclusions, as different datasets with the same summary statistics can lead to vastly different insights and implications.

3. What is Pearson's R?

Pearson's correlation coefficient, often referred to as Pearson's R or simply Pearson's correlation, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It's a widely used method to assess how well two variables move together in a linear fashion.

The Pearson's correlation coefficient ranges between -1 and +1:

- A correlation of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A correlation of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A correlation of 0 indicates no linear relationship between the variables.

The formula to calculate Pearson's correlation coefficient between two variables X and Y is:

$$r = \Sigma((x_i - \bar{x})(y_i - \bar{y})) / \sqrt{(\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2)}$$

Where:

r: Pearson's correlation coefficient

xi: Each individual value of variable X

$\bar{x}$: Mean of variable X

yi: Corresponding individual value of variable Y

$\bar{y}$: Mean of variable Y

In essence, Pearson's correlation coefficient evaluates how much the actual pairs of data points deviate from their respective means. The numerator of the formula represents the "cross-deviations" (how much both variables deviate from their means together), and the denominator normalizes the values by their individual standard deviations.

Some important points to note about Pearson's correlation coefficient:

- Pearson's R only measures linear relationships. It might not accurately reflect the relationship if the association between the variables is nonlinear.
- Correlation does not imply causation. A strong correlation between two variables does not necessarily mean that one causes the other.
- Outliers can greatly influence the correlation coefficient, potentially leading to misleading interpretations. It's important to visualize data and consider the context.
- Correlation does not capture the magnitude of differences or the scale of the variables; it only assesses the strength and direction of the linear relationship.

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

Scaling in the context of data preprocessing refers to the process of transforming the features (variables) of your dataset to a similar scale.

Why Scaling is Performed:

- Algorithm Sensitivity: When features have different scales, the algorithm might give more weight to larger-scaled features, leading to suboptimal performance.
- Convergence: Scaling can help optimization algorithms converge more quickly and efficiently..
- Interpretability: Scaling ensures that the coefficients or importance values assigned to features are comparable and make sense in terms of the impact they have on the outcome.
- Regularization: Some regularization techniques, such as L1 and L2 regularization, assume that features are on similar scales. Scaling helps these techniques work effectively.
- Distance Metrics:. Scaling ensures that distances are computed accurately.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

- This method scales features to a specific range, typically between 0 and 1.
  Formula: X_normalized = (X - X_min) / (X_max - X_min)
- Keeps the distribution and relationships between data points intact.
- Useful when you need all features to be on the same scale but want to preserve the original distribution.

Standardized Scaling (Z-score Scaling):

- This method standardizes features to have a mean of 0 and a standard deviation of 1.
  Formula: X_standardized = (X - X_mean) / X_std
- Centers the data around zero, making it suitable for algorithms that assume a normal distribution.
- Useful when you want to transform the data into a standard normal distribution and deal with outliers.

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

An infinite VIF value typically arises due to perfect multicollinearity, where one predictor variable is a perfect linear combination of other predictor variables in the model. Perfect multicollinearity results in a rank-deficient design matrix, causing the inverse of the matrix to be non-existent, and consequently leading to an infinite VIF value for one or more variables. This happens because the formula for calculating VIF involves matrix inversion, and when perfect multicollinearity is present, the matrix is singular and cannot be inverted.

For example, consider a simple scenario with three predictor variables (X1, X2, X3) where X3 is a linear combination of X1 and X2:

X3 = 2 * X1 + 3 * X2

In this case, the design matrix would be rank-deficient, leading to an infinite VIF value for X3.

It's important to note that infinite VIF values are a clear indicator of a severe multicollinearity problem, and they warrant careful consideration and corrective action. In practice, if you encounter infinite VIF values, you should review your data and model to identify and address the underlying

multicollinearity issue. This might involve removing one of the correlated variables, reconsidering the model specification, or using techniques like regularization to mitigate multicollinearity's impact on coefficient estimates.

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

A **Q-Q plot**, short for Quantile-Quantile plot, is a graphical tool used to assess the similarity between the distribution of a sample of data and a theoretical distribution, typically the normal distribution. It is a powerful visualization technique to determine whether the data follows a particular distribution, such as the normal distribution, and to identify departures from that distribution.

Here's how a Q-Q plot works:

- **Data Preparation**: Collect a sample of data that you want to analyze and compare to a theoretical distribution. For example, if you're interested in checking whether your data is normally distributed, you would collect your dataset.
- **Sorting and Ranking**: Sort the data in ascending order and assign each observation a rank, such that the smallest observation gets rank 1, the second-smallest gets rank 2, and so on.
- **Theoretical Quantiles**: For each observation, calculate the theoretical quantile that corresponds to its rank. The theoretical quantile is derived from the chosen theoretical distribution. For a normal distribution, these quantiles can be calculated using the standard normal distribution (z-scores).
- **Creating the Q-Q Plot**: Plot the observed data quantiles on the y-axis and the corresponding theoretical quantiles on the x-axis. If the data follows the theoretical distribution closely, the points in the Q-Q plot should roughly form a straight line. Deviations from a straight line indicate departures from the theoretical distribution.

Use and Importance of Q-Q Plot in Linear Regression:

Q-Q plots are particularly useful in linear regression and other statistical analyses for the following reasons:

- **Normality Assessment**: In linear regression, the assumption of normality of residuals is important for accurate statistical inference. By plotting the residuals' quantiles against the expected quantiles from a normal distribution, you can visually check if the residuals are approximately normally distributed. Departures from the straight line in the Q-Q plot might indicate non-normality.
- **Detection of Outliers**: Q-Q plots can help identify outliers or extreme values in your data. Outliers might cause deviations from the expected straight line in the plot.
- **Model Validity**: Assessing normality of residuals and identifying potential outliers helps ensure the validity of your linear regression model's assumptions and results.
- **Model Improvements**: If non-normality or outliers are detected through the Q-Q plot, you might consider data transformations or applying robust regression techniques to improve the model's performance and reliability.