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?

Answer: Following are the observations done based on categorical variables:

- among seasons during the spring there is significant drop in demand of bikes
- *2019 saw more demand of bikes than 2018
- in a year's time the demand increased gradually from january to june and then it decreased till december
- demand was low during the holidays
- demand is high on clear day but low on snow and rainy day

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

Answer: When there are 'n' number of values in a categorical column, we can create 'n' number of dummy variables as columns. But just n-1 number of dummy variables are sufficient enough to express all the n variables . Hence it is ideal to remove any one of those n dummy variables. Removal of one dummy variable helps in saving memory and computational expense.

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

Answer: When pair plots are drawn to all the numerical variables, temp column has highest correlation with target variable.

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

Answer: I have validated the assumption of Linear Regression Model based on below assumptions

1. Normality of error terms - Error terms are normally distributed
2. Multicollinearity check - There is no significant multicollinearity among variables observed
3. Linear relationship validation - Linearity is visible among variables

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

Answer: The top 3 features contributing significantly are:

1. Temperature
2. Year
3. Light_snowrain

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, meaning changes in the independent variable(s) correspond to proportional changes in the dependent variable.

The equation for simple linear regression, involving one independent variable, is:

$$Y=\beta0+\beta1X+\epsilon$$

Here, Y represents the dependent variable, X is the independent variable, β0 is the y-intercept, β1 is the slope (showing how much Y changes for a unit change in X), and ϵ is the error term, accounting for variability not explained by X.

The objective of linear regression is to find the best-fitting line through data points by minimizing the sum of squared errors (differences between observed and predicted values), a method known as Ordinary Least Squares (OLS).

Key assumptions include:

- **Linearity**: The relationship between XXX and YYY is linear.
- **Independence**: Errors are independent.
- **Homoscedasticity**: Constant variance of errors across all levels of XXX.
- **Normality**: Errors are normally distributed.

Linear regression is evaluated using metrics like R-squared, which indicates the proportion of variance in the dependent variable explained by the independent

variables. It's widely applied in fields like economics, biology, and engineering for predicting outcomes and understanding relationships.

2. Explain the Anscombe's quartet in detail

Answer: Anscombe's quartet consists of four datasets, each with 11 pairs of x and y values. Despite having nearly identical simple descriptive statistics—such as the mean, variance, correlation, and linear regression line—the datasets display vastly different patterns when graphed.

Each dataset shares the following statistical properties:

- **Mean of x**: 9 for all datasets.
- **Mean of y**: Approximately 7.5 for all datasets.
- **Variance of x**: Approximately 11.
- **Variance of y**: Approximately 4.1.
- **Correlation between x and y**: Approximately 0.82.
- **Linear regression line**: Nearly $y=3.00+0.5xy = 3.00 + 0.5xy=3.00+0.5x$.

However, visualizing the datasets reveals unique characteristics:

1. **Dataset 1**: A typical linear relationship with data points near a straight line.
2. **Dataset 2**: A clear non-linear relationship, forming a curve.
3. **Dataset 3**: A linear trend with one outlier that skews the regression line.
4. **Dataset 4**: Almost all xxx values are the same except for one, creating a vertical line influenced by a single outlier.

Anscombe's quartet demonstrates the importance of data visualization, revealing patterns and outliers that summary statistics alone may obscure, highlighting the limitations of relying solely on numerical summaries.

3. What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is represented by the symbol r and ranges from -1 to 1.

- **Value Interpretation**:

    - An r value of **+1** indicates a perfect positive linear relationship, where increases in one variable correspond to proportional increases in the other.
    - An r value of **-1** indicates a perfect negative linear relationship, where increases in one variable correspond to proportional decreases in the other.
    - An r value of **0** suggests no linear relationship between the variables.
- **Calculation**: Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. This standardization makes r dimensionless, allowing for easy comparison across different datasets.
- **Usage**: It is widely used in statistics to assess the degree of linear association between two variables, helping to understand and predict relationships in data. However, it only captures linear relationships and can be misleading if the relationship is non-linear or if there are outliers.

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

Answer: Scaling is the process of adjusting the range of feature values in a dataset to ensure they are on a similar scale, which helps improve the performance of machine learning models.

Why Scaling is Performed: Scaling is done to ensure that no single feature dominates others due to its larger scale, which can impact model training, especially for distance-based algorithms (e.g., k-nearest neighbors, SVMs) and gradient descent-based methods (e.g., linear regression, neural networks).

Difference Between Normalized and Standardized Scaling:

- **Normalized Scaling**: Rescales data to a fixed range, typically [0, 1], based on the minimum and maximum values. It's useful when you want to compare data points relatively.
- **Standardized Scaling**: Transforms data to have a mean of 0 and a standard deviation of 1 (z-score normalization), useful when data follows a normal distribution or for algorithms that assume a standard distribution.

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

Answer: A Variance Inflation Factor (VIF) value becomes infinite when there is perfect multicollinearity in a dataset, meaning one predictor variable is an exact linear combination of one or more other predictors. This situation causes the denominator in the VIF formula (which involves the coefficient of determination, R^2, of a predictor variable regressed on all others) to become zero, leading to division by zero, and thus an infinite VIF.

Perfect multicollinearity can occur when:

- A variable is duplicated or is a direct linear transformation of another (e.g., measuring both height in inches and centimeters).
- There is an exact relationship between multiple variables (e.g., including both temperature in Celsius and Fahrenheit).

When VIF is infinite, it indicates redundancy among variables, making it impossible to separate their individual effects in a regression model.

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 assess if a dataset follows a particular theoretical distribution, most commonly a normal distribution. It plots the quantiles of the data against the quantiles of the specified distribution.

Use and Importance in Linear Regression:

In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot helps verify this assumption by comparing the distribution of residuals to a normal distribution. If the residuals lie approximately along the 45-degree reference line in the Q-Q plot, it indicates that they are normally distributed, supporting the model's validity. Deviations from this line suggest non-normality, which could imply potential issues with the model, such as skewness, heavy tails, or the presence of outliers, potentially impacting model accuracy and inference.