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?

    - Number of Bike booking is more in fall season compared to winter , spring and summer season.
    - Number of booking increased in 2019 compared to 2018
    - Number of bookings is comparatively more when the weather is clear
    - Irrespective of the holiday or workingday the number of bookings remains same.
    - Saturday, wednesday and thursday contributes more bookings
    - Number of bike booking is comparatively more in the month of August, September and October.
2. Why is it important to use drop_first=True during dummy variable creation?
    drop_first=True is used to
    - avoid multicollinearity and to simplify the model interpretation.
    - reduce the dimensionality.

    If the number of categorical variable is p then the number of dummy variable to be created is p-1
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    Temp variable has the highest correlation with the target variable count
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

    - Used pair plot to validate the linear relationship between dependent and independent variables
    - Used Durbin-Watson value to validate the independence of residual
    - Used VIF value to validate the no perfect multicollinearity
    - Used distplot on residual to validate the error terms are normally distributed with mean zero
    - Used scatter plot to validate the homoscedasticity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    Temperature
    Light Rain(Weather sit)
    Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.
    Two types:
    - Simple linear regression
    - Mulit linear regression

Simple linear regression:

Predicting target variable using one independent variable

Formula:

Y=mx+b

m- slope

b-intercept

Multiple Linear regression:

Predicting the target variable using more than one independent variable

Formula:

Y = b+m1x1+m2x2….mnxn

b-intercept

m-slope

n- number of independent variables

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, but have very different distributions and appear very different when plotted .

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where -1 represents a perfect negative linear relationship, +1 represents a perfect positive linear relationship, and 0 represents no linear relationship.

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

xi and yi are the individual data points of the two variables.

¯x¯ and ¯y¯ are the means of the respective variables.

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

Scaling is a preprocessing step in data preparation that involves transforming the numerical features of a dataset to a common scale. It ensures that all features have similar magnitudes which is important for many machine learning algorithms to work effectively and produce accurate results.

Scaling is performed for 2 reasons:

- Equalizing Feature Magnitudes
- Improving Convergence and Performance

Normalized Scaling:

In normalized scaling, each feature is scaled independently to have a minimum and maximum value within a specified range, usually [0, 1].

Normalized scaling is particularly useful when the distribution of the data is not necessarily Gaussian (normal), and you want to bring all features to a similar range.

Standardized Scaling:

In standardized scaling (also called z-score normalization), each feature is scaled to have a mean of 0 and a standard deviation of 1.

Standardized scaling assumes that the data follows a Gaussian distribution and is useful when you want to center the data around zero and ensure that the features have comparable variances.

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

The occurrence of infinite Variance Inflation Factor (VIF) values typically indicates a problem known as perfect multicollinearity in the context of multiple linear regression. Perfect multicollinearity happens when one or more independent variables in a regression model are perfectly correlated, leading to numerical instability in the calculations.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, typically the normal distribution. It provides a visual comparison between the quantiles of the observed data and the quantiles expected under the theoretical distribution. This plot helps to identify deviations from the assumed distribution and assess the normality of the data.

Use of Q-Q Plot:

Normality Assessment: One common assumption in linear regression is that the residuals (differences between observed and predicted values) should follow a normal distribution. Q-Q plots are used to visually assess whether this assumption holds true. If the residuals are approximately normally distributed, they should roughly follow a straight line on the Q-Q plot.

Importance of Q-Q Plot:

Assumption Checking: Linear regression models rely on certain assumptions, and violation of these assumptions can lead to biased or unreliable results. Normality of residuals is one such assumption. Q-Q plots provide an intuitive and graphical way to check whether the assumption of normality is met.

Detecting Departures: Deviations from the expected straight line on a Q-Q plot can indicate departures from normality. If the plot shows a curve, skewness, heavy tails, or outliers, it suggests that the data might not be normally distributed. Identifying these departures early allows you to consider alternative modeling techniques or transformations.

Model Validity: Ensuring that the residuals are normally distributed is crucial for the validity of statistical inference, hypothesis testing, and confidence interval estimation. If the residuals deviate significantly from normality, the p-values and confidence intervals obtained from the regression analysis may not be accurate.

Outlier Detection: In addition to assessing normality, Q-Q plots can help detect outliers in the data. Outliers might cause deviations from the expected straight line pattern in the plot.