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?

Results from Bar charts, Scatter Plot and Box Plots is:
1)Number of registrations are more in fall summer
2) The demad of bike is less in the month of spring when compared with other seasons

3)The number of bike registrations was higher in September 2019 compared to other months that year, whereas in 2018, June witnessed the highest number of bike registrations.
4)The number of bike registrations on mon, tue and wed is low.
5)When there is Clear, Few clouds, Partly cloudy, Partly cloudy then there are many bike registrations done
6)Number of bike registrations are more in 2019
7)There is no significant change in bike demand with workign day and non working day.

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

When creating dummy variables from categorical data, drop_first=True is an important parameter to consider. It is specifically used in the process of one-hot encoding to avoid the **dummy variable trap and multicollinearity** issues. It saves **degrees of freedom, improves model efficiency**, and ensures reliable and interpretable results in statistical analysis or machine learning models.

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

Looking at the pair-plot **Temp** has highest positive correlation with target variable count

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

After building a linear regression model on the training set, you need to validate its assumptions to ensure the model's reliability and accuracy. Here are three key steps for validating the assumptions of linear regression:

Residual Analysis: Calculate the residuals (the differences between the actual target values and the predicted values) for the training data. Plot a scatter plot of the residuals against the predicted values. The plot should not show any discernible patterns or trends; otherwise, it indicates that the model might not capture all the relevant information in the data.

Normality of Residuals: Check the normality of the residuals by creating a histogram or a Q-Q plot. The residuals should follow a roughly normal distribution around zero. If the residuals deviate significantly from normality, it might suggest that the model assumptions are violated.

Homoscedasticity: Assess the homoscedasticity (constant variance of residuals) by plotting the residuals against the predicted values or independent variables. The scatter plot should show a relatively constant spread of points without a funnel shape. Heteroscedasticity, where the spread of residuals changes with the predicted values, can indicate that the model's assumptions are not met.

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

Spring season : -0.4947

Temperature : 0.3929

Mix+Few clouds : -0.3543

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm predicting a continuous target based on one or more independent variables. It finds the best-fitted line (simple) or hyperplane (multiple) representing the relationship.

The model is **Y = β0 + β1*X + ε. β0** is intercept, β1 is slope, and ε is error. The goal is to minimize sum of squared errors (SSE) or mean squared error (MSE). Ordinary Least Squares (OLS) fits the model by adjusting coefficients. Evaluation metrics include MSE, mean absolute error (MAE), and R-squared. R-squared gauges model's fit to data, higher values indicate better fit. Linear regression provides a simple approach to modeling relationships and making predictions.

Example: Predicting house prices based on features like area, bedrooms, and location.

2) Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four datasets with identical stats but different patterns. Created by Anscombe, it emphasizes data visualization's role in understanding stats. Despite similar stats, quartet's plots vary significantly, showing summary stats' limitations.

Datasets I, II, III, IV each contain 11 points (X, Y). Notable properties:

Dataset I: Linear relationship suited for linear regression.

Dataset II: Non-linear quadratic pattern, unfit for simple linear regression.

Dataset III: Mostly linear, outlier alters regression line.

Dataset IV: Two clusters with weak linear relationships, reveals insights when plotted together.

Takeaway: Visual exploration is vital. Summary stats can mislead; data visualization uncovers underlying patterns. Anscombe's quartet cautions against relying solely on summary stats, highlights data visualization's value.

Example: Analyzing car mileage based on engine displacement and horsepower.

3) What is Pearson’s R?

Pearson's R measures linear correlation between continuous variables. R ranges -1 to +1:

r = +1: Perfect positive linear relationship.

r = -1: Perfect negative linear relationship.

r = 0: No linear relationship.

Formula: **r = (Σ((X_i - X_mean) * (Y_i - Y_mean))) / (n * X_std * Y_std)**

Example: Height and weight correlation in a population.

Pearson's R assesses linear association strength and direction. Limitation: Doesn't capture non-linear relationships. Other measures like Spearman's rank correlation or Kendall's tau are used for that.

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

Scaling transforms features to a common scale, ensuring comparable magnitudes. Some ML algorithms are sensitive to feature scale. Scaling prevents bias due to varying magnitudes.

Normalized Scaling (Min-Max Scaling): In normalized scaling, also known as Min-Max scaling, the values of features are transformed to a specific range, typically between 0 and 1. The formula to perform normalized scaling for a feature X is:

X_normalized = (X - X_min) / (X_max - X_min)

where X_min is the minimum value of the feature, and X_max is the maximum value.

Example: Image pixel intensity scaling.

Standardized Scaling (Z-score Scaling):

In standardized scaling, the values of features are transformed to have a mean of 0 and a standard deviation of 1. This method is also known as Z-score scaling. The formula for standardized scaling is:

X_standardized = (X - X_mean) / X_std

where X_mean is the mean of the feature, and X_std is the standard deviation.

Example: Exam scores with different scales.

Scaling ensures fairness. Normalized scaling confines values to a specific range, while standardized scaling centers on mean and scales by std deviation. Choice depends on dataset and algorithm.

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

Infinite VIF occurs due to perfect **multicollinearity**. This arises when an independent variable can be precisely predicted by others. This disrupts ordinary **least squares (OLS)** estimation used in VIF calculation. VIF is reciprocal of tolerance. Tolerance is the unexplained variance proportion in a variable by other variables. Perfect multicollinearity leads to zero tolerance, yielding infinite VIF.

Practically, infinite VIF signifies a redundant predictor, rendering model estimation ineffective. Solution: Remove correlated variables or use regularization (e.g., Ridge/Lasso regression). Addressing multicollinearity ensures meaningful linear regression results.

Example: Predicting house price using both area and number of bedrooms.

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

A Q-Q plot assesses dataset normality by comparing observed quantiles to a theoretical distribution. In linear regression:

Normality Assessment: Q-Q plots check if residuals follow a normal distribution. Straight points imply normality; deviations suggest departures.

Residual Diagnostics: Q-Q plots identify non-normality patterns, like heavy tails or skewness. Non-normal residuals can bias regression coefficients.

Model Assumptions: Q-Q plots validate linear regression assumptions of normal and constant variance errors. Violations impact test validity.

Decision Making: A well-behaved Q-Q plot boosts confidence in regression assumptions. Deviations hint at further investigation or data transformation.

In essence, Q-Q plots diagnose residual normality, aiding in assessing model assumptions and regression reliability.

Example: Checking if residuals follow a normal distribution in predicting exam **scores.**