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?**

    1. Temperature - Increase in temp, increases the demand of bike rentals.
    2. Year - 2019 fared lot better than 2018 and hence increasing in trend bikes Rentals
    3. Bike rentals are more in Holidays than on working days / Non-Holidays.
    4. High Demand in September - The season of Fall, has a better demand than any other season
    5. There are no Bike rentals in Heavy Snow Rainy climate.

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

    Using drop_first=True during dummy variable creation avoids the dummy variable trap of redundancy by preventing multicollinearity. This ensures that the resulting dummy variables are linearly independent, as one category is used as a baseline and not represented explicitly.

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

    Temperature columns i.e. “temp” and “atemp” have highest correlation with the target variable.

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

    The assumptions were validated by checking the R-squared value to assess the model's goodness of fit, ensuring VIF values were below limited value to confirm low multicollinearity, and analysing residual plots for constant variance and normal distribution.

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

    - Year
    - Temperature
    - Holiday
    - Spring and Fall season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a method used to find the relationship between one dependent variable (the outcome we want to predict) and one or more independent variables (the inputs we use to make predictions).

The formula for this relationship is written as:

$Y=b0+b1X1+b2X2+\cdots+bnXn + E$

- Y: The dependent variable (what we are predicting).
- X1,X2,...,Xn : The independent variables (the predictors).
- b0: The intercept (the starting point of the line when all predictors are zero).
- b1,b2,...,bn: The coefficients (how much the dependent variable changes when the independent variables change).
- E : The error term (the difference between the predicted and actual values).

The main goal of linear regression is to find the best values for b0,b1,...,bn that minimize the difference between the actual and predicted values of Y. This is done by minimizing the sum of the squared errors, making the line of best fit as close as possible to all data points.

Following assumptions exists in Linear Regression -

i. There is a linear relationship between X and Y

ii. Error terms are normally distributed with mean zero(not X, Y)

iii. Error terms are independent of each other

iv. Error terms have constant variance (homoscedasticity)

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four datasets with identical statistical properties, such as means, variances, and correlation coefficients, yet they differ significantly in their graphical representations. These datasets demonstrate that relying solely on summary statistics can be misleading, as different data distributions can produce the same numerical results. Visualizing data is crucial to understanding its true nature and ensuring accurate analysis.

3. What is Pearson's R?

Pearson correlation coefficient is a measure of the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. It quantifies

how well the values of one variable predict the values of another, with higher absolute values indicating a stronger linear relationship.

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

Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range of features in a dataset. It ensures that all features contribute equally to the analysis, preventing any single feature with a larger range from disproportionately influencing the model's performance.

Reasons for Scaling:

1. Improves Model Performance: Some machine learning algorithms, such as gradient descent-based algorithms, are sensitive to the range of input features. Scaling helps them converge faster and perform better.
2. Equal Contribution of Features: When features have different scales, those with larger ranges can dominate the model's output. Scaling ensures that each feature contributes equally to the result.
3. Improves Accuracy: It can enhance the accuracy and performance of models, especially those that rely on distance calculations

Difference Between Normalized Scaling and Standardized Scaling:

- Normalized Scaling: This technique rescales the data to a fixed range, usually [0, 1] or [-1, 1]. It is useful when you want to ensure that all features are on the same scale without considering the distribution of the data. Normalization is particularly useful when the data has different units or scales.
- Standardized Scaling: This technique centres the data by subtracting the mean and scales it to have a unit variance. Standardization is useful when the data follows a Gaussian (normal) distribution and when it is important to consider the distribution shape. It helps to handle outliers by reducing their influence on the model

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

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the independent variables in a regression model. This occurs when one predictor variable can be perfectly predicted from a linear combination of the other predictors. In mathematical terms, this means the correlation between one variable and the others is exactly ±1, causing the denominator in the VIF calculation ($1 - R^2$) to become zero. As a result, dividing by zero leads to an infinite VIF value, indicating that the regression coefficients are not

uniquely determined, and the model cannot accurately separate the effects of the collinear variables.

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, often the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution

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

1. Assessing Normality of Residuals: In linear regression, one of the assumptions is that the residuals (errors) are normally distributed. A Q-Q plot can help visually assess if this assumption is met by comparing the distribution of the residuals with a normal distribution.
2. Identifying Outliers: Deviations from the straight line in a Q-Q plot can indicate the presence of outliers in the data. Points that lie far from the line suggest that the data contains values that are not consistent with the assumed distribution.
3. Detecting Skewness: If the points on the Q-Q plot form an S-shaped curve, this indicates that the data may be skewed. The direction of the curve indicates whether the skewness is positive or negative.
4. Assessing Homoscedasticity: While not a primary tool for this, Q-Q plots can sometimes provide hints about heteroscedasticity (non-constant variance) if the residuals deviate systematically from the line.
5. Model Evaluation: By examining the Q-Q plot, one can determine if the linear regression model is appropriate for the data. Significant departures from the expected distribution may suggest that the model's assumptions are violated, indicating a need for model modification or alternative approaches.