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? (3 marks)

   Season: when it is spring number of bookings are least compared to the summer, fall and winter.

   'Yr': in 2019 we can see more number of people were renting the bike compared to 2018.

   'Weathersit': We can see whenever weather is mild the bookings tend to be more compared to when the weather is harsh.

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

   This will drop one of the dummy variable since with n-1 dummy variables are enough to explain all the possible categories. 'n' is number of categories/states in the categorical variable. (eg : if n is 3 then that variable has 3 states, hence 2 dummy variables are enough)

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Categorical: Yr, Season, Weathersit. Numerical: atemp/temp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   One way would be calculating the r2 value for y_trained against y_pred(this value is predicted on X_trained values).

   Second way would be plotting the (y_train vs X_train) and (Y_predict vs X_train), both the graphs shall have least divergence or less RSS.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   season_3, weathersit_3, season_2, Yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, meaning the change in the dependent variable is proportional to the change in the independent variables. The model is represented by the equation y = mx+c, where m is the coeffient, c is the error term, y is dependant variable, x is independent variable. The goal is to find the best-fitting line (or hyperplane in higher dimensions) by minimizing the sum of the squared differences between observed and predicted values, known as the least squares criterion.

2. Explain the Anscombe's quartet in detail. (3 marks)

   Anscombe's quartet consists of four datasets that share nearly identical summary statistics, such as mean, variance, correlation, and linear regression lines. However, when plotted, they exhibit strikingly different patterns.

   - Dataset 1: Displays a simple linear relationship, with points closely clustered around a straight line.
   - Dataset 2: Shows a nonlinear relationship, with points forming a clear curve, indicating that a linear model might not be suitable.
   - Dataset 3: Contains a nearly linear distribution of points, with a single outlier that significantly influences the regression line and correlation.
   - Dataset 4: Features a cluster of points with one extreme outlier, affecting the regression line and correlation despite the majority of data points being constant.

   Anscombe's quartet demonstrates that similar statistical properties can mask very different data distributions. It highlights the need for graphical representations to uncover true data patterns, relationships, and outliers that simple statistics can overlook.

3. What is Pearson's R? (3 marks)

   Pearson's R, also known as the Pearson correlation coefficient, 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 indicates a perfect positive linear relationship: as one variable increases, the other variable increases proportionally.
   - -1 indicates a perfect negative linear relationship: as one variable increases, the other decreases proportionally.
   - 0 indicates no linear relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Scaling is a process in data preprocessing where the features of a dataset are transformed to fit within a specific range or distribution. It is performed to ensure that no single feature dominates others due to differences in scale, which can significantly impact the performance of machine learning algorithms, especially those based on distance metrics or gradient descent.

   - Normalized scaling (or min-max scaling) rescales the features to a range, typically [0, 1], using the formula:

     x' = (x-xmin)/(xmax-xmin)

     This method preserves the relationships between values, proportionally adjusting them.

   - Standardized scaling (or z-score normalization) transforms the features to have a mean of 0 and a standard deviation of 1.

   The key difference is that normalization rescales data to a fixed range, while standardization adjusts data to a common scale with a mean of 0 and a standard deviation of 1.

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

(3 marks)

A VIF becomes infinite when R2 equals 1, indicating perfect multicollinearity. This situation arises when a predictor variable is a perfect linear combination of other predictors in the model, meaning there is exact redundancy. As a result, the denominator in the VIF formula becomes zero, leading to an infinite VIF. This indicates that the model cannot uniquely estimate the coefficients of the predictor variables, making the regression results unreliable.

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

(3 marks)

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

Use and Importance in Linear Regression:

- Normality Check: In linear regression, residuals (errors) should ideally follow a normal distribution. A Q-Q plot helps assess this assumption by showing if residuals deviate from normality.
- Model Validation: By comparing the distribution of residuals to a normal distribution, the Q-Q plot helps validate the regression model. Deviations from the diagonal line suggest non-normal residuals, which could indicate issues with the model or data.
- Diagnosing Issues: If the points deviate significantly from the line, it may signal problems like outliers or model misspecification, prompting further investigation and potential adjustments.