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)

Categorical variables like 'weathersit', 'seasons' have impact on the dependent variable in linear fashion and has both positive and negative correlation. For example, the bike rental count is low in spring and high in fall season. There is also a linear increase in the dependent variable with the 'year' categorical variable as there is an year on year increase in the bike rental count. Variable like 'weekday' and 'workingday' has no impact on the dependent variable.

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

It avoids multi-collinearity by removing one of the dummy variables which would act as the baseline after removal and avoids a perfect correlation amongst them. It also improves the interpretability as the coefficients are compared to the baseline (dropped dummy variable). It also improves the efficiency of the model to some extent based on the total number of variables as there will be lesser variables to train.

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

From the pair plot, atemp has the highest correlation with the target variable as the regression line will have a better fit except for a few outliers. While temp also has similar correlation the width of the data is narrower in case of atemp indicating a possible better fit for a regression model.

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

The linear regression model has been validated using a scatter plot with predicted vs actual values on test data along with the regression line to see that the line fits the data well. There shall be no multi-collinearity between the selected features and has been verified with the VIF values of the features and found to be in the acceptable range (<=5). The residuals, difference between actuals and predicted should be normally distributed and the peak at 0 which has been confirmed by the distribution plot which shows a perfect normal distribution of residuals.

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

Based on the final model, we could conclude that the temperature has the highest positive correlation indicating a higher demand with the increase in temperature. There is a high negative correlation between the bike demand and the light snow rain weather situation indicating if the weather is light rain, snowy then the demand for bike is less. The year variable indicates an increase in demand for shared bikes year on year and has high positive correlation.

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used to predict a continuous target variable based on one or more predictor variables. The algorithm assumes a linear relationship between the target and predictor variables. The goal is to find the best-fitting line that minimizes the difference between the predicted and actual values.

The equation of the line is represented as: y = mx + c, where 'y' is the target variable, 'x' is the predictor variable, 'm' is the slope, and 'c' is the y-intercept. The algorithm estimates the optimal values for 'm' and 'c' by minimizing a cost function, typically the mean squared error (MSE) which measures the average squared difference between predicted and actual values.

Once the model is trained, it can be used to predict the target variable for new input data. Techniques, such as gradient descent is used to find the optimal values for the model parameters. It's important to assess the model's performance using metrics like R-squared and adjusted R-squared. Regularization methods like Ridge and Lasso are used to prevent overfitting and improve model generalization.

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

Anscombe’s quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation, and regression line), but they differ greatly when visualized. The datasets highlight how relying only on statistical summaries can be misleading, as graphs reveal key differences like outliers or non-linear relationships. The quartet demonstrates the importance of visualizing data alongside statistical analysis to gain a full understanding and avoid misinterpretations.

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

Pearson correlation coefficient (or Pearson's R), is a measure of the linear relationship between two variables. It quantifies how closely the values of one variable are related to the values of another on a scale from -1 to 1.

- +1: Perfect positive correlation (as one variable increases, the other increases in a linear fashion).
- -1: Perfect negative correlation (as one variable increases, the other decreases in a linear fashion).
- 0: No linear correlation (the variables don't have a linear relationship).

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

Scaling is the process of adjusting feature values to a similar range, ensuring that no feature is prioritized due to its size. It's important for improving model performance, especially in algorithms like linear regression and gradient descent, which are sensitive to scale of the features.

- Normalization (Min-Max Scaling) resizes values to a fixed range, typically [0, 1]. It's best for keeping relative relationships intact while compressing data into a small range.
- Standardization (Z-score Scaling) adjusts data so that it has a mean of 0 and a standard deviation of 1. It's useful for features with different distributions and units, ensuring they follow a normal distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between two or more predictor variables in a regression model.

Perfect Multicollinearity occurs when the relationship between two variables is so strong that one can be exactly predicted from the other(s). This makes the denominator in the VIF formula zero, leading to an infinite value.

The VIF formula is:

$VIF = 1/ (1-R^2)$, If $R^2 = 1$ (perfect correlation), the denominator becomes zero, making VIF infinite.

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 assess if a dataset follows a particular theoretical distribution, such as a normal distribution. The Q-Q plot compares the quantiles of the observed data against the quantiles of the theoretical distribution. If the data follows the assumed distribution, the points on the plot will fall approximately along a straight line.

A Q-Q plot of the residuals can help visually check the linear regression assumption that the residuals (errors) are normally distributed. If the points deviate significantly from the straight line, it suggests that the residuals may not be normally distributed.

Q-Q plots help assess the validity of the normality assumption, to identify the outliers.