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)

   Ans.
   Categorical variables such as Year (yr), season, weathersit have significant effect on the dependent variable "cnt"

   Year '2019' has more demand for BoomBikes
   Seasons 'summer' and 'fall' has comparatively more demand than the season 'spring' and 'winter'
   Weather type 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' has less demand for BoomBikes

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

   Ans.
   By using **drop_first=True** we avoid issue of Multi collinearity between the dummy variables created for categorical variables
   Interpretability of the dummy variables will be improved
   By dropping one of the columns we avoid the complexity of relation between independent variables

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

   Ans.
   'temp' and 'atemp' columns has high correlation with target variables

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

   Ans.
   a. Error terms are Normally distributed concentrated/centered around 0
   b. Multi collinearity within the permissible limits as VIF values are less than 5
   c. There is a Linear relation between dependent and independent variables

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

   Ans.
   Year (Yr) with co efficient value 0.2367,' atemp' with co efficient value 0.4365 and 'weather3' with co efficient value -0.2822

General Subjective Questions

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

Ans.

Linear regression is modeling a relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It is a supervised learning algorithm used for regression

Detailed explanation of the linear regression algorithm:

1. Simple Linear Regression and Multiple Linear Regression:

Simple Linear Regression: When there is only one independent variable, it's called simple linear regression. The equation for simple linear regression is:

$Y=b0+b1*X+\varepsilon$

Y is the dependent variable.

X is the independent variable.

b0 is the intercept (where the line crosses the Y-axis).

b1 is the slope (the change in Y for a unit change in X).

Multiple Linear Regression: When there are multiple independent variables, it's called multiple linear regression. The equation becomes:

$Y=b0+b1*X1+b2*X2+...+bn*Xn+\varepsilon$

2. Model Training: To find the best-fit line (or hyperplane in the case of multiple linear regression), the model aims to minimize the sum of squared differences between the predicted values and the actual values. This is typically done using the method of least squares, where the objective is to minimize the residual sum of squares (RSS) or the mean squared error (MSE).

3. Coefficient Estimation: The coefficients (intercept and slopes) are estimated during model training.

4. Making Predictions: Once the coefficients are estimated, you can use the linear equation to make predictions for new data points

5. Model Evaluation: To assess the model's performance, you typically use metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared (coefficient of determination). These metrics help measure how well the model fits the data and how accurately it makes predictions.

6. Assumptions of Linear Regression: It's important to check whether the assumptions of linear regression are met, including linearity, independence of errors, homoscedasticity, normality of residuals, and absence of multicollinearity.

In summary, linear regression is a simple yet powerful algorithm used for modeling the relationship between dependent and independent variables by fitting a linear equation to the data. It's widely applied in various fields, including economics, finance, and machine learning.

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

Ans.

Anscombe's quartet consist of four datasets, having similar statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we plot on graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the

drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Ans.
Pearson's R is a coefficient which defines the linear correlation between two variables.
Value ranges between -1 and 1

Pearson's R between 0 and 1: When one variable changes, the other variable changes in the same direction.
Pearson's R value is 0: There is no linear correlation

Pearson's R between 0 and -1: When one variable changes, the other variable changes in the different direction.

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

Ans.
A data preprocessing step applied to independent variables to normalize the data within particular range.

As the data received has different columns with different units, different magnitude, different ranges. Scaling is done to bring all variables to same level of magnitude

Normalized scaling brings all the data ranges between 0 and 1

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Ans.
If there is a perfect correlation between two variables VIF value will be infinite, as the correlation between two variables increases VIF value increases

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans.
Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential

As you build your machine learning model, ensure you check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, you might want to check the distribution of your feature variable and consider transforming them into a normal shape.