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)

   The dependent variable has a positive relationship with the month of June, July , Aug & Sept.
   Summer, Fall and Winter has a higher median Cnt values.
   When the weather is clear, the bike company makes more revenue.
   2019 has a higher impact on the dependent variable.

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

   The categorical variable having multiple levels can be replaced by dummy variable.We have observed that if there are N levels of the categorical variable then N-1 variable are able to capture the essence of the categorical variable so one of the variable (the first one) can be dropped from the analysis data set.

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

   The variable temp (atemp) has the highest correlation with the dependent variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   After building the Linear Regression model on the training set, we tested the model on the test data set and checked for the R2 value. There is not much difference between the R2 values for the train data set (85.2% adjusted R2= 84.5%) and test data set (R2 = 81.9%). So we are confident that the model is predicting the results on the test data set within the same confidence level as on training data set.

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, the top 3 features contributing towards explaining the demand of the shared bikes are temp(0.4680), Light Snow (-0.2536) and wind speed (-0.1898). The data from 2019 also had a coefficient of (0.2314) but that is a historic dataset.

General Subjective Questions

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

Concept

Linear regression is a method to model the relationship between a dependent variable and one or more independent variable.

Mathematical Interpretation

In Linear Regression, A line (in case of single independent variable) or Hyperplane is fitted on the data.

The equation for the same is

Linear Regression

$Y = B0 + B1X + B2X…+BnX + e$

Where Y = Dependent or Output variable

B0 = Intercept or the constant.

B1,B2…Bn = Coefficients of the independent variables

e= Error or residual

Assumption

While fitting a lint or hyper plane using linear regression of a data set, following assumption are made

a) Linearity : Relationship between Independent variable and target variable is linear.
b) Independence : Observations are independent of each other.
c) Homoscedasticity : the error or residuals are constant
d) Normality : The residuals are normally distributed
e) No Multicollinearity : Independent variables are not correlated with each other.

Model Fit

The final objective is to fit a line or hyperplane and identify the constant and coefficients of the independent variable in such a way that the sum of squares of the error of actual values and predicted values is minimum.

Evaluating the Model

The model can be evaluated based on various statistical parameters like R2, Adjusted R2 etc.

Writing the code and segregating the data set

Programming languages like python, R etc can be used to evaluate the Intercepts and coefficients of the independent variable within the statistical significance by segregating the past data into the training set and test set.

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

   Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean , variance, correlation) but are graphically very different.

   Anscombe's quartet serves as a powerful reminder of the importance of data visualization in statistical analysis. Despite having the same summary statistics, the four datasets reveal vastly different patterns and relationships when visualized. This underscores the limitations of relying solely on statistical summaries and highlights the necessity of visual inspection for accurate data interpretation.

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

   Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It is a statistic that quantifies the degree to which two variables are linearly related. The coefficient is named after Karl Pearson, who developed it.

   Interpretation of Pearson's r

   The value of r ranges from -1 to 1.

   - r = 1: Perfect positive linear relationship.
   - r = -1: Perfect negative linear relationship.
   - r= 0: No linear relationship.
   - Values between 0 and ±1 indicate the strength and direction of the linear relationship.

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

   Scaling is important preprocessing technique to standardize the independent variables. Scaling is performed on the independent variable because the algorithms may be sensitive to the magnitude of data and model performed optimally when the data is in a standard range post scaling. Scaling ensures that all features contribute equally to the distance calculations and optimization process, improving the performance and training stability of the model.

Improves Model Performance: Scaling can improve the performance of algorithms that are sensitive to the scale of the data, such as gradient descent-based algorithms.
Convergence Speed: For optimization algorithms, like gradient descent, scaling can lead to faster convergence.
Equal Contribution: Ensures that all features contribute equally to the analysis. Features with larger ranges can dominate the learning process if scaling is not performed.
Improved Accuracy: Helps in achieving better accuracy by preventing the model from being biased towards certain features.
Reduced Risk of Over fitting: Scaling can help reduce the risk of over fitting, especially in regularization methods where different ranges can lead to different penalties.

Types of scaling

Normalized Scaling = (X-Xmin)/(Xmax - Xmin)
Standardized Scaling = (X-mu)/Std dev

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

   VIF = 1 /(1-R2)

   So if VIF = infinite which means 1-R2 = 0 or R2 =1.
   R2 = 1 indicate that the independent variable is perfectly correlated with the other independent variables. To address infinite VIF, you can remove one of the perfectly correlated variables or combine them into a single predictor.

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 to help assess if a dataset follows a particular distribution, such as the normal distribution. It plots the quantiles of the data against the quantiles of the specified theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line

   Interpretation of Q-Q plot

   - Straight Line: If the points form a straight line, the data follows the theoretical distribution.
   - S-Shaped Curve: Indicates skewness. For normality, a rightward S-shape indicates positive skew, and a leftward S-shape indicates negative skew.
   - Upward/Downward Curves: Indicate heavy tails or light tails. Data points above the line on the right end and below the line on the left end indicate heavy tails, while the opposite pattern indicates light tails.