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)

    a. Season has a strong impact – Count of rentals is very low in Spring and count is high during fall.
    b. Weathersit also has a strong impact – Count of rentals is very low during wet days(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and count is high when the weather is Clear(Clear, Few clouds, Partly cloudy, Partly cloudy).
    c. Working day or any weekdays doesn't have much impact on Count of rentals.

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

    a. It is important to use drop_first=True during dummy variable creation to avoid multicollinearity.
    b. When we create dummy variables for a categorical variable with 'n' categories, we create 'n-1' dummy variables.
    c. If we don't drop the first variable, the model becomes redundant and causes issues with interpretation.

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

- Numerical value 'atemp' has the highest correlation with target variable 'cnt'.

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

    a. Plotted a histogram of residuals and checked if it is a normal distribution
    b. Checked for multicollinearity using VIF values. It is in acceptable range below 5 for all chosen variables of the model built.
    c. A scatter plot of residuals vs. predicted values should not show a clear pattern or funnel shape indicating heteroscedasticity.

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

- Temp(temperature), yr(Year), Wet(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) are the top 3 features contributing significantly towards explaining the demand of the shared bikes. This is obtained by looking at the absolute high co-efficient values.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    a. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. It is a statistical

method that is used for predictive analysis. Linear regression makes predictions for numeric variables such as sales, salary etc. It finds how the value of the dependent variable is changing according to the value of the independent variable.

b. The steps involved are
    i. Reading, Understanding & Visualizing the data (Import Data, Create relevant Plots)
    ii. Prepare the data for modelling (train-test split, scaling)
    iii. Create & Train the model
    iv. Residual Analysis
    v. Predictions & Evaluation using test set

2. Explain the Anscombe's quartet in detail. (3 marks)
    a. Anscombe's quartet are 4 datasets created by Francis Anscombe. It illustrates the danger on relying numerical summaries alone when looking at a dataset.
    b. The summary statistics such as mean, variance, correlation and regression co-efficient are same for all 4 datasets.
    c. But when plotted in a graph they are considerably different
    d. The Anscombe's quartet emphasizes **data visualization first** approach.
    e. It shows the necessity of combining statistical analysis with graphical exploration for more correct data interpretation.
3. What is Pearson's R? (3 marks)
    a. Pearson's R quantifies how strongly two variables are related and whether the relationship is positive or negative. The correlation coefficient ranges between -1 and +1.
    b. +1 indicates strong positive correlation, -1 indicates strong negative correlation and 0 indicates no linear correlation. The variables can be related but not linearly.
    c. Pearson's R is used to identify relevant features that have a strong linear correlation with the target variable
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    a. Scaling is a preprocessing technique in machine learning that transforms numerical data to a common range.
    b. Scaling of variables is an important step because if the feature variables are at different scale the co-efficient of the variables will also be in a very different scale making our calculations & interpretations difficult. So to have a better interpretability we need to scale the variables.
    c.

Normalized scaling(Min-Max)	Standardised scaling
Compresses the data between 0 & 1. Takes care of outliers.	Converts data so that the mean is 0 and standard deviation of 1.
Calculated as (x - x min)/x max – x min)	Calculated as (x – x mean) / sigma

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

- The reason for VIF value being infinite is when features are highly correlated with each other indicating multicollinearity. Formula to calculate is VIF = $1/1-R^2$. When there is perfect multicollinearity R^2 will be 1, that results VIF being infinite.

(3 marks)

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

a. The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.
b. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
c. Linear regression assumes that the residuals (errors) are normally distributed. A Q-Q plot of the residuals helps to assess if this assumption hold true. If the points lie close to the 45-degree line, the residuals are likely normally distributed.

(3 marks)