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?
    a. Season has significant impact on the Demand(Cnt) as the demand is high in 'Fall' and 'Summer' . The max of Winter is almost same as median of Fall and Summer.
    b. The demand seems to increase over the years, 2019 has more demand compared to 2018. Median for 2018 is around 4000 vs 6000 for 2019.
    c. Weather seems to also have impact as there is high demand in 'clear weather' and less demand during 'Snow'

2. Why is it important to use drop_first=True during dummy variable creation?
    a. Reduces an extra column as remaining columns will be able to cover all scenarios.
    b. It reduces correlation created between variables.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

    Registered is highly correlated, however since cnt is a sum of registered and casual, temperature(temp) is the next highly correlated.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    - ✓ The assumptions of Linear Regression are validated by doing residual analysis.
    - ✓ The residuals of each data set is calculated by decreasing the predicted value from the actual value.
    - ✓ A distribution plot is plotted to check if
        a. the errors are following normal distribution
        b. a random pattern of residuals supports a linear model.
        c. both the sum and the mean of the residuals are equal to zero
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Year
    b. Temperature
    c. WeatherSit (Lightsnow)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - Linear Regression algorithm is a supervised Regression algorithm which predicts the output of continuous variables based on one or more dependent variables.
    - For example, prediction of a house price based on different parameters like locality, number of bedrooms, bathrooms, furnishing status, house age etc
    - There are two types of Linear Regression
        - Simple
        - Multiple
    - The equation for Multiple Linear regression is y = β0+ β1X1+ β2X2+…. βnXn.
        - X1, X2, X3,… Xn is independent variable
        - B0 is intercept
        - β1, β2, --- βn is slope
    - The purpose is to find the best Fit Line equation that can predict the dependent variable based on the independent variables.
    - Cost function
        - Mean Squared Error (MSE) cost function is employed, which calculates the average of the squared errors between the predicted values and actual values.
    - Assumptions of Linear Regression
        - Linearity - The independent and dependent variables have a linear relationship with one another.
        - Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
        - Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant.
        - No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity is found by using VIF.
        - Normality: The residuals should be normally distributed.
    - Evaluation Metrics
        - Coefficient of Determination (R-squared) – This metric indicates how much variation the developed model can explain or capture. It is always in the range of 0 to 1.
        - Adjusted R2 - accounts the number of predictors in the model and penalizes the model for including irrelevant predictors that don’t contribute significantly to explain the variance in the dependent variables.

2. Explain the Anscombe’s quartet in detail. (3 marks)
    - Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when graphed.
    - Anscombe’s Quartet exhibits diverse patterns in scatter plots, illustrating the importance of visualizing data for meaningful insights beyond numerical summaries.
    - 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)
    - ✓ The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

      The following tables gives the coefficient, strength and direction

Pearson coefficient	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

- ✓ The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.
- ✓ The Pearson correlation coefficient is a good choice when the following are true:
    - o Both variables are quantitative
    - o The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
    - o The data have no outliers. A scatterplot is one way to check for outliers—look for points that are far away from the others.
    - o The relationship is linear between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - ✓ Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale.
    - ✓ The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.
    - ✓ Scaling types
        - i. Min-Max scaling is a Normalization technique in which values are rescaled so that they end up between 0 and 1.
        - ii. Standardization is where the values are around the mean with a unit standard deviation. This means that the mean of the feature becomes zero, and the distribution has a unit standard deviation. Even if there are outliers in the data, they will not be affected by standardization.

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

- ✓ VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- ✓ A large value of VIF indicates that there is a correlation between the variables
- ✓ VIF formula is 1/(1-Ri)2
- ✓ Since VIF is dependent on R2, R2 can be 1, when there is high correlation between the variables or when there may be problem of overfitting as it is able to explain the variation 100%
- ✓ When R-square is 1 then VIF is 1/(1-1) i.e. 1/0 i.e infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   - ✓ Quantile-Quantile (Q-Q) plot determines if a dataset follows a certain probability distribution or whether two samples of data are from the same population.
   - ✓ Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
   - ✓ A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
   - ✓ Interpretation of Q-Q plot
     - o If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
     - o Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.
   - ✓ Importance of Q-Q plot in Linear Regression.
     - o This plot is used to if the residuals of the model are normally distributed. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then the residuals have a non-normal distribution.
     - o To check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To check this, create a Q-Q plot for the residuals of the model and compare them with the normal distribution.