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?
   Several categorical variables—season, month, year, weekday, working day, and weathersit—significantly influence the dependent variable 'cnt.' These variables exhibit a notable correlation with 'cnt,'. Their impact is visualized through both bar plots and box plots.
2. Why is it important to use drop_first=True during dummy variable creation?
   The purpose of dummy variable creation is to represent 'n' categorical levels as 'n-1' new columns, indicating the existence of each level (0 or 1). Employing drop_first=True ensures alignment with 'n-1' levels, reducing correlation among dummy variables. For instance, with 3 levels, drop_first will exclude the first column.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Among the numerical variables, the 'temp' and 'atemp' exhibit the highest correlation with the target variable 'cnt,' as observed in the pair-plot.
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Validation of Linear Regression models involves assessing assumptions such as Linearity,Normality of error, Homoscedasticity, and Multicollinearity after constructing the model on the training set.
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   In the final model, the three most impactful features in explaining the demand for shared bikes are temperature, year, and season.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Linear regression is a predictive modeling technique revealing the relationship between the dependent (target variable) and independent variables (predictors). It depicts the linear association, showcasing how the dependent variable changes with variations in the independent variable. If there's a single input variable (x), it's termed simple linear regression; if multiple, it's termed multiple linear regression. The model aims to determine optimal values for a0 and a1, forming the best-fit line with minimal error. Techniques like RFE, Mean Squared Error (MSE), or cost functions aid in obtaining these values.
2. Explain the Anscombe’s quartet in detail.
   Anscombe’s Quartet comprises four datasets with nearly identical simple descriptive statistics, yet each exhibits unique characteristics that can deceive a regression model if constructed hastily. Despite sharing statistical observations, these datasets differ significantly in distributions and scatter plot appearances. The quartet serves as a reminder to prioritize graph plotting before analysis and model construction. Each dataset, though statistically similar, presents distinct features, such as linear relationships, non-linearity, outliers, and high-leverage points, challenging the assumption of a one-size-fits-all regression model.
3. What is Pearson’s R?
   The Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   It's a preprocessing step crucial for aligning features that have varying magnitudes, units, and ranges. Without scaling, algorithms might give undue weight to certain high-value features, leading to inaccurate modeling. Normalized scaling employs minimum and maximum feature values, while standardized scaling uses mean and standard deviation for normalization. Normalization scales values between (0,1) or (-1,1), whereas standardization doesn't bound values to a specific range. Normalization is sensitive to outliers, unlike standardization. Normalization is preferred when feature scales differ, while standardization is suitable for a normal distribution.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   VIF (Variance Inflation Factor) elucidates the correlation between an independent variable and the other independent variables. When a VIF value exceeds 10 or even 5, it indicates significant multicollinearity. A VIF approaching infinity is a consequence of perfect correlation between variables, resulting in an R2 value of 1. Resolving this involves eliminating one of the variables causing this perfect multicollinearity
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   A Q-Q plot, a probability plot, compares two probability distributions by plotting their quantiles against each other. It's a graphical tool aiding the assessment of whether a dataset aligns with theoretical distributions like Normal, exponential, or Uniform distributions. Q-Q plots ascertain similarity between distributions, with a more linear plot indicating likeness. In linear regression, where multivariate normality is crucial, Q-Q plots help validate this assumption along with scatter plots. In regression, Q-Q plots confirm if training and test datasets share the same distribution from the population. This plot's advantages include its adaptability to sample sizes and its ability to detect distributional aspects like location shifts, scale changes, symmetry alterations, and outlier presence. Q-Q plots assess similarity in distribution, location, scale, distribution shape, and tail behavior between datasets.