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?

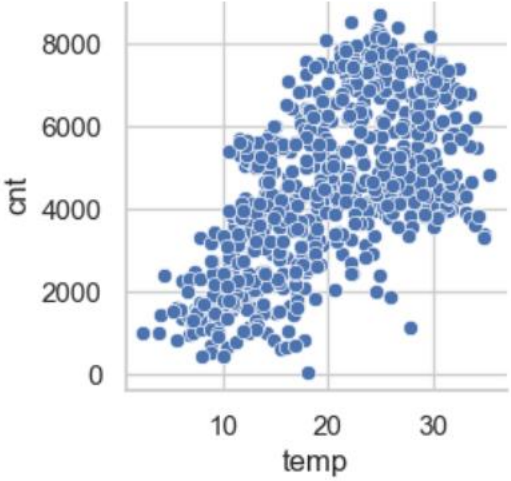
   Ans: From the analysis of the categorical variables from the dataset, we could infer that bikes are rented more in seasons-'fall' followed by 'summer', and more number in 2019 than 2018, more in the months of 'June' to 'September'. Also they were rented more when weathersit was 1: Clear, Few clouds, Partly cloudy, Partly cloudy

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

   Ans: To remove extra column (created during creation of dummy variables) we need to use drop_first=True, to remove the redundancy. Else we get the problem of multicollinearity which results in high VIF.

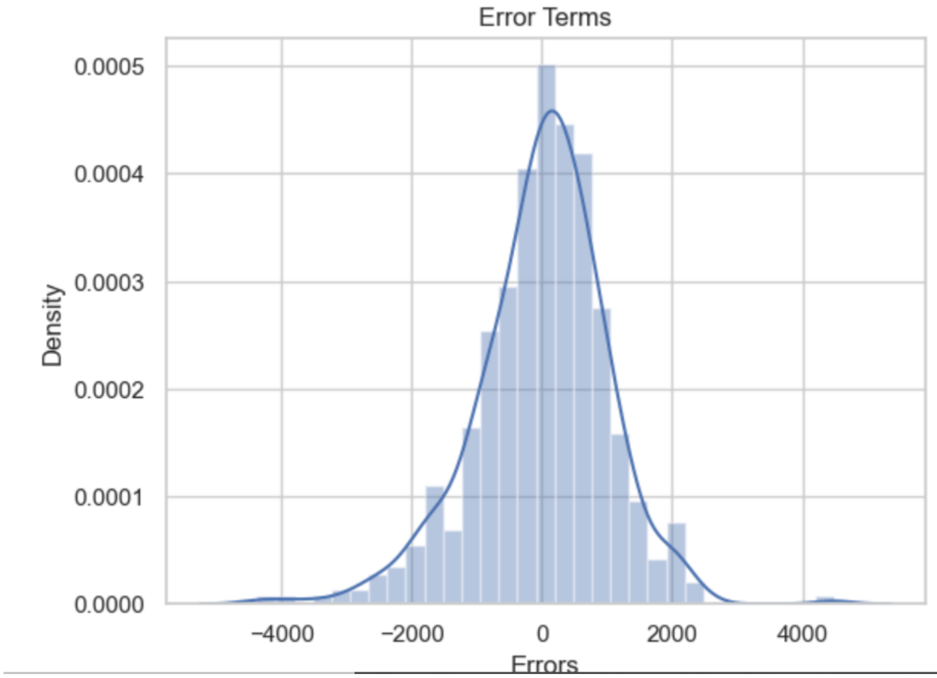
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Ans: Temp variable has highest correlation with the target variable.



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

   Ans: The distribution of residuals should be center around 0 and it should be a normal distribution. We tested this by running a distplot of residuals and it resulted in a normal distribution.



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

   Ans: The variables, yr, weathersit ->Light Snow, Light Rain + Thunderstorm + Scattered clouds, and season->spring are top 3 features contributing significantly towards the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a machine learning algorithm which helps in prediction of dependent variables if one or more independent variables are provided. When there is only one independent variable, it is called simple linear regression. When there are multiple independent variables, it is called multiple linear regression. The main goal is to find the best fitted line that minimizes the difference between the observed values and the predicted values. This difference is measured by the sum of square of residuals.

To find the best line we use Ordinary Least Squares (OLS) method. Assumptions of linear regression:

Linearity: The relationship between dependent and independent variables should be linear.

Independence: The errors should be independent of each other.

Homoscedasticity: The errors should have constant variance at all levels of the independent variables.

Normality: The errors should be normally distributed.

No multicollinearity: In case of multiple regression, the independent variables should not be too highly correlated with each other.

Once we fit the linear regression model, we evaluate its performance metrics such as, R-square, Adjusted R-squre and Mean squared error.

2. Explain the Anscombe’s quartet in detail.
   Ans: Anscombe's Quartet consists of four datasets with nearly identical summary statistics (mean, variance, correlation and linear regression line) but different distrubutions and patterns when graphed. It illustrates the importance of visualizing data before analysis. Each dataset shows different characteristics- A linear relation, non linear relation, a linear relationship with an outlier and a set heavily influenced by a single point. The Quartet emphasizes that data visualization and context are crucial for accurate data interpretation.

3. What is Pearson’s R?
   Ans: Pearson's correlation coefficient, measures the strength and direction of the linear relationship between two variables. Its value ranges from -1 to 1, where 1 indicates a perfect positive linear correlation and -1 indicates a perfect negative linear correlation and 0 indicates no linear correlation. Pearson's R is calculated by taking the covariance of the two variables and diving it by the product of their standard deviations. It's commonly used in statistics to determine how closely two variables are related.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling is a data preprocessing technique used to adjust the range and distribution of features in a dataset. It ensures that each feature contributes equally to the model's performance, particularly for algorithms that compute distances between data points.

   Scaling is performed to ensure equal contribution between all the features in a model preventing features with larger ranges from dominating each other. It also helps algorithms converge faster during optimization. It also standardizes the interpretation of model coefficients. It is used to enhance model performance.

   Normalized scaling: Rescales data to a fixed range usually [0,1] or [-1,1].

   Standardized scaling: Rescales the data to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity. When VIF is infinite, it indicates perfect multicollinearity, meaning one predictor variable is a perfect combination of one or more other predictor variables.It could also that the variable is redundant or is highly linearly dependent with other variable(s).

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

Ans: A Q-Qplot is a graphical tool used to compare the distribution of a dataset to a theoritical distribution, often the normal distribution. In a Q-Qplot, the quantities of the sample data are plotted against the quantities of the theoritical distribution. If the data follows the theoritical distribution, the points will lie approximately along a straight line.

The use and importance of a Q-Q plot in linear regression:

Assessing normality: A Q-Qplot helps check if the residuals are normally distributed or not by plotting the quantities of the residuals against the quantiles of a normal distribution.

Identifying Deviations: It can identify skewness or heavy tails. Points deviating form straight line indicate departures from normality which could affect the validity of statistical tests and confidence intervals.

Detecting Outliers: Outliers can be spotted in q Q-Qplot as points that are far from the diagonal

line.
Model Diagnostics: By using Q-Q plots, you can diagnose and improve linear regression model.