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?

Answer: The categorical variables have indeed effect on the dependent variable 'cnt' (bike rental count). From basic EDA and model evaluation (based on p value and VIF) we have eliminated some categorical values and the remaining variables have effect on the target variable.

The features which have effect on the 'cnt' variable are:

Months(July and September), day(Sunday), Weather(light snow condition, Misty cloud condition), Season(Winter, summer, spring), year and holiday.

Their correlation coefficient and effect with statistics such as P-value is calculated and presented while building the model.

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

Answer: During dummy variable creation, the dataframe will have a dummy variable created for every category. But we need only n-1 dummy variables for the model as the first category can be dropped and considered as the base state. This is why use drop_first=True should be used, the default is 'False'.

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

Answer: 'registered' variable. The highest correlation in the pair plot is for the 'registered' variable. But it is irrelevant since it depends on the 'cnt' variable. The actual feature which is highly correlated is the 'temp' variable.

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

Answer: The assumptions are validated with residual analysis on the training data.

-The error terms should be normally distributed.

-The mean value of the error terms should be 'zero'.

Both are validated after building the model.

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

   Answer: Temperature, Weather (Light snow condition the most) and Year are the top three features.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression algorithm is used to predict continuous target variable from independent feature variables. The linear regression algorithm assumes linear relationship between the target variable and the independent variables. In simple linear regression the model fits a straight line that minimizes the error terms which is usually the sum of square of difference between the predicted value and the actual value (Residuals). When there is more than one feature, the straight line becomes a hyper plane which fits through the training data points and minimizes the cost function or error terms.

Even though there are other fancy models in the industry, linear regression models are still relevant. Because this model provides interpretability.

Simple linear regression: $Y = \beta_0 + \beta_1 X$

Multiple linear regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Where the β_0 is the intercept and the other β s are the coefficient.

The betas are selected by choosing the line that minimizing the squared distance between each Y value and the line of best fit. The betas are choosing such that they minimize this expression:

$\sum_i (y_i - (\beta_0 + \beta_1 X))^2$

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

2. Explain the Anscombe's quartet in detail

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. The numerical variables should be scaled within a common range. Otherwise, the model will try to fit more importance to the variables with higher numerical value. Scaling is bringing the feature range to a common scale. There are two type of scaling, namely, normalized and standard scaling.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

The formula is: (X-Xmin) / (Xmax – Xmin)

The maximum value is 1 and minimum is 0.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

The formula is: (X- σ) / mu

Where σ = standard deviation of the feature values and mu is their mean.

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

   Answer: The VIF value is infinity when there is perfect correlation. That means, if a variable is just another name for the exact same variable, then the VIF is infinity. If a variable is just a scaled version of another feature variable, then the VIF is infinity. A very high VIF indicates that the variable is highly dependent on other feature variables.

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

Answer: The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In linear regression, it is a tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. The q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.