Assignment-based Subjective Questions

Q-1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Below are the effect on the dependent variable (Cnt):

- **season:** Bike renting is high in season summer and fall.
- **yr:** Bike renting is high in year 2019.
- **mnth:** Bike renting is high in month May-Oct.
- **weekday:** Bike renting is high on Sat, Wed, Thu, Fri.
- **weathersit:** Bike renting is high in clear weather.
- **holiday:** Bike renting is high if there is no holiday
- **workingday:** Bike renting is high in working days

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

Answer: drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

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

Answer: The **temp** variable has the highest correlation with the target variable.

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

Answer: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

Answer: The top 3 features contributing significantly towards the demand of the shared bikes are the **temperature**, the **year** and the **Light Snow** variables.

--Next Page --

General Subjective Questions

Q-1: Explain the linear regression algorithm in detail.

Answer: Linear Regression is an ML algorithm used for **supervised learning**. It **helps** in **predicting** a **dependent variable(target)** based on the given **independent variable(s)**. The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables.

There are two types of linear regression- simple linear regression and multiple linear regression.

Simple linear regression is used when a single independent variable is used to predict the value of the target variable.

Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable.

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A **positive linear relationship** is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

Q-2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of **four data sets** that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each **dataset** consists of **eleven points**. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

Q-3: What is Pearson's R?

Answer: Pearson's Correlation **Coefficient** is used to establish a **linear relationship between two quantities**. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between **-1** and **+1**.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

--Next Page --

Q-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is a technique performed in **pre-processing** during building a machine learning model **to standardize the independent feature variables** in the dataset in a fixed range.
The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.
The difference between **normalization** and **standardization** is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Answer: The value of VIF is infinite when there is a **perfect correlation between the two independent variables**. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Answer: The quantile-quantile (Q-Q) plot are used to **plot quantiles** of a **sample distribution** with a **theoretical distribution** to determine if any **dataset** concerned **follows** any distribution such as **normal, uniform or exponential distribution**. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.

A Q–Q plot is used to **compare** the **shapes of distributions**, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.