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:
   There are 2 main categorical variables – season and weathersit.
   ***Season** – My final model suggests that an increase in temperature leads to more bike rentals. Although summer itself is not considered significant in my model but we can infer from the temperature that there will be more rentals in summer as compared to spring, fall and winter.*
   ***Weathersit** – Bike rentals increase in clear weather and drop in rainy, snow or cloudy weather.*
2. Why is it important to use **drop_first=True** during dummy variable creation?
   Answer:
   The number of dummy variables is n-1 where n is the number of possible values of the parent variable. For example, there are 4 seasons – summer, winter, spring & fall. We need only 3 variables to map all seasons –
   - *Spring is 100*
   - *Summer is 010*
   - *Winter is 001*
   - *Fall is 000*

   While calling pd.get_dummies(), it automatically generates as many dummy variables as values of the parent variable. So it is important to use drop_first=True to drop 1 dummy variable.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer:
   Temp has the highest correlation with the target variable
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer:
   I validated the model in 2 ways –
   a. *Calling scikit-learn's r2_score method on the test set of the target variable and the predicted set based on model*
   b. *Drawing a scatter plot on the predicted set vs test set and checking if a linear relationship is evident*
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 towards explaining the demand for rental bikes are –
   a. *Year – Demand grows from 2018 to 2019*
   b. *Temperature – Demand grows with rise in temperature*
   c. *Clear (Weather condition) – Demand grows in clear weather conditions*

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer:
   Linear Regression is a statistical model that maps the relationship between 1 dependent variable and 1 or more independent variables. It is called a linear regression because the model follows an equation like a linear equation.
   A single linear equation looks like below. β0 is a constant term also called intercept.

   $$y=\beta0+\beta1X+\epsilon$$

   - *y: Dependent variable (target).*
   - *XX: Independent variable*
   - *β0: Intercept*
   - *β1: Coefficient (slope)*
   - *ϵ: Error term*

   Linear regression can also be multiple where there more than 1 independent variables. Its equation looks like this –

   $$y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$$

2. Explain the Anscombe's quartet in detail.
   Answer:
   Anscombe's quartet consist of 4 datasets which have identical statistical properties like Mean, variance, correlation, linear regression line and R squared, but when the datasets are plotted in a graph they look very different. The 4 plots are –
   a. *Linear relation*
   b. *Non-linear relation*
   c. *Linear relation with outlier*
   d. *Vertical straight line with 1 outlier*
3. What is Pearson's R?
   Answer:
   Pearson's R or Pearson correlation coefficient is a measure of the strength and direction of relationship between 2 variables.
   - *The values of Pearson's R always fall between -1 and 1.*
   - *Higher absolute values indicate strong linear relation, and lower absolute values indicate weak linear relation.*
   - *Positive values indicate that as 1 variable increases the other one increases too.*
   - *Negative values indicate that when 1 variable increases the other decreases.*
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:
In machine learning scaling is done to bring all features to the same scale irrespective of their actual units or range of values.
Scaling improves model performance. It ensures that any feature doesn't dominate in a model because of larger values or different unit of measurement.

- ***Normalized Scaling*** *– It is also known as Min Max scaling.*
  The formula looks like this - Xnorm = (X−Xmin) / (Xmax−Xmin)
  Transformed values range between 0 and 1
- ***Standardized Scaling*** *– It scales data based on mean and standard deviation. Tarnsformed values have no specific range.*

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   VIF can become infinite when R-squared is 1 since VIF = 1/ (1- R^2).
   R-squared can become 1 when a target variable is perfectly linearly correlated to a combination of several independent variables indicating multicollinearity. It can also happen if you fail to drop 1 dummy variable.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer:
   In linear regression, the Q-Q plot is used to check if the residuals (the differences between observed and predicted values) are normally distributed, which is an important assumption of linear regression models. If residuals are normally distributed, the points on the Q-Q plot will follow the 45-degree reference line.