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:
   With reference to the “Detailed statistical analysis on the above variables” section in the notebook -
   - Season - 32% bike bookings during fall season, 28% during summer, lowest in spring.
   - Month – Highest booking in June to September, least in January, in both years
   - Holiday - 97.6% bikes were rented on a non-holiday
   - Weekday: No effect
   - Working day : Around 70% booking on working days
   - Weather Situation : 69% booking on 'Clear, Few clouds, Partly cloudy, Partly cloudy' days, no booking on a 'Heavy Snow/Rain/Hail/Fog' day

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer: drop_first=True is important as it allows us to avoid creation of an extra column while creating dummy variables. Thus reducing the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Answer: Both the numeric variables ‘temp’ & ‘atemp’ has highest correlation of 0.63 with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer:
   With reference to the following sections in the notebook:
   - Linearity Relationship check - Linearity was well-preserved for both temp and windspeed features.
   - Multicollinearity check - No or negligible Multicollinearity observed.
   - Residual Analysis
     - Normal distribution of error terms, around mean 0
     - Homoscedasticity is well preserved, Constant deviation of points from the zero line, Homoscedasticity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of shared bikes?
   Answer:
   - "temp" (Temperature) with the coefficient of 0.5499
   - Light Snow/Rain (weathersit =3) with the coefficient of -0.2880
   - 'yr' (Year) with a coefficient of 0.2331

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer: Linear Regression algorithm, a supervised learning technique that performs a regression task to predict the target value based on the independent variable(s). Here we train one model to predict the behaviour of the target variable based on some data.

   There are two types of linear regression models:
   a. Simple linear regression – used when the number of independent variable is 1.
      The mathematical equation behind the Linear Regression is:

      $$Y = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

   b. Multiple linear regression – used when the number of independent variables is more than 1 and single variable is not good enough to predict the target.
      The mathematical equation behind the Linear Regression is:

      $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

   In the above equations:
   Y = target variable
   X_1 - X_p = independent variables
   β0 = intercept value of the line
   β1 - βp = slope/coefficient value of X which represents the increase of coefficients times of target variable when there is a unit increase in the independent variable (X)
   ε = random error

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. It comprises four data sets that have nearly identical simple descriptive statistics but with very different distributions and appear very different when graphed.

These four data set plots have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

The four datasets can be described as:

i. Dataset 1: fits the linear regression model pretty well.
ii. Dataset 2: does not fit linear regression model on the data quite well as the data is non-linear.
iii. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
iv. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson's R?

Answer: Pearson's correlation, also known as Pearson's R is a numerical summary indicating the strength of the linear association between the variables and it is denoted by r.

Positive correlation coefficient - when the variables tend to go up and down together

Negative correlation coefficient - when the variables tend to go up and down in opposition with low values of one variable associated with high values of the other.

This correlation coefficient varies between -1 and +1 where:

- r = 1 => data is perfectly linear with a positive slope
- r = -1 => data is perfectly linear with a negative slope
- r = 0 => there is no linear association
- r > 0 < 5 => there is a weak association
- r > 5 < 8 => there is a moderate association
- r > 8 => there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer: Scaling is a technique used to transform the data to fit into a specific scale.

   Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges. If scaling is not done then the algorithm only takes magnitude into account and not units hence results into incorrect model. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   Normalized scaling brings all of the data in the range of 0 and 1.
   Formula to this is –
   $x = x\text{-}min(x) / (max(x) – min(x))$

   Standardized scaling 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.
   Formula to this is –
   $x = x – mean(x) / sd(x)$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   If there is perfect multicollinearity between independent variables then the VIF value becomes infinity. It shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) = 1/0 = infinity.
   In case of perfect multicollinearity between two variables, we need to drop one of the variables from the dataset that is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer:
   Q – Q plot known as Quantile – Quantile plot, is used to check the normality of data (reference line and value distribution). When all the values lie on the reference line then the distribution is normal.

A Q-Q plot helps 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.

Importance of Q-Q Plot in Linear Regression:

i. Two datasets/sample can be of different size.
ii. Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.
iii. It helps assessing if the residual of the model is normally distributed.