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: Analysis done is as mentioned below:
   - Fall season have maximum booking.
   - Booking has increased from 2018 to 2019 in each season.
   - Number of bookings for each month increased from 2018 to 2019.
   - In the mid of the month, there is more booking as compared to start and end month of the year.
   - Clear weather has most booking.
   - Booking increased in each type of weather in 2019 as compared to 2018.
   - Booking seemed to be almost equal either on working day or non-working day. But the count increased from 2018 to 2019.
   - 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

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

   Ans: When creating dummy variables (one-hot encoding) from categorical variables, setting drop_first=True in libraries like pandas or scikit-learn is important for several reasons:
   - Avoiding Multicollinearity
   - Reducing Redundancy
   - Enhancing computational efficiency

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

   Ans: temp variable 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?

   Ans: Validating the assumptions of linear regression is a crucial step to ensure that the model is appropriate for the data and that the results can be interpreted reliably. Here are common assumptions and ways to validate them after building the model on the training set:

   - Error should be normally distributed.
   - There is should be insignificant multicollinearity among variables.
   - Linearity should be visible among variables.
   - There should be no visible pattern in residual values.

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

   Ans: Top 03 features are given below:
   - Light_snowrain
   - Sep
   - Misty

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans: Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The algorithm assumes a linear relationship between the input features and the target variable. In simple linear regression, there is only one predictor variable, while in multiple linear regression, there are multiple predictor variables.

   Two types:
   - Simple Linear Regression
   - Multiple Linear Regression

   Assumptions of Linear Regression:
   Linearity: The relationship between the variables is assumed to be linear.
   Independence: Residuals (errors) should be independent of each other.
   Homoscedasticity: Residuals should have constant variance across all levels of predictor variables.
   Normality of Residuals: Residuals should be approximately normally distributed.
   No Perfect Multicollinearity: Predictor variables should not be perfectly correlated.
   Applications:
   Linear regression is widely used in various fields, including economics, finance, biology, and social sciences, for tasks such as predicting sales, estimating stock prices, and analyzing the impact of variables on an outcome.

   Limitations:
   Linear regression has limitations, such as assuming a linear relationship, sensitivity to outliers, and vulnerability to multicollinearity. It may not perform well in complex, non-linear relationships between variables.

   Despite its limitations, linear regression is a fundamental and interpretable algorithm, providing insights into the relationships between variables in a straightforward manner.

2. Explain the Anscombe’s quartet in detail.

   Ans: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. This dataset was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical exploration and visualization in understanding and interpreting data. The quartet is designed to illustrate the limitations of relying solely on summary statistics (mean, variance, correlation, etc.) without visually inspecting the data.

3. What is Pearson’s R?

   Ans: The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
   The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

   Ans: Scaling is a preprocessing step in machine learning that involves transforming the features of a dataset to a standard range or distribution. The goal of scaling is to ensure that all features contribute equally to the model training process and prevent certain features from dominating due to differences in their scales or units. Scaling is particularly important for algorithms that are sensitive to the scale of input features, such as gradient-based optimization algorithms used in many machine learning models.

   Key Differences:

   - Range:
     Normalized Scaling: Scales values to a range between 0 and 1.
     Standardized Scaling: Scales values with a mean of 0 and a standard deviation of 1.

   - Formula:
     Normalized Scaling: Involves the minimum and maximum values of the variable.
     Standardized Scaling: Involves the mean and standard deviation of the variable.

   - Sensitivity to Outliers:
     Normalized Scaling: Sensitive to outliers as it is influenced by the range of values.
     Standardized Scaling: Less sensitive to outliers due to its reliance on the mean and standard deviation.

   - Interpretability:
     Normalized Scaling: Preserves the relative differences between values.
     Standardized Scaling: Provides a more interpretable scale with a mean of 0 and standard deviation of 1.

- Algorithm Suitability:
  Normalized Scaling: Suitable for algorithms where the scale of features is important (e.g., distance-based algorithms).
  Standardized Scaling: Suitable for algorithms where the relative importance of features is more significant than their absolute values.

The choice between normalized and standardized scaling depends on the characteristics of the data and the requirements of the specific machine learning algorithm being used.

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

   Ans: If there is perfect correlation between two independent variables, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. In case of perfect correlation, R2 equal to 1, which lead to 1/(1-R2) equal to infinity. To solve this we need to drop one of the variables from the dataset which 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.

   Ans: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the observed data against the quantiles expected from a chosen theoretical distribution. The Q-Q plot is particularly useful for visually inspecting the distribution of residuals in the context of linear regression.

   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. Q–Q plots can be used to compare collections of data, or theoretical distributions.

   Importance of Q-Q Plot in Linear Regression:
   - Assumption Checking:
     Q-Q plots are crucial for checking the normality assumption of residuals in linear regression. If the residuals are not normally distributed, it may impact the reliability of statistical inferences and predictions.

   - Detecting Outliers:
     Outliers or heavy tails in the distribution of residuals can be detected visually in a Q-Q plot. Identifying outliers is important for assessing the impact of influential points on the regression model.

   - Model Diagnostics:
     Q-Q plots are part of a set of diagnostic tools used to assess the overall fitness of a linear regression model. Along with other diagnostics like residual plots, leverage plots, and Cook's distance, Q-Q plots help identify potential issues with the model.

   - Guidance for Transformations:
     If the Q-Q plot suggests a departure from normality, it may indicate the need for transformations or other adjustments to improve the model's performance.