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: To check the effect of the categorical variables on the target, bar plots were created which revealed the following:
   - **season** has a significant effect on the target, the demand is highest in the autumn whereas it is lowest in spring.
   - **holiday** also has some significant effect on the target, the demand is significantly less on holidays.
   - **weekday and workingday** do not seem to have any significant effect on the target.
   - **weathersit** has a significant impact on the demand with the lowest demand on the day when it is light precipitation and the highest demand on a clear day.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer: The **drop_first = True** is a parameter used in **pd.get_dummies()** method, which creates the dummy variables for the categorical variables present in the dataset. It is essential to use this argument otherwise it will result in multi-collinearity issues. Let us suppose that we have a categorical variable with 4 different possible values. If we create dummy variables for this feature without using the drop_first argument, we will get 4 different columns. However one of the columns among the four is redundant since we can represent the four levels using the combination of the other three variables. Not using the drop_first will result in high multi-collinearity in the data, as the fourth variable is easily determined using the combination of the three variables.
   Using drop_first = True will ensure that we have not used redundant features and result in a better adjusted R squared value, which penalizes the model as more features are introduced.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Answer: The feature **atemp (feels like)** has the highest correlation with the target variable with a correlation coefficient of 0.631.
   The feature **temp** has a correlation coefficient of 0.627 with the target which is almost similar to that of the atemp with the target, this is largely due to the fact that temp and atemp have almost a perfect positive correlation coefficient with each other (0.99).

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answers: The assumptions of the Linear Regression after building the model on training data were validated by the following methods:
   1. Multi-collinearity: It was addressed by checking the VIF after creating the model and removing the variables iteratively where VIF was greater than 5.

2. Normality of error terms: This was checked by creating a histogram/KDE plot of the residuals, this was also validated by Q-Q plot where the theoretical quantiles and the residual quantiles almost form an overlapping straight line which points to the fact that the residuals are normally distributed.
3. Auto-correlation: It was checked by plotting the error terms against the date that there is no apparent seasonality and/or pattern(s) in the error terms.
4. Homoscedasticity: It was checked by plotting residuals against the target. There is no apparent funnel shape or pattern in the plot.
5. Linearity: There is a strong linear relationship between the temp(atemp) and the target. Linear regression necessitates the presence of a linear relationship of one or more features with the target.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Answers: The features with the most contribution to the target (demand for shared bikes) will be the ones with the highest absolute value of coefficients. These features are:
   - temp: 2.957
   - weathersit_light_preci_thunderstorm: -2.16
   - year: 2.05

General Subjective Questions:

1. **Explain the Linear Regression algorithm in detail.**

   Answer: Linear Regression is one of the most popular Machine Learning algorithms. It is a supervised Machine Learning algorithm that is, the model is trained on known ground truth data. As the name suggests, it is of the class regression among the supervised machine learning algorithms, where the target is a continuous variable. This algorithm assumes a linear relationship of the target with one or more independent variables, or in other words, the target can be explained using a linear combination of the features. The goal of the algorithm is to find a line that best represents this relationship.
   To achieve the same, the algorithm estimates the coefficients of the features and the intercept. The algorithm tweaks these parameters by using one or more techniques. We can choose to minimize the sum of squared errors (the sum of squares of the difference between the actual and predicted value) and choose the model with the lowest MSE/RMSE.
   We can also choose to use an optimization algorithm called the Gradient Descent to arrive at the best possible set of parameters. In Gradient Descent, the model parameters are iteratively adjusted and updated in a direction that minimizes the cost function. Since

the cost function of Linear Regression is always a bowl-shaped function, therefore there are no local minima.
However, this algorithm's predictions are reliable when key assumptions are found to be true. Following are the assumptions of Linear Regression:

- The most basic assumption is that there is a linear relationship between the dependent variable with the independent variables since this algorithm is only able to model the linear relationship by finding the best-fit line.
- This algorithm assumes the variance of the residuals (difference between predicted and actual values) is fairly constant. This assumption is validated when the predicted values are plotted against the residuals. This is also referred to as homoscedasticity.
- It assumes that the residuals are normally distributed and centred around 0.
- This algorithm assumes that the data points are independent of each other.
- It assumes that the independent variables are not highly correlated with each other. High correlation among the features will not allow correct interpretation of the coefficients of the features, or in other words, we may not be able to estimate the individual effects of the independent variables. This assumption is validated by checking the VIF (variance inflation factor) values. A high VIF value indicates higher multi-collinearity. A VIF > 10, is considered bad, but it is also good to investigate where VIF > 5.

The above assumptions are also some of the drawbacks of this algorithm. Linear Regression is heavily influenced by outliers. The algorithm also suffers from underfitting and overfitting. Since the algorithm tries to establish the linear relationship, the model may not be able to find and learn the non-linear patterns in the data, leading to underfitting as the model is too simple. On the other hand, if we use polynomial terms of the features to capture these non-linear relationships, the model can easily overfit the training data.

To address overfitting, we can use regularization techniques such as L1 and L2 regularization, or increase training data size. To address underfitting, we can use polynomial terms of the features to increase model complexity.

2. **Explain the Anscombe's quartet in detail.**
   Answer: Created by a statistician named Anscombe, Anscombe's quartet is a set of four datasets having almost identical statistical properties but showing large differences among each other when these are plotted. This was created to bring to light the importance of visual plotting of the data and that the statistical properties of the data may not fully capture its nature, patterns or outliers.
   With respect to Linear Regression, we can say that we cannot solely rely on the descriptive statistics of the dataset to determine whether the algorithm will be a good fit or not.
   In this Quartet, the first dataset fits the Linear Regression model well, whereas the model could not fit on the second dataset well due to the non-linearity in the data.

The third dataset contains some outliers that the Linear Regression model could not handle and the fourth dataset contains significant outliers which are also not handled by the Linear Regression model. The quartet thus also brings to light the inability of the Linear Regression algorithm to handle the outliers.

3. **What is Pearson's R?**
   Answer: Pearson's R or Pearson correlation coefficient determines the direction and the strength of the linear relationship between the two variables. The value of this coefficient ranges from -1 to +1.
   - If the value of R is positive, it indicates a positive linear relationship, which means as one variable increases, the other one also increases.
   - If the value of R is negative, it indicates a negative linear relationship, which means as one variable increases, the one decreases and vice-versa.
   - A value of R close to either +1 and -1 indicates that the relationship is strong whereas a value close to 0 indicates that the relationship is weak.
   - A spurious correlation can also exist between the two variables, where a value of R is fairly high but the two variables are not even remotely related.

   The value of R can be influenced by outliers and the range and the scale of the data. This coefficient can only measure the linear relationships and therefore if the relationship between the two variables is non-linear, it will not be able to capture it.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Answer: Scaling refers to the preprocessing/transformation/adjusting of the numerical feature values so that they are all on a similar scale.
   Scaling is performed so that we can have a fair comparison of the features that are on different scales. It makes the interpretation of the machine learning model easier as it will lead to comparable coefficients. It also helps machine learning algorithms to converge quickly to the optimal solution.
   Normalized scaling or min-max scaling scales the feature values to the range of 0 to 1. It preserves the proportions/relative positions within the variable and can be affected by outliers.

$$normalized\_new(x) = \frac{old(x) - min(x)}{max(x) - min(x)}$$

   Standardized scaling or standardization transforms the values of the variables in such a way that the transformed values have a mean of 0 and a standard deviation of 1. This method of scaling is useful when the absolute values and distribution of variables are important. In comparison to normalized scaling, it is less affected by outliers.

$$standardized\_new(x) = \frac{old(x) - mean(x)}{std(x)}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Answer: We have observed that the value of VIF(variance inflation factor) can become infinite sometimes. This case arises when we observe a perfect/near-perfect correlation or relationship between two or more independent variables.
   Example: If we have two features/independent variables X1 and X2 in our data in such a way that X1 can be perfectly predicted by a linear combination of X2 or vice-versa, this points to the fact that one of these features is redundant and the model is not gaining any extra meaningful information if we include both of the features.

   When we have a feature that can be perfectly estimated by a linear combination of one or more features, this means that the variance in that variable is perfectly explained, or in other words, the R-squared value will be 1.

   From the formula of VIF: $VIF = \frac{1}{1 - R_i^2}$,

   The value of 1 for R-squared will make the denominator equal to 0 and therefore VIF will become infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical method that helps in comparing the quantiles of the observed data against the quantiles of the expected distribution. More specifically, in linear regression, it is used to check if the residuals follow a normal distribution, which is one of the main assumptions for the algorithm.
   If the residuals follow a normal distribution, the points on the Q-Q plot will fall along a straight line. If these points deviate significantly from the straight line, it suggests the residuals are not normal and therefore violating the assumption of linear regression. This plot, therefore, helps us in taking key decisions about changes and transformations that we need to make in our Linear Regression model so that the assumptions are validated and thus makes sure that the p-values and the inferences made from the model are valid.
   This plot is also used to check for the skewness in the residuals as well, which is observed when the points on the plot bend in a particular direction, thus giving us more idea about the nature of the errors.