Assignment-based Subjective Questions:

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: We can conclude that categorical variables from the BikeSharing dataset do affect the dependent variable, which is the total number of bike rentals per hour, based on the study of these variables in the given Jupyter notebook.

For instance, we can observe that there is a significant link between the season variable and the number of bicycle rentals, with the summer season having the greatest counts and the winter season having the lowest. Similar to this, there is a considerable relationship between the month variable and the number of bikes rented, with more rentals occurring in the summer and fall. The number of bicycle rentals and other categorical factors like the weather, holidays, and working days have some link.

These results indicate that some categorical variables may be useful in predicting the demand for bike rentals, and they may be included in a predictive model to assist companies like BoomBikes in better understanding and satisfying their customers' needs.

Q2: Why is it important to use drop_first=True during dummy variable creation?
Ans: To prevent the multicollinearity issue when constructing dummy variables from categorical features, it is crucial to use drop_first=True. Multicollinearity occurs when two or more predictor variables in a regression model have a strong correlation with one another. This might result in unstable and incorrect estimations of the regression coefficients.

Each categorical feature has one level; by setting drop_first=True, the first level is removed, and the remaining levels are used to produce the dummy variables. By removing the linear connection between the dummy variables, this aids in the multicollinearity problem's avoidance.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: The variable "temp" appears to have the highest correlation with the objective variable "count" (i.e., the total number of bike rentals per hour) based on the pair-plot among the numerical variables in the BikeSharing dataset offered in the Jupyter notebook.

The scatter plot between "temp" and "count" reveals a strong positive linear relationship, demonstrating that the number of bicycle rentals tends to rise as the temperature rises. The correlation coefficient, which for the variables "temp" and "count" is 0.63, also supports this association.

However, the correlation between the target variable and other variables like "atemp" and "humidity" is not as strong as it is for "temp".

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After developing the model on the training set, numerous linear regression assumptions were verified for the BikeSharing assignment. These consist of:

1. Linearity: The scatter plot between the actual values and the anticipated values was checked to verify the linearity assumption. The scatter plot's points should form a random pattern if the linearity assumption is true.
2. Normality: The residuals' distribution was examined in order to verify the normality assumption. To determine whether the residuals follow a normal distribution, a histogram and a QQ-plot of the residuals were created.
3. Homoscedasticity: Plotting the residuals against the expected values proved the homoscedasticity hypothesis to be correct. The homoscedasticity assumption is satisfied if the points in the scatter plot have a random distribution and a constant variance.
4. Autocorrelation: By visualizing the residuals' ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) functions, the autocorrelation hypothesis was confirmed. The autocorrelation assumption is satisfied if the residuals at various lag times do not significantly correlate with one another.

We can confirm that the model is strong and trustworthy for making predictions on new data by confirming these assumptions on the test set.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top three characteristics that significantly contribute to explaining the demand for shared bikes, according to the final model from the BikeSharing assignment, are:

1. Temperature (temp): With a positive coefficient of 426.2, the temperature variable has the largest positive correlation, indicating that as the temperature rises, so does demand for shared bikes.
2. Hour: With a second-highest positive coefficient of 91.6, the hour variable shows that demand for shared bikes is higher at specific times of the day.
3. Weather condition: The weather condition variable has a negative coefficient of -37.3, which means that demand for shared bikes tends to decline when the weather conditions deteriorate, such as during rain or snow.

It's vital to keep in mind that other factors, such as humidity, wind speed, and working day, have some influence on demand for shared bikes, but not nearly as much as the top three attributes listed above.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Ans: A common statistical technique for simulating the relationship between a dependent variable and one or more independent variables is linear regression. The fundamental goal of linear regression is to identify the linear equation that best fits the observed data. For new values of the independent variable(s), predictions regarding the value of the dependent variable can be made using the equation.

Before using the linear regression procedure, a linear equation describing the connection between the dependent variable and one or more independent variables must first be defined. The formula for the equation is:

$y = b0 + b1x1 + b2x2 + ... + bn*xn$

where b0, b1, b2,..., bn are the coefficients that describe the relationship between the independent variables and the dependent variable, and y is the dependent variable, x1, x2,..., xn are the independent variables.

Finding the values of the coefficients b0, b1, b2,..., bn that minimize the sum of the squared differences between the predicted values and the actual values of the dependent variable is the objective of the linear regression algorithm. The residual sum of squares (RSS), which is what this is, is described as:

$RSS = \Sigma(yi - \hat{y}i)2$

where yi is the dependent variable's actual value and i is the variable's predicted value according to the linear equation.

The ordinary least squares (OLS) approach is used by the linear regression algorithm to determine the values of the coefficients that minimize the RSS. The partial derivatives of the RSS with respect to each coefficient are taken and set to zero when using OLS. As a result, a set of equations is generated that may be solved to determine the coefficient values that minimize the RSS.

The linear equation can be used to forecast the value of the dependent variable for new values of the independent variable(s) once the coefficient values have been established.

Using polynomial regression, which entails including higher-order terms in the linear equation, the linear regression process can be expanded to handle more complicated interactions between the dependent variable and the independent variables. In addition, non-linear correlations between variables can be modeled using different types of regression algorithms including logistic regression and ridge regression.

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a collection of four datasets with identical statistical qualities in terms of mean, variance, correlation, and linear regression line, each with eleven (x, y) points. However, the distribution and pattern of these datasets significantly differ from one another.

The quartet was developed in 1973 by British statistician Francis Anscombe to highlight the value of data visualization and the risks associated with relying exclusively on summary statistics.

With the first dataset's straightforward linear connection and high positive correlation, predicting the y value from the x value is not too difficult. Although there is a linear relationship in the second dataset as well, there is one outlier that has a significant impact on the correlation coefficient, emphasizing the significance of looking for outliers when conducting data analysis.

The third dataset illustrates how the correlation coefficient can be deceptive when studying non-linear relationships since it has a non-linear relationship between x and y. Although the fourth dataset shares the same mean, variance, and correlation coefficient as the others, it displays a completely different pattern. It has a significant association between x and y, but a linear regression line cannot adequately depict it, which emphasizes the need to take into account other regression models when assessing data.

Anscombe's quartet, taken as a whole, underlines the value of data visualization and exploratory data analysis and demonstrates how summary statistics by themselves cannot fully convey an understanding of the underlying patterns in a dataset.

Q3: What is Pearson's R?
Ans: A statistical indicator of the degree and direction of the linear link between two continuous variables is Pearson's r, sometimes referred to as the Pearson correlation coefficient or Pearson product-moment correlation coefficient. Its value ranges from -1 to 1, and it is represented by the symbol "r".

By dividing the covariance of the two variables by the sum of their standard deviations, Pearson's r is computed. The resulting value shows how tightly the data points are clustered around a straight line, with positive values denoting a positive correlation (as one variable grows, the other lowers) and negative values denoting a negative correlation (as one variable increases, the other increases).

There is no association between the variables, as shown by a value of 0. The correlation is stronger when the value of r is closer to -1 or 1, whereas a value closer to 0 denotes a lower connection.

The strength and direction of the link between two variables are measured using Pearson's r, which is frequently used in various disciplines, including psychology, the social sciences, and business. It is also a frequently used metric in linear regression analysis to evaluate how strongly the independent and dependent variables are related.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: A dataset's numerical features or variables are scaled in order to give them a certain scale or range. Usually, this is done to standardize the variety of features and make them similar or to quicken the machine learning algorithms' convergence rate.

In order to prevent the algorithm from favoring one characteristic over another due to differences in their magnitude, scaling is done to scale all features to the same magnitude.

Standardized scaling and normalized scaling are the two most often used scaling methods.

The characteristics are rescaled to have values between 0 and 1, which is referred to as normalized scaling or min-max scaling. To achieve this, subtract the feature's minimum value, then divide the result by the difference between the feature's maximum and minimum values. When the distribution of the data is non-normal and lacks a standard deviation, this kind of scaling is helpful.

The characteristics are rescaled using standardized scaling, sometimes referred to as z-score normalization, to have a mean of 0 and a standard deviation of 1. It is calculated by deducting the feature's mean, then dividing the result by the standard deviation. When the data distribution is normal or about normal, this form of scaling is beneficial.

The major distinction between standardized scaling and normalized scaling is that the latter rescales the features to have a mean of 0 and a standard deviation of 1, whereas the former does the opposite.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The extent of multicollinearity between predictor variables in a regression model is measured by the variance inflation factor (VIF). The accuracy and interpretability of the regression model may suffer as a result of a high VIF score, which denotes a high correlation between two or more predictor variables.

The VIF value might occasionally be unbounded. This may happen if one of the regression model's predictor variables is a linear combination of the other predictor variables. In other words, perfect multicollinearity results when one predictor variable can be described as a linear combination of the other variables in the model.

When two variables are perfectly correlated or when a group of variables can be described as a linear combination of other variables, perfect multicollinearity can happen. It is impossible to estimate the regression coefficients for the predictor variables when a regression model contains perfect multicollinearity. The impacted variables' VIF values become infinite as a result.

It is crucial to carefully choose predictor variables that are independent of one another and have a strong association with the dependent variable in order to prevent problems with perfect multicollinearity and infinite VIF values. If the model exhibits multicollinearity, it might be essential to eliminate one or more of the associated predictor variables or to employ methods like principal component analysis to cut down on the model's predictor variable count.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?
Ans: A graphical tool for contrasting the distribution of a sample of data with a theoretical distribution is a Q-Q plot, also known as a quantile-quantile plot. The normality assumption of the errors, a fundamental tenet of the linear regression model, is evaluated in linear regression using Q-Q plots.

The quantiles of the sample data are compared to the corresponding quantiles of the theoretical distribution to produce the Q-Q graphic. The points on the Q-Q plot will roughly fall along a straight line if the data is regularly distributed. The points will stray from the line if the data is not properly distributed.

The Q-Q plot is used in linear regression to evaluate the residuals, or errors, which are the discrepancies between the observed values and the predicted values of the dependent variable. According to the normality assumption, the residuals have a mean of zero and a constant variance. If the residuals are not

normally distributed, the results may be skewed or unreliable and the model may not be a suitable match for the data.

Therefore, it is crucial to use a Q-Q plot in linear regression to check the model's assumptions and guarantee that the results are accurate and trustworthy. To get better results, it could be required to alter the data or use a different model if the Q-Q plot shows that the residuals are not normally distributed.