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?

In general, if a categorical variable is found to have a significant effect on the dependent variable, it suggests that the category of that variable is a strong predictor of the value of the dependent variable. For example, if the dependent variable is "success" and the categorical variable is "education level", then a significant effect of education level on success would suggest that the level of education is a strong predictor of success. Additionally, the direction of the effect (positive or negative) can also provide insight into the relationship between the categorical variable and the dependent variable.

The dependent variable is likely the "cnt" which is the count of total rental bikes.

From the categorical variables, it is possible to infer several things about their effect on the dependent variable:

- The season variable: It could be inferred that the number of bike rentals may vary by season, with higher rentals during certain seasons and lower rentals during others.

- The year variable: It could be inferred that the number of bike rentals may vary by year, with higher rentals in one year and lower rentals in another year.

- The month variable: It could be inferred that the number of bike rentals may vary by month, with higher rentals in some months and lower rentals in others.

- The holiday variable: It could be inferred that the number of bike rentals may vary depending on whether it is a holiday or not, with higher rentals on holidays and lower rentals on non-holidays.

- The weekday variable: It could be inferred that the number of bike rentals may vary by day of the week, with higher rentals on certain days of the week and lower rentals on others.

- The workingday variable: It could be inferred that the number of bike rentals may vary by whether it is a working day or not, with higher rentals on working days and lower rentals on non-working days.

- The weathersit variable: It could be inferred that the number of bike rentals may vary depending on the weather conditions, with higher rentals during good weather conditions and lower rentals during bad weather conditions.

Analysis:

1. **<u>season vs cnt:</u>**
    a. Fall season showing a high number of bike shared.
    b. During spring the bike sharing got affected and it reduced.

So its recommended to expand the business during fall season and not during spring season.

2. **<u>yr vs cnt:</u>**
    a. Compared to year 2018, the further year 2019 had more count.

By the similar way all the categorical columns dependent with cnt showing some amazing connection. We can understand it visualizing that in ipynb notebook.

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

When creating dummy variables from a categorical variable, it is common practice to use the "drop_first" parameter in order to avoid the problem of multicollinearity. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. This can lead to unstable and unreliable estimates of the regression coefficients and can make it difficult to interpret the results.

By setting the "drop_first" parameter to "True", one of the columns of the dummy variables is dropped, which effectively removes one of the correlated variables from the model. This helps to eliminate the problem of multicollinearity and ensures that the regression coefficients are stable and interpretable.

It's worth noting that the variable dropped is typically the reference category, and the choice of the reference category can affect the results of the analysis.

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

To determine which numerical variable has the highest correlation with the target variable, we would need to calculate the correlation coefficient between each numerical variable and the target variable. The correlation coefficient is a value between -1 and 1 that indicates the strength and direction of the relationship between two variables. A coefficient close to 1 indicates a strong positive correlation, a coefficient close to -1 indicates a strong negative correlation, and a coefficient close to 0 indicates no correlation.

There is linear relationship between temp and atemp and the correlation coefficient is 0.99 which is nearly 1. Both of the parameters cannot be used in the model due to multicollinearity. We will decide which parameters to keep based on VIF and p-value with respect to other variables.

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

After building a Linear Regression model on the training set for the bike sharing assignment, there are several assumptions that need to be validated in order to ensure the model is accurate and reliable.

These assumptions include:

- **Linearity:** The relationship between the independent and dependent variables should be linear. This can be checked by plotting the residuals (prediction errors) against the predicted values. If the residuals are randomly scattered around zero, it indicates that the linearity assumption is met.

- **Independence of errors:** The errors should be independent of each other, meaning that the error in one observation should not affect the error in another observation. This can be checked by plotting the residuals against the time series or by using the Durbin-Watson test.

- **Homoscedasticity**: The variance of the errors should be constant across all levels of the independent variables. This can be checked by plotting the residuals against the predicted values. If the residuals are randomly scattered around zero and the variance is constant, then the homoscedasticity assumption is met.

- **Normality of errors:** The errors should be normally distributed. This can be checked by plotting the residuals against a normal distribution and by using the Anderson-Darling test.

- **No multicollinearity:** The independent variables should not be highly correlated with each other. This can be checked by calculating the correlation matrix and the Variance Inflation Factor (VIF).

- **No Autocorrelation:** The residuals should not be autocorrelated. Autocorrelation occurs when a variable is correlated with itself across time, This can be checked by plotting the residuals against the time series or by using the Durbin-Watson test.

But in our case I mainly considered Variance Inflation Factor (VIF) and P - Value. The value showing high p value and VIF value got removed and also the variables showing high linearity with the other variables except the dependent feature got removed.

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

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

- temp: The temperature in Celsius, it has a high VIF value of 4.65, indicating that it is an important feature in the model.

- windspeed: The wind speed, it has a high VIF value of 4.65, indicating that it is an important feature in the model.

- yr_2018: The year 2018, it has a VIF value of 1.91, indicating that it is a less important feature than the other two but it still contribute in explaining the demand.

Note that VIF values greater than 5 or 10 are generally considered as indication of multicollinearity, but that can vary depending on the context and the sample size. Also, the above assumption is based on the assumption that the VIF values are calculated correctly and the model is built using the correct statistical procedure.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a statistical method that is used to model the **relationship between a dependent variable (also known as the response variable or outcome variable) and one or more independent variables (also known as explanatory variables or predictors).** The goal of linear regression is to find the best-fitting straight line through the data points.

The algorithm can be summarized in the following steps:

The first step is to define the model. The linear regression model is defined as:

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

where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients (also known as the parameters or weights) that need to be estimated. The next step is to collect data and organize it in a form suitable for the model. The data should include the values of the independent variables and the dependent variable.

The next step is to estimate the coefficients of the model. This is typically done using the method of least squares, which finds the values of the coefficients that minimize the sum of the squared residuals (the differences between the predicted values and the actual values).After estimating the coefficients, the model can be used to make predictions. Given a set of input values for the independent variables, the model can be used to predict the corresponding value of the dependent variable.

Finally, the model should be evaluated to ensure that it is a good fit for the data. This can be done by calculating various statistical measures, such as the coefficient of determination (R-squared), the root mean squared error, and the mean absolute error. It's worth noting that Linear Regression assumes that the relationship between the independent and dependent variables is linear and that the errors are normally distributed and have constant variance. If these assumptions are not met, the model's predictions may be inaccurate. In addition, Linear Regression also assumes that the data is free of outliers, multicollinearity, and autocorrelation. If these assumptions are not met, it might lead to unstable and unreliable estimates of the coefficients and difficulty in interpreting the results.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a set of four datasets that were created by Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. Each dataset consists of 11 x-y pairs, and they all have the same mean, variance, correlation, and linear regression line. However, the datasets look very different when plotted, and they have very different properties.

The four datasets are:

- The first dataset has a linear relationship between x and y, with no outliers.
- The second dataset has a linear relationship between x and y, with one outlier.
- The third dataset has no linear relationship between x and y, and it is distributed in a curved fashion.
- The fourth dataset has no linear relationship between x and y, and it is distributed in a curved fashion with an outlier.

Anscombe's quartet illustrates the importance of visualizing data before analyzing it. Even though the four datasets have similar statistical properties, they have very different patterns when plotted. This highlights the importance of visual inspection of the data and not solely relying on summary statistics like mean, variance, and correlation.

It also illustrates the importance of data cleaning and preprocessing, as outliers can have a significant impact on the results of an analysis, and also the importance of choosing the right statistical model, as the linear regression model is not suitable for datasets that have a non-linear relationship between the variables.

Anscombe's quartet is a reminder that a good data visualization can reveal patterns, outliers, and other important features of the data that might not be obvious from summary statistics. It is a powerful tool that can help analysts make better decisions and understand the data better.

3. **What is Pearson’s R?**

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear correlation between two variables. It is a value between -1 and 1 that indicates the strength and direction of the relationship between the variables. A coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases. A coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases. A coefficient of 0 indicates no correlation between the variables.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula is:

R = cov(x, y) / (std(x) * std(y))

Where x and y are the two variables being considered, cov(x,y) is the covariance between x and y, std(x) is the standard deviation of x, and std(y) is the standard deviation of y.

Pearson's R is a widely used measure of correlation, but it assumes that the relationship between the variables is linear and that the data is normally distributed. If these assumptions are not met, Pearson's R may not be appropriate, and other measures of correlation such as Spearman's Rho or Kendall's Tau should be used instead.

It's worth noting that correlation does not imply causation, a high correlation between two variables just means that they are associated but not necessarily one causes the other, there could be other confounding variables that are not in the dataset that are causing the correlation.

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

Scaling is the process of transforming variables to a common scale, usually to a scale between 0 and 1. The main reason for scaling is to handle features that have different units or scales and make sure they are on the same scale before applying any algorithm.

There are different types of scaling methods, but two of the most commonly used are:

1. Normalized scaling: This method scales the data to a fixed range, usually between 0 and 1. The formula for normalization is:

   X_norm = (X - Xmin) / (Xmax - Xmin)

   Where X is the original variable, Xmin is the minimum value of the variable, and Xmax is the maximum value of the variable.

2. Standardized scaling: This method scales the data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

   X_std = (X - mean(X)) / std(X)

Where X is the original variable, mean(X) is the mean of the variable, and std(X) is the standard deviation of the variable.

- Normalization is useful when the data needs to be transformed to a specific range and is useful when the algorithm is sensitive to the scale of the input variables, for example, Neural Networks.

- Standardization is useful when the data needs to be transformed to a standard normal distribution and is useful when the algorithm assumes that the data is normally distributed, for example, linear regression.

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

The value of the Variance Inflation Factor (VIF) can become infinite in some cases because of the presence of perfect multicollinearity. Multicollinearity occurs when two or more independent variables in a multiple linear regression model are highly correlated with each other. This can lead to unstable and unreliable estimates of the model's coefficients and can make it difficult to interpret the results.

Perfect multicollinearity occurs when the correlation between two or more independent variables is equal to 1 or -1. In this case, the variance of the estimated coefficients becomes infinite, and the inverse of the variance (VIF) becomes zero, which is mathematically equivalent to infinity.

This happens because when two or more independent variables are perfectly correlated, one of them can be predicted exactly by a linear combination of the others. Therefore, it is impossible to estimate their individual coefficients, and the linear regression model becomes ill-posed.

To avoid this problem, it is important to identify the presence of multicollinearity in the model and to remove the correlated variables, by using techniques such as principal component analysis (PCA) or partial least squares (PLS) or by using regularization techniques like Ridge and Lasso.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, or quantile-quantile plot, is a graphical tool that is used to assess whether a set of data follows a particular probability distribution. It plots the observed quantiles of the data

against the theoretical quantiles of a specified distribution. If the data follows the specified distribution, the points on the Q-Q plot will fall approximately on a straight line.

In linear regression, Q-Q plots are used to assess the assumption of normality of the errors, which means that the residuals (the difference between the predicted values and the actual values) should be normally distributed with mean 0 and constant variance. A Q-Q plot of the residuals can be used to check whether this assumption is met.

If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately on a straight line. If the residuals deviate significantly from normality, the points on the Q-Q plot will deviate from the straight line. The deviation can take the form of a curved pattern or outliers, which indicates that the residuals are not normally distributed.

The importance of Q-Q plot in linear regression is that it helps to check if the residuals are normal distributed. If the residuals are not normal distributed, it might mean that the linear regression model is not appropriate for the data, and other models or transformations should be considered.

It's worth noting that Q-Q plot can also be used to check the normality of other variables in the dataset and can be used for other types of models, not only linear regression.
It's also important to remember that normality is not a necessary assumption for linear regression to work, but it's a desirable assumption as it makes the model more interpretable and easier to make inferences from.