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]

The analysis of categorical variables and their relationship with the dependent variable ('cnt', total number of bike rentals) revealed several insights, indicating how different factors may influence bike rental demand:

Season: The total number of bike rentals varies significantly across seasons. This suggests that the season has a substantial effect on bike rental demand. Typically, bike rentals are higher in seasons with milder weather (such as summer and fall) compared to seasons with harsher weather conditions (like winter).
Year ('yr'): The analysis indicated a noticeable increase in bike rentals from 2018 to 2019. This trend suggests that the demand for bike rentals is growing over time. The 'yr' variable, therefore, is significant in predicting the demand for shared bikes, capturing the year-on-year growth in bike-sharing usage.
Month ('mnth'): Monthly trends in bike rentals show clear patterns, with peaks during warmer months, likely corresponding to favorable weather conditions and possibly more leisure activities or commuting by bike. This indicates that the month can be a significant predictor of bike rental demand.
Weather Situation ('weathersit'): The total number of bike rentals also varies with different weather situations..
Weekday/Working Day: While not explicitly highlighted in the analysis, these variables could also influence bike rental demand.

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

[ANS]

Using `drop_first=True` is a best practice for dummy variable creation because it helps avoid multicollinearity, reduces redundancy without losing information, enhances computational efficiency, and improves model interpretability.

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

[ANS]

From the pair plot among the numerical variables in relation to the target variable ('cnt'), it appears that:

- Numerical variables like 'temp', 'atemp', 'hum', and 'windspeed', 'temp' and 'atemp' (the actual and "feels like" temperatures) show a notable positive correlation with 'cnt'. This indicates that higher temperatures, which are generally associated with more comfortable weather, tend to correlate with increased bike rentals.

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

[ANS]

To validate the assumption we perform residual analysis of the error terms by plotting histogram plot using seaborn library and found that the distribution is normally distributed

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

[ANS]

1. `yr` : indicates a significant year-on-year increase in bike rental demand, reflecting possibly the growing popularity of the bike-sharing service.

2. `temp` : Higher temperatures are positively associated with increased bike rental demand, likely due to more favorable weather conditions for biking.
3. `atemp` (Feeling temperature in Celsius): Similar to `temp`, the 'feels like' temperature also positively influences bike rental demand, indicating that people's perception of comfortable weather conditions plays a significant role in their decision to rent bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

[ANS]

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form, simple linear regression, involves a single independent variable and fits a line (y = mx + b) to the data, where `y` is the dependent variable, `x` is the independent variable, `m` is the slope of the line, and `b` is the y-intercept.

In multiple linear regression, where there are two or more independent variables, the equation expands to y = b0 + b1x1 + b2x2 + ... + bnxn, with `b0` being the intercept and `b1, b2, ..., bn` being the coefficients of each independent variable `x1, x2, ...,` `xn`. These coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.

The goal of linear regression is to find the best-fitting line through the data points that minimizes the sum of the squared differences between the observed and predicted values (residuals). This is typically achieved using the least squares method, which calculates the best-fit line by minimizing the sum of the squares of the residuals.

Linear regression is widely used for prediction and forecasting, where its ease of use and interpretation makes it a popular choice across various fields, from economics and finance to the natural sciences.

2. Explain the Anscombe's quartet in detail.

[ANS]
Anscombe's quartet comprises four distinct datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven points (x, y) and was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance and limitations of descriptive statistics, and to encourage the use of visualizations when analyzing data.

Key Points of Anscombe's Quartet:

Statistical Similarities:

- All four datasets have nearly the same mean and standard deviation for both the x and y variables.
- They all have nearly the same correlation between x and y.
- Linear regression lines fitted to the datasets result in nearly identical equations.

Graphical Dissimilarities:

- When plotted, each dataset shows a drastically different distribution and relationship between x and y.
  - Dataset I follows a linear relationship with minimal variance around the regression line, fitting the assumptions of linear regression well.
  - Dataset II reveals a clear non-linear relationship (curvilinear), where a linear model is inappropriate despite the linear correlation.
  - Dataset III shows a linear relationship with one clear outlier affecting both the slope of the regression line and the correlation.

- Dataset IV consists of x-values that are almost the same for all points except one outlier, showing how a single outlier can significantly affect the regression model.

Significance:

Anscombe's quartet is a powerful illustration of why it is crucial not only to rely on statistical properties when analyzing data. It underscores the importance of visual data exploration to detect underlying patterns, outliers, or anomalies that might not be evident from summary statistics alone. The quartet teaches that similar statistical numbers can lead to misleading conclusions if the context of the data and graphical analyses are ignored. It's a foundational lesson in data analysis, emphasizing the need for a comprehensive approach that includes both statistical and graphical examination.

3. What is Pearson’s R?

[ANS]
Pearson's R, also known as Pearson's correlation coefficient or simply Pearson's, is a statistic that measures the linear correlation between two variables, X and Y. Its value ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship, and
- -0 indicates no linear relationship between the variables.

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

Pearson’s correlation coefficient is most appropriate for measurements taken from an interval scale, which is where the distance between any two adjacent units of measurement (or 'ticks') is the same anywhere along the scale. Pearson's R is widely used in the sciences and finance as a measure of the strength and direction of a linear relationship between two continuous variables. It is a key tool in regression analysis, helping to understand how closely two variables are related and whether there is a potential predictive relationship between them. However, it's important to note that Pearson's R only captures linear relationships; it may not accurately reflect the strength of non-linear relationships.

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

[ANS]
Scaling is a technique used in data preprocessing to adjust the range of variable values. The process involves transforming the scales of features to a level playing field, making it easier for algorithms to converge more quickly and perform better. Scaling is particularly important in algorithms that calculate distances between data points or when the features vary widely in magnitudes, units, or range.

Why is Scaling Performed?

Algorithm Efficiency: Many machine learning algorithms perform better or converge faster when features are on a similar scale, particularly those that use gradient descent as an optimization technique (e.g., linear regression, logistic regression) or that rely on measuring distances between points (e.g., k-nearest neighbors, k-means clustering).
Fair Comparison: Scaling ensures that features contribute equally to the model performance and outcome, preventing variables with larger ranges from dominating those with smaller ranges.
Improved Accuracy: Normalizing or standardizing data can improve the accuracy of predictive models by giving equal weight to all features, thus allowing the model to learn better patterns from the data.

Difference between Normalized Scaling and Standardized Scaling

Normalized Scaling (Min-Max Scaling):

- Normalization adjusts the data so that the scale ranges from 0 to 1.
- It is useful when you need a bounded range for your data, but it does not handle outliers well.

Standardized Scaling (Z-score Normalization):

- Standardization transforms the data to have a mean of 0 and a standard deviation of 1, converting the feature values to z-scores.
- It is less sensitive to outliers than normalization and is suitable when you want to preserve the distribution of your data, making it ideal for algorithms that assume normality.

Choosing between normalization and standardization depends on the specific needs of your data and the model you are using. Standardization is generally preferred for most machine learning scenarios, especially when the assumption of normality underlies the model's algorithm.

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

[ANS]

A VIF value can become infinite due to 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 to compare two probability distributions by plotting their quantiles against each other. If both distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line y = x. When used in the context of linear regression, a Q-Q plot typically compares the quantiles of the residuals (the differences between observed and predicted values) to the quantiles of a standard normal distribution.