Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on the analysis of the categorical variables and the provided conclusions, we can infer the following effects on the dependent variable (bike rental count):

- **Season:** The bike rental count seems to be relatively consistent across different seasons. There isn't a significant difference in the distribution of data across the seasons, suggesting that seasonality might not be a strong predictor of the bike rental count.
- **Year:** The analysis reveals that 2019 witnessed a better trend in terms of bike rental compared to 2018. This suggests that the year, specifically the year 2019, had a positive impact on bike rentals, possibly due to factors like increased awareness, marketing campaigns, or improved bike infrastructure.
- **Month:** The data provides insights for all months, which can help to identify seasonal variations within each year. A deeper analysis of month-wise variations might unveil specific months with higher or lower bike usage, possibly connected to holidays or weather patterns.
- **Holiday:** The analysis revealed 21 holiday days. Studying the influence of holidays on bike rentals can help uncover whether they lead to increased or decreased rental counts, offering valuable insights for business planning.
- **Weekday:** Analyzing weekday trends can help identify whether weekdays or weekends have higher bike usage, potentially revealing different user profiles (e.g., commuters during weekdays and leisure users during weekends).
- **Working day:** Similar to weekdays, studying working days can provide a deeper understanding of user behavior and their dependency on working days versus non-working days.
- **Weather:** The presence of good weather (clear, few clouds) positively influenced bike rentals. This implies that favorable weather conditions are a significant factor in user decisions to rent bikes. Poor weather conditions, on the other hand, might be associated with reduced bike usage.

In Summary: The categorical variables like year, holiday, weekday, working day, and weather condition contribute to a certain extent to the dependent variable (bike rental count). Especially weather condition and year show positive influence on bike rental count.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

It helps in avoiding multicollinearity issues. When creating dummy variables for categorical features, we often include one less dummy variable than the total number of categories. This is done to prevent multicollinearity, where the independent variables are highly correlated. For instance, if we have a categorical variable 'Color' with values 'Red', 'Green', 'Blue', we can create two dummy variables: 'Red' and 'Green'. If a row has 'Red' = 1 and 'Green' = 0, it implicitly indicates that the color is 'Blue'. Using drop_first=True ensures that we drop the first dummy variable and avoid this redundancy, thus reducing multicollinearity. This makes the model more robust and easier to interpret, as it reduces the chances of unstable coefficient estimates and helps in proper model estimation.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?   (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Looking at the pair-plot among the numerical variables, 'temp' has the highest correlation with the target variable 'cnt'. This is visible in the scatter matrix plot, where 'temp' and 'cnt' show a strong positive linear relationship.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

To validate the assumptions of linear regression after building the model on the training set, I performed the following steps:

1. **Residual Analysis:**

- I plotted a histogram of the residuals (the difference between the actual and predicted values) to check for normality. A normally distributed error term is a key assumption of linear regression.

- I also examined the scatter plot of the predicted values against the residuals to look for heteroscedasticity (non-constant variance of errors). Constant variance of errors is another key assumption.

2. **Multicollinearity Check (VIF):**

- I calculated the Variance Inflation Factor (VIF) for the independent variables to detect multicollinearity (high correlation between predictor variables). High VIF values indicate that the variables are highly correlated, which can lead to unstable coefficient estimates.

3. **Statistical Significance of Coefficients:**

- I examined the p-values for the coefficients of the independent variables in the model summary. Low p-values (typically less than 0.05) indicate that the corresponding variable is statistically significant and contributes to explaining the variation in the dependent variable.

By examining these aspects, I was able to assess whether the assumptions of linear regression were met and determine if the model was appropriate for the dataset.

The model was deemed suitable for prediction if the residuals were normally distributed, the variance of errors was constant, the variables were not highly correlated, and the coefficients were statistically significant.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on the model's coefficients, the top 3 features contributing to the demand of shared bikes are:

3. atemp (feeling temperature)
4. yr_0 (whether year is 2018)

5. windspeed

These features have a relatively larger magnitude in their coefficients, indicating a more significant influence on predicting the demand of shared bikes.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression Algorithm Explanation

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

It assumes a linear relationship between the variables, meaning that a change in the independent variable(s) will result in a proportional change in the dependent variable.

How it Works:

1. Data Representation: The algorithm starts with a dataset containing observations of the dependent and independent variables.

2. Model Definition: It defines a linear equation that represents the relationship between the variables. The equation has coefficients (weights) for each independent variable, and an intercept term.

3. Cost Function: It calculates the difference between the actual dependent variable values and the predicted values from the linear equation. This difference is called the "cost function" or "loss function".

4. Optimization: The algorithm's goal is to minimize this cost function by finding the best possible values for the coefficients and the intercept term.

5. Gradient Descent: Often, gradient descent is used to optimize the model. Gradient descent is an iterative process that adjusts the coefficients slightly in the direction that reduces the cost function.

6. Model Evaluation: Once the algorithm has converged (i.e., the cost function is minimized), the model is considered trained.

7. Prediction: The trained model can then be used to predict the dependent variable value for new or unseen data points.

Key Concepts:

- Dependent Variable: The variable you are trying to predict (e.g., house price).

- Independent Variable: The variable(s) used to predict the dependent variable (e.g., house size, number of bedrooms).

- Coefficients: Weights assigned to each independent variable, indicating their influence on the dependent variable.

- Intercept: A constant term added to the equation, representing the baseline value of the dependent variable when all independent variables are zero.

- Cost Function: Measures the difference between actual and predicted values.

- Gradient Descent: An iterative method used to find the optimal coefficients.

- R-squared: A statistical measure that indicates how well the model fits the data.

Applications:

Linear regression is widely used in various fields, including:

- Finance: Predicting stock prices.

- Marketing: Analyzing the effectiveness of advertising campaigns.

- Healthcare: Predicting patient outcomes.

- Engineering: Modeling physical phenomena.

Advantages:

- Simple and easy to implement.

- Easy to interpret the results.

- Relatively fast to train.

- Can be used for both prediction and inference.

Limitations:

- Assumes a linear relationship between variables, which might not always be the case.

- Sensitive to outliers in the data.

- Can be affected by multicollinearity (high correlation between independent variables).

In summary, linear regression is a powerful and widely used algorithm for modeling linear relationships between variables. It is a valuable tool for making predictions and gaining insights into the underlying relationships within data.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as the mean, variance, correlation, and linear regression line. However, when these datasets are plotted on a graph, they reveal significantly different patterns and relationships between the variables. This highlights the importance of visualizing data in addition to relying solely on summary statistics.

Key aspects of Anscombe's Quartet:

- Identical Descriptive Statistics:
  - The four datasets share very similar means, variances, correlation coefficients, and linear regression lines.
- Distinct Visualizations:
  - When plotted, the datasets show diverse relationships between the variables. Some show a linear relationship, while others showcase non-linear patterns or outliers.
- Implications for Data Analysis:
  - It warns against making inferences based on only summary statistics. Graphical representation provides a richer understanding of data distribution and patterns.
- Importance of Visualization:
  - Visualization is a crucial part of data analysis to identify potential problems in the data, outliers, and the appropriateness of the chosen model.
- Robustness and Outliers:
  - Anscombe's Quartet demonstrates that some statistical measures, like correlation, can be sensitive to outliers or non-linearity.

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables.

It indicates the strength and direction of the linear association between the variables.

A value of +1 represents a perfect positive linear correlation, -1 represents a perfect negative linear correlation, and 0 represents no linear correlation.

In the context of the provided code, Pearson's R is used in the pairwise_corr function to calculate the correlation matrix for a given set of columns in the DataFrame.

This correlation matrix helps us identify the linear relationships between various features, such as temperature, humidity, and wind speed, as they relate to bike usage counts.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique used to transform the features of a dataset into a specific range or distribution. It aims to address the issue of features having vastly different scales, which can negatively impact the performance of many machine learning algorithms.

Scaling is performed for several reasons:

- Improved algorithm performance: Many algorithms, such as k-nearest neighbors and support vector machines, are sensitive to the scale of features. Features with larger ranges can dominate the learning process, leading to biased results. Scaling ensures that all features contribute equally to the model.
- Faster convergence: Scaling can accelerate the training process of some algorithms, especially those based on gradient descent. When features have similar scales, the algorithm can reach the optimal solution more efficiently.
- Enhanced interpretation: Scaling can improve the interpretability of the results. For example, when features are scaled to a similar range, the magnitude of the coefficients in a linear regression model provides a clearer indication of the relative importance of each feature.

- Preventing numerical instability: Scaling can prevent numerical instability in algorithms that involve calculations based on feature values. For example, algorithms that use Euclidean distance might be affected by features with very large values, leading to incorrect results.

Difference between normalized scaling and standardized scaling:

Normalized scaling (Min-Max Scaling): This technique transforms the features by linearly scaling them to a fixed range, typically between 0 and 1. It works by subtracting the minimum value of the feature from each value and then dividing by the range (maximum value minus minimum value). * Formula: X_scaled = (X - X_min) / (X_max - X_min) * Effect: Preserves the original distribution of the feature and is suitable for algorithms that are sensitive to the range of features, such as neural networks.

Standardized scaling (Z-score normalization): This technique transforms the features by subtracting the mean and dividing by the standard deviation. This results in a distribution with a mean of 0 and a standard deviation of 1. * Formula: X_scaled = (X - X_mean) / X_std * Effect: Makes the data follow a standard normal distribution and is often preferred when the distribution of the features is not known or assumed to be normal. Suitable for algorithms that are not sensitive to the range of features, such as linear regression.

In Summary:

- Scaling is a crucial preprocessing step in Machine Learning to ensure that all features contribute equally to the model, resulting in improved performance and faster convergence.
- Normalization is used when we want to preserve the original distribution and when the algorithm is sensitive to feature range.
- Standardization is preferred when we want to ensure a standard normal distribution and when the algorithm is not sensitive to feature range.

The choice between normalization and standardization depends on the specific dataset, the features, and the machine learning algorithm being used.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

When a variable is perfectly correlated with other variables in the dataset, the variance inflation factor (VIF) becomes infinite.

This happens because it is impossible to estimate the effect of a variable on the target if that variable is completely explained by other variables.

In such cases, the model would be unstable and unable to make accurate predictions.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific distribution, often a normal distribution.

It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

In linear regression, Q-Q plots are crucial for checking the assumption of normality of residuals. Residuals are the differences between the actual observed values and the predicted values from the regression model.

1. **Normality of Residuals:** If the residuals follow a normal distribution, the points on the Q-Q plot will roughly fall along a straight diagonal line. Deviations from this line suggest departures from normality.

2. **Identifying Outliers:** Q-Q plots can help identify outliers, which are data points that deviate significantly from the overall pattern. Outliers can have a substantial impact on regression results and should be investigated.

3. **Model Validity:** If the residuals are not normally distributed, it can indicate that the linear regression model might not be the best fit for the data, and other techniques or transformations may be necessary.

Importance:

* **Validating Assumptions:** Q-Q plots are a valuable tool to check the key assumption of normality in linear regression, ensuring that the model's results are reliable and meaningful.

* **Improving Model Performance:** By identifying departures from normality and outliers, Q-Q plots can help improve the accuracy and robustness of the linear regression model.

* **Understanding Data:** Q-Q plots provide insights into the distribution of the residuals and can help guide the choice of appropriate statistical methods or data transformations for analysis.