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? (3 marks)

Answer:

From the analysis of the categorical variables in the dataset, we can infer the following effects on the dependent variable (cnt), which represents the demand for shared bikes:

- Season:
  - Spring (negative coefficient): The demand for shared bikes tends to decrease during the spring season compared to the reference season.
  - Summer (positive coefficient): The demand increases during the summer season.
  - Winter (positive coefficient): There is also an increase in demand during the winter season compared to the reference season (presumably autumn or fall).
- Month:
  - July (negative coefficient): The demand decreases in July.
  - September (positive coefficient): The demand increases in September.
- Weekday:
  - Sunday (negative coefficient): The demand is lower on Sundays compared to the reference day
  - Saturday (not included in the final model): If included, it would indicate the effect of Saturdays on demand. Since it's not in the final model, it might have had a negligible or redundant effect.
- Weather Situation:
  - Light Rain or Light Snow (negative coefficient): The demand significantly decreases during light rain or light snow conditions.
  - Misty and Cloudy (negative coefficient): The demand also decreases under misty and cloudy conditions, but to a lesser extent than during light rain or snow.
- Holiday (negative coefficient): The demand for shared bikes tends to be lower on holidays compared to non-holidays.
- Year (positive coefficient): The demand for shared bikes increased in the subsequent year compared to the base year.

Inferences:

- Seasonality: There is a clear seasonal trend where summer and winter see an increase in bike demand, while spring sees a decrease.
- Monthly Variation: Certain months like July see a dip in demand, whereas September sees a rise, possibly due to weather conditions, vacation periods, or other factors.
- Weekly Patterns: Demand tends to be lower on Sundays, which might be due to fewer people commuting to work or school.
- Weather Conditions: Adverse weather conditions like rain, snow, and cloudy weather decrease bike demand significantly, indicating that people prefer to avoid biking in such conditions.
- Holiday Effect: Fewer people use shared bikes on holidays, likely because there is less commuting.

- Yearly Trend: The overall demand for shared bikes has increased over the years, suggesting growing popularity or increased adoption of bike-sharing services.

These inferences can help tailor marketing strategies, operational planning, and service improvements to align with user behavior and external factors impacting bike demand.

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

Answer: Using drop_first=True during dummy variable creation is important for several reasons:

- Avoiding Multicollinearity:
    - When creating dummy variables, each categorical variable is transformed into a set of binary (0/1) columns. For example, a categorical variable with 4 categories will be transformed into 4 binary columns.
    - By setting drop_first=True, one category is dropped, removing the perfect multicollinearity. This ensures that the model can be properly estimated
- Interpretability:
    - Dropping the first dummy variable creates a reference category. The coefficients of the remaining dummy variables can then be interpreted relative to this reference category.
    - For instance, if a categorical variable "Season" with categories "Spring", "Summer", "Fall", and "Winter" is transformed into dummy variables with "Spring" as the dropped category, the coefficients for "Summer", "Fall", and "Winter" will indicate the effect of these seasons relative to "Spring".
- Model Efficiency:
    - Dropping one dummy variable reduces the number of columns (features) in the dataset. This can lead to a more efficient model in terms of computation and memory usage.
    - It simplifies the model by reducing redundancy without losing any information about the categorical variable.

Conclusion:

Using drop_first=True is a best practice in regression modeling with categorical variables because it helps avoid multicollinearity, makes the model more interpretable by providing a reference category, and enhances model efficiency.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer: Based on the pair-plot and also correlation matrix the highest correlation with the target variable is **temperature (Temp)**.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: After building the linear regression model on the training set, I validated its assumptions by these checkpoints: -

- Linearity of the Relationship:
    - Checked using scatter plots and residual plots.
    - Observed vs. predicted values should form a roughly straight line.
    - Residuals should exhibit no discernible pattern when plotted against predicted values.
- Independence of Errors:
    - Validated using the Durbin-Watson test.
    - A Durbin-Watson statistic close to 2 suggests no autocorrelation in residuals.
- Homoscedasticity:
    - Assessed by plotting residuals against predicted values.
    - Residuals should have a constant spread across all levels of predicted values, without any funnel-shaped pattern.
- Normality of Errors:
    - Evaluated through histograms
    - Residuals should follow a bell-shaped distribution in histograms
- No Multicollinearity:
    - Checked using Variance Inflation Factor (VIF).
    - VIF values above 10 (or sometimes 5) indicate high multicollinearity among predictors.

By systematically validating these assumptions, I ensure the reliability and validity of the linear regression model, thereby enhancing the confidence in its predictions and inferences.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top three features contributing significantly to explaining demand for shared bikes are those with the highest coefficients or importance scores. These are: -

- Temperature (temp): With a coefficient of 0.491, temperature has the highest positive impact on bike demand. This suggests that as temperature increases, the demand for shared bikes also tends to increase.
- Year (yr): With a coefficient of 0.233, the year variable indicates a positive impact on bike demand over time. This could imply that bike-sharing demand has been increasing over the years.
- Weather Situation - Light Rain or Light Snow (weathersit_Light Rain or Light Snow): Despite being negatively correlated, with a coefficient of -0.290, this weather situation variable still contributes significantly to explaining bike demand. It suggests that during light rain or snow, there is a decrease in bike demand.

These features stand out as the top contributors based on their coefficients in the final linear regression model.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Answer: Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable (also known as the target variable) and one or more independent variables (also known as predictor variables). It assumes a linear relationship between the independent variables and the dependent variable. Linear regression aims to find the best-fitting straight line that describes the relationship between the variables.

Key Components of Linear Regression:

- **Dependent Variable (Y):**
  - The variable we want to predict or explain. It is denoted as Y.

- **Independent Variable(s) (X):**
  - The variables that are used to predict the dependent variable. It can be one or more variables. Denoted as X1, X2, ..., Xr.

- **Coefficients (θ1, θ2, ..., θr):**
  - The parameters or weights associated with each independent variable. These coefficients determine the slope of the regression line.

- **Intercept (θ0):**
  - The constant term that represents the value of the dependent variable when all independent variables are zero.

Working of Linear Regression:

Linear regression works by fitting a linear equation to the observed data points. The general equation for a simple linear regression model with one independent variable is:

$$Y=\theta 0+\theta 1X1$$

For a multiple linear regression model with multiple independent variables, the equation becomes:

$$Y=\theta 0+\theta 1X1+\theta 2X2+...+\theta rXrY$$

The goal of linear regression is to estimate the coefficients (θ0, θ1, θ2, ..., θr) that minimize the difference between the observed values of the dependent variable and the values predicted by the model.

Steps in Linear Regression:

- **Data Collection:** Gather data on the dependent variable and independent variables.
- **Data Preprocessing:** Clean the data, handle missing values, and perform feature engineering if needed.
- **Model Training:** Use the training data to estimate the coefficients (θ0, θ1, θ2, ..., θr) that minimize the error between the observed and predicted values.
- **Model Evaluation:** Assess the performance of the model using evaluation metrics such as mean squared error (MSE), R-squared (R2), etc.
- **Prediction:** Use the trained model to make predictions on new data.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but appear very different when plotted graphically. This quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and not relying solely on summary statistics.

Explanation:

Imagine you have four different datasets, each containing several pairs of X and Y values. When you look at these datasets, you might calculate summary statistics like the mean, variance, and correlation coefficient to describe the relationship between X and Y.

Now, Anscombe's quartet takes these four datasets, and interestingly, despite having very different- looking scatter plots, they all share almost the same statistical properties! For example, each dataset has the same:

- **Mean of** X **and** Y**:** The average value of both X and Y is almost identical across all datasets.
- **Variance of** X **and** Y**:** The spread or variability of data points around the mean is also very similar

  in all datasets.

- **Correlation Coefficient:** The strength and direction of the linear relationship between X and Y are nearly the same for all datasets.

Why It's Important:

Anscombe's quartet highlights the limitations of relying solely on summary statistics without visualizing the data. It demonstrates that datasets with the same summary statistics can exhibit vastly different relationships between variables when plotted graphically.

Real-World Implications:

- **Data Visualization:** It emphasizes the importance of visualizing data to gain deeper insights beyond summary statistics.
- **Statistical Analysis:** It cautions against drawing conclusions based solely on summary statistics, as they may not capture the full complexity of the data.
- **Decision Making:** In real-world scenarios, decision-makers should not rely solely on statistical summaries but should also consider visual representations to understand the underlying patterns in the data.

  In essence, Anscombe's quartet teaches us that seeing is believing, and it's crucial to explore data visually to truly understand its nuances and patterns.

3. What is Pearson's R? (3 marks)

   Answer: Pearson's r, also known as Pearson's correlation coefficient, is a measure that tells us how two sets of data are related to each other in a linear way. It ranges from -1 to 1.

   - If r=1, it means the two sets of data have a perfect positive linear relationship, meaning as one set of data increases, the other set also increases in a perfectly straight line.
   - If r=−1, it means the two sets of data have a perfect negative linear relationship, meaning as one set of data increases, the other set decreases in a perfectly straight line.
   - If r=0, it means there is no linear relationship between the two sets of data.

   For example, if we look at the relationship between the number of hours a student studies and the grade they get on a test, Pearson's r can tell us if there's a strong positive correlation (meaning more studying leads to higher grades), a strong negative correlation (meaning more studying leads to lower grades), or no correlation at all. It helps us understand how closely two sets of data move together.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Answer:

   Scaling is the process of transforming the values of variables to a specific range or distribution. It's like resizing or reshaping the data so that it's easier to compare and analyze.

   Why Scaling is Performed:

   Comparability: Scaling makes it easier to compare variables that have different units or ranges.

   Algorithm Performance: Many machine learning algorithms perform better when the features are scaled. Scaling helps algorithms converge faster and prevents some features from dominating others.

   Difference Between Normalized Scaling and Standardized Scaling:

   Normalized Scaling: Also known as Min-Max scaling, it transforms the data to a range between 0 and 1. Each value is subtracted by the minimum value in the dataset and then divided by the difference between the maximum and minimum values.

Standardized Scaling: Also known as z-score scaling, it transforms the data so that it has a mean of 0 and a standard deviation of 1. Each value is subtracted by the mean of the dataset and then divided by the standard deviation.

Simple application:

Normalized Scaling: Imagine you have a dataset of test scores ranging from 60 to 90. Normalized scaling would transform these scores so that 60 becomes 0 and 90 becomes 1, with all other scores falling in between.

Standardized Scaling: Using the same test score example, standardized scaling would transform the scores so that the mean is 0 and the spread of scores is consistent, regardless of the original range of scores. It's like putting all the scores on the same scale, centered around the average score.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: When the value of Variance Inflation Factor (VIF) becomes infinite, it indicates that there's a perfect linear relationship between one of the predictor variables and the combination of the other predictor variables.

In simple terms, it's like having one predictor variable that can be perfectly predicted by a combination of other variables. This situation is known as multicollinearity, where one variable is almost a duplicate or a perfect linear combination of another variable(s).
Multicollinearity can cause issues in regression analysis because it makes it difficult for the model to estimate the unique effect of each predictor variable on the target variable. It can lead to unreliable coefficient estimates and inflated standard errors, affecting the overall accuracy and interpretability of the model.

In practical terms, infinite VIF values serve as a warning sign that the variables in the model are highly correlated with each other and might need to be addressed, either by removing redundant variables or by using techniques like regularization to mitigate the multicollinearity issue.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to compare the distribution of a sample of data to a theoretical distribution, typically a normal distribution.

How it Works:

- In a Q-Q plot, the quantiles (ordered values) of the sample data are plotted against the quantiles of the theoretical distribution.
- If the sample data follows the theoretical distribution closely, the points in the plot will form a straight line.
- Deviations from the straight line indicate differences between the sample distribution and the theoretical distribution.

Use and Importance in Linear Regression:

• **Normality Assumption:** Q-Q plots are used to assess whether the residuals (errors) from a linear regression model are normally distributed. This is important because many statistical tests and techniques, including linear regression, assume that the residuals are normally distributed.

- **Model Validity**: A linear regression model is valid when the residuals are normally distributed. If the Q-Q plot shows deviations from a straight line, it suggests that the residuals do not follow a normal distribution, indicating potential issues with the model's assumptions.
- **Identifying Outliers**: Q-Q plots can also help identify outliers or data points that deviate significantly from the expected distribution. Outliers can affect the model's accuracy and may need to be investigated further.

  In simple terms, a Q-Q plot helps us check if the residuals of a linear regression model look like they come from a normal distribution. If they do, it suggests that the model is valid and reliable. If not, it's a sign that the model might need adjustments or further investigation.