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:

From the analysis of the categorical variables in the dataset, we can infer the following about their effect on the dependent variable (**`cnt`**):

- **Year (yr)**: The year (yr) has a **strong positive effect** on bike rentals. The coefficient of 0.2344 indicates that rentals significantly increased in 2019 compared to 2018. This suggests either increasing popularity of the bike rental service over time or other year-specific factors driving more usage in 2019.
- **Season (season)**: The seasons have **varying effects** on bike rentals:
    - **Spring** (season_Spring) has a **negative effect** with a coefficient of -0.1152, meaning fewer rentals occur during this season.
    - **Winter** (season_Winter) has a **positive but smaller effect** (0.0423), suggesting a moderate increase in bike rentals during winter months.
    - This implies that summer and fall (the reference seasons) likely see higher bike rentals compared to spring and winter.
- **Weather conditions (weathersit)**: Weather conditions also show a significant impact:
    - **Mist/Cloudy weather** (weathersit_Mist Cloud) has a **negative effect** on rentals (coef = -0.0791).
    - **Light Snow or Rain** (weathersit_Light Snow / Rain) has a **stronger negative effect** (coef = -0.2801), indicating that poor weather conditions drastically reduce bike rentals.

Overall, categorical variables like year, season, and weather conditions clearly influence bike rental patterns, with bike usage increasing over time, varying across seasons, and decreasing under unfavorable weather conditions.

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

Ans:

Using `drop_first=True` when creating dummy variables helps avoid multicollinearity or **prevent confusion in the model** and makes it easier to understand.

- **Prevents the Dummy Variable Trap**: When creating dummy variables, if all categories are included, the information from one category can be inferred from the others, leading to perfect multicollinearity (i.e., one variable is a linear combination of others). By dropping the first category, we remove this redundancy and avoid this issue, which allows the model to run correctly.
- **Improves Model Interpretability**: By dropping the first category, the remaining categories are compared against the dropped reference category. This helps in easier interpretation of the coefficients, as they represent the effect relative to that reference category.

In essence, `drop_first=True` ensures a valid model by eliminating unnecessary multicollinearity and enhances 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 analysis, the variables `temp` (temperature) and `atemp` (feeling temperature) exhibit the highest correlation with `cnt`, the target variable representing the count of total rental bikes. The scatter plots in the visualization show a clear, tight linear pattern between both `temp` and `atemp` with `cnt`, indicating that as temperatures rise, so does the usage of rental bikes. This strong positive correlation suggests that warmer conditions likely encourage more people to rent bikes.

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

Ans:

To ensure my linear regression model was reliable, I checked several key assumptions after building the model on the training set:

- **Linearity**: I verified that the relationship between the predictors and the target variable (`cnt`) was linear. This was done by looking at the scatter plots of the residuals against the predicted values. These plots should not show any particular pattern; the points should be randomly spread around a central line.
- **Normality of Residuals**: I needed to make sure that the residuals (the differences between observed and predicted values) were normally distributed. To do this, I

looked at the histogram and the density plot of the residuals, checking if they approximated a bell curve shape.
- **Independence of Residuals**: It's important that the residuals are independent of each other. I used the Durbin-Watson statistic to test for autocorrelation in the residuals, where a value close to 2 suggests that there is no autocorrelation.
- **Multicollinearity**: Finally, I checked for multicollinearity, which happens when predictor variables are too highly correlated with each other. I used Variance Inflation Factors (VIF) to assess this. A VIF above 5 would indicate problematic multicollinearity, which could distort the regression coefficients and weaken the statistical power of the model.

By addressing these checks thoroughly, I made sure that the model's assumptions were met, thereby enhancing the trustworthiness and accuracy of my regression analysis.

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

Ans:

Based on the final model results, the top three features contributing significantly towards explaining the demand for shared bikes are:

- **Temperature (`temp`)**: With a coefficient of 0.4507, temperature has a strong positive impact on bike rentals, indicating that higher temperatures lead to an increase in bike usage. This is the highest among all the coefficients, highlighting its crucial role in influencing bike rental demand.
- **Year (`yr`)**: The coefficient for the year variable is 0.2344, which shows a significant positive effect, reflecting that bike rentals were higher in 2019 compared to 2018. This suggests a year-over-year growth in bike usage.
- **Weather Situation - Light Snow or Rain (`weathersit_Light Snow / Rain`)**: This variable has a coefficient of -0.2801, indicating a strong negative impact on bike rentals. Adverse weather conditions such as light snow or rain substantially decrease the demand for bike rentals.

These three factors are the most significant predictors of bike rental demand according to the model, with temperature playing the most critical role in encouraging higher rental activity, followed by the effect of year indicating growth or trend changes, and the significant deterrent effect of poor weather conditions.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- **Definition:**
    - Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. The simplest form is simple linear regression, which involves one independent variable, while multiple linear regression involves two or more.

- **What is Linear Regression?**
    - Linear regression is a method used to **predict a value** (the dependent variable) based on other known values (independent variables). It tries to draw a **straight line** through data points that best represents the relationship between these values.
    - For example, if you want to predict someone's salary based on their years of experience, you can use linear regression to find the relationship between **experience** (independent variable) and **salary** (dependent variable).

- **The Equation:**
    - Linear regression can be represented by a simple formula:
      $Y = \beta 0 + \beta 1\, X1 + \beta 2\, X2 + ... + \beta n\, Xn + \epsilon$
        - **Y**: The value we want to predict (e.g., salary).
        - **X**: The known value (e.g., years of experience).
        - $\beta 0$: The intercept, which is where the line crosses the Y-axis (it's the value of Y when X is 0).
        - $\beta 1, \beta 2, ..., \beta n$: The slope, which tells us how much Y changes for every unit increase in X (e.g., how much salary increases for each additional year of experience).
        - ϵ : This is the error term (residual), which accounts for the difference between the predicted and actual values.

- **Goal of Linear Regression:**
    - The main goal is to find the **best-fitting line** through the data. This line **minimizes the errors** (the difference between the actual data points and the predicted values on the line). To do this, linear regression uses a method called **Ordinary Least Squares (OLS)**.
    - **OLS** works by minimizing the sum of the squared differences between the observed data points and the line (hence the term "least squares").
- **Example:**
    - Consider a group of people, showing their **years of experience** and **salary**. By plotting this data on a graph, linear regression helps you draw a straight line that shows the general trend – people with more experience tend to have higher salaries. Once this line is drawn, you can predict someone's salary based on how many years of experience they have.
- **Key Assumptions:**
    - **Linearity**: The relationship between the input (X) and output (Y) should be straight. If the relationship curves, linear regression might not work well.
    - **Independence**: Each data point should be independent of the others (e.g., one person's salary shouldn't depend on another's).
    - **Equal Variance**: The spread of errors (differences between actual and predicted values) should be roughly the same across all data points. If some points have much bigger errors, it can lead to problems.
    - **Normal Distribution of Errors**: The errors (the differences between the predicted and actual values) should follow a normal distribution.
- **Benefits:**
    - **Simple and Interpretable**: It's easy to understand how each independent variable (e.g., years of experience) impacts the dependent variable (e.g., salary).
    - **Fast**: It’s computationally efficient and works well for small to medium-sized datasets.
- **Limitations:**
    - If the data doesn’t meet the assumptions (like linearity or normal distribution of errors), the model may not perform well.

    - Linear regression can't handle complex relationships (for example, non-linear relationships) between the variables as well as some other machine learning methods.

- **Conclusion:**

    - Linear regression is a great tool for modeling simple relationships between variables. It helps predict a continuous value based on other known values by drawing a straight line through the data points. However, it works best when its assumptions are met and when the relationship between the variables is relatively straightforward.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a set of four different datasets created by statistician Francis Anscombe in 1973. These four datasets all have nearly identical statistical properties, such as:

- The average (mean) of the X and Y values
- The spread (variance) of the X and Y values
- The correlation between X and Y
- The equation of the line that best fits the data

Despite these similarities in statistical measures, when plotted on a graph, the datasets look very different.

The Four Datasets:

- **Dataset 1**: The points form a straight line, making it appropriate to fit a linear regression (line of best fit).
- **Dataset 2**: The points form a curve, so a straight line wouldn't provide an accurate representation.
- **Dataset 3**: Most points line up well, but there is a single outlier that affects the line of best fit.
- **Dataset 4**: Almost all the X values are the same except for one outlier, which makes the line of best fit misleading.

Key Insights from Anscombe's Quartet:

- **Statistics alone can be misleading**. Just by looking at the average, variance, or correlation, it might seem like all four datasets are the same. However, plotting the data reveals significant differences.
- **Visualizing data is crucial**. Graphs help to identify patterns such as curves or outliers that may not be apparent from the statistical measures alone. They provide a clearer understanding of the data.
- **Outliers can distort results**. A single extreme point can influence the outcome of statistical analyses, such as shifting the line of best fit. It is essential to consider the impact of outliers on the analysis.

Conclusion:

Anscombe's quartet demonstrates the importance of visualizing data alongside statistical analysis. It highlights that relying solely on numbers can lead to incorrect conclusions and that graphs offer a more accurate picture of the data's true nature.

3. What is Pearson's R?

Ans:

Pearson's R, also known as the **Pearson correlation coefficient**, is a measure of the **strength and direction** of the linear relationship between two variables. It tells how strongly two variables are related and whether the relationship is positive or negative.

How it works:

- Pearson's R gives a value between **-1 and 1**.
    - A value of **1** means there is a **perfect positive relationship**, meaning as one variable increases, the other increases at a constant rate.
    - A value of **-1** means there is a **perfect negative relationship**, meaning as one variable increases, the other decreases at a constant rate.
    - A value of **0** means there is **no linear relationship** between the two variables.

Example:

- If Pearson's R is **0.8**, it indicates a strong **positive** relationship, meaning both variables tend to increase together.

- If Pearson's R is **-0.7**, it indicates a strong **negative** relationship, meaning one variable tends to decrease when the other increases.

Conclusion:

Pearson's R is useful for understanding how two variables move together and whether the relationship is strong, weak, positive, or negative. It is widely used in data analysis to quantify the degree of association between variables.

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

Ans:

- **What is Scaling?**
    - Scaling is the process of adjusting the values of data so that different features or variables fall within a similar range. This ensures that no single feature dominates simply because it has larger numbers. For example, if one feature is in kilometers and another in meters, scaling brings them to a common level for better analysis.
- **Why is Scaling Performed?**
    - Scaling improves how machine learning models work, especially when different variables have different units or ranges. It ensures that all variables contribute equally to the model.
    - It prevents features with larger numbers from dominating the learning process, which can lead to biased results.
    - For algorithms that calculate distances (like k-nearest neighbors or clustering), scaling ensures that all features contribute fairly to the distance calculation.
- **Difference between Normalized Scaling and Standardized Scaling:**
    - **Normalized Scaling**:
        - Normalization adjusts the values so that they fall within a specific range, typically between **0 and 1**. This is useful when the range of values is important.
        - **Formula**:

$$X' = \frac{X_{max} - X_{min}}{\text{----------}}$$

$$X - Xmin$$

- This formula scales each value of X based on the minimum and maximum values of the data.

- **Standardized Scaling**:
  - Standardization adjusts the data so that the mean is **0** and the standard deviation is **1**. This centers the data and spreads it evenly.
  - **Formula**:

$$X' = \frac{X - \mu}{\sigma}$$

  - This formula centers the data by subtracting the mean (μ) and scales it by dividing it by the standard deviation (σ).

- **Conclusion:**

  - Scaling is essential for preparing data in machine learning, allowing features to contribute equally to the model. Normalization scales data to a specific range (like 0 to 1), while standardization centers data around a mean of 0 with a standard deviation of 1.

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

Ans:

A **VIF (Variance Inflation Factor)** value can become **infinite** when there is **perfect multicollinearity** between two or more variables. This means that one variable can be **perfectly predicted** by the others, creating a situation where the variables are too strongly related. When this happens, the model struggles to distinguish between the effects of these variables, leading to an infinite VIF.

- **Why does this happen?**

  - **Multicollinearity** occurs when variables are highly correlated, and one can be explained by others with little to no error.
  - The VIF formula measures how much a variable's variance is inflated due to multicollinearity. When there is perfect multicollinearity, the denominator in the VIF formula becomes **zero**, causing the result to be **infinite**.

- **VIF Formula:**

$$VIF = \frac{1}{1 - R^2}$$

  Where:
  - R^2 is the **coefficient of determination** from regressing one variable on the others.
  - If $R^2 = 1$ (indicating perfect correlation), the denominator becomes zero, resulting in an **infinite VIF**.

- **Example:**

  - If two variables, **X1** and **X2**, are perfectly correlated (e.g., X2 = 2 * X1), then R^2 for one variable when regressed on the other will be **1**. Plugging this into the formula leads to division by zero, resulting in an **infinite** VIF.

- **Conclusion:**

  - An infinite VIF signals **perfect multicollinearity**, where one variable can be exactly predicted by others. This indicates a problem in the model, suggesting that one or more variables need to be removed or modified to resolve this issue.

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

Ans:

A **Q-Q plot** (quantile-quantile plot) is a type of graph used to compare the **distribution** of a dataset to a **theoretical distribution**, typically the **normal distribution**. It plots the **quantiles** of the data against the **quantiles** of the theoretical distribution to check if the data follows that distribution.

- **Use in Linear Regression:**

  - In linear regression, a key assumption is that the **residuals** (the differences between the actual and predicted values) should be **normally distributed**.
  - The Q-Q plot is used to check whether the residuals follow a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will lie along a **straight diagonal line**.

- **Importance of a Q-Q Plot:**
    - **Assumption Check**: It helps verify the normality assumption in linear regression. If the points deviate from the straight line, it suggests that the residuals are not normally distributed, which may affect the accuracy of the model.
    - **Detecting Outliers**: A Q-Q plot can reveal outliers or data points that do not fit the assumed distribution, which could influence the results of the regression.
    - **Model Validation**: By confirming the normality of residuals, the Q-Q plot helps ensure that the linear regression model is valid and that the conclusions drawn from it are reliable.

- **Conclusion:**
    - A Q-Q plot is a graphical tool used to check if the residuals of a linear regression model are normally distributed. It is important for validating the assumptions of the model and ensuring its accuracy.