Name: Sanchita Patil

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?

1. Season:
Peak Season: Fall (3) shows the highest average bike rentals, indicating peak demand.
Low Season: Spring (1) has the lowest average bike rentals.

2. Year:
The bike-sharing service likely saw increased popularity or improved market conditions in 2019 (indicated as 1), making it a more favorable year for bike rental operations compared to the previous year 2018(indicated as 0).

3. Holiday:
On average, there are slightly fewer bike rentals on holidays compared to non-holiday days (holiday = 1 mean: 3735 bikes vs. holiday = 0 mean: 4530 bikes).

4. Weekday:
Weekend, particularly Saturdays, tend to have higher bike rental demand. Sundays (weekday 6) exhibit higher variability, suggesting less predictable rental patterns, possibly influenced by varying weekend activities or weather conditions. There is a noticeable trend where weekdays closer to the weekend (Thursday to Saturday) generally have higher average bike rental counts compared to weekdays early in the week (Monday to Wednesday).

5. Working day:
Working days tend to have higher average bike rentals, suggesting that commuting and work-related travel may contribute significantly to bike rental demand during weekdays.

6. Weather Condition:
The visual reveals that weather conditions significantly influence bike rental demand. Clear or mostly clear days (weathersit = 1) generally see the highest average and maximum bike rentals, likely due to favorable outdoor conditions encouraging bike usage. In contrast, days with mist or cloudy weather (weathersit = 2) show moderate demand, while days with light snow or rain (weathersit = 3) experience lower average and maximum rentals.

7.Month:
June (mnth = 6) emerges as the month with the highest average bike rentals, likely due to favorable weather conditions and increased outdoor activities during summer. In contrast, January (mnth = 1) and February (mnth = 2) have the lowest average bike rentals, correlating with colder weather and potentially less favorable outdoor conditions.

Name: Sanchita Patil

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

By setting drop_first = True, we are telling Python to drop the first dummy variable that it creates for each categorical feature. This action breaks the perfect correlation between dummy variables and prevents multicollinearity. Essentially, dropping one dummy variable ensures that each variable captures unique information without redundancy.

Example:

- Without drop_first = True : We have dummy variables for "spring," "summer," "fall," and "winter." If all are 0, then it must be "spring."
- With drop_first = True : You have dummy variables for "summer," "fall," and "winter." If all are 0, the model knows it's "spring" without needing an explicit "spring" dummy.

Conclusion:

Using drop_first = True when creating dummy variables helps the model work more effectively by avoiding redundancy and ensures each variable contributes unique information. This practice enhances the accuracy and reliability of predictions made by the model, making them more suitable for real-world applications like predicting bike rentals based on seasonal patterns.

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

Looking at the pair-plot and correlation matrix among the numerical variables, registere d has the highest correlation(0.945411) with the target variable

Name: Sanchita Patil

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

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

1. **Checking Linearity:** I examined scatter plots of the dependent variable against each independent variable to ensure that the relationships are approximately linear.
2. **Assessing Normality of Residuals:** I plotted a histogram and a Q-Q plot of the residuals to verify if they follow a normal distribution. This helps ensure that the errors are normally distributed, which is important for valid statistical inference.
3. **Testing for Homoscedasticity:** I utilized scatter plots of standardized residuals against predicted values to check for constant variance (homoscedasticity) of the residuals across different levels of predicted values.
4. **Detecting Multicollinearity:** I calculated Variance Inflation Factors (VIFs) for the independent variables to identify and address multicollinearity issues, ensuring that predictors are not highly correlated with each other.
5. **Examining Autocorrelation:** I performed the Durbin-Watson test on the residuals to detect autocorrelation, which assesses whether residuals are independent over time or across observations.

Based on these assessments, all assumptions of Linear Regression were found to be satisfied in my model, confirming the reliability of the regression analysis and the validity of the statistical inferences drawn from it.

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 provided and considering the coefficients and statistical significance (P-values), the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **Average Temperature (avg_temperature):** This variable has a coefficient of 0.4592 and a very low P-value ($P < 0.001$), indicating a strong positive relationship with bike demand. As average temperature increases, the demand for shared bikes tends to increase as well.
2. **Year (year):** The coefficient for year is 0.2360 with a very low P-value ($P < 0.001$), indicating that over the years, there has been an increasing trend in bike demand. This suggests that bike sharing has become more popular or accessible over time.
3. **Weather Situation (weathersit_Light Rain/Snow):** This categorical variable has a coefficient of -0.2514 with a very low P-value ($P < 0.001$), indicating a significant negative impact on bike demand during light rain or snow. As the weather

worsens, particularly with light rain or snow, the demand for shared bikes decreases.

General Subjective Questions

1. Explain the linear regression algorithm in detail

What is Linear Regression?

Linear regression helps us understand how one variable (let's call it y) changes with respect to another variable (let's call it X). It assumes that y can be expressed as a straight line function of X, plus some random error.

Formula:

The formula for a simple linear regression, where we have one predictor variable (X) and one outcome variable (y), can be written as:

$y = \beta 0 + \beta 1 \cdot X + \epsilon$

- y : The outcome or dependent variable we want to predict.
- X : The predictor or independent variable that influences y.
- β0 : Intercept, where the line crosses the y-axis. ()
- β1: Slope, which tells us how much y changes for a unit change in x.
- ϵ : Random error or noise that the model cannot explain.

Requirements and Assumptions:

1. **Linearity:** The relationship between X and y should be linear, meaning when X changes, y changes proportionally.
2. **Independence:** The observations (data points) should be independent of each other. One observation should not influence another.
3. **Homoscedasticity:** The variance of the residuals (the differences between observed and predicted values) should be constant across all levels of X. In simpler terms, the spread of points around the line should be similar everywhere.
4. **Normality:** The residuals should follow a normal distribution. This means that most residuals are close to zero, and fewer residuals are far from zero.

Name: Sanchita Patil

5. **No Multicollinearity:** If using multiple predictor variables, they should not be highly correlated with each other. This can make it difficult to separate their individual effects on y.

Example:

Imagine you want to predict how much someone weighs based on their height. Linear regression helps you find the best straight line that shows how weight changes with height. The line has an intercept (where the line crosses the weight axis) and a slope (how much weight changes for every inch of height).

You gather data points (people's heights and weights) and draw this line. The model also considers some random factors that might affect weight (like diet or genetics), which we can't predict perfectly.

To use linear regression, you need data that shows how xxx (like height) affects yyy (like weight). The method assumes that this relationship is a straight line, the data points are independent, the errors are random and follow a normal pattern, and that the variables you use are not too similar to each other.

In essence, linear regression is like drawing the best-fitting line through your data points to understand and predict relationships between variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a fascinating example in statistics that highlights the importance of visualizing data rather than relying solely on summary statistics. It consists of four datasets that have nearly identical simple descriptive statistics (like means, variances, correlations) but are completely different when plotted graphically.

Anscombe's quartet is named after the statistician Francis Anscombe. It consists of four sets of data points (datasets A, B, C, and D) that, at first glance, appear to have very similar statistical properties such as means, variances, correlations, and regression lines. However, when you plot these datasets and visually examine them, they reveal vastly different patterns and relationships.

Key Points:

1. **Similar Statistics:** Each dataset in Anscombe's quartet has:
    - Same mean and variance for both x and y variables.
    - Same correlation coefficient between x and y.
    - Same linear regression line equation (slope and intercept).
2. **Different Patterns:** Despite their similar summary statistics, the datasets tell different stories when plotted:
    - **Dataset A:** Forms a simple linear relationship.

- **Dataset B:** Also forms a linear relationship but with an outlier that influences the regression line.
- **Dataset C:** Has a non-linear relationship where one outlier significantly affects the regression line.
- **Dataset D:** Appears to have no clear relationship when plotted, with one outlier dramatically changing the linear regression.

3. **Implications:** Anscombe's quartet illustrates that summary statistics alone (like means and correlations) can be misleading. It emphasizes the importance of data visualization to understand relationships and patterns in data. What may seem like identical datasets statistically can behave very differently when plotted visually.

Why It's Important:

- Anscombe's quartet challenges the notion that summary statistics provide a complete understanding of data. It underscores the need to visualize data to detect outliers, non-linear relationships, and other nuances that summary statistics may miss.
- It highlights the danger of over-relying on statistical measures without verifying through graphical exploration.

In essence, Anscombe's quartet teaches us that "seeing is believing" in data analysis—visualizing data can reveal insights that numbers alone cannot capture. It's a powerful reminder to always explore and interpret data with both statistical analysis and visual examination.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as r, is a measure that tells us how closely two variables are linearly related to each other. In simple terms, it quantifies the strength and direction of the relationship between two variables.

Understanding Pearson’s r:

1. **Measuring Relationship :** Pearson’s r is used to determine how much one variable changes when another variable changes. For example, if we're looking at height and weight, Pearson’s r would tell us how much weight tends to change as height changes.
2. **Strength of Relationship:** The value of r ranges from -1 to +1:
   - **Positive r** (closer to +1): Indicates that as one variable increases, the other variable tends to increase as well. For instance, as height increases, weight also tends to increase.
   - **Negative r** (closer to -1): Indicates that as one variable increases, the other tends to decrease. For example, as temperature increases, sales of winter coats tend to decrease.

- **r = 0**: Means there is no linear relationship between the variables; changes in one variable do not predict changes in the other.

3. **Graphical Representation:** When you plot data points for two variables, Pearson's rrr tells you how tightly these points cluster around a straight line:
   - If r is close to +1 or -1, the points cluster closely around a line, indicating a strong linear relationship.
   - If r is close to 0, the points are scattered randomly, suggesting no strong linear relationship.
4. **Assumptions:** Pearson's r assumes that the relationship between variables is linear (changes in one variable are directly proportional to changes in the other), and that the data points are not influenced by outliers or other non-linear patterns.
5. **Interpreting r:** The magnitude of r indicates how strong the relationship is:
   - $r = 1$ or $r = -1$: Perfect positive or negative linear relationship.
   - $r = 0$: No linear relationship.
   - $0 < |r| < 0.30$: Weak relationship.
   - $0.3 \leq |r| < 0.70.3$: Moderate relationship.
   - $|r| \geq 0.7$: Strong relationship.

In summary, Pearson's correlation coefficient r helps us understand how closely two variables move together.

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

Scaling in the context of data refers to the process of transforming variables to a comparable range. It's done to ensure that different variables contribute equally to the analysis and to improve the performance of certain algorithms.

Why Scaling is Performed:

- **Equal Contribution:** Variables may have different scales (e.g., age in years vs. income in thousands), which could skew analysis towards variables with larger scales. Scaling ensures each variable contributes equally.
- **Algorithm Performance:** Many machine learning algorithms perform better or converge faster when features are on a similar scale. Scaling helps algorithms that use distance-based metrics or require normalization.
- **Interpretability:** Scaling can make it easier to interpret the coefficients or weights derived from models.

Types of Scaling:

Type	Description	Example
Normalized Scaling	Scales data to a [0, 1] range. Transforms data proportionally within the minimum and maximum values of each variable.	Age (18-65 years) normalized to [0, 1]: 30 years → (30 - 18) / (65 - 18) = 0.28

Name: Sanchita Patil

Standardized Scaling	Centers data around 0 with a standard deviation of 1 (Z-score). Useful for algorithms that assume normally distributed data.	Income ($20,000-$200,000) standardized: $50,000 → (50,000 - mean) / standard deviation

Key Differences:

- **Normalized Scaling:** Adjusts values to a specific range (e.g., [0, 1]). It's useful when the distribution of the data does not follow a Gaussian (normal) distribution. This type of scaling retains the original distribution shape but compresses or stretches it to fit within the specified range.
- **Standardized Scaling:** Transforms data to have a mean of 0 and a standard deviation of 1. It's suitable for algorithms that assume normally distributed data or benefit from the scale of the data. Standardized data has zero mean and unit variance, making comparisons between different variables more straightforward.

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

Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors in a regression model. When the VIF for a predictor variable is infinite, it indicates an extremely high level of collinearity with other variables in the model.

Example:

Imagine you have a group of friends, and you're trying to predict how happy each friend is based on factors like their age, income, and how much they exercise. Now, let's say two of your friends, John and Sarah, are very similar in age, income, and exercise habits. When you try to predict happiness, the model gets confused because it's hard to tell if changes in happiness are due to John's age, income, or exercise, or if it's because Sarah's values are so similar.

In statistical terms, this confusion is called collinearity. VIF measures how much this confusion affects your predictions. When VIF is infinite for a variable, like John's age, it means that this confusion is so severe that the model can't separate out the effects of age from the effects of income or exercise. In other words, the information about one variable (like age) is redundant because it's already perfectly explained by the other variables (like income and exercise).

In practical terms, infinite VIF usually happens when one or more variables in your model are almost perfectly predictable from the others. This can make your regression model unreliable because it becomes difficult to trust the individual effects of each predictor on the outcome. To solve this problem, you might need to remove one of the

highly correlated variables or find a way to combine them into a single variable that captures their shared information more effectively.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a certain theoretical distribution, such as the normal distribution. Here's a simple explanation of its use and importance in linear regression:

Example

Imagine you have a list of numbers that represent something you're studying, like heights of people in a town. A Q-Q plot helps you see if these numbers look like they might come from a specific type of distribution, like if they're normally distributed (which means most people are around the average height, with fewer very tall or very short people).

Use and Importance in Linear Regression:

In linear regression, we often assume that our data follows a normal distribution. This assumption is important because it affects how we interpret the results of our regression analysis. Here's why Q-Q plots are useful:

1. **Checking Normality**: Before using linear regression, we typically check if our data (like heights) is normally distributed. A Q-Q plot helps us visually compare our data's distribution to a normal distribution. If the points on the plot form a straight line, it suggests our data is close to being normal.
2. **Model Assumptions**: Linear regression works best when our data meets certain assumptions, including normality. If our data doesn't look normal on a Q-Q plot, it might mean our model won't give us accurate predictions.
3. **Adjusting Data**: If our data isn't normal, we might need to transform it (like using logarithms) to make it fit better with linear regression. Q-Q plots help us decide if these transformations are needed.

Working:

- **Plotting**: In a Q-Q plot, we plot the quantiles (or positions) of our data against the quantiles of a theoretical normal distribution. Each point on the plot represents how well our data matches the normal distribution.
- **Interpretation**: If our data points fall along a straight line, it means our data is close to normal. If they don't, it suggests our data might be skewed (like having too many people of one height) or have outliers (like very tall or short people).

In essence, a Q-Q plot is like a check-up for our data before using it in linear regression. It helps us see if our data fits our assumptions, like being normally distributed. By using Q-

Name: Sanchita Patil

Q plots, we can make sure our linear regression analysis gives us the most accurate predictions possible. If our data doesn't look normal, we know we might need to adjust it to get better results.