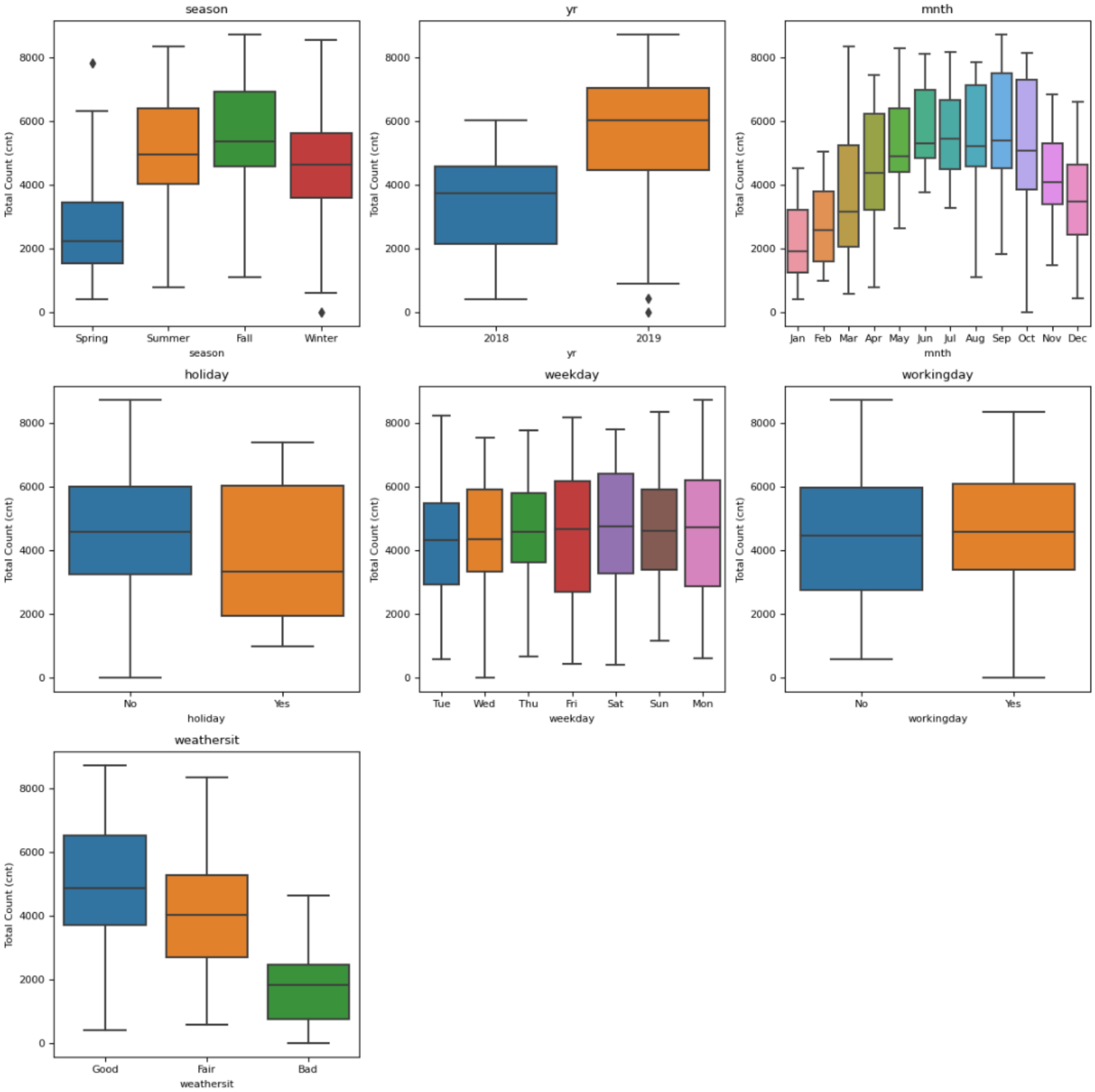
Assignment-based Subjective Questions

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:



Above plots shows Box plots visualizing the relationship between several categorical features and the total count (`cnt`), presumably of bike rentals. Here's a breakdown of the insights:

- **`season`:** Bike rentals are highest in Fall and Summer, with lower counts in Spring and significantly lower counts in Winter. There is one outlier of a very high count in Spring.
- **`yr`:** Bike rentals are considerably higher in 2019 than in 2018. A few low outliers exist for 2019.
- **`mnth`:** This plot clearly shows seasonality. Rentals increase from January to a peak in June-September, then decrease again through the end of the year. This aligns with the `season` plot.

- **`holiday`:** The median rental count is slightly higher on non-holidays. However, the interquartile range (IQR) is wider for non-holidays, indicating more variability in rental counts on regular days.
- **`weekday`:** There's not a strong, visually apparent relationship between the day of the week and rental counts. The distributions look relatively similar across weekdays.
- **`workingday`:** The median rental count is higher on working days compared to non-working days. This might seem counterintuitive but could be due to factors such as commuting or other workday-related bike usage.
- **`weathersit`:** Rental counts are highest in "Good" weather, followed by "Fair" weather, and drastically lower in "Bad" weather. This is as expected.

Summary:

- Seasonality is a Major Factor: The `season` and `mnth` plots highlight the strong seasonal influence on bike rentals.
- Year-over-Year Growth: Rentals appear to have increased from 2018 to 2019.
- Weather is Important: `weathersit` has a substantial impact, with bad weather significantly reducing rentals.
- Weekday Effect is Weak: There's no clear pattern based on the day of the week.
- Working Day vs. Non-Working Day: Higher median rentals on working days suggest potential commuting or work-related usage.

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:

Using `drop_first=True` when creating dummy variables in regression models is crucial for avoiding the dummy variable trap, also known as perfect multicollinearity. Reasons are as follows:

1. **The Dummy Variable Trap:**
   When we create dummy variables for a categorical feature, we introduce redundancy. If the categorical feature has n categories, we only need n-1 dummy variables to represent it fully. If we create n dummy variables, the model becomes perfectly multicollinear.

   Example: Suppose we have a feature "Season" with four categories: Spring, Summer, Fall, and Winter. If we create dummy variables for all four seasons, we'll have:

   `season_Spring`
   `season_Summer`
   `season_Fall`
   `season_Winter`

The problem is that if we know the values of three of these dummy variables, we automatically know the value of the fourth. For instance, if `season_Spring`, `season_Summer`, and `season_Fall` are all 0, then `season_Winter` must be 1. This perfect linear dependence creates the dummy variable trap.

2. **Consequences of the Dummy Variable Trap:**

- Unstable Coefficients: The model cannot uniquely estimate the coefficients for all the dummy variables. The coefficients become unstable and can vary wildly with small changes in the data.
- Inflated Standard Errors: Standard errors of the coefficients become very large, making it difficult to determine statistical significance.
- Model Interpretation Issues: It becomes impossible to isolate the effect of each individual category.

3. **How `drop_first=True` Solves the Problem:**

   By setting `drop_first=True`, we drop one of the dummy variables, effectively removing the redundancy. In the "Season" example, we might drop `season_Spring`. Now, we have:

   `season_Summer`
   `season_Fall`
   `season_Winter`

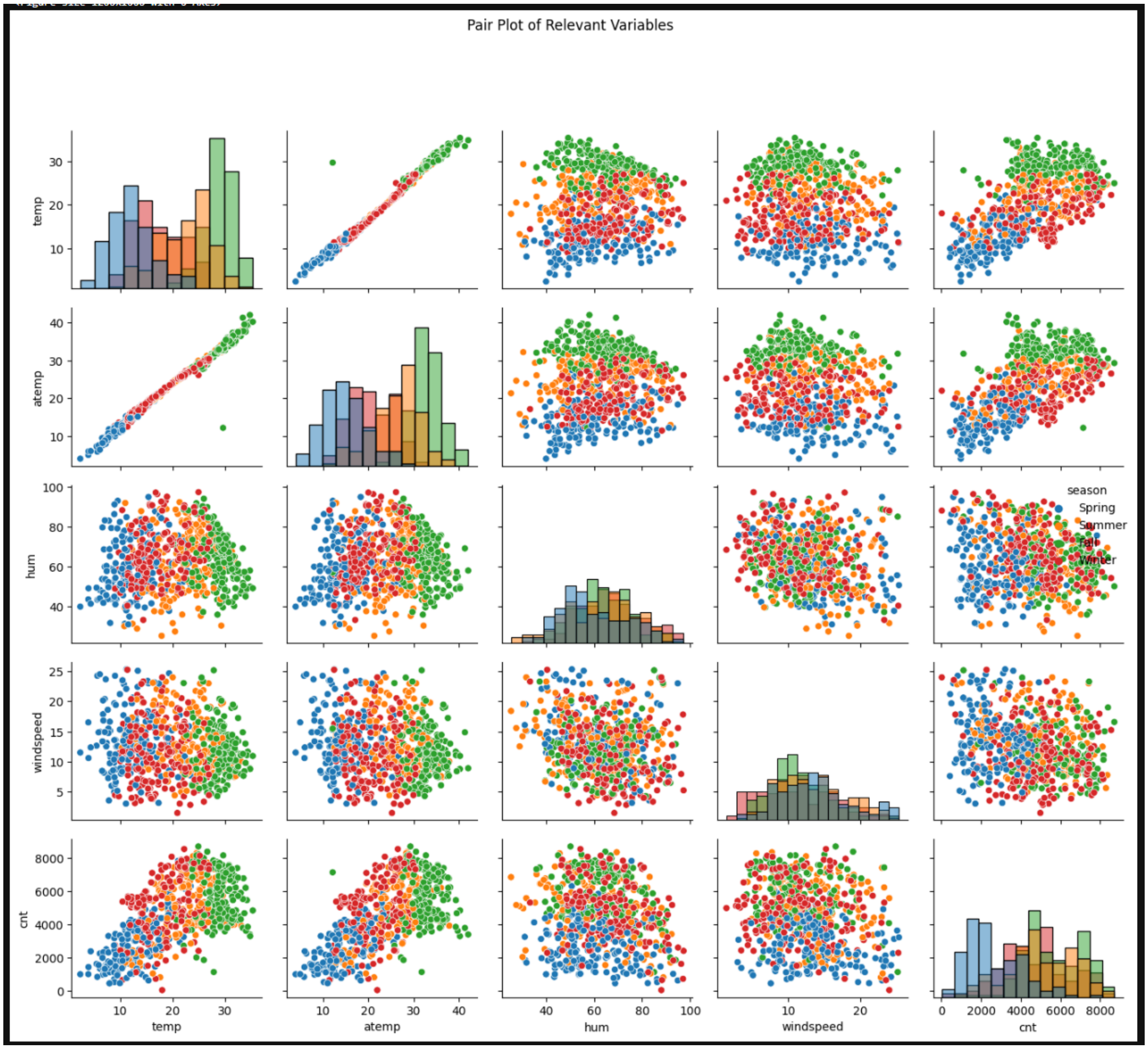
   If all three of these are 0, it implicitly means that the observation belongs to the dropped category (Spring). This removes the perfect multicollinearity and allows the model to estimate coefficients reliably.

4. **Interpretation with `drop_first=True`:**

   The coefficients of the remaining dummy variables are interpreted as the difference in the outcome variable compared to the dropped category (the reference category). So, `season_Summer` would represent the difference in bike rentals between Summer and Spring.

   In summary, using `drop_first=True` is essential for preventing multicollinearity issues when creating dummy variables, leading to more stable, interpretable, and statistically valid regression models.

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:



The pair plot shows **'temp' and 'atemp'** having the strongest positive linear correlations with cnt (bike rentals). It's difficult to definitively say which one is the highest just from visual inspection, as they appear very similar.
But as per the calculation of correlation coefficients (e.g., Pearson's correlation) both has similar and strong correlation with cnt.
(Since temp and atemp are so highly correlated, in the next step while doing analysis, 'atemp' was removed by me to avoid multicollinearity issues in the regression model.

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:

1. Linearity: We examined the pair plot, which helps visualize the relationship between the predictors and the target variable (`cnt`). While the pair plot provides an initial visual check, we would ideally create scatter plots of the predicted values vs. the residuals. A random pattern around the horizontal axis (no clear trend) indicates linearity. If patterns are present (e.g., a curve), it suggests a non-linear relationship, and transformations of variables or non-linear models might be needed.

2. Homoscedasticity (Constant Variance of Residuals): Similar to the linearity check, we'd examine the scatter plot of predicted values vs. residuals. A consistent spread of residuals across the range of predicted values indicates homoscedasticity. If the spread varies (e.g., a fanning or cone shape), it suggests heteroscedasticity. Addressing heteroscedasticity could involve transformations of the target variable or using weighted least squares regression.

3. Normality of Residuals: We use a combination of visual inspection (histograms and Q-Q plots) and statistical tests (like the Omnibus and Jarque-Bera tests reported in the model summary) to assess the normality of residuals. A normally distributed histogram and a Q-Q plot where points fall close to the diagonal line indicate normality. If the residuals deviate substantially from normality, it might affect the reliability of hypothesis tests. Transformations of the target variable or robust regression methods can sometimes help. We previously noted in the model summary that the Omnibus and Jarque-Bera tests indicated potential non-normality of the residuals, which is a point to address.

4. No or Little Multicollinearity: We used the Variance Inflation Factor (VIF) to check for multicollinearity among the predictor variables. As discussed earlier, while `temp` had a VIF slightly above 5, it was retained due to its importance. The other variables had acceptable VIF values. We also removed `atemp` due to its high correlation with `temp` to prevent multicollinearity.

5. Independence of Errors (No Autocorrelation): The Durbin-Watson statistic, reported in the model summary, helps assess autocorrelation. A value around 2 suggests no significant autocorrelation. If autocorrelation is present, it indicates that the errors are not independent, and techniques like autoregressive models might be more appropriate.

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:

Based on the provided OLS Regression Results, the top 3 features contributing most significantly to explaining the demand for shared bikes are:

1. yr_2019: This variable has the highest positive coefficient (0.2284) and a very low p-value (0.000), indicating a strong positive relationship with bike rentals. Being in the year 2019 significantly increases the predicted bike count compared to the other year (presumably 2018, which would be the reference category if `drop_first=True` was used for the `yr` variable).

2. temp: Temperature has the second highest positive coefficient (0.5353) and a p-value of 0.000. This indicates that higher temperatures are strongly associated with increased bike rentals.

3. season_Winter: This has a coefficient of 0.1630 and a p-value of 0.000. This suggests that during winter, bike rentals are significantly higher compared to the reference season (likely spring, if `drop_first=True` was used for season). This might seem counterintuitive, but it's important to remember that this is the effect of winter after controlling for temperature. It could be that other factors related to winter (e.g., fewer holidays, different working patterns) lead to increased bike usage even after accounting for the lower temperatures.

It's important to note that `weathersit_Bad` has a large negative coefficient (-0.3127), meaning that bad weather significantly reduces bike rentals. However, in terms of positive contributions to demand, the top three are `yr_2019`, `temp`, and `season_Winter`.

The correlation heatmap provides additional information about the relationships between the predictors. It confirms the strong positive correlation between `temp` and `cnt`, and shows that `yr_2019` also has a positive correlation with `cnt`. It also highlights the negative correlations between `windspeed`, `weathersit_Bad`, and `cnt`. However, the regression coefficients are more informative for determining the relative importance of predictors, as they account for the combined effects of all variables in the model.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer:

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (the target) and one or more independent variables (predictors) by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane in multiple linear regression) that minimizes the difference between the predicted values and the actual values.

Here's a breakdown of the algorithm:

1. Assumptions: Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
- Normality: The errors are normally distributed.
- No or Little Multicollinearity: Predictor variables are not highly correlated with each other.

2. The Linear Equation: The core of linear regression is the linear equation:

- For simple linear regression (one predictor): `$y = \beta_0 + \beta_1 x + \varepsilon$`
- For multiple linear regression (multiple predictors): `$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{\square} x_{\square} + \varepsilon$`

Where:
`$y$` is the dependent variable (target).
`$x_1, x_2, ..., x_{\square}$` are the independent variables (predictors).
`$\beta_0$` is the intercept (the value of `$y$` when all `$x$` are 0).
`$\beta_1, \beta_2, ..., \beta_{\square}$` are the coefficients (slopes) representing the change in `$y$` for a one-unit change in each corresponding `$x$`.
`$\varepsilon$` is the error term (the difference between the predicted and actual values).

3. Cost Function (Ordinary Least Squares): The goal is to find the values of the coefficients (β_0, β_1, ...) that minimize the error. The most common method is Ordinary Least Squares (OLS), which minimizes the sum of squared errors (SSE):

`$SSE = \Sigma(y_i - \hat{y}_i)^2$`

Where:

`$y_i$` is the actual value of the target for the i-th observation.
`$\hat{y}_i$` is the predicted value of the target for the i-th observation.

4. Optimization: Various optimization algorithms can be used to find the coefficients that minimize the SSE. The most common are:

- Gradient Descent: Iteratively adjusts the coefficients in the direction of the negative gradient of the cost function.
- Normal Equation: A closed-form solution that directly calculates the optimal coefficients using linear algebra.

5. Model Evaluation: Once the model is trained, its performance is evaluated using metrics such as:

- R-squared: Measures the proportion of variance in the target variable explained by the model.
- Adjusted R-squared: Similar to R-squared but penalizes the inclusion of unnecessary variables.
- Root Mean Squared Error (RMSE): Measures the average difference between the predicted and actual values.
- Mean Absolute Error (MAE): Another measure of prediction error, less sensitive to outliers than RMSE.

6. Prediction: After training and evaluation, the model can be used to predict the target variable for new data points by plugging the values of the predictors into the linear equation.

In summary, linear regression aims to find the best-fitting linear relationship between predictors and a target variable by minimizing the sum of squared errors. The resulting model can then be used for prediction and to understand the influence of different predictors on the target.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Anscombe's quartet is a set of four datasets that illustrate the importance of visualizing data and not relying solely on summary statistics. Each dataset consists of eleven (x, y) pairs, and they share several identical statistical properties, including:

Mean of x: 9
Variance of x: 11
Mean of y: 7.50
Variance of y: 4.125
Correlation between x and y: 0.816

Linear regression line: y = 3 + 0.5x

Despite these identical summary statistics, the datasets are visually very different. Here's a description of each:

1. Dataset I: This dataset shows a clear linear relationship between x and y, with the points scattered fairly evenly around the regression line. This is the kind of dataset that linear regression is well-suited for.

2. Dataset II: This dataset shows a clear non-linear relationship. The points follow a curve, not a straight line. While the calculated linear regression line still "fits" in terms of minimizing the sum of squares, it's clearly not an appropriate model for this data.

3. Dataset III: This dataset has a perfect linear relationship except for one outlier. This outlier significantly influences the regression line, pulling it upwards. This illustrates the sensitivity of linear regression to outliers.

4. Dataset IV: This dataset is even more extreme. All but one of the x values are the same. The single data point with a different x value completely determines the slope of the regression line. This demonstrates how a single influential point can drastically affect the results.

Key Takeaways from Anscombe's Quartet:

- Visualization is Essential: Summary statistics alone can be misleading. Visualizing the data is crucial for understanding the true relationship between variables and for identifying potential issues like non-linearity, outliers, or influential points.
- Don't Rely Solely on Statistical Tests: While statistical tests like R-squared and p-values are useful, they should not be interpreted blindly. They can be misleading if the underlying assumptions of the statistical method are violated.
- Explore Your Data: Before applying any statistical method, it's essential to explore the data thoroughly. This includes creating scatter plots, histograms, and other visualizations, as well as calculating descriptive statistics.

Anscombe's quartet serves as a powerful reminder that data analysis should be a holistic process that involves both statistical analysis and careful exploration of the data itself. Relying solely on numbers can lead to incorrect conclusions and misinterpretations.

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear association between two continuous variables. In simpler terms, it tells us how strongly two things are related, assuming that relationship is a straight line.

Here's a breakdown of what it means:

- Linear Association: It measures the linear relationship, meaning how well the data points fit a straight line. If the relationship is curved or non-linear, Pearson's r might not be the best measure.

- Strength: The value of Pearson's r ranges from -1 to +1.

    - +1: Indicates a perfect positive linear correlation. As one variable increases, the other increases proportionally.
    - 0: Indicates no linear correlation. There's no linear relationship between the variables.
    - -1: Indicates a perfect negative linear correlation. As one variable increases, the other decreases proportionally.

- Direction: The sign (+ or -) of Pearson's r indicates the direction of the relationship.

- Significance: While the value of r tells you the strength and direction, we also need to consider its statistical significance. A significant r value means the correlation is unlikely to have occurred by chance alone. This is usually determined using a p-value.

Example: If we're studying the relationship between hours of study and exam scores, a Pearson's r of +0.8 might indicate a strong positive correlation: more study hours are associated with higher scores. A Pearson's r of -0.2 might indicate a weak negative correlation, suggesting that there is a slight trend of lower scores with more study hours, although this is likely due to chance. A Pearson's r of near 0 would suggest no linear relationship. Keep in mind that correlation doesn't equal causation; even a strong correlation doesn't prove that one variable causes changes in the other.

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:

Scaling, in the context of data analysis and machine learning, refers to the process of transforming the features (variables) of a dataset to a similar range of values. It's a crucial preprocessing step that often significantly improves the performance of many machine learning algorithms.

Several reasons make scaling essential:

- Improved Algorithm Performance: Many algorithms, particularly distance-based algorithms like k-Nearest Neighbors (k-NN) and support vector machines (SVMs), are sensitive to the scale of features. If one feature has a much larger range of values than others, it will dominate the distance calculations, leading to inaccurate results. Scaling ensures that all features contribute equally to the algorithm's decision-making.

- Faster Convergence: Gradient descent-based algorithms (like those used in neural networks) often converge faster when features are on a similar scale. This is because the gradients are more balanced, preventing the algorithm from oscillating wildly during optimization.

- Improved Interpretability: Scaling can sometimes make it easier to interpret the results. For example, if all features are scaled to a 0-1 range, it becomes easier to compare their relative importance.

- Preventing Dominance of Certain Features: Features with larger magnitudes can overshadow features with smaller magnitudes, leading to biased results. Scaling mitigates this issue.

Normalized Scaling vs. Standardized Scaling:

Both normalized and standardized scaling are common techniques, but they achieve this in different ways:

1. Normalized Scaling (Min-Max Scaling):

- Method: This scales features to a specific range, usually between 0 and 1. The formula is:

  `x_scaled = (x - x_min) / (x_max - x_min)`

  where `x` is the original value, `x_min` is the minimum value in the feature, and `x_max` is the maximum value.

- Effect: It preserves the relative distances between data points. The smallest value becomes 0, and the largest becomes 1, with all other values proportionally scaled in between.

- When to use: Useful when we want to maintain the original distribution of the data and the relative magnitudes between data points are important.

2. Standardized Scaling (Z-score Normalization):

- Method: This scales features to have a mean of 0 and a standard deviation of 1. The formula is:

  `x_scaled = (x - μ) / σ`

  where `x` is the original value, `μ` is the mean of the feature, and `σ` is the standard deviation.

- Effect: It transforms the data to follow a standard normal distribution (approximately). Outliers might have a larger influence.

- When to use: Preferred when the data is not normally distributed or when outliers could significantly skew the results of min-max scaling. It's generally more robust to outliers than min-max scaling.

The best scaling method depends on the specific dataset and the machine learning algorithm being used. Experimentation is often necessary to determine which method works best.

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:

We've observed that sometimes the Variance Inflation Factor (VIF) can be infinite. This happens when there's perfect multicollinearity among the predictor variables in a regression model. Perfect multicollinearity means that at least one predictor variable is a perfect linear combination of other predictor variables. In simpler terms, one variable can be exactly predicted from the others.

The VIF is calculated as `1/(1-R²)`, where R^2 is the R-squared value from a regression of one predictor variable on all the other predictor variables. When there's perfect multicollinearity, the R^2 becomes 1, leading to a denominator of 0 and making the VIF infinite. This indicates a serious problem in the model because we can't separate the individual effects of the perfectly correlated predictors. The model becomes unstable and unreliable because we can't estimate the regression coefficients accurately. We need to address the multicollinearity, perhaps by removing redundant predictors or using techniques like Principal Component Analysis (PCA) before proceeding with the analysis.

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:
We use a Q-Q (quantile-quantile) plot to visually assess whether a dataset follows a particular theoretical distribution, often a normal distribution. It compares the quantiles of the data to the quantiles of the theoretical distribution.

In a Q-Q plot:

- The x-axis represents the quantiles of the theoretical distribution (e.g., the quantiles of a

standard normal distribution).
- The y-axis represents the quantiles of the observed data.

If the data follows the theoretical distribution, the points in the Q-Q plot will fall approximately along a straight diagonal line. Deviations from this line suggest departures from the theoretical distribution.

Use and Importance in Linear Regression:

In linear regression, we often assume that the residuals (the differences between the observed and predicted values) are normally distributed. This assumption is crucial for the validity of many statistical tests associated with linear regression, such as hypothesis tests for the regression coefficients. We use a Q-Q plot of the residuals to check this assumption:

1. Assessing Normality of Residuals: We create a Q-Q plot of the residuals. If the points fall close to the diagonal line, it suggests that the residuals are approximately normally distributed. Significant deviations from the diagonal line indicate non-normality.

2. Identifying Outliers: Points far from the diagonal line in the Q-Q plot might represent outliers in the data. These outliers can heavily influence the regression results.

3. Guiding Model Improvement: If the Q-Q plot reveals substantial non-normality, it suggests that the linear regression model may not be appropriate for the data. We might need to consider transformations of the response variable or predictors, use a different type of regression model (e.g., robust regression), or investigate potential outliers.

Importance: A Q-Q plot provides a valuable visual check for the normality assumption in linear regression. While other tests of normality exist, the Q-Q plot offers a clear graphical representation that allows us to quickly identify potential problems and assess the severity of deviations from normality. This helps ensure the reliability and validity of our linear regression model and its associated inferences.