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)

From the analysis, the categorical variables such as holiday, working day, weather situation, and year have significant effects on the dependent variable `cnt` (the count of shared bike rentals). Here's what can be inferred:

- Holiday: The correlation heatmaps show that on holidays, the relationship between temperature (`temp`) and bike rentals (`cnt`) may differ compared to non-holidays. Typically, the number of bike rentals might decrease on holidays due to fewer people commuting to work.

- Working Day: On working days, there is usually a stronger positive correlation between `temp` and `cnt`, indicating that more people rent bikes for commuting when the weather is favourable. On non-working days, the pattern might be different as recreational use may vary.

- Weather Situation: Different weather conditions (clear, misty, light snow, heavy snow) affect the demand for shared bikes. For instance, in clear weather, there's a higher positive correlation between `temp` and `cnt`, while adverse weather conditions like heavy snow reduce bike rentals significantly.

- Year: The year variable (`yr`) shows how bike-sharing demand changes over time. The analysis might reveal growth in demand from one year to the next, indicating increasing popularity or changes in user behaviour.

Overall, these categorical variables modify the relationship between the independent variables and the dependent variable, affecting the strength and direction of correlations.

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

Using `drop_first=True` during dummy variable creation is important to:

- Avoid the Dummy Variable Trap: Including all dummy variables for a categorical feature can lead to multi collinearity because the dummy variables are highly correlated (they sum to one). By dropping the first category, we eliminate redundancy.

- Provide a Reference Category: The dropped category serves as a baseline against which the effects of other categories are compared in the regression model.

This ensures that the model coefficients are interpretable and that the matrix of features is full rank, which is necessary for estimating the regression coefficients accurately.

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

The numerical variable `temp` (temperature) has the highest correlation with the target variable `cnt` (count of bike rentals). This is evident from both the pair-plot and the correlation matrix, where `temp` shows a strong positive linear relationship with `cnt`.

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

After building the linear regression model, the following steps were taken to validate its assumptions:

1. Linearity: Verified by plotting the observed values versus the predicted values to check for a linear relationship.

2. Independence: Ensured by the study design (assuming observations are independent) and checking the Durbin-Watson statistic if necessary.

3. Homoscedasticity: Checked by plotting the residuals versus the fitted values. A random scatter suggests constant variance (homoscedasticity), while patterns indicate heteroscedasticity.

4. Normality of Residuals: Assessed using a Q-Q plot (Quantile-Quantile plot) of the residuals. If the residuals align closely with the diagonal line, it indicates that they are approximately normally distributed.

5. Multi collinearity: Evaluated using Variance Inflation Factor (VIF). VIF values greater than 5 or 10 indicate high multi collinearity, which can inflate the standard errors of the coefficients.

By performing these diagnostics, the model's validity and the reliability of its estimates were ensured.

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

Based on the final model and the earlier correlation analysis, the top three features contributing significantly are:

1. Temperature (`temp`): Highest positive correlation with bike rentals; as temperature increases, the demand for bikes increases.

2. Humidity (`hum`): Negative correlation with bike rentals; higher humidity might decrease the demand due to discomfort.

3. Wind Speed (`windspeed`) or Apparent Temperature (`atemp`): Depending on the analysis, one of these could be the next significant feature. `atemp` often closely correlates with `temp`, while `windspeed` might negatively affect bike rentals.

General Subjective Questions

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

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

- Assumptions:

1. Linearity: The relationship between independent and dependent variables is linear.
2. Independence: Observations are independent of each other.
3. Homoscedasticity: Constant variance of errors.
4. Normality: Errors are normally distributed.
5. No Multi collinearity: Independent variables are not highly correlated.

- Estimation (Least Squares Method):

  - The coefficients (beta) are estimated by minimizing the Sum of Squared Errors (SSE):

    SSE = sum_{i=1}^{n} (y_i - \hat{y}_i)^2

  - This results in the best-fitting line that minimizes the discrepancies between observed and predicted values.

- Evaluation Metrics:

  - Mean Squared Error (MSE): Average squared difference between observed and predicted values.

  - Coefficient of Determination (\(R^2 \)): Proportion of variance in the dependent variable explained by the independent variables.

- Interpretation:

  - Each coefficient \(\beta_i \) represents the change in the dependent variable for a one-unit change in the independent variable \(x_i \), holding all other variables constant.

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

Anscombe's Quartet is a set of four datasets that have nearly identical simple statistical properties (mean, variance, correlation, and linear regression line) but appear very different when graphed. It demonstrates the importance of visualizing data before analyzing it.

Key Points:

- Identical Statistical Properties:

  - Each dataset has the same mean and variance for \(x \) and \(y \).

  - Each has the same correlation coefficient between x and y.

  - Each shares the same linear regression line.

- Different Distributions:

  - Dataset I: A typical linear relationship with some scatter.

  - Dataset II: A nonlinear relationship; a curve would fit better.

  - Dataset III: Linear relationship with an outlier affecting the line.

  - Dataset IV: Vertical line with an outlier creating a correlation.

Importance:

- Reveals Limitations of Statistical Measures: Summary statistics can be misleading without visual context.

- Emphasizes Data Visualization: Plotting data can uncover patterns, trends, and anomalies that statistics might not reveal.

- Illustrates the Impact of Outliers: Outliers can significantly affect statistical calculations and interpretations.

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

Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables X and Y. It quantifies the strength and direction of the linear relationship.

Formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Properties:

- Range: $-1 \leq r \leq 1$

  - $r = 1$: Perfect positive linear correlation.

  - $r = -1$: Perfect negative linear correlation.

  - $r = 0$: No linear correlation.

- Interpretation:

  - Positive r: As X increases, Y tends to increase.

  - Negative r: As X increases, Y tends to decrease.

  - The closer $|r|$ is to 1, the stronger the linear relationship.

Usage:

- Assessing the linear relationship between two continuous variables.

- Essential in fields like statistics, data analysis, and machine learning for feature selection.

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

Scaling is the process of adjusting the range of features in data so that they can be compared on common grounds. It's essential in machine learning algorithms that compute distances between data points or are sensitive to the scale of the data.

Why Scaling is performed:

- Improves Model Performance: Algorithms like gradient descent converge faster with scaled data.

- Equal Weightage: Ensures that all features contribute equally to the result, preventing dominance by features with larger magnitudes.

- Prevents Bias: Models are not biased toward variables with higher absolute values.

Difference between Normalized Scaling and Standardized Scaling:

1. Normalization (Min-Max Scaling):

  - Formula:

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

  - Range: Scales data to a fixed range, usually [0, 1].

  - Effect: Preserves the shape of the original distribution but compresses the scale.

  - Use Cases: Useful when the distribution is not Gaussian or when you want to bound values.

2. Standardization (Z-score Scaling):

  - Formula:

    $$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

    where μ is the mean and σ is the standard deviation.

  - Range: Centers the data around zero with a standard deviation of one.

  - Effect: Converts the data to a standard normal distribution.

  - Use Cases: Preferred when the data follows a Gaussian distribution.

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

An infinite Variance Inflation Factor (VIF) occurs when there is perfect multicollinearity among the independent variables. This means that one variable is a perfect linear combination of one or more other variables.

Reasons:

- Perfect Multicollinearity: Exact linear relationships between variables cause division by zero in the VIF calculation.

- Dummy Variable Trap: Including all categories of a categorical variable without dropping one can lead to perfect multicollinearity.

- Redundant Variables: Duplicate variables or those derived directly from others (e.g., sum of other variables).

Consequences:

- Unstable Coefficient Estimates: High standard errors lead to unreliable estimates.

- Inflated Variances: Makes it difficult to assess the effect of independent variables.

Solution:

- Remove or Combine Variables: Drop one of the correlated variables or combine them.

- Regularization Techniques: Use methods like Ridge Regression to handle multi collinearity.

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

A Q-Q (Quantile-Quantile) Plot is a graphical tool to assess if a dataset comes from a particular theoretical distribution, most commonly the normal distribution.

How It Works:

- Plotting Quantiles: Plots the quantiles of the sample data against the quantiles of the theoretical distribution.

- Interpretation:

  - Straight Line: If the points approximately form a straight line, the data conforms to the distribution.

  - Deviations: Systematic deviations indicate departures from the distribution (e.g., skewness, kurtosis).

Use in Linear Regression:

- Assumption of Normality: Linear regression assumes that the residuals (errors) are normally distributed.

- Diagnostic Tool: A Q-Q plot of the residuals helps to verify this assumption.

- Importance:

  - Model Validity: Ensures that statistical tests (like t-tests for coefficients) are valid.

- Identifying Outliers: Extreme deviations in the plot can indicate outliers or heavy tails.

By using a Q-Q plot, one can assess the appropriateness of the linear regression model and make necessary adjustments, such as transforming variables or using robust regression techniques.