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)

   Ans:

   ### yr (yr1_2019)

   #### Statistical Significance:

   The p-value for yr is 0.000, which is much less than the common alpha level of 0.05. This indicates that the year variable is highly statistically significant in predicting the total bike rentals (cnt).

   #### Effect Size:

   The coefficient for yr is 0.2404, meaning that, all else being equal, an increase in the year results in an increase of about 0.2404 units in the total bike rentals. This positive effect implies that bike rentals have increased over the years.

   #### Multicollinearity:

   The VIF value for yr is 1.94, which is below the common threshold of 5. This suggests that multicollinearity is not a concern for this variable.

   ### weathersit (weathersit1_Light Snow/Rain & weathersit1_Mist/Cloudy)

   #### Statistical Significance:

   The p-value for weathersit is 0.000, indicating that the weather situation is also highly statistically significant in predicting the total bike rentals.

   #### Effect Size:

   The coefficients for LIGHT SNOW and MIST/CLOUDY are -0.2449 & -0.0596 respectively, meaning that, all else being equal, a less favourable weather situation results in a decrease of units in the total bike rentals. This negative effect indicates that worse weather conditions lead to fewer bike rentals.

   #### Multicollinearity:

   The VIF value for weathersit is 1.06 & 1.42, which are below the threshold of 5, indicating that multicollinearity is not a significant concern for this variable.

   #### Summary

   Both yr and weathersit are significant predictors of bike rentals.

   yr has a positive effect on bike rentals, suggesting an increase in rentals over time.

   weathersit has a negative effect, indicating fewer rentals in less favourable weather conditions.

   The VIF values suggest that multicollinearity is not an issue for these variables.

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

Ans: Using `drop_first=True` when creating dummy variables is a best practice in regression analysis to avoid multicollinearity, improve model stability, and simplify the interpretation of model coefficients.

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

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

   ### Absence of Multicollinearity

   Checked if predictors are not highly correlated with each other. A value of Variance Inflation Factor (VIF) for each predictor of 5 or above indicate multicollinearity, meaning the independent variables are highly correlated.

   ### Normality of Residuals

   Performed residual analysis by plotting distribution of error terms (residuals) and checked if the graph normal and centred around zero.

   ### Patterns in error terms

   Checked if the residuals exhibit any pattern when plotted against training dataset or predicted values using scatter plot. There shouldn't be any.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Ans:
   atemp – coefficient of 0.5450
   weathersit – coefficient of -0.2449
   yr – coefficient of 0.2404

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear regression is a fundamental and widely used statistical method to model the relationship between a dependent variable (target or outcome variable) and one or more independent variables (predictors or features). The primary objective of linear regression is to establish a linear relationship between the dependent and independent variables by fitting a line that best represents the data points.
   Ans:

   ### Simple Linear Regression:

   - Models the relationship between two variables by fitting a linear equation to observed data.
   - Equation: $Y = \beta_0 + \beta_1 X + \varepsilon$
     - Y: Dependent variable
     - β_0: Intercept (constant term)
     - β_1: Slope (coefficient of the independent variable)

- X: Independent variable
- ε: Error term (residual)

Multiple Linear Regression:

- Models the relationship between one dependent variable and multiple independent variables.
- Equation: Y = beta_0 + beta_1 X_1 + beta_2 X_2 + … + beta_n X_n + ε
  - X_1, X_2, …, X_n: Independent variables
  - beta_1, beta_2, …, beta_n: Coefficients of the independent variables
  - ε: Error term (residual)

Steps in Linear Regression:

1. Data Collection:
   - Gather data containing the dependent variable and one or more independent variables.
2. Exploratory Data Analysis (EDA):
   - Visualize data using scatter plots, histograms, etc.
   - Summarize data using descriptive statistics.
   - Identify patterns, relationships, and anomalies.
3. Data Preprocessing:
   - Handle missing values.
   - Encode categorical variables.
   - Normalize or standardize numerical variables if needed.
   - Split data into training and test sets.
4. Model Specification:
   - Define the form of the regression model (simple or multiple).
5. Parameter Estimation:
   - Use methods like Ordinary Least Squares (OLS) to estimate the coefficients (beta_0, beta_1, …, beta_n).
   - OLS minimizes the sum of the squared differences between observed and predicted values.
   - Equation to minimize: $\sum$ (Y_i - Y_i_pred)^2)
     - Y_i: Actual value
     - Y_i_pred: Predicted value (Y_i_pred = beta_0 + beta_1 X_i1 + beta_2 X_i2 + … + beta_n X_in)
6. Model Fitting:
   - Fit the linear regression model to the training data.
7. Model Evaluation:
   - Assess the performance of the model using metrics like:
   - R-squared (R^2): Proportion of the variance in the dependent variable that is predictable from the independent variables.
   - Adjusted R-squared: Adjusted for the number of predictors in the model.
   - Mean Squared Error (MSE): Average of the squares of the errors.
   - Root Mean Squared Error (RMSE): Square root of MSE.
   - Perform residual analysis to check assumptions:
   - Linearity: Relationship between the predictors and the outcome is linear.

- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of residuals.
- Normality: Residuals are normally distributed.

8. Prediction:
   - Use the fitted model to make predictions on new data.

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

Ans:

Anscombe's Quartet consists of four datasets that are designed to illustrate the importance of graphing data before analysing it statistically. Each of the four datasets has nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. The quartet demonstrates how statistical properties can be misleading when not accompanied by visualizations.

Despite these identical statistical properties, the datasets are different in distribution and should be visualized to fully understand their nature.

The Four Datasets:

1. Dataset I:
   - A typical linear relationship between x and y.
   - The data points are close to the regression line.
2. Dataset II:
   - A clear nonlinear relationship.
   - The data points form a curve.
3. Dataset III:
   - An outlier affects the linear relationship.
   - One data point significantly deviates from the others, influencing the regression line.
4. Dataset IV:
   - Vertical alignment of points except for one point.
   - All data points except one have the same x value, and the outlier strongly influences the correlation and regression line.

Visualization and Analysis:

Dataset I:

- Description: Linear relationship with slight random noise.
- Plot: A scatter plot shows a clear linear trend with data points scattered around the regression line.

Dataset II:

- Description: Nonlinear relationship.
- Plot: A scatter plot reveals a clear curve, indicating a quadratic or other nonlinear relationship, which is not captured by the linear regression line.

Dataset III:

- Description: Linear relationship with an influential outlier.
- Plot: A scatter plot shows most data points forming a linear pattern, but one outlier significantly deviates from the line, affecting the slope and intercept of the regression line.

Dataset IV:

- Description: Vertical alignment of points with one outlier.
- Plot: A scatter plot shows almost all data points with the same x value and only one point with a different x value. This outlier heavily influences the correlation and regression line.

Importance of Anscombe's Quartet:

1. Illustrates the Limitations of Descriptive Statistics:
   - Shows that datasets can have identical statistical properties but different distributions.
   - Highlights the need for visualizing data to understand its true nature.
2. Demonstrates the Impact of Outliers:
   - Emphasizes how a single outlier can significantly affect the results of statistical analyses, such as correlation and regression.
3. Encourages Data Visualization:
   - Advocates for the use of scatter plots and other visual tools to complement statistical summaries.
   - Helps to identify patterns, relationships, and anomalies that descriptive statistics alone cannot reveal.
4. Promotes Critical Thinking in Data Analysis:
   - Encourages analysts to question and verify statistical findings by examining the underlying data.
   - Reminds that context and visual inspection are crucial in drawing accurate conclusions from data.

Conclusion:

Anscombe's Quartet serves as a powerful reminder that relying solely on summary statistics can be misleading. Visualizing data is essential for uncovering the true nature of relationships and patterns within the data, ensuring a more accurate and comprehensive analysis.

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

Pearson's R, also known as Pearson's 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 between the two variables. The coefficient is denoted as r and ranges from -1 to +1.

Ans:

Formula for Pearson's R:

$$\frac{\sum (Xi - X\_mu)(Yi - Y\_mu)}{\text{SQRT}(\sum (Xi - X\_mu)^2 \sum (Yi - Y\_mu)^2)}$$

Where Xi and Yi are the individual sample points.

X_mu and Y_mu are the means of X and Y respectively.

Limitations:

1. Linear Relationship: Pearson's R only measures linear relationships and does not capture non-linear relationships.

2. Outliers: Highly sensitive to outliers, which can distort the correlation.
3. Assumption of Normality: Assumes that the data is approximately normally distributed.
4. Homogeneity of Variance: Assumes that the variance of the variables is constant.

Conclusion:

Pearson's R is a fundamental tool in statistics and data analysis for quantifying the linear relationship between two variables. However, it should be used with caution, taking into account its assumptions and limitations, and always complemented with data visualization to get a complete understanding of the relationship between the variables.

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

Ans:

What is Scaling?

Scaling is the process of transforming the range of a feature to a standard scale, without distorting differences in the ranges of values. It's a crucial step in data preprocessing for many machine learning algorithms that compute distances between data points. Scaling ensures that features contribute equally to the model's learning process.

Why is Scaling Performed?

1. Improving Model Performance:
   Algorithms that use gradient descent optimization (like linear regression and neural networks) converge faster with scaled features.
2. Ensuring Comparability:
   Different features in a dataset may have different units and scales (e.g., age in years, income in dollars). Scaling brings all features to the same level, making the model's coefficients more interpretable.
3. Preventing Dominance:
4. Features with larger ranges can dominate the learning process, leading to biased results. Scaling prevents this by ensuring no single feature overwhelms the model.

Types of Scaling

1. Normalized Scaling (Min-Max Scaling):

- Formula: $X_{\text{scaled}}=X_{\text{max}}-X_{\text{min}}X-X_{\text{min}}$
- Range: Transforms the data to a fixed range, usually [0, 1].

2. Standardized Scaling (Z-Score Normalization):

- Formula: $X_{\text{scaled}}=\sigma X-\mu$
- Range: Transforms the data to have a mean of 0 and a standard deviation of 1.

Key Differences between Normalized Scaling and Standardized Scaling:

1. Range:

- Normalized Scaling: Transforms data to a fixed range, typically [0, 1].
- Standardized Scaling: Transforms data to have a mean of 0 and a standard deviation of 1.

2. Sensitivity to Outliers:

- Normalized Scaling: Sensitive to outliers since it depends on the minimum and maximum values in the data.

- Standardized Scaling: Less sensitive to outliers but still affected; outliers can influence the mean and standard deviation.

3. Use Cases:

- Normalized Scaling: Preferred when you want to preserve the relationship between data points and know the data is bounded.

- Standardized Scaling: Preferred when the data is normally distributed or when using algorithms that assume data is centred around zero.

Conclusion:

Scaling is a critical preprocessing step in machine learning that ensures features contribute equally to the learning process. Normalized scaling and standardized scaling are the two primary methods, each with its own use cases and characteristics. Choosing the right scaling method depends on the specific requirements of the machine learning algorithm and the nature of the data.

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

Ans: The value of VIF can become infinite due to the following reason:

Perfect Multicollinearity:

- If one predictor variable is a perfect linear combination of one or more other predictors, then R_i^2 will be 1.

- When R_i^2 = 1, the denominator of the VIF formula becomes zero (1 - R_i^2 = 0), leading to an infinite VIF.

- This indicates perfect multicollinearity, meaning that the predictor can be perfectly predicted from other predictors.

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

Ans: A Q-Q (Quantile-Quantile) plot is a graphical tool 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.

How to Interpret a Q-Q Plot

- Straight Line: If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately on a straight line.

- Deviation from Straight Line:

- Concave Curve: Indicates that the data has heavier tails than the theoretical distribution.

- Convex Curve: Indicates that the data has lighter tails than the theoretical distribution.

- S-Shape: Suggests that the data is skewed.

Importance of Q-Q Plot in Linear Regression

1. Normality Assumption:

- Linear regression assumes that the residuals (errors) of the model are normally distributed.

- A Q-Q plot can be used to check this assumption. If the residuals are normally distributed, the points will lie on a straight line.

2. Identifying Outliers:

   - Q-Q plots can help identify outliers in the data. Points that deviate significantly from the line indicate potential outliers.

Interpretation of the Q-Q Plot

- Straight Line: If the points lie on a straight line, the residuals are normally distributed.
- Deviations: Any significant deviations from the straight line suggest departures from normality, indicating potential issues with the model assumptions.

Conclusion

A Q-Q plot is a powerful diagnostic tool in linear regression to check the normality of residuals, identify outliers, and assess model assumptions. Ensuring that the residuals are normally distributed is crucial for the validity of the linear regression model and its statistical inferences.