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)**

The analysis of the categorical variables in the dataset reveals the following effects on the dependent variable (***cnt***), which represents the demand for shared bikes:

Season:

- ***Spring*** (negative coefficient): Demand for shared bikes decreases during the spring season compared to the reference season.
- ***Summer*** (positive coefficient): Demand increases during the summer season.
- ***Winter*** (positive coefficient): Demand also increases during the winter season compared to the reference season (presumably autumn or fall).

Month:

- **July** (negative coefficient): Demand decreases in July.
- **September** (positive coefficient): Demand increases in September.

Weekday:

- **Sunday** (negative coefficient): Demand is lower on Sundays compared to the reference day.
- **Saturday** (not included in the final model): If included, it would indicate the effect of Saturdays on demand. Since it's not in the final model, it might have had a negligible or redundant effect.

Weather Situation:

- **Light Rain or Light Snow** (negative coefficient): Demand significantly decreases during light rain or light snow conditions.
- **Misty and Cloudy** (negative coefficient): Demand also decreases under misty and cloudy conditions, but to a lesser extent than during light rain or snow.

Holiday (negative coefficient): Demand for shared bikes tends to be lower on holidays compared to non-holidays.

Year (positive coefficient): Demand for shared bikes increased in the subsequent year compared to the base year.

Inferences:

- **Seasonality**: There is a clear seasonal trend where summer and winter see an increase in bike demand, while spring sees a decrease.
- **Monthly Variation**: Certain months, like July, see a dip in demand, whereas September sees a rise, possibly due to weather conditions, vacation periods, or other factors.
- **Weekly Patterns**: Demand tends to be lower on Sundays, which might be due to fewer people commuting to work or school.

- **Weather Conditions:** Adverse weather conditions such as rain, snow, and cloudy skies significantly reduce bike demand, indicating that people tend to avoid biking in such conditions.
- **Holiday Effect:** Shared bike usage decreases on holidays, likely due to reduced commuting needs.
- **Yearly Trend:** There is an overall increase in demand for shared bikes over the years, suggesting growing popularity or increased adoption of bike-sharing services**.**

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

   Using **drop_first=True** during dummy variable creation is important to avoid the issue of multicollinearity in regression models.

   Multicollinearity

   Multicollinearity occurs when two or more predictor variables in a model are highly correlated, making it difficult to determine the individual effect of each predictor on the response variable. In the context of dummy variables, it can lead to a situation called the "***dummy variable trap***."

   Dummy Variable Trap

   While creating dummy variables for a categorical feature with ***k***-categories, it will end up with ***k***-dummy variables (one for each category). If all the k-dummy variables are included in the regression model along with an intercept term, the dummy variables will be perfectly collinear. This means one dummy variable can be perfectly predicted from the others, leading to multicollinearity.

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

   Based on pair-plot, ***"temperature"*** has the highest correlation with target variable.

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

   After constructing the linear regression model on the training set, I validated its assumptions through the following steps:

   1. **Linearity of the Relationship:**
      - Utilized scatter plots and residual plots to check.
      - The observed versus predicted values should ideally form a straight line.
      - Residuals should show no clear pattern when plotted against predicted values.

2. **Independence of Errors:**
    - Confirmed using the Durbin-Watson test.
    - A Durbin-Watson statistic near 2 indicates no autocorrelation in the residuals.
3. **Homoscedasticity:**
    - Assessed by plotting residuals versus predicted values.
    - Residuals should have a constant spread across all levels of predicted values, without any funnel-shaped patterns.
4. **Normality of Errors:**
    - Evaluated using histograms.
    - Residuals should follow a bell-shaped distribution in histograms.
5. **No Multicollinearity:**
    - Checked using the Variance Inflation Factor (VIF).
    - VIF values exceeding 10 (or sometimes 5) suggest significant multicollinearity among predictors.

      By systematically validating these assumptions, I ensure the linear regression model's reliability and validity, thereby increasing confidence in its predictions and inferences.

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

The top three features significantly contributing to the demand for shared bikes, based on their coefficients or importance scores in the final model, are:

- **Temperature (temp):** With a coefficient of 0.491, temperature has the most significant positive impact on bike demand. This indicates that as temperature rises, the demand for shared bikes also increases.
- **Year (yr):** With a coefficient of 0.233, the year variable shows a positive impact on bike demand over time. This suggests that bike-sharing demand has been growing annually.
- **Weather Situation - Light Rain or Light Snow (weathersit_Light Rain or Light Snow):** Although negatively correlated with a coefficient of -0.290, this weather condition still significantly influences bike demand. It indicates a decrease in bike demand during light rain or snow.

These features are the most significant contributors to bike demand based on their coefficients in the final linear regression model.

General Subjective Questions

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

Linear regression is a fundamental algorithm in machine learning used for predicting a continuous target variable based on one or more input features.

Linear Regression Algorithm

1. **Objective:** Linear regression aims to model the relationship between a dependent variable Y and one or more independent variables X. The goal is to find the best-fitting line (or hyperplane) that predicts Y from X by minimizing the error.

2. **Mathematical Model:** For a simple linear regression (one independent variable), the model can be represented as:

$$Y=\beta0+\beta1X1+\epsilon$$

Y – Dependent variable
X - Independent variable
β1 – *slope of the line*
β0 – *intercept of the line*
ϵ – *error term (difference between predicted and actual values)*

For multiple linear regression (more than one independent variable), the model extends to:

$$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$$

X1,X2,...,Xn are the independent variables.
β1,β2,...,βn are the coefficients for these variables.

3. **Fitting the Model:** The parameters β0,β1,...,βn are estimated using the method of ***Least Squares***, which minimizes the sum of squared differences between the observed values and the values predicted by the model.
The cost function for this is:

$$J(\beta) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}))^2$$

where ***m*** is the number of observations, and yi is the actual value of Y for the i-th observation.

4. **Evaluation:**
Coefficient of Determination (R^2): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.

Mean Squared Error (MSE): Measures the average of the squares of the errors. Lower values indicate better model performance.

5. **Assumptions**: Linear regression makes several assumptions about the data
   Linearity: The relationship between the independent and dependent 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 of Errors: The residuals (errors) are normally distributed.

6. **Application:** Linear regression is used in various fields such as finance, economics, and social sciences for tasks like forecasting, risk assessment, and trend analysis.

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but different distributions and appearances when plotted. It was introduced by Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it.

Overview:

1. **Datasets:** Anscombe's quartet consists of four datasets, each containing 11 data points with two variables: x and y. Despite having similar summary statistics, the datasets show different patterns when plotted.
2. **Statistical Properties:** All four datasets have nearly identical properties:
   - Mean of x: 9.0
   - Mean of y: 7.5
   - Variance of x: 11.0
   - Variance of y: 4.12
   - Correlation between x and y: 0.816
   - Regression Line: The least-squares fit line for each dataset has the same slope (0.5) and intercept (3.0).
3. **Differences in Plots:**
   - **Dataset 1**: Shows a linear relationship with a clear, straight line fit.
   - **Dataset 2**: Also shows a linear relationship, but with a vertical outlier that significantly affects the fit.
   - **Dataset 3**: Shows a nonlinear relationship with a curve, though it still has the same summary statistics.
   - **Dataset 4**: Displays a clear outlier that distorts the fit line, demonstrating how a single point can impact regression analysis.

4. **Importance**:
   Anscombe's quartet highlights that statistical summaries alone can be misleading and that visualizing data through plots is crucial for understanding underlying patterns. It emphasizes that different datasets can have identical statistical properties but different distributions, which can significantly affect analysis and conclusions.

Real-World Implications:

Data Visualization: It emphasizes the importance of visualizing data to gain deeper insights beyond summary statistics.

Statistical Analysis: It cautions against drawing conclusions based solely on summary statistics, as they may not capture the full complexity of the data.

Decision Making: In real-world scenarios, decision-makers should not rely solely on statistical summaries but should also consider visual representations to understand the underlying patterns in the data.

Anscombe's quartet demonstrates that relying solely on statistical measures without examining the data visually can lead to incomplete or incorrect interpretations. It underscores the importance of using data visualization techniques to reveal the true nature of the data.

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

Pearson's R, also known as the ***Pearson correlation coefficient***, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of their association.

Pearson's R is a statistical measure that calculates the degree to which two variables move in relation to each other. It is represented by the symbol rrr and ranges from -1 to 1.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where:

- x_i and y_i are individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

Interpretation:

- **r = 1**: Perfect positive linear correlation; as one variable increases, the other variable increases proportionally.

- **r = -1**: Perfect negative linear correlation; as one variable increases, the other variable decreases proportionally.
- **r = 0**: No linear correlation; changes in one variable do not predict changes in the other variable.

Pearson's R assumes:

- **Linearity**: The relationship between the variables is linear.
- **Normality**: Both variables are normally distributed.
- **Homoscedasticity**: The variance of the errors is consistent across all levels of the independent variable.

Limitations:

- **Nonlinearity**: Pearson's R is not suitable for capturing nonlinear relationships.
- **Outliers**: Sensitive to outliers, which can significantly affect the correlation coefficient.

Pearson's R is used in various fields, including social sciences, finance, and medicine, to understand and quantify the linear relationships between variables, guide predictions, and inform further analysis.

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

Scaling is a preprocessing technique in data analysis and machine learning that adjusts the range and distribution of features to make them more suitable for modeling. It ensures that different features contribute equally to the analysis, especially when they have different units or scales.

Purpose of scaling:

- **Uniformity**: Brings features to a common scale, which is crucial for algorithms sensitive to the magnitude of features (e.g., gradient descent-based methods).
- **Performance**: Enhances the performance and convergence speed of many machine learning algorithms.
- **Interpretability**: Ensures that features with different units or ranges do not disproportionately affect the model.

Types of Scaling

1. **Normalized Scaling**: Rescales the feature values to a specific range, typically [0, 1]
    - Useful when the goal is to transform features to a bounded range, which can be particularly useful for algorithms that rely on distances (e.g., k-nearest neighbors).
    - All feature values are adjusted proportionally between the minimum and maximum values

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

2. **Standardized Scaling (Z-score Normalization)**: Transforms the data to have a mean of 0 and a standard deviation of 1.
    - Useful when features need to be centered around zero with unit variance. This is particularly important for algorithms that assume data is normally distributed (e.g., linear regression, logistic regression)
    - The feature values are rescaled based on their statistical properties, which makes the distribution of data more Gaussian-like

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

where μ is the mean of the feature, and σ is the standard deviation.

Key Differences:

- **Range**:
    - **Normalized Scaling**: Transforms data to a fixed range (e.g., [0, 1]).
    - **Standardized Scaling**: Centers data around 0 with a standard deviation of 1, so the data has no fixed range.
- **Impact on Distribution**:
    - **Normalized Scaling**: Maintains the original shape of the distribution but within a specific range.
    - **Standardized Scaling**: Alters the shape of the distribution to have zero mean and unit variance, which may normalize skewed distributions.

Application:

- **Normalized Scaling**: Often used for algorithms requiring bounded input ranges or when features are in different units.

- **Standardized Scaling**: Commonly used for algorithms that rely on distance metrics or assume normally distributed data.

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

The ***Variance Inflation Factor (VIF)*** measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity. A high VIF indicates that a feature is highly correlated with other features in the model. When the value of VIF is infinite, it usually points to a serious issue in the data.

Reasons for Infinite VIF:

- **Perfect Multicollinearity:** It occurs when one predictor variable is an exact linear combination of other predictor variables. In such cases, the correlation between features is perfect, leading to an undefined or infinite variance for the coefficient estimates**.**

  Ex: If two features where one is a direct multiple of the other (e.g., X2=2×X1), the VIF for one of these features will be ***infinite***.
- **Singular Matrix:** In the context of regression, the design matrix (matrix of features) becomes singular if it has linearly dependent columns. A singular matrix cannot be inverted, which is required for calculating the VIF. The inversion failure leads to an infinite VIF because the variance of the regression coefficients cannot be computed.
- **Redundant Features:** Redundant features are variables that provide duplicate or redundant information. When these features are included in the model, they lead to collinearity issues, which can cause the VIF to become very large or infinite.

  Ex: Including both temperature in Celsius and temperature in Fahrenheit as features in the same model would cause perfect multicollinearity.

Simply, an ***infinite VIF*** typically results from perfect multicollinearity, where predictor variables are linearly dependent on each other, leading to a singular matrix and failure in the calculation of variance inflation. Identifying and addressing such issues usually involves removing or combining redundant features to ensure that the predictors are linearly independent.

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 used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset's distribution with the quantiles of a theoretical distribution.

- A Q-Q plot is a scatter plot where each point represents a quantile of the sample data plotted against the corresponding quantile of the theoretical distribution.
- The x-axis represents the quantiles of the theoretical distribution, while the y-axis represents the quantiles of the empirical data.

Construction of QQ Plot:

Empirical Quantiles: Calculate the quantiles of the sample data.

Theoretical Quantiles: Calculate the quantiles from the theoretical distribution (e.g., normal distribution).

Plot Points: Plot each pair of empirical and theoretical quantiles on the Q-Q plot.

Interpretation:

Straight Line: If the points lie approximately on a straight line (usually the line y=x), the sample data is approximately normally distributed.

Curvature: Deviations from the line suggest deviations from normality. For example, an S-shaped curve indicates heavy tails, while a bowing shape indicates light tails.

Use and Importance in Linear Regression

1. **Assumption Checking:**
   - **Normality of Residuals**: A Q-Q plot of the residuals can help verify this assumption.
   - **Model Diagnostics**: If residuals are not normally distributed, it can indicate that the model is not appropriately capturing the underlying data structure, potentially requiring transformations or different modeling approaches.
2. **Model Validity:**
   - **Goodness of Fit**: Assessing the normality of residuals helps ensure that the assumptions of linear regression are met, leading to more reliable inferences and predictions.
   - **Influence of Outliers**: Outliers or influential points can distort the residual distribution. A Q-Q plot can help identify such points by showing discrepancies from the normal distribution.
3. **Transformation Decisions:**
   - **Improving Normality**: If the residuals are not normally distributed, transformations (e.g., logarithmic, square root) might be applied to the dependent variable or features to better meet the normality assumption.
4. **Diagnostics for Statistical Tests:**

- **Accuracy of Statistical Inferences**: Many statistical tests and confidence intervals in linear regression assume normally distributed residuals. Checking this assumption helps ensure the validity of these tests.

Simply, A Q-Q plot is a valuable diagnostic tool in linear regression for checking the normality of residuals, which is a key assumption of the model. By visualizing how the empirical quantiles of residuals compare to theoretical quantiles, it helps assess model validity, detect deviations from normality, and guide appropriate model adjustments or transformations.

Submitted: Uma Maheswari K