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)

   Answer: We can infer the following:

   a. **Season**: The highest bike demand is in the fall, followed by summer and winter, with spring showing the lowest demand.
   b. **Weather Situation**: Most bike counts occur in clear weather with a total of rides, significantly outpacing the counts during misty/cloudy conditions and light snow/rain. This indicates that clear weather conditions are strongly preferred for bike usage, while adverse weather conditions like light snow or rain lead to drastically lower demand for bike-sharing services.
   c. **Holidays**: Most bike counts occur on days that are not holidays.
   d. **Working day:** Fewer people are booking bikes on weekends and holidays compared to weekdays.
   e. **Weekday:** The bike counts across weekdays show relatively consistent usage.
   f. **Month:** Bike demand shows a clear upward trend from January to August, peaking in the summer months (June to August). After August, there is a gradual decline, with demand dropping by December. This pattern suggests that bike usage is highest during the warmer months (spring and summer) and decreases during the colder months, likely due to weather conditions.

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

   Answer: Using **drop_first=True** in dummy variable creation helps prevent multicollinearity by avoiding the **dummy variable trap**. When creating dummies for a categorical variable with **n** categories, you get **n** dummy variables. However, including all **n** can make one variable perfectly predictable from the others, leading to redundancy.

   By dropping the first category, you create **n−1** dummy variables, with the dropped category serving as the **reference**. This ensures the model remains stable and interpretable, with each dummy variable showing the effect relative to the reference category.

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

   Answer: The numerical variable **temperature (temp)** shows the strongest correlation with the target variable **bike rentals (cnt).**

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

Answer:
1. Linearity: Checked if residuals are randomly scattered around zero. A clear pattern indicates non-linearity.
2. Independence of Errors: Checked for autocorrelation in residuals.
3. Homoscedasticity: Looked for constant variance. If the spread of residuals changes with fitted values, heteroscedasticity may be present. A consistent spread across the residual plot indicates homoscedasticity.
4. Normality of Residuals: Visualized the distribution of residuals. They are following a normal distribution.
5. No Multicollinearity: Calculated VIF for each predictor. Values above 5-10 indicate multicollinearity issues.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:
1. Temperature (temp): With a coefficient of **0.5181**, it has the strongest positive impact, indicating that higher temperatures greatly increase bike demand.
2. Year: The coefficient of **0.2325** suggests a positive trend in bike usage over time, showing that demand has been increasing from 2018 to 2019.
3. Weather Situation (Light Snow/Rain): With a coefficient of **-0.2872**, this feature significantly reduces demand, indicating that adverse weather conditions discourage bike usage. Riders prefer clear weather for bike riding.

General Subjective Questions

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

Answer:
Linear regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (target). Here's a detailed explanation of the linear regression algorithm:

1. Basic Concept:

- The goal of linear regression is to find the best-fitting linear relationship between the independent variables (X) and the dependent variable (Y). This relationship is typically represented as:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$$

- Here, Y is the predicted value, β0 is the intercept, β1,β2,...,βn are the coefficients for each predictor X1,X2,...,Xn, and ϵ represents the error term.

2. Assumptions:

Linear regression relies on several key assumptions:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: The residuals (errors) are independent of each other.
- **Homoscedasticity**: The residuals have constant variance at all levels of the independent variables.
- **Normality**: The residuals are normally distributed (important for inference).
- **No multicollinearity**: Independent variables should not be too highly correlated with each other.

3. Model Fitting:

- **Objective**: The goal is to find the coefficients (β\betaβ) that minimize the difference between the observed values and the values predicted by the model.
- **Cost Function**: This is usually done by minimizing the **Mean Squared Error (MSE)**, defined as:

$$\text{MSE} = \boxed{\frac{1}{n}\sum_{i=1}^{n}}\boxed{(Y_i - \hat{Y}_i)}\boxed{^2}$$

Mean Error Squared

- **Optimization**: Techniques like **Ordinary Least Squares (OLS)** are used to find the best coefficients. OLS finds the coefficients that minimize the MSE.

4. Gradient Descent:

- In some implementations, especially with larger datasets, **gradient descent** may be used to optimize the coefficients iteratively. The algorithm adjusts the coefficients in the direction that reduces the cost function, using the gradient (the slope of the cost function).

5. Model Evaluation:

- After fitting the model, it's important to evaluate its performance:
    - **R-squared**: Measures the proportion of variance in the dependent variable explained by the independent variables.
    - **Adjusted R-squared**: Adjusts R-squared for the number of predictors, preventing overfitting.

- **Residual Analysis**: Analysing residuals helps check the assumptions of the model.

6. Prediction:

- Once the model is trained and validated, you can use it to make predictions on new data:

$$Y\hat{}=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn$$

7. Interpretation:

- Each coefficient $\boldsymbol{\beta i}$ represents the change in the dependent variable $\boldsymbol{Y}$ for a one-unit increase in the corresponding predictor $\boldsymbol{Xi}$, holding all other variables constant.

8. Limitations:

- **Linearity Assumption**: If the relationship is not linear, linear regression may perform poorly.
- **Sensitivity to Outliers**: Linear regression can be heavily influenced by outliers.
- **Overfitting**: Using too many predictors can lead to overfitting, especially with limited data.

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

Answer:
Anscombe's quartet is a set of four datasets that illustrate the importance of data visualization and the limitations of relying solely on statistical summary measures. Developed by the statistician Francis Anscombe in 1973, the quartet consists of four distinct datasets that have nearly identical statistical properties but differ significantly in their distributions and relationships. Here's a detailed breakdown:

1. Structure of Anscombe's Quartet:

The quartet includes four datasets, each containing 11 pairs of (x, y) values. The datasets are typically labeled as **Anscombe's I, II, III, and IV**. Here are their key characteristics:

- **Identical Summary Statistics**:
    - For all four datasets, the following statistics are the same:
        - Mean of x: 9
        - Mean of y: 7.5

- Variance of x: 11
- Variance of y: 4.125
- Correlation between x and y: 0.816
- Linear regression line: $y=3+0.5xy = 3 + 0.5xy=3+0.5x$

2. Visualization:

Despite sharing the same statistical properties, the datasets exhibit distinct patterns when plotted. This difference emphasizes the necessity of visualizing data before drawing conclusions. Here's a brief overview of each dataset:

- **Dataset I**:
  - Appears to show a strong linear relationship between x and y. The points are roughly distributed around a straight line.
- **Dataset II**:
  - Also shows a linear relationship, but one outlier significantly affects the y-value for a specific x-value. When the outlier is removed, the relationship appears more linear.
- **Dataset III**:
  - This dataset has a quadratic relationship. Although it appears to have a correlation coefficient similar to the others, the relationship is clearly non-linear.
- **Dataset IV**:
  - Here, the data points form a vertical line with an outlier. The presence of the outlier skews the correlation and linear regression, misleadingly suggesting a strong relationship.

3. Importance:

Anscombe's quartet serves several critical purposes in statistics and data analysis:

- **Data Visualization**:
  - It highlights the importance of plotting data. While summary statistics might suggest a similar relationship, visual inspection reveals vastly different patterns.
- **Understanding Relationships**:
  - It teaches that correlation does not imply causation. In datasets II, III, and IV, the apparent linear correlation masks underlying relationships.

- **Modeling Decisions**:
    - The quartet reminds analysts to choose appropriate models based on the data's characteristics. Linear regression may not be suitable for all datasets, particularly those with non-linear relationships.

4. Applications:

The insights from Anscombe's quartet are widely applicable in fields such as:

- **Statistics and Data Science**: Emphasizing the need for exploratory data analysis (EDA) before modeling.
- **Machine Learning**: Reminding practitioners to visualize data to understand relationships and potential non-linear patterns.
- **Research and Reporting**: Encouraging researchers to accompany statistical summaries with visualizations to provide a clearer understanding of the data.

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

Answer:
Pearson's R, or the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables.

Key Features:

1. **Range**:
    - R ranges from −1 to 1:
        - **+1**: Perfect positive correlation
        - **-1**: Perfect negative correlation
        - **0**: No correlation
2. **Interpretation**:
    - Values close to 1 or −1 indicate a strong relationship; values near 000 suggest a weak relationship. The sign indicates the direction.
3. **Assumptions**:
    - Assumes linearity, normality of data, and homogeneity of variance.
4. **Limitations**:
    - Sensitive to outliers and only captures linear relationships.

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

Answer:
Scaling is a technique used to adjust the range of independent variables (features) so they fall within a specific range, often to improve the performance of machine learning algorithms. Some algorithms are sensitive to the scale of the features, and scaling ensures all features contribute equally.

Why is Scaling Performed?

- **Improves Model Performance**: Algorithms like gradient descent, k-NN, SVM, and neural networks are sensitive to feature scales. Without scaling, features with larger ranges can dominate.
- **Speeds Up Convergence**: For optimization algorithms (like gradient descent), scaling helps speed up convergence by ensuring the algorithm moves evenly along all dimensions.
- **Prevents Numerical Instability**: With unscaled data, some algorithms may suffer from instability, particularly when large values lead to overflow.

Difference Between Normalized and Standardized Scaling:

1. **Normalized Scaling**:
    - **Definition**: Normalization scales the data to a range of [0,1] (or [−1,1]), based on the min and max values of each feature.
    - **When to Use**: Useful when the distribution is not Gaussian (normal) or when you know the bounds of your data.
    - **Formula**: Adjusts values based on min-max scaling (finds where each value stands between the minimum and maximum).
2. **Standardized Scaling**:
    - **Definition**: Standardization scales the data to have a mean of 0 and a standard deviation of 1, making the data follow a standard normal distribution.
    - **When to Use**: Best when the features have a Gaussian-like distribution or for algorithms that assume normality, like linear regression or logistic regression.
    - **Formula**: Adjusts values based on their z-scores (how far they are from the mean in terms of standard deviations).

Example Use:

- **Normalization** is often used in constrained data spaces like image pixel values.
- **Standardization** is more common for general predictive models that assume normally distributed data.

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:
   The **Variance Inflation Factor (VIF)** measures how much the variance of a regression coefficient is increased due to multicollinearity among predictors. A VIF value of infinity typically occurs in the following situations:
   1. Perfect Multicollinearity:
- This happens when one predictor variable is a perfect linear combination of other predictor(s). For example, if you have two variables X1and X2such that X2=2X1, the correlation is perfect, leading to infinite VIF for either variable.
   2. Redundant Variables:
- Including a variable that is highly correlated with one or more existing variables can cause VIF to spike. This redundancy results in perfect or near-perfect linear relationships, inflating the variance.
   3. Singular Matrix:
- In cases where the design matrix (matrix of predictors) is singular (not invertible), it indicates that there is perfect multicollinearity. This results in undefined or infinite VIF values because the regression model cannot be accurately estimated.
   Implications:
   When you encounter infinite VIF values, it signals that the model may suffer from multicollinearity, which can lead to unreliable coefficient estimates and difficulties in interpreting the effects of predictors.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer:
A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, typically the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution.

Uses of a Q-Q Plot:

1. **Assessing Normality**:

- Q-Q plots are primarily used to check if the residuals of a regression model are normally distributed. This is an important assumption for many statistical tests and regression models.

2. **Identifying Deviations**:
   - The plot helps visualize deviations from normality. If the points closely follow the reference line, the data is likely normally distributed. Deviations from the line can indicate skewness or the presence of outliers.

Importance in Linear Regression:

1. **Model Assumptions**:
   - Linear regression assumes that the residuals (errors) are normally distributed. A Q-Q plot provides a quick visual check of this assumption, helping to validate the model's appropriateness.
2. **Influence of Outliers**:
   - It can help identify outliers or influential points that may affect the regression results. Significant deviations from the line could indicate problematic data points.
3. **Enhancing Model Reliability**:
   - Ensuring that the residuals are normally distributed increases the reliability of hypothesis tests (e.g., t-tests for coefficients) and confidence intervals generated from the model.
4. **Informing Transformations**:
   - If the residuals are not normally distributed, a Q-Q plot can indicate whether a transformation (e.g., log or square root) might be necessary to meet the assumptions of linear regression.