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?

- The fall season experiences a high number of bike rentals.
- June, August, and October have a high number of bike rentals.
- People prefer to rent bikes when the weather is good.
- People tend to avoid renting bikes on weekends.
- People do not prefer renting bikes on holidays.
- Bike rentals were higher in 2019 than in 2018.

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

- To avoid multicollinearity: Creating dummy variables for categorical data without excluding one category can lead to multicollinearity. By omitting one category, we prevent this issue because the excluded category's information is implicitly contained within the remaining categories.
- Interpretability: Omitting one category (usually the reference category) enhances the interpretability of the model coefficients, making them more intuitive.
-

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

- The registered variable has a high correlation with the target variable, cnt.

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

- We utilized residual analysis to validate the assumptions of linear regression.
- We plotted a histogram of the error terms and found that the "Error Distribution" is normally distributed around 0, indicating that our model has appropriately handled the assumption of error normality.
- Assumption of Independent Error Terms: There is virtually no relationship between the residuals and the predicted values.
- Homoscedasticity: The variance appears consistent at both ends of the fitted line.

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

The top 3 features are:

- **temp**
- **weathersit_bad**
- **yr**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm: A Brief Explanation

Linear regression is a statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. The goal is to predict the value of the dependent variable based on the values of the independent variables.
Key Points:

1. Equation: The basic linear regression model is represented by the equation:
   - Simple Linear Regression (one predictor): $y=β0+β1x+ϵ$
   - **Multiple Linear Regression** (multiple predictors): $y=β0+β1x1+β2x2+\cdots+βnxn+ϵ$
   - Here, y is the dependent variable, x1,x2,…,xn are the independent variables, β0 is the intercept, β1,β2,…,βn are the coefficients, and ϵ\epsilonϵ is the error term.

2. Assumptions:
   - Linearity: The relationship between the dependent and independent variables is linear.
   - Independence: The residuals (errors) are independent.
   - Homoscedasticity: The residuals have constant variance.
   - Normality: The residuals are normally distributed.

3. Steps:
   - Data Collection and Preparation: Gather and preprocess data.
   - Model Specification: Choose the dependent and independent variables.
   - Parameter Estimation: Use methods like Ordinary Least Squares (OLS) to estimate the coefficients that minimize the sum of squared residuals.
   - Model Fitting: Fit the model using the estimated coefficients.
   - Validation: Check model assumptions and evaluate performance using metrics like R-squared, Adjusted R-squared, MSE, and RMSE.
   - Interpretation: Understand the meaning of the coefficients in the context of the data.
   - Prediction: Use the model to make predictions on new data.

4. Model Evaluation:
   - R-squared: Indicates the proportion of variance in the dependent variable explained by the independent variables.
   - Adjusted R-squared: Adjusts the R-squared value for the number of predictors.
   - Residual Analysis: Examine residual plots to check for independence, homoscedasticity, and normality.
   - Goodness-of-Fit: Evaluate how well the model fits the data using error metrics.

Linear regression is a powerful yet simple tool for predictive modeling, providing insights into relationships between variables while ensuring interpretability and ease of use.

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a set of four distinct datasets that have nearly identical statistical properties yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. Here are the key points about Anscombe's Quartet:

Key Properties
Each of the four datasets in Anscombe's Quartet has the following nearly identical statistical properties:

Mean of x and y: The mean values of the x and y variables are the same across all datasets.
Variance: The variance of the x and y variables are the same across all datasets.
Correlation: The correlation coefficient between x and y is almost identical for all datasets.
Regression Line: The linear regression line $y=β0+β1x$ is similar for all datasets, with almost identical intercepts and slopes.
Residuals: The sum of the squared residuals is nearly the same for all datasets.

The Four Datasets
Dataset I: Appears as a linear relationship with some scatter, fitting well with a linear regression model.
Dataset II: Forms a clear curve, indicating a non-linear relationship which a linear model cannot capture accurately.
Dataset III: Contains an outlier which significantly influences the linear regression line, skewing the interpretation.
Dataset IV: All data points are aligned except for one extreme outlier, which drastically affects the correlation and regression results.
Importance
Visual Examination: Anscombe's Quartet underscores the critical role of visualizing data. Despite identical statistical summaries, the datasets reveal different patterns, relationships, and outliers when plotted.
Model Appropriateness: It highlights the importance of choosing the right model for data analysis. Blindly applying statistical models without visual inspection can lead to misleading conclusions.
Outlier Influence: Demonstrates how outliers can disproportionately affect statistical measures and the importance of identifying and understanding their impact.
Conclusion
Anscombe's Quartet serves as a powerful reminder that statistical metrics alone can be insufficient for understanding data. Visualization is an essential tool in data analysis, providing insights that numbers alone may not reveal.

3. What is Pearson’s R?

"Pearson's r," also called a Pearson correlation coefficient is a statistic that quantifies the strength and direction of the linear relationship between two continuous variables. It measures how well the data points of two variables fit on a straight line. Pearson's correlation coefficient ranges from -1 to 1.

1. When r is close to 1, it indicates a strong positive linear relationship. This means that as one variable increases, the other tends to increase as well.

2. An r value of 0 suggests no linear relationship between the variable

3. When r is close to -1, it indicates a strong negative linear relationship. This means that as one variable increases, the other tends to decrease, and vice versa.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a preprocessing technique in data analysis and machine learning that transforms the features (variables) of a dataset to a common scale or range. It is primarily done to address issues related to the differing scales of variables, which can affect the performance of various machine learning algorithms.

Variables in a dataset may have different measurement units and scales. Some variables may have values in a small range, while others may have values in a much larger range. Scaling ensures that all variables contribute equally to the analysis or modelling process.

1. In normalized scaling, the data is scaled to a specified range, typically [0, 1]. This is done by subtracting the minimum value of the variable from each data point and then dividing by the range. normalized scaling is useful when you want to preserve the original range of the data, and you're not concerned about the distribution's shape.

2. In standardized scaling, the data is transformed to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable from each data point and dividing by the standard deviation.

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

A VIF of infinity can occur when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be exactly predicted from a linear combination of the other independent variables.

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

The Q-Q plot is designed to help you visually compare the quantiles of your data to the quantiles of a theoretical distribution, which can reveal deviations from the expected distribution.
a Q-Q plot is a valuable tool for assessing the distribution of data, especially in the context of linear regression. It helps evaluate the normality assumption, detect skewness and outliers, and guide model improvement if deviations are observed.