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)

Analysis reveals a seasonal variation in bike rental counts, with peaks during spring and summer, followed by a decline in fall and winter.

- A year-over-year comparison indicates an uptick in bike rentals in 2019 compared to 2018.
- The period from June to September marks the highest demand for bike rentals, whereas January experiences the lowest.
- Holiday periods see a reduction in bike rental demand compared to non-holiday times.
- Weekday analysis shows negligible variation in bike rental demand.
- The distinction between working and non-working days does not significantly affect bike rental demand.
- Weather conditions significantly influence bike rental counts, with the highest numbers observed during clear or partly cloudy weather, followed by misty conditions, and lower counts during light snow or rain.

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

Employing drop_first=True is crucial as it prevents the generation of redundant columns when creating dummy variables, thereby minimizing the risk of multicollinearity among these variables.

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

Temperature (temp) and feels-like temperature (atemp) exhibit a strong positive correlation, suggesting they convey overlapping information.

- The variables representing total bike count, casual, and registered users are also highly positively correlated.

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

The model's efficacy is assessed using the R-squared value, or Coefficient of Determination, which stands at 0.814. This indicates that 81.4% of the variance in the dependent variable can be explained by the model's independent variables.

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 primary factors driving the demand for shared bikes, as per the final model, include temperature, weather conditions (weathersit), and the year.

General Subjective Questions

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

Linear regression is a fundamental algorithm in statistics and machine learning for predicting a quantitative outcome. It's widely used due to its simplicity and interpretability. The goal of linear regression is to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Simple Linear Regression

In simple linear regression, there is one independent variable and one dependent variable The goal is to find the line that best fits the data points. The line is described by the equation

.

Multiple Linear Regression

Multiple linear regression extends simple linear regression to include two or more independent variables. This allows for more complex models that can capture relationships between multiple predictors and the target variable.

Finding the Best-Fit Line

The process of "fitting" the regression line to the data involves finding the values of the coefficients that minimize the difference between the observed values and the values predicted by the model. This is typically done using a method called **Least Squares**, which minimizes the sum of the squared differences between observed and predicted values.

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven points (x, y) and was constructed by the statistician Francis Anscombe in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Let's delve into the details:

The Quartet

The four datasets in Anscombe's quartet are usually labeled I, II, III, and IV. Despite their differences in distribution and appearance when graphed, they share the following statistical properties (approximately):

Mean of x: The average value of the x-variable is about the same across all four datasets.
Mean of y: The average value of the y-variable is also nearly the same across all datasets.
Variance of x: The variance (a measure of spread) of the x-variable is similar for each dataset.
Variance of y: The y-variable also has a similar variance across the datasets.
Correlation between x and y: The correlation (a measure of the strength and direction of a linear relationship) between x and y is the same in all four datasets.
Linear regression line: When a linear regression model ($y = mx + b$) is fitted to each dataset, the slope (m) and y-intercept (b) of the resulting line are the same.
Coefficient of determination (R^2): This statistic measures the proportion of the variance for the dependent variable that's predictable from the independent variable. It is the same for all datasets.

3. What is Pearson’s R?

Pearson's r, also known as the Pearson product-moment correlation coefficient (PPMCC) or simply the Pearson correlation coefficient, is a measure of the linear correlation between two variables X and Y
. It quantifies the degree to which a relationship between two variables can be described by a linear equation. The value of Pearson's r
ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship, and
- **0** indicates no linear relationship between the variables.

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

Scaling is a crucial preprocessing step in data analysis and machine learning that involves adjusting the scale of the data so that different features (variables) contribute equally to the analysis, model training, or distance computations. This process is essential when the data set contains features with different units of measurement or scales because most machine learning algorithms perform better or converge faster when features are on a relatively similar scale.

Why is Scaling Performed?

Scaling is performed for several reasons:

Algorithm Efficiency: Many algorithms that use distance calculations, such as k-nearest neighbors (KNN) and k-means clustering, are sensitive to the scale of the data. Scaling ensures that all features contribute equally to the result.
Improved Gradient Descent: Algorithms that use gradient descent as an optimization technique (e.g., linear regression, neural networks) converge faster when features are on the same scale.
Balanced Feature Contribution: Ensures that no single feature dominates the model due to its scale, allowing the model to learn more evenly from all features.
Requirement Fulfillment: Some algorithms, like Support Vector Machines (SVM) and Principal Component Analysis (PCA), explicitly require features to be scaled for optimal performance.

Normalized Scaling vs. Standardized Scaling

Although both normalization and standardization are scaling techniques, they differ in their methods and outcomes:

Normalization (Min-Max Scaling):

- Normalization rescales the features to a fixed range, typically 0 to 1, without distorting differences in the ranges of values or losing information about outliers.

Standardization (Z-score Normalization):

- Standardization rescales data to have a mean (μ) of 0 and a standard deviation (σ) of 1. The outcome of standardization is that the features will be centered around the mean, with a unit standard deviation.

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

VIF sometimes is infinite when an independent variable is perfectly linearly correlated with one or more independent variables. Since $\mathrm{R2}$ would be equal to 1 in this case (indicating perfect prediction), the denominator in the VIF formula becomes 0 ($1-\mathrm{R2}=01-R_2=0$), making the VIF infinite. This situation indicates that the variable can be perfectly predicted by the other variables in the model, violating the assumption that independent variables should be independent of each other.

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

A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a set of data potentially came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution. It compares the quantiles of the data to the quantiles of the theoretical distribution, providing a visual means to evaluate how well the data matches the expected distribution.

Structure of a Q-Q Plot

- **X-axis (Theoretical Quantiles):** Represents the quantiles from the theoretical distribution. If assessing normality, these would be the quantiles from the standard normal distribution.
- **Y-axis (Sample Quantiles):** Represents the quantiles from the data being tested.
- **Plot Points:** Each point on the plot represents a quantile from the dataset matched with the corresponding quantile from the theoretical distribution.

Using Q-Q Plots in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to assess the normality of the residuals. Residuals are the differences between the observed values and the values predicted by the regression model. The assumptions of linear regression include the idea that these residuals are normally distributed. A Q-Q plot can help visually confirm this assumption or identify deviations from it.

Importance of Q-Q Plots in Linear Regression

- **Normality Assessment:** One of the key assumptions of linear regression is that the residuals (errors) follow a normal distribution. A Q-Q plot is a direct way to visually check this assumption. If the points lie approximately along a straight line, the residuals can be considered normally distributed.
- **Identifying Skewness and Kurtosis:** Deviations from the straight line in a Q-Q plot can indicate skewness (if one tail is longer or fatter than the other) or kurtosis (if the tails are heavier or lighter than the normal distribution).
- **Outlier Detection:** Q-Q plots can also help in identifying outliers. Points that deviate significantly from the line might indicate outliers in the data.