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)
   - Bike rentals are more popular on weekdays, specifically Tuesday (3), Wednesday (4), and Friday (6).
     Bike rentals are more frequent during certain seasons, notably summer (2) and fall (3).
     Increased bike rentals are noted during specific weather conditions (1: Clear, Few clouds, Partly cloudy).
     Bike rentals exhibited higher numbers in the year 2019 (1) compared to 2018 (0).

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 multicollinearity in regression models. It helps prevent perfect multicollinearity by dropping one of the dummy variables, making the model more stable and interpretable.

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

   - It is observed that temperature one has the highest correlation with the target variable.

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

   - I followed a comprehensive approach to validate the assumptions of Linear Regression after constructing the model on the training set. Firstly, I computed the R-squared and adjusted R-squared values to assess the model's goodness of fit. Next, I calculated the Variance Inflation Factors (VIFs) to identify potential multicollinearity issues among predictors.
     Furthermore, I conducted thorough residual analysis on the training data. This involved plotting histograms of error terms to check for normality and evaluating the spread of residuals around zero. Additionally, I assessed homoscedasticity by examining the residuals' distribution across predicted values.
     To gain a holistic perspective, I executed model evaluation by visually comparing the predicted values (y_pred) against the actual values (y_test). This allowed me to gauge the model's predictive performance and understand how well it captures the underlying relationships.
     By systematically performing these steps, I rigorously validated the assumptions of Linear Regression, ensuring that linearity, independence, normality, and homoscedasticity were met, bolstering the reliability and credibility of the model's outcomes.

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 contributing significantly towards explaining the demand for shared bikes are:
     temp (temperature): With a coefficient of 0.4180, temperature has a substantial positive impact on bike rentals. As temperature increases, more people tend to rent bikes.
     yr (year): With a coefficient of 0.2349, the year has a notable positive influence on bike rentals. This suggests that bike rentals have increased over the years.
     Light_Snow&Rain: With a coefficient of -0.2989, the presence of light snow or rain has a significant negative effect on bike rentals. People are less likely to rent bikes during such weather conditions.
     These features have the highest coefficients in absolute value, indicating their stronger impact on predicting bike rental demand. Other features like workingday, windspeed, and seasonal variables (spring, summer, winter) also contribute significantly to explaining bike rental demand, as evident from their respective coefficients and p-values.

General Subjective Questions

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

Linear regression is a widely used statistical algorithm for modelling the relationship between a dependent variable and one or more independent variables. It is a fundamental technique in machine learning and statistics, finding applications in various fields, including finance, economics, healthcare, and social sciences. The primary goal of linear regression is to find the best-fitting straight line that represents the relationship between the variables.

To begin with linear regression, a dataset containing observations of the dependent variable and corresponding values of the independent variables is required. Each observation represents a unique data point, and the first step is to visualize the data through scatter plots or other graphical representations to gain insights into the relationship's nature. The scatter plots can provide a visual sense of how the dependent variable changes concerning the independent variables, which helps in identifying any underlying patterns or trends.

The linear regression model is mathematically represented as $Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn$, where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, and β0, β1, β2, ..., βn are the coefficients representing the slope and intercept of the line. The objective is to estimate these coefficients in a way that minimizes the difference between the predicted values (obtained from the model) and the actual values of the dependent variable in the dataset. This difference is known as the cost or loss, and one common cost function used in linear regression is the mean squared error (MSE) or the sum of squared errors (SSE).

To find the optimal coefficients, an optimization algorithm like gradient descent is employed. The gradient descent iteratively updates the coefficients by moving in the direction of steepest descent of the cost function. This process continues until convergence, where further updates do not lead to significant improvement in the cost. The choice of optimization algorithm and learning rate can affect the speed and accuracy of convergence, requiring careful tuning.

Model evaluation is crucial to ensure the model's performance and generalizability. Various metrics like R-squared (coefficient of determination) is used to assess how well the model fits the data. A higher R-squared value and lower error metrics indicate a better-performing model.

With the optimized model, new or unseen data with the independent variables can be fed into the model to predict the dependent variable's values. These predictions can be valuable for making informed decisions or understanding the behaviour of the dependent variable concerning changes in the independent variables.

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

Anscombe's quartet is a famous set of four datasets, each containing 11 paired data points, designed by the statistician Francis Anscombe in 1973. What makes these datasets intriguing is that despite having nearly identical summary statistics, such as means, variances, correlations, and linear regression coefficients, they exhibit vastly different patterns when graphed.

The quartet serves as a compelling reminder of the importance of data visualization in the data analysis process. While summary statistics can provide valuable insights, they may not reveal the full story hidden within the data. By visually inspecting the datasets and plotting them, we can uncover distinct patterns, including linear, quadratic, or even outlier-dominated relationships.

This powerful demonstration underscores the dangers of relying solely on summary statistics for drawing conclusions about data. Failing to visually explore the data can lead to misleading interpretations and erroneous assumptions. Instead, embracing exploratory data analysis through graphs and visualizations allows analysts to gain deeper insights into the data distribution, identify anomalies, and make more informed decisions in statistical modelling and hypothesis testing.

The implications of Anscombe's quartet extend beyond the field of statistics. It is a timeless reminder of the significance of data visualization in understanding complex datasets across various disciplines, including data science, machine learning, and decision-making. By incorporating visualizations into our data analysis toolkit, we can better comprehend the underlying data patterns, appreciate data nuances, and avoid potential pitfalls in drawing accurate and reliable conclusions. Ultimately, Anscombe's quartet stands as a testament to the power of visual exploration in extracting meaningful insights from data.

3. What is Pearson’s R? (3 marks)

Pearson's R, also known as Pearson correlation coefficient or simply correlation coefficient, is a statistical measure used to quantify the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where -1 represents a perfect negative correlation, 0 indicates no correlation, and 1 denotes a perfect positive correlation.

The coefficient is commonly used in data analysis to assess the degree of association between variables. A positive correlation implies that as one variable increases, the other tends to increase as well, while a negative correlation indicates that as one variable increases, the other tends to decrease. Pearson's R is a valuable tool for understanding the relationships between variables and is widely employed in various fields, including economics, social sciences, and machine learning.

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

Scaling is a data preprocessing technique used in various machine learning algorithms to bring all the features or variables to a similar numerical scale. It ensures that no single feature dominates the model due to its larger magnitude, and all features contribute equally during the learning process.

Scaling is performed to standardize the range of variables, which is crucial for algorithms that rely on distance calculations, such as k-nearest neighbours or gradient descent optimization in neural networks. By scaling the data, we prevent certain features from overshadowing others and enable the algorithm to converge faster and find better solutions.

There are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized scaling, also known as min-max scaling, scales the data to a specified range, usually between 0 and 1. It calculates the scaled value for each data point by subtracting the minimum value and dividing by the range (maximum value - minimum value) of the feature. Normalization is ideal when the data does not have a Gaussian distribution and has outliers.

2. Standardized scaling, also called z-score normalization, scales the data to have a mean of 0 and a standard deviation of 1. It calculates the scaled value for each data point by subtracting the mean and dividing by the standard deviation of the feature. Standardization is suitable when the data has a Gaussian distribution and when features have different units or scales.

In summary, scaling is performed to ensure all features have comparable scales in machine learning models. Normalized scaling rescales the data to a specific range, while standardized scaling standardizes the data to have a mean of 0 and a standard deviation of 1. The choice between the two depends on the nature of the data and the requirements of the machine learning algorithm being used.

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

The occurrence of infinite values in the Variance Inflation Factor (VIF) typically happens due to perfect multicollinearity in the data. VIF is a measure used to assess the level of multicollinearity between predictor variables in a multiple linear regression model. It quantifies how much the variance of a regression coefficient is inflated due to the correlation with other predictors.

When two or more predictor variables are perfectly correlated (i.e., their relationship can be expressed by an exact linear equation), it becomes impossible for the regression model to estimate the individual coefficients accurately. As a result, VIF becomes infinite for the variable involved in the perfect multicollinearity.

Perfect multicollinearity is problematic because it hinders the interpretation of the regression coefficients and can lead to unreliable and unstable model results. To address this issue, one needs to identify and handle the correlated variables appropriately, either by removing one of the correlated variables or through other feature engineering techniques like principal component analysis (PCA) or regularization methods. By doing so, the VIF values can be brought back to manageable levels, making the regression model more robust and interpretable.

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 plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution (usually a normal distribution). It helps visualize whether the data deviates from the expected distribution.

In linear regression, Q-Q plots are essential for validating the assumption of normality of the residuals. Residuals are the differences between the observed values and the predicted values from the regression model. If the residuals follow a normal distribution, the Q-Q plot will show the points roughly following a straight line, indicating a good fit to the assumed normal distribution.

Deviation from the straight line suggests non-normality in the residuals, which can have implications for the reliability of the linear regression model. Departures from normality might indicate that the model assumptions are violated, leading to biased estimates and unreliable inferences.

By using Q-Q plots in linear regression, we can detect potential issues with the model and, if needed, apply appropriate data transformations or consider different modelling approaches to achieve better results and reliable conclusions. It's a valuable tool to ensure the validity of regression assumptions and improve the model's performance.