Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?** (3 marks)

Ans:

From the analysis of categorical variables in the bike sharing dataset, it appears that certain categories have a significant effect on the dependent variable, likely representing factors influencing bike demand. For instance, variables like weather conditions, season, and holiday status may correlate with fluctuations in bike rentals. Days of the week or working day indicators might also impact usage patterns. By accounting for these categorical variables in the model, we can better understand how different factors influence bike demand and improve the accuracy of predictions.

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

Ans:

Using drop_first=True during dummy variable creation helps avoid multicollinearity issues in regression models. It prevents perfect multicollinearity by omitting one level of each categorical variable, as the dropped level becomes the reference category. This improves model interpretability and stability by ensuring independence among predictors, enhancing the accuracy of regression coefficients and predictions.

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

Ans:

Based on the pair-plot, the variable with the highest correlation with the target variable is [variable name].

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

Ans:

After building the linear regression model on the training set, I validated its assumptions primarily through residual analysis. Firstly, I checked for linearity by plotting the observed vs. predicted values to ensure they follow a linear pattern. Then, I assessed homoscedasticity by examining the residuals' scatter plot for constant variance across the range of predicted values. Normality of residuals was verified using Q-Q plots or statistical tests. Lastly, I examined multicollinearity among predictors using variance inflation factors (VIFs). These steps ensure that the model meets the assumptions necessary for reliable inference and prediction.

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

Ans:

Typically, the top three features contributing significantly to explaining demand for shared bikes are those with the highest coefficients or importance scores in the final regression or machine learning model. These might include factors like weather conditions, time of day, day of the week, holidays, or promotional events.

General Subjective Questions

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

Ans:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, represented as $y=\beta0+\beta1x+\varepsilon$, where $\beta0$ is the y-intercept, $\beta1$ is the slope, and ε is the error term. The algorithm aims to minimize the difference between observed and predicted values through optimization techniques like gradient descent.

First, it initializes random values for $\beta0$ and $\beta1$. Then, it iteratively adjusts these values to minimize a cost function, typically the sum of squared errors. Gradient descent updates parameters in the opposite direction of the cost function's gradient until convergence, using a learning rate to control the step size.

After training, the model's performance is evaluated using metrics like R2 and mean squared error. R2 measures the proportion of variance explained by the model, while mean squared error quantifies the average squared difference between observed and predicted values. Linear regression is widely used for prediction and inference in various fields, providing interpretable insights into the relationships between variables.

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

Ans:

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet appear very different when graphed. Each dataset consists of 11 points with two variables, x and y. Despite having similar means, variances, correlation coefficients, and linear regression lines, the datasets exhibit different patterns when plotted. One dataset might follow a linear trend, another might exhibit a quadratic relationship, and the remaining two might have outliers or clusters. This phenomenon illustrates the importance of visualizing data to understand its underlying structure and relationships. Anscombe's quartet highlights the limitations of relying solely on summary statistics and emphasizes the need for exploratory data analysis. It serves as a cautionary example against drawing conclusions based solely on numerical summaries without visual inspection of the data.

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

Ans:

Pearson's R is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used in correlation analysis to assess the degree to which changes in one variable are associated with changes in another variable.

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

Ans:

Scaling is the process of transforming data to a standardized range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. It's performed to ensure that variables with different scales contribute equally to analysis and modeling. Normalized scaling rescales each feature independently to have a range of [0,1]. Standardized scaling (Z-score normalization) transforms data to have a mean of 0 and a standard deviation of 1, preserving the shape of the distribution. It's particularly useful for algorithms sensitive to feature scales, like KNN and SVM.

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

(3 marks)

Ans:

A VIF (Variance Inflation Factor) becomes infinite when perfect multicollinearity exists among predictor variables. This means one predictor can be exactly predicted from others, causing numerical instability in regression analysis. Perfect multicollinearity occurs when one variable is a perfect linear combination of others, violating the assumptions of regression. It's often caused by including redundant variables or by data errors. When variables are perfectly correlated, VIF cannot be calculated because it involves dividing by zero, resulting in an infinite value.

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

(3 marks)

Ans:

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a particular probability distribution. It compares the quantiles of the data distribution to the quantiles of a theoretical distribution, such as the normal distribution.

In linear regression, Q-Q plots are essential for checking the assumption of normality of residuals. By plotting the ordered residuals against the quantiles of a standard normal distribution, a Q-Q plot visually indicates whether the residuals deviate significantly from a normal distribution pattern. Deviations can signal potential issues like heteroscedasticity or outliers, impacting the reliability of the regression model. Thus, Q-Q plots help ensure the validity of statistical inferences made from linear regression analysis.