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?

A - The categorial variables played a vital role as their impact on the dependent variable is the highest. Change in these variables can significantly affect the dependent variable. Moreover, 'atemp', 'year', and 'month' among these categorical variables exhibit the highest correlation with the dependent variable, highlighting their importance in predictive modelling.

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

A - Using drop_first=True omits one level of each categorical variable in dummy variable creation, preventing multicollinearity and enhancing model interpretability by avoiding redundancy in the dataset, especially in regression analysis and machine learning models.

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

A – The highest correlation among the numeric variables is the ‘atemp’ variable.

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

A - I performed Residual Analysis on the data. Additionally, I plotted the histogram of the error terms to observe their distribution. While building the model, I set a threshold for the P-Value less than 0.05 and the VIF less than or equal to 5. Consequently, the model was built using only those specific variables.

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

A - Based on the final model, ‘temperature’, ‘year’, ‘month’ are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail?

A - Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets. Supervised learning has two types:

Classification: It predicts the class of the dataset based on the independent input variable. Like the image of an animal is a cat or dog?

Regression: It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

The algorithm tries to find the best-fitting line that minimizes the difference between the predicted values and the actual values observed in the training data. This is achieved by adjusting coefficients through a process called ordinary least squares(OLS), where the algorithm calculates the optimal values for these coefficients to minimize the sum of squared errors between the predicted and actual values.

Once the model is trained, it can make predictions for new data points by applying the learned coefficients to the input variables. Linear regression is widely used due to its simplicity and interpretability.

2. Explain the Anscombe's quartet in detail?

A - Anscombe's quartet comprises four datasets with identical statistical properties but different visual representations when plotted. Each dataset consists of 11 pairs of x and y values. When plotted, three of the datasets appear to follow a linear relationship, while the fourth exhibits a non-linear pattern. Despite these apparent differences, all four datasets have the same mean, variance, correlation coefficient, and regression line when calculated using traditional statistical methods.

Despite having the same mean, variance, correlation coefficient, and regression line, the datasets exhibit distinct patterns when graphed, ranging from linear to non-linear relationships. This demonstration underscores the necessity of visualizing data to uncover underlying patterns and outliers, supplementing traditional statistical analysis for a more comprehensive understanding.

3. What is Pearson's R?

A - Pearson's correlation coefficient (often denoted as Pearson's r) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It 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.

Pearson's r is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used in various fields, including statistics, economics, psychology, and biology, to assess the degree of association between two variables and to identify patterns in data.

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

A - Scaling is a preprocessing technique used in data analysis and machine learning to transform the features of a dataset to a similar scale. Scaling is performed to ensure that features with larger magnitudes do not dominate those with smaller magnitudes during model training.

Normalized scaling rescales features to a specified range, often between 0 and 1, preserving their relative relationships. Standardized scaling transforms features to have a mean of 0 and a standard deviation of 1, centering the data around the mean and scaling it by the standard deviation.

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

A - In regression analysis, the Variance Inflation Factor (VIF) measures the degree of multicollinearity among predictor variables. When the VIF is infinite, it indicates that the presence of perfect multicollinearity, wherein one predictor variable can be exactly predicted by a linear combination of other predictor variables in the model.

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

A - The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. In linear regression, a straight line in the plot indicates that residuals follow a normal distribution, ensuring the validity of regression analysis.

The importance of a Q-Q plot lies in its ability to detect departures from normality in the residuals, which can affect the validity of statistical inference and predictions made by the regression model.