Assignment-based Subjective Questions

Q.1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Ans: e would first examine the distribution of satisfaction levels within each type of product to see if there are any noticeable patterns.
- Next, we could conduct a chi-square test of independence to determine if there's a significant association between the type of product and satisfaction level.
- Additionally, we might visualize the relationship using stacked bar plots to see how satisfaction levels vary across different product types.
- Finally, we could calculate effect size measures such as Cramer's V to quantify the strength of the association between the type of product and satisfaction level.

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

- **Multicollinearity**: When you have dummy variables representing categorical variables with more than two categories, including all dummy variables in the model can lead to multicollinearity. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other. This can cause issues with the estimation of coefficients and can make the interpretation of the model less reliable.
- **Dummy Variable Trap**: Including all dummy variables in the model without dropping one can lead to the dummy variable trap. The dummy variable trap occurs when the presence of redundant dummy variables (i.e., when all dummy variables are included) causes perfect multicollinearity. This can result in the model being unidentifiable and can lead to incorrect estimation of coefficients.

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

- In my experiment, I found that variables such as humidity (**Hum**), apparent temperature (**atemp**), and the binary indicators for January (**jan**), working day (**workingday**), and Saturday (**Saturday**) exhibit the highest correlation with the dependent variable labeled as 'high' or 'VIP'.
- Upon further analysis, these variables demonstrate strong associations with the outcome of interest compared to other variables in the dataset.

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

- Validating linear regression assumptions post-training is crucial for model reliability. Steps include analyzing residuals for randomness, checking normality, verifying homoscedasticity, confirming linearity, assessing residual independence, and identifying outliers and influential points.

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

- **Temperature (temp)**: It has the highest positive coefficient of 0.6651, indicating that an increase in temperature leads to a significant increase in bike demand.
- **Year (yr)**: With a coefficient of 0.2408, the "yr" feature suggests that bike demand has been increasing over the years.
- **Winter**: The "winter" feature has a coefficient of 0.1444, indicating that the winter season has a positive effect on bike demand, possibly due to holiday activities or seasonal preferences.

General Subjective Questions

Q.1. Explain the linear regression algorithm in detail.

- **Purpose**: Models the relationship between a dependent variable and one or more independent variables.
- **Assumption**: Assumes a linear relationship between variables.
- **Parameter Estimation**: Estimates coefficients to minimize the difference between observed and predicted values.
- **Model Evaluation**: Assessed using metrics like $R2$ and mean squared error.
- **Assumption Checking**: Ensures linearity, independence, homoscedasticity, and normality of residuals.
- **Interpretation**: Coefficients indicate strength and direction of relationships between variables.

Q.2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of four datasets, each containing 11 (x, y) pairs, designed to have nearly identical simple descriptive statistics. Despite their statistical similarities, the datasets exhibit very different patterns when visualized.
- This highlights the importance of data visualization and the limitations of relying solely on summary statistics for understanding data relationships.

Q.3. What is Pearson's R?

Pearson's *r* is a measure 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.

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

- Scaling is the process of transforming data to a common scale to ensure that different features have similar ranges or distributions. It's performed to make the features comparable and prevent features with larger magnitudes from dominating those with smaller magnitudes during model training.
- **Normalized Scaling**: In normalized scaling, also known as min-max scaling, the data is scaled to a fixed range, typically between 0 and 1. Each feature's values are transformed proportionally, preserving the original distribution.
- **Standardized Scaling**: In standardized scaling, also known as z-score normalization, the data is transformed to have a mean of 0 and a standard deviation of 1. This method maintains the shape of the original distribution while cantering the data around the mean and adjusting its spread.

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

- IF becomes infinite when there is perfect multicollinearity among the independent variables in the regression model. This happens when one variable can be perfectly predicted by a linear combination of others, leading to singularities in the computation of VIF due to a zero determinant in the matrix.

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

- Q-Q plot is a graphical tool used to assess whether a dataset follows a specific probability distribution, such as the normal distribution. In linear regression, Q-Q plots are important for checking the normality assumption of residuals. They provide a visual comparison between the observed quantiles of residuals and the expected quantiles under the assumption of normality, helping to diagnose deviations from this assumption.