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?**

Ans:

- Seasonal Effects (e.g., spring, winter, fall, summer): Different seasons can significantly impact bike rental behaviours. Typically, seasons with milder weather like summer and fall might show higher rental rates due to more favourable biking conditions.
- Monthly Effects (e.g., January, December, July, November, September): Specific months might reflect unique seasonal characteristics not fully captured by broad seasonal categories. For instance, December and January might see fewer rentals due to colder weather and holiday periods, while September could experience an increase, possibly due to favourable weather and the start of the school season or return to work routines after summer holidays.
- Weather Conditions (e.g., clear, light snow, mist + cloudy): Weather conditions play a crucial role. Clear weather typically boosts bike rentals as seen in broader data trends and model's positive coefficient for clear days. Adverse weather conditions like light snow and mist or cloudy weather are likely to deter bike rentals, as reflected by the negative coefficients.

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

Ans:

Using drop_first=True during dummy variable creation helps avoid multicollinearity (where variables are too highly correlated), simplifies the model by reducing the number of variables, and improves interpretability by setting a baseline category for comparison.

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

Ans:

The temp variable 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?**

Ans:

Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

Ans:

The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

General Subjective Questions

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

Ans:

Linear regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a continuous dependent variable (outcome). It assumes that there is a linear relationship between the independent variables and the dependent variable. In simple linear regression, there is only one independent variable, whereas in multiple linear regression, there are multiple independent variables.

Assumptions of Linear Regression:

Linear regression relies on several assumptions, including:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The errors are independent of each other.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
4. Normality: The errors are normally distributed.
5. No multicollinearity: The independent variables are not too highly correlated with each other.

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

Ans:

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics (e.g., mean, variance, correlation) but exhibit markedly different patterns when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and to caution against relying solely on summary statistics.

The Datasets:

1. Dataset I:
   - Contains linear relationships between variables.
   - Has a correlation coefficient of approximately 0.816.
   - Simple linear regression fits well.
2. Dataset II:
   - Similar to Dataset I but with one outlier.
   - The outlier significantly affects the slope and correlation.
   - Demonstrates the influence of outliers on regression analysis.
3. Dataset III:
   - Exhibits a non-linear relationship between variables.
   - Regression analysis may not capture the underlying pattern.
4. Dataset IV:
   - Appears to have no relationship between variables.
   - The correlation coefficient is close to zero.
   - However, there is a perfect relationship between X and Y in subsets of the data, leading to misleading conclusions if not visualized properly.

3. What is Pearson's R?

Pearson's correlation coefficient is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables.

Pearson's R ranges from -1 to 1.

- A correlation of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A correlation of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A correlation of 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?**

Ans:

Scaling is a preprocessing technique used in machine learning to standardize the range of independent feature variables in a dataset. It involves transforming the values of the features so that they all fall within a similar scale or range.

Scaling is performed in machine learning to ensure that all feature variables in the dataset are on a similar scale. This is done to prevent features with larger magnitudes from dominating the learning algorithm and to improve the performance, stability, and interpretability of the model.

Normalization scales data to a fixed range, typically between 0 and 1, preserving the original distribution. Standardization transforms data to have a mean of 0 and a standard deviation of 1, making it less sensitive to outliers and suitable for algorithms assuming Gaussian-distributed data.

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

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped to define a working model for regression.

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

Ans:

The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any

distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.