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?

   From the analysis of categorical variables like season, weather situation, month, and weekday, the below observations are made

   - Season: - Summer and fall typically predict higher rentals compared to winter and spring, indicates seasonal impact on bike-sharing demand.
   - Weather: - Clear weather had higher bike rentals.
   - Month: - warmer months had higher bike rentals
   - Weekdays: - Weekdays had higher bike rentals vs weekends were lower.

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

   Using drop_first=True during dummy variable creation is crucial because it helps to

   - Avoid Multicollinearity: By dropping the first dummy variable, we prevent the creation of redundant features that could lead to multicollinearity in the regression model.
   - Full Rank Matrix: This approach ensures the resulting matrix is of full rank which helps in linear regression.

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

The numerical variable temperature (temp) had the highest correlation with the target variable bike rental counts (cnt), indicates that temperature is a strong predictor of bike rental demand.

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

The assumptions of Linear Regression were validated as follows,

- Linearity: Scatter plots of predicted values v/s residuals were examined to ensure no pattern, indicating a linear relationship.
- Homoscedasticity: The same scatter plots were used to check for constant variance of residuals across all levels of the predicted values
- Normality of Residuals: A histogram and Q-Q plot of the residuals were created to check if the residuals were approximately normally distributed.
- Multicollinearity: Variance Inflation Factor (VIF) was calculated for each feature to ensure no high multicollinearity among the independent variables.

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

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes were

- Temperature (temp): Higher temperatures were resulting higher bike rental counts.
- Year (yr): Second year showed increase compared to first year.
- Humidity (hum): Higher humidity reduced the demand (Negative correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The algorithm aims to find the best-fitting straight line through the data points, which can be used to make predictions. The independent variable is also known as the **predictor variable**. And the dependent variables are also known as the **output variables.**

- **Linear relationship:** A linear relationship must exist between the independent and dependent variables. To determine this relationship, we can create a scatter plot—a random collection of x and y values—to see whether they fall along a straight line.
- **Residual independence:** A residual is the difference between the observed data and the predicted value.
- **Normality:** Graphing techniques like Q-Q plots determine whether the residuals are normally distributed. The residuals should fall along a diagonal line in the center of the graph.
- **Homoscedasticity:** Homoscedasticity assumes that residuals have a constant variance or standard deviation from the mean for every value of x.

Type of linear regressions:

- **Simple linear regression:** Simple linear regression is defined by the linear function: $Y= \beta0*X + \beta1 + \varepsilon$.
- **Multiple linear regression:** In multiple linear regression analysis, the dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows: $Y= \beta0*X0 + \beta1X1 + \beta2X2+\ldots\ldots \beta nXn+ \varepsilon$

2. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- 1 indicates a perfect positive linear correlation
- -1 indicates a perfect negative linear correlation
- 0 indicates no linear correlation.

Formula to calculate Pearson's R

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The Pearson correlation coefficient can also be used to test whether the relationship between two variables is significant.

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

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Scaling is performed on the data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

  It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalized scaling brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- Normalization is useful when the scale of the features is arbitrary, while standardization is preferred when the features have different units or distributions.

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

The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity occurs when one predictor variable is an exact linear combination of one or more other predictor variables.

This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Q-Q plot is also a crucial diagnostic tool in linear regression to ensure that the residuals are normally distributed, to validate the assumptions of the regression model and ensures the better results.