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: Lets discuss each categorical variable and there effect on dependent variable :

1. mnth (i.e. month) – Month have a good corelation with the dependent variable as change in month shows significant change in the dependent variable. Its found that bike rental ratio get higher in month of September and October.

2. Season- It's found that bike rental ratio goes up in summer and fall season and it can be said that the best season to plan a trip is summer and fall and spring can be avoided in spring.

3. yr(Year) – The data states that there is a significant hike in bike rentals from 2018 to 2019.

4. Holiday- Bike rentals are higher on holidays and we can see high bike rentals ratios.

5. Weekday – From the data we can see that on Fridays and Saturday there is a significant hike in bike rentals as compared on other days of the week.

6. Weathersit - Weather conditions significantly influence bike usage patterns. The highest ridership occurs on clear or partially cloudy days, indicating a strong preference for favorable weather. Interestingly, even during light rainy days, the number of registered users remains relatively high, suggesting that the bikes are likely being utilized for daily commutes to workplaces. However, we lack data for days with heavy rain or snow, so it's challenging to draw conclusions regarding bike usage under such extreme weather conditions.

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

Ans - Using drop_first=True in dummy variable creation is important to prevent multicollinearity in regression. It simplifies interpretation, stabilizes the model, and prevents mathematical issues caused by perfect multicollinearity by dropping one reference category, making coefficients more intuitive and reliable.

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

Ans- "The variable 'temp' exhibits the strongest correlation with the target variable, boasting a correlation coefficient of 0.99."

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

Ans - After building a Linear Regression model on the training set, validate its assumptions:

1. Check residuals for linearity, constant variance, and normality.
2. Ensure residuals are independent (no autocorrelation).
3. Detect multicollinearity among predictors.
4. Identify outliers and influential observations.
5. Evaluate model fit statistics.
6. Perform cross-validation to assess generalization.
7. Use statistical tests for heteroscedasticity and autocorrelation.
8. Apply transformations if assumptions are not met.

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

Ans : According to our final model, the three most influential predictor variables affecting bike bookings are as follows:

1.Temperature (temp) - With a coefficient value of '0.5634,' it indicates that a one-unit increase in the temperature variable results in a corresponding increase of 0.5634 units in bike bookings.

2.Weather Situation 3 (weathersit_3) - Showing a coefficient value of '-0.3070,' this suggests that compared to Weathersit1, a one-unit increase in Weathersit3 leads to a decrease of 0.3070 units in bike bookings.

3.Year (yr) - With a coefficient value of '0.2308,' it indicates that a one-unit increase in the year variable results in an increase of 0.2308 units in bike bookings.

General Subjective Questions

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

Ans : *Linear Regression:*

Linear regression is a technique for modelling the relationship between a dependent variable (target) and one or more independent variables (features or predictors).

Simple Linear Regression:
- In simple linear regression, we aim to find a straight line that best fits the data with one predictor and one target.
- The model's hypothesis function is $Y = \beta_0 + \beta_1 X$.
- It minimizes the mean squared error (MSE) to estimate the parameters β_0 and β_1.

Multiple Linear Regression:
- For multiple linear regression, there are multiple predictors, allowing for more complex relationships.
- The hypothesis function is an extension of the simple linear regression equation with multiple coefficients.
- It still minimizes the MSE to estimate all parameters.

Model Evaluation:
- The model's quality is assessed using metrics like R-squared (R^2) to measure how well the model explains the variance in the data.
- Assumptions include linearity, normality of residuals, homoscedasticity, no multicollinearity, and independence of residuals.

Linear regression is widely used for tasks like predicting prices, economic analysis, and understanding relationships between variables.

2. Explain the Anscombe's quartet in detail.

Ans -

Anscombe's quartet is a collection of four datasets that share identical descriptive statistics, including means, variances, R-Squared values, correlations, and linear regression lines. However, when you create scatter plots for these datasets, they reveal distinct visual patterns. This quartet was crafted by statistician Francis Anscombe in 1973 to underscore the significance of data visualization and the potential misinterpretation of relying solely on summary statistics.

Each of the four datasets in Anscombe's quartet consists of 11 pairs of x and y values. When you graph them, each dataset appears to have a unique relationship between x and y, featuring different patterns of variation and diverse strengths of correlation. Despite these apparent distinctions, all four datasets share the same summary statistics, including the same means and variances for x and y, correlation coefficients between x and y, and linear regression lines.

Anscombe's quartet serves as a demonstration of the importance of exploratory data analysis and highlights the limitations of relying solely on summary statistics. It underscores the crucial role of data visualization in identifying trends, outliers, and other critical details that may not be evident from summary statistics alone.

3. What is Pearson's R?

Ans -

The Pearson coefficient, also known as the Pearson correlation coefficient or Pearson product-moment correlation coefficient, measures the strength of the relationship between two variables measured on an interval or ratio scale.

Understanding the Pearson Coefficient:
To calculate the Pearson coefficient, variables X and Y are plotted on a scatter plot. A linear relationship must exist for the coefficient to be meaningful; a non-linear scatter plot is not useful. The closer the scatter plot resembles a straight line, the stronger the association. Numerically, the Pearson coefficient ranges from -1 to +1, similar to correlation coefficients used in linear regression. A value of +1 signifies a perfect positive relationship, indicating both variables move in the same direction. Conversely, a value of -1 indicates a perfect negative relationship, where one variable increases as the other decreases. A coefficient of 0 implies no correlation between the variables.

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

Ans - Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centred around 0 or in the range (0,1) depending on the scaling technique.

Feature scaling is a critical step in data preprocessing to address disparities in magnitudes, values, or units. Without feature scaling, machine learning algorithms may give undue importance to larger values and underestimate smaller values, irrespective of their unit of measurement.
Standardization is divided by the standard deviation after the mean has been subtracted. Data is transformed into a range between 0 and 1 by normalization, which involves dividing a vector by its length.

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

Ans - A Variance Inflation Factor (VIF) becoming infinite typically occurs when there is perfect multicollinearity in the dataset. Perfect multicollinearity means that one or more independent variables can be expressed as a linear combination of others, rendering them redundant in the regression model. This perfect relationship leads to an issue in the calculation of VIF because VIF relies on matrix inversion, and in the presence of perfect multicollinearity, the matrix becomes singular or non-invertible.

In simpler terms, when one or more predictors are perfectly correlated or can be predicted exactly from others, it breaks down the mathematical calculation for VIF because it cannot handle the presence of perfect multicollinearity. As a result, the VIF for the affected variables becomes infinite.

To address this issue, it's essential to identify and resolve multicollinearity in the dataset. This can involve:

1. Removing one of the correlated variables if it doesn't have a substantial impact on the model.
2. Combining the correlated variables into a composite variable.
3. Reconsidering the choice of predictors or collecting additional data to reduce multicollinearity.

Resolving multicollinearity not only helps in preventing infinite VIF values but also improves the stability and interpretability of your regression model.

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

Ans -

Q-Q Plot in Linear Regression:

A Q-Q (Quantile-Quantile) plot is a graphical tool that plays a crucial role in assessing the normality assumption of residuals in linear regression. The normality of residuals is a fundamental assumption in linear regression, and Q-Q plots are instrumental in evaluating this assumption. Here's why Q-Q plots are valuable in the context of linear regression:

1. Normality Assumption:
   Linear regression assumes that the residuals (the differences between the observed and predicted values) follow a normal distribution. This assumption is vital for conducting valid statistical tests and constructing reliable confidence intervals. Q-Q plots help us determine whether this assumption holds.

2. Visual Assessment:
   Q-Q plots provide a visual means of assessing the normality of residuals. They compare the quantiles of the observed residuals to the quantiles expected from a theoretical normal distribution. When the residuals are normally distributed, the points in the Q-Q plot will fall along a straight line. Deviations from this line can indicate non-normality.

3. Skewness and Outliers:
   By examining a Q-Q plot, you can detect skewness and outliers in the residuals. If the points in the plot deviate from a straight line, it suggests that the distribution of residuals is skewed or has heavy tails. These deviations can be indicative of issues in the linear regression model, such as nonlinearity or the presence of influential observations.

4. Model Validation and Improvement:

Q-Q plots are a fundamental step in the validation of a linear regression model. If the residuals deviate significantly from a normal distribution, it may be necessary to consider data transformations or robust regression techniques to address the issue. Addressing deviations from normality can lead to a more accurate and robust model.