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: Variables - workingday and weekday didn't have much effect on the dependent variable. The other categorical variables -
   a) season implied that fall had highest demand of bikes followed by summer, winter and spring
   b) holidays has less demand rather than non-holidays
   c) weathersit – category 1 had the most demand of bikes followed by 2 and 3
   d) mnth implied that months from June - September had the highest demand of bikes
   e) yr implied that demand of bikes is rising every year

2) Why is it important to use **drop_first=True** during dummy variable creation?
   Ans: This is because the number of dummy variables that needs to be created is (predictor – 1). Therefore, we use **drop_first=True** to always create dummy variable one less than the number of predictor for that variable.

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: **temp** has the highest corelation

4) How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: Performed Residual Analysis to check if the Error Terms is normally distributed.

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: **year, temp,** and **weathersit**

General Subjective Questions

1) Explain the linear regression algorithm in detail.
   Ans: Linear regression is a statistical method for modeling the relationship between a dependent variable (denoted as 'y') and one or more independent variables (denoted as 'x'). The goal of linear regression is to

find the best-fit line that can represent the relationship between the independent variables and the dependent variable. This line is known as the regression line or the line of best fit.

The equation of the regression line is of the form: y = mx + c

The objective of linear regression is to find the values of m and c that minimize the sum of the squared differences between the actual y values and the predicted y values. This is known as the least-squares method.

2) Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a group of four datasets that have identical statistical properties, but very different visual representations. They were created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in statistical analysis.

Each of the four datasets in Anscombe's quartet consists of 11 x-y pairs. The x-values are identical in all four datasets and range from 4 to 14. The y-values, however, are different for each dataset, and have been chosen so that the datasets have identical summary statistics.

Here are the four datasets in Anscombe's quartet:

Dataset I:
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset II:
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y = [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]

Dataset III:
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset IV:
x = [8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 19]
y = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89, 12.5]

Despite having identical summary statistics (i.e., the mean, variance, correlation coefficient, and regression line for each dataset are the same), the datasets in Anscombe's quartet have very different visual representations when plotted on a graph.

Dataset I appears to have a linear relationship between x and y, and the regression line fits the data well.

Dataset II has a similar regression line to Dataset I, but the data points are much more scattered.

Dataset III appears to have a non-linear relationship between x and y, with a curved regression line that fits the data well.

Dataset IV appears to have no relationship between x and y, except for one outlier point that has a strong influence on the regression line.

Anscombe's quartet highlights the importance of data visualization in statistical analysis, and demonstrates that summary statistics alone are often insufficient for understanding complex relationships between variables. By visualizing the data, we can identify patterns and trends that would be impossible to detect by looking at summary statistics alone.

3) What is Pearson’s R?
Ans: Pearson's R is a statistical measure that describes the linear correlation between two continuous variables. It is also known as the Pearson correlation coefficient or simply the correlation coefficient.
Pearson's R is a value that ranges from -1 to 1, where:
    1. A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear fashion.
    2. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear fashion.
    3. A value of 0 indicates no correlation, meaning that there is no linear relationship between the two variables.

The formula for Pearson's R is as follows:
r = (nŒ£xy - Œ£xŒ£y) / sqrt[(nŒ£x^2 - (Œ£x)^2)(nŒ£y^2 - (Œ£y)^2)]

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a data preprocessing technique used to transform the values of a variable to a specific range or distribution. It involves adjusting the values of one or more variables to make them more comparable or to make their interpretation easier.

Scaling is performed for a number of reasons, including:

1. Comparability: When comparing variables that have different units or scales, scaling can be used to bring them to a common scale for easier comparison.
2. Model performance: Many machine learning algorithms work better when the input features are on a similar scale, so scaling is often performed to improve model performance.
3. Interpretation: Scaling can help to make the interpretation of the data easier and more meaningful.

The difference between normalized scaling and standardized scaling is in the range of the resulting values. Normalized scaling will produce values between 0 and 1, while standardized scaling produces values that may be positive or negative, and that have a mean of 0 and standard deviation of 1.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: In some cases, the value of VIF can be infinite. This happens when the coefficient of determination (R-squared) is equal to 1 for the model that includes the predictor variable. When R-squared equals 1, it means that the predictor variable is a perfect linear combination of other predictor variables in the model. As a result, the VIF for that variable becomes infinite.
This situation typically arises when one of the predictor variables is a linear combination of the other predictor variables, which can occur due to various reasons such as including an interaction term between two highly correlated variables or adding a variable that is derived from other variables. In such cases, the infinite VIF indicates that the predictor variable can be dropped from the model without losing any information.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare the distribution of a sample of data to a theoretical probability distribution. The Q-Q plot plots the quantiles of the data against the corresponding quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie on a straight line.
In linear regression, Q-Q plots are used to check the normality assumption of the residuals, which is one of the assumptions of the linear regression

model. The residuals are the differences between the observed values and the predicted values from the regression model. If the residuals are normally distributed, the Q-Q plot of the residuals will show a straight line, indicating that the residuals follow a normal distribution.

The importance of Q-Q plots in linear regression is that they help to verify the assumption of normality of residuals, which is critical for accurate inference and interpretation of the regression results. If the residuals are not normally distributed, this can indicate a violation of the linear regression assumptions and suggest that the model may not be appropriate for the data. In such cases, appropriate steps such as transforming the data or using non-linear regression techniques may be necessary.

Overall, Q-Q plots are a useful tool for assessing the normality assumption of residuals in linear regression and can help to ensure that the regression model is appropriate for the data.