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?

- Bike demand in the fall is the highest.
- Bike demand takes a dip in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is high in the months from May to October.
- Bike demand is high if weather is clear
- Bike demand is almost similar throughout the weekdays.
- Bike demand doesn't change whether day is working day or not.

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

It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables. For Example: We have three variables: A, B and C. We can only take 2 variables as A will be 1-0, B will be 0-1, so we don't need C as we know 0-0 will indicate C. So we can remove it.

It is also used to reduce the collinearity between dummy variables.

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

'atemp' and 'temp' both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

a. **Linearity:** The relationship between the dependent (bike count) and independent variables is linear and is explained by a straight line equation.
b. **Independence:** The observations are independent of each other.
c. **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
d. **Normality:** The errors follow a normal distribution, as we can see in the histogram of the error plot.
e. **No multicollinearity:** The independent variables are not highly correlated with each other, as can be seen in the correlation graph plotted

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

1. Temperature
2. Weathersit
3. Season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a popular and widely used algorithm in machine learning and statistics for modeling the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting linear relationship between the independent variables. The resulting linear equation can be used to predict the value of the dependent variable for new or unseen input data.

Linear regression models the relationship between the dependent variable and independent variables using a linear equation. In simple linear regression, the equation is y = b0 + b1*x, where y is the dependent variable, x is the independent variable, b0 is the intercept, and b1 is the coefficient for the independent variable. In multiple linear regression, the equation expands to y = b0 + b1x1 + b2x2 + ... + bn*xn, where x1, x2, ..., xn are the independent variables, and b1, b2, ..., bn are their respective coefficients. The coefficients determine the impact of each independent variable on the dependent variable.

Once the model is obtained, it is evaluated to assess the performance. Common metrics include the coefficient of determination (R-squared), measuring the variance explained by the model, and the VIF, used to assess multicollinearity in a regression model.
Once the model is trained and evaluated, it can make predictions on new data by calculating the dependent variable's value based on the learned coefficients.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties but exhibit significantly different patterns when graphically visualized. It highlights the importance of data visualization and the limitations of relying solely on summary statistics.

The four datasets in Anscombe's quartet have the following characteristics:

1. **Dataset I:** It represents a simple linear relationship between the x and y variables. When plotted, the data points align almost perfectly along a straight line.
2. **Dataset II:** It also demonstrates a linear relationship, but with an outlier. This dataset highlights the impact of outliers on the summary statistics, such as the mean and correlation coefficient.
3. **Dataset III:** It exhibits a non-linear relationship between x and y, resembling a quadratic curve. The regression line for this dataset would not be a good fit, despite having a similar mean and correlation coefficient as the first dataset.
4. **Dataset IV:** This dataset consists of an outlier that significantly affects the linear relationship between x and y. When the outlier is removed, the linear relationship vanishes, emphasizing the importance of inspecting and understanding the individual data points.

Anscombe's quartet shows that relying solely on summary statistics can be misleading, and visualizing data can offer valuable insights into the nature of the relationships and the quality of the model fit. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as Pearson's R or simply as R, is a statistical measure that quantifies the linear relationship between two continuous variables. It measures the strength and direction of the linear association between the variables.
Pearson's R ranges from -1 to +1. The value of R indicates the degree of correlation:

- If R = 1, it represents a perfect positive correlation, indicating that the variables have a strong linear relationship and increase or decrease together.

- If R = -1, it represents a perfect negative correlation, indicating that the variables have a strong linear relationship, but they move in opposite directions.
- If R = 0, it represents no linear correlation, indicating that there is no linear relationship between the variables.

Pearson's R is widely used in various fields, including statistics, social sciences, economics, and data analysis. It provides a measure of the linear relationship between variables and can help assess the strength and direction of association.

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

Scaling is a pre-processing technique used in data preparation, primarily in machine learning and data analysis. It involves transforming the numerical values of variables to a standardized range or distribution. Scaling is performed to ensure that all variables are on a comparable scale, eliminating potential bias or dominance of certain variables in the analysis.

The main reasons for scaling are:

1. To bring variables to a similar range.
2. To avoid the dominance of variables.

The two common scaling techniques are:

1. **Min-max scaling** – This scales the variables to a range between 0 and 1. It subtracts the minimum value of the variable and divides by the range (maximum value minus minimum value). This maintains the relative distribution of the data.
2. **Standardized scaling** – This transforms variables to have a mean of 0 and a standard deviation of 1. It subtracts the mean of the variable and divides by the standard deviation. This centers the data around zero and creates a standard deviation-based scale.

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

In some cases, the value of the Variance Inflation Factor (VIF) can become infinite. This occurs when perfect multicollinearity is present in the regression model. Perfect multicollinearity, where independent variables are perfectly predicted by linear combinations of other variables, can lead to an infinite value of the Variance Inflation Factor (VIF). This occurs when dividing by zero in the VIF calculation. Perfect multicollinearity causes issues in regression models, including unstable coefficient estimation and difficulty interpreting variable effects. To address this, the variables causing the multicollinearity need to be identified and handled, such as removing correlated variables or using techniques like principal component analysis (PCA). While perfect multicollinearity is rare, high multicollinearity still poses concerns and may require mitigation.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution. It compares the quantiles of the observed data with the quantiles expected from the theoretical distribution. In linear regression, Q-Q plots are used to check the assumption of normality for the residuals. By examining the plot, we can determine if the residuals are normally distributed. If the points on the plot form a straight line, it suggests that the residuals follow a normal distribution, which is desired for linear regression. Deviations from the line indicate departures from normality, which can help identify issues like skewness or outliers. Q-Q plots help guide model improvements, such as considering data transformations or exploring alternative regression models. Overall, Q-Q plots provide a visual and quantitative assessment of normality assumptions in linear regression, aiding in understanding the model's performance and potential areas for enhancement.