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?

- More people rent bikes during spring and summer, and fewer during fall and winter.
- In 2019, more bikes were rented compared to 2018.
- The most popular months for bike rentals are June to September, while January is the least popular.
- Fewer people rent bikes during holidays.
- Bike rentals are consistent throughout the weekdays.
- There is no difference in bike rentals between working and non-working days.
- The highest bike rentals occur during clear or partly cloudy weather, followed by misty or cloudy weather, and then light snow or light rain weather.

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

- When creating dummy variables, an extra column is created for each category.
- This can lead to multicollinearity and affect the accuracy of statistical models.
- Using drop_first=True helps to reduce the number of columns created and the correlations among the dummy variables.
- By dropping the first column, we can avoid the dummy variable trap.
- This improves the accuracy of our models.

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

- The temp and atemp variables are highly positively correlated, meaning they carry similar information.
- The total_count, casual, and registered variables are highly positively correlated, indicating they measure similar aspects of bike usage.
- The pair-plot suggests that temp and atemp have 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?

- To validate the assumptions of Linear Regression, I used the R-squared or Coefficient of Determination.
- The R-squared value was 0.81 on average, which means that the predictor can explain 81% of the variance in the target variable that is contributed by the independent variables.

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

- Temperature
- Weathersit
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fit line that describes the relationship between the variables.

The algorithm works by first defining a linear equation that describes the relationship between the dependent variable and the independent variables. The equation takes the form:

y = b0 + b1x1 + b2x2 + ... + bn*xn

where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that determine the slope and intercept of the line.

The algorithm then uses a method called least squares to find the values of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values of the dependent variable. This is done by calculating the partial derivatives of the sum of squared errors with respect to each coefficient, and then setting them equal to zero to find the values that minimize the error.

Once the coefficients have been calculated, the algorithm can be used to make predictions for new values of the independent variables. The predicted value of the dependent variable is calculated by plugging the new values of the independent variables into the linear equation.

Linear regression is a simple and powerful algorithm that can be used to model a wide range of relationships between variables. It is widely used in fields such as economics, finance, and engineering to make predictions and inform decision-making. However, it is important to note that linear regression assumes a linear relationship between the variables, and may not be appropriate for data that exhibits non-linear relationships.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four datasets that have the same statistical properties but look different when plotted. It was made to show that just looking at summary statistics is not enough to understand data. The quartet proves that even if datasets have the same statistical properties, they can still have different patterns of variation. By visualizing the data, we can see important information that we might miss if we only look at summary statistics.

3. What is Pearson's R?

- Pearson's R is a statistical measure that shows how strong and in what direction the linear relationship is between two variables.
- It ranges from -1 to 1, where -1 means a perfect negative correlation, 0 means no correlation, and 1 means a perfect positive correlation.
- It is commonly used in statistics, economics, and psychology to measure the strength of the relationship between two variables.
- Pearson's R assumes that the relationship between the variables is linear and that the variables are normally distributed.
- Despite its limitations, it is still a widely used and powerful statistical measure.

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

- Scaling is a technique used to transform variable values to a specific range.
- It is done to make variables comparable and prevent larger values from dominating the analysis.
- There are two types of scaling: normalized scaling and standardized scaling.
- Normalized scaling transforms values to a range between 0 and 1.
- Standardized scaling transforms values to have a mean of 0 and a standard deviation of 1.

Normalized scaling:

x_norm = (x - min(x)) / (max(x) - min(x))

where x is the original value, x_norm is the normalized value, min(x) is the minimum value of x, and max(x) is the maximum value of x.

Standardized scaling:

x_std = (x - mean(x)) / std(x)

where x is the original value, x_std is the standardized value, mean(x) is the mean value of x, and std(x) is the standard deviation of x.

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

- VIF (Variance Inflation Factor) measures how much multicollinearity exists in a set of regression variables.
- If one or more independent variables in a regression model are perfectly collinear, the VIF value can be infinite.
- Perfect collinearity means that the variables are linearly dependent on each other.
- To avoid this problem, it is important to remove one of the collinear variables from the model.

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

- A Q-Q plot is a graph used to compare a sample of data to a theoretical distribution.
- In linear regression, Q-Q plots are used to check if the residuals are normally distributed, which is important for linear regression.
- Q-Q plots help us identify any deviations from normality and take steps to address them.