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?

The bike rental count shows an increase during the spring and summer seasons, followed by a decrease during the fall and winter seasons.

The demand for rental bikes increased in 2019 compared to 2018.

The months of June to September experience the highest demand for rental bikes, while January has the lowest demand.

Bike demand is lower during holidays compared to non-holidays.

The demand for rental bikes is consistent throughout the weekdays.

There is no significant difference in bike demand between working days and non-working days.

The highest bike rental counts are observed 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?

Using drop_first=True when creating dummy variables is important because it helps to reduce the number of columns created and the correlations among the dummy variables. When creating dummy variables, an extra column is created for each category, which can lead to multicollinearity and affect the accuracy of statistical models. By dropping the first column, we can avoid the dummy variable trap and improve 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, indicating that they carry similar information.

The total_count, casual, and registered variables are highly positively correlated, suggesting that they are measuring similar aspects of bike

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

I used the R-squared, or Coefficient of Determination, which is 0.81 on average in our case. This means that the predictor is able to 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 set of four datasets that have identical statistical properties but different visual representations. It was created to demonstrate the importance of visualizing data and not relying solely on summary statistics. The quartet shows that even datasets with identical statistical properties can have very different patterns of variation, and that visualizing the data can reveal important insights that are not apparent from summary statistics alone.

3. What is Pearson's R?

Pearson's R is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, with values of -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating 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 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 data pre-processing technique used to transform the values of variables to a specific range. It is performed to ensure that the variables are comparable and to prevent variables with larger values from dominating the analysis. There are two common types of scaling: normalized scaling and standardized scaling. Normalized scaling transforms the values to a range between 0 and 1, while standardized scaling transforms the 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?

The VIF (Variance Inflation Factor) is a measure of the degree of multicollinearity in a set of regression variables. Sometimes, the value of VIF can be infinite when one or more independent variables in a regression model are perfectly collinear, meaning that they 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 graphical technique used to compare the distribution of a sample of data to a theoretical distribution. In linear regression, Q-Q plots are used to check the assumption of normality of the residuals, which is a key assumption of linear regression. The use of a Q-Q plot is important because it allows us to identify any deviations from normality and take appropriate steps to address them.