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?

Answer :

- Highest Bikes were rented on Friday.
- Highest bikes were rented when the weather was clear and not cloudy or rainy.
- Highest Bikes were rented in the month of August.
- Highest bikes were rented in the Fall Season.
- Highest Bikes were rented on working days and not on holidays or non working days.
- The Saturday and Sunday being non working days the count of bikes rented on these days is almost negligible for these days compared to other days. But surprisingly there is an interesting observation, that whenever Wednesday is a holiday or non working day, more bikes are being rented compared to when Wednesday is a working day.

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

Answer:

It is important to use drop_first=True during dummy variable creation because it prevents multicollinearity. Multicollinearity is a problem that occurs when two or more independent variables are highly correlated.

Using drop_first=True will drop one of the dummy variables for each categorical variable. This ensures that the dummy variables are not perfectly correlated with each other

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

Answer:

The columns **temp** and **atemp** are highly correlated.

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

Answer:

- **Linearity**: The relationship between the independent variables and the dependent variable should be linear. This can be checked by plotting the data and looking for a linear trend.
- **Residual Analysis**: Examined for consistent variance by graphing residuals against predicted values. Additionally, visualized residuals against individual predictors to detect any potential patterns
- **Multicollinearity**: Conducted an analysis of the correlation matrix and Variance Inflation Factors (VIF) to identify instances of strong correlation among predictor variables.
- **Homoscedasticity**: The variance of the residuals should be constant across all values of the independent variables. This can be checked by plotting the residuals against the fitted values and looking for a constant variance

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

Answer:

Judging from the coefficients of the independent variables, the three most significant variables are **yr(year)**, **atemp**, **season_winter**

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer:

Linear regression is a widely used supervised machine learning algorithm that aims to establish a relationship between a dependent variable (also known as the target) and one or more independent variables (also known as features or predictors). The goal of linear regression is to create a linear equation that can predict the dependent variable based on the values of the independent variable.

Linear equation is of the form:

y = mx + b

where:

- y is the dependent variable
- m is the slope of the line
- b is the y-intercept
- x is the independent variable

Linear regression is a powerful tool that can be used to predict a wide variety of values. However, it is important to remember that the linear regression algorithm only works if the assumptions of the model are met. If the assumptions are not met, then the predictions may not be accurate.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a set of four data sets that have the same mean, standard deviation, correlation coefficient, and regression line. However, the four data sets look very different when plotted. This shows that the mean, standard deviation, correlation coefficient, and regression line are not enough to fully understand the relationship between two variables.Anscombe's quartet is a reminder that we should not rely on summary statistics alone when evaluating a data set. We should also look at the data visually to get a better understanding of the relationship between the variables.

Anscombe's quartet was first published in 1973 by Francis Anscombe. Anscombe was a statistician who was interested in the limitations of using summary statistics to understand data. He created the quartet to show that the mean, standard deviation, correlation coefficient, and regression line can be misleading if the data is not normally distributed

3. What is Pearson's R?

Answer:

Pearson's r is a statistical measure of the linear correlation between two variables. It is a coefficient that ranges from -1 to 1, where:

A value of -1 indicates a perfect negative correlation

A value of 0 indicates no correlation

A value of 1 indicates a perfect positive correlation

The higher the value of Pearson's r, the stronger the linear relationship between the two variables. For example, if Pearson's r is 0.7, then there is a strong positive correlation between the two variables. This means that as the value of one variable increases, the value of the other variable also tends to increase.

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

Answer:

Scaling is a technique used to transform data so that it has a standard distribution. This can be useful for a number of reasons, including:

- Making the data more comparable
- Improving the performance of machine learning algorithms
- Improving the interpretability of machine learning models

There are two main types of scaling:

Normalized scaling: Normalized scaling is a type of scaling where the features are divided by their standard deviation. This means that the features will have a mean of 0 and a standard deviation of 1.

Standardized scaling: Standardized scaling is a type of scaling where the features are subtracted from their mean and then divided by their standard deviation. This means that the features will have a mean of 0 and a standard deviation of 1, but they will also be centered around 0.

The main difference between normalized scaling and standardized scaling is that normalized scaling does not center the features around 0

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

Answer:

The variance inflation factor (VIF) is a measure of how much the variance of an estimated regression coefficient is inflated due to collinearity. Collinearity is when two or more independent variables are highly correlated. When there is collinearity, the VIF can be infinite.

There are a few reasons why the VIF can be infinite:

Perfect collinearity: If two independent variables are perfectly correlated, then the VIF for each variable will be infinite. This is because the variance of the estimated regression coefficient for each variable will be zero.

Near-perfect collinearity: If two independent variables are nearly perfectly correlated, then the VIF for each variable will be very high. This is because the variance of the estimated regression coefficient for each variable will be very small.

Multicollinearity: If there are multiple independent variables that are correlated with each other, then the VIF for each variable can be high. This is because the variance of the estimated regression coefficient for each variable will be inflated by the correlation between the other independent variables.

If the VIF is infinite, then it means that the independent variables are perfectly or nearly perfectly correlated

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

Answer:

A Q-Q plot is a graphical tool used to assess whether or not a set of data follows a particular distribution. In the context of linear regression, a Q-Q plot can be used to assess whether or not the residuals (the difference between the actual values and the predicted values) follow a normal distribution.

The use and importance of a Q-Q plot in linear regression is that it can help to identify potential problems with the model. If the residuals do not follow a normal distribution, then this can indicate that the model is not fit well to the data