## Assignment based questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

In the case study the categorical variables are: season, month, yr, weekday, holiday and weathersit. From the final model, it is significant that year, workingday and season has positive effect on the dependent variable count (cnt). Hence final model building shows growth in R-squared and Adjusted R-squared value for year, season, workingday etc.

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

Answer:

It is advisable to use drop_first = True during dummy variable creation because it helps in eliminating the redundant feature from the dataset. Hence it reduces the correlation among dummy variables.

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

Answer:

The independent variables temp and atemp both have high correlation with the target variable cnt.

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

Answer:

The four assumptions of linear regression are:

a. Linearity: This assumption is tested using pair-plot. From the plot we can conclude that target variable cnt has linear relationship with independent variables temp and atemp.
b. Homoscedasticity: There is no visible pattern in the scatter plot between residual and predicted value of the training set data.
c. Independence: The error terms are independent of each other.
d. Normality: Error terms are normally distributed. We tested this by plotting the residuals.
e.

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

Answer:

Based on final model, the top 3 features are: yr, atemp and winter. These features are significantly contributing towards the demand of the shared bike.

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a type of supervised machine learning. In this technique, models are trained based on labelled data i.e. the output variable is predicted based on some inut variables. The linear regression is used when there is some sort of linear relationship between dependent and independent variables and finds out the best fit linear relationship based on labelled trained data. The linear regression model aims to find best fit line and the optimal values for intercept and coefficients to minimize the errors. Linear regression is of two types:

a. Simple Linear regression:
   In this technique, we can predict the dependent variable based on single independent variable. It is denoted by equation: y=mx+c where m is the slope and c is the intercept.
b. Multiple Linear regression: In this technique, the dependent variable is predicted based on multiple independent variables. The equation is denoted by:
   y = c + m1x1+m2x2+….+ mnxn where c is the intercept and m1,m2,m3..mn are the slopes or coefficients.

2. Explain the Anscombe's quartet in detail.

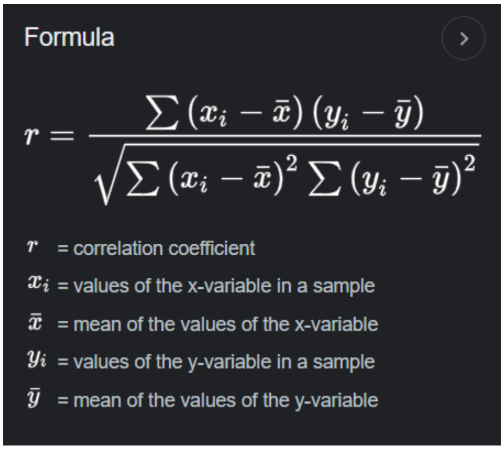
Answer:

Anscombe's quartet consists of four datasets which are nearly same in terms of simple descriptive statistics (i.e. mean, standard deviation, correlation seems to be similar), but there distributions appear very different when plotted in a scatter plot. This is important from data visualization perspective. The data points must be plotted to see the sample distribution before applying any algorithm to build the model.

3. What is Pearson's R?

Answer:

Pearson correlation coefficient, also known as Pearson R, is a statistical term which measures the strength between two continuous variables as well as their direction i.e. they are positively correlated or negatively correlated. The value of Pearson's R lies between 1 to -1. The formula is denoted by:



Formula

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

$r$ = correlation coefficient

$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Answer:

Feature scaling is a technique to standardize the independent feature variables in a fixed range. It is a crucial part of data preparation since it affects the performance of a machine learning model.

Normalization is a scaling technique where the values of the independent variables are rescaled between 0 and 1. It is also known as MinMax scaling.

Standardized scaling is a technique where the independent variables are rescaled around their mean with a unit standard deviation.

Outliers in the data are taken into consideration in case normalized scaling, whereas the standardized scaling does not affect the outliers.

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

Answer:

Infinite VIF value indicates perfect correlation between two independent variables. In this case, the R-square value of the independent variable is 1 which leads to infinite VIF.

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

Answer:

Q-Q plot is also known as Quantile-Quantile plot. This is used to plot the quantiles of the sample distribution against the quantiles of a theoretical distribution. It helps to determine if a dataset follows a specific probability distribution like normal, uniform or exponential distribution.

Importance of Q-Q plot in linear regression:

i. It helps to determine if the train and test sets are derived from same distribution.

ii. It helps to identify if the residuals follow a normal distribution which is an assumption of linear regression.

iii. It helps to determine the skewness of the distribution.