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 :

1. Demand for bikes has increased from 2018 to 2019. 100 % hike against 2018.
2. Month plays very important role as Jan- Jun there is very high demand for the bikes.
3. Demand for bikes is high during the fall season and next will be summer, winter, and spring seasons.
4. Demand for bikes increase with weather.
5. Demand on working days is slightly higher. However it is not that much correlated.

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

Answer:

If you use drop_first=TRUE than it will create N-1 columns which will create and we can encode categorical variable with less dimensions.

n dummy variable creation if we will not use drop_first then it will create all separate columns for all levels of categorical variables but when we set it as True then N-1 columns will create where N is total level of categorical variable. Thus, we can encode categorical variables with less dimensions.

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

Answer:

Temp and atemp has equally high correlation with target variable cnt

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

Answer :

1. check linearity in between dependent and independent variables
2. Calculate residuals (y_train-y_train_pred
3. Check variance of residual

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

Answer : weathersit, year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is a machine learning technique that establishes a statistical relationship between a dependent variable and one or more independent variables. Its purpose is to predict the value of the dependent variable based on the value of one or more independent variables, assuming a linear relationship between them.

The objective is to find the best-fit line that fits the data points, called the regression line. This line is represented by the equation

y=mx+c, where y is the dependent variable, x is the independent variable, m is the slope, and c is the y-intercept.

There are two types of linear regression:

1. simple linear regression and

2. multiple linear regression. In simple linear regression, only one independent variable is involved, while multiple linear regression involves more than one.

We use the OLS method provided by the statsmodel.api library or several other libraries such as sklearn to implement linear regression. The coefficients b0, b1, b2, ..., bn represent the contribution of each independent variable in multiple linear regression equation y = b0 + b1x1+b2x2…+bnxn.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a collection of four datasets that share identical statistical characteristics, yet exhibit striking dissimilarities when graphed. The quartet underscores the significance of visual data analysis and urges against depending solely on statistical summaries. It illustrates that summary statistics can be deceptive when dealing with intricate data. It is crucial to use graphical analysis to comprehend the data's nature and any underlying connections. Anscombe's quartet serves as a compelling reminder to statisticians and data analysts to not rely solely on summary statistics, but to thoroughly explore and visualize the data before drawing conclusions or making predictions.

3. What is Pearson’s R? (3 marks)

Answer:

Pearson's R, also called Pearson's correlation coefficient, is a statistical measure that evaluates the linear relationship between two continuous variables. It is a numerical value that ranges from -1 to +1, where a value of +1 indicates a perfect positive correlation, a value of -1

indicates a perfect negative correlation, and a value of 0 suggests no correlation between the two continuous variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling in machine learning refers to the process of transforming data to fit within a specific scale, which is done to ensure that all features are on a similar scale. This can improve the performance of algorithms and predictions during modelling. Two common methods of scaling are normalization and standardization. Normalization scales all numeric variables within the range of 0-1, while standardization scales all numeric variables to have a mean of 0 and a standard deviation of 1. Normalization is useful when the distribution of the data is unknown or when it does not follow a Gaussian distribution. On the other hand, standardization is useful when the data follows a Gaussian distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

If 2 predictor variables have perfect correlation then the VIF is infinite. This means one variable can be explained perfectly using a linear relation with other. The value is infinte as the R2 correlation between the variables becomes 1 and the formula for VIF 1/1-R = 1/0 = infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

The Q-Q plot (quantile-quantile plot) is a visual method for determining whether two sets of data originate from a population with the same distribution, such as normal, exponential, or uniform. It is a probability plot that compares two probability distributions by plotting their quantiles against each other. Q-Q plots are utilized in linear regression to verify whether the residuals conform to a normal distribution. If the residuals follow a normal distribution, it suggests that the linear regression model is accurate and that its assumptions have been satisfied. However, if the residuals do not conform to a normal distribution, it indicates that there might be an issue with the model, and additional investigation is required.