Assignment-based Subjective Questions

Question 1:

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

Answer:

1. The demand for bikes is highest during the fall season and lowest during the spring season.
2. September month has the highest demand for bikes and January has the lowest.
3. There is a highest demand for bikes on Sunday and lowest demand on Tuesdays.
4. Demand for bikes has grown when compared count between years 2018 and 2019.
5. Not much difference can be seen when demand comparison is made between working day vs non-working day.
6. Demand for bikes is highest when it is clear weather and lowest when it is cloudy.
7. Demand for bikes is higher during holiday when compared to non-holiday

Question 2:

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

Answer:

We create dummy variables to show categories using 0s and 1s. For example if 3 dummy variables are considered. The first one is 1, and the others are 0. If the second and third are 0, we automatically know the first is 1. Similarly, if the third is 1, then we know the first and second are 0. This is why using 'drop_first=True' is important—it helps us get rid of one of the dummy variables and keeps things clear. This helps in preventing multicollinearity making the model stable and accurate in predictions.

Question3:

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

Answer:

Variables ‘temp’ and ‘atemp’ have the highest correlation with the target variable count (cnt).

Question 4:

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

Answer:

On completion of model creation over training set, we consider the test data. It is prepared by keeping the common columns between training and test sets. Create a scatter plot for train and test data and check if the points fall on a straight line or not.

Question 5:

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

Answer:

Year, Temperature and Weather are the top 3 significant contributors which explain the demand of shared bikes.

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Answer:

Linear Regression serves as a machine learning model utilized to analyze diverse datasets and create an optimal model for the data.

- Purpose is to determine the relationship between two variables.
- The process of implementing linear regression involves several steps:

A. We begin by comprehending the data through the data dictionary.
B. Subsequently, we employ Exploratory Data Analysis to visualize the data.
C. Once we grasp the data, we proceed to prepare it by transforming categorical columns into binary columns using dummy variables.
D. The data is then split into training and testing sets.
E. Working with the training dataset, we conduct multiple linear regression tasks to construct an ideal model that maximizes the Adjusted R-squared percentage.
F. Similarly, we apply this model to the test dataset, retaining only the columns present in the training dataset—effectively keeping the common columns between the two.
G. We compare the linear regression lines of the training and test data using scatter plots.
H. Additionally, we compare their R-square values to assess whether the training and test data provide the best fit for the model

Question 2:

Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a collection of four distinct datasets that have nearly identical mean, variance and correlation values but have different distributions and appearances when put on a graph. Each dataset consists of 11 (x, y) data points.

The significance of Anscombe's quartet is to in demonstarate the importance of visualizing data and not relying solely on summary statistics. Despite having similar statistical properties, these datasets have vastly different patterns, and this tells us that data visualization is important for understanding the underlying relationships and making deductions..

Question 3:

What is Pearson's R?

Answer:

Pearson's R, or Pearson's correlation coefficient, is a measure of how closely two things are related in a straight-line manner. It tells us if one thing goes up, does the other tend to go up too, or does it go down? The value ranges from -1 to 1. If it's close to 1, they both go up together (positive relationship). If it's close to -1, as one goes up, the other goes down (negative relationship). If it's close to 0, there's not much of a connection between them. It's a way to put a number on how connected two things are."

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Question 4:

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

Answer:

Scaling is a process used to standardize categorical independent variables, making the data comparable within a specific range.

➢ This process is applied on collected data that encompasses features with significant differences in sizes, measurement and units. Without scaling, the algorithm considers these differences, potentially leading to inaccuracies in modelling.

➢ Normalizing scaling adjusts the data to fall within the range of 0 to 1.

➢ Standardized scaling involves replacing values with their z-scores.

Question 5:

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

Answer:

This happens when the predictors are so closely related that we can predict one from another in a perfect way When this happens, the VIF becomes very large and shows as infinity. To solve this, we need to check if our predictors are too similar and make adjustments so they provide distinct and useful information.

Question 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 (Quantile-Quantile plot) is a graph that helps us check if a dataset follows a normal distribution. In linear regression, we need to check if the residuals are normally distributed. In Linear regression we assume normal distribution of residuals. If the Q-Q plot shows a straight line, it indicates the residuals are approximately normally distributed, validating a key assumption of linear regression. If not, adjustments to the model or data transformation might be needed for more accurate predictions.