Assignment-based subjective Questions

Ques -1

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

Ans-1

Categorical variables season, weathersit, holiday, mnth and yr helped us to infer following details –

Season – spring had least bike rental counts and summer had maximum bike rental counts.

Weathersit – No bike rentals during unfavorable weather conditions for example during heavy rains and snow. And Bike rentals were high during clear weather conditions for example clear and partly cloudy weather conditions.

Holiday – Bike rentals were reduced during Holidays.

Yr – 2019 had more bike rentals than 2018.

Mnth – Bike rentals were highest for September and least for December.

Ques -2

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

Ans-2

It is important to drop the first column in order to avoid dummy variables from being redundant.

Ques-3

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

Ans-3

Temp and atemp are the two variable which are having highest correlation with the target variable.

Ques-4

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

Ans-4

Residual distribution should follow normal distribution and should be centred around zero i.e mean should be around zero.

Ques -5

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

Ans-5

Top 3 features which are contributing significantly towards explaining the demand for shared bikes are as below –

1 . temp coefficient

2 . yr coefficient

3. weathersit coefficient

General Subjective Questions

Ques – 1

Explain the linear regression algorithm in detail.

Ans -1

Linear regression is a supervised machine learning algorithm that is used for the prediction of numeric values. Linear regression is the most basic form of the regression analysis.

Regression is most commonly used predictive analysis model. Linear regression is based on following equation –

Y = mx + c

Regression is mainly divided into 2 forms –

1. Simple Linear regression
2. Multiple linear regression

Ques -2

Explain the Anscombe's quartet in detail.

Ans-2

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties. but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Ques-3

What is Pearson's R?

Ans -3

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

Ques – 4

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

Ans -4

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Ques-5

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

Ans-5

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

Ques-6

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

Ans-6

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.