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? (3 marks)

Ans:

Categorical variables with two levels may be directly entered as predictor or predicted variables in a multiple regression model.

It can affect the significance tests of the regression coefficients. Significance tests are based on the standard errors of the coefficients, which depend on the variance and covariance of the predictors.

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

Ans :

Using drop_first=True helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans :

Temp has highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans :

Checked for P-value and VIF,
p-Value should be less than 0.05 and VIF should be less than 5.
Dropped those columns which were not satisfying the conditions considering highly co-related columns
Repeated the procedure.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans :-

Based on final model top three features contributing significantly towards explaining the demand are:
year (0.233)
holiday(-0.098)
Temperature (0.491)

General Subjective Questions

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

Ans :

A linear regression is one of the easiest statistical models in machine learning. Understanding its algorithm is a crucial part of the Data Science Certification's course curriculum. It is used to show the linear relationship between a dependent variable and one or more independent variables.

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

Y=mX+bY=mX+b

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the YY-intercept. If X = 0,Y would be equal to bb.

Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

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

Ans :

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, 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.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

3. What is Pearson's R? (3 marks)

Ans :

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is

known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship
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
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Ans :
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
In the business world, "normalization" typically means that the range of values are "normalized to be from 0.0 to 1.0". "Standardization" typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean. However, not everyone would agree with that.

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

Ans :
In statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It

quantifies the severity of multicollinearity in an ordinary least square's regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. Cuthbert Daniel claims to have invented the concept behind the variance inflation factor, but did not come up with the name.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Ans :
InStatistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.