

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. The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019

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

Ans. drop_first=True is important to use as that extra column created during dummy variable creation.

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

Ans. The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if we consider all the features.

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

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

General Subjective Questions

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

Ans. Linear regression is machine learning algorithm on supervised learning. It is sort of regression where target values is based on independent variable. Linear regression performs the task to predict a dependent variable value Y based on an independent variable X.

$$Y=m_1+m_2x$$

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.

RMSE is cost function of 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.

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

Ans. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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.

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.

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

Ans. If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.