

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)

Answer: Bike demand takes a dip in spring.

Bike demand in year 2019 is higher as compared to 2018.

Bike demand is high in the months from May to October.

Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.

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

Answer: It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate unfurnished. So we can remove it

It is also used to reduce the collinearity between dummy variables.

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

Answer: `atemp` and `temp` both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

Answer: According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

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

Answer: Based on final model top three features contributing significantly towards explaining the demand are:

Temperature (0.552)

Weather: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)

Year (0.256).

General Subjective Questions

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

Answer: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

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

Answer: 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)

Answer: The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

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

Answer: 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.

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.

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

Answer: 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)

Answer: A Q-Q plot is used to compare the shapes of distributions.