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:

1. Categorical variables such as season, yr, season are variables which can be put into a groups based on characteristics. These are used to predict the dependent variable which is the demand for the bikes.

2. Fall season has the highest bookings and Spring has the lowest

3. There are more bookings in the year 2019 than 2018

4. Over all spread in the month plot is reflection of season plot.

5. Clear weather is more optimal for renting

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

Answer

A variable with n levels can be represented by n-1 dummy variables.Even if we remove the first column then, we can represent the data.

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

Answer

temp had the highest corellation coefficient of all variables.

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

Answer

Assumptions were made by plotting the distributions. Normal distribution.

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

Answer

The most effecting variables for the model are:

Temp

Year

Winter

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

Answer

A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables.

Data is divided into a test and training data.

Train data is divided into independent and dependent variables

The linear model is fitted using the training data set.

Incase of multiple features, the predicted variable is a hyperplane instead of a line

The predicted variable is compared with test data and assumptions are made

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

Anscombes quartet comprises of four data sets that have nearly identical simple descriptive statistics but quite different distribution when visualized graphically. The simple statistics consists of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be different from one another.

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

Pearson's R measures the strength of two variables. It is the covariance of two variables divided by the product of their standard deviation. The values are from +1 to -1

- 1 mens total positive linear correlation where one variable is directly proportional to the other variable
- Value of 0 means no correlation
- -1 means a total negative correlation. One variable is inversely proportional to the other variable

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

Scaling of a variable is performed to keep a variable in certain range. Scaling is a per-processing step in linear regression analysis. The reason we scale a variable is to make the computation of gradient descent faster. If the data has small variables(0-1) and big variables(0-1000) than the tike taken by the gradient descent algorithm will become huge.

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

If R Square is 1 then VIF becomes infinite, which means that there is a perfect correlation between the features.

$VIFi = 1/1-Ri^2$

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

(3 marks)

A Q-Q plot is a scatter plot of two sets of quantiles against each other. Its purpose is to check if the two data sets of data are from the same distribution. If the data is from same source then the plot will appear as a line.