Assignment-based Subjective Questions

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

Answer :Categorical variables were holiday, year, season,mnth,weekday,working day and weathersit. These variables have following impact on dependant variable demand.

1.The bike demand was higher in 2019 than 2018

#2. the bike demand did not have much difference even if its working day or holiday or any particular day of week

3The demand for bikes was highest during fall season particularly in Sept and Oct

4. The spike in demand for bikes was higher on Clear Few clouds days

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

Answer: If we don't drop the first column then the dummy variables will be correlated. drop_first=True deletes extra column created during dummy variable creation.

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

Answer: The bike demand has highest correlation to temp and atemp.

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

Answer:

1. I did residual analysis by plotting the error terms to check for normal distribution

2. I dropped variables based on VIF and p-values to avoid overfitting problem

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

Answer: The top 3 variables contributing to demand are year, temperature and Winter

General Subjective Questions

Explain the linear regression algorithm in detail.
Answer: Linear regression is a supervised machine learning algorithm ,that analyses relationship between dependant variable and given set of independent variables i.e increase or decrease of one or more independent variable will also change the dependant variable.

Mathematically it is represented by Y=mx+c where y is dependant variables, xis independent variable, c is constant, m is slope of regression line

2 types of linear regression :simple linear regression and multiple linear regression. Simple linear regression is represented by Y=m0+m1x+c where m0 is yintercept ,m1 is model parameter, c is error coefficient .Multiple linear regression is represented by Y=m0+m1x1+m2x2+m3x3 where x1,x2 and x3 are independent variable.

Explain the Anscombe's quartet in detail.
Answer: Anscombe's Quartet includes four data sets that have identical statistical features, but have a very different distribution and look totally different when plotted on a graph. It is used to determine effect of outliers and other influential observations on statistical properties

What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us " can we draw a line graph to represent the data? "

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is

not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values. • Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. • Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Answer: When R2=1, since VIF=1/(1-r2). R2 = 1 shows perfect correlation between independent variables.To solve this we need to drop a variable which will reduce multicollinearity.

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

Answer:Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.Quantiles are points in your data below which a certain proportion of data fall. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.Q-Q plots are used to find the whether its normal distribution for a random variable.In linear regression Q-Q plot can be used to decide if test and train data came from same population.