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. Holiday/working day - factor does not impact much on bike usage.
2. weekday - Median of people using bike on weekday is very close across all the days.
3. season - People using bike varies with season, as it can be clearly seen that the number of people uses bike during "Fall" season is more. This could be a good predictor.
4. Weathersituation - People using bike on Clear Weather day is more when compared to other weathers and also no one used bike when it rains heavily. This could also be a good predictor.
5. Year - Bike usage has increased considerably in 2019 when compared to 2018.

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

Answer:
When we are creating dummy variable for a certain variable X with y different values, we always has to create y-1 variables. Hence in order to implement that we use drop_first = True.
The reason for y-1 variable is we can represent all the different y values with y-1 values, hence we do that.

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

Answer:
tmp/atmp has equal and high correlation with cnt.

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

Answer:
1. I checked if there is any linear relationship between x and y for few variables of x on the train dataset.
2. Check for homoscedasticity(residuals are plotted against the dependant variables to see if there is equal distribution of variance)
3. Check for normal distribution of residual errors.

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

Answer:
Tmp
Season
year

General Subjective Questions

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

Answer

Linear regression is used to predict a variable(dependant) using an independent variable. If there is only one indpendant variable which predicts a dependant variable then it is called simple linear regression. If there are multiple variables which predicts a dependant variable then it is multiple linear regression. Usually linear regression fits as straight line equation.

Y = mx+ C also represented as (y = Beta(0) + Beta(1) x) incase of simple linear regression for multiple it is.

y = Beta(0) + Beta(1) x + Beta(2) x + …

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

Answer

This theorem explains the importance of visualising data before making any decisions or applying any models. This was first introduced in 1973 with the help of for different datasets having same statistical values, however when plotted, each dataset generated different kind of plot. Hence before attempting to interpret a model, its always advised to analyse the data through visualisation.

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

Answer

It is a method to measure a correlation. It is a number between -1 and 1 which measures the strength and direction (positive or negative) correlation between two variables.

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

Answer

Scaling is performed on the variables where the measuring scales are different. Suppose if a column has values mentioned in range of 0 to 100 and other column with values in range of 100 to 100000. When we build a model, these variables have different range scales. Hence it is important to bring all the variables in a comparable scale. If we don't have comparable scale, some of the coefficients may not fit the model properly. Hence we use standardization or normalization for scaling the variables.

X – variable
Normalisation = (x-xmin)/(xmax – xmin) – all the values will be between 0 and 1.
Standardization = (x-mu)/sigma – this does not fit the values between any two ranges.

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

Answer

VIF for certain variables were infinity. Which means there was a perfect correlation between few independent variables. In this case there would have been R2 = 1 , which lead to VIF = 1/1-R2 as infinity. To handle that we may need to drop one of the variable or we need to combine such variables.

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

Answer

QQ means quantile to quantile plot. This plot helps to determine, how the datasets are distributed. This is used in linear regression to see if there is a normal distribution pattern for the residuals.