BIKE SHARED ASSIGNMENT USING LINEAR REGRESSION

By
Gowtham Ravichandran

SUBJECTIVE QUESTION

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

Ans: From the categorical variables,
- 2019 has the more bike usage compared to 2018.
- LowRainandSnow and MistandCloudy the lower bike due to climate scenerios
- Spring and Winter has the hard the climate so lower bike usage
- July has high temperature and Sep, Mar of spring or snow seasons

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

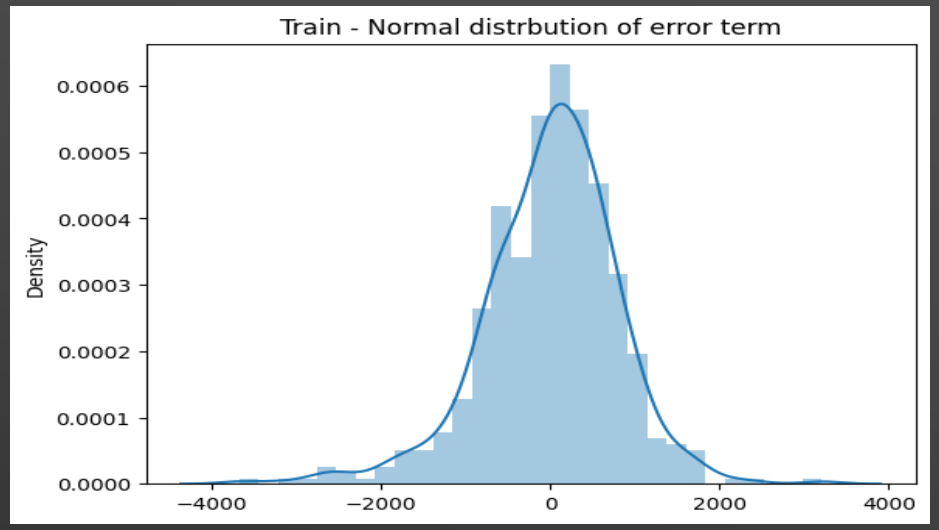
Ans: drop_first = True, then it will drop the first category. So if you have K categories, it will only produce K – 1 dummy variables . The first category is easily able to identified using the other category data and it act as redundant.

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

Ans: from the pair plot the ' atemp' has the highest correlation with the target variable 'cnt'

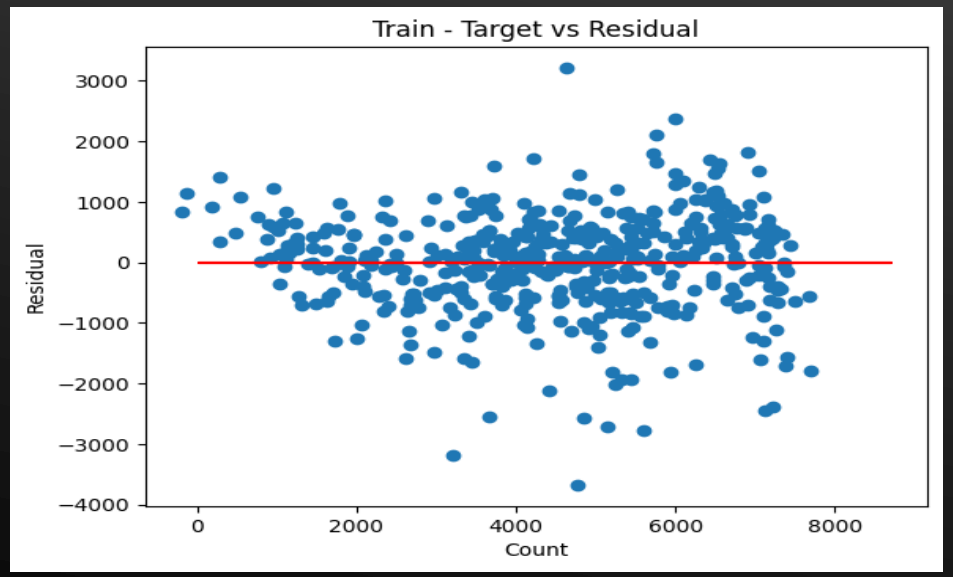
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: 1. Mulitcolinearity between variables.
2. Mean of disrtibuted is 0 and the distribution of errors is normal and plot the histogram of the errors



3.The variance of errors is constant across all levels of the independent variables, plot the residuals versus the predicted values of y is show below

4.Error terms are not be dependent on one another. No auto correlation was calculated by Durbin-Watson_statistic: 2.002 and in range beween 1.50 to 2.50

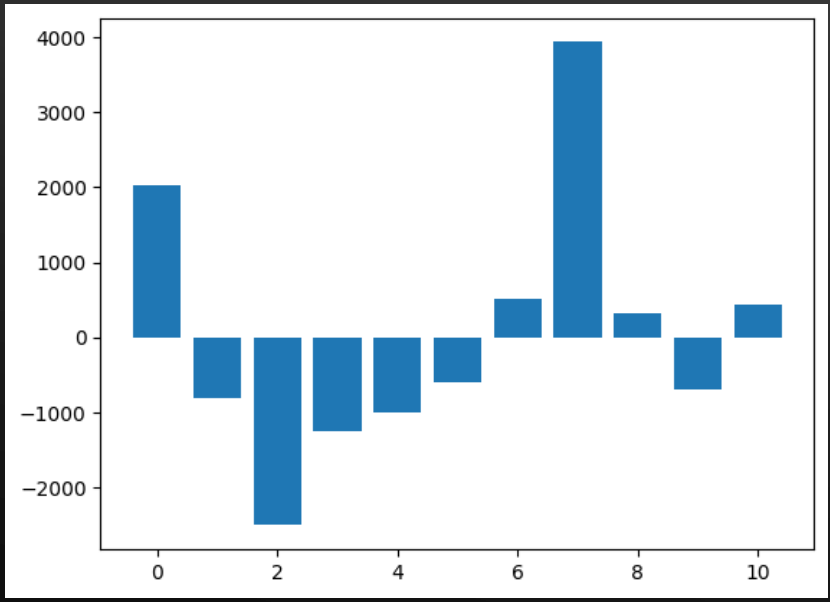


5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Based on the coefficent and below three feature analysis. The first three features are

1. temp
2. 2019
3. Sep

Feature: 2019, Score: 2033.64173
Feature: holiday, Score: -819.74379
Feature: LowRainandSnow, Score: -2510.57190
Feature: windspeed, Score: -1254.66710
Feature: spring, Score: -1010.25520
Feature: Jul, Score: -606.66553
Feature: Sep, Score: 520.28944
Feature: temp, Score: 3937.52655
Feature: Mar, Score: 319.21269
Feature: MistandCloudy, Score: -703.98768
Feature: winter, Score: 440.59056



General Subjective Questions

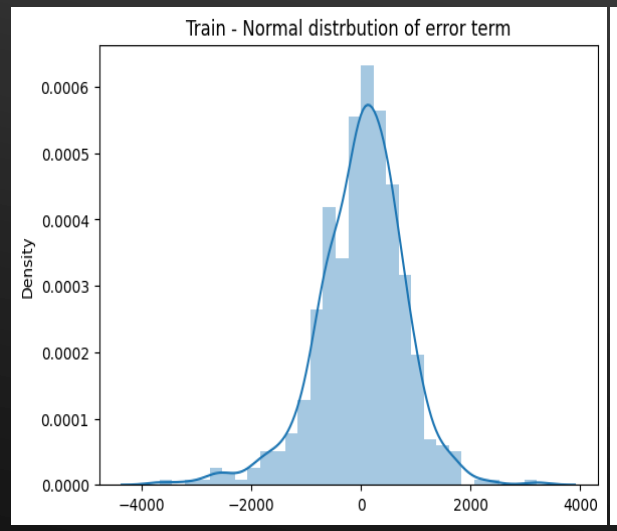
1. Explain the linear regression algorithm in detail?

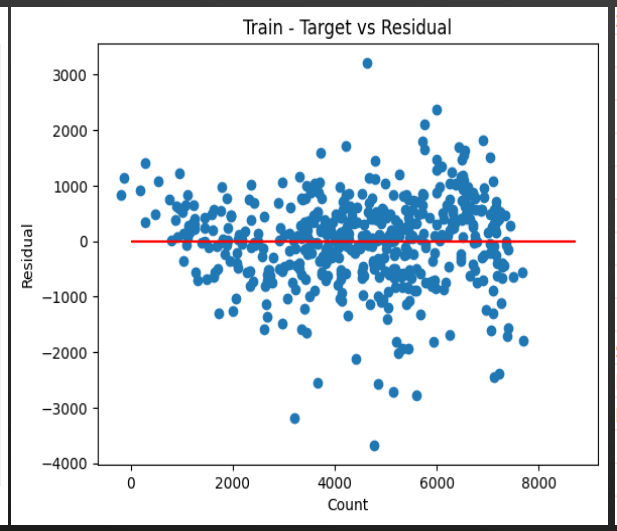
Ans: Linear Regression is supervised classifier. It used to predict or forecast the output for the continuous variables. It predict the value using the historical data of the independent variables.

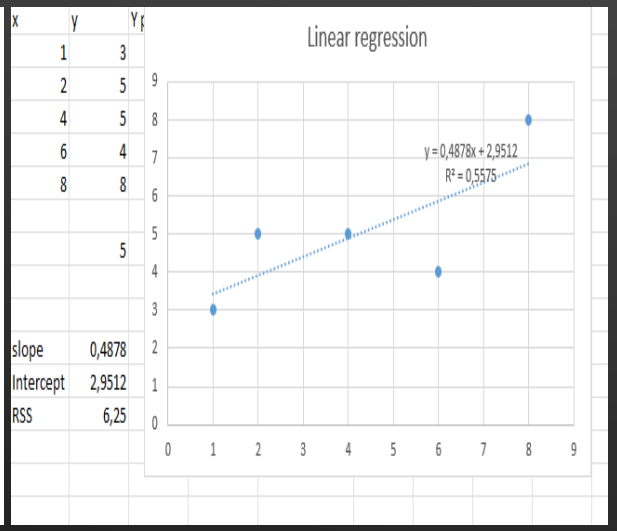
The Linear Regression coefficient of independent variables and try to fit the best fit lines using the least square method for set of paired data

Assumption of linear regression :

1. Mulitcolinearity between variables.
2. Mean of disrtibuted is 0 and the distribution of errors is normal and plot the histogram of the errors.
3.The variance of errors is constant across all levels of the independent variables, plot the residuals versus the predicted values of y is show below. Check for homoscedasticity
4.Error terms are not be dependent on one another. No auto correlation was calculated in range beween 1.50 to 2.5





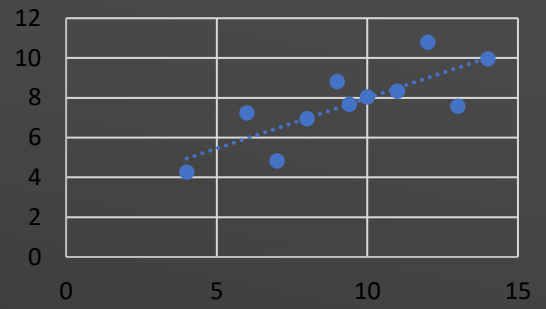


2. Explain the Anscombe's quartet in detail?

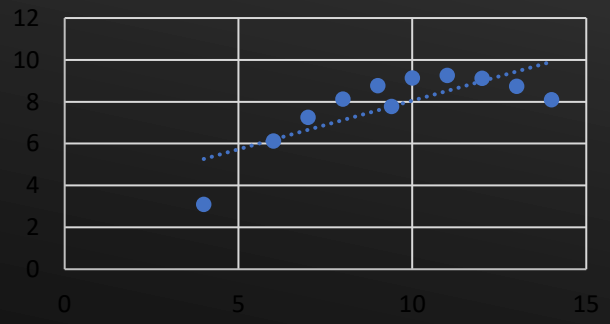
Ans: Anscombe's quartet consist of the four dataset which has the similar statical properties while plotting it shows different in the visualization.
The below dataset shows similar mean and standard deviations, but while plot it show different

	x1	y1	x2	y2	x3	y3	x4	y4
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,7	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,8	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
Mean	9,4	7,679	9,4	7,777	9,4	7,673	9,1	7,562
STD	3,204164	2,037817	3,204164	1,911701	3,204164	2,037875	3,478505	2,129736

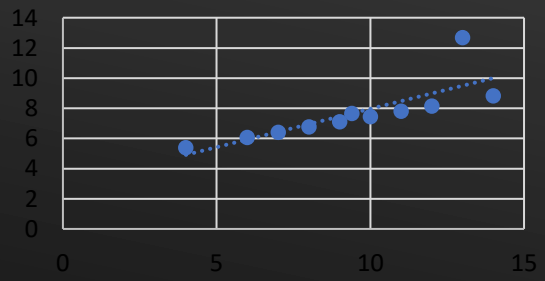
Dataset-1



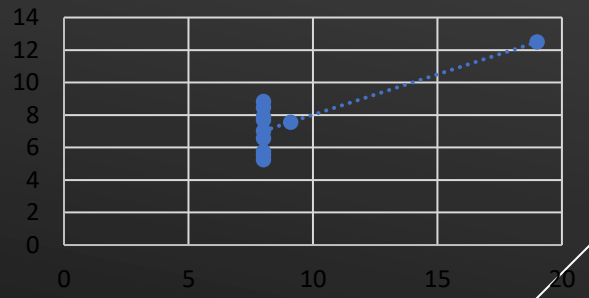
Dataset-2



Dataset-3



Datatset-4



Dataset-1- show the linear relationship between x and y.
Dataset-2- show nonlinear relationship with x and y.
Dataset-3- prefect linear for all the data expect one which is outliers
Dataset-4- one high point to produce a high correlation coefficient

3. What is Pearson's R?

Ans: Pearson's R is the coefficent that used to measure the strength between different variables and their relationship

+1 – Perfect postivie relationship between two variables . One increase and other also will increase
-1 – Perfect negative relationship between two variables . One increase and other will decrease.
0 – no relationship between varibales.

Its range from the +1 to -1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

The scaling is used to normalized the variables into different scale into similar scale. so that computation and memory fo algorthim can be efficently optimized.

Normalization	Standardization
The scale value between range of 0 to 1	It centers data around the mean and scales to a standard deviation of 1
Sensitive to outliers	Less sensitive to outliers
Shape of the original distribution is not changed	Changes the shape of the original distribution
It may not preserve the relationships between the data points	it will preserve the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

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

Ans: The perfectly correlated variables has the R squared value as 1.
The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

substitue in this formula, 1 divisible by 0 . Any value divisble by 0 is infinity.

6. what is a q-q plot? explain the use and importance of a q-q plot in linear regression?

Ans: Q–Q plot (quantile-quantile plot) is a probability plot, that comparing two probability distributions by plotting their quantiles against each other in similar distribution.

A plot using (x,y). The x in plot denotes distribution of the first half quantile and y denotes the distribution of the second half quantile.

In the linear regression, when we have training and test data set received separately and we can confirm using Q-Q plot that both the data sets are from populations with same distributions.