LINEAR REGRESSION ASSIGNMENT

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Ans :

i) The rent of the bike have increased if there is a clear weather.

ii) During the year 2019 the bike rentals have become popular and rent increased when compared to year 2018.

iii) Considering the data when compared to other months the bike rentals have increased in Sep month. Company should expand the business during this time.

iv) Bike rentals have increased more during the non-holidays than holidays

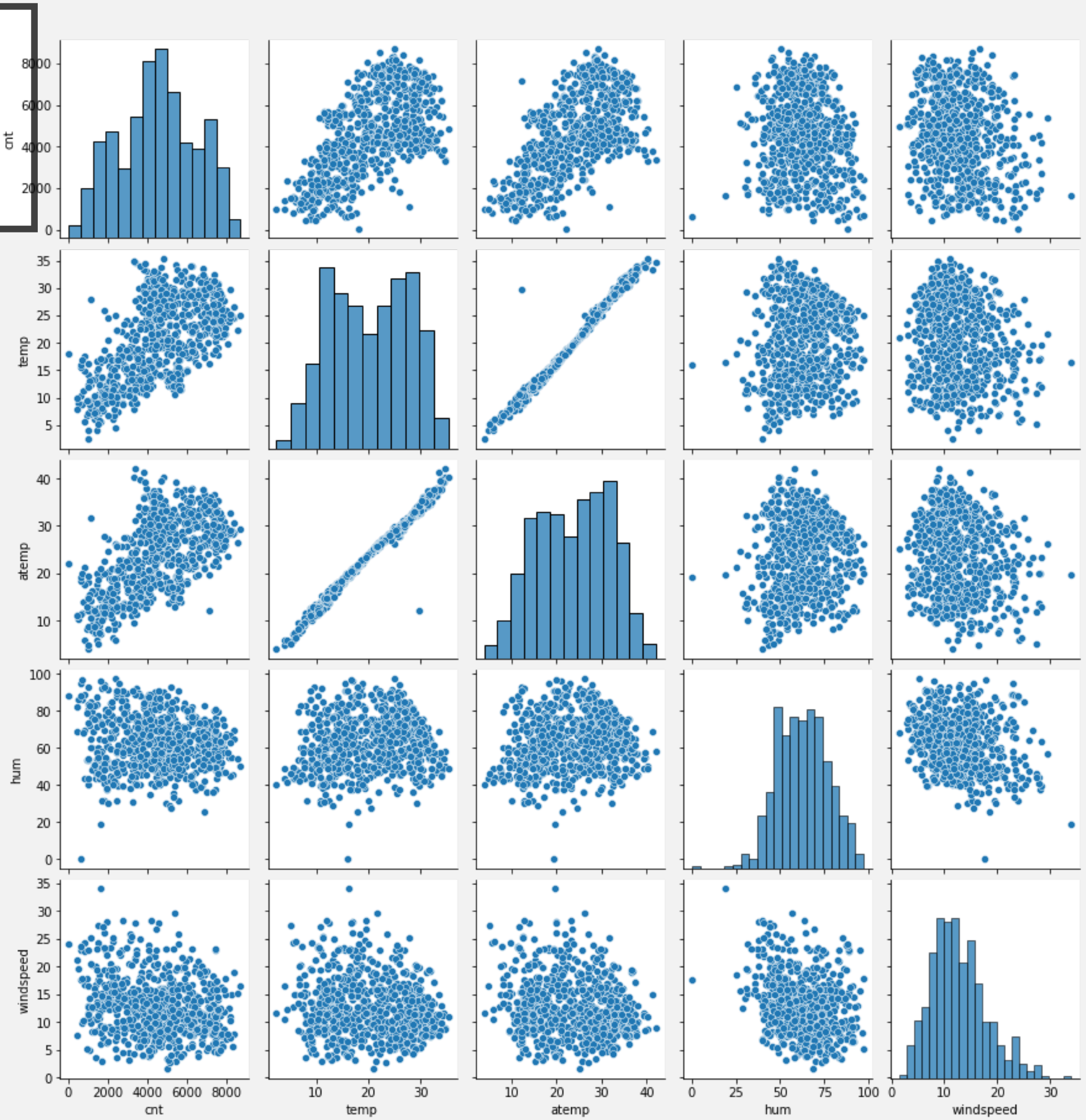
2 Q) Why is it important to use drop_first=True during dummy variable creation ?

Ans : A variable with n levels can be represented by n-1 dummy variables. So , if we remove first column then also we can represent the data. And to avoid redundant columns.

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Ans :Upon observing the “cnt column” has high correlation with “temp column”



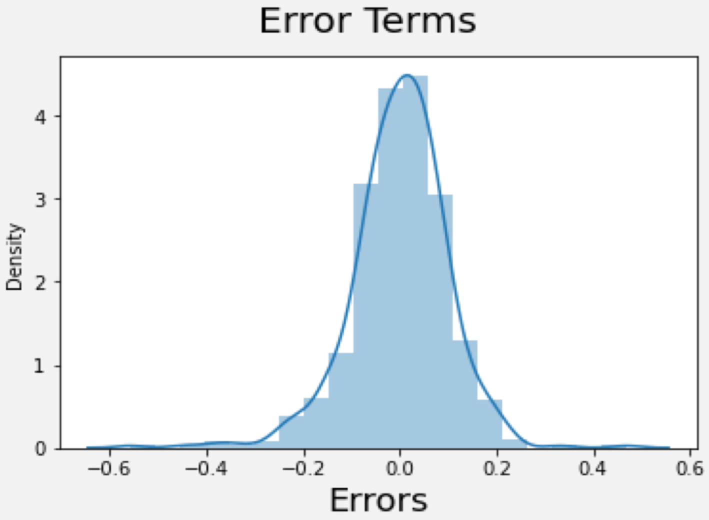
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

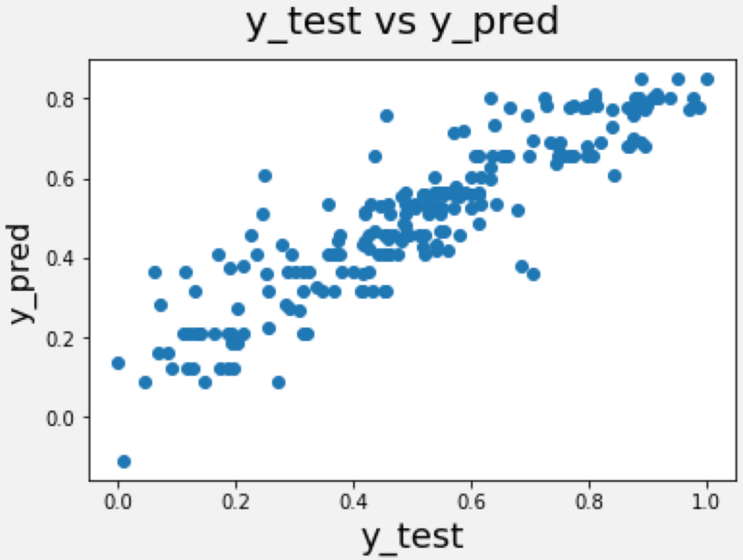
4 Q) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans :

1) Residual errors follows Normal Distribution.

2) Maintain Linear Relation b/w Dependent variable (Test & Pred)





ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

Ans : The below are the top 3 features contributing towards explaining the demands of the shared bikes.

i) Next Year i.e. 2019 (0.2469)

ii) Month August (0.1538)

iii) Month September (0.1937)

GENERAL SUBJECTIVE QUESTIONS

1 Q) Explain the linear regression algorithm in detail?

Ans : Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.

Linear regression can be classified into two types:

i. Simple Linear Regression: Simple linear regression explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$Y=\beta0+\beta1X1 +\epsilon$

ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.

Formula for the Multiple Linear Regression:

$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:

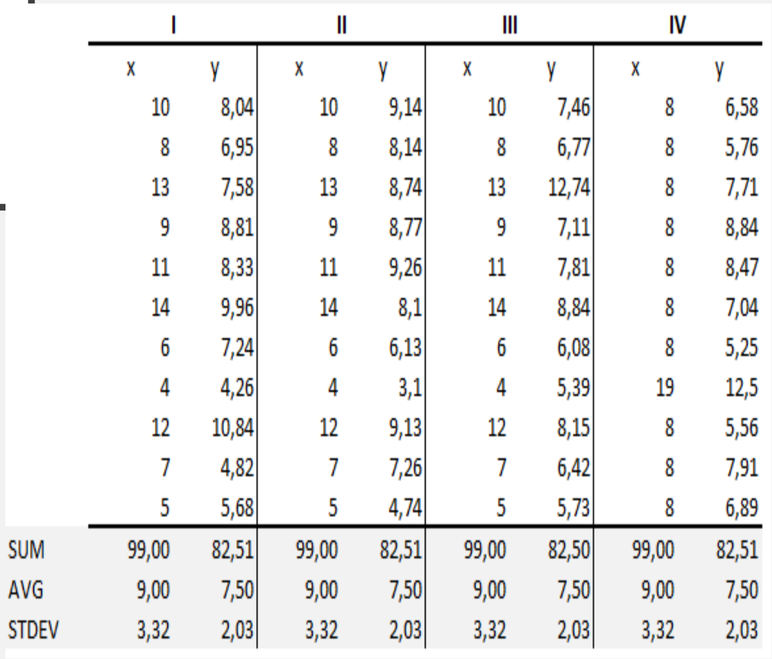
- Gradient descent
- Differentiation

statsmodels or SKLearn libraries in python can be used for the linear regression.

GENERAL SUBJECTIVE QUESTIONS

2 Q) Explain the Anscombe's quartet in detail

Ans : Statistician Francis developed Anscombe. This is a method which contains four datasets, each containing eleven (x, y) pairs. The key thing to note about these datasets is that they share the same descriptive statistics. Each graph speaks a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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	12,74	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	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Dataset I appears to have clean and well-fitting linear models.

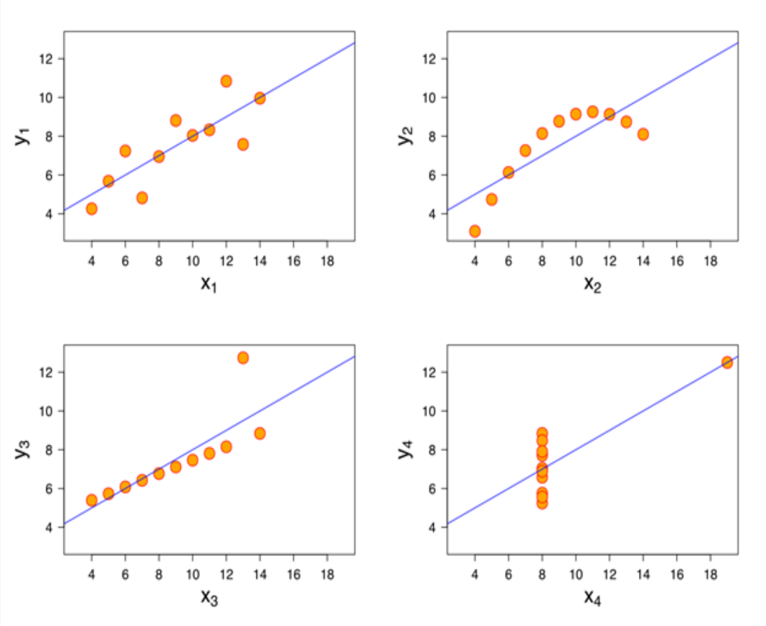
· Dataset II is not distributed normally.

· In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

· Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

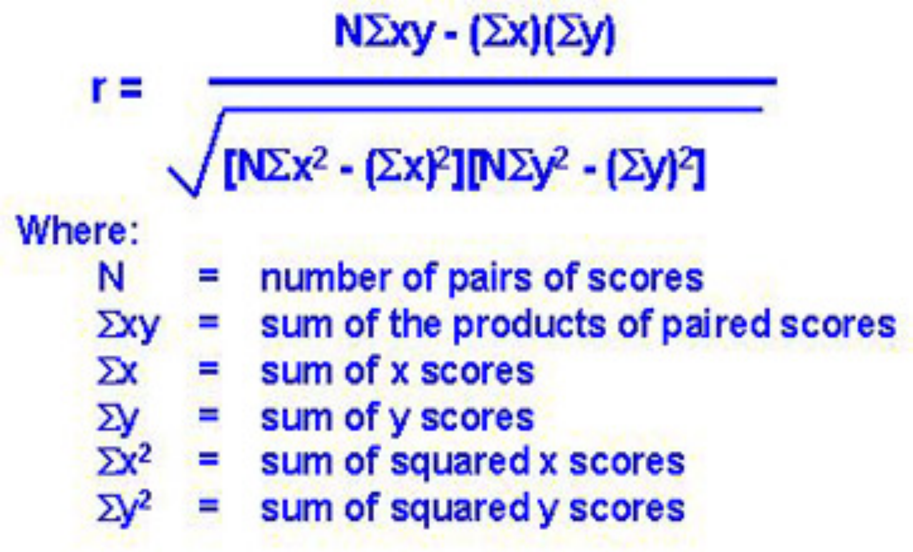


GENERAL SUBJECTIVE QUESTIONS

3 Q) What is Pearson’s R?

Ans : Karl Pearson developed Pearson’s R it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by ‘r’. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Mathematically, Pearson’s correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a “product moment”, i.e. the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Formula :



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

GENERAL SUBJECTIVE QUESTIONS

4 Q) What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

Ans : Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single/same range.

The two most discussed scaling methods are Normalization and Standardization. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Formula of Normalized scaling: x = x-min(x) / max(x) – min(x)

Formula of Standardized scaling: x = x-min(x) / sd(x)

GENERAL SUBJECTIVE QUESTIONS

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

Ans : The value of VIF is calculated by the below formula: $VIF_i = 1 / (1 - R_i^2)$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

GENERAL SUBJECTIVE QUESTIONS

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

Ans : The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

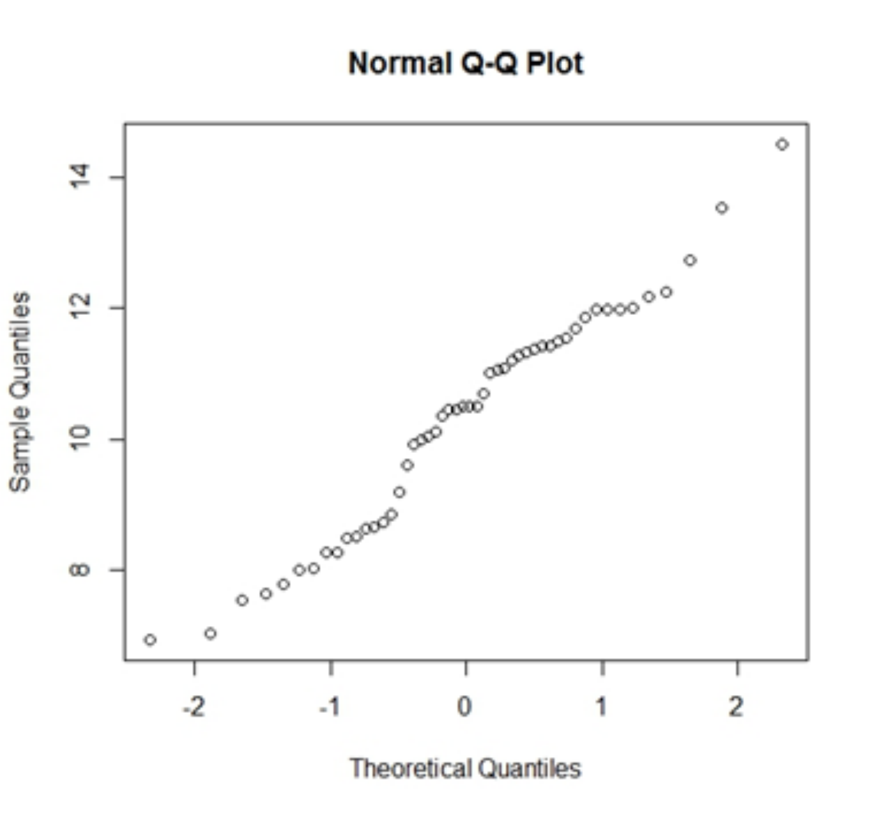
Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.



Thank You!