ASSIGNMENT -LINEAR REGRESSION

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?

- In seasons, fall season has shown the highest demand followed by summer.
- In months, it was observed that the demand is comparatively less in January, February, November and December.
- The demand is higher when the weather is clear and no demand at all when the weather is heavy rain.
- All of the weekdays have similar medians and doesn't exhibit any special pattern.
- For non-holidays the demand is high and falls in closer distribution compared to holidays.
- Demand has increased from 2018 to 2019 significantly.

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

While creating dummy columns, the columns for all the values of variables gets created and we can eliminate the column for one of the variables since that variable can be taken into account from the other dummy column itself. In total, n-1 variables (where n is total unique values of the considering column for which dummy variables are being created for) will have their dummy columns.

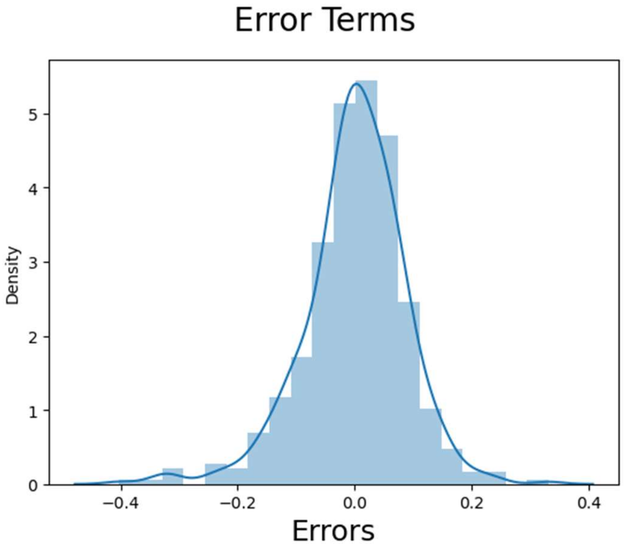
So, it is essential to drop 1 column form all the 'n' columns. Anyone of the created dummy columns can be dropped and drop first drops the first column.

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

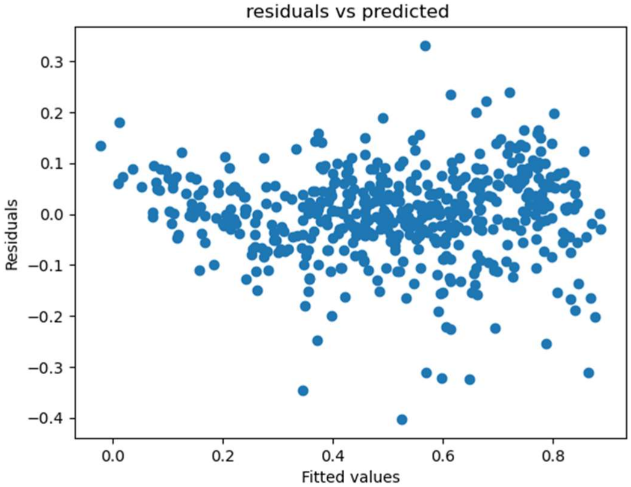
Ignoring registered, casual and instant like columns which are only present as a data but cannot be the cause of the demand, "temp and atemp" has the highest correlation. While we can consider both of them as one since they are highly correlated to each other.

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

Distribution Plot of residuals



Plot of residuals vs predicted values



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

- Temperature (increment in demand)- co efficient = 0.5098
- Weather situation-Light snow, Light rain…. (Decrement in demand)- co efficient = -0.2488
- Year (increment in demand)- co efficient = 0.2305

ASSIGNMENT -LINEAR REGRESSION

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail?

Linear regression algorithm is used for prediction of dependant variable based on the given independent variables. Linear regression comes under Supervised Learning since it uses set of data (usually called 'Training dataset') for finding relationship of independent variables with the variable to be predicted in the end, so that, the algorithm can be applied for predicting the dependant variable for other datasets.

Linear regression works on the principle of Best fit line. Linear regression tries to find a linear relationship between a pair of independent and dependent variables and it is continued for all the in dependent variables. The final model includes the equation involving all the significant variables for predicting the dependant variables.

Residual is the difference between actual value of dependent variable and predicted value of dependent variable by the algorithm. The values of Sum of squares of these residual becomes the deciding factor for specifying how reliable the algorithm is.

The algorithm can be fine-tuned by checking parameters like r-squared, adjusted r-squared for fine tuning the algorithm.

2. Explain the Anscombe's quartet in detail.

Anscombe quartet is a set of 4 graphs depicting the data with same statistical properties line mean,median but different types of distributions, the quartet illustrates the importance of graphing the data and also presents the kind of influence outliers can have on the data and provides a case that only statistical properties won't be able to describe that data completely.

For example for the data

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	12.74	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

Average Value of x = 9 (for all x1,x2,x3,x4)

Average Value of y = 7.50(for all y1,y2,y3,y4)

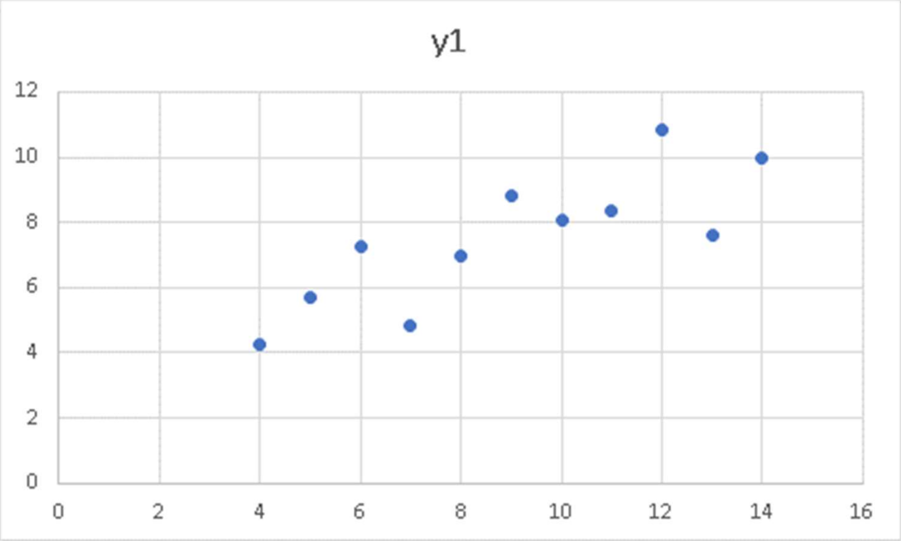
Variance of x = 11(for all x1,x2,x3,x4)

Variance of y =4.12(for all y1,y2,y3,y4)

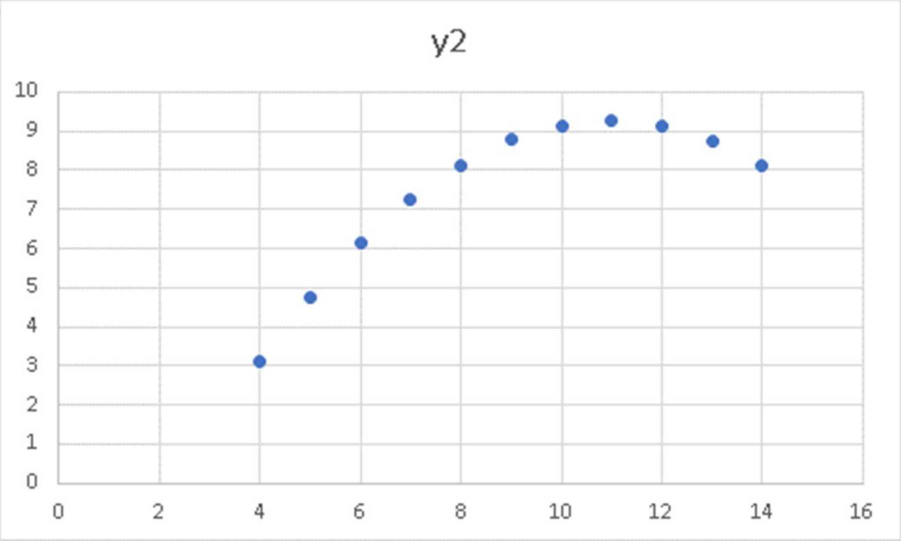
Correlation Coefficient = 0.816

But while plotting the graph

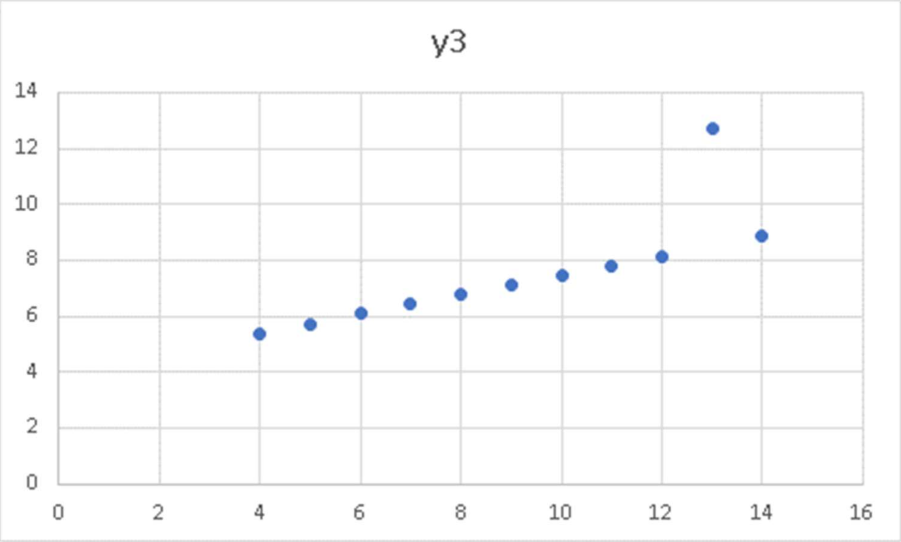
x1 vs y1



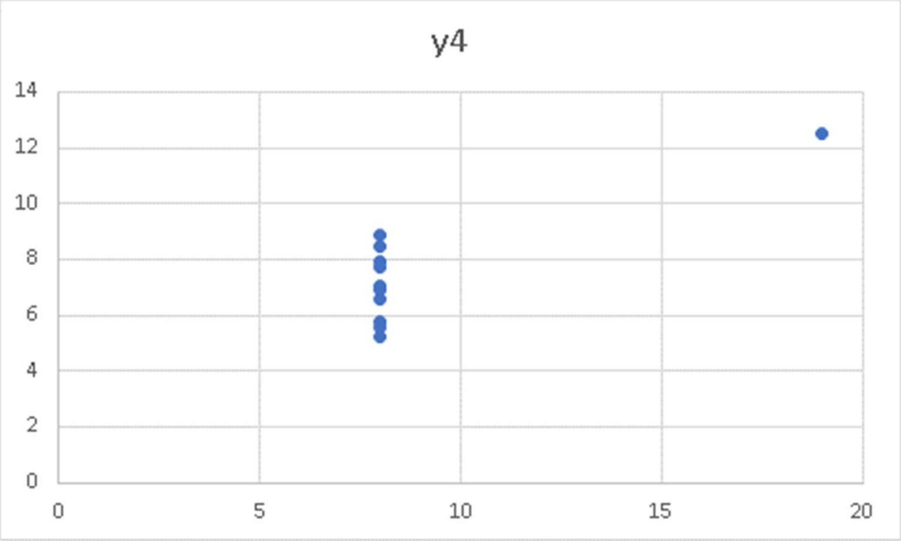
X2 vs y2



X3 vs y3



X4 vs y4



So, above data and graphs illustrates the importance of graphing the data.

3. What is Pearson’s R?

Pearson’s R is a number between “-1 “ to “+1” which indicates the linear correlation between variables.

The range of “0” to “-1” indicates that the variables are negatively correlated and when one of them decreases, the other one increases. The correlation increases from “0” to “-1” where 0 correlation implies the variables are not correlated at all.

The range of “0” to “+1” indicates that the variables are positively correlated and when one of them decreases, the other one decreases. The correlation increases from “0” to “+1” where 0 correlation implies the variables are not correlated at all.

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

   Scaling is an act of normalising data to a certain range. It is a step of pre-processing. When all of the variables exist in similar range, performing linear regression becomes less tedious.

   Normalised scaling uses maximum and minimum values for scaling while Standardised scaling uses mean and standard deviation values for scaling.

   Normalised scaling= (X-X_min)/(X_max-X_min)

   X is the value of variable while X_min and X_max indicates minimum and maximum value respectively.

   Standardised scaling= (X-X_mean)/(std. deviation)

   X_mean is the mean of X variable values; std. deviation is for the X_variable

   Normalised scaling gets affected by outliers since maximum and minimum contains outliers most of the time and they are included in the formula of calculation while Standardised scaling won’t be affected by outliers. Hence, preferred.

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

   VIF= 1/1-(r^2)

   When the variables have perfect collinearity, (r^2)=1, then VIF = 1/0= infinite.
   So, if VIF is infinite, then the variables are perfectly correlated.

   So, the variable in question can be completely explained by other variables in the dataset.

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

   Q-Q plots are is plotting data of different quantiles together and if they are from same distribution, they should roughly fall in a straight line. This is useful to determine,

   i. if they are from the same distribution or not.

   ii. their scaling

   iii. the distribution shape both data.