Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1.

- The demand has risen from 2018 to 2019, with the median increasing nearly 1.5 times,
- From the seasons column it is evident that Fall and summer experience higher demand than the rest of the seasons while in spring, the demand is at lowest,
- From the months' column we observed that the demand in initial months is low which then increases and reaches the peak around September from where on it again starts to decrease, with January being the month with the least demand,
- Also the demand for bikes on holidays is lower as compared to the same for non-holidays,
- When the weather is clear, the demand is comparatively more than when the weather is misty or cloudy, whereas it gets plummeted if it's precipitating.

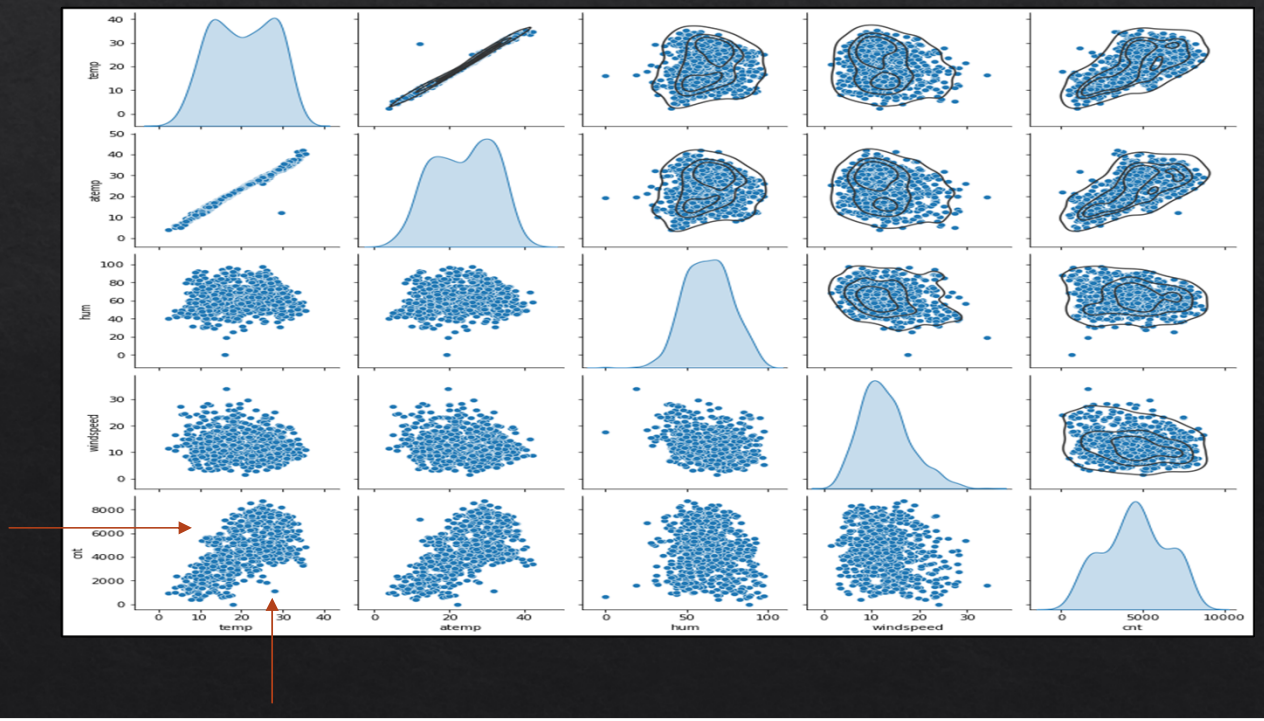
Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans2. When we create dummy variables using pd.get_dummies function, we are converting the categorical column into simple yes-no variables, whose values are in *encodings of 1 for yes and 0 for no*, and now this function creates as many extra columns as there are values in that particular column, so in case of a simple event say like a coin toss where we need to know the outcome, since there are only two possible outcomes so if it is not tail its a given that it would be heads and vice versa, this same logic works for n number of values where we only require (n-1) number of those values to figure out the occurrence of all n values in the form of dummy variables (encodings 1 and 0 for yes and no respectively) so it makes more sense to load only those required n-1 number of columns in our dataset for n values, this is the function of drop_first = True, the lack of which would result in the presence of redundant data.

This redundant data makes it harder for our Algorithm/model to converge or fit, and this may even result in overfitting or higher VIF values among other variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans3. Temp (Temperature at that time) or atemp (Temperature as felt at that time) has the highest correlation with the target variable, and both can practically be considered a single entity.



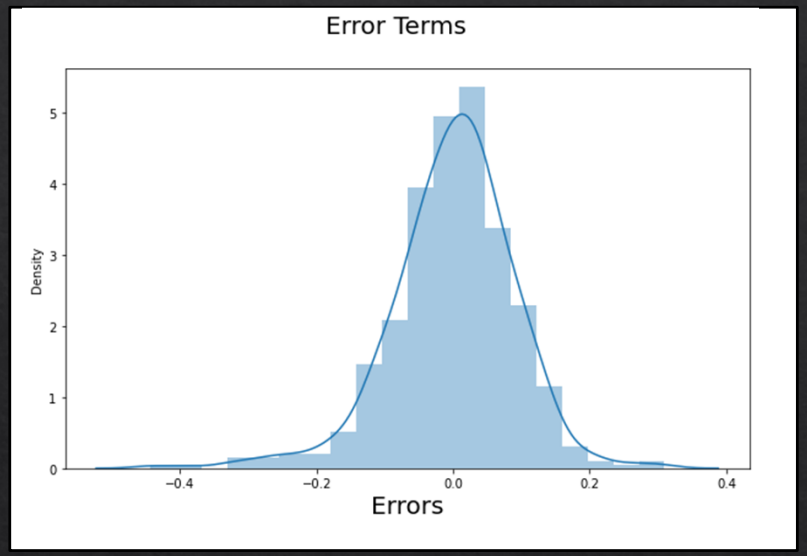
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans4. The Linear Regression consists of following assumptions,

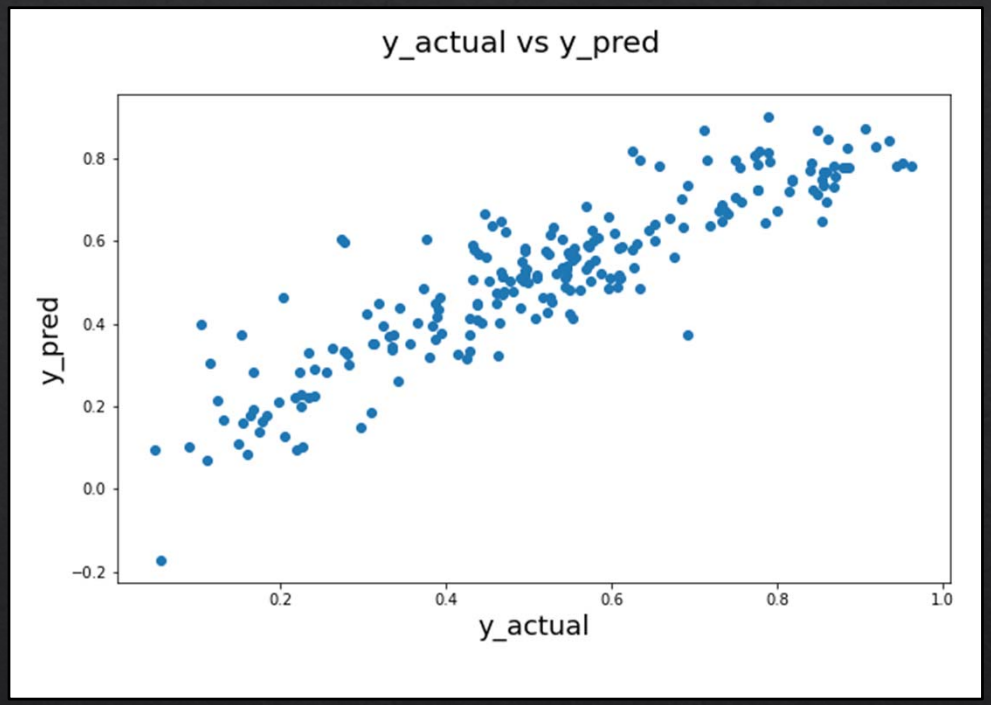
- Linear Relationship → It states that the relationship between independent and dependent variables has to be linear, which can easily be validated using a scatter plot in case of one independent variable or pair plots for multiple independent variables, and we did the same,
- Multivariate Normality → It states that all error terms must be distributed normally with a mean equal to 0, which can easily be verified using a Q-Q plot or by plotting the distribution of error terms, and we used a distribution plot,
- Multicollinearity → According to these assumptions, all independent variables should not be highly correlated to each other, which can easily be verified using a correlation matrix or heatmaps same, and also VIF can be used to identify Multicollinearity, we used VIF, lower the VIF less correlated the variables are with each other, and the upper limit specified by us is 5,
- Homoscedasticity → It states that all residuals must be equal across the regression line, and the same can be verified using a scatter plot between error terms and regression line, and We did the same.

(cont….)

Multivariate Normality



Homoscedasticity



Multicollinearity →

	Features	VIF
2	atemp	4.99
3	windspeed	3.83
5	season_Winter	2.62
0	yr	2.06
4	season_Summer	2.04
10	mnth_Nov	1.81
6	mnth_Aug	1.59
12	weathersit_Misty	1.57
7	mnth_Dec	1.41
11	mnth_Sep	1.35
9	mnth_Jan	1.29
8	mnth_Feb	1.26
13	weathersit_Precipitating	1.09
1	holiday	1.06

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans5. So based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are -:

- `$atemp$` or the temperature that we feel, contributes highest in demand for bikes, and with a unit increase in atemp, it is worth noticing that demand increases by 0.4432 units, given rest have kept constant,
- `$weathersitPrecipitating$` or when its either raining/snowing/thunderstorms outside the demands get reduced, and it is worth noticing that demand decreases by 0.2875 units for a unit increase in the former, given rest have kept constant,
- `$yr$` contributes significantly to the demand for bikes, with a unit increase in a year, and it is worth noticing that demand increases by 0.2349 units with a unit increase in a year, given that the rest have kept constant, meaning the demand for bikes is rising annually.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

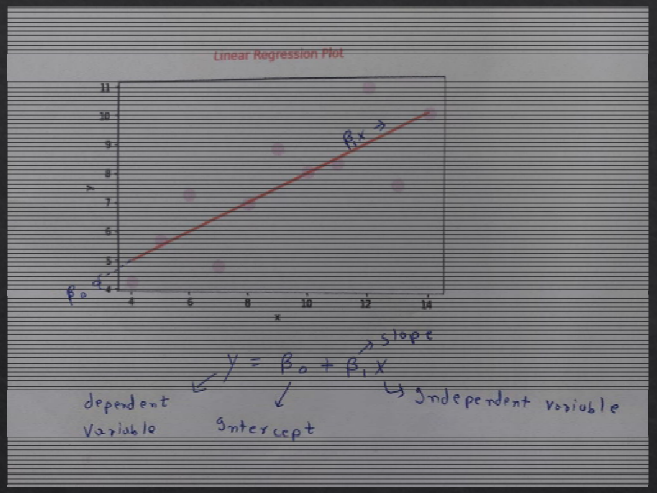
Ans1. Linear Regression is a regression technique based upon Machine learning wherein prediction is made upon target variable using independent variables. It not only provides us with a prediction but also with the strength of those predictions and the significance of those predictors, The most simple algorithm is the regression equation with one independent variable, in which case the equation becomes,

$y = B_0 + X^* B_1$, here y is our dependent variable, X is one independent variable, B_0 is intercept and B_1 is the coefficient (slope) of X.

when the number of independent variables increases, the more complex the equation becomes, and the product of those independent variables with their coefficients (slopes) is added on the right side of the equation.

(cont.....)

Simple Linear regression plot, with regression line in red and datapoints in pink



In general regression is done in following ways

1. Simple Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Unsupervised Regression

Assumptions of Linear Regression :-

1. Linear Relationship → dependent and independent variables must follow linear relationship,
2. Multivariate Normality →all error terms must be distributed normally with mean equal to 0,
3. Multicollinearity → all independent variables must not be highly correlated to each other,
4. Homoscedasticity → all residuals must be equal across the regression line (equal variance).

- Linear Regression models are in most cases based upon the least square approach, where the squared error (sum of the difference between all predicted and actual values) have kept at a minimum, and it can be done by either Gradient descent method, which minimizes the cost function (optimization) or by simply differentiating the equation and equating it with zero (Minima),
- Then R squared value is used to determine the accuracy of the fit of regression line, it is also called as coefficient of determination, which is simply the ratio of Explained variation to the Total variation,
- But since the R squared method does not take into account the number of features used, it is best to use the adjusted R square method, which penalizes for number of features.

Q2. Explain the Anscombe's quartet in detail.

Ans2. In 1973 statistician Francis Anscombe demonstrated the need for visualization or graphically representation of the data by using identical or data with relatively same descriptive statistics, and when he plotted this dataset, he observed that these plots were following a completely different distribution from each other,

```
In [14]: # Anscombe's Quatret

         df
```

Out[14]:

	x1	y1	x2	y2	x3	y3	x4	y4
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

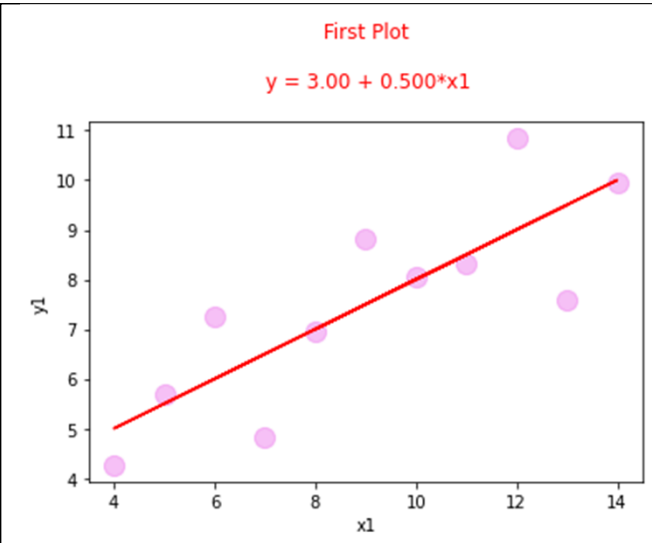
(CONT….)

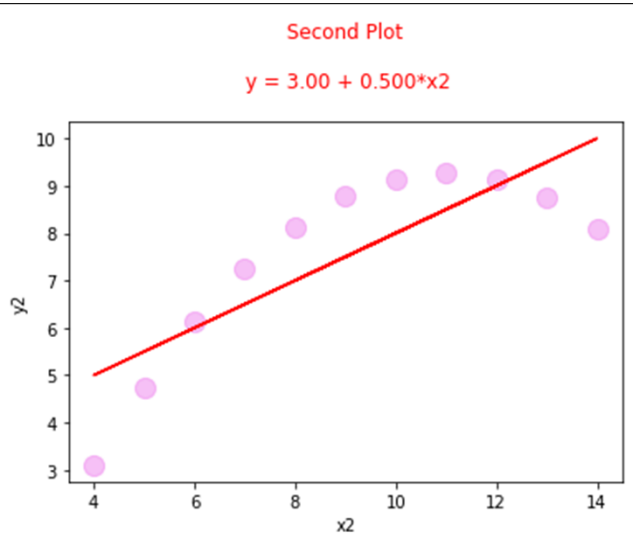
Here the four datasets seem to have identical descriptive summary, with each data set having 11 points and properties as follows :-

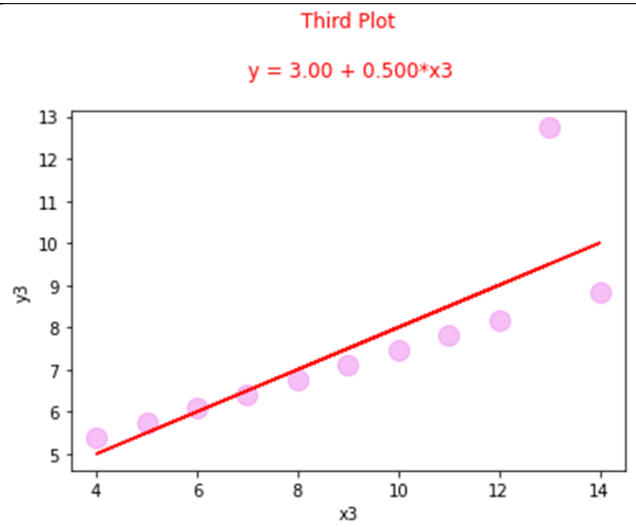
```
In [15]: df.describe()
```

Out[15]:

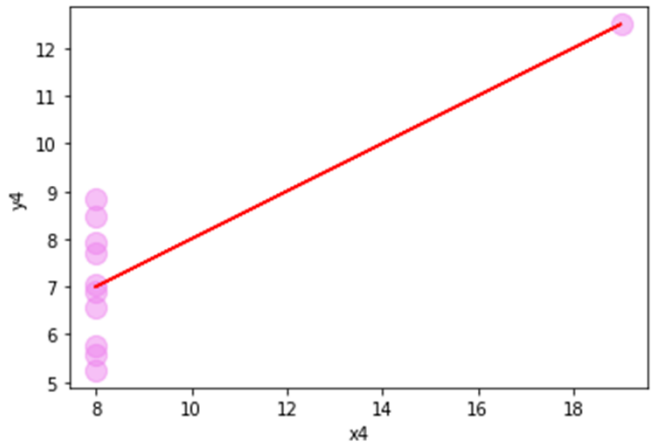
	x1	y1	x2	y2	x3	y3	x4	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	9.000000	7.500909	9.000000	7.500000	9.000000	7.500909
std	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
min	4.000000	4.260000	4.000000	3.100000	4.000000	5.390000	8.000000	5.250000
25%	6.500000	6.315000	6.500000	6.695000	6.500000	6.250000	8.000000	6.170000
50%	9.000000	7.580000	9.000000	8.140000	9.000000	7.110000	8.000000	7.040000
75%	11.500000	8.570000	11.500000	8.950000	11.500000	7.980000	8.000000	8.190000
max	14.000000	10.840000	14.000000	9.260000	14.000000	12.740000	19.000000	12.500000







- The first plot signifies that X and y follows a simple linear relationship,
- The second plot signifies there is some non-linear relationship,
- Third plot, though follow a linear relationship, it's completely different from others,
- The Fourth plot shows there is no relationship between variables.



Q3. What is Pearson's R?

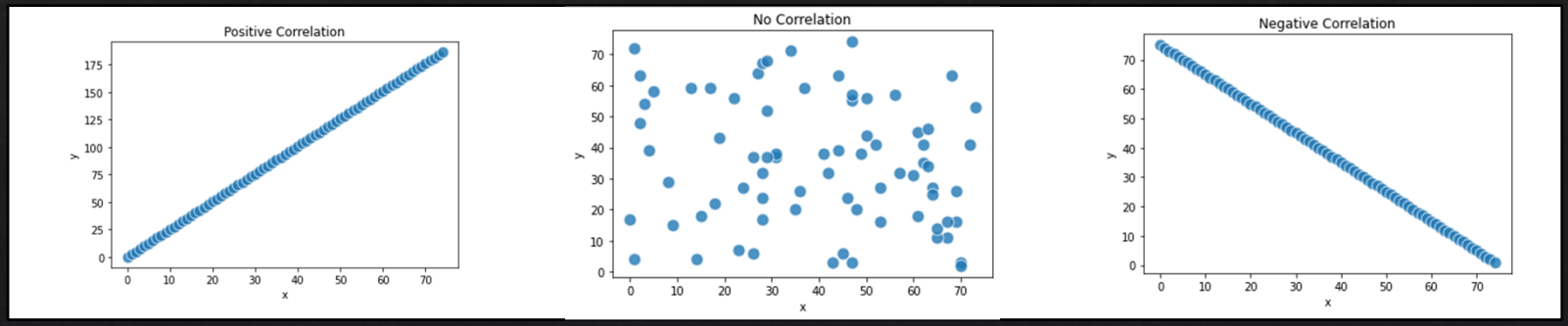
Ans3. Pearson correlation (Pearson's R) is also known as the coefficient of correlation, and it signifies how much linearly related two numerical variables are, by quantifying the relationship between them, it ranges from -1 to 1, with the negative sign signifying the inverse relation between the two variables,

It is measured by taking the ratio of covariance of two variables and the product of their standard deviation, The closer the value is to 1 or -1, the stronger is the correlation between these variables,

The closer the value is to 0, the weaker is the correlation between these variables.

$$\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}}$$

← The Formula to calculate Pearson's R



Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4. Scaling is an intermediatory step taken while processing the data for model building, done essentially to ensure that all the features in our dataset are comparable and straightforward to interpret by transforming the values.

While scaling, we normalize the dataset within a specific limit since our dataset may contain features having inconsistent dimensions regarding other features, an example of which could be that in a dataset, a feature called 'temperature' has values that range from 2 to 36, at the same time values for a feature called 'demand of bikes for renting' ranges from 2000 to 8500, these values are incomparable therefore our algorithm will take a long time to converge, to avoid this scenario we scale both these features and normalize their values thus reducing the time taken by our model to converge.

Also scaling does not have any effect on the accuracy or the statical significance of the model.

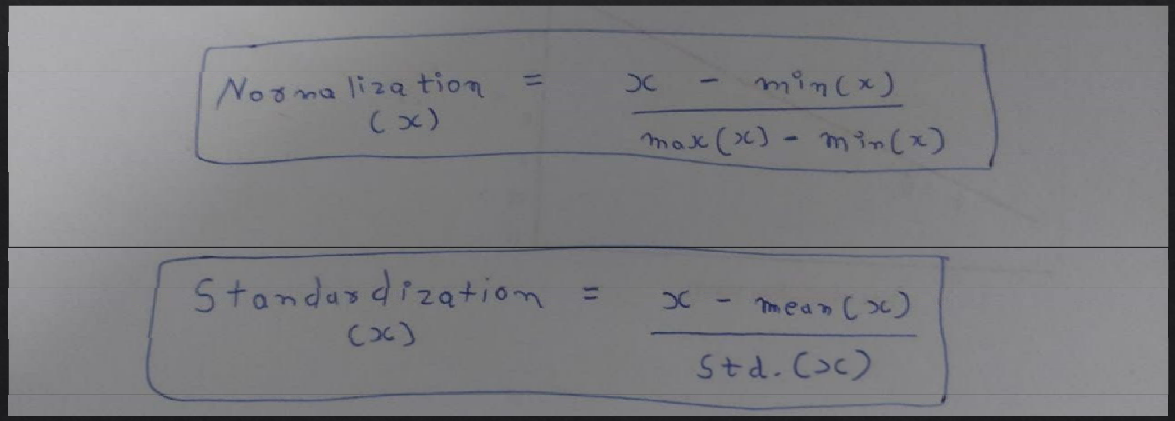
Scaling is achieved in two ways -:

1. Normalization / MinMax scaling and
2. Standardization

(cont...)

- Difference between Normalization and Standardization

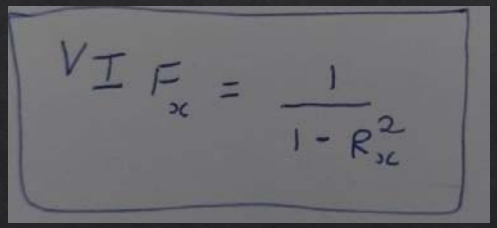
Normalization	Standardization
Basically, rescales the values between 0 and 1	Basically rescales the values such that mean is 0 and standard deviation is 1
Should be used mostly when dataset does not follow normal distribution	Should be used mostly when dataset follow normal distribution
Outliers affected	Outliers not affected



Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans5. VIF is a scale that helps us determine the multicollinearity among all the independent variables present in our dataset by providing a coefficient quantifying the same.

VIF is also called as reciprocal of the tolerance, and is calculated for say `x' feature as,



here, R^2 represents the coefficient of determination of the `x` feature on all other independent variables.

The higher the value of this coefficient for a feature, the more is the multicollinearity of that particular feature with other independent variables.

In some cases, the VIF of a certain feature can become infinite, which indicates that there exists perfect collinearity between that feature and the rest of the independent variables, meaning that using the rest of the variables, we can explain all the variance of data in that particular feature, thus making that same feature redundant.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans6. Q-Q plot or quantile-quantile plot is a graphical representation of probability distributions between two quantitative variables, with the idea behind a Q-Q plot in linear regression is to determine whether the test and train subsets belong to the same dataset or population given that both these subsets must have an identical scale consistent with the dataset.

Q-Q plots also tells us about the type of theoretical distribution from which our datasets / subsets have emerged.

- This plot helps in identifying outliers, changes in the shape of distribution or its symmetry or even its location,
- This plot does not rely on size, therefore it can be used on any number of samples with varying sizes,

A Q-Q plot is made using plotting the quantiles of one subset against another in a scatterplot.

They are generally made for test subset and predicted variables to ascertain that they both are part of the same distribution, which makes Q-Q plots very important in linear regression.

A good Q-Q plot is one in which all quantile points lie close to or on the line passing through the origin of the same plot with an angle of 45 degrees from either axis.