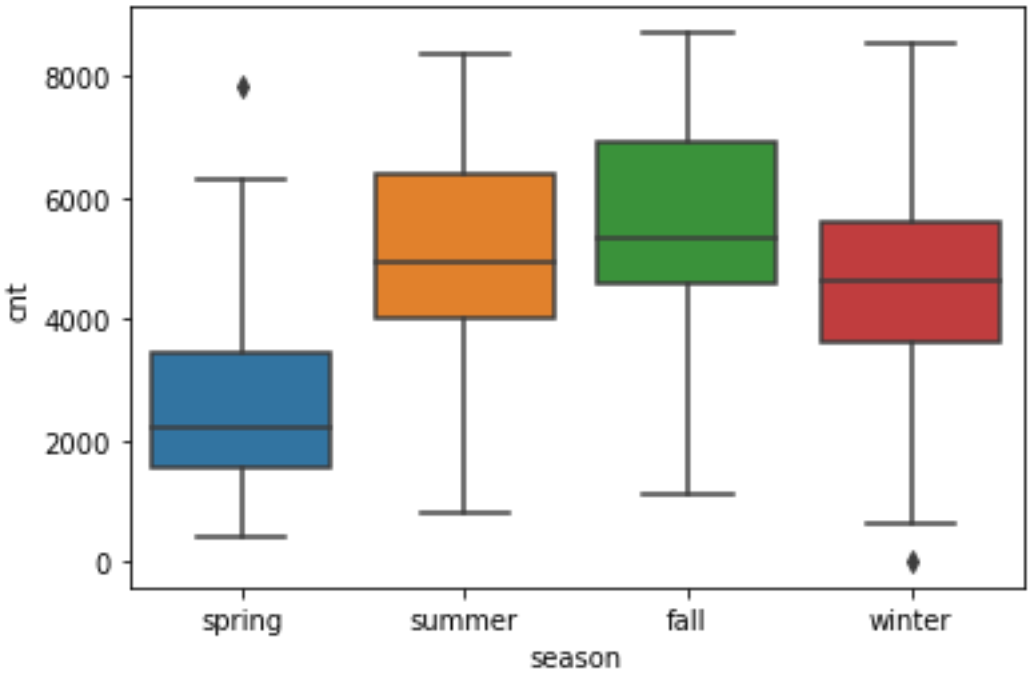
Assignment-based Subjective Questions

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

Ans:

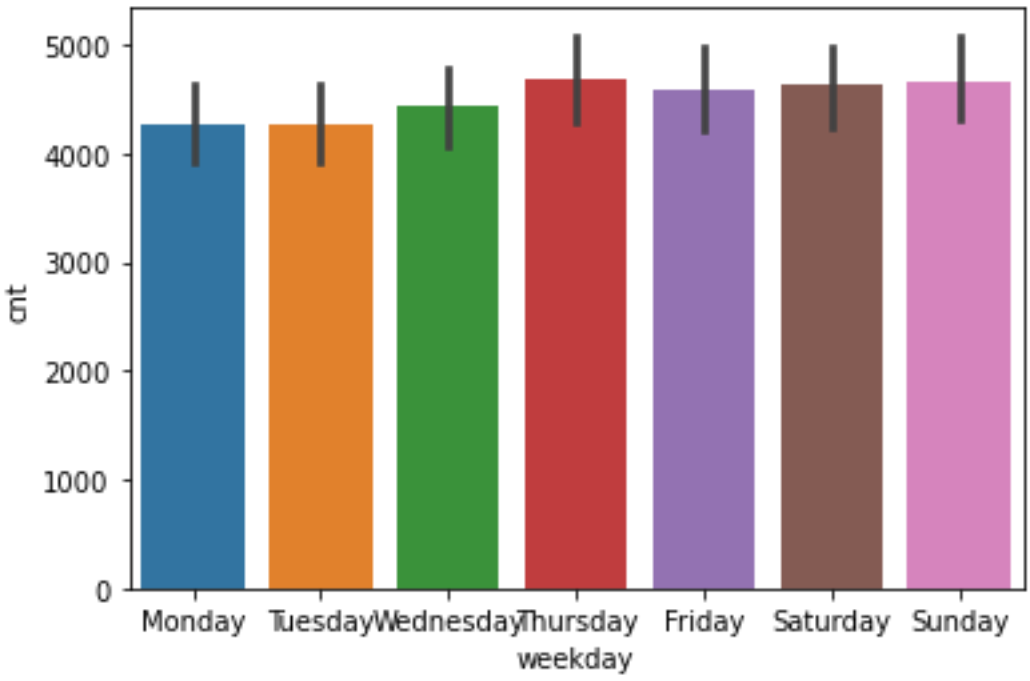
- Impact of categorical variable ‘season’:

Count of rental bikes is the highest in fall season followed by the Summer and the Winter. In the Spring time demand is the least.



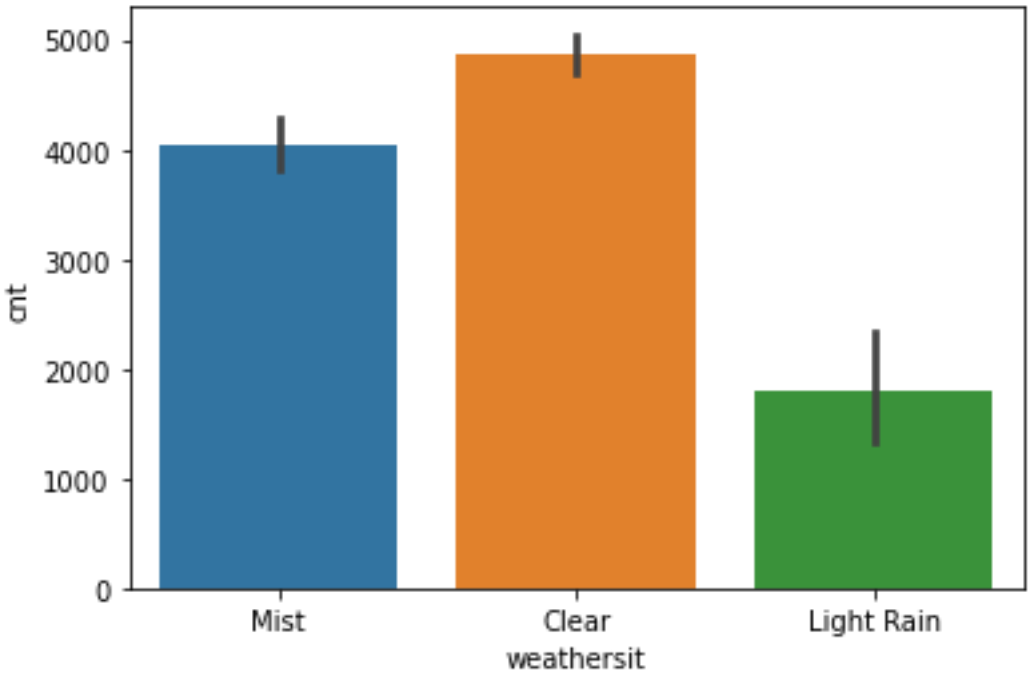
- Impact of categorical variable ‘weekday’:

The count of rental bikes is slightly higher on Thursday than other days. On Friday , Saturday and Sunday the count is almost same. Monday and Tuesday has little less demand. Avg. Demand of rental bike lies between 4000 to 4500.

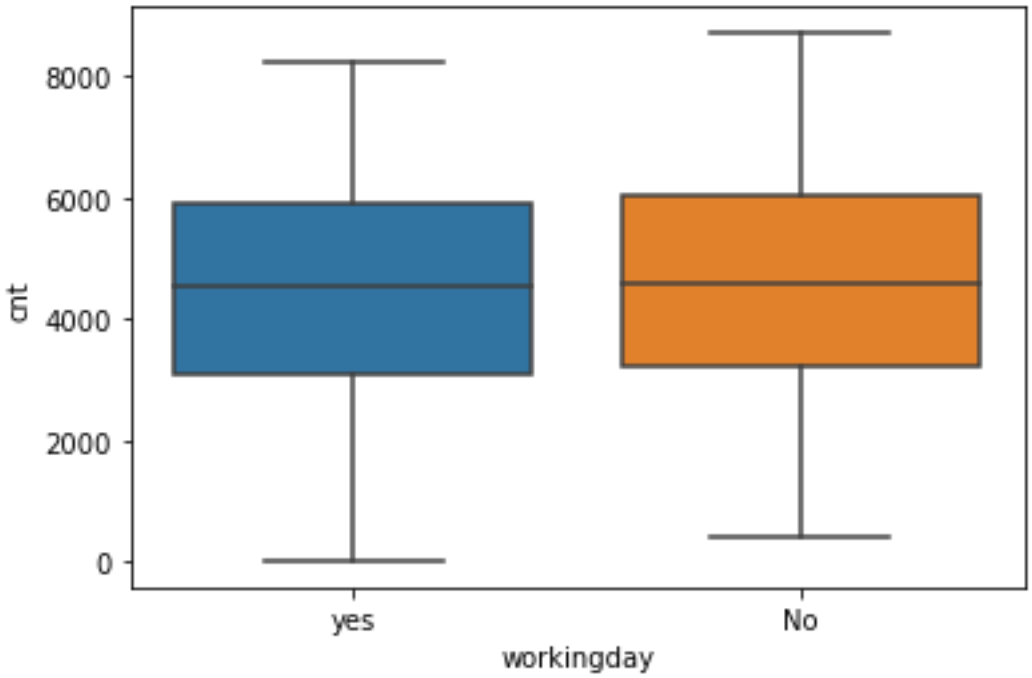


- Impact of categorical variable ‘weathersit’:

Count of rental bike goes highest in clear weather and decreases in rainy day.

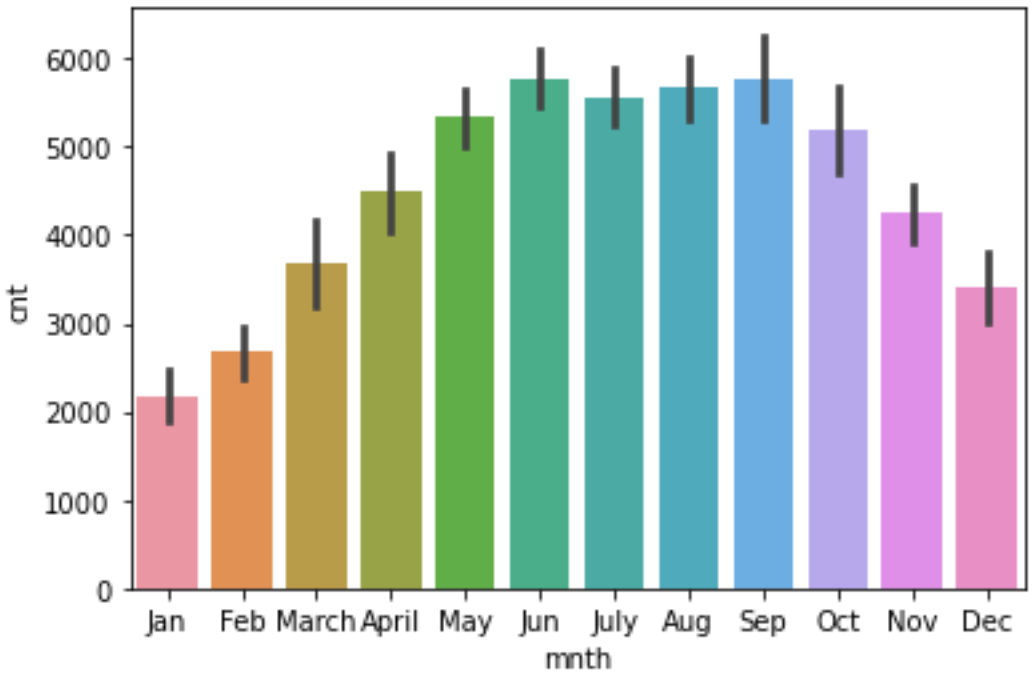


- Impact of categorical variable ‘workingday’:



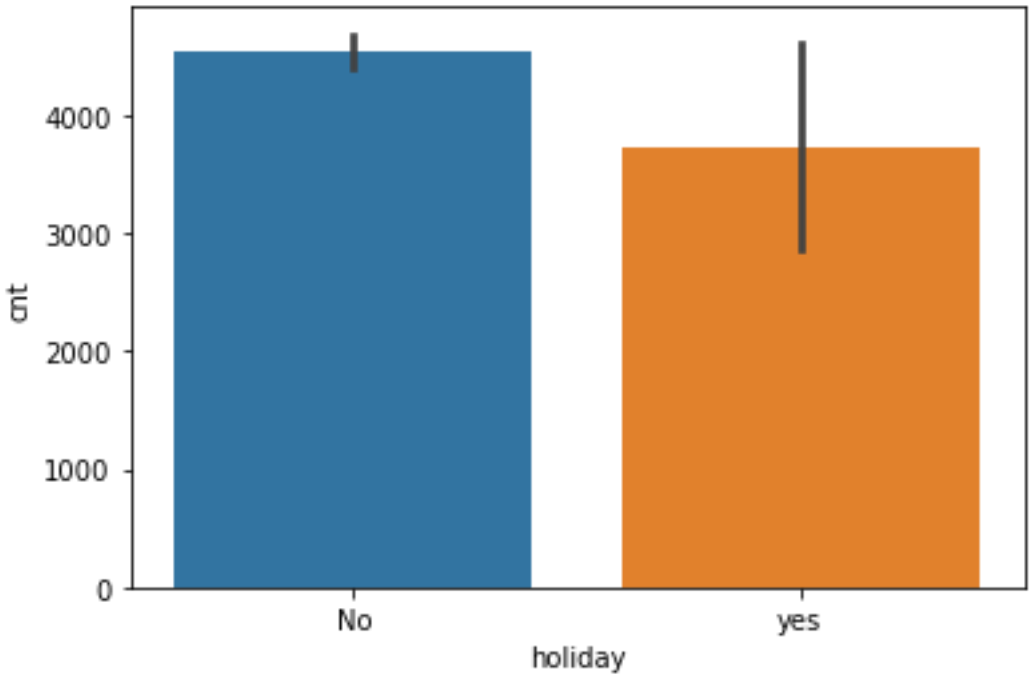
There is slightly high demand of shared bikes on non-working day

- Impact of categorical variable ‘mnth’:



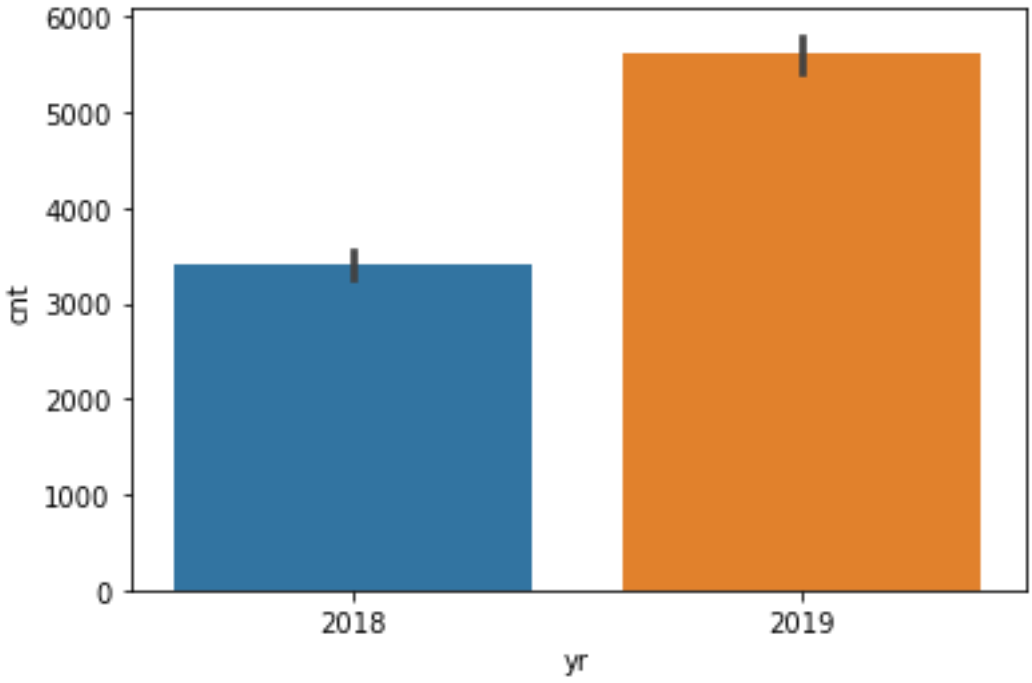
Demand straightly increases up to month of Jun then deteriorate but again rise is shown from July to August and then again deteriorate continuously till December with lowest value.

- Impact of categorical variable ‘holiday’:



Demand is high on non-holidays,

- Impact of categorical variable ‘yr’:



Demand increased sharply in the year 2019.

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

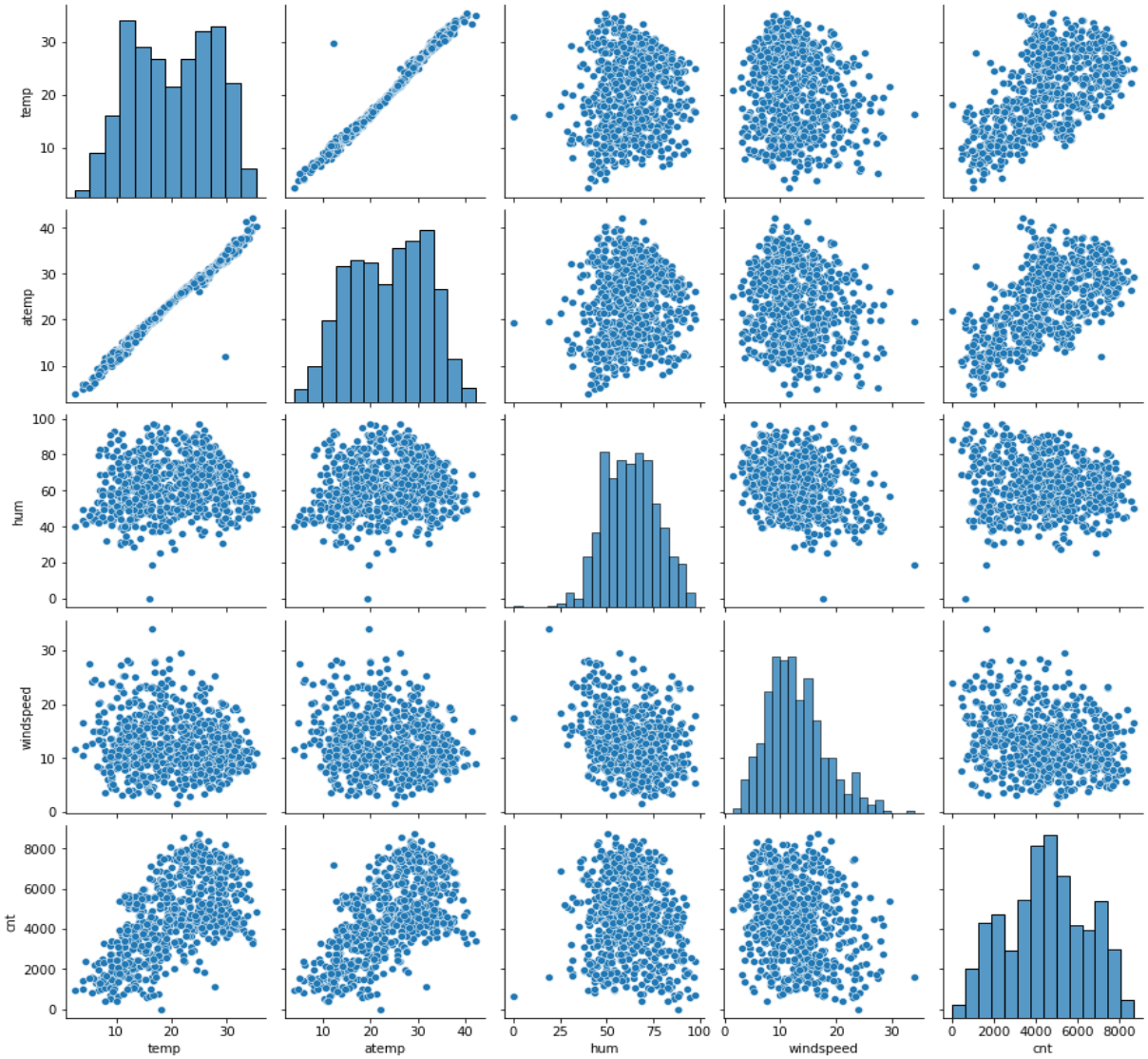
   Ans:

   This command is used to create (level -1) number of dummy variables Instead of creating same number of dummy variables as many as levels are there under a categorical variable . Thus the number of unnecessary column can be reduced.

   Exmp: Suppose there are 3 levels named men, women and children under category variable.10 for men & 01 for women is assigned then automatically 00 will denote rest one. Thus creating two dummies 0 & 1 for 3 levels easily can reduce one column for another dummy variable.

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

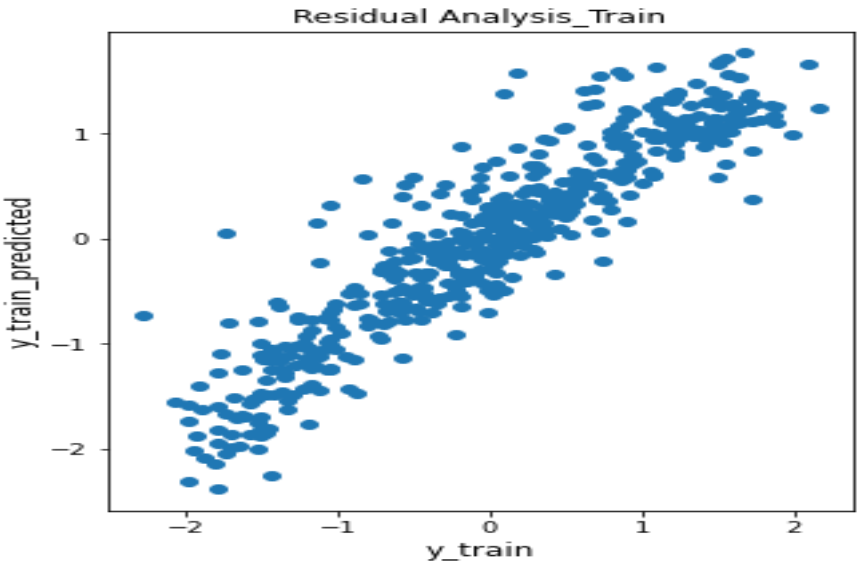
   Ans: ‘temp’ and ‘atemp’ both variables have the highest correlation with target variable.



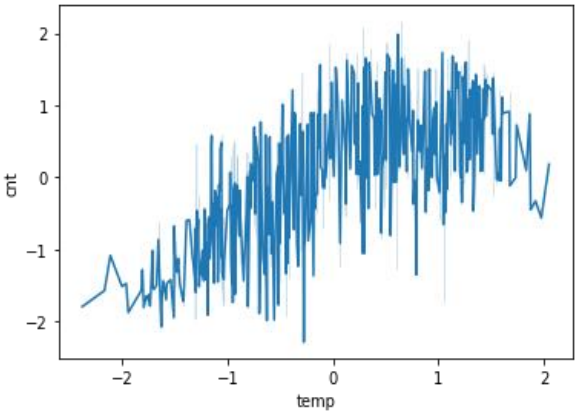
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

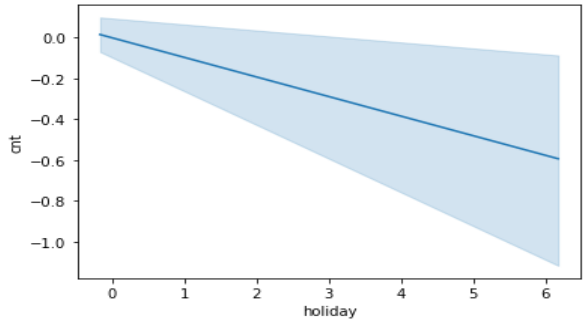
Ans:

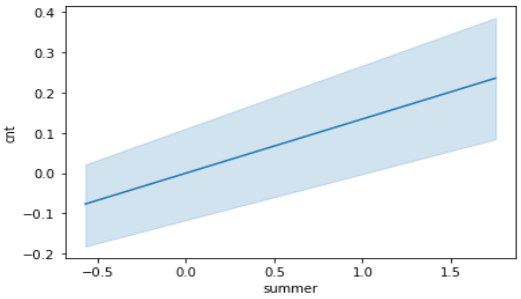
1. As Error points are randomly and almost uniformly distributed i.e. homoscedasticity has been found.

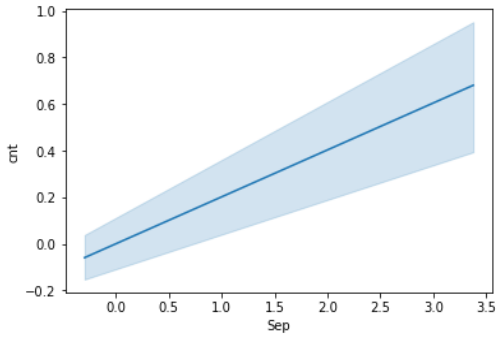


2. Linear relationship found within dependent and independent variables.

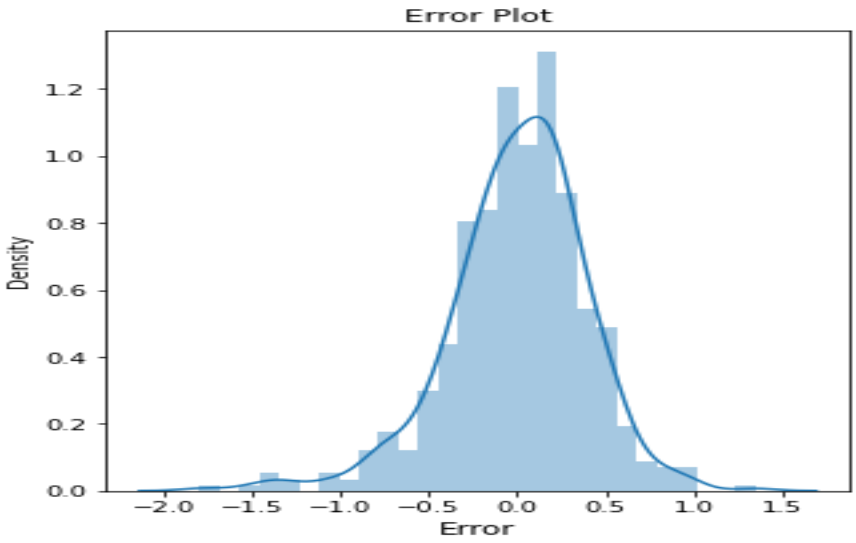








3. Error terms are independent of each other.
4. Error terms are normally distributed.



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

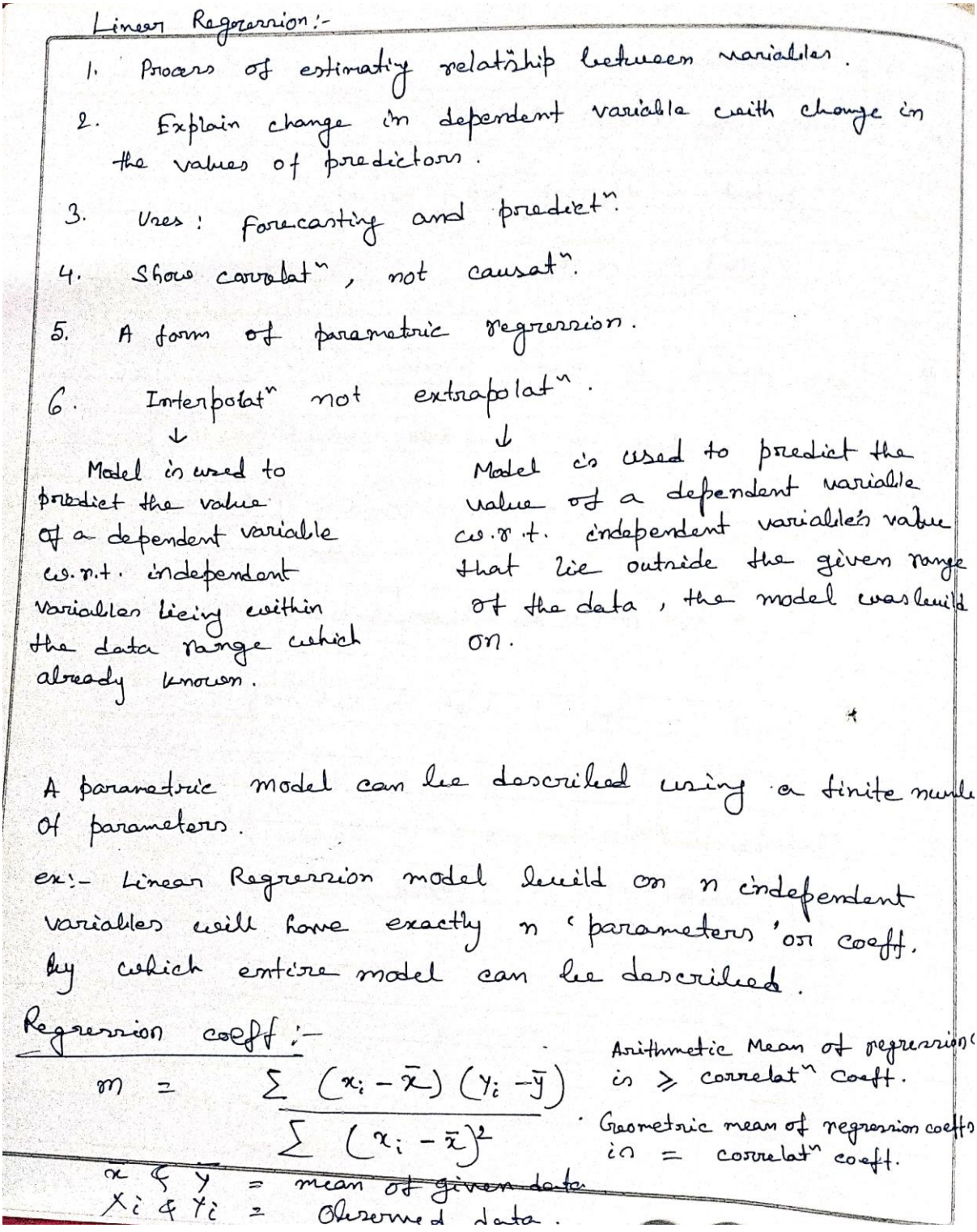
Ans:

Temperature, Year and Season are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:



Linear Regression:-

1. Procers of estimating relationship between variables.
2. Explain change in dependent variable with change in the values of predictors.
3. Uses: forecasting and predictⁿ.
4. Show correlatⁿ, not causatⁿ.
5. A form of parametric regression.
6. Interpolatⁿ not extrapolatⁿ.

↓ Model is used to predict the value of a dependent variable w.r.t. independent variables lieing within the data range which already known.

↓ Model is used to predict the value of a dependent variable w.r.t. independent variables value that lie outside the given range of the data, the model was build on.

A parametric model can be described using a finite numb of parameters.

ex:- Linear Regression model build on n independent variables will have exactly n 'parameters' or coeff. By which entire model can be described.

Regression coeff:-

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Arithmetic Mean of regression is ≥ correlatⁿ coeff.

Geometric mean of regression coeffs is = correlatⁿ coeff.

$\bar{x}$ & $\bar{y}$ = mean of given data

x_i & y_i = Observed data.

Assumptions of Linear Regression :-

1. Linear Relationship between X & Y.
2. Error terms are normally distributed (not X, Y)
3. " " " independent of each other.
4. " " have constant variance. (homoscedasticity)

Linear Regression :- → Supervised method.

→ Linear relationship between a dependent & one or more independent variables.

→ Used to find best fit line among the data.

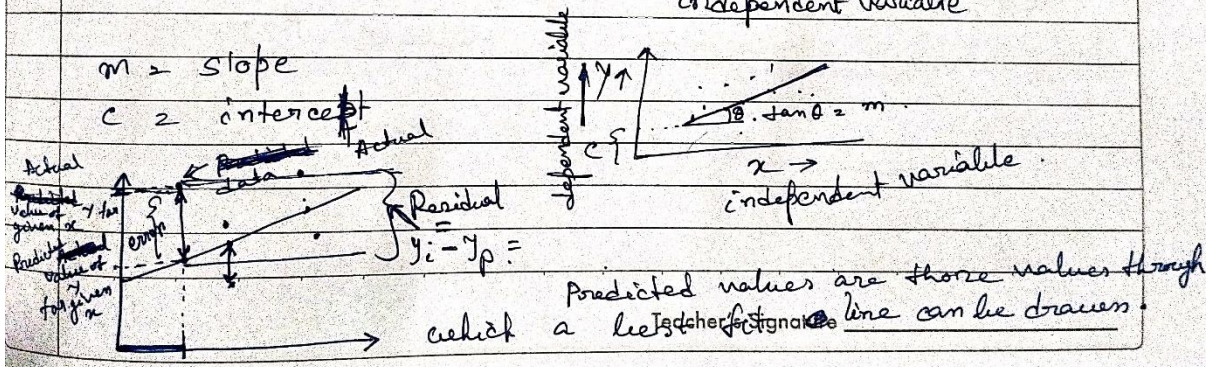
→ " " " the coefficients of a line (m & c).

$y = mx + c$. from plot of actual data to find the values of predicted data. using ~~$y_p = mx$~~

predicted value (y_p) = Actual value of independent variable $\times m + c$.

m = slope

c = intercept



which a best fit straight line can be drawn.

Predicted values are those values through

Linear Regression: → Use Least square method.

$r_i = y_i - y_p$

Residual Sum of Squares (RSS) $\sum_{i=1}^{n} r_i^2$.

$y_p = mx_i + c$.

m & c derived from the plot of x_i and y_i.

y_{actual}.

$$RSS = (y_1 - c - mx_1)^2 + (y_2 - c - mx_2)^2 + \ldots (y_n - c - mx_n)^2$$

$$\boxed{RSS = \sum_{i=1}^{n} (y_i - c - mx_i)^2}$$ = SSE Sum of Square Error.

$$= \sum_{i=1}^{n} (y_{i_{actual}} - y_{predicted})^2$$

2. Explain the Anscombe's quartet in detail.

Date

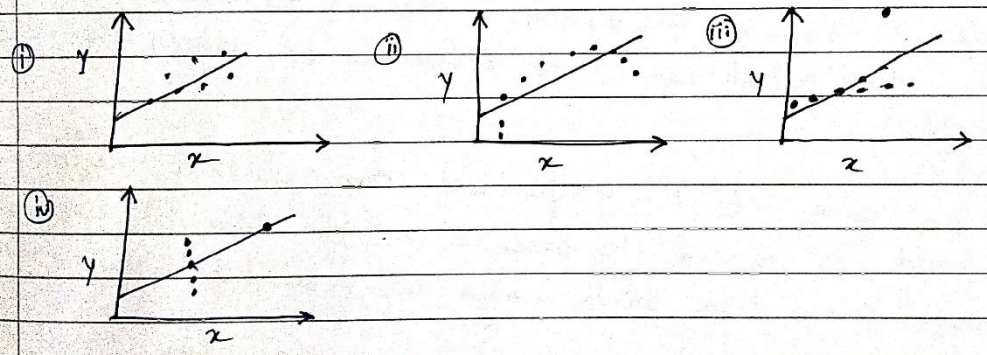
Expt. No. Page No.

Anscombe's quartet :- This consists of four data sets which have nearly identical simple descriptive statistics but very different distribution appeared when graphically plotted.

Therefore, Anscombe's quartet implies the importance of graphical analysis of data before the statistical analysis and model building applying any algorithm.

In this quartet, constructed by statistician Francis Anscombe, 11 (x, y) points are present for each dataset. Each dataset has almost identical statistic summary.

When the points are plotted graphical views are as follows:-



(i) The 1st scatter plot represents the linear relationship where y can be modelled as gaussian with mean linearly dependent on x.

(ii) The 2nd plot implies that the relationship between x & y is not linear so pearson correlation coeff. is not relevant. Something more efficient than general regression and its corresponding coeff. of correlation would be more appropriate.

Teacher's Signature

(iii) In the 3rd case though the relat^n between x & y is linear still a robust regression is needed to model it as there is outlier which has great influence to lower the correlat^n coeff.

(iv) In case of 4th graph we can see that the one high-leverage point is enough to produce a high correlat^n coeff. even though other data points show no relat^{n}ship between them.

Thus the quartet shows how can the basic statistic properties make us fool based on the effect of outliers and other inadequencies in data. So graphical analysis is must before applying any algorithm and building model.

3. What is Pearson's R?

Pearson's coeff :-

Pearson coeff. is a correlation coeff. used in statistic to determine the value of the linear relationship between two variables. It is denoted by 'r'.
This rely on the hypothesis that the given data can be represented by a straight line. This coeff. is the rescaled version of regression coeff.

Mathematical Expression :- $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$

x_i = values of x variable in a sample
y_i = " " y " " " "
$\bar{x}$ = mean of the values of x variable.
$\bar{y}$ = " " " " " "

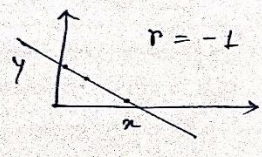
Teacher's Signature ——— .

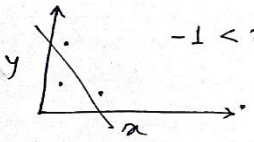
Value of r varies in the range of -1 to $+1$.

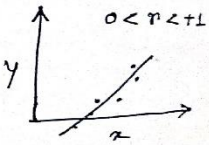
if $r = -1$ then it means all data points lie on a line and the value of y decreases as the value of x increases or vice-versa.

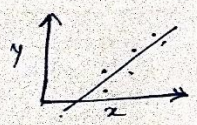
if $r = +1$ then it means all data points lie on a line and the value of y increases as the value of x increases.

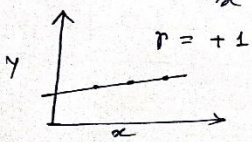
if $r = 0$ then there is no linear relationship between variables.

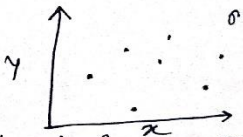












Then this coeff. determine the strength of linear relat" as well as direct". The sign is determined by the regression slope (m).

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

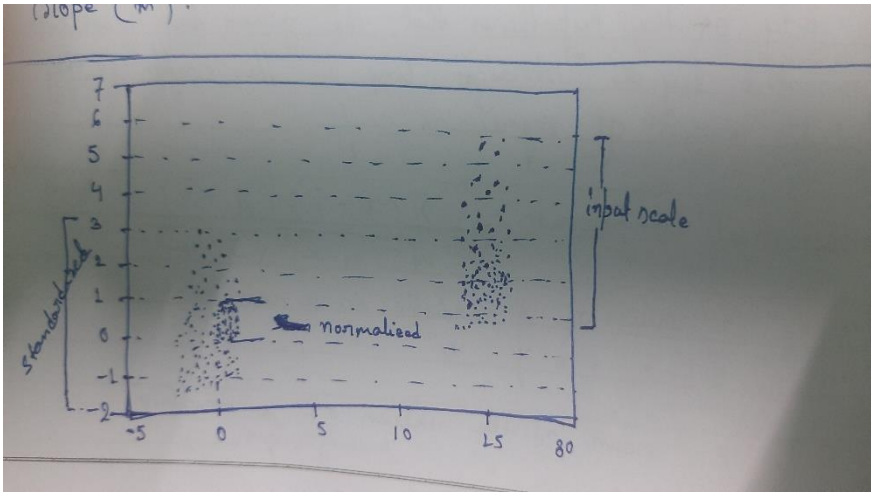
Ans:
Scaling is the process of changing the range of data of a given dataset.
There are two methods of scaling:

1. Normalization or Min-Max scaling
2. Standardization (mean-0, sigma-1)

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model, might be very large or very small as compared to the other coefficients. This can affect model's performance. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.
Due to rescaling optimization in the backend becomes faster also.

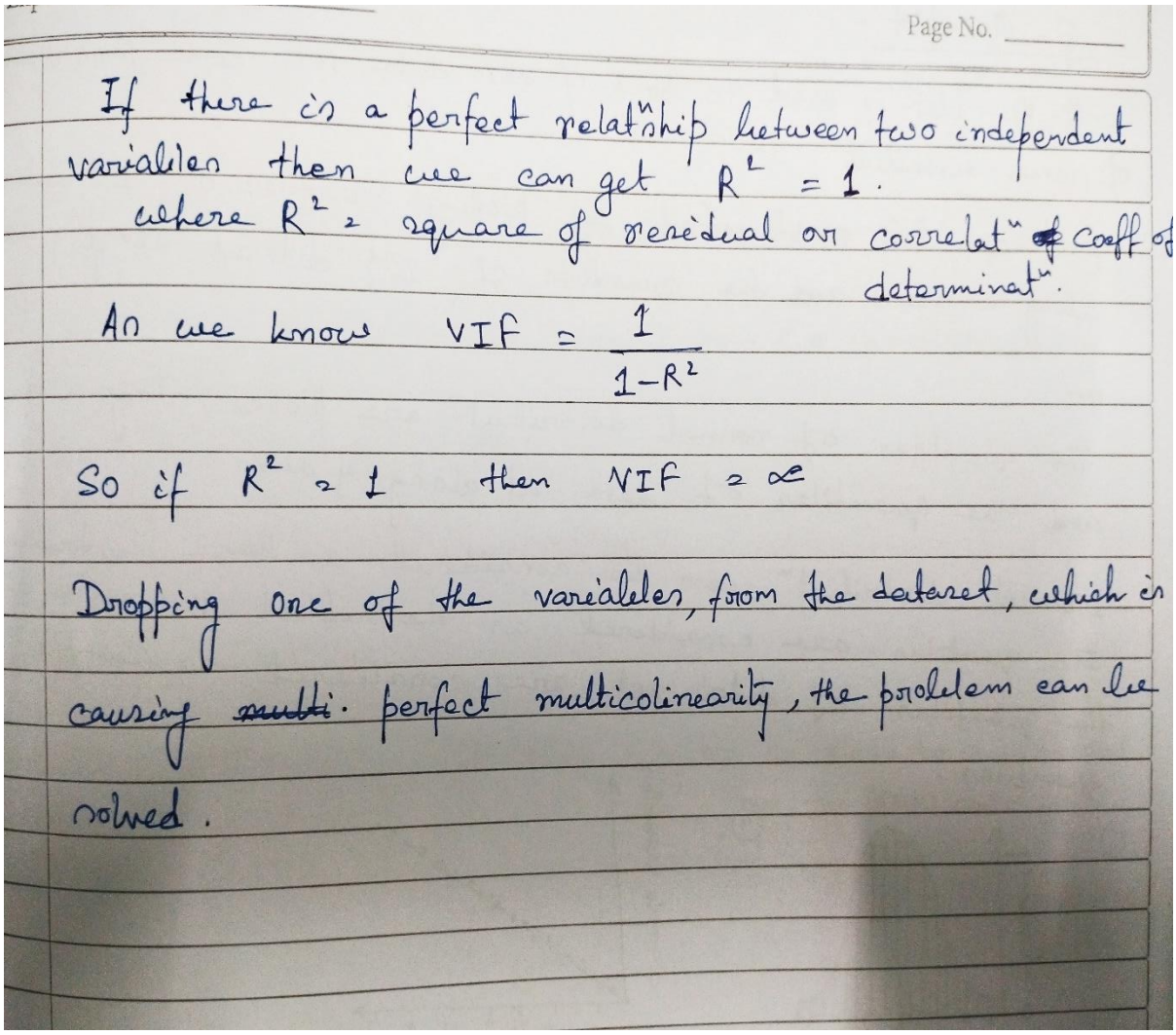
Standardization method doesn't compress data within a particular range as Normalization method does.

Standardization method is preferred over normalization in case of such data which have outliers



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

Ans:



If there is a perfect relationship between two independent variables then we can get $R^2 = 1$.

where R^2 = square of residual or correlation coeff of determination.

As we know $VIF = \frac{1}{1-R^2}$

So if $R^2 = 1$ then $VIF = \infty$

Dropping one of the variables, from the dataset, which is causing ~~multi~~. perfect multicolinearity, the problem can be solved.

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

Ans:

Q-Q plot :-

Q-Q plot is used to show if two data sets come from the populat of same distribut^n.

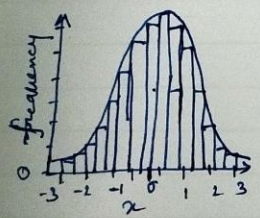
plot can be constructed by plotting quantiles of 1st dataset along x axis and the quantiles of 2nd dataset along y axis.

or.

The quantiles of normal distribut^n are plotted along x axis and the quantiles of data set along y axis.

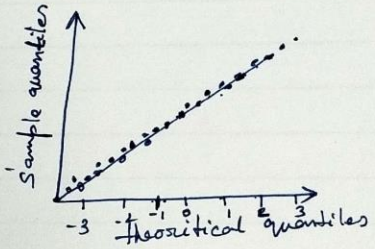
Normal distribut^n can be considered as base distribut^n. Its quantiles are considered as theoritical quantiles and the quantiles of data set are considered as sample quantiles.

(i)



Histogram.

(ii)



Normal Q-Q plot.

Example :- Figure (i) is showing the distribut^n of a set of some sample points. The distribut^n is almost normal distribut^n. So plotting the sample quantiles against quantiles of normal distribut^n (fig ii) get a straight line or can say the both quantiles match up along a straight line.

Therefore, it can be said that the sample comes from a normal distribut^n.

Some advantages are there in using Q-Q plot. Like:-

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

2. Many distributional aspects can be simultaneously tested. Like,
   a) Shifts in location.
   b) Shifts in scale.
   c) changes in symmetry.
   d) presence of outlier.

Interpretation:-

1. If all points of quantiles lie on or close to a straight line constructed at angle 45°, then the two data sets are coming from same distribution.

2. Otherwise different distribution is followed by the samples.

Use of importance in case of linear regression:-

a) In case of linear regression, the train & test dataset can be examined by Q-Q-plot to determine if those are coming from popilation of same distribution.

b) It helps to determine whether the residuals are normally distributed or not. If those are not aligne with the line with angle 45° then those are not following normal distribution.

Teacher's Signature _____