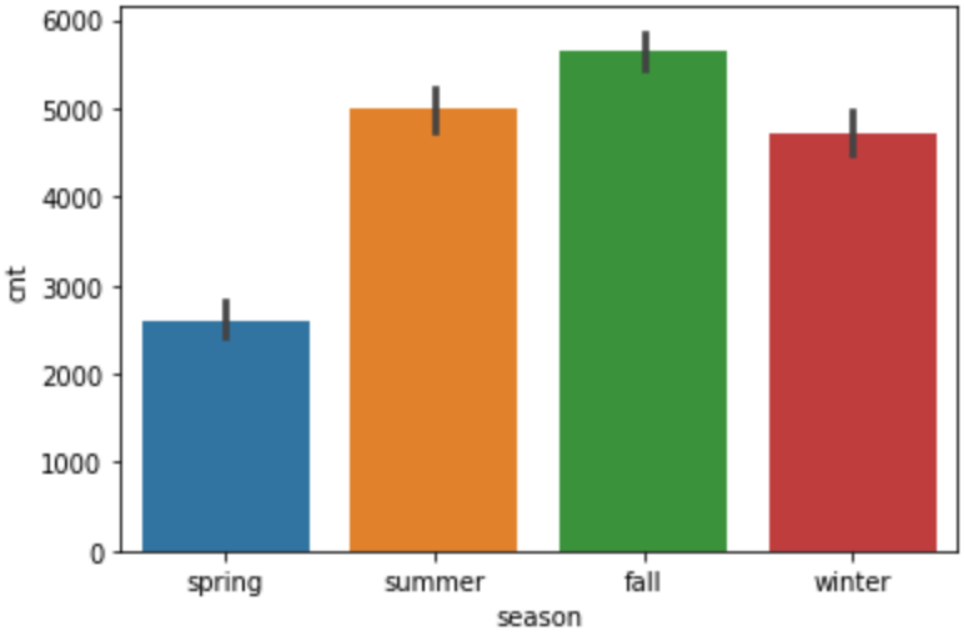
Assignment-based Subjective Questions

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

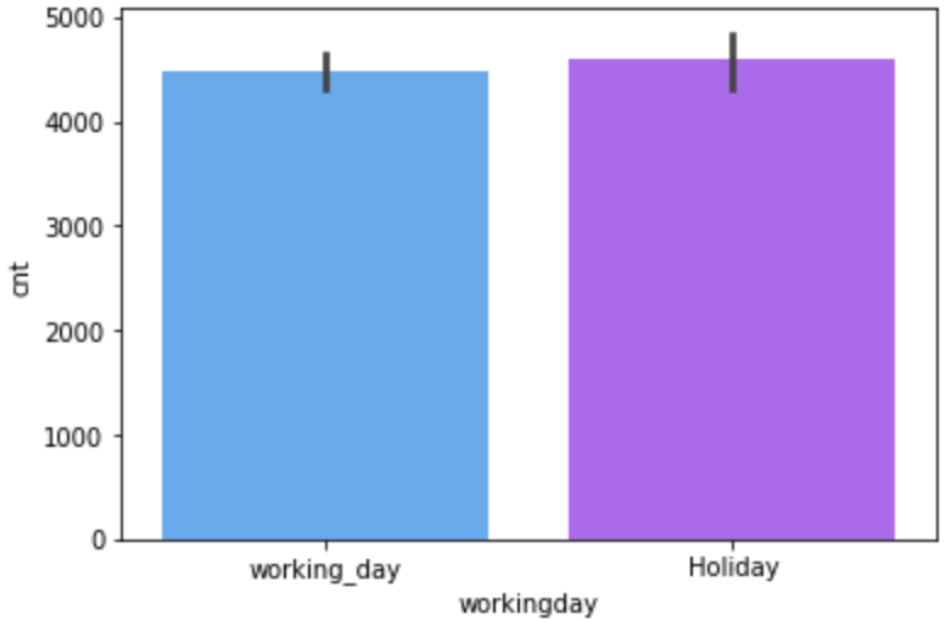
Ans: The categorical variables in the given dataset seems to have great effect on dependent variable “cnt”. They are given as numerical values in the given dataset which needs to be transitioned first to actual categorical values using maps and then we need to get dummy data to be used finally in our model.

The category features are ['season', 'holiday', 'mnth', 'weekday', 'workingday', 'weathersit']. After mapping to actual values, below graphs depict the relation between category feature and the dependent variable

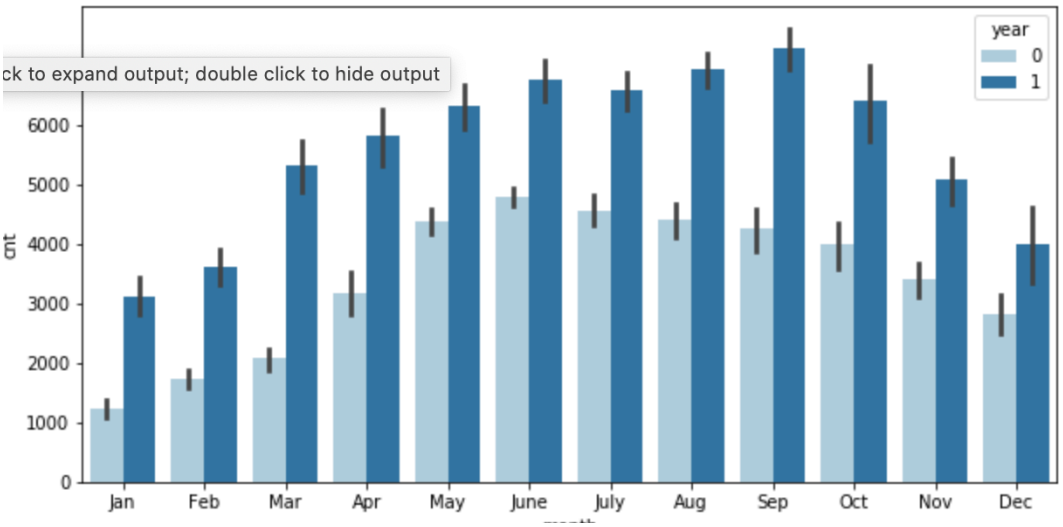
Season

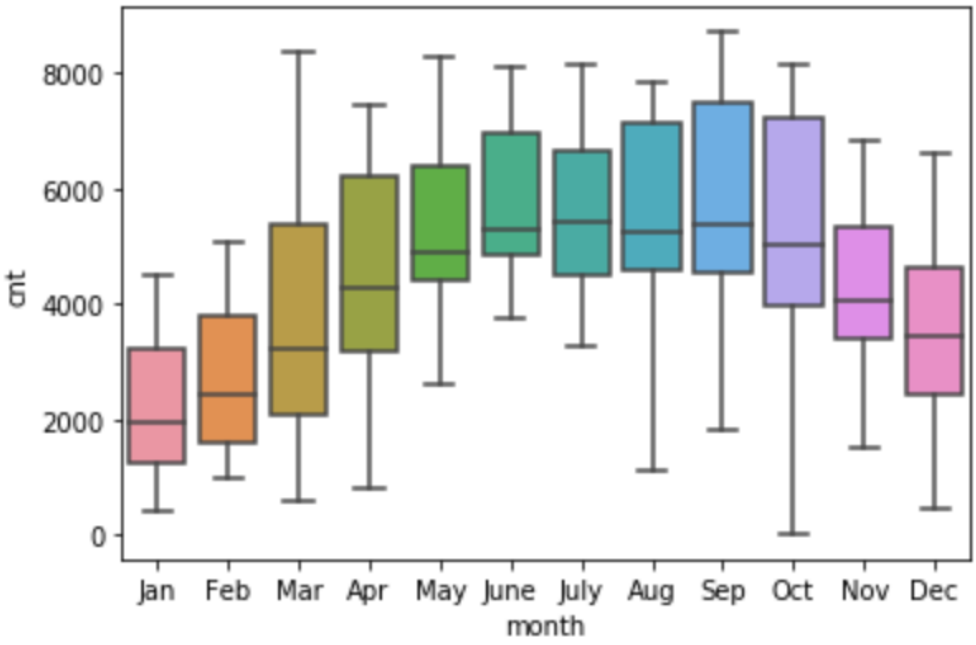


Holiday

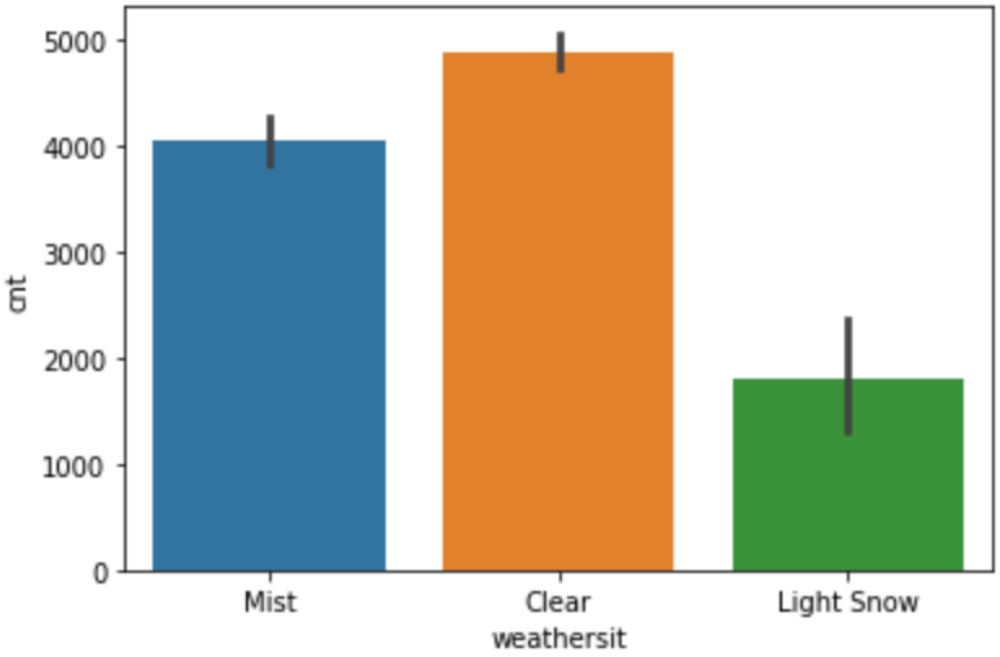

5000
4000
3000
2000
1000
0
cnt
working_day
Holiday
workingday

Month


ck to expand output; double click to hide output
year
0
1
6000
5000
4000
3000
2000
1000
0
cnt
Jan
Feb
Mar
Apr
May
June
July
Aug
Sep
Oct
Nov
Dec



Weathersit



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

Ans: Not dropping the first column will reflect dummy variables to be highly correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. Also possibility is getting the redundant feature.

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

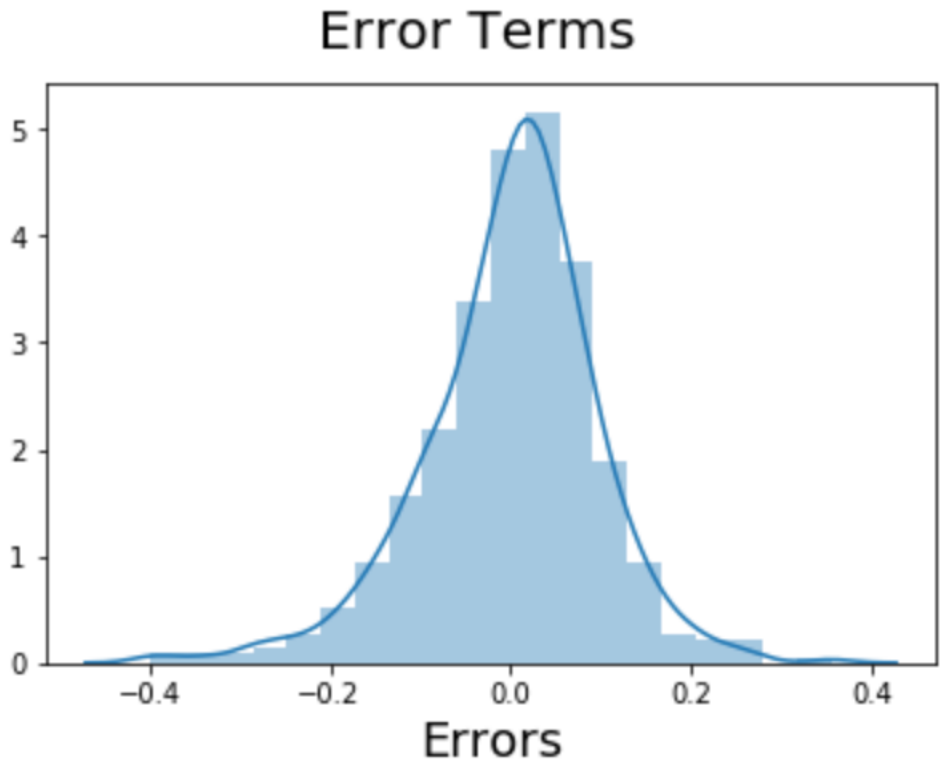
Ans: temp seems to have highest correlation with the target variable.

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

Ans: The assumption of Linear Regression will include:

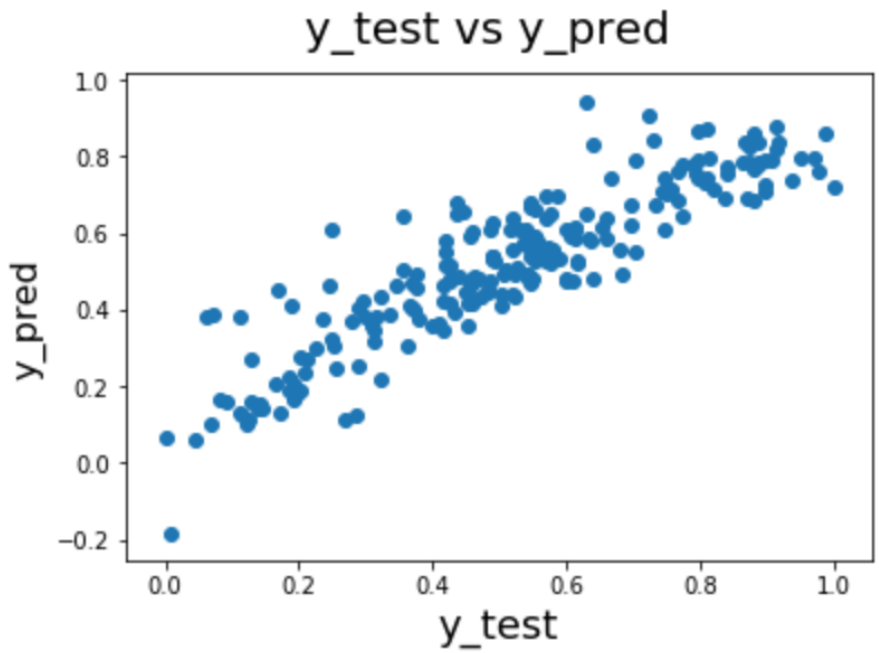
- The error terms are normally distributed.
- The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.
- The predicted values have linear relationship with the actual values

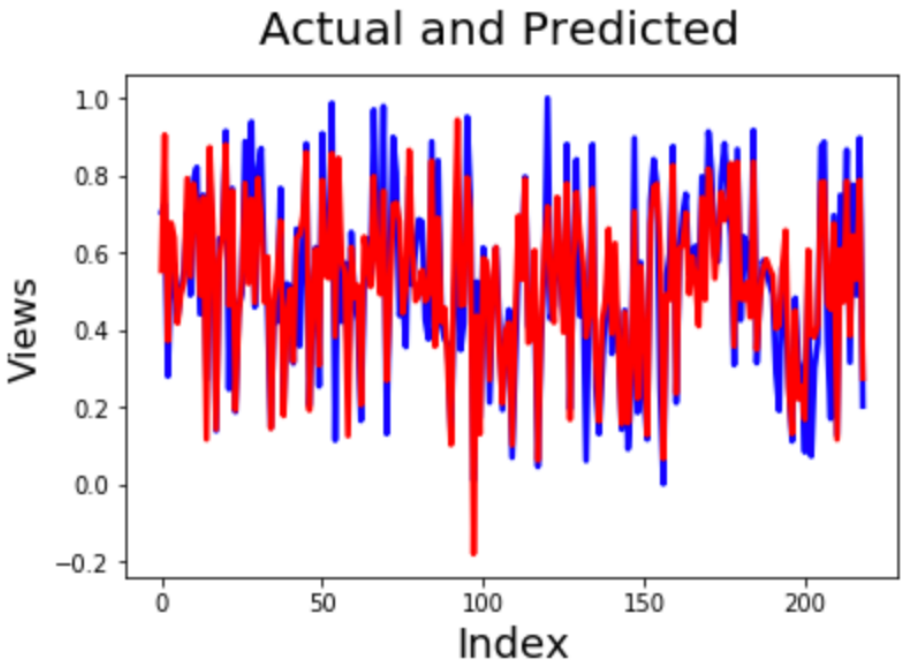
For the point 1 of error terms being normally distributed, plotted the histogram to check if it comes to be normally distributed. Below is the graph



For point 2, training/test data being no overfit/underfit, I checked for Rsquared, Ajusted Rsquared, p values and VIF values of variables.

For point 3, created a scatter plot between y test and y predicted to check if shows in a linear manner





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

Ans: Temperature, humidity and year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

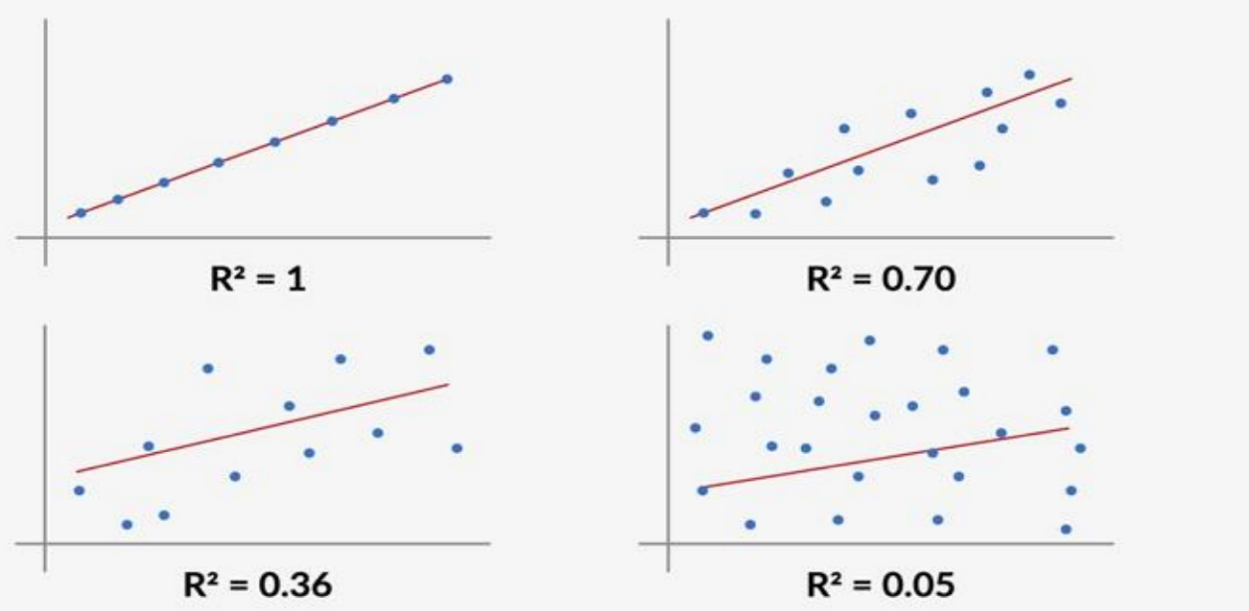
Ans: Regression is a method of modelling a target value based on independent predictors. This method is used for finding out cause and effect relationship between variables.
Linear regression is a linear model i.e. a model that assumes a linear relationship between the input variables(x) and the single output variable(y).
When there is a single input variable (x), the method is referred to as simple linear regression.
When there are multiple input variables, its referred as multiple linear regression.

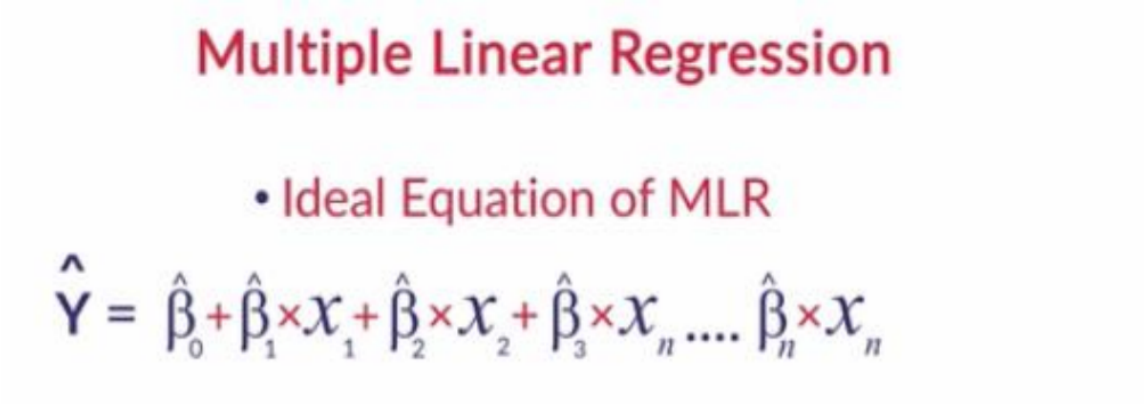
The strength of the linear regression model can be assessed using:

1. R squared

Mathematically, it is represented as: R^2 = 1 - (RSS / TSS). R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
RSS-> Residual Sum of Squares, TSS-> Total sum of Squares. TSS gives us the deviation of all the points from the mean line

Physical significance of R squared

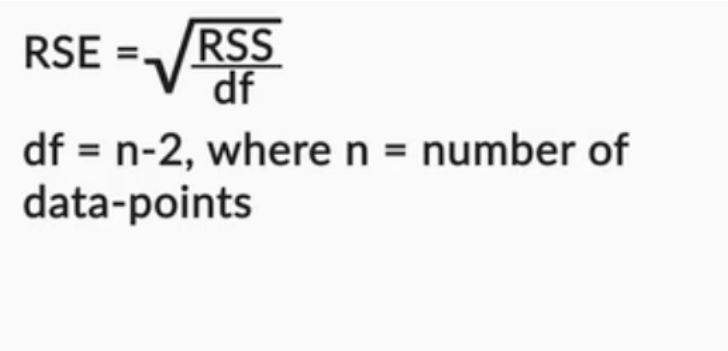




The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimizing the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:

- Differentiation
- Gradient descent

2. RSE (Residual Squared Error):

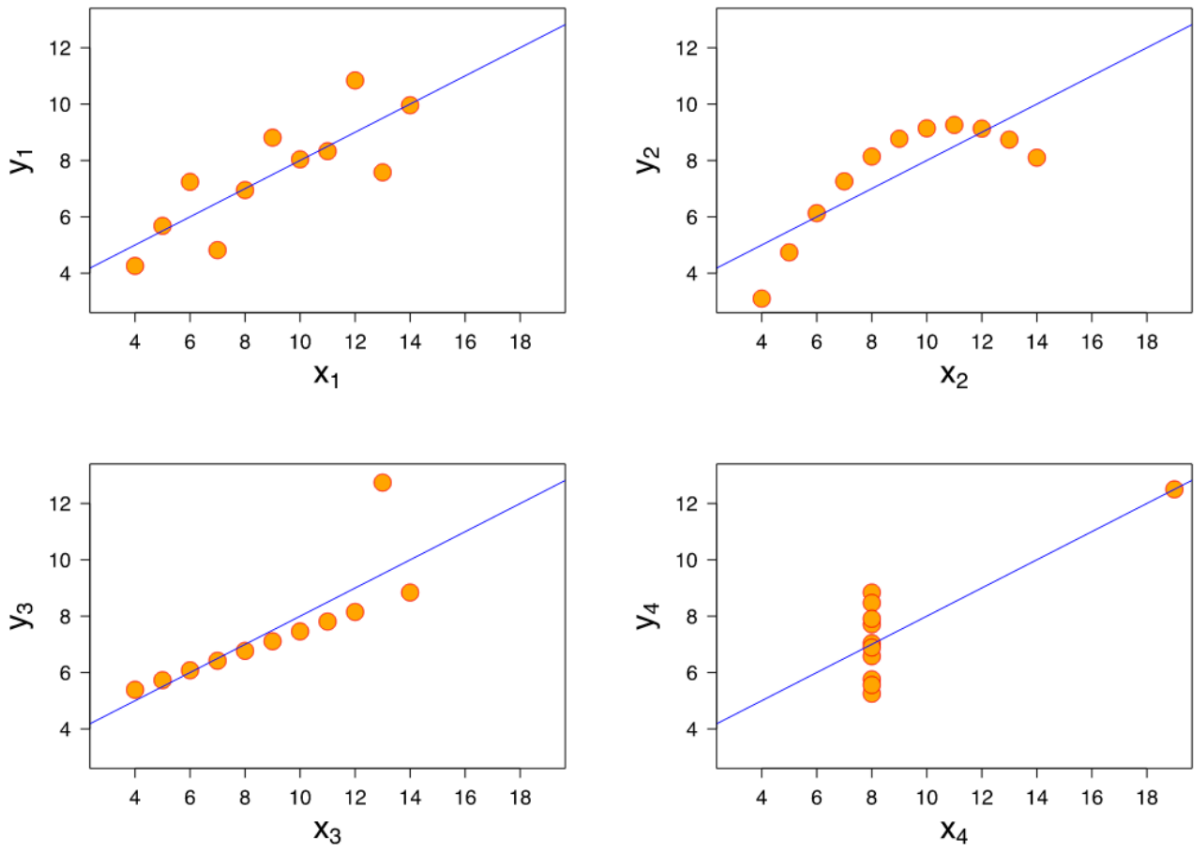


RSE is an absolute quantity measure while R squared is a relative measure of linear regression model metric and hence latter is more preferred approach.

2. Explain the Anscombe's quartet in detail.

Ans: It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Example:

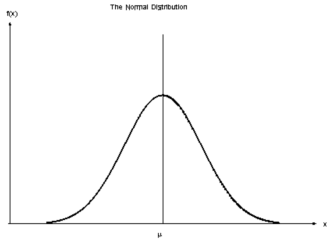


The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R?

Ans: Pearson's R also referred as Pearson correlation coefficient (PCC) or the bivariate correlation, is a statistic that measures linear correlation between X and Y. It has a value between +1 and -1. A value of +1 is total positive correlation, 0 is no linear correlation and -1 is total negative linear correlation.
Assumption while using Pearson r:

1. Variables should be normally distributed.



2. There should be no significant outliers.

3. Each variable should be continuous.
4. The 2 variables should have linear relationship.
5. Homoscedascity: meaning equal variances, meaning, the point should be on best fit line in scatter plot.

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

Ans: Scaling is a method used to normalize the range of independent variables or features of data, also known as normalization. It is a technique to standardize the independent variables present in the data in a fixed range.
Scaling features helps ease of interpretation. It just affects the coefficients and none of other parameters like p- value, r-squared.
Standardization brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax or Normalized scaling, on the other hand, brings all the data in the range of 0-1.
The formulae used in the background for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

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

Ans: VIF is variance inflation factor. VIF calculates how well one independent variable is Explained by all other independent variables combined. VIF will be infinite when R squared value is equal to 1 meaning perfect fit.

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

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform

distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

It helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.