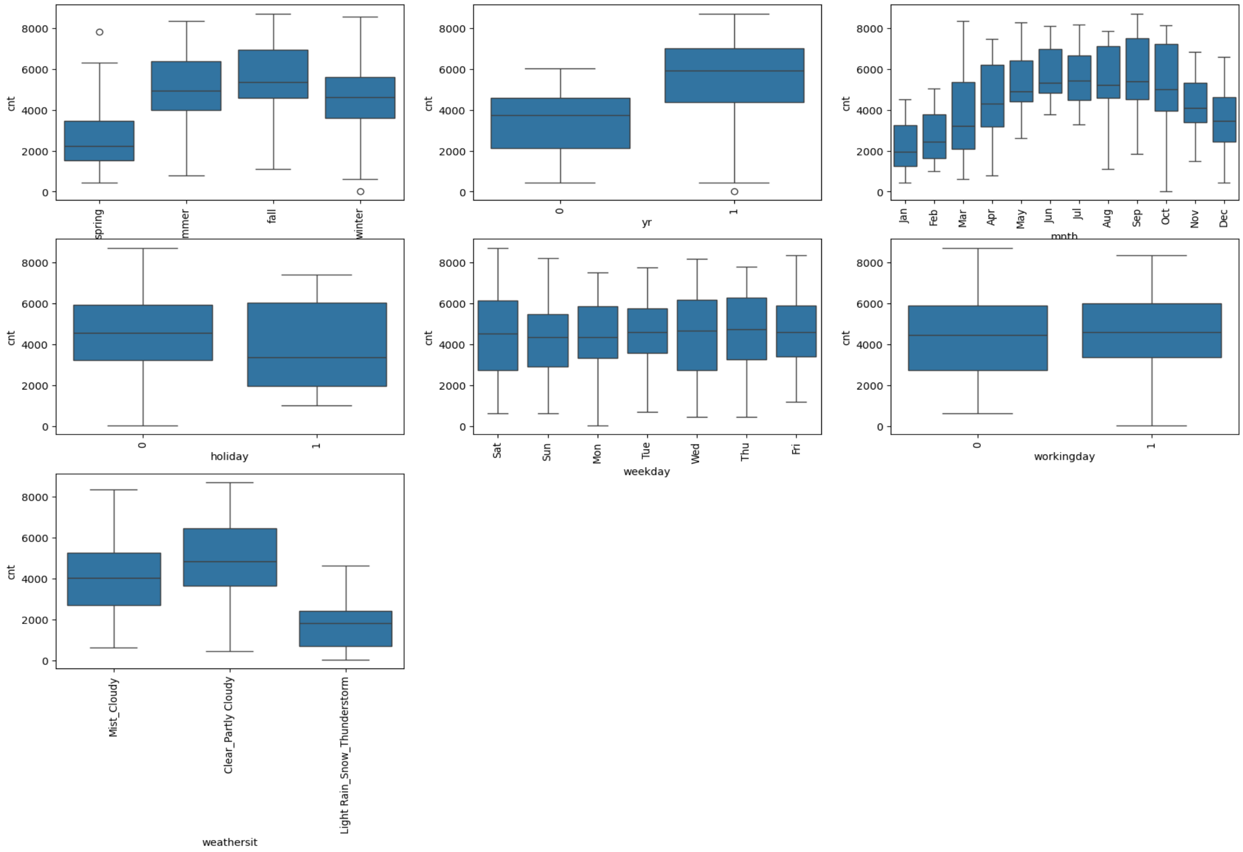
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer :

- Spring season is less preferred as compared to other seasons. Fall sees highest demand.
- Bike renting in 2019 is higher than that of 2018.
- Most of the demands seems to be during mid of the year (May – Oct). January observe lowest demand.
- Clearly, bike renting is more during non-holiday days.
- Demand of Bikes for rent remains almost constant throughout weekdays.
- There is no significant variance in bike renting in weekdays & working day.
- Most of the booking is done when weather is clear or partly cloudy, followed by Mist weather.
  - There are negligible booking when there is light snow/rain.
  - There is no booking at all when there is Heavy Rain/Thunderstorm.

Box Plot For Categorical Variables :



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

Answer :

Drop_first = True remove first level to get p-1 dummies out of p categorical variables. This is to reduce extra column which is created when dummy function is called.

Also, it is helps in reducing collinearity between dummy variables.

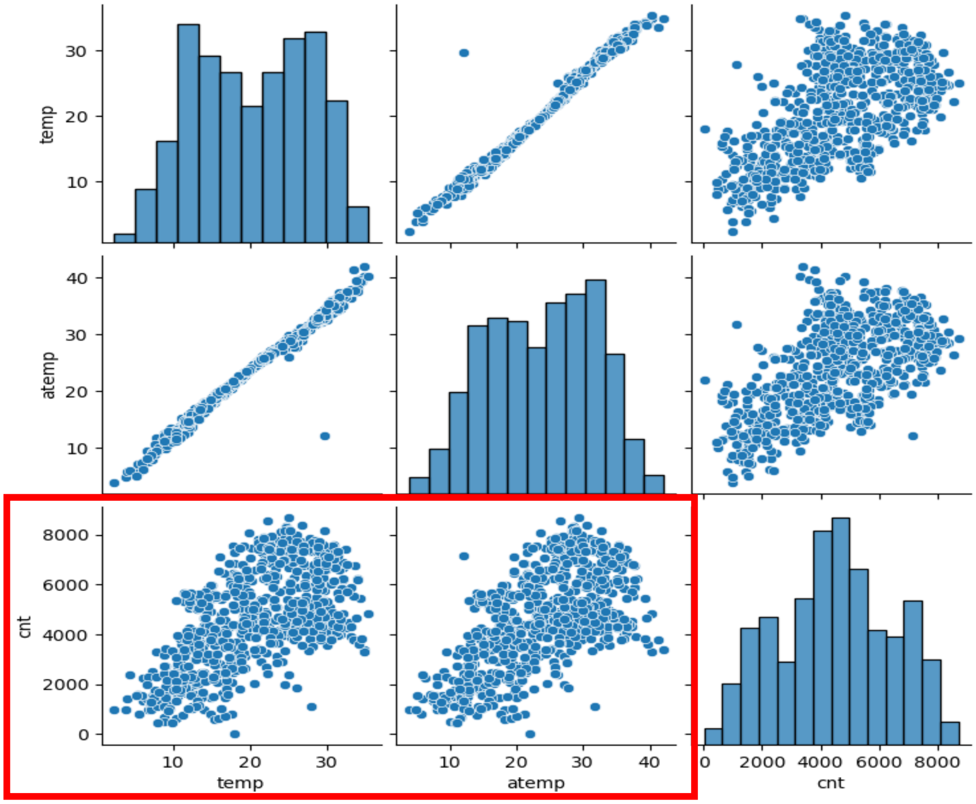
For example, if we had a column as "Gender" that has values as "Male" & "Female", get_dummies() will create 2 columns, Gender_Male & Gender_Female. If we delete any of the columns, eg Gender_Male, then its value can still be derived from other column (i.e Gender_Female).

Gender	**Gender_Male**	**Gender_Female**
Male	**1**	**0**
Female	**0**	**1**

Hence, if we have categorical variable with p-levels, then p-1 columns can represent the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

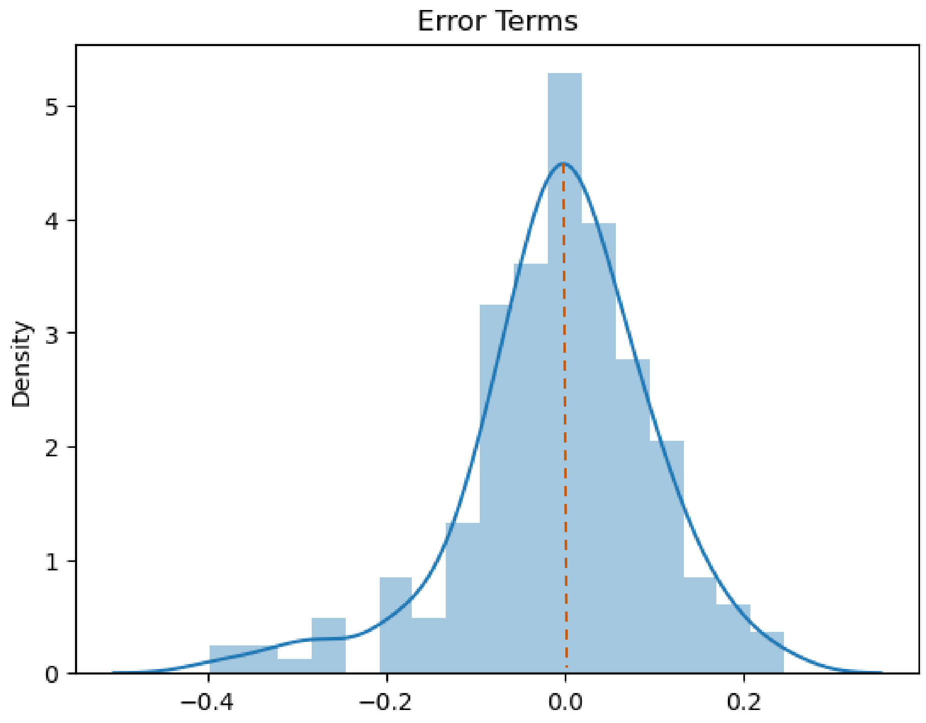
Answer : temp & atemp have highest correlation with target variable.



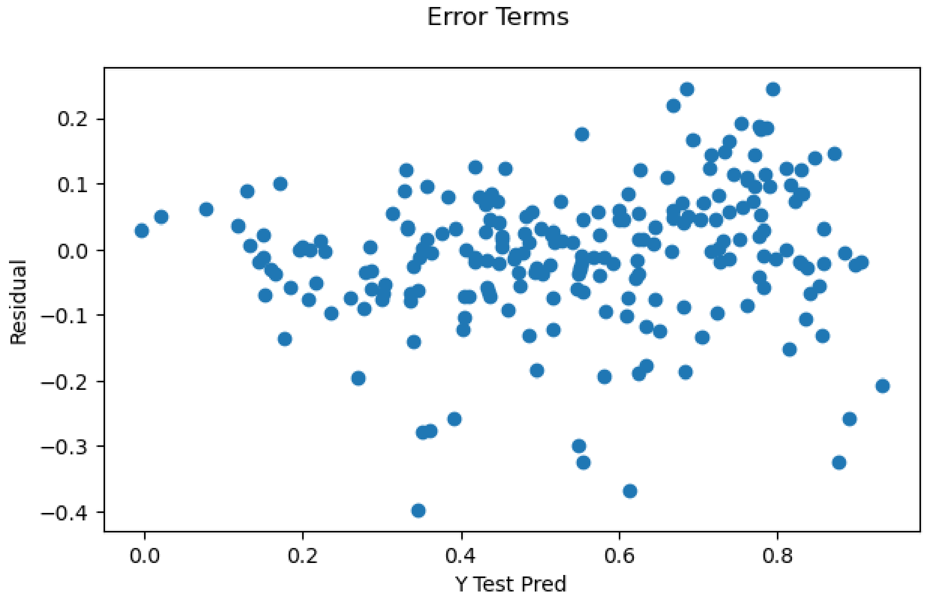
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer :

- Linearity in relation among independent & predicted variable.
- Random distribution of Error Term
- Homoscedasticity
- Mean of Error Term is very close to 0.
- Low VIF (low multi collinearity)



--> Error Term has mean at Zero, showing normal distribution



--> Error Terms are random in nature. No dependency on X or Y variable is observed

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer :
Based on final model, below 3 variables have signification contribution:

- Temperature (var name - temp, Weightage - 0.452)
- Year (var name - yr,Weightage - 251)
- Weather (var name - Light Snow, Light Rain + Thunderstorm + Scattered Clouds, Weightage - - 0.232)

<u>Other Coefficients:</u>

yr	**0.251**
workingday	0.037
temp	**0.452**
hum	-0.200
windspeed	-0.178
season_spring	-0.099
season_winter	0.096
mnth_Dec	-0.090
mnth_Feb	-0.051
mnth_Jan	-0.073
mnth_Jul	-0.048
mnth_Nov	-0.081
mnth_Sep	0.061
weekday_Sat	0.038
weathersit_Light Rain_Snow_Thunderstorm	**-0.232**
weathersit_Mist_Cloudy	-0.049

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer :
Linear Regression Algorithm is a machine learning algorithm, more specifically supervised learning, algorithm that learn from the given dataset and map the data points to the most optimized linear function. It is later used to predict the output for new dataset.

On the basis of task & nature of output, Machine learning models can be classified into following three types:
1. Regression: The output variable to be predicted is a continuous variable, e.g. scores of a student
2. Classification: The output variable to be predicted is a categorical variable, e.g. incoming emails as spam or ham
3. Clustering: No predefined notion of label allocated to groups/clusters formed, e.g. customer segmentation for generating discounts

Regression and Classification fall under supervised learning methods
Clustering falls under unsupervised learning methods – in which there is no predefined notion of labels.

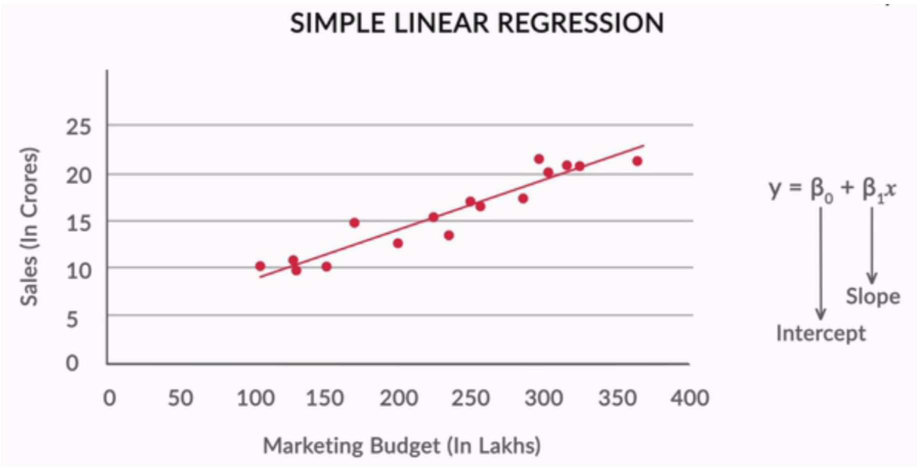
It is form of predictive modelling technique which tells us relation between target variable & independent variable.

There are 2 types of Linear Regression:

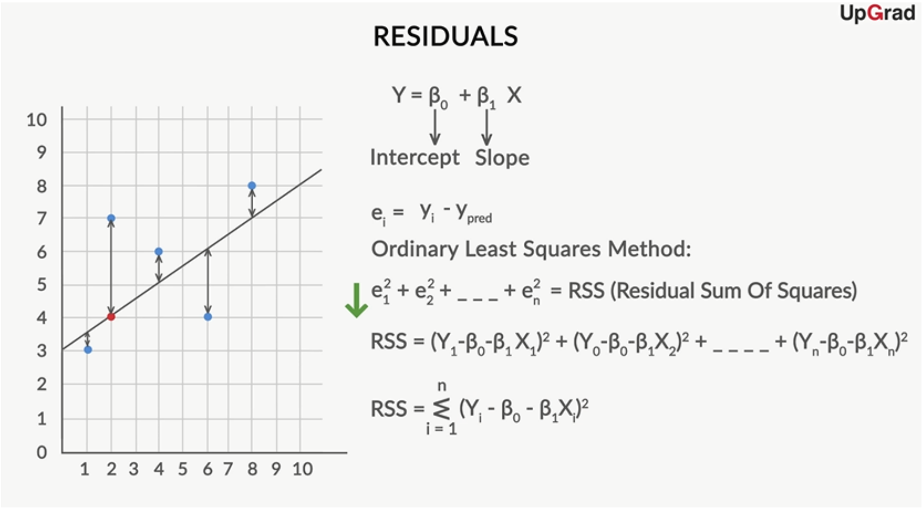
1) **Simple Linear Regression ::** Only one independent variable is present & the model has to find out its liner relationship with target variable.

   Formula :: $Y = \beta_0 + \beta_1.X$
   β_0 - Intercept
   β_1 - Slope

   Pictorial Representation :



Best Fit Line : The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



RESIDUALS

$Y = \beta_0 + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

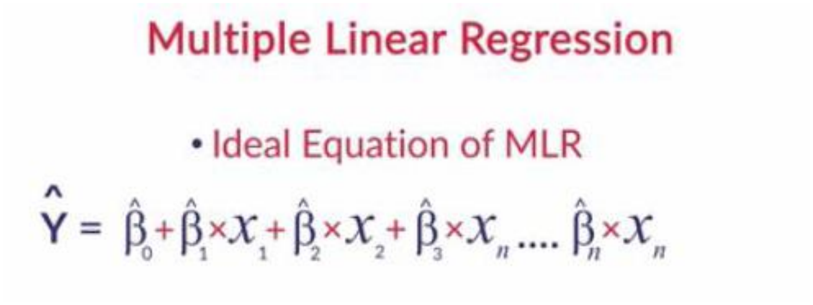
$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_0 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

The strength of the linear regression model can be assessed using 2 metrics:
1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

2) **Multiple Linear Regression ::** There are more than once independent variable for the model to find relationship.

Formula ::

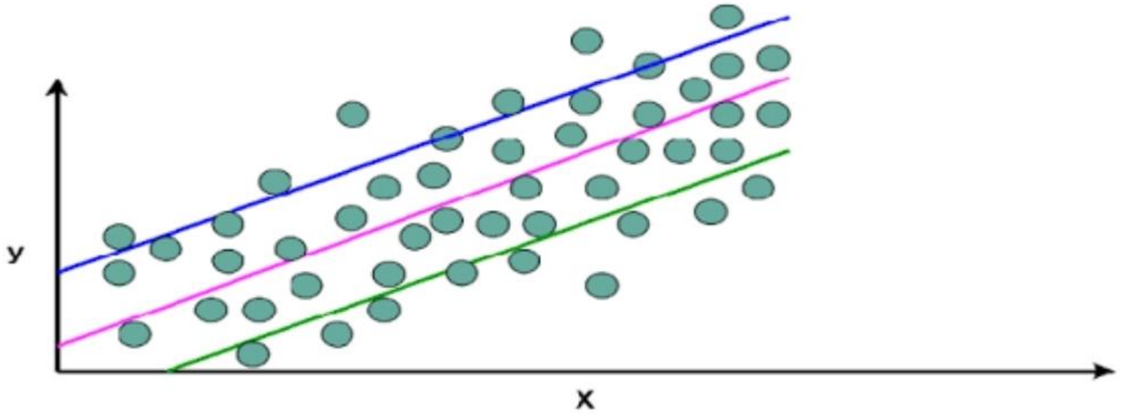


Multiple Linear Regression

- Ideal Equation of MLR

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times x_1 + \hat{\beta}_2 \times x_2 + \hat{\beta}_3 \times x_n .... \hat{\beta}_n \times x_n$$

Pictorial Representation :



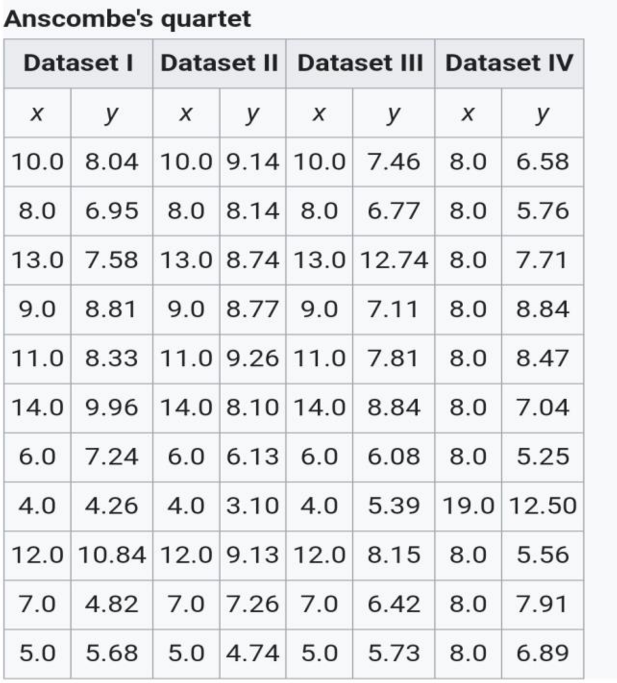
Multiple Linear Regression

---

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer :

Anscombe's quartet comprises of 4 datasets that have nearly identical simple descriptive statistics, yet have very different distribution & appear very different when graphed. Each dataset consists of eleven *(x,y)* pairs as follows :

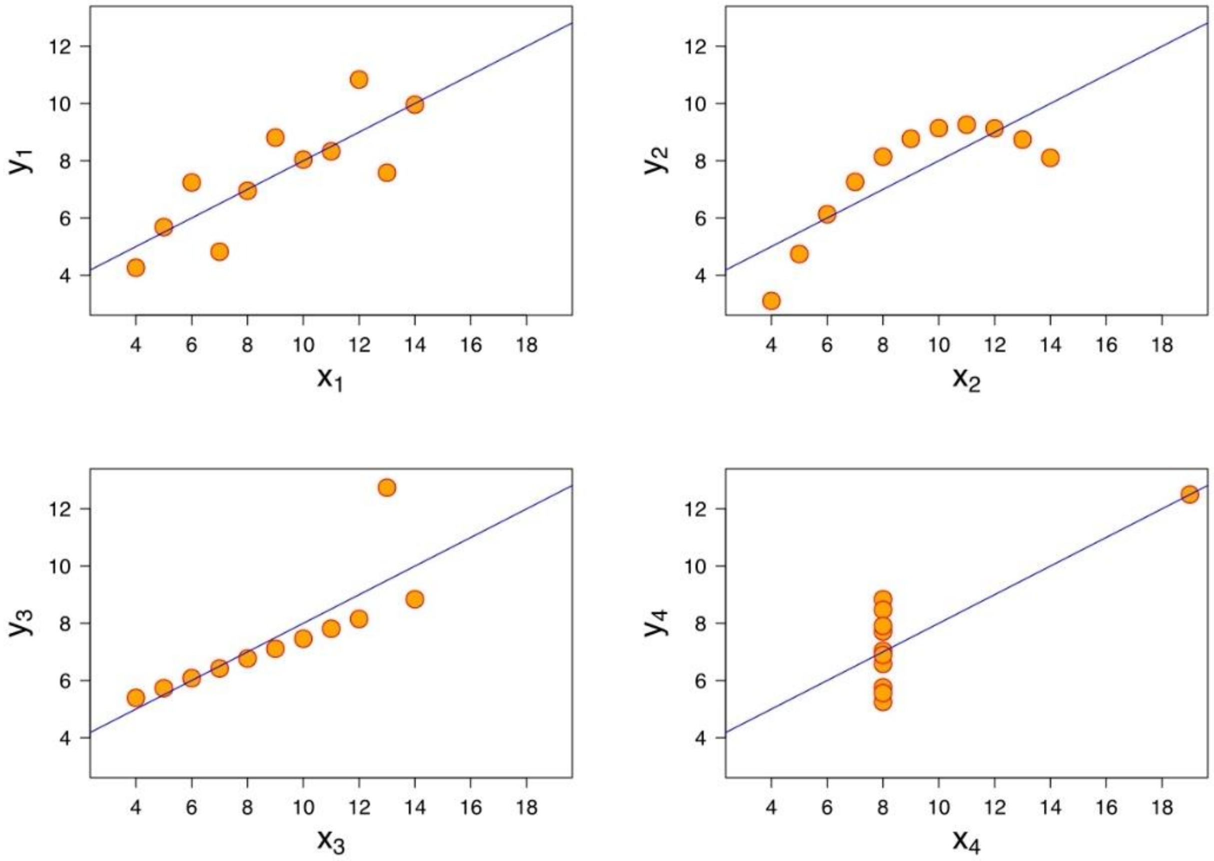


Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Here we see identical descriptive summary:
Mean of all X value is 9 in each data set.
Variance of all X is 11 in every data set.
Mean of all Y value is 7.50 in each data set.
Variance of all Y is 4.127 in every data set.
The correlation between x and y is 0.816 for each dataset
A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

As per above observation, these four datasets appear to be similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Dataset 1 consists of a set of points that appear to follow a rough linear relationship with some variance.
Dataset 2 fits a neat curve but doesn't follow a linear relationship.
Dataset 3 looks like a tight linear relationship between *x* and *y*, except for one large outlier.
Dataset 4 looks like *x* remains constant, except for one outlier as well.

While the descriptive statistics of Anscombe's quartet may appear uniform, the visualization reveals distinct patterns. It shows necessity of combining statistical analysis with graphical exploration for correct data interpretation.

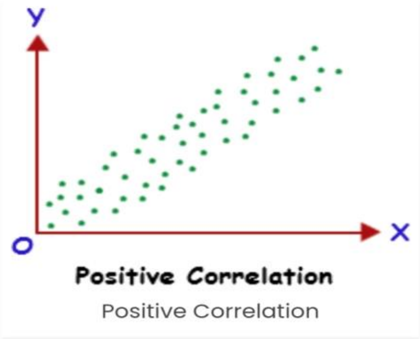
3. What is Pearson's R? (3 marks)

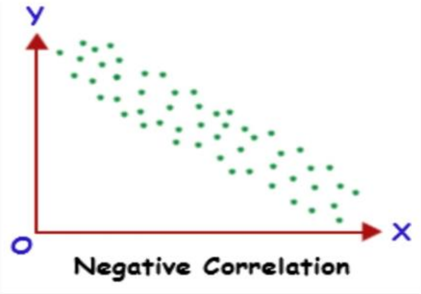
Answer :

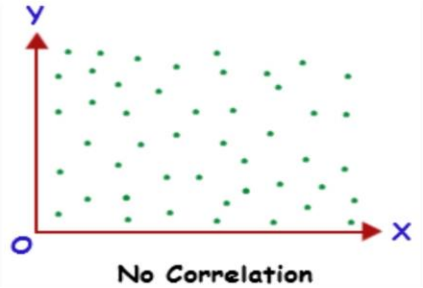
The correlation between the data set is done by Pearson's R. It is a measure of how well they are related. Its complete name is Pearson Product Moment Correlation (PPMC).

It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data?

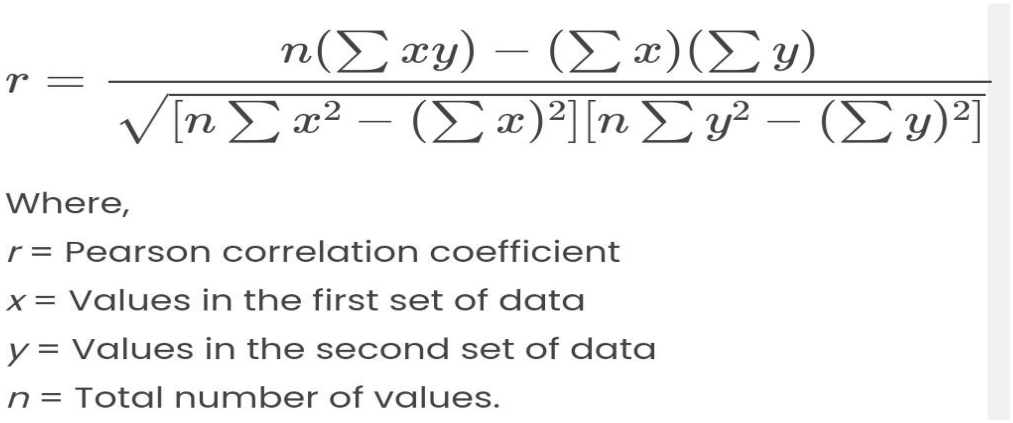
The value of Pearson’s R lies between -1 to +1.
0 means data is not related
-1 means data has negative correlation
+1 mean data has positive correlation







Formula:
The Pearson’s correlation coefficient is denoted by letter “r”. Formula is given by:



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

Where,

r = Pearson correlation coefficient

x = Values in the first set of data

y = Values in the second set of data

n = Total number of values.

Example:

Pearson correlation is used in many of real life situations.

For example, for an investor who wishes to diversify the portfolio, Pearson's Correlation Coefficient can be useful. Calculation from scatter plots of historical returns between pair of assets (equity-bond, equity-commodity, large cap-small cap etc) will produce Pearson's coefficients to assist the investor in assembling the portfolio based on risk & return parameters.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

In statistics, scale refer to the range or spread of values in the datasets. We use scaling to make data more manageable & comparable, especially when dealing with variables that have different units or vastly different ranges.

It is basically a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm by putting all numbers on similar scale.

Why is scaling performed:

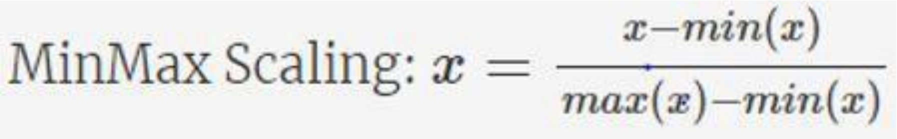
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence there is chance of incorrect modelling.

To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like F-statistic, p-values, R-squared, etc.

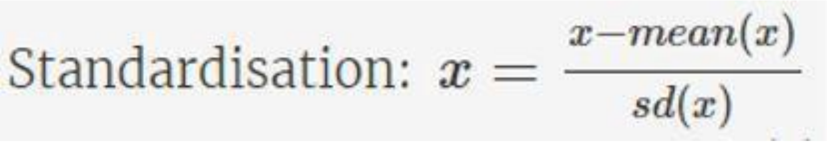
Normalized Scaling Vs Standardized Scaling:

a. Normalized Scaling brings all of the data in the range of 0 and sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

b. Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). sklearn.preprocessing.scale helps to implement standardization in python.



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)
Answer :
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

VIF (Variance Inflation Factor)
VIF = 1/ (1-R2)
Hence, when R Square reaches 1, VIF become infinity

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗 _1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

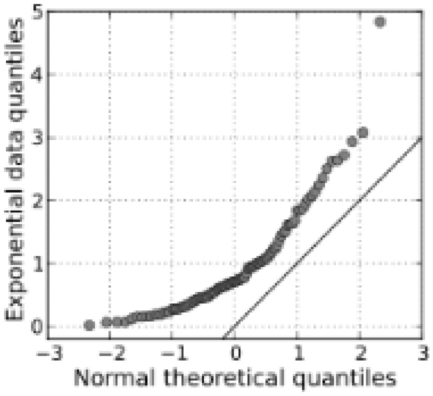
X_2=C+ α_1 X_1+α_3 X_3+⋯

〖VIF〗 _2=1/(1-R_2^2)

if all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

A Q-Q plot is a scatter plot 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.

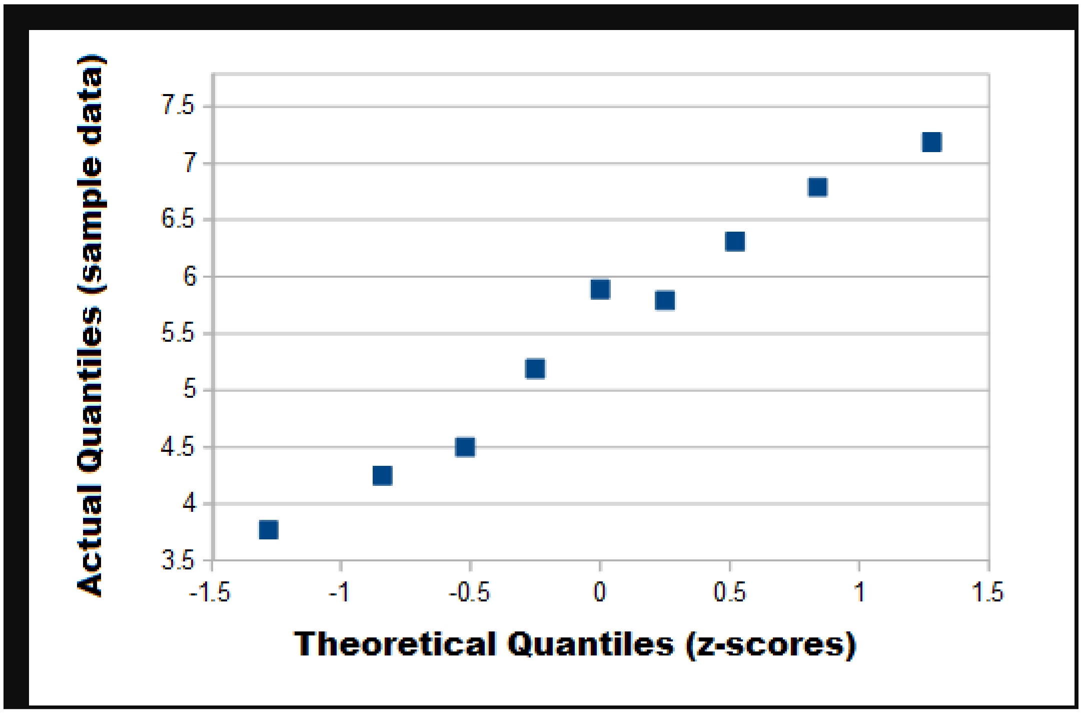


A Q Q plot showing the 45 degree reference

The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q Q plot;

if the two data sets come from a common distribution, the points will fall on that reference line.



Quantile-Quantile (Q-Q) plot 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.

Advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.