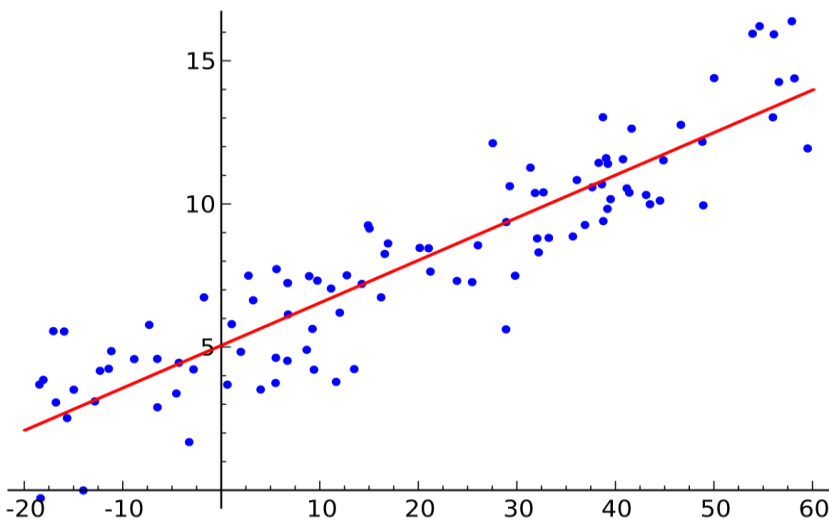


Explain the linear regression algorithm in detail.

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

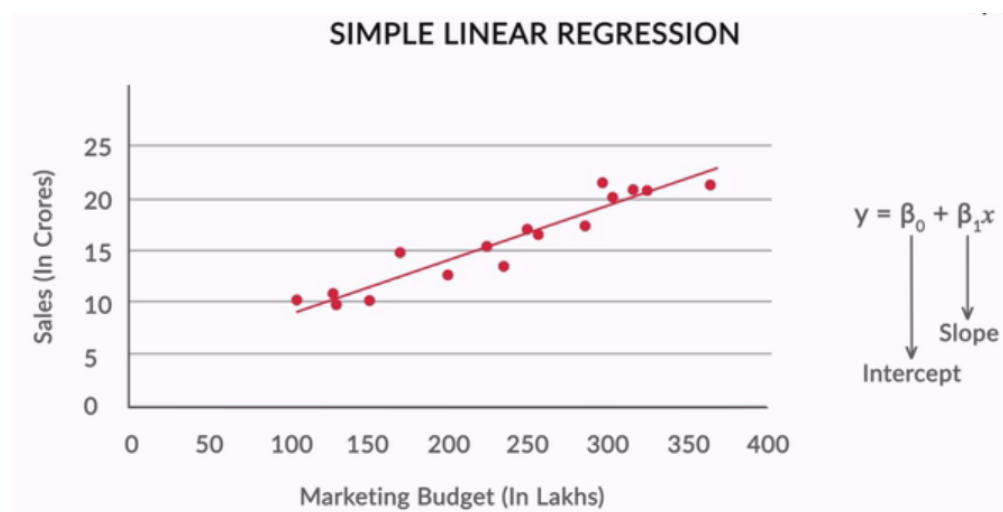
Simple: Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.



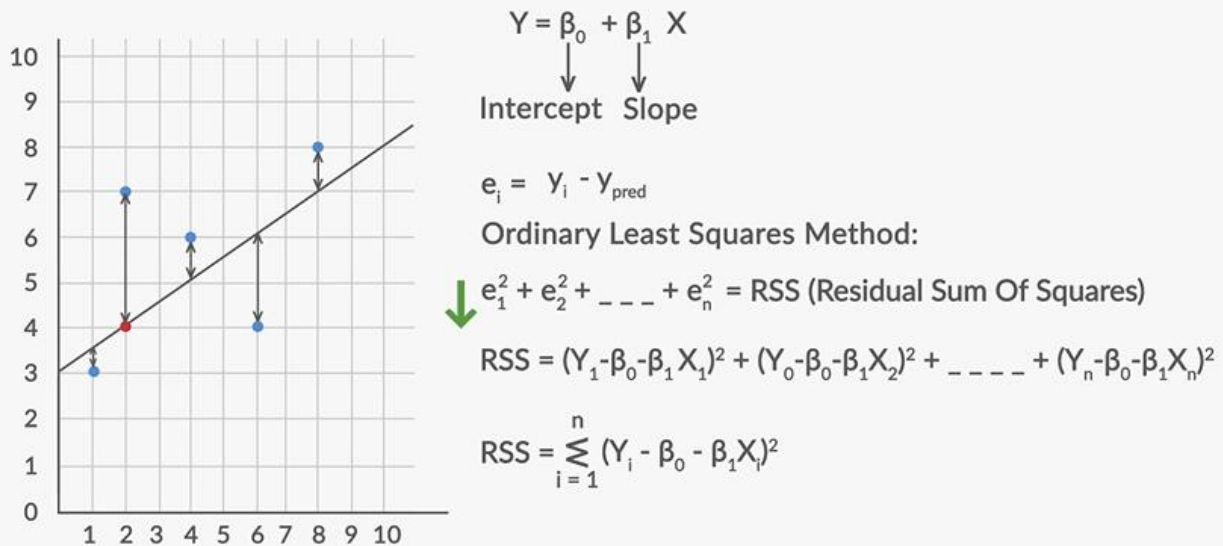
Regression Line: The standard equation of the regression line is given by the following expression

$$Y = \beta_0 + \beta_1 X$$

Where β_0 : - Intercept and β_1 : Slope



Best Fit Line: The best-fit line is found by minimising 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:



Strength of Linear Regression Model:

The strength of the linear regression model can be assessed using 2 metrics:

1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

1. R2 or Coefficient of Determination

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.

R2 Formula

•
$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where

RSS= Residual sum of square

TSS= Sum of errors of the data from mean

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

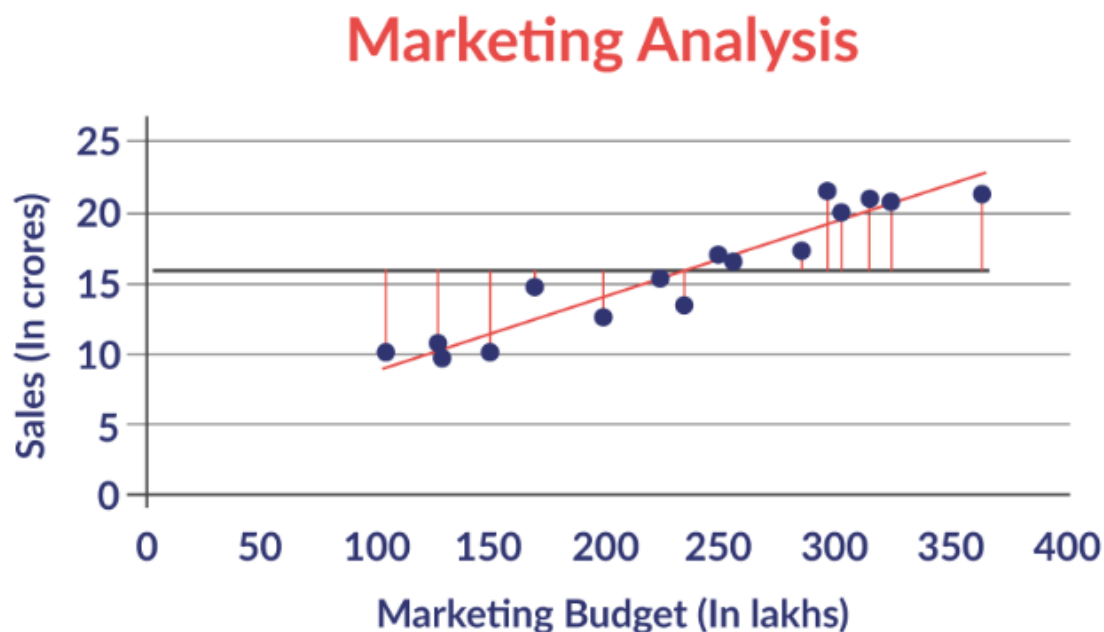
$$RSS = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Importance of RSS/TSS:

If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points as shown below. This is the worst possible approximation. TSS gives us the deviation of all the points from the mean line.



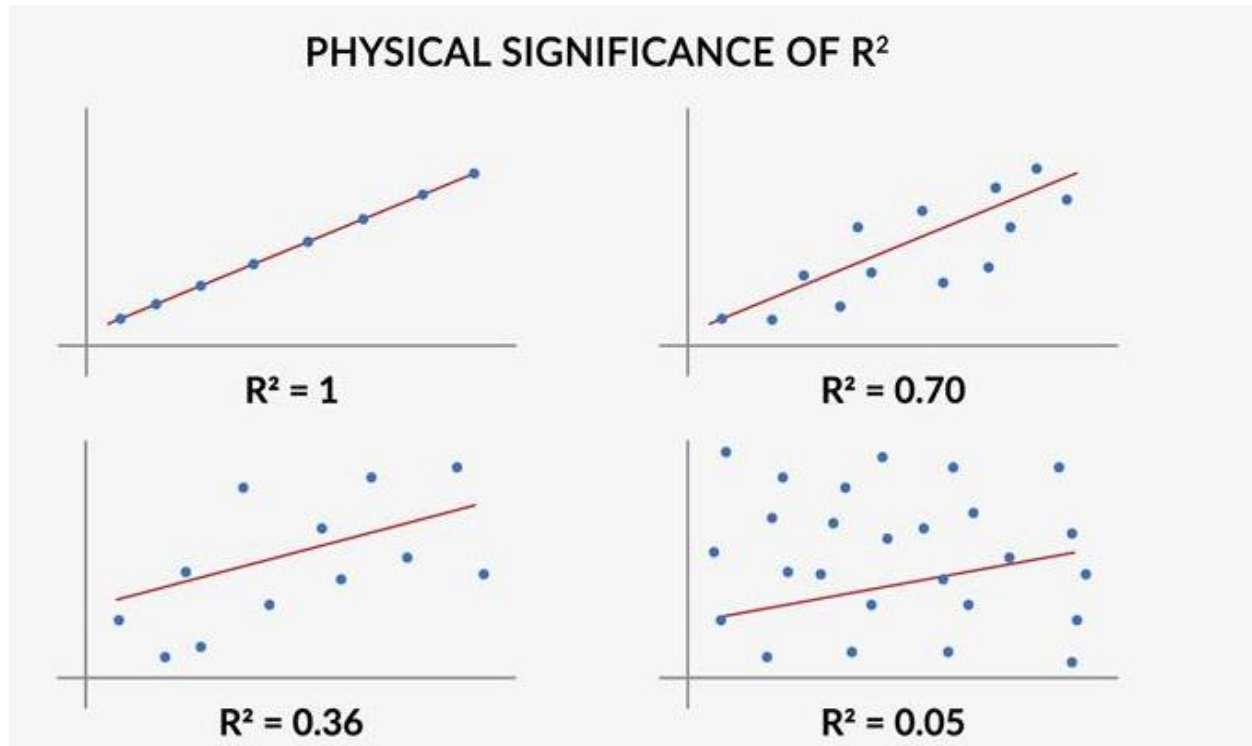
Trying to reinforce this understanding of R^2 visually, look at the 4 graphs of marketing data and compare the corresponding R^2 values.

In Graph 1: All the points lie on the line and the R^2 value is a perfect 1

In Graph 2: Some points deviate from the line and the error is represented by the lower R^2 value of 0.70

In Graph 3: The deviation further increases and the R^2 value further goes down to 0.36

In Graph 4: The deviation is further higher with a very low R^2 value of 0.05



Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables.

The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X .

In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, Radio marketing, and Newspaper marketing.

The simple linear regression model is built on a straight line which has the following formula:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Multiple linear regression also uses a linear model that can be formulated in a very similar way.

• Ideal Equation of MLR

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

The multiple linear regression explains the relationship between one continuous dependent variable (y) and two or more independent variables (x1, x2, x3... etc).

• Sales Prediction Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Tv marketing} + \hat{\beta}_2 \times \text{Internet marketing} + \hat{\beta}_3 \times \text{New paper marketing}$$

Dummy Variables: The categorical variables need to be converted to numeric form to be used in regression modelling. Thus, you create dummy variables.

For two level variables, you change the levels into 1 and 0 where 1 is one level and 0 is indicating another. In our case, for basement variable 1 indicates the presence of a basement and 0 indicates its absence. But when you directly convert the factor variable to a numeric type, the factor value of that variable is replaced by levels of variable, which is called coercion.

R-squared vs Adjusted R-squared: We then built a model containing all variables and saw the summary of the results. In multiple variable regression, adjusted R-squared is a better metric than R-squared to assess how good the model fits the data. R-squared always increases if additional variables are added into the model, even if they are not related to the dependent variable. R-squared thus is not a reliable metric for model accuracy. Adjusted R-squared, on the other hand, penalises R-squared for unnecessary addition of variables. So, if the variable added does not increase the accuracy adequately, adjusted R-squared decreases although R-squared might increase.

Multicollinearity: It may be that some variables could have some relation amongst themselves; in other word, the variables may be highly collinear to each other. A simple way to detect collinearity is to look at the correlation matrix of the independent variables as shown.

Correlation Matrix



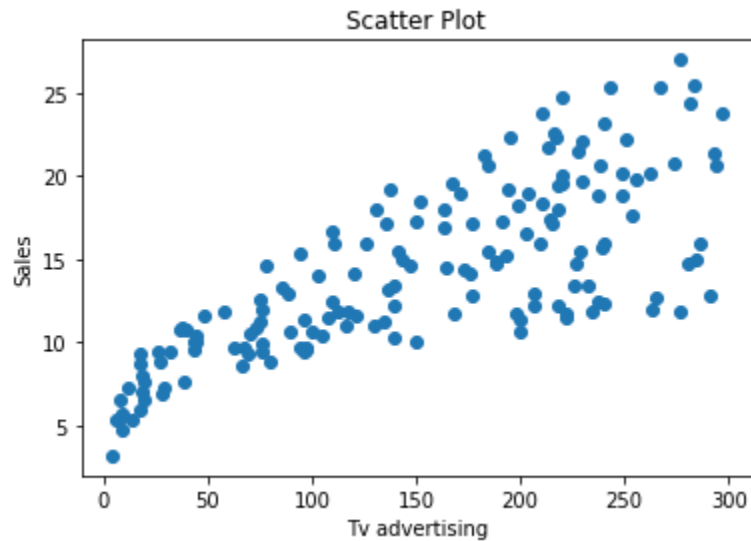
	Var1	Var2	Var3	Var4	Var5
Var1	1	-0.08071	0.098675	0.014625	0.061913
Var2	-0.08071	1	-0.10168	0.37678	0.103062
Var3	0.098675	-0.10168	1	0.049934	0.119171
Var4	0.014625	0.37678	0.049934	1	0.002249
Var5	0.061913	0.103062	0.119171	0.002249	1

A large value in this matrix would indicate a pair of highly correlated variables. Unfortunately, not all collinearity problems can be detected by the inspection of the correlation matrix. It is possible for collinearity to exist between three or more variables even if no pair of variables has a high correlation. This situation is called multicollinearity.

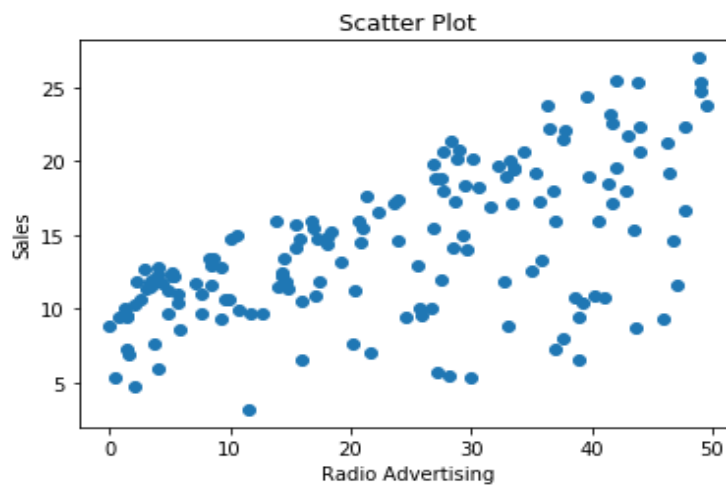
What are the assumptions of linear regression regarding residuals?

Answer: These Assumptions which satisfied while building a linear regression model produces a best fit model for the given set of data.

1. **Linear Relationship between the features and target:** According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

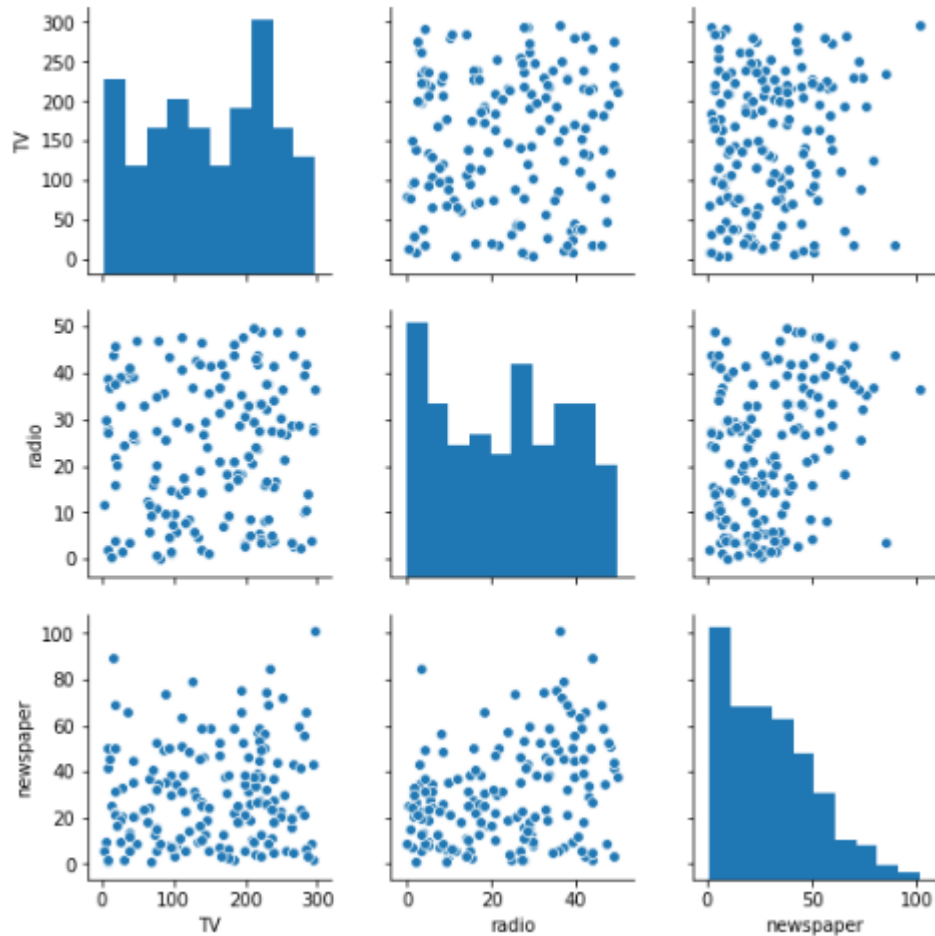


Scatter plot of the feature TV vs Sales tells us that as the money invested on Tv advertisement increases the sales also increases linearly.



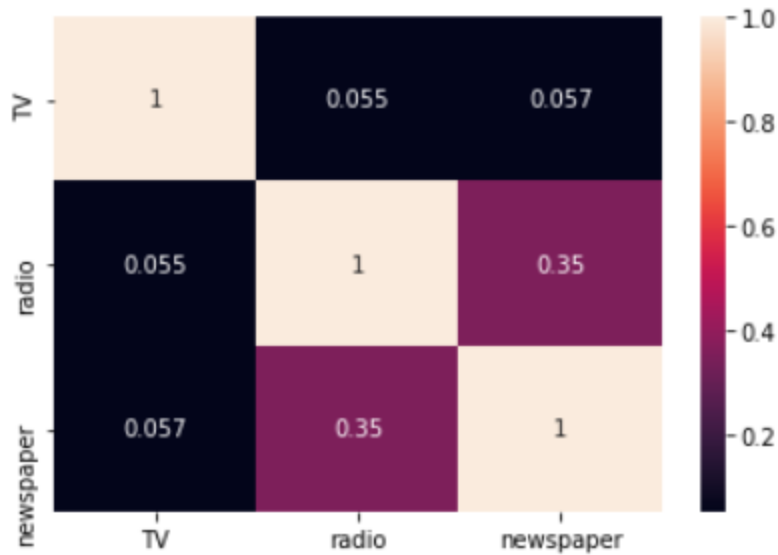
Scatter plot which is the feature Radio vs Sales also shows a partial linear relationship between them, although not completely linear.

2. **Little or no Multicollinearity between the features:** Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features.



The above pair plot shows no significant relationship between the features.

3. **HeatMap (Correlation Matrix):** This heatmap gives us the correlation coefficients of each feature with respect to one another which are in turn less than 0.4. Thus the features aren't highly correlated with each other.



Why removing highly correlated features is important?

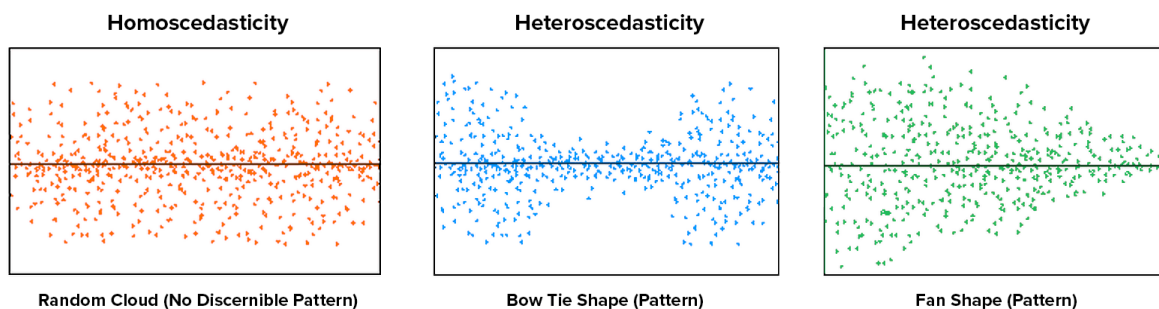
The interpretation of a regression coefficient is that it represents the mean change in the target for each unit change in an feature when you hold all of the other features constant. However, when features are correlated, changes in one feature in turn shifts another feature/features. The stronger the correlation, the more difficult it is to change one feature without changing another. It becomes difficult for the model to estimate the relationship between each feature and the target independently because the features tend to change in unison.

How multicollinearity can be treated?

If we have 2 features which are highly correlated we can drop one feature or combine the 2 features to form a new feature, which can further be used for prediction.

3. Homoscedasticity Assumption:

Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.



Homoscedasticity vs Heteroscedasticity

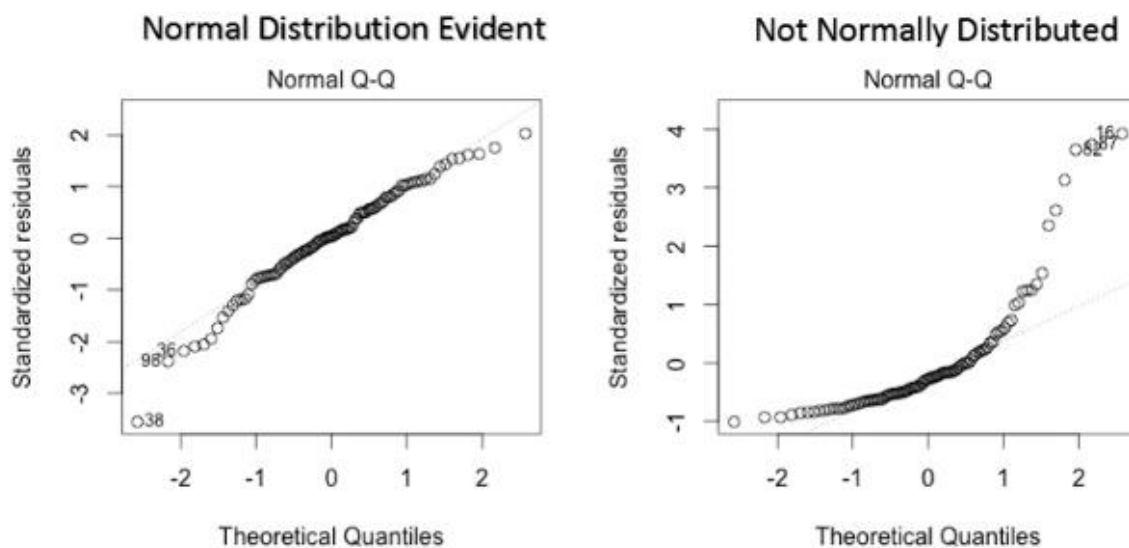
The leftmost graph shows no definite pattern i.e constant variance among the residuals, the middle graph shows a specific pattern where the error increases and then decreases with the predicted values violating

the constant variance rule and the rightmost graph also exhibits a specific pattern where the error decreases with the predicted values depicting heteroscedasticity

4. Normal distribution of error terms:

The fourth assumption is that the error(residuals) follow a normal distribution. However, a less widely known fact is that, as sample sizes increase, the normality assumption for the residuals is not needed. More precisely, if we consider repeated sampling from our population, for large sample sizes, the distribution (across repeated samples) of the ordinary least squares estimates of the regression coefficients follow a normal distribution. As a consequence, for moderate to large sample sizes, non-normality of residuals should not adversely affect the usual inferential procedures. This result is a consequence of an extremely important result in statistics, known as the central limit theorem.

Normal distribution of the residuals can be validated by plotting a q-q plot.



Q. What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation is "R" value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words, Coefficient of Determination is the square of Coefficient of Correlation.

R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850 ^a	.723	.690	4.57996
a. Predictors: (Constant), weight, horsepower				
b. Dependent Variable: mpg				

Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R square value i.e. .723 (or 72.3%). R square is simply square of R i.e. R times R.

Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R square is a better term. You can explain R square for both simple linear regressions and also for multiple linear regressions.

Q. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY**, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

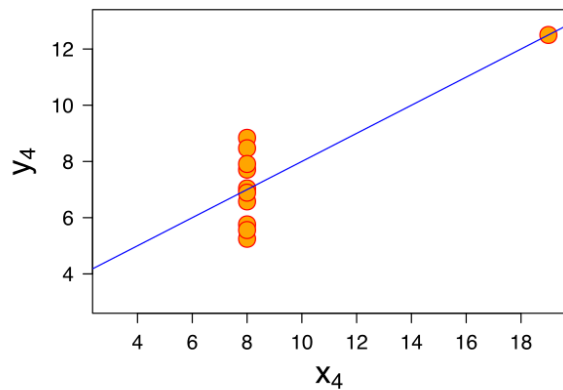
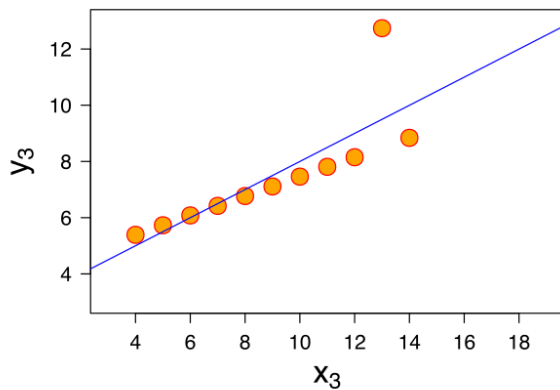
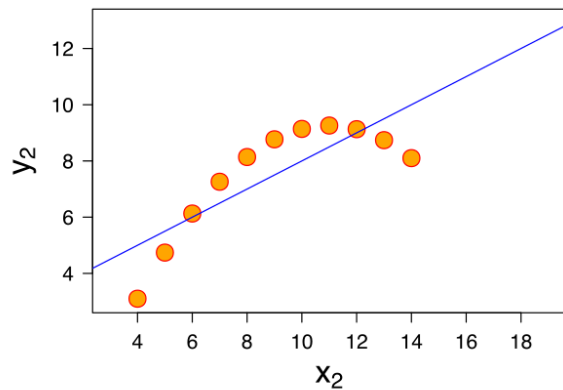
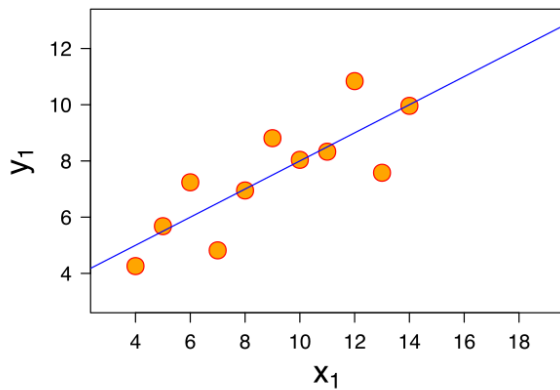
The summary statistics show that the means and the variances were identical for x and y across the groups :

Mean of x is 9 and mean of y is 7.50 for each dataset.

Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Dataset I appear to have clean and well-fitting linear models.

Dataset II is not distributed normally.

In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Q. What is Pearson's R?



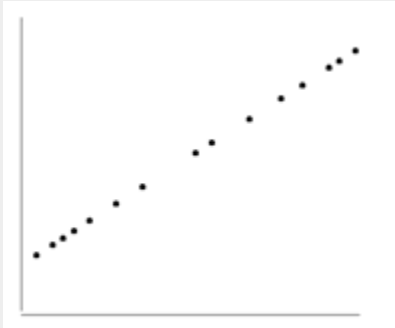
Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

$r = -1$		data lie on a perfect straight line with a negative slope
$r = 0$		no linear relationship between the variables
$r = +1$		data lie on a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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

Feature scaling can vary our result a lot while using certain algorithms and have a minimal or no effect

Why Scaling

Most of the times, our dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.

How to Scale Features

There are common methods to perform Feature Scaling.

Standardisation:

Standardisation replaces the values by their Z scores.

$$x' = \frac{x - \bar{x}}{\sigma}$$

This redistributes the features with their mean $\mu = 0$ and standard deviation $\sigma = 1$. `sklearn.preprocessing.scale` helps us implementing standardisation in python.

Mean Normalisation:

$$x' = \frac{x - \text{mean}(x)}{\text{max}(x) - \text{min}(x)}$$

This distribution will have values between **-1 and 1** with $\mu=0$.

Standardisation and **Mean Normalization** can be used for algorithms that assumes zero centric data like **Principal Component Analysis(PCA)**.

Min-Max Scaling:

$$x' = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}$$

This scaling brings the value between 0 and 1.

Unit Vector:

$$x' = \frac{x}{||x||}$$

Scaling is done considering the whole feature vector to be of unit length.

Min-Max Scaling and Unit Vector techniques produces values of range [0,1]. When dealing with features with hard boundaries this is quite useful.

For example, when dealing with image data, the colors can range from only 0 to 255.

When to Scale

Rule of thumb any algorithm that computes distance or assumes normality, scale your features!!!

Some examples of algorithms where feature scaling matters are:

1. k-nearest neighbors with an Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.
2. Scaling is critical, while performing Principal Component Analysis(PCA). PCA tries to get the features with maximum variance and the variance is high for high magnitude features. This skews the PCA towards high magnitude features.
3. We can speed up gradient descent by scaling. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.
4. Tree based models are not distance based models and can handle varying ranges of features. Hence, Scaling is not required while modelling trees.
5. Algorithms like Linear Discriminant Analysis(LDA), Naive Bayes are by design equipped to handle this and gives weights to the features accordingly. Performing a features scaling in these algorithms may not have much effect.

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

The **variance inflation factor (VIF)** quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

How the VIF is computed

The *standard error* of an *estimate* in a *linear regression* is determined by four things:

- The overall amount of noise (error). The more noise in the data, the higher the standard error.
- The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a *scale* effect).
- The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.
- The extent to which a predictor is correlated with the other predictors in a model.

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the *R-squared* statistic of the regression where the predictor of interest is predicted by all the other predictor variables (). The *variance inflation* for a variable is then computed as:

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

Some statistical software use *tolerance* instead of VIF, where tolerance is:

$$1 - R^2 = \frac{1}{VIF}.$$

The VIF can be applied to any type of predictive model (e.g., CART, or deep learning). A generalized version of the VIF, called the *GVIF*, exists for testing sets of predictor variables and generalized linear models.

How to interpret the VIF

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem (e.g., if estimating *price elasticity*), whereas in straightforward predictive applications very high VIFs may be unproblematic.

If one variable has a high VIF it means that other variables must also have high VIFs. In the simplest case, two variables will be highly correlated, and each will have the same high VIF.

Where a VIF is high, it makes it difficult to disentangle the relative importance of predictors in a model, particularly if the standard errors are regarded as being large. This is particularly problematic in two scenarios, where:

The focus of the model is on making inferences regarding the relative importance of the predictors. The model is to be used to make predictions in a different data set, in which the correlations may be different.

The higher the VIF, the more the standard error is inflated, and the larger the confidence interval and the smaller the chance that a coefficient is determined to be statistically significant.

Q. What is the Gauss-Markov theorem?

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

Gauss Markov Assumptions

There are five Gauss Markov assumptions (also called conditions):

- Linearity: the parameters we are estimating using the OLS method must be themselves linear.
- Random: our data must have been randomly sampled from the population.
- Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
- Exogeneity: the regressors aren't correlated with the error term.
- Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

Q. Explain the gradient descent algorithm in detail.

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

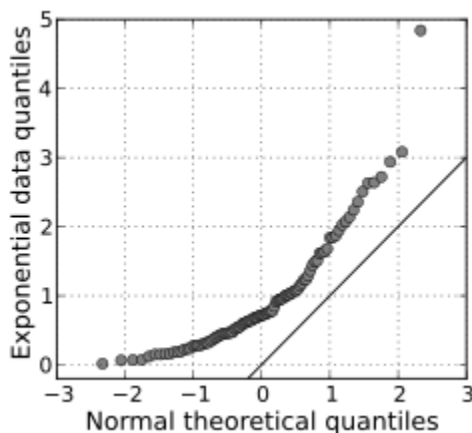
Batch Gradient Descent: This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.

Stochastic Gradient Descent: This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.

Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where $b < m$ are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

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

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.



A Q Q plot showing the 45 degree reference line.

The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot**. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

How to Make a Q Q Plot

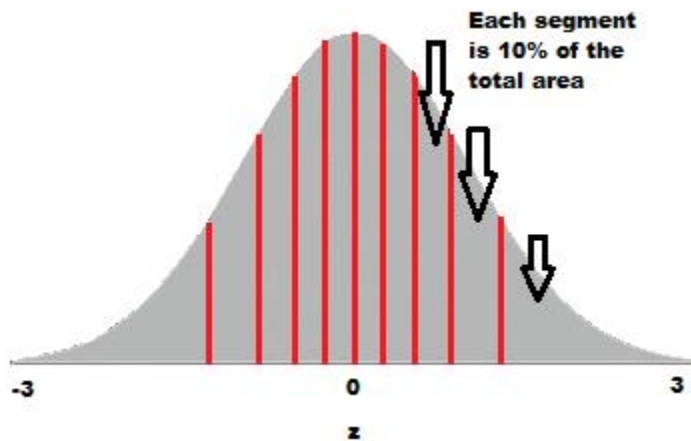
Sample: Do the following values come from a normal distribution?

7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: **Order the items from smallest to largest.**

3.77
4.25
4.50
5.19
5.89
5.79
6.31
6.79
7.19

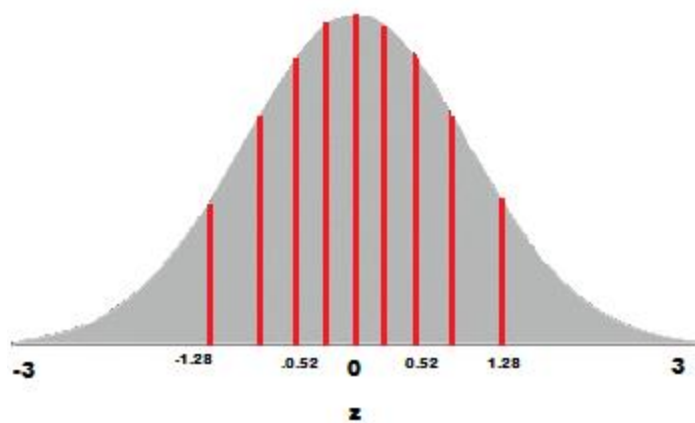
Step 2: **Draw a normal distribution curve.** Divide the curve into $n+1$ segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because $100\% / 10 = 10\%$).



Step 3: **Find the z-value (cut-off point) for each segment** in Step 3. These segments are *areas*, so refer to a z-table (or use software) to get a z-value for each segment.

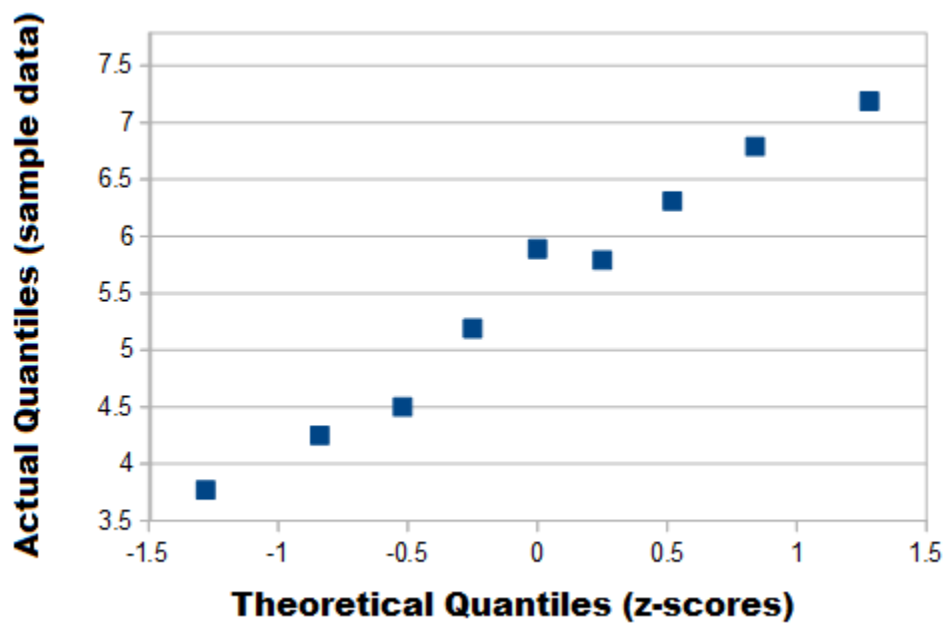
The z-values are:

10% = -1.28
20% = -0.84
30% = -0.52
40% = -0.25
50% = 0
60% = 0.25
70% = 0.52
80% = 0.84
90% = 1.28
100% = 3.0



A few of the z-values plotted on the graph.

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3). I used Open Office for this chart:



The (almost) straight line on this q q plot indicates the data is approximately normal.