1. **Explain the linear regression algorithm in detail.**

Linear Regression model is a model building technique which helps in creating a model to predict the value of dependent variable using single or multiple independent variable. In simpler terms, we have 'y' as dependent variable and 'x' as independent variable. So with the help of Linear Regression we can predict the value of 'y' using 'x'. Linear Regression model finds the best fit line on given data of 'y' and 'x'. Taking a real life example, we want to predict the price of house based on its locality, so the linear regression model finds the best fit line on data available of house price and locality.

The best fit line means, it will create the equation of line that is best to predict the value of dependent variable. The equation of line is y =mx + c, where m is the slope and c is the intercept. Linear Regression model will calculate the best value of intercept ('c') and slope ('m')
Linear Regression can only be utilize in case we have historical data available. Also Linear Regression model will not help us to predict the categorical variable. Whenever building the Linear Regression Model, all the data should be in numerical format. We need to convert the categorical variable into numeric format using different methods available

Linear Regression model can be used to predict the value in the range of the values of observations that have been used for building the model .It cannot predict the value when the input value is out of range of data point provided while creating the model. This is called Extrapolation. In technical terms Linear Regression Model can be used in case of Interpolation and not in case of extrapolation.

There are certain assumption before building the model which are mentioned below

- There should be a linear relationship between dependent variable(y) and dependent variable(x)
- Error terms should follow the normal distribution with mean equal to zero. Error terms is the difference between predicted value and actual value.
- All the error terms should be independent of each other the next value of 'y' should not be dependent on previous value of 'y'.
- Error terms should have a constant variance. Suppose difference of error term at x=5 and x = 10 is 0.5 then difference error term of x = 10 and x = 15 should also be 0.5. The error terms must follow this pattern

2. **What are the assumptions of linear regression regarding residuals?**

When Linear Regression model predict the value of dependent variable there is always chance that there should be some difference between actual value and predicted value of dependent variable. In case there is no difference between actual value and predicted value then there is case of over fitting, this also implies model is not perfect
The difference between the predicted value of the dependent variable (*y*) and the actual value is called the **residual** (*e*). There are certain assumptions regarding the residual value which are as follows

- It is assumed that all the residuals are normally distributed
- When we take the mean of all the residuals it should be around zero. This also means residuals are normally distributed around zero.
- Residuals should have a constant variance. Suppose difference residual at x=5 and x = 10 is 0.5 then difference of error term of x = 10 and x = 15 should also be 0.5. All the residuals must follow this pattern. This assumption is also known as the assumption of homogeneity or homoscedasticity.
- All the residual should be independent of each other.

3. **What is the coefficient of correlation and the coefficient of determination?**

The coefficient of correlation is used to indicate the relation between the independent variable and the dependent variable. The coefficient of correlation is has a range of -1.00 to +1.00. If the coefficient is zero then the independent variable and dependent variable is not related. If the coefficient is -1 then the independent variable and dependent variable is negatively co-related (If the value of one variable increase then the value of other variable will decrease). If the coefficient is 1 then the independent variable and dependent variable is positively co-related (If the value of one variable increase then the value of other variable will also increase.)

The coefficient of determination is a used to indicate the percentage change in the amount of the dependent variable that is "explained by" the changes in the independent variables. The coefficient of determination is the square of the correlation (r) between predicted y value and actual values. The coefficient of determination ranges from 0 to 1. The coefficient of determination is denoted by R^2. An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R^2 of 1 means the dependent variable can be predicted without error from the independent variable

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

Where N is number of observation, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

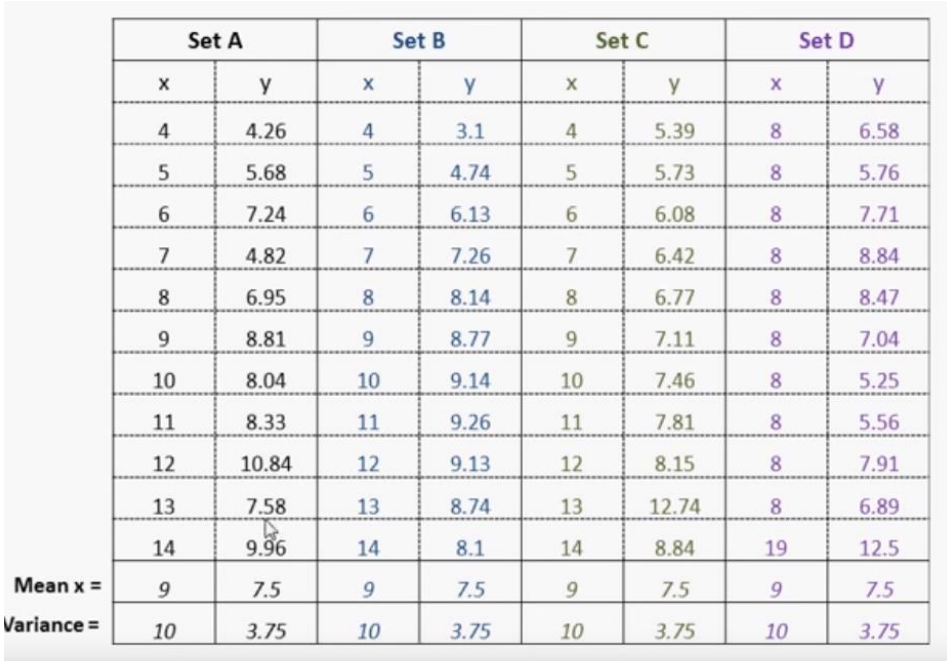
If there are multiple dependent variable used in model then we should calculate the Adjusted R^2

$$\text{Adjusted } R2=1-(1-R2)(N-1)N-p-1$$

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.

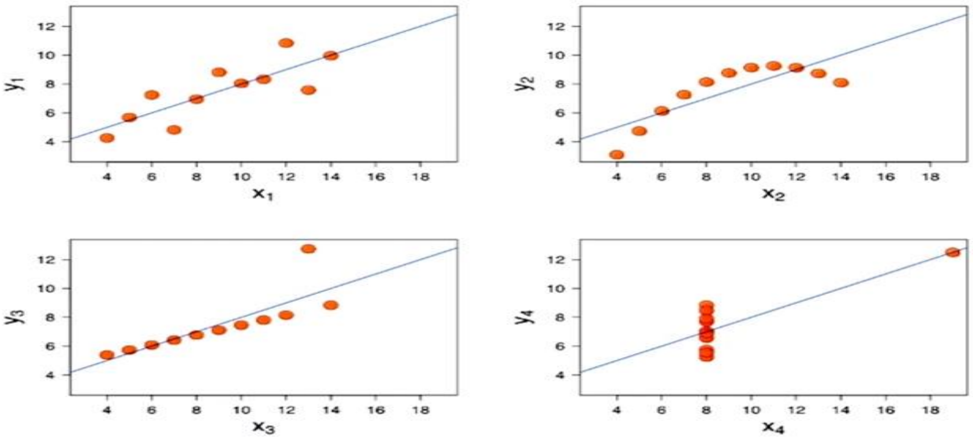
4. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a group of four data set that have identical statistical properties. But when we plot the graph of all four data set they look totally different.



	Set A		Set B		Set C		Set D	
	x	y	x	y	x	y	x	y
	4	4.26	4	3.1	4	5.39	8	6.58
	5	5.68	5	4.74	5	5.73	8	5.76
	6	7.24	6	6.13	6	6.08	8	7.71
	7	4.82	7	7.26	7	6.42	8	8.84
	8	6.95	8	8.14	8	6.77	8	8.47
	9	8.81	9	8.77	9	7.11	8	7.04
	10	8.04	10	9.14	10	7.46	8	5.25
	11	8.33	11	9.26	11	7.81	8	5.56
	12	10.84	12	9.13	12	8.15	8	7.91
	13	7.58	13	8.74	13	12.74	8	6.89
	14	9.96	14	8.1	14	8.84	19	12.5
Mean x =	9	7.5	9	7.5	9	7.5	9	7.5
Variance =	10	3.75	10	3.75	10	3.75	10	3.75

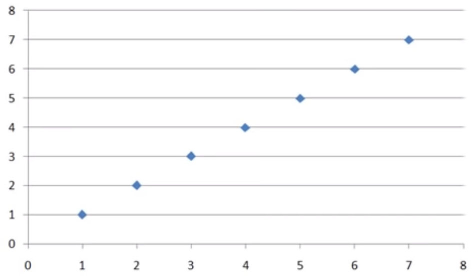
In the above data set we have four set of data A,B,C,D. All the four data set have same mean and variance. Let plot the above data



As it is evident from the above graph's we ned to visualize the data for better understanding of data. Statistict is just a tool for analysis, they should be supported by anecdoctal analysis and visualisation

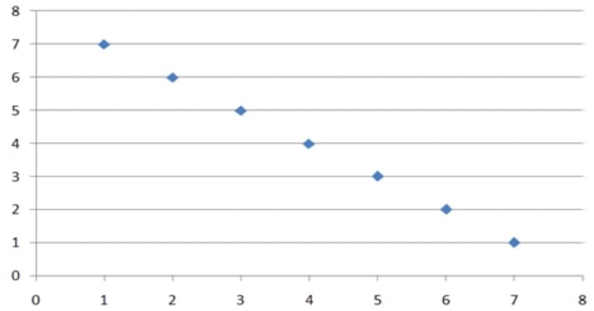
5. What is Pearson’s R?

Pearson’s R is used to measure the strength of linear relationship between two variable. Pearson’s R is always between -1 and 1. If r is equal to 1 then both the variable are positively correlated (if one variable increase other variable will also increase)



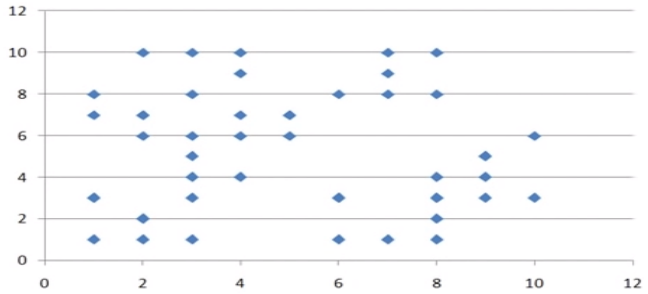
r = 1.0

If r is equal to -1 then both the variable are positively correlated (if one variable increase other variable will decrease)



r = -1.0

If r is equal to zero then there is no relation between the variable



r = 0

The formula for Pearson is a follow

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

For the Pearson *r* correlation, both variables should be normally distributed.
This correlation coefficient is designed for linear relationships and it might not be a good measure for if the relationship between the variables is non-linear. The other correlation coefficient is Spearman's R which is used to determine the correlation if the relationship between the variables is not linear.
So even though, Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. If the relation between two variable is non linear we should look at Spearman's R instead or Pearson's R. It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

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

Scaling is a method used to normalize the range of independent variables. Most of the times, dataset will contain features highly varying in magnitudes, units or range. If one column is in range of 1-100 and other column is in range of 1000- 100000 there can be computational problem for machine learning algorithm. To overcome and speedup the computational power of algorithm we need to bring all the value within same range, this can be achieved by scaling. There are various scaling technique available for scaling, but most commonly used techniques are ***'Normalization*** and ***'Standard Scaling'.*** In Standard Scaling we replace the value by their Z Score

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

Standard Scaling redistributes all the value with their mean **$\mu = 0$** and standard deviation **$\sigma = 1$**

Rescaling(Min Max Normalization) can be calculated by below formula

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

MinMax Scaling bring all the values between 0 and 1.

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

When we have multiple independent variable used in model to predict the value of dependent variable, there is always a chance that there can be some type of relation between independent variable this can lead to multicolinearity problem. To check, how much the independent variable are related we can use Variance Inflation Factor. Detecting multicollinearity is important because while it does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a multiple regression model. VIF can be calculated by below formula

$$\mathrm{VIF}_i = 1/1 - R_i^2$$

Here, VIF_i is the value of VIF for the ith variable, R_i^2 is the R^2 value of the model when that variable is regressed against all the other independent variables.

If the value of VIF is high for a variable, In simple terms, the variable is linearly dependent on some other variables.. The other independent variable are able to expain the variable having high VIF Value. Below table denote the ideal value for VIF

VIF	CONCLUSION
1	No multicollinearity
1-5	Moderate
5 OR GREATER	Severe

8. **What is the Gauss-Markov theorem?**

Ordinary Least Squares (OLS) method is widely used to estimate the parameters of a linear regression model. For the validity of OLS estimates, there are assumptions made while running linear regression models.

- It is assumed that all the residuals are normally distributed
- When we take the mean of all the residuals it should be around zero. This also means residuals are normally distributed around zero.
- Residuals should have a constant variance. Suppose difference residual at x=5 and x = 10 is 0.5 then difference of error term of x = 10 and x = 15 should also be 0.5. All the residuals must follow this pattern. This assumption is also known as the assumption of homogeneity or homoscedasticity.
- All the residual should be independent of each other.

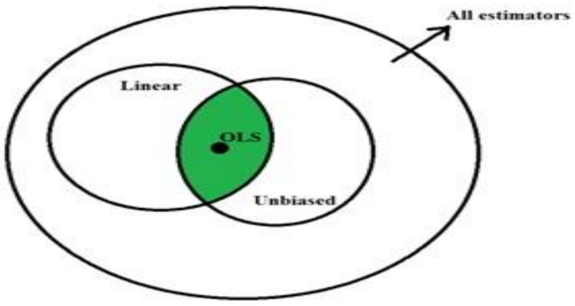
These assumptions are extremely important because violation of any of these assumptions would make OLS estimates unreliable and incorrect.

The Gauss-Markov Theorem

Let the regression model be: $Y=\beta_o+\beta_i X_i+\varepsilon$

Let β_o and β_i be the OLS estimators

According to the Gauss-Markov Theorem, under the assumptions $_5$ of the linear regression model, the OLS estimators β_o and β_i are the Best Linear Unbiased Estimators (BLUE). This theorem tells that one should use OLS estimators not only because it is unbiased but also because it has minimum variance among the class of all linear and unbiased estimators.



OLS has minimum variance of all estimators that lie in the green region (provided assumptions A1 to A6 are satisfied)

Property 1: Unbiasedness

If you look at the regression equation, you will find an error term associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number. Therefore, before describing what unbiasedness is, it is important to mention that unbiasedness property is a property of the estimator and not of any sample. Consider a simple example: Suppose there is a population of size 1000, and you are taking out samples of 50 from this population to estimate the population parameters. Every time you take a sample, it will have the different set of 50 observations and, hence, you would estimate different values of βo and βi. The unbiasedness property of OLS method says that when you take out samples of 50 repeatedly, then after some repeated attempts, you would find that the average of all the βo and βi from the samples will equal to the actual (or the population) values of βo and βi. In layman's term, if you take out several samples, keep recording the values of the estimates, and then take an average, you will get very close to the correct population value. If your estimator is biased, then the average will not equal the true parameter value in the population.

Property 2: Best: Minimum Variance

- If the estimator is unbiased but doesn't have the least variance – it's not the best!
- If the estimator has the least variance but is biased – it's again not the best!
- If the estimator is both unbiased and has the least variance – it's the best estimator.

9. Explain the gradient descent algorithm in detail.

Optimization is always the goal whenever we are building a machine learning model. Since Gradient descent is an optimisation algorithm, it help us to optimize our model. In linear regression, the model targets to get the best-fit regression line to predict the value of y based on the given input value (x). While training the model, the model calculates the cost function which measures the Root Mean Squared error between the predicted value (pred) and true value (y). The model targets to minimize the cost function.

To minimize the cost function, the model needs to have the best value of θ_1 and θ_2. Initially model selects θ_1 and θ_2 values randomly and then itertively update these value in order to minimize the cost function untill it reaches the minimum. By the time model achieves the minimum cost function, it will have the best θ_1 and θ_2 values. Using these finally updated values of θ_1 and θ_2 in the hypothesis equation of linear equation, model predicts the value of x in the best manner it can.
The below formula is used to find θ_1 and θ_2

$$J(\theta_0,\theta_1)=\sum_{N_{i=1}}(y_i-y_i(p))^2$$

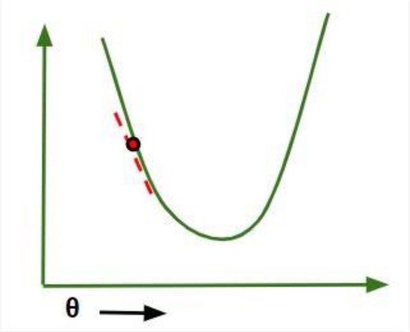
After having initial value of θ_1 and θ_2 we need to minimize their value which can be achieved by Iterative form solution. We can minimize θ_1 and θ_2 by applying the partial derivates method on above function to find the minimized θ_1 and θ_2

$$\theta_1=\theta_0-\eta\ (\partial/\partial\theta J(\theta))$$

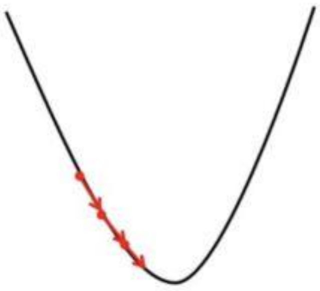
Where η is known as the learning rate, which defines the speed at which we want to move towards negative of the gradient.

The choice of correct learning rate is very important as it ensures that Gradient Descent converges in a reasonable time. :

- If we choose η **to be very large**, Gradient Descent can overshoot the minimum. It may fail to converge or even diverge.



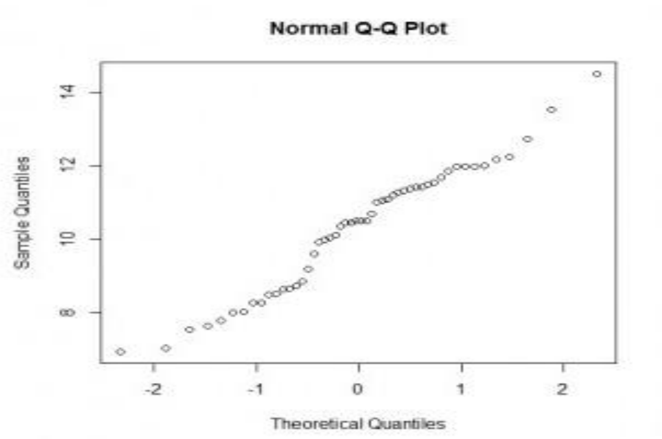
- If we choose η to be very small, Gradient Descent will take small steps to reach local minima and will take a longer time to reach minima.



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

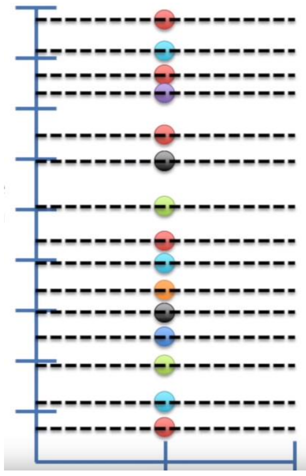
Quantile is basically a fraction where certain values fall below that quantile. For example Median is a quantile where 50% of data falls below 50% quantile and 50% of data above 50%quantile. This implies median divides the data in two equal parts. The basic purpose of QQ Plots is to find whether two set of distribution comes from same distribution.

A Q-Q plot is a scatterplot 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.



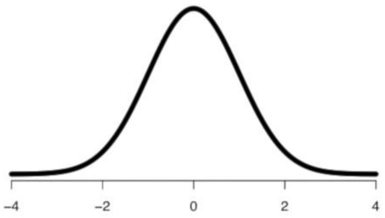
Below are the steps to create the QQ Plot

1. Give each data its own quantile



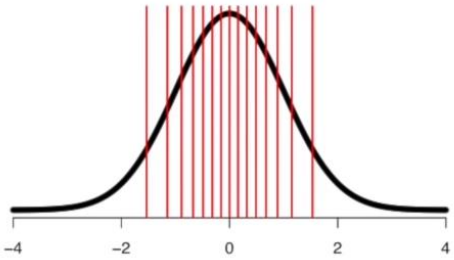
Step 1: Give each point its own quantile.

2. Get a normal curve

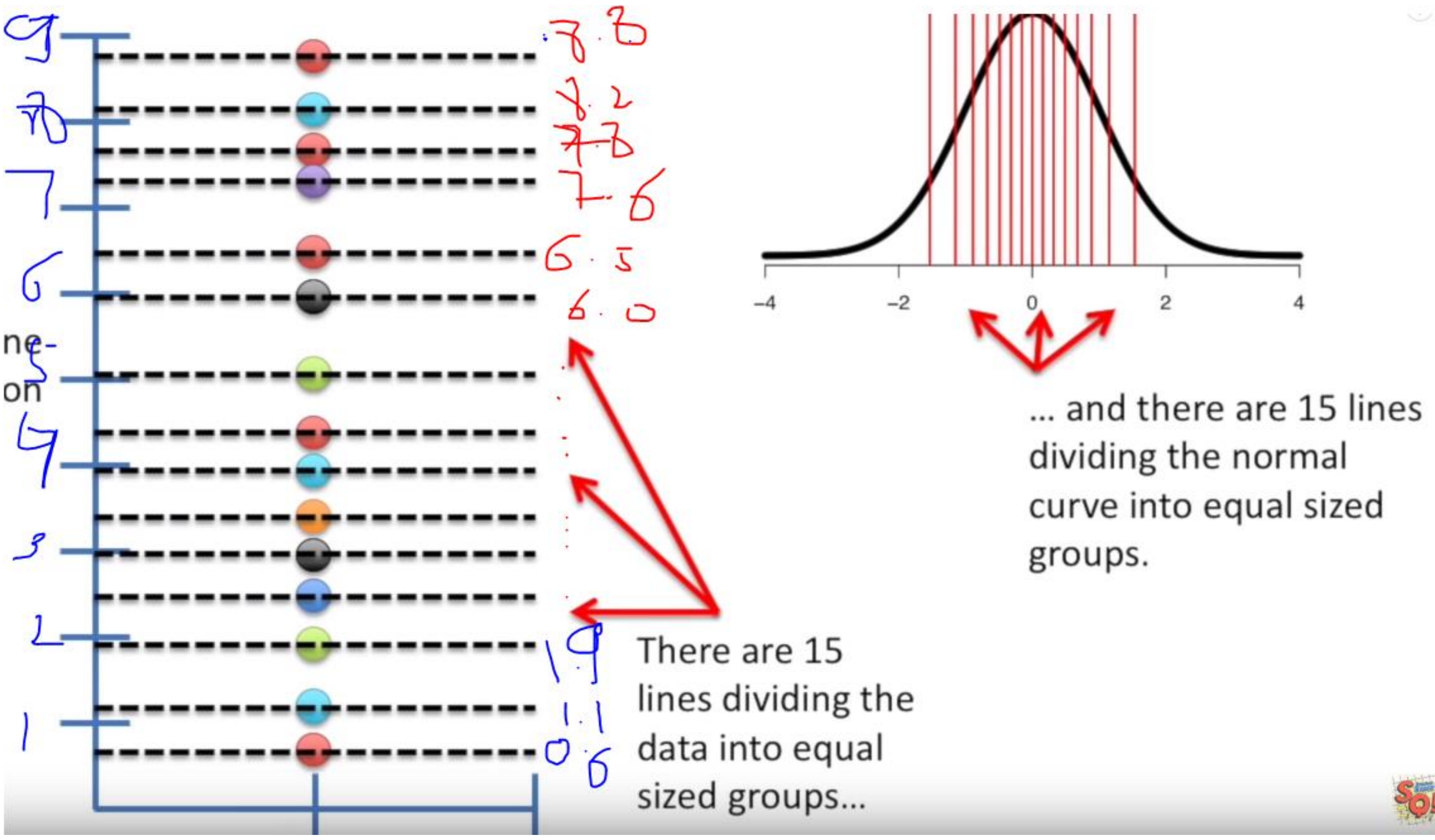


Step 2: Get yourself a normal curve (any normal curve will do)

3. Add the same number of quantile to the curve as we created for data



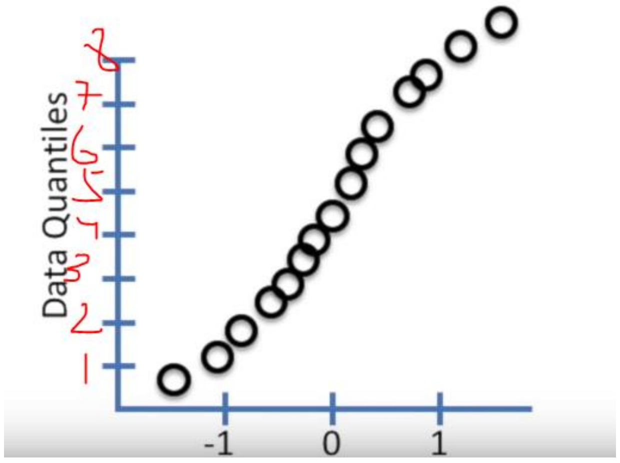
Step 3: Add the same number of quantiles to the curve as you created for the data.



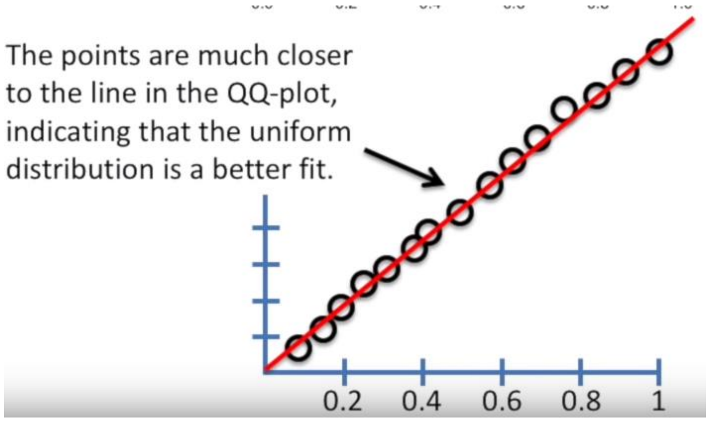
There are 15 lines dividing the data into equal sized groups...

... and there are 15 lines dividing the normal curve into equal sized groups.

4. Plot QQ Graph. On X axis we take the quantile we get from normal distribution and on y axis we take the quantile which we get from data



5. Now we will fit straight line through the dots. If the data is normally distributed most of the points would be on line. But that is not the case so the data does not follow normal distribution
When wew tried QQ plot for normal distribution below is QQ plot for uniform distribution



The points are much closer to the line in the QQ-plot, indicating that the uniform distribution is a better fit.

QQ plot is used in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don’t, residuals doesn’t have follow normal distribution. This implies that for small sample sizes, you can’t assume your estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid.