1. What is data normalization? How is it different from database normalization (1st/2nd/3rd)?

Normalization is the process of organizing data in a database. This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency.

Put simply, data normalization **ensures uniformity in how your data looks, reads, and can be utilized**—across all of the records in your customer database. This is done by standardizing the formats of specific fields within your customer database.

**1NF**

The first normal form, otherwise known as 1NF, has the following rules:

* Uniquely Identify each row (with a primary key)
* Multiple value columns must be separated, so that there is only one value in each column per row
* Consistent data type must be enforced for each column

# 2NF

Second normal form, or 2NF, has only one rule:

* Every non-key column in a table must depend on the value of the key

# 3NF

Third normal form, or 3NF, again only has one rule:

* Every non-key column must ONLY depend on the value of the key

1. What is a distribution? What are the uses for frequency and probability distribution?

A statistical distribution, or probability distribution, **describes how values are distributed for a field**. In other words, the statistical distribution shows which values are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution.

While **a frequency distribution gives the exact frequency or the number of times a data point occurs, a probability distribution gives the probability of occurrence of the given data point**. When the number of test cases are large, the frequency distribution and the probability distributions are similar in shape.

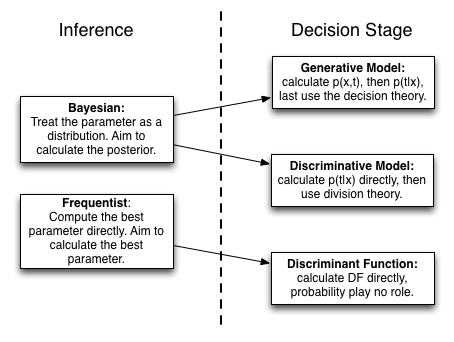
In Statistics, the probability distribution **gives the possibility of each outcome of a random experiment or event**. It provides the probabilities of different possible occurrences. Also read, events in probability, here. To recall, the probability is a measure of uncertainty of various phenomena.

1. What is a decision? How's it different from inference?

Decision theory, in statistics, a set of quantitative methods for reaching optimal decisions. A solvable decision problem must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their consequences.

In computability theory and computational complexity theory, a decision problem is a problem that can be posed as a yes–no question of the input values. An example of a decision problem is deciding whether a given natural number is prime.

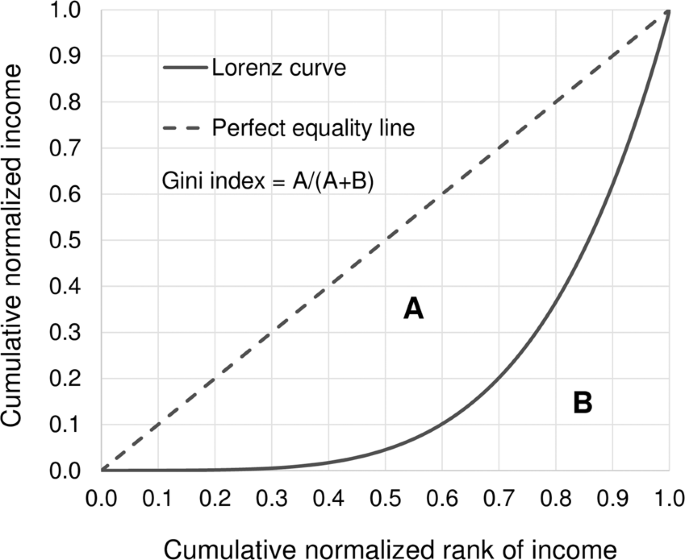
Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data.



1. Google- what is Gini in probability, and explain in your own terms

The Gini index is a measure of the distribution of income across a population. A higher Gini index indicates greater inequality, with high-income individuals receiving much larger percentages of the population's total income.

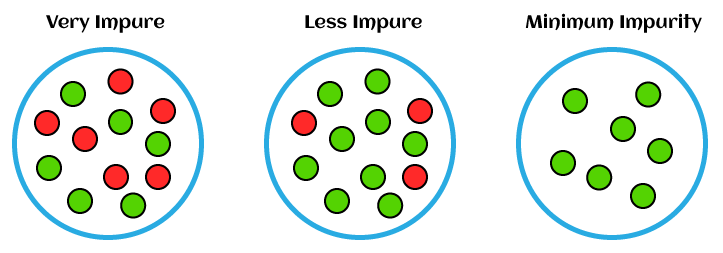
The Gini index is calculated as the ratio of the area between the perfect equality line and the Lorenz curve (A) divided by the total area under the perfect equality line (A + B).



1. What is entropy?

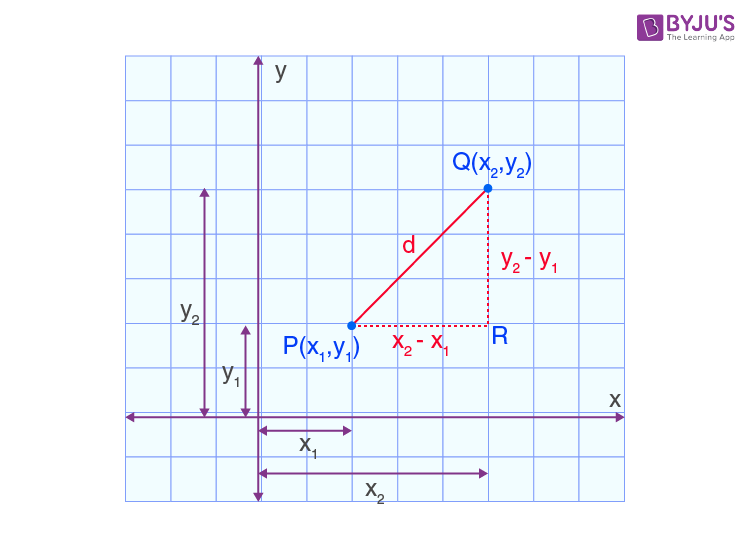
Entropy is a scientific concept as well as a measurable physical property that is most commonly associated with a state of disorder, randomness, or uncertainty.

Information Entropy or Shannon's entropy quantifies the amount of uncertainty (or surprise) involved in the value of a random variable or the outcome of a random process. Its significance in the decision tree is that it allows us to estimate the impurity or heterogeneity of the target variable



1. What is euclidean distance?

In Mathematics, the Euclidean distance is defined as **the distance between two points**. In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.



1. What's the difference between correlation and covariance?

Covariance and correlation are two terms that are opposed and are both used in statistics and regression analysis. Covariance shows you how the two variables differ, whereas correlation shows you how the two variables are related.

This means that as x increases in value, y decreases. For example, a calculated covariance value of -38.15 means that x and y have an inverse relationship in which x increases as y decreases in value. A covariance value of 0 means that x and y have no relationship to one another. Covariance measures the direction of a relationship between two variables, while correlation measures the strength of that relationship. Both correlation and covariance are positive when the variables move in the same direction, and negative when they move in opposite directions.

1. What is mean squared error?

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

There is no correct value for MSE. Simply put, the lower the value the better and 0 means the model is perfect.

