Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 42
RFM and Market Basket Analysis (Contd.)

(Refer Slide Time: 00:33)

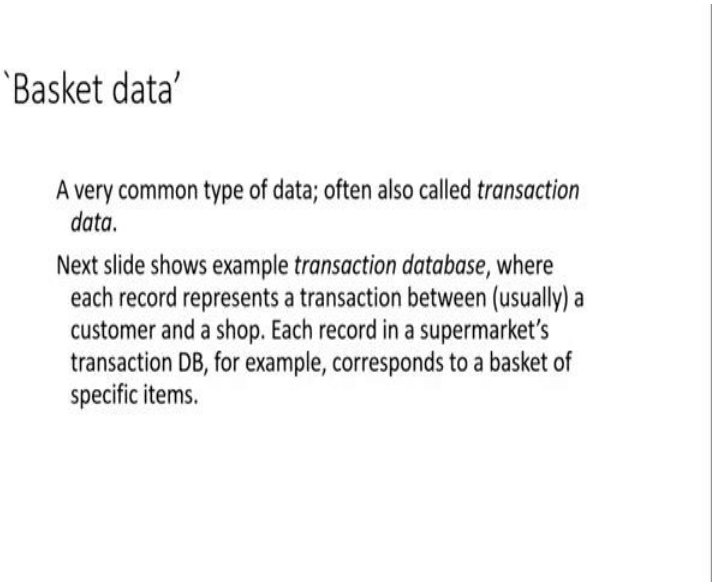
Hello everybody, welcome to marketing analytics course, this is Doctor Swagato Chatterjee from VGSOM IIT Kharagpur who is taking this course and we are in week 8 and session 3, video lecture number 3 also and we will be discussing in this case Market Basket analysis. So, in the earlier videos of this week we have discussed about RFM analysis that is one part of rate analytics which is mainly used in retail but also used in other context.

On the other hand, Market Basket Analysis is something that is used only in retail, so majorly used in ecommerce but also used in the (())(00:53) motor vehicle retail stores. So, this is also generated this particular that the data used in this particular analysis is also generated from the scanner data, what is scanner? The data that is generated when you scan some item in the billing counter that is something called scanner data.

So, we go with a basket, the basket has multiple products and people actually the people in the counter scans the products that are there in our basket and that particular information along with your customer ID probably the loyalty card number, etc. the date, time etc. the person's name who is there in the counter etc. gets stored in the ERP.

When we actually analyze that data after collecting it from ERP and to find out that, what kind of products goes together well and what kind of offer I can make, what kind of product bundles I can make, these are something which we can find out from Market Basket analysis.

(Refer Slide Time: 01:58)



So, the first question is what is a basket data? A very common type of data is basket data, which is often also called transaction data. So, the data that is the, that is coming from our transaction is called transaction data. In the next slide we will show how the transaction database looks like, where each record each particular row represents transactions between usually a customer and a shop.

So, each record in a super markets transaction database for example, corresponds to a baskets specific items, so it is a particular baskets item what are the products that you have bought in one particular visit in a retail store is something that we find out in a transaction data.

Now, often times ecommerce purchase, see the problem, why do we buy lots of products together? Because I cannot go to a retail stores multiple times there is a, if it is a (())(03:04) motor retail store, then there is a cost in terms of going to that particular store. So, if I go multiple times, I will incur that cost multiple times for Kirana stores small mom and pop stores that is there close by to your home, you generally go multiple times and you probably visit multiple times in a week and probably buy at max two items, three items, four items at a time probably sometimes less than that.

But in a in a in a bigger retail store where which is a little bit away from your home it is a supermarket or it is a hypermarket kind of store, there the cost of travel from your home to

that place is high and that is why you plan for that purchase visit and when you plan for this purchase visit you actually note it down, what are the different kinds of products that you are going to purchase in that purchase visit.

And then you go there sometimes you purchase those items which are there in the least, sometimes you do not purchase and purchase something else, which is not there in the list which you purchase based on the impulse buying or based on then and there purchase decision moment of truth based purchase decision making you buy those products.

Now, market basket becomes a basket only when you purchase multiple products and that is more prominent for supermarkets, less prominent for Kirana stores or mom and pop stores. On the other hand, in the case of ecommerce at one point of time, if you remember, we used to purchase so there was a delivery fee and that is why because there was a delivery fee for each item sorry for each purchase transaction.

That is why what we used to do is that we use to buy lots of products together in a basket. Similar things we still do in let us say ecommerce farms, which are food-based ecommerce farm, for example let us say Zomato or Swiggy, when you order food for each visit you have to pay, each transaction you have to pay 15 rupees, 20 rupees as delivery charges.

So, that is why you tend to buy lots of products food products from one restaurant because if you buy it from different restaurant and different transaction happens then different delivery charges will be there. So, that is why for this kind of context, you will have to see the basket kind of a data, where multiple products are there in one transaction.

But let us say for Amazon if you have taken the prime subscription of Amazon or let us say some other subscription for Flipkart or for various ecommerce form where the delivery charges are minimal in those kinds of situations people actually buy one item, two item or often they buy one item at a go.

So, there the basket is not created. So, if the basket is not created, then there is another way you have to deal with the recommendations or the market baskets, so cannot get a basket there, because people do not buy multiple items in one transaction. You buy even book separately, because the purchases are very not very the planned, sometimes it is very impulse or then and there you purchase you decide that I will purchase this and you purchase.

So, often times the level of motivation level of engagement that you can see in a (())(06:33) motor retail store is not so much visible in ecommerce form. And if that is the case, then this Market Basket Analysis might not work in that way. So, this is something that was there in the initial days, ecommerce form used to use to follow Market Basket Analysis quite a bit, but slowly it is going down but still it is an important topic and that is why we will discuss.

(Refer Slide Time: 07:04)

ID	apples,	beer,	cheese,	dates,	eggs,	fish,	glue,	honey,	ice-cream
1	1	1		1			1	1	
2			1	1	1				
3		1	1			1			
4		1				1			1
5					1		1		
6						1			1
7	1			1				1	
8						1			1
9			1		1				
10		1					1		
11					1		1		
12	1								
13			1			1			
14			1			1			
15								1	1
16				1					
17	1					1			
18	1	1	1	1				1	
19	1	1		1			1	1	
20					1				

So, what is basket data? A very common type of data which has transaction data basically and this is how it looks like. So, let us say, each row here is a transaction ID each row and what are the columns? The first column is the transaction ID basically and the second column is apple, beer, cheese, dates, eggs, fish, glue, honey, ice cream, so certain food items that are kept here. And blank means that that part, in that particular thing that is not been purchased and one means it has been purchased.

So, now if I see the first row that means the first row is ID number 1, where apple was bought, beer was bought, dates were bought, glue was bought and honey was bought, so these are the 5 items that was bought the other items has not been bought. The second case, cheese, dates and eggs have been bought. In the third case beer and cheese has been bought, now see just image in the situation that you have multiple products, thousands of products in an ecommerce form all in a (())(08:07) motor store and you have millions of customers.

Now, if you create this kind of a metrics for those kind of every transaction then these data is a huge data, so ideally it is very difficult to do this analysis for the whole data set at a go, so what we do is? We generally wrap it up, for example, we do it for only did dairies dairy

products, only food products or only apparel products and then also transactions which happened in this month only, because sometimes month versus month the transaction pattern can change depending on what offers are going on, depending on what kind of atmosphere, what kind of season it is, various other things might impact.

So, that is why we reduce the number of transactions based on segmenting by now week or month and etc. And then we reduce the number of products also the number of columns in that way by reducing the number of categories, so this is something that we do.

(Refer Slide Time: 09:12)

Now, if we go ahead, what is Market Basket Analysis? Any analysis that is done with this market basket data is called Market Basket Analysis is as simple as that. Now, if I want to think about it in a little bit detailed way, it is the input is the least of purchases by purchaser and we do not have names here, so we do not identify which customer it is, we just have the customer transaction ID that is all. And identify purchase pattern, so this the job of this market basket analysis is to identify mark purchasing pattern.

What items tend to be purchased? So, this are the certain questions that we tried to answer, what items tend to be purchased together. So, for example, steak and potatoes or beers and pretzels, or in our case alcohol and snacks items, which will have that alcohol or it is bread and butter.

These guys are bought together always when you buy bread you jam or butter something you buy. So, these which items tend to be purchase together, sometimes these are obvious, some

of the things are not so obvious, so finding out that not so obvious things is the job of Market Basket Analysis.

What items or purchase sequentially? For example, let us say if you buy a house then you will buy a furniture obviously, there is a sequence. If you buy a car you will later buy tires or buy probably petrol or buy lubricants, so these are some of or let us say if you buy a computer later you will probably buy the spare parts or accessories or internet connections.

So, these are all Wi-Fi routers, these are all basically sequential purchase. So, sometimes we try to see that whether the same data set we can I can make based on whether some things has been bought and something has been bought within that week or within that month, generally cloud them in one transaction ID. What items tend to be purchased by season that is also something which can be answered by Market Basket Analysis. So, in majorly we will focus on not first question.

(Refer Slide Time: 11:19)

So, what is we categorize customer purchase behaviour by doing this and then identify actionable information, so that is something which is very important that you have to identify actionable information, some information based on which you can do something, we can work on it. So, we create purchase profiles that is an actionable information if I know that, this is one purchase profile that is another purchase profile, segmenting the rather than segmenting the customers, we are segmenting the transactions that can be done.

And we can find out the profitability of each purchase profile. If I can break it in segments of transactions, which segment is more profitable I can try find that out. And we can use for

marketing this information, how? In layout or catalogs. So, it is a classic example I told in my in one of our retail marketing class where students that we have you have you ever checked that why does people put the product which are used for cleaning utensils, let us say Vim bar or some gels and etc.

So, the gels or Vim bar and etc. which is cake based products, cake based dishwashing products are not kept just beside crockery items, high end crockery items. Beside high end crockery items you put gel based thing and then the cake based thing you put it in the place of soaps and etc. or let us say for the detergents and etc. are kept.

So, ideally this is a Market Basket Analysis result that you have analyzed that people who use to buy crockery which is high end they also buy something in a sequence they buy something which they feel that is good for their hand, good for the product, that the crockery that they bought and etc. So, they buy a high end, more pricey, more quality, giving or more value giving product which is let us say the gel version of dishwashing products.

Now, on the other hand, those who do not buy crockery, who buy normal let us say utensils, which is metallic utensils, they might want to they might be okay with buying a Vim bar kind of product as well. So, that is how you position the product, because you know that these two guys combined with each other, this guy two guys might go with each other.

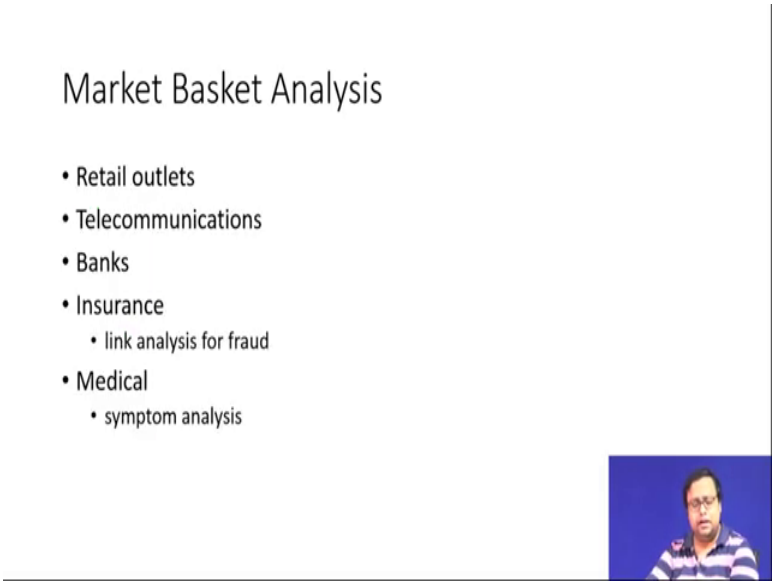
So, the layout catalogs of your it can be a (())(14:02) motor store, it can be a ecommerce store also, ecommerce store it is very common that if you see in Amazon, you are seeing one particular product and they are in the bottom it has a bundle is given that whoever purchased this also purchase have that and these two products together is 100 rupees off something like that. So, at the bottom you come to know all of these thing in details are given.

So, that is also our example of layout or catalog billing. Select products for promotion, which products will be used for promotion that can come up from Market Basket analysis. As I just told, product placement, space allocation, these things can also come out as a result of market basket analysis.

(Refer Slide Time: 14:44)

So, Steve Schmidt, president of ACNielsen US, has told that Market Basket benefits are selection of promotions, merchandising strategy, sensitive to price, like Italian entrees, pizza, pies, Oriental entrees and orange juice, these are also important. Then uncover customer spending patterns this also can be done and john promotional opportunities, like combining them, bundling them, etc. can be done using this analysis technique.

(Refer Slide Time: 15:14)



Why it is used? It is used in retail outlets mainly as I told it is also used in telecommunications, it is also used in banks, it is also used in TV bundles. Have you seen that, the there is the bundle that has been created by Tata sky for sports lovers, or for regional product lovers or Hindi movie lovers or English movie lovers.

So, they create different kind of product bundles at different pricing for them that is also one example of Market Basket Analysis you can find out that whoever watches these also watches that and that is why and they have chosen these kind of channels together when they have created their own combination, so that is why I am also offering this kind of a combination. In banks, in insurance, in medical, this can be used.

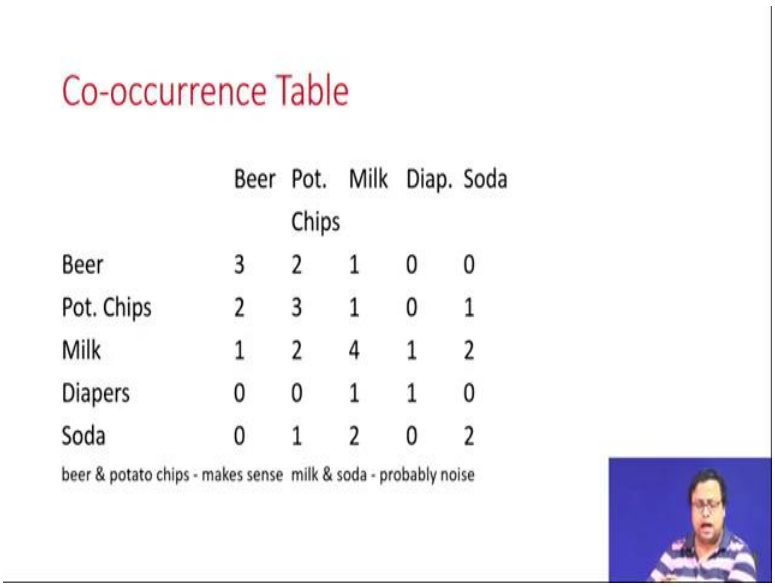
(Refer Slide Time: 16:03)

So, chain store age executive in 1995, it has been used and then customer shop on personal needs, not on product grouping, so this is also that is a important factor. Initially we associate products by category and then what percentage of his category was in each market basket that is something that we analyze, but that we have to keep in mind that customer shop based on their personal needs, not on product groupings.

(Refer Slide Time: 16:31)

So, let us say these are my 5 customers, 6 customers data, these are the basically transaction data not customer data and these are the products that has been bought together.

(Refer Slide Time: 16:44)



Now, if I put it in this way, see that every time beer has been bought, out of them there are beer and pot. Chips combination is 2, out of this six one, beer and potato chips happening together there were two such cases. Beer are at milk happening together there were one such cases, soda and milk happening together there were two such cases and so on. And most common product is milk. See milk and milk happening together is for this most common product is milk.

So, you just see that beer and potato chips is making sense, but milk and soda happening together probably there is no underlying meaning, why milk and soda will come together? I do not know. So, that is something that you have to think about that why milk and soda is coming together, so probably a noise. So, sometimes you have to identify the noises as well.

(Refer Slide Time: 17:41)

What are the profiles that you get? Some of the very classic profiles are beauty conscious, health conscious, sport conscious, sub consciousness, unconsciousness, also are probably less aware or less carefree kind of likely smokers, casual drinker can be. Then new family, illness over a counter I means like when you buy a Paracetamol or when you buy a some Metrogyl or Flagyl kind of products which is used for your stomach upset and etc. that is illness over the counter product, people buy like anything.

So, let us see we are going for a travel, you want to buys this over the counter products, stock it, so that during the travel if you face something, so these kinds of products comes with together, like some digestive items, some products which related to pain, some product which is related to basic first stage your fever all of these products can come together.

Then home handyman products or sentimental products, there can be certain purchase profiles which is focused on kids or pets, gardening, certain hobbies. Automotive is also an hobby, TV or stereo enthusiast also an hobby, then seasonal or traditional kind of products, homemakers, there are lots of purchase profile I will not read them up there and lots of purchase profiles that can be created and this is something that regularly the marketing managers do to find out that purchasing profiles.

(Refer Slide Time: 19:15)

For example, in a beauty conscious purchasing profile, what are the products that can come together? One example can be considered, cotton ball, hair dye, cologne and nail polish, these are some of the products which can come very easily together in a beauty conscious one.

(Refer Slide Time: 19:35)

So, now you have to design that out of this purchase profile if I can find out from Market Basket lots of purchase profile which one will I push and which will I not push, so ideally I will push that one which will give me better profit. For example, kids fashion is a high margin purchase profile 15.24 dollar is something that I make up of it from each transaction, so I will put there.

But let us a smoker, the purchase profile is 2.88 dollar 1 will not push their, student or home office, they are highly price sensitive they take lot of time before purchasing anything, so I will not push them. So, we generally actually create a from upper purchase profile to lower purchase profile we sought them up based on the profitability of their purchase profile, the average margin or average profit that I generate from each purchase profile and based on that we push some people which we push do not push some other profiles, so that is something that is also an application of market basket analysis.

(Refer Slide Time: 20:40)

Now, if other things that then Affinity Positioning like coffee and coffee makers is close proximity it can be put or cross selling can be done. So, if you are buying cold medicine, I can also ask you to buy, digestive medicine or orange juice or something like that. So, if you Monday night Football kiosks on Monday pm that can also be shown. So, this kind of things can be done, cross selling and also be done.

(Refer Slide Time: 21:05)

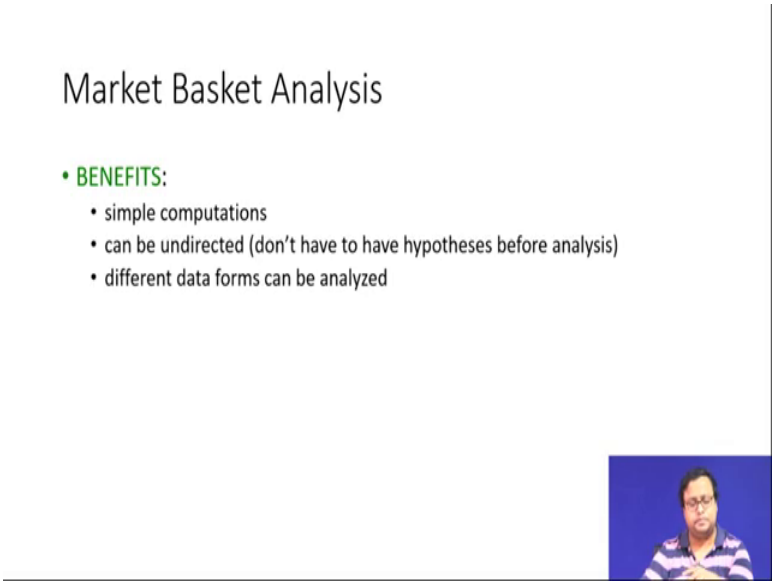
Now, I have talked about lots of good things about Market Basket Analysis, but every good thing will have a bad thing as well, there are some corns as well, what are the limitations of market basket analysis? So, some of the limitations of Market Basket Analysis can be handled by priority minority, recommendation engine and etc. But still it is very easy that is why it is more used, it is more fast and easy than recommendation engine.

So, it takes over 18 months to implement that is one of the major. So, you have to create lots of data, if you have lots of data from the past, no problem but you have to have the data. And Market Basket Analysis only identifies hypotheses which need to be tested. For example, the one that I have just shown that soda and milk coming together.

It is a hypothesis that they this is a noise or let us say beer and pretzels come together, beer and potato chips come together, it is an hypothesis that they should come together you have to test that using various predictive analytics technique to further implement, but this hypothesis generation is also important.

Measurement of impact needed that is also something is a limitation you do not know how to measure them. Difficult to identify the product groupings. And complexity grows exponentially as the data goes up, because it is again you have to deal with the whole metrics. If you have to deal with the whole metrics, is the cost column numbers goes on the row numbers goes up, you are in a dilemma, you are increasing the complexity exponentially they and that is where the problem is.

(Refer Slide Time: 22:42)



The benefits is, simple computation. Can be undirected that means do not have to have hypothesis before analysis, you can create hypotheses after the analysis and different data forms can be analyzed.

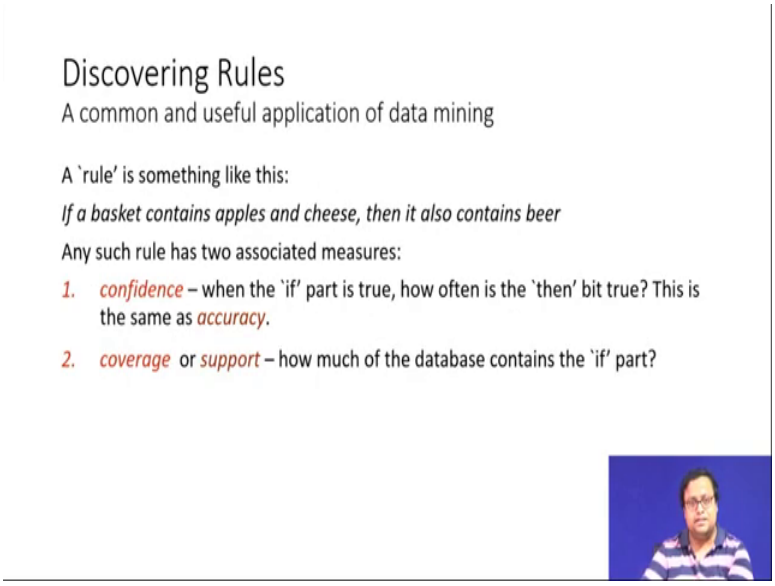
(Refer Slide Time: 22:56)

Now let us come to the numbers. Our example transactions database has 20 records of supermarket transactions, from a supermarket that only sells 9 things. One month in a large supermarket with five stores spread over a reasonably sized city might easily yield a database of 20 million baskets, each containing a set of products from a pool of around 1000. So, to understand what is what that means? That means that I am doing a small problem, real life problems are not like that.

So, we used to see a picture in Facebook and other social media websites that the data analysis that we show in the classroom are basically a puppy and the data analysis that you actually do in the corporate are basically dinosaurs or very big monsters and that is true actually.

So, the one that I will show you in the class because I have limitation of time and etc. that will be 20 records of 9 products, but in real life situation you will have 20 million records and 1000 products and that will probably you have to break your head on that. So, that is, so that is a part of the story.

(Refer Slide Time: 24:05)



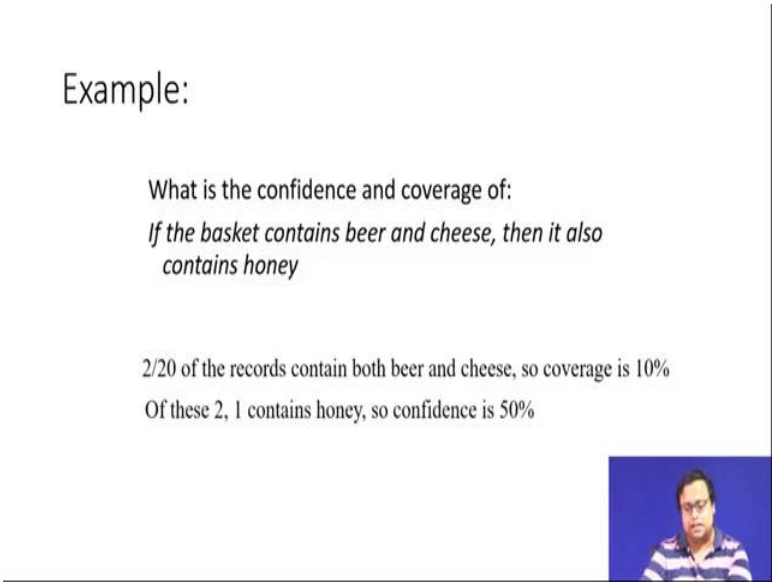
So, how to discover the rules? I told till now, I have spent a lot of time to say that there will be association rules, two products should come together, so how to create? And rule, a common and useful application of data mining, it is a rule is something like this, this is how the rule will sound like.

If a basket contains apple and cheese, then it also contains beer. So, the condition of the rule is the if part, we call it the confidence, confidence is? When how when if part is true, or how often is then be true. So, given that if is true, how often the then is true. So, probability of the then part given that the if part, this is confidence, this is the same as accuracy.

And the other another nomenclature is coverage or support that means, how much of the database contents the if part? That means, how what is the probability of the if part? So, in this particular rule that has been written here, probability of the apples and cheese, probability

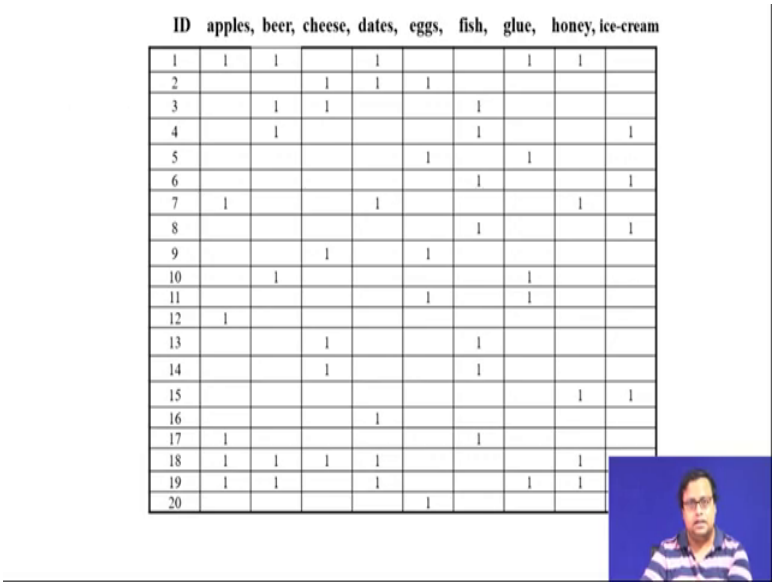
and apple and cheese occurring together is coverage or support. Given that that occurs, what is the probability that beer will occur? Is called confidence.

(Refer Slide Time: 25:30)



For example, what is the confidence and coverage of, if a basket contains beer and cheese, then it also contains honey. So, 2 by 20 records contain both beer and cheese, in that data set that has shown, so coverage is 10 percent. And only out of these 2 only 1 contains honey, so that is why confidence is 50 percent.

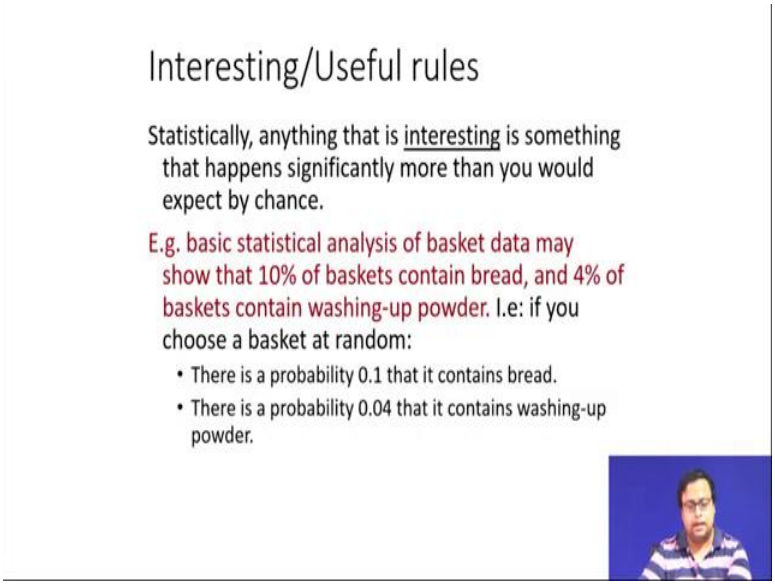
(Refer Slide Time: 25:50)



ID	apples,	beer,	cheese,	dates,	eggs,	fish,	glue,	honey,	ice-cream
1	1	1		1			1	1	
2			1	1	1				
3		1	1			1			
4		1				1			1
5					1		1		
6						1			1
7	1			1				1	
8						1			1
9			1		1				
10		1					1		
11					1		1		
12	1								
13			1			1			
14			1			1			
15								1	1
16				1					
17	1					1			
18	1	1	1	1				1	
19	1	1		1			1	1	
20					1				

So, this is the data set you can check it up.

(Refer Slide Time: 25:53)

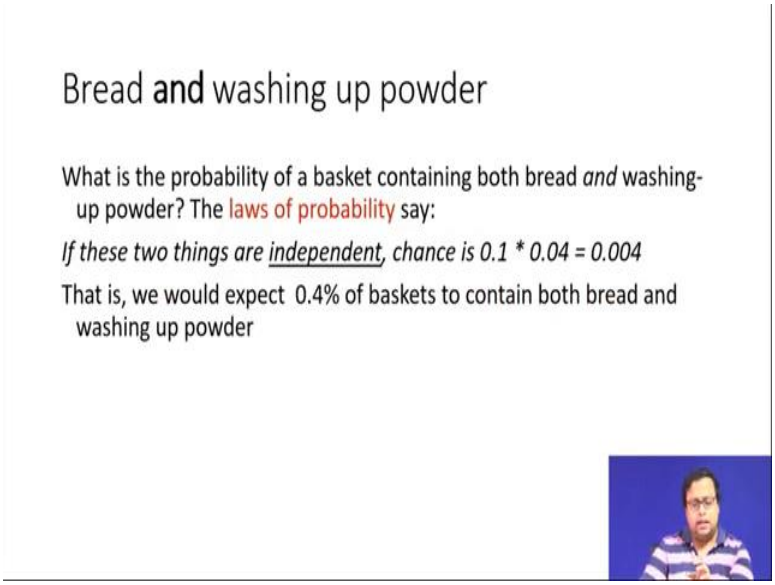


Interesting and useful rules are what? You have to focus on statistically anything that is interesting is something that happens significantly more than you would expect by chance. So, if a rule says that by chance if there was no rule, it happens x amount of time, if there is a rule if there is a condition in which this particular thing suits up then it is interesting.

For example, let us say in general nobody purchases, let us say beer in India, let us assume that in a particular place, nobody purchases beer, but when it rains lots of beer purchase happens. So, that that given that rain is happening beer purchase goes up that rule is interesting, if only that that is more than what is expected commonly.

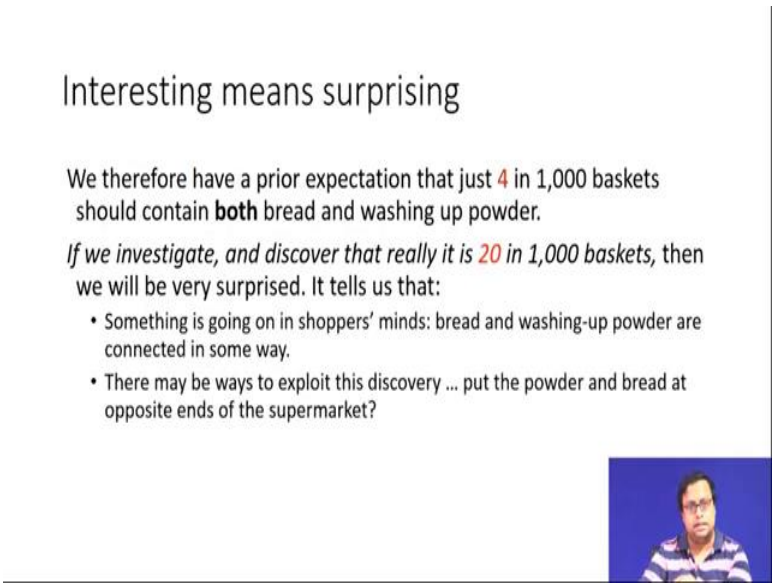
So basic statistical analysis of basket data may show that 10 percent of baskets content bread and 4 percent of baskets content washing powder. That is, if you choose a basket randomly if you just close a basket, the probability of having bread is point 1 and probability of washing powder is 0.04, very low that means you might have bread more commonly than washing powder.

(Refer Slide Time: 27:13)



Now, what is the probability of the basket containing both bread and washing powder? Ideally that should be further lower 0.1 into 0.04 if they are independent to each other, so 0.004. Now, what is we would expect 0.4 percentage of baskets to contain both bread and washing powder in case of independence.

(Refer Slide Time: 27:37)

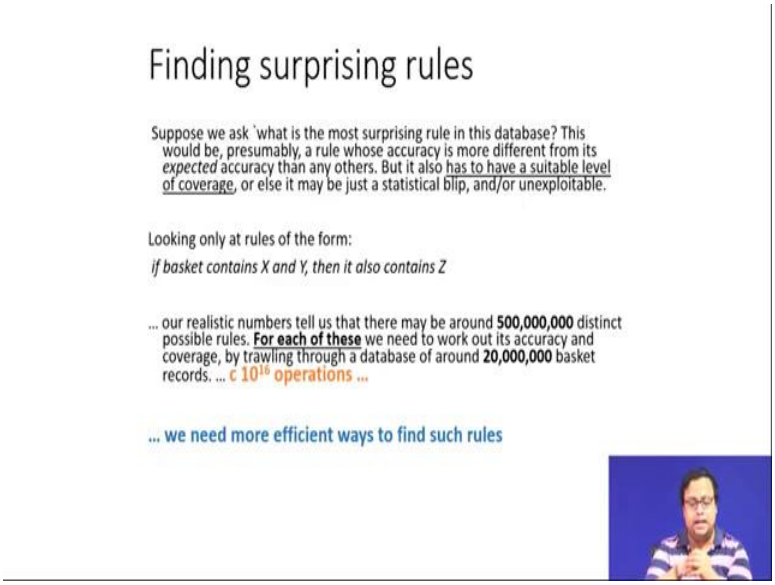


Now, interesting means surprising, now by chance if it is not the case, we therefore have a prior expectation that just 4 in 1000 baskets should contain both bread and washing powder makes sense, because washing powder is 4 percent, bread is 10 percent, if both occurring together if they have no relationship with each other is, 4 percent into 10 percent answer 40 by 10,000 that means 4 in 1000.

Now, if we investigate and discover that 20 in 1000 baskets are actually having bread and washing powder together then that is a surprise element. It tells us something is going on consumers mind that bread and washing powder are connected in some way that is a hypothesis that we have to build. There may be ways to exploit this discovery, put the powder and bread at opposite ends of the supermarket that means what?

You if around, let us say 2 percent people, if they there is a combination that whoever buys these also might buy also that if you put them two opposite corner in the supermarket, then you were making the person walk down the aisles of the supermarket and while he will travel more within the supermarket, the more he travels, the more he buys, so that kind of strategy people take to make sure that the revenue goes up.

(Refer Slide Time: 29:05)

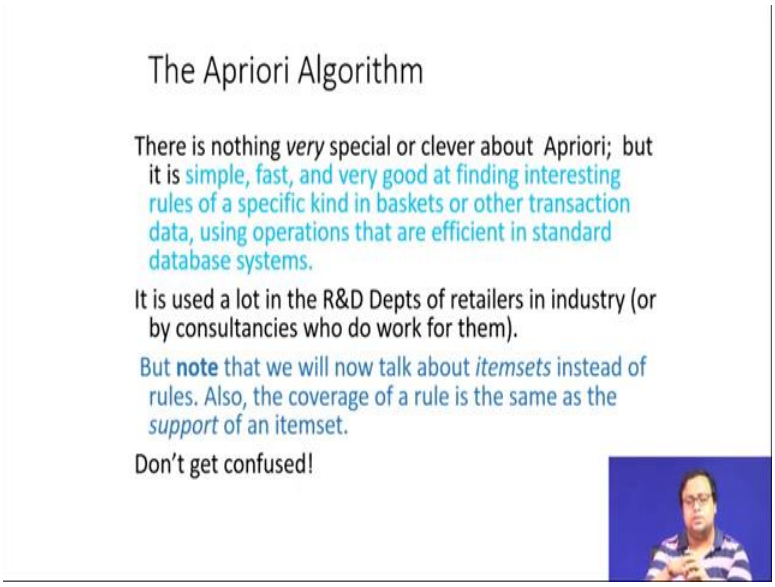


So, finding surprising rules. Suppose we ask, what is the most surprising rule in the database? So, that would be presumably a rule whose accuracy is more different from his expected accuracy than any others. But it also has to be have a suitable level of coverage. So, first of all, accuracy is important to be surprising, but coverage is also important, or else it may be just a statistical blip and not an expert level.

So, looking only at rules of the form. If basket contains x and y it also contains z. Our realistic numbers tell us that there may be around 500 million distinct possible rules, if there are 2 million transactions, 20 million transactions of 1000 products, in around 500 million rules is of obvious, it can happen that many rules. For each of these we need to walk this accuracy and coverage.

And by trawling through the database of around, as I told 20 million basket records. So, that is a huge operation to create all the coverage and the coverage and support for each of them and confidence for each of them. So, we need more efficient ways to find such rules, we cannot do it for all these combinations.

(Refer Slide Time: 30:23)



So, what do we do we use a Apriori algorithm? What is Apriori algorithm? There is nothing very special or clever about Apriori, but it is simple, fast and very good at finding interesting rules on a specific kind in baskets or other transaction data. Using operations that are efficient in standard data base system.

So, it is used a lot in R&D departments of retailers in industry. But note that we will now talk about item sets instead of rules, I am not talking about rules and I am talking about the combination of items. And also, the coverage of a rule is same as the support of that item set. So, a coverage means, how many times the if part happens, there is also similar to the support of the item set.

So, let us see, do not get confused, let us see, what is the algorithm? So, actually, we will deal with this algorithm part in the next video, let us take a break on this particular thing. So, all I discussed till now is, what is the usage of Market Basket Analysis, what is the basic algorithm and now we will use a data mining technique to make it easier, more durable. Let us discuss about that algorithm in the next video. Thank you very much for being with me, I will come back in the next video.