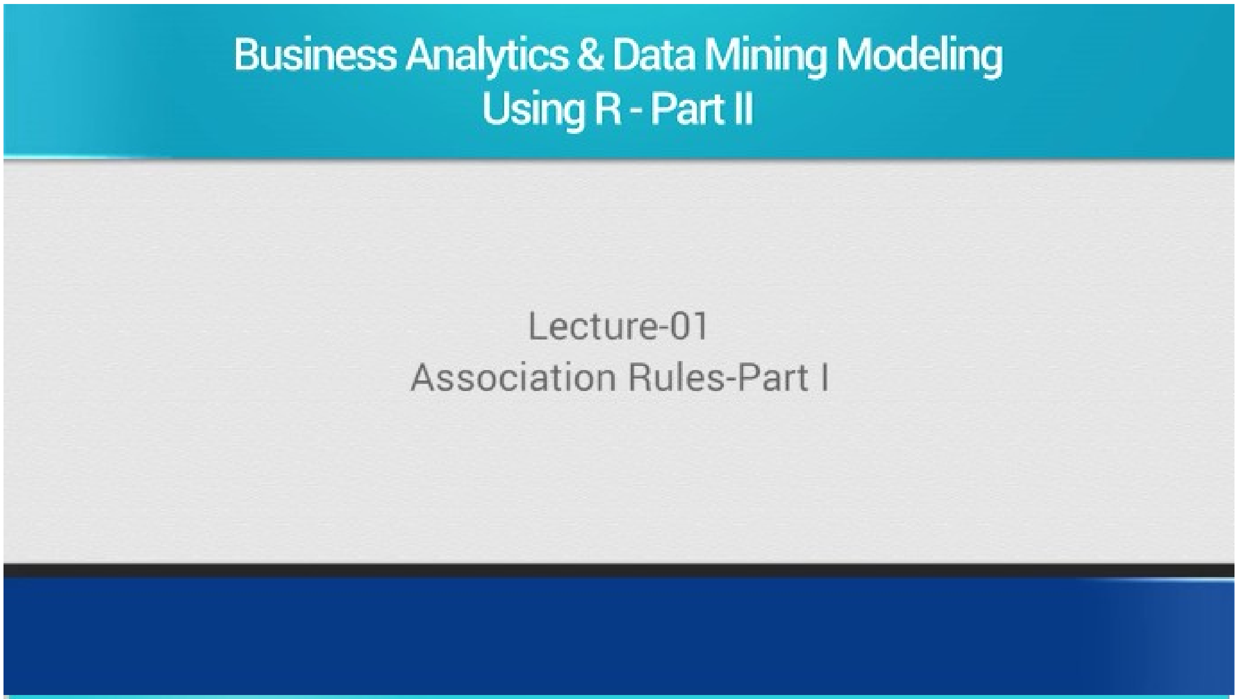
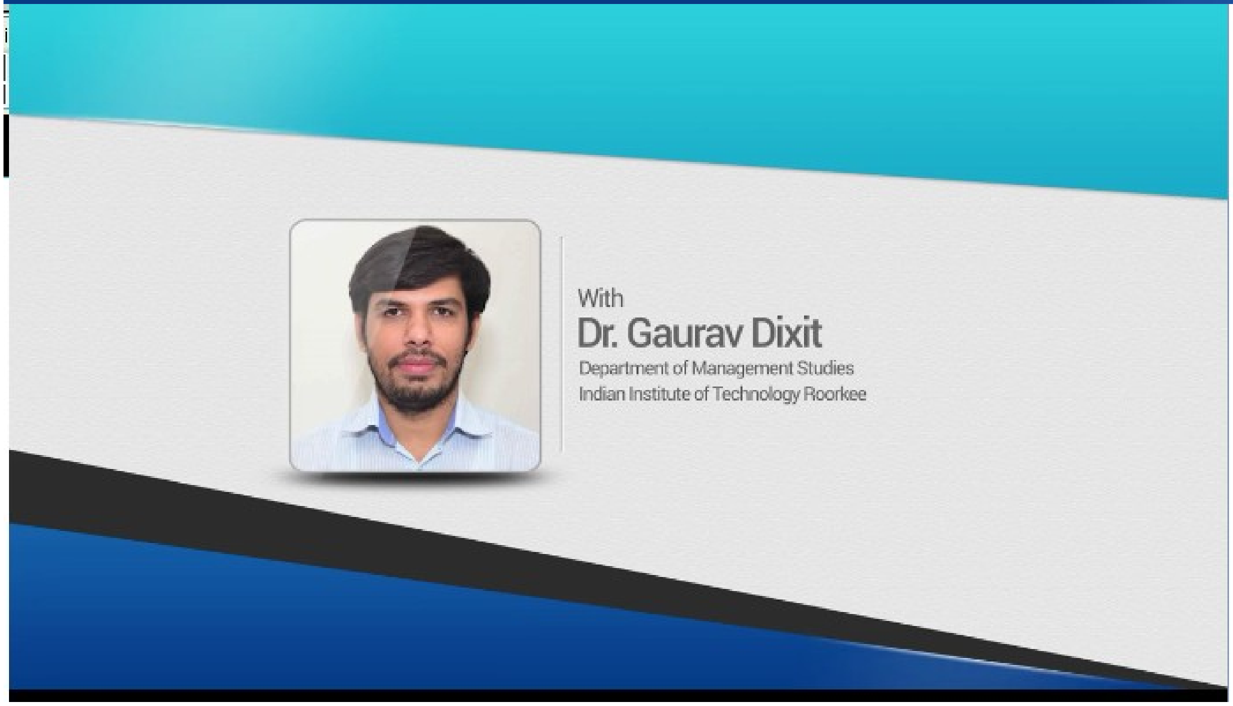
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
NPTEL
NPTEL ONLINE CERTIFICATION COURSE
Business Analytics & Data Mining Modeling
Using R – Part II
Lecture-01
Association Rules – Part 1
With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee



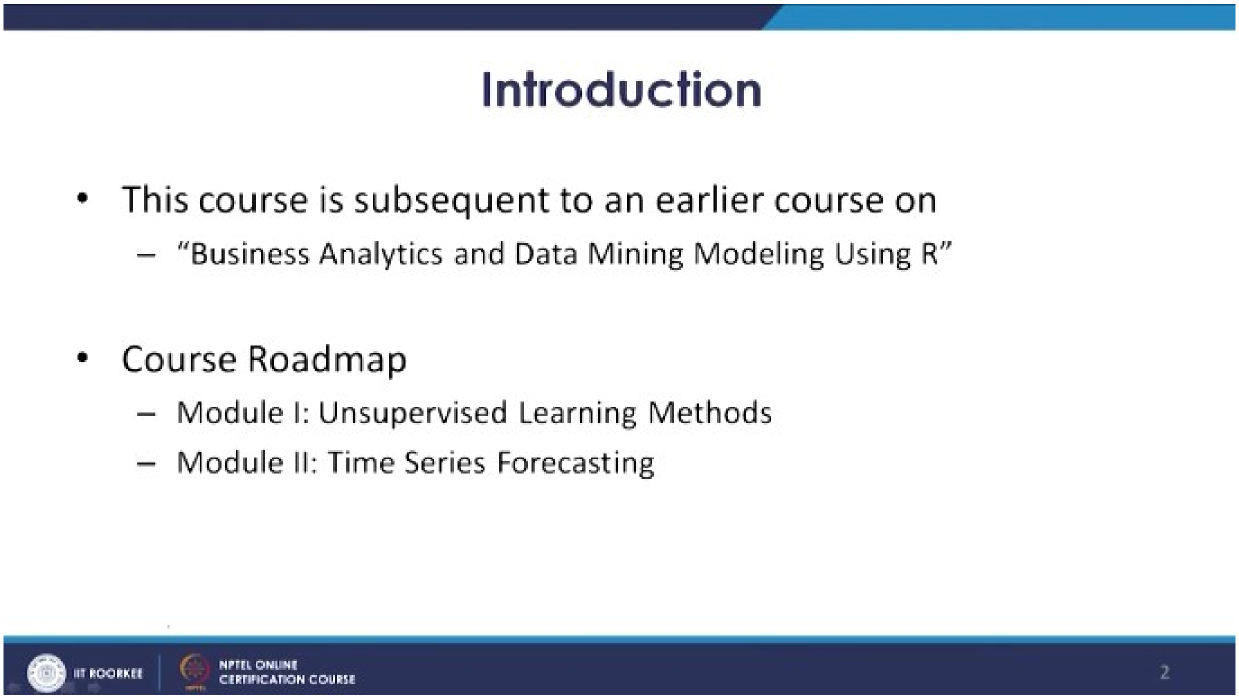
Business Analytics & Data Mining Modeling Using R - Part II

Lecture-01
Association Rules-Part I



With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee

Welcome to the course Business Analytics and Data Mining Modeling Using R Part 2, so this is the very first lecture, and this particular course is subsequent when earlier course on business

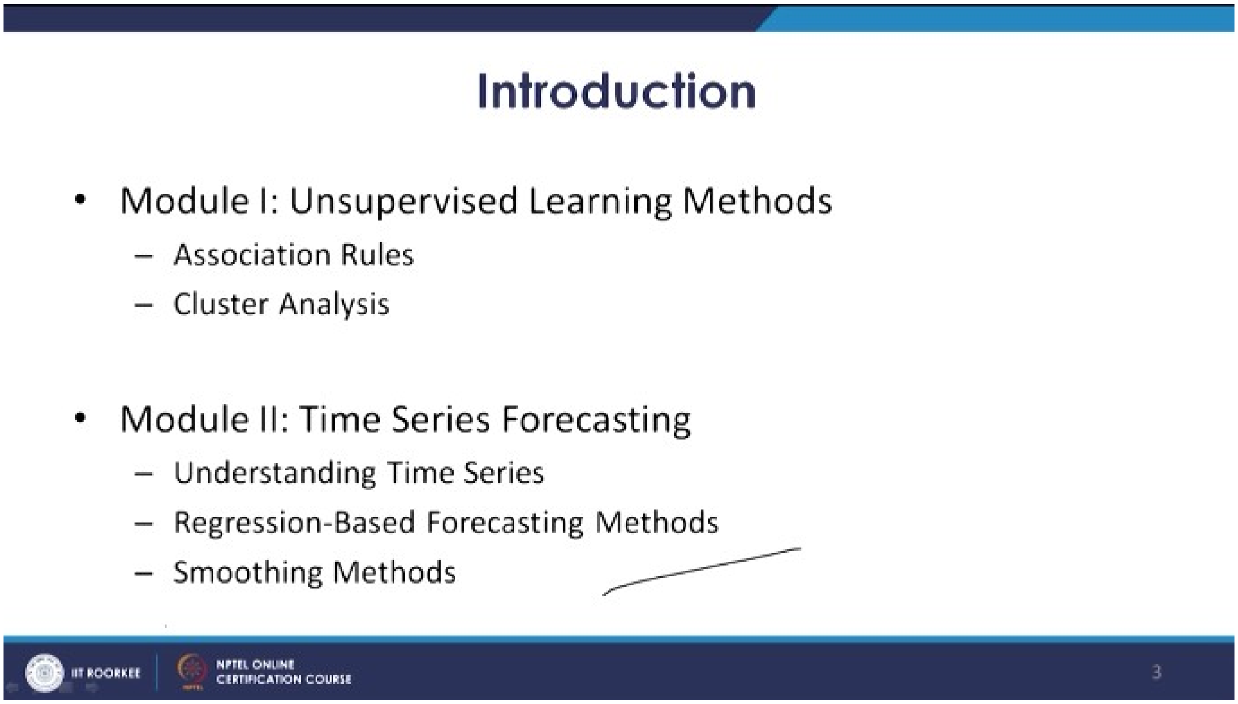


Introduction

- This course is subsequent to an earlier course on
  - "Business Analytics and Data Mining Modeling Using R"
- Course Roadmap
  - Module I: Unsupervised Learning Methods
  - Module II: Time Series Forecasting

analytics and data mining modeling using R, this is part 2 of the previous course, so it is highly recommended that you go through the previous course video lectures, and there you can get, you can finalize yourself with the basic concepts that are required for this course, you'll also get the supplementary lectures on R and basic statistics, so it is highly recommended that you go through the previous course that is called business analytics and data mining modeling using R, this is the, this particular course is part 2 of the previous course, and we are now right now in the very first lecture of this.

So 2 particular modules that we want to cover in this part, the first one is unsupervised learning methods, and the second one is time series forecasting, so the previous course we were able to cover the basic data mining part and then supervise learning methods.

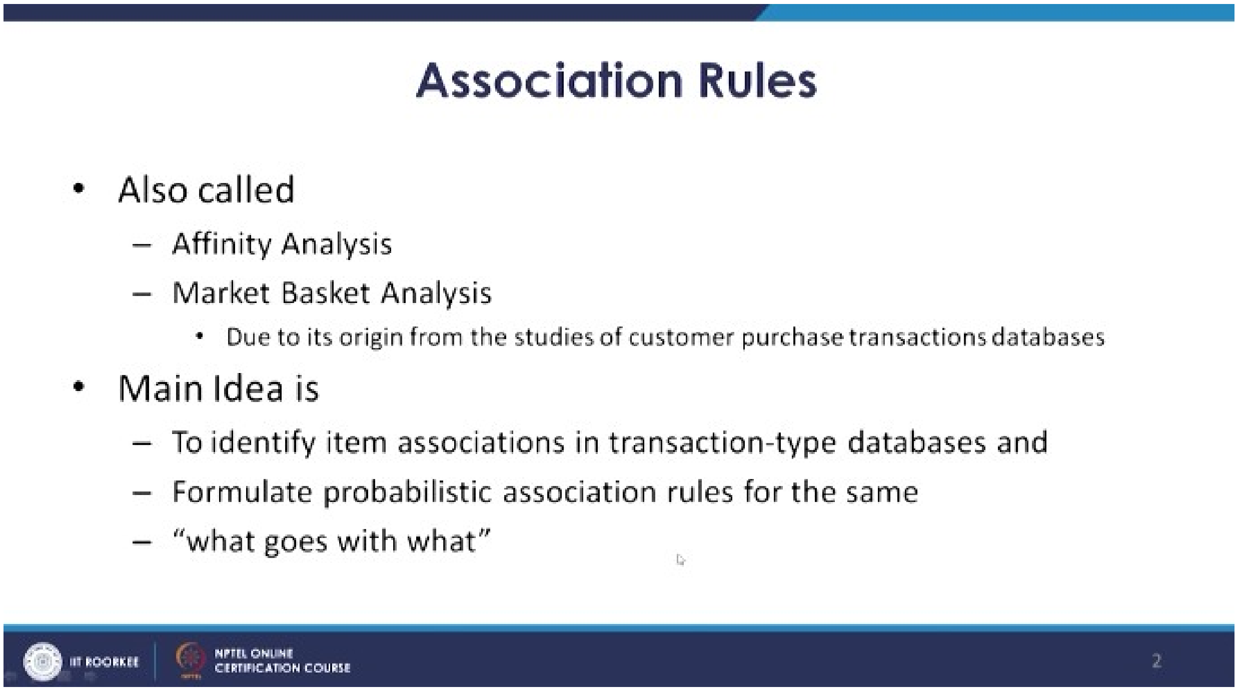


Introduction

- Module I: Unsupervised Learning Methods
  - Association Rules
  - Cluster Analysis
- Module II: Time Series Forecasting
  - Understanding Time Series
  - Regression-Based Forecasting Methods
  - Smoothing Methods

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 3

So now we will move to unsupervised learning methods, in this particular part of the course. So let’s move forward, so in unsupervised learning methods the first module that we have, we’ll discuss these two techniques, first one is association rules, the second analysis is cluster analysis, so these are the two techniques unsupervised you know learning methods that we are going to discuss and do a modeling using different data sets and the software that we use R, and the second module is about time series forecasting so there we have divided into three part, so first one is understanding time series so where we’ll try and understand different components of time series, the second one is where we focus on regression based forecasting methods, and the third one is where we focused on smoothing methods. So with this let’s move forward to our first technique that is association rules.



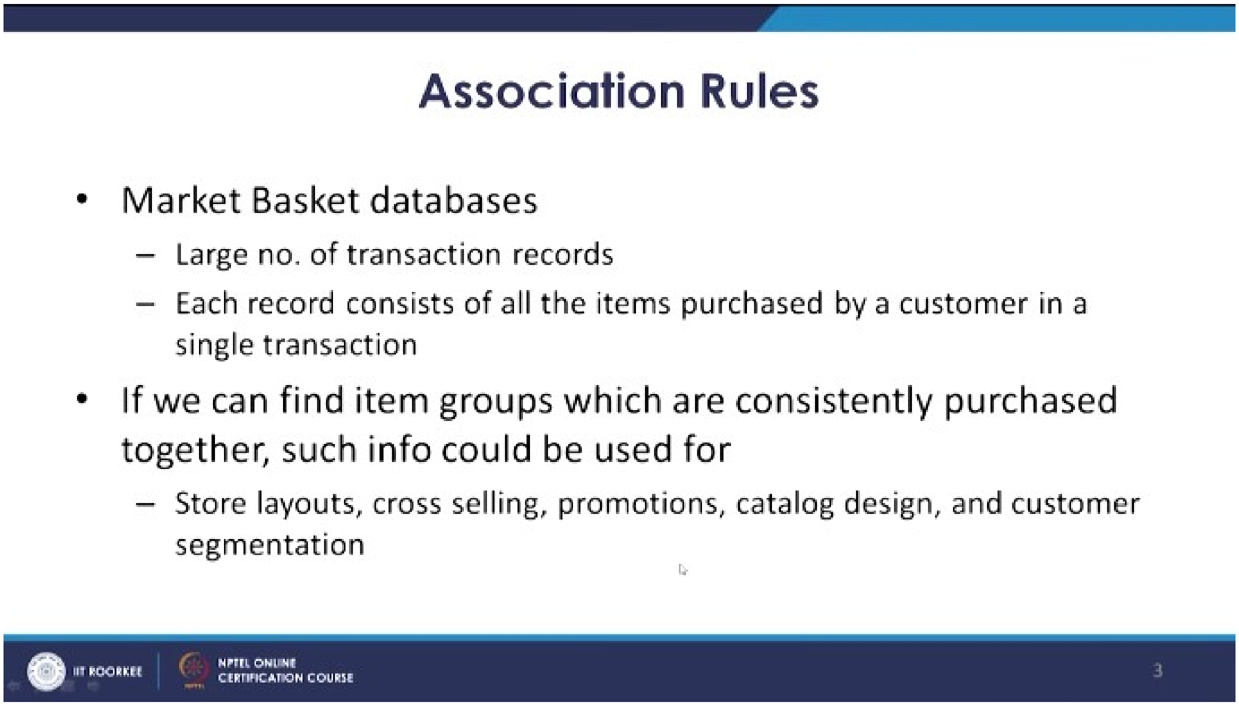
Association Rules

- Also called
  - Affinity Analysis
  - Market Basket Analysis
    - Due to its origin from the studies of customer purchase transactions databases
- Main Idea is
  - To identify item associations in transaction-type databases and
  - Formulate probabilistic association rules for the same
  - "what goes with what"

IIT ROORKEE
NPTEL ONLINE
CERTIFICATION COURSE
2

So let's discuss first technique, first unsupervised techniques that we want to cover, that is the association rules, so association rules this particular technique is also called affinity analysis, it is also called market basket analysis, and the mainly due to its origin from the studies that happened on customer purchased transaction databases, so that actually, so this particular technique is quite popular in the marketing domain, and it is called market basket analysis there.

So let's understand the main idea behind this particular technique association rules, so what we try to achieve in this technique is and this method is we try to identify item associations in transaction type databases, and then we formulate probabilistic association rules for the same, so the main idea is from the, looking at, we look at the transactions and we try to find out what goes with what, so different you know groupings between items association dependency between items and from that we try to formulate certain association rules and try to understand which particular items or purchased together, or happen to occur in the transactions together.

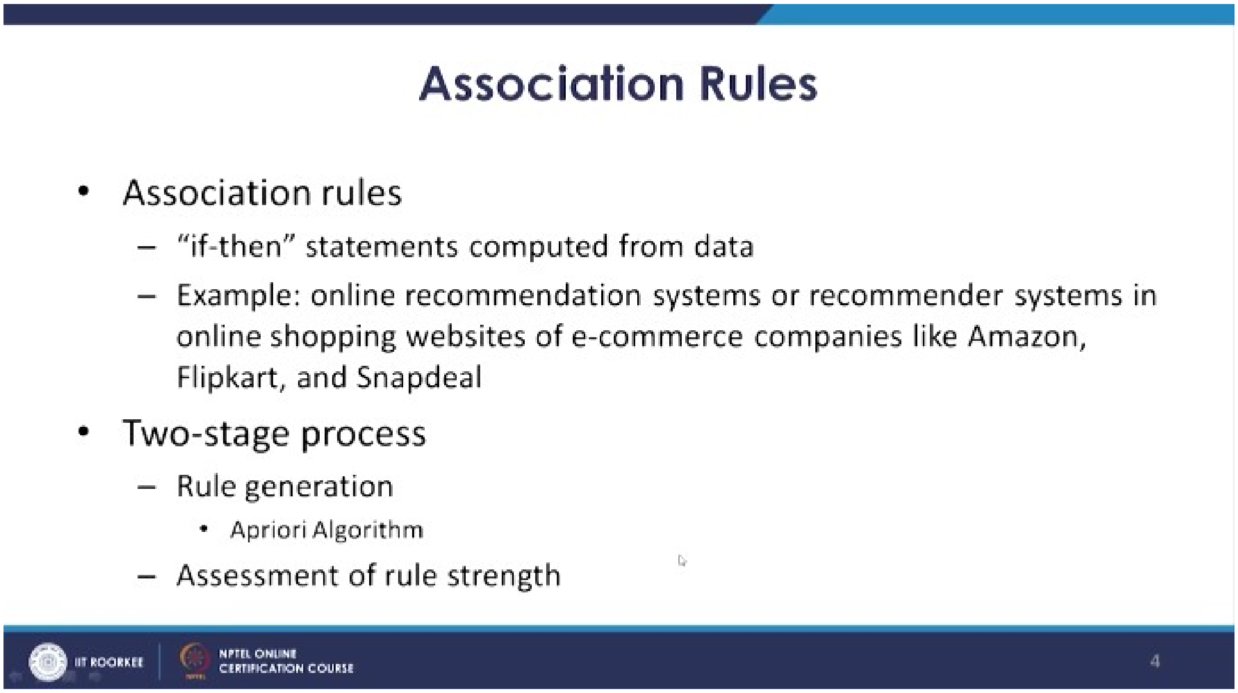


Association Rules

- Market Basket databases
  - Large no. of transaction records
  - Each record consists of all the items purchased by a customer in a single transaction
- If we can find item groups which are consistently purchased together, such info could be used for
  - Store layouts, cross selling, promotions, catalog design, and customer segmentation

So market basket databases they you know typically consists of large number of transaction data records, so any retail store or super market where you see the barcode scanners so you know the different items that particular customer you know purchases, those are actually recorded and it transaction is created, so all those items they you know, in a particular transaction they become a single transaction and when there are number of transaction because of the qualification of this technologies there are number of, millions of records, transaction that are available, so the idea is to analyze these records and to identify association or dependencies between these items, so as you can see market basket databases are about large number of transaction records, and each record consist of all the items purchased by a customer in a single transaction, so any single transaction the number of items that are purchased so that becomes a one record, one transaction and similar transaction from other customers, so those large number of transaction then can be analyzed using this particular technique association rules and we can formulate some of these rules which can later on be used for, as you can see in the second point for store layouts, cross selling, promotions, catalogue design and customer segmentation, so as you can see the second point as we talk about, if we can find item groups which are consistently a purchased together this formation can be used for these purposes, for example store layout as we talk about the, we can design the store layouts in an optimal fashion because we will happen to know that which items are being purchased together and therefore these store layouts can be designed to you know take advantage of that information.

Similarly cross selling, so we'll also know if you know two particular items three, four particular item group is being purchased together we can have create number of cross selling opportunities, that is why you would see many times you know a baskets some items are grouped together when you visit super markets and retailer stores, many items are grouped together, so this is the precisely the idea, the cross selling is the idea, so sometimes they also attached promotional offer, they also all give you some discounts, so how those items are you know selected for this kind of discounts or cross selling is actually based on more often they're not based on association rules.



Association Rules

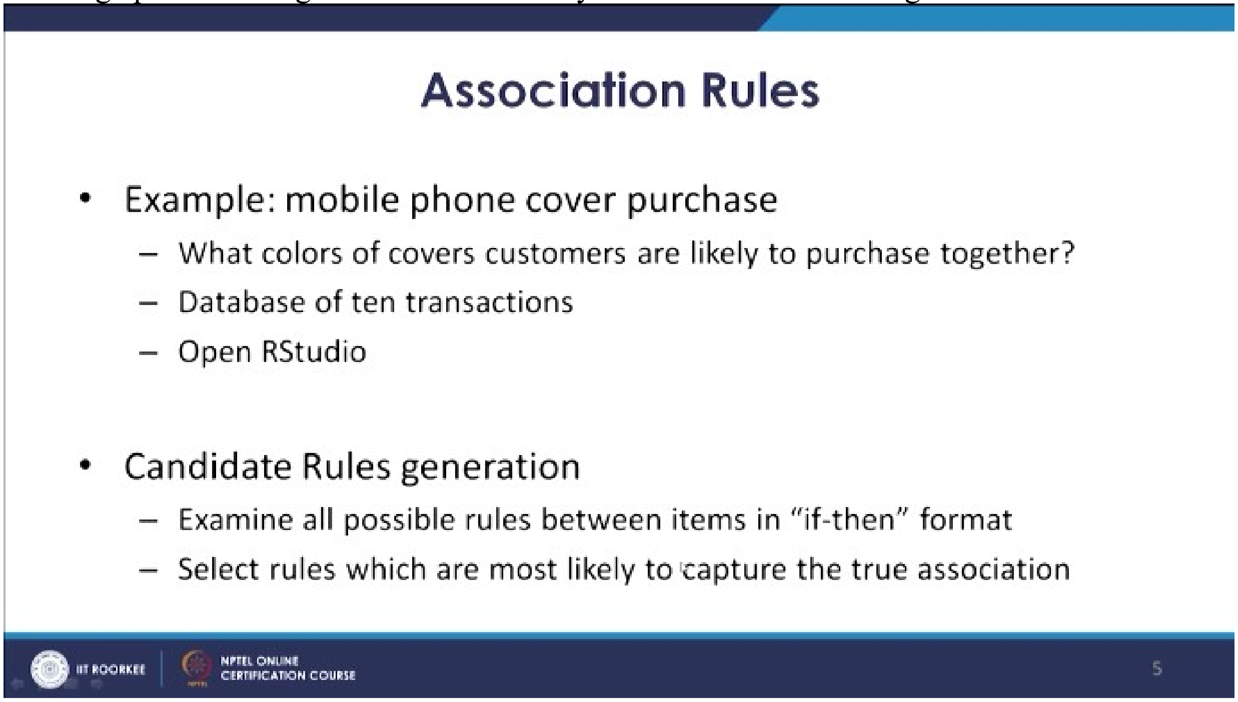
- Association rules
  - “if-then” statements computed from data
  - Example: online recommendation systems or recommender systems in online shopping websites of e-commerce companies like Amazon, Flipkart, and Snapdeal
- Two-stage process
  - Rule generation
    - Apriori Algorithm
  - Assessment of rule strength

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
4

Similarly catalogue design, right, so sometimes you would see in the catalogue design itself you know some items are grouped together you know, those two items are purchased then you know certain discount or certain you know promotion or you know is done, all the catalogue the way items are put together that itself is based on you know information that we get from association rules, so catalogue designing and the customer segmentation, so these associations between items you know they also in a way tell us the buying pattern of customers, so using these buying patterns also we can do our customer segmentation, so you can see a number of areas, number of issues where association rules can really be useful and used, so we talk about the association rules so typically they are you know expressed using if then a statements, so if then a statements are computed from data, so just like in the previous course we talked about classification and decision, classification and regression trees, card, algorithm that we discussed in the previous course, so there also we add certain decision rules, similarly in this case, in this particular technique association rules also we have this, if then kind of a statements, so these statements this being the unsupervised learning methods, these statement, these rules are directly computed from the data that we learned through coming lectures.

So another example for association rules is the online recommendation systems or recommender systems that you see in online shopping websites of e-commerce companies like Amazon, Flipkart, and Snapdeals, so whenever you’re looking for a certain item and you’re examining that particular item you’re on that page, so if you scroll down and you would see that a number of other items are also being searched, so they say you know customer who purchased this item also purchased some of these items, so that conjunction you know for a particular item and this items that conjunction is created on those shopping websites, so that is also based on association rules, so that is one example how, you know, where you can understand how association rules can be used.

So if you look at the association rules you know it’s typically done through a two stage process, so first stage is called rule generation, so in this particular stage we’ll also discuss the apriori algorithm and the second stage is assessment of rule strength, so in the first stage we tried to generate all the possible rule, all the possible combinations and then we assess the strength, so the rules with higher strength are probably going to be selected for implementation, so basically two stage process rule generation followed by assessment of rule strength.

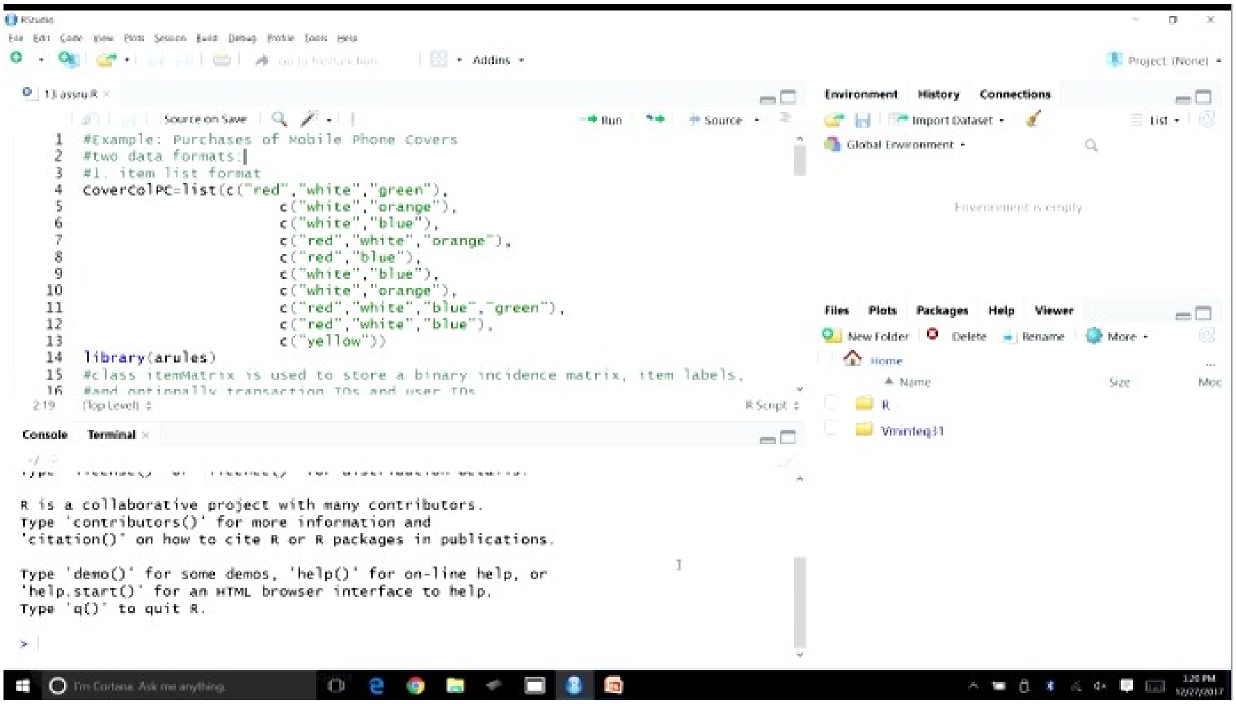


Association Rules

- Example: mobile phone cover purchase
  - What colors of covers customers are likely to purchase together?
  - Database of ten transactions
  - Open RStudio
- Candidate Rules generation
  - Examine all possible rules between items in “if-then” format
  - Select rules which are most likely to capture the true association

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 5

So while we are discussing some concept related to association rules will you know discuss them through this particular example, mobile phone cover purchase, so in this particular example the main idea is that what colors of covers customers are likely to purchase together, so we all have mobile phones and sometimes we also go purchase the cover for the same and then different colors of, different covers with different colors are available, so sometimes customer like to purchase you know different, you know covers with different colors so in this particular example we are trying to, you know we would try to identify the covers which are going to be, which are typically purchased together, so in this case we have this database of 10 transactions, so let’s look at this particular data set, this is one example that we are going to use in our discussion.



```
#Example: Purchases of Mobile Phone Covers
#two data formats:
#1. item list format
CoverColPC=list(c("red","white","green"),
                c("white","orange"),
                c("white","blue"),
                c("red","white","orange"),
                c("red","blue"),
                c("white","blue"),
                c("white","orange"),
                c("red","white","blue","green"),
                c("red","white","blue"),
                c("yellow"))
library(arules)
#class itemMatrix is used to store a binary incidence matrix, item labels,
#and optionally transaction IDs and user IDs
```

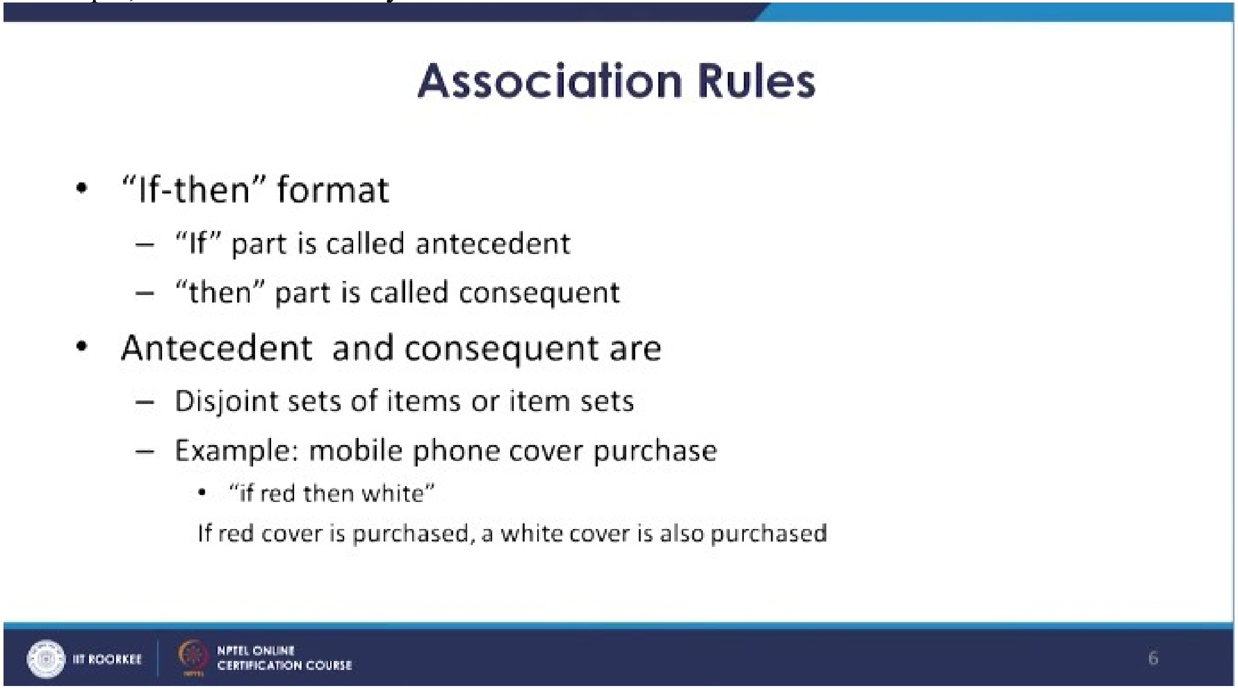
```
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

So as you can see in the example purchase of mobile phone covers, so you can see in the R studio here, so you can see we have 10 transaction in the list you can see there, first one is red, white, green, and the second one is white orange, so each one is, actually one transaction, so you know some customer purchased these three mobile phone covers, color red, white and green, the second transaction where the white and orange you know colored covers were purchased, then the third one is where white and blue, fourth one red, white, orange, then red, blue, so in this fashion we have this, so this is an artificial example, so this particular example that we are going to use for our discussion in this particular technique, for this particular technique, so 10 databases as you can see.



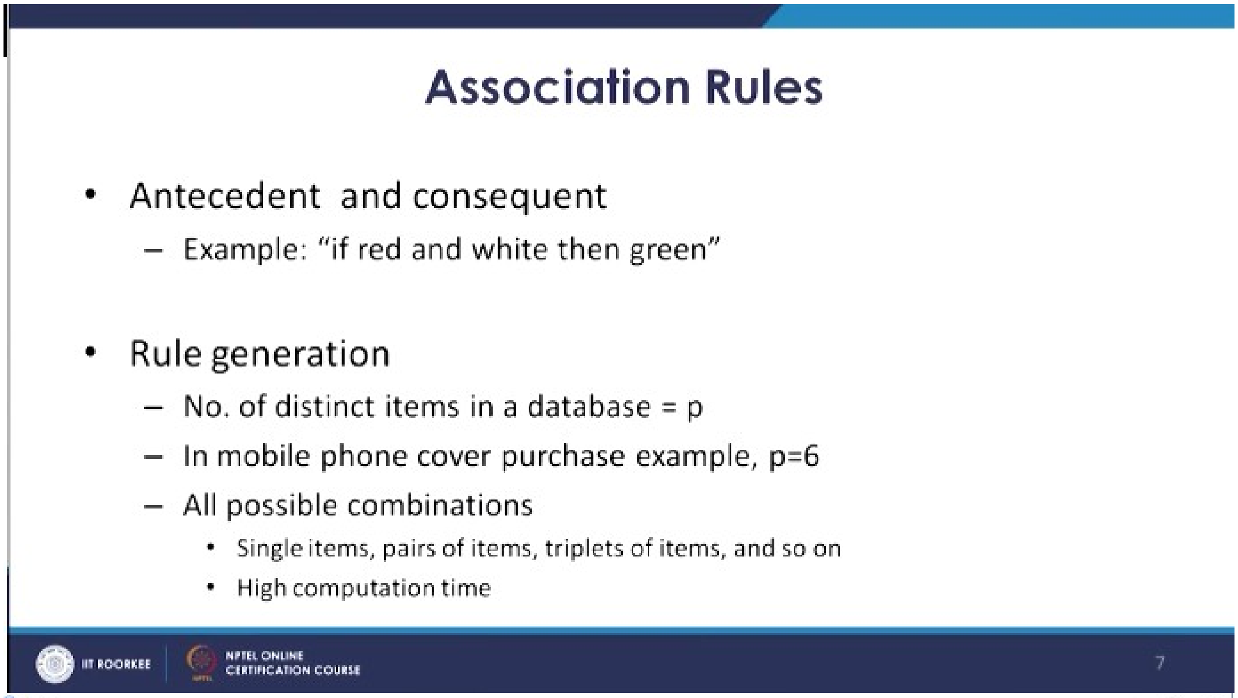
Association Rules

- "If-then" format
  - "If" part is called antecedent
  - "then" part is called consequent
- Antecedent and consequent are
  - Disjoint sets of items or item sets
  - Example: mobile phone cover purchase
    - "if red then white"

    If red cover is purchased, a white cover is also purchased

So we look at the first process in association rules that is candidate rules generation, so in this case we typically examine all possible rules between items in and if then format, so once that happens then we select rules which are most likely to capture the true association, so we don't want to, we want to you know wide random chance occurrences, so we would like to select items which are having some sort of true dependence, some sort of true associations, so that being the idea, so first we need to examine all possible rules, and then later on select the most you know, these rules with higher strength, so let's understand this format, if then format so in this format the if part is also called antecedent and the then part is called consequent, so if you look at these two parts antecedent and consequent, so they are sensibly disjoint sets of items or item sets, so for example if we say that you know if there is a transaction and which involves the mobile phone covers of these 3 colors red, white and green, so if we say that red and white you know we can you know put in antecedent item set, and the green will go into the consequent item set, so as you can see that antecedent and consequent items are they are disjoint that means they have no items in common.

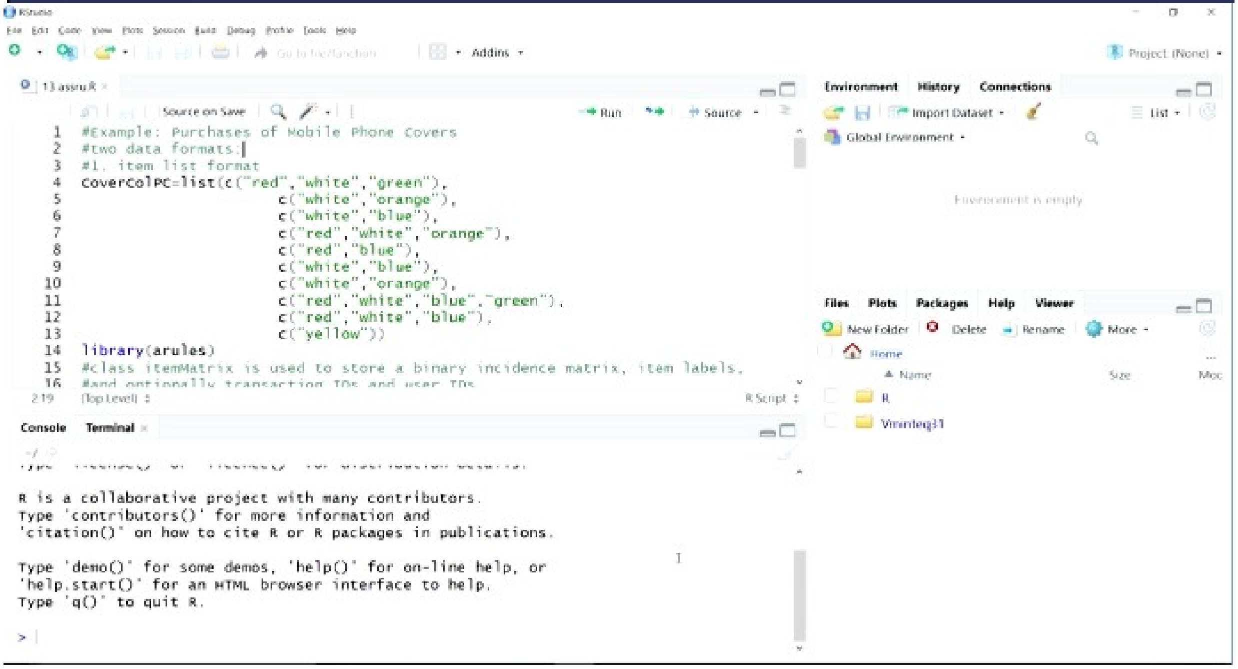
So while we are looking for these rules, we have to make sure that antecedent and consequent if then format that we use to express these rules, so antecedent and consequent item set this would not have any item in common, right, so as you can see here you know through an example so if mobile phone cover purchase example so you can see if red then white, so this essentially means if red cover is purchased a white cover is also purchased, so here the antecedent item set is having just one item that is red cover, and the consequent item set is having just one item that is white cover, so the antecedent and consequent items they are disjoint, they have no items in common, essentially we are trying to say that if a particular item is being purchased and other item is also being purchased.



Association Rules

- Antecedent and consequent
  - Example: "if red and white then green"
- Rule generation
  - No. of distinct items in a database = p
  - In mobile phone cover purchase example, p=6
  - All possible combinations
    - Single items, pairs of items, triplets of items, and so on
    - High computation time

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
7



```r
#Example: Purchases of Mobile Phone Covers
#two data formats:
#1. item list format
CoverColPC=list(c("red","white","green"),
                c("white","orange"),
                c("white","blue"),
                c("red","white","orange"),
                c("red","blue"),
                c("white","blue"),
                c("white","orange"),
                c("red","white","blue","green"),
                c("red","white","blue"),
                c("yellow"))
library(arules)
#class itemMatrix is used to store a binary incidence matrix, item labels,
#and optionally transaction IDs and user IDs
```

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Another example could be if red and white, then green, so that is again if we go back to R studio, so you can see the first one red, white, green so this particular example is can be expressed as if red and white then green, we can also say that if red and green then white, so however we have to make sure that these sets are disjoint, however you know which, later on as we understood that the strength of these rules are to be assessed, so we'll know which a particular rule is having highest strength and probably those rules are going to be selected later on, so one example could be if red and white then green, so in this case as you can see the antecedent item set is having two items that is red and white, and the consequent item set is having just one item that is green.

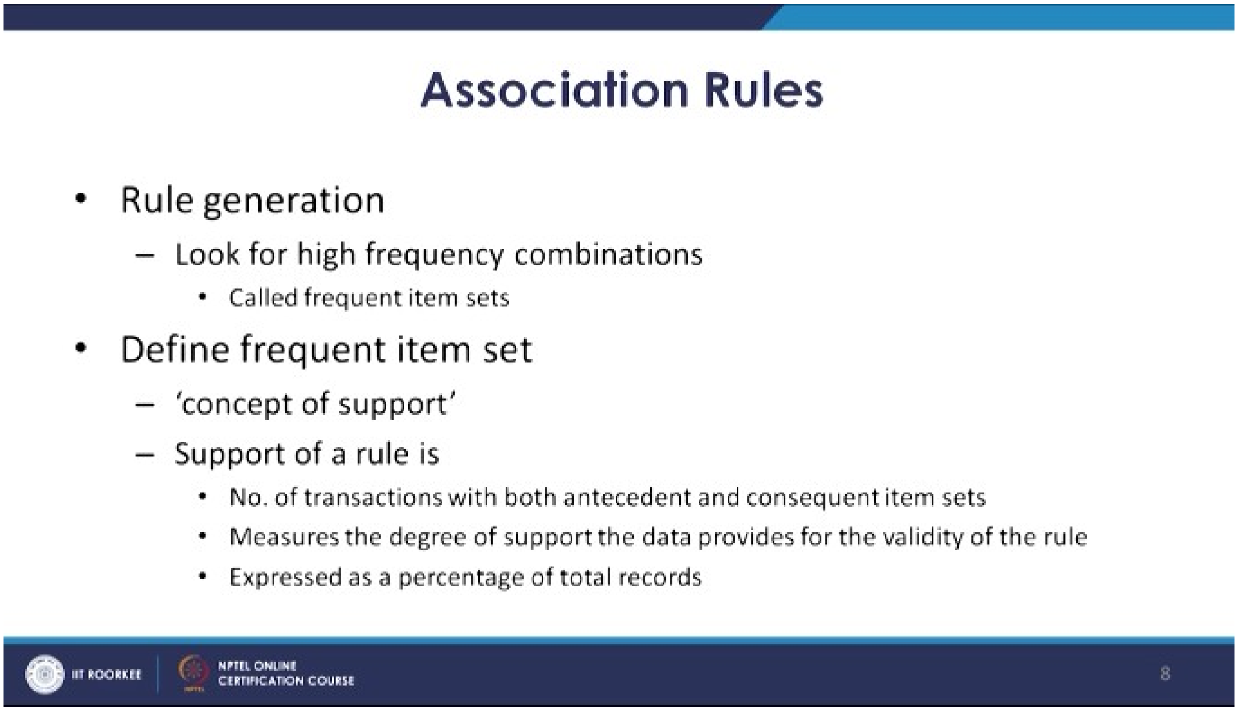
So let’s discuss a bit more on the rule generation process, so if you look at the rule generation process first we need to identify the number of distinct items in a database, for example this mobile phone purchase database that we have, we have 10 transactions, but out of all these 10 transactions you know how many distinct items are there, right, so because many items would be common across you know these transactions, so we need to identify the number of distinct items, so let say if this items are you know number of these items, number of distinct items are P, then you know in this particular case if we go back to R studio, and if we try to identify, try to find out the number of distinct item set, here you can see that there are 6 distinct item set in this case, you can see red, white, green and then orange, blue and then yellow, so last transaction you can see yellow is also there, so these are the 6 color that we have, so number of, so though we have 10 transactions, so number of distinct item set, number of distinct items are 6.

If we think about a database which is having millions of transactions, so you can see it becomes a quite you know, quite a computing, quite a higher you know computing intensive exercise to find out the number of distinct items.

So if we look at you know, once we know that number of distinct items that are there in database, we need to you know find out all possible combinations, so if we look at you know just this example where you know mobile phone cover purchase example, 6 distinct items, so we need to look at, we need to examine single items, pairs of items, triplets of items, and so on, so if we you know understand this process so all those sets you know different number of elements in them, different number of items in them are to be examined, so that would actually require a higher computation time, so therefore we need to find the solution for this, otherwise the computation time would be much higher, and we won’t be able to apply this particular technique.

So what is typically done is, we look for high frequency combinations, so instead of looking for all possible combination as an alternative solution we can look for just you know high frequency combinations, so these combinations are also called frequent item sets, so instead of all possible item sets we can look for these frequent item sets.

Now the next question is going to be how do we define what is a frequent item set, so how this particular you know frequent item set is define, depends on the concept of support, so this concept, the concept or support is used to define what is a frequent item set.



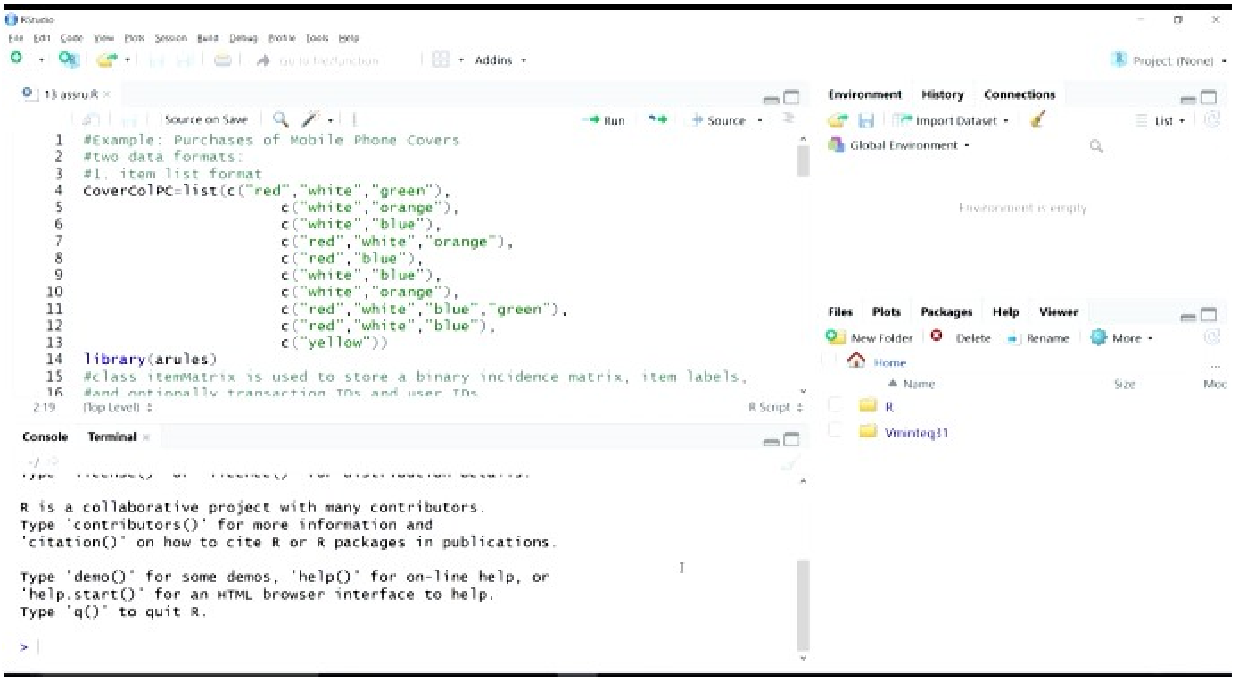
Association Rules

- Rule generation
  - Look for high frequency combinations
    - Called frequent item sets
- Define frequent item set
  - 'concept of support'
  - Support of a rule is
    - No. of transactions with both antecedent and consequent item sets
    - Measures the degree of support the data provides for the validity of the rule
    - Expressed as a percentage of total records

So let's understand what we mean by support, so as you can see support of a rule is, number of transactions with both antecedent and consequent item sets, so we have to see the number of transaction in our databases which are having both antecedent and consequent item set, for example if we go back to the example that we discussed, if red and white then green, so as we can see that red and white are the, red and white you know get the part of the antecedent item set, and green is the part of the consequent item set.

Now we will have to look at the transactions in the database where the antecedent item set that is red and white is present and also the consequent item set that is having just one item green should also be present, so both antecedent item set consisting of red and white, and the consequent item set consisting of just green, both should be present.

So this particular number will give us the idea of support of that particular rule, so if this becomes the rule, if red and white then green, then the support of rule you know, in a sense number of transactions that are you know, number of transaction with both antecedent and consequent item sets, so that will actually give us an indication about, indication about the support that a particular rule enjoys, so another way to measure the support of rule is the degree of support that data provides for the validity of the rule, so this support of a rule also measures the degree of support the data provides for the validity of the rule, so let's say you know our mobile phone, you know mobile phone cover example, there are 10 transaction and out of those 10 transaction, 4 transactions are searched where both the antecedent and consequent item sets are present, then the support of that particular rule is going to be, you know, determine by that number, so typically support of rule is expressed as a percentage of total records, so if we have an example, for example, previous example that we discussed if red and white then green so in that case if you know there are 4 transactions in the database where both the antecedent and consequent item sets are present, then you know out of those, you know there are 4 transaction out of total 10 transaction then the support of this particular rule can be express in percentage term and this is going to be the 40% because of the 4 is 40% of the total number of transaction that is 10, so that would indicate the support that this particular rule enjoys.



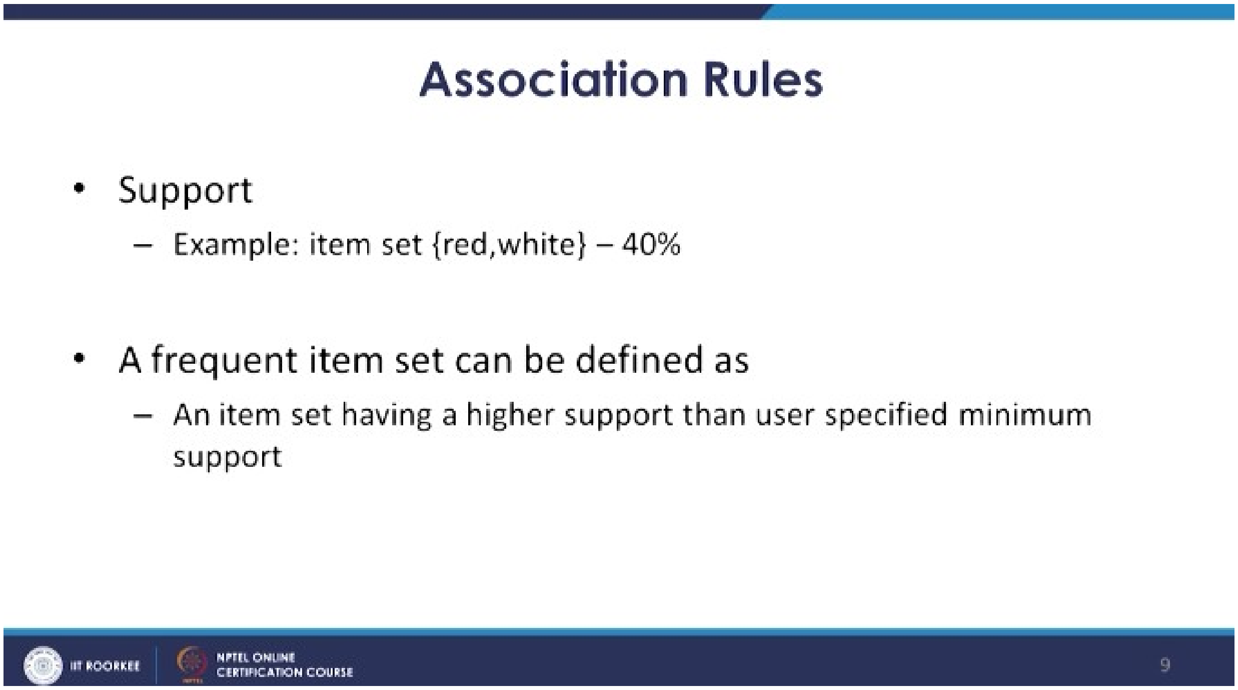
```r
#Example: Purchases of Mobile Phone Covers
#two data formats:
#1. item list format
CoverColPC=list(c("red","white","green"),
                c("white","orange"),
                c("white","blue"),
                c("red","white","orange"),
                c("red","blue"),
                c("white","blue"),
                c("white","orange"),
                c("red","white","blue","green"),
                c("red","white","blue"),
                c("yellow"))
library(arules)
#class itemMatrix is used to store a binary incidence matrix, item labels,
#and optionally transaction IDs and user IDs
```

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

So as we can see here the example item set red and white, so this is 40% if we go back to our R studio and look at the database, so you can see let's find out how many transactions are there, where you know this particular you know red and white both are present, so we can see the first transaction red, white, green, so red and white are present, the second one you know not present, just the white, third one just the white, and the fourth one again red, white, orange, so here also we can see red and white, the fifth one red and blue, so just the blue, so sixth one white and blue again not present, seventh one the white and orange, then we have red, white, blue, green, so here also we can see red and white are present, so this is the third transaction where red and white both are present, then the ninth transaction we can see red, white and blue, so this is the fourth transaction where red and white both are present, then the last transaction is you know yellow, so out of these 10 transaction we can find out, we can easily spot 4



Association Rules

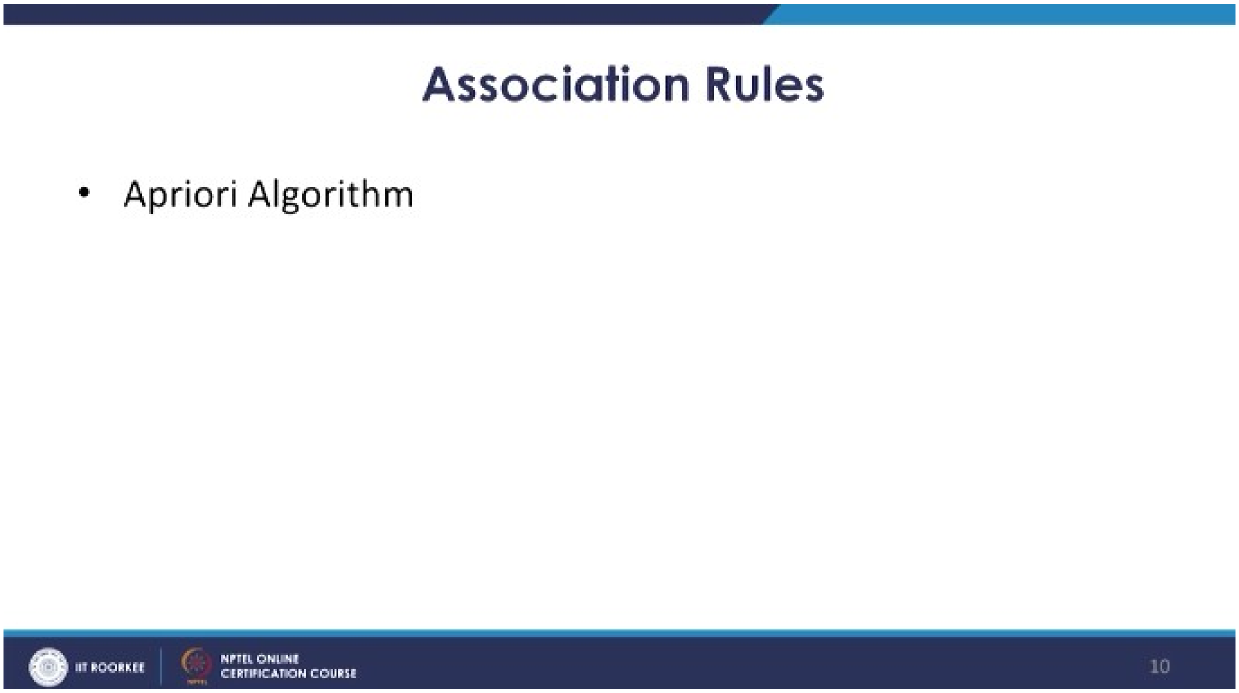
- Support
  - Example: item set {red,white} – 40%
- A frequent item set can be defined as
  - An item set having a higher support than user specified minimum support

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 9

transactions, where both red and white are present. So this gives us the idea about the support of this particular rules, so the item set is red and white, then the support of you know rule that, rule is that we talked about if red colored cover is purchase then white is purchased as well, so support for this particular rule is 40% as we saw in the data set.

So now once we have computed the support of rule, how do we you know again we go back to our previous question, how do we define a frequent item set, so once we are able to compute the support for a particular rule, so if frequent item set can be defined as an item set having a higher support then usual specified minimum support, so user has to specify what is going to be the minimum support level for a particular you know, for a particular item set to be conceded as a frequent item set, so if the support for a particular rule, particular item set is higher than that then that is going to be considered as frequent item set, and these frequent item sets are then going to be you know, used in the rule generation process, so if we go back to our R studio the example, so if we look at the transactions here, so we have total 10 transactions, and let’s say the rule that we considered earlier the red and white, the item set that we considered earlier red and white, so there if the, you know, minimums user specified support is 3, then this particular item set is going to be considered as frequent item set, however if the user specified support is 5, then of course the you know, the support for this is 4 or the 40%, so that comes below that user specified minimal support, so then in that case this particular rule won’t be considered as a you know high frequent item set.

So in this fashion we can also find out you know about the other rule, for example the other example that we discussed red, white and green, so if red and white you know if they become the part of the antecedent item set, and green becomes the part of the consequent item set, and if we look at the transactions here, so we find just two transaction where all these 3 items are present, so first transaction red, white and green, and then we can see 8 transaction where red, white, blue, green, so there are two transactions, so if the minimal support level is 3 then of course this particular item set is going to come below that, so therefore it is not going to be conceded as frequent item set.

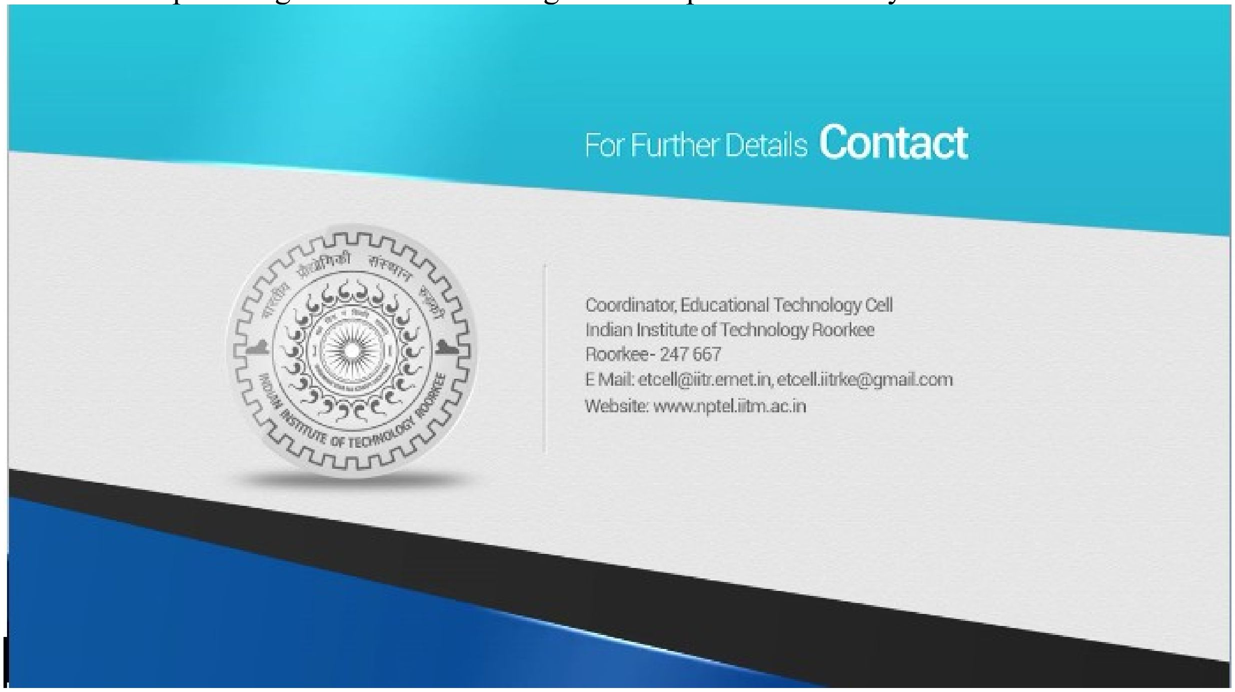


Association Rules

- Apriori Algorithm

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

So once we are able to, we're able to identify a frequent item set given the user specified minimal support level, our next step takes us to the discussion of Apriori algorithm, which is the key part of the rule generation process, so we'll stop here and in the next session we'll discuss the Apriori algorithm and the rule generation process. Thank you.



For Further Details **Contact**

Coordinator, Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee- 247 667
E Mail: etcell@iitr.ernet.in, etcell.iitrke@gmail.com
Website: www.nptel.iitm.ac.in

For Further Details Contact
Coordinator Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee – 247 667
E Mail:-etcell@iitr.ernet.in, iitrke@gmail.com

Website: www.nptel.iitm.ac.in

Acknowledgement

Prof. Ajit Kumar Chaturvedi
Director, IIT Roorkee

NPTEL Coordinator

IIT Roorkee
Prof. B. K Gandhi

Subject Expert

Dr. Gaurav Dixit
Department of Management Studies
IIT Roorkee

Produced by

Mohan Raj.S

Graphics

Binoy V.P

Web Team

Dr. NibeditaBisoyi
Neetesh Kumar
Jitender Kumar
Vivek Kumar
Dharamveer Singh
Gaurav Kumar

An educational Technology cell
IIT Roorkee Production