

Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:10
Lecture:38

TRENDAHUB CASE ON RFM

Now, we are going to see a live case, which is a case of customer segmentation that we are going to do using RFM. So we are going to club the customers as per their RFM scores. Initially we are going to see how to calculate the R,F,M values. So calculation of RFM and R,F,M values can be done through various methods or various sources, like we can calculate using simple Excel calculations or it can be done through SQL queries or it can be done through programming using Python or R etc.

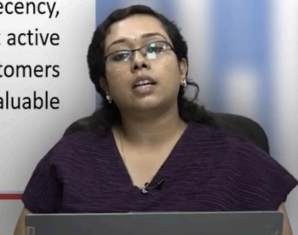
So basically what we are doing is, we are extracting the recency score, frequency score, monetary score of each customer. Then we are scaling it because that is how, you know if we are giving to a clustering algorithm to cluster the customers, then it might, it will perform very efficiently only if it is scaled properly. So R, F and M should have similar scales in order to you know perform efficiently if we are clustering the customers. So initially what we are doing is, after calculating the R, F and M values, we will be scaling the R,F and M values and then we will be finding the RFM values which is nothing but R multiplied by F multiplied by M if we are not giving any weights to R,F and M. But it depends on business, if they want to give weights or does not want to give weights.

TRENDAHUB: A Case Of Customer Segmentation using RFM

A retail company called "TrendHub" wants to conduct an RFM analysis based on their raw data. The dataset includes information about various aspects of their sales transactions, such as product details, customer information, pricing, and logistics. TrendHub operates several stores, each identified by its IssuingPlantCode. They sell a variety of branded products (Brand) across different styles, colors, and sizes. The transactions include sales (TransType) with transaction prices (TranPrice) and quantities (TrnQty). Each sale is associated with a timestamp (TrnTime). Additionally, tax details, discounts, and transporter information are included.

The company is interested in RFM analysis to categorize its customers based on Recency, Frequency, and Monetary value. They want to understand which customers are the most active and valuable, enabling them to tailor marketing strategies effectively, reward loyal customers (Loyalty Points), and improve their overall sales performance. This analysis will provide valuable insights for TrendHub's future business decisions and customer engagement strategies.

BUSINESS INTELLIGENCE & ANALYTICS




Usually business might give more privilege or more importance to R values than F values then least to monetary values. But this need not be the case with everyone because some business might be needing the monetary value which might be increasing the profits to the business. So it might be giving a more weightage to M value as well. So that is, in this case we are not giving weightages but it is up to you if you want to give or not. So I will read out the case, so that we can get into the calculations after this.

So a retail company called TrendHub wants to conduct an RFM analysis based on their raw data. So initially, we have a raw data which might have missing values, which might have null values, which might be having very absurd values which we have to clean and preprocess before we move into the actual RFM analysis. So the data set contains information about various aspects of the sales transactions such as product details, customer information, pricing and logistics. So it has various aspects about how the customer has done the sales transactions. So it has the customer ID, the products that they have purchased, if it was an actual sale or if the customer has returned the product and the price of the product and the date on which it was ordered, everything that is what the raw data that we have.

So TrendHub operates several stores, each identified by its issue plan code. They sell a variety of branded products. So we also have the brand value, brand name across the raw data. So they sell a variety of branded products across different styles, colors, sizes. The transaction includes sales, that is the transaction type and the transaction prices and the quantities.

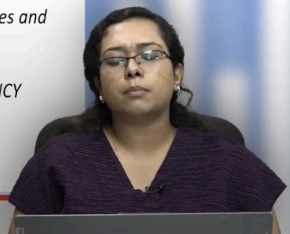
Each sale is associated with a time stamp, that is the time in which the transaction was done. Additionally, there are many details like tax details, discounts, the transporter information, etc. But do we need all these for RFM analysis? Not at all. We just need three data. So you have to imagine which three data are we going to use for RFM.

So the company is interested in RFM analysis to categorize its customers based on recency, frequency and monetary value. They want to understand which customers are the most active and valuable, enabling them to tailor marketing strategies effectively, reward loyal customers with loyalty points and also improve their overall sales performance. This analysis will provide valuable insights for TrendHub's future business decisions and customer engagement strategies. So what we are having with us is a raw data that has to be cleaned, preprocessed and then it will be used for analysis of RFM. Once RFM is done, we will also see how we cluster these customer segments using K-Means clustering that you have already learned, so that we can see how customers, various kinds of customer segments are allotted to different clusters and we will see what cluster statistics are there and then we will try to match these clusters with whatever segments that we have studied before this.



DATA CLEANING AND DATA PREPARATION

- Treating the outliers
 - Transaction price can never be negative → removed all the records with a negative Transaction price.
 - Transaction Quantity can never be negative → removed all the rows with a negative Transaction Quantity.
- Returns are considered as negative values since it has to be subtracted if the customer returns the product
- Pivot tables are applied in excel and the following are obtained:
 - *Sum of total value gives the total monetary transactions performed by a particular custom (both sales and returns)-MONETARY*
 - *Count of transactions quantity gives the total number of visits the of customer to the store- FREQUENCY*
 - *Days since last order(Current date-most recent transaction) -RECENCY*



BUSINESS INTELLIGENCE & ANALYTICS

So the first thing that we do whenever we get a raw data is to identify the columns that are needed for the analysis that we will be doing. So there are so many columns that are not needed. For example, the brand name, the item type or the transport mode, that are not needed for our RFM analysis. So we need to select the columns that are actually going to be used in our analysis. So the first thing that we will be using is the transaction

price.

So transaction price is the value that the customer has spent for purchasing a product. So can price be negative? Price can never be negative, right. So what we do is whatever data that has negative transaction price needs to be removed. So if it is a simple Excel, then you can just use the filter option to filter out all the values that have negative sign in it so that we can select those transactions whose price are above 0. The next one of next step of treating the outliers is the transaction quantity.

So we have to multiply the price and quantity to get the overall monetary value of a purchase. So the transaction quantity also can never be negative. So we have to remove all the rows which are having negative transaction quantity as well. So if we see the transaction type, there is also a column. I will show you the raw data.

So we have a column which is called the transaction type in which there are two types. One is sale and one is return. So if you are buying anything from say, Flipkart or Amazon, what you will do is, once you purchase something, it comes under the sales data. But if you are not happy with the product or if the product has come broken or something, then you will go and return the product, right. So in turn, you have made 0 monetary value there, if you have returned the product.

So what we have to do is, we have to assign negative values wherever the transaction type is returned. So that in total, it will come to the actual monetary purchase that the customer has made. So in this example that I am going to show you, I am using Excel analysis to show you how the RFM calculations are done and especially I am using the pivot tables so that, you know for efficient analysis and quick analysis. So pivot tables, it will give you the sum, count, all those kinds of function in just one click. So if you do not know how to work with pivot tables, I urge you to go and read on how pivot tables can be used for easier calculations.

It need not always be the case that you use Excel for RFM analysis especially if the data set is huge and it is very complicated, then you can go for other techniques like, you know SQL also. So both Excel as well as SQL has predefined queries that will help you group these customers into quintiles or 5 groups or you know deciles like 10 groups depending on how many groups you want to segment the customers in.

So we have one column which is known as the total value column. Total value is the total monetary value of purchases that a customer has made during a fixed time period. That can be the entire duration past, entire life cycle of his purchase or it can be a prefix number of years like 5 years.

So sum of total value gives us all the sales value in plus sign and return values in negative sign. So it counts both, adds setup and gives the total monetary value. So that is what we are assigning as monetary.

And the next one is frequency. So how do we get what the frequency value is? We will count the number of transactions that the customer has undergone over the past. So count of transaction quantity gives the total number of visits the customer has come to the store.

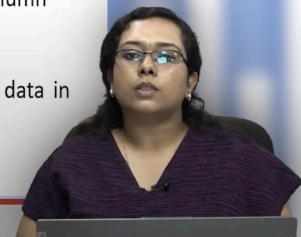
Then the last one is recency value and how can we calculate the recency? There might be many dates on which the customer has come to the store, right. So we have to take the max of all the dates of a particular customer. Say a customer has come before 3 years. That means we have to calculate today's date minus that max of order date so that we will get the number of day, we will get the difference in terms of number of days.

So that number of days is what we call as recency because he has come before, say it is 5000. That means he has come before 5000 days. So that was the most recent transaction that the customer has done. So that is how we get the recency value. So as of now we have discussed how we get the monetary, frequency and recency values.

But when you think all these will be in various scales, different scales and next step is to group and scale them. It can be from 1 to 5 or 1 to 10, 1 being the least and 10 being the most or 1 being the least and 5 being the highest value the R, F and M can take. So we will go into how we do that.

GROUPING AND SCALING

- Since R,F,M values have to be scaled for better efficiency during clustering and ease of comparative analysis, we scale it into values from 1-10
- Recency value obtained by grouping *Days since last order* into 10 deciles and then scaling them from 1-10
- Reverse scaling done for Recency since the most recent customer needs to be given the highest R value
- Frequency values have been obtained by grouping *Count of txn qty* into deciles and then scaling it to 1-10
- Monetary values have been obtained by grouping and scaling *Sum of total value* column
- $RFM=R * F * M$
- Next step is customer segmentation which will be done with this preprocessed data in python



So the next topic is grouping and scaling. So since now we have R, F and M values but those are not scaled similar to each other, they have to be scaled for better efficiency especially when we give it as input for K means clustering or different kinds of clustering algorithms.

So also it would give us a sense of ease when we compare between R, F and M just by directly looking into it. So in this case, we are scaling it from 1 to 10 but if you want you can scale into 1 to 5 as well. So how do we get the recency? We get the recency by grouping days since last order into 10 deciles and then scaling from 1 to 10. So days since last order, how do we get it? That is the max of the order date minus today's date or today's date minus maximum of the order date or the most recent order date that a customer has. So we will get the difference in terms of number of days.


So if it is 5000, then the 5000 has to be scaled from 1 to 5 depending on how many records we have. Let us take two customers. One customer has made the most recent purchase before 3 years. Another customer has made the most recent purchase before 1 year. So if we subtract the max of the order date from today's date then which customer has the highest value? The customer who has come before 3 years because say, it is 800 days that is the difference between today's date and his last order date.

Another customer just came before 1 year. So his recency might be 300. But actually recency should be high for the customer who has been more recent. So the value 5 should be given to the customer who has been recent and not the customer who has come before 3 years. So what we have to do is, we have to do reverse scaling only for recency because frequency and monetary value will not need that reverse scaling.

So reverse scaling has to be done so that the most recent customer gets the highest R value. So that we can do in Excel itself and then we go to the frequency value. So frequency value can be obtained by grouping, count of transaction quantity. So count of transaction quantity is nothing but the number of transaction that customer has done over the past lifetime. And then we similarly scale the frequency value also into buckets of 1 to 10.

Then the last one is the monetary value which can be obtained by grouping and scaling sum of total value. So I have already told you how to calculate the sum of total value, keeping in mind if it is a sales transaction or a return transaction and then we get the sum of total value column which can be grouped into 10 buckets and then scaled from 1 to 10. So the final thing that we are going to do is calculation of RFM which is nothing but R into F into M unless you want to give a weight to any of the 3 matrices, like recency you want to give 3 or frequency you want to give 2 and monetary 1, then you can assign

the weights as well. So once we have got the RFM value of each customer, what we have to do is, we have to know how the customer segmentation happens. So the customer segmentation or the customer clustering can be done using one of the clustering techniques.



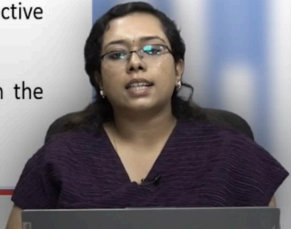
NTILE FUNCTION IN SQL

- The SQL NTILE function is a valuable tool for RFM (Recency, Frequency, Monetary) analysis, aiding in segmenting and ranking customers into distinct groups based on their behavior.
- This function divides a dataset into equal parts, creating quantiles or percentiles, which can be very useful for categorizing customers according to their RFM scores.

```
CREATE TABLE RFM AS SELECT customerid, recencydays, NTILE(5) OVER (ORDER BY recencydays) AS Recency FROM Transactions;
```

- The NTILE function assigns a numerical ranking to customers within their respective segments, reflecting their RFM scores in comparison to others.
 - A customer in the "top 20%" segment will have a higher NTILE rank than one in the "bottom 20%."

BUSINESS INTELLIGENCE & ANALYTICS



So we have learnt how to do it in K-means clustering. So once we have got the R, F and M values, what we are going to do is we have to feed them into the K-means clustering algorithm, which will cluster it automatically into how many ever number of clusters that you are choosing in it. So we will be going into that in the next session. So we saw that in Excel, we can group or bucket the customers into 10 or 5 groups. So there are commands in Excel that help you do that. We will be going into that shortly. But can we do this in SQL as well? Yes, the answer is yes. That is the ntile function helps you to group or segment the customers into distinct groups based on their behavior. This function divides a complete data set into equal parts, say 5 equal parts or 10 equal parts creating quantiles or percentiles. So, which will be useful for categorizing the customers according to the RFM score

So the code or the syntax for the ntile function is create table RFM. So what are we doing here? We are creating a new table RFM which will be selecting the values or selecting the columns which are customer ID, recency days. So we are selecting 2 columns which is, one is customer ID obviously and the recency days is nothing but the days, we have got the value in the form of days right by subtracting the max of order value from the today's date. So we have got the recency in terms of days, that is what we are selecting here and then we are ntile, using ntile function to group it into 5 percentiles or 5 segments, that is why we have given 5 over order by recency days. So we are ordering

it, you know in ascending order by recency days, as recency. So we are giving the alias as a recency value from the transaction table.

So transaction table is the raw data that we have. So we are giving the new name as recency, for the value that we are going to calculate. So how are we getting that? By grouping the entire recency days value into 5 buckets and whichever buckets it is falling into, say first bucket then it is given the recency as 1. If it is falling in the third bucket then the recency value will be 3 and similarly. So it will be having the recency value just, only within 1 to 5 values 1, 2, 3, 4 and 5. So similar function is there in Excel as well, we will see that.

So in the ntile function, the ntile function assigns a numerical ranking to the customers within their respective segments reflecting their RFM scores in comparison to the others. So it will fall from 1 to 5. So the customer in the top 20 segment, that is the customer who has the 5 recency value, have a higher ntile rank than the bottom 20. So the highest 20 percent will be having 5 as the recency, the lowest 20 percent will be having 1 as the recency value. So that is the use of ntile function in SQL. So we will go into the analysis of how we do, how we perform all this with the raw data that we are having with us.

So this is just a screenshot of an example. This is not the raw data. This is just an example to show you how we group it. So initially we have the customer ID, then we have the most recent order date. So that is subtracted from the current date, in order to get the days since last order, that is the third column. So the person who has 288 in the first row in the days since last order, is he more recent or less recent? He is less recent than the second guy who is having 67, because before 67 days he has made a purchase. The first guy has made a purchase before 288 days. But if you consider this as the recency, then what happens is, we are giving higher value to the guy who has come very long back. So what we have to do here is, we have to reverse scale the recency values.

So what we do is, the 288, the guy having value as 288 should be given less priority than the guy who has 67 as the recency value. So what we do is, we reverse scale it. Here it is given as recency value 7 for the first guy and 1 for the next guy. So this is wrong and this should be interchanged in the reverse scaling process that we will see when we are dealing with the raw data.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		
	Barcode	Brand	StyleCode	Color	Size	Trans Typ	TranPrice	TrnQty	TrnTime	Issuingplc	Document	DocumentDa	Tax Perce	Tax Value	Disc.Amount	Disc Per	Transport	TransportWay	Billnc	AdidasDa	SI_no	Tran	
1	0912090702	NIKE	312083-061	NA	7	Sale	2030	1	07:55	2030	1	14-Nov-06	12.5	225.56	145	6.67						1	Regula
2	0912090702	NIKE	312083-061	NA	7	Sale	2030	1	07:55	2030	1	14-Nov-06	12.5	225.56	145	6.67						1	Regula
3	0912090702	NIKE	312083-061	NA	7	Sale	2030	1	07:55	2030	1	14-Nov-06	12.5	225.56	145	6.67						1	Regula
4	8902358307	MTV	MM0997	D	L	Return	489	1	04:39	489	1	30-Nov-06	4	18.81	0	0						1	Revers
5	8902358307	MTV	MM0997	D	L	Return	489	1	04:39	489	1	30-Nov-06	4	18.81	0	0						1	Revers
6	8902358307	MTV	MM1229	C	M	Return	594	1	04:39	594	1	30-Nov-06	4	22.85	0	0						2	Revers
7	8902358307	MTV	MM1229	C	M	Return	594	1	04:39	594	1	30-Nov-06	4	22.85	0	0						2	Revers
8	8262203015	NIKE	593423-101	NA	M	Return	80.5	2	04:39	161	1	30-Nov-06	4	3.1	23	12.5						3	Revers
9	8262203015	NIKE	593423-101	NA	M	Return	80.5	2	04:39	161	1	30-Nov-06	4	3.1	23	12.5						3	Revers
10	8902565640	REEBOOK	C045F001-OW	NA	L	Return	699	1	04:39	699	1	30-Nov-06	4	26.88	0	0						4	Revers
11	8902565640	REEBOOK	C045F001-OW	NA	L	Return	699	1	04:39	699	1	30-Nov-06	4	26.88	0	0						4	Revers
12	9700000158	LEVIS	RL0700	NA	42	Sale	700	1	07:59	700	2	14-Nov-06	4	26.92	0	0						1	Regula
13	9700000158	LEVIS	RL0700	NA	42	Sale	700	1	07:59	700	2	14-Nov-06	4	26.92	0	0						1	Regula
14	890256731	REEBOOK	1-143044	NA	09	Return	1490	1	05:51	1490	2	10-Dec-06	12.5	165.56	0	0						1	Revers
15	890256731	REEBOOK	1-143044	NA	09	Return	1490	1	05:51	1490	2	10-Dec-06	12.5	165.56	0	0						1	Revers
16	8262203013	NIKE	593422-100	NA	L	Sale	80.5	2	08:02	161	3	14-Nov-06	4	3.1	23	12.5						1	Regula
17	8262203013	NIKE	593422-100	NA	L	Sale	80.5	2	08:02	161	3	14-Nov-06	4	3.1	23	12.5						1	Regula
18	0912089942	NIKE	SX1165-102	NA	M	Sale	115.5	1	08:02	115.5	3	14-Nov-06	4	4.44	16.5	12.5						2	Regula
19	0912089942	NIKE	SX1165-102	NA	M	Sale	115.5	1	08:02	115.5	3	14-Nov-06	4	4.44	16.5	12.5						2	Regula
20	8902358358	MTV	MD0223	W	L	Return	0	1	04:03	0	3	30-Dec-06	4	0	299	100						5	
21	8902358358	MTV	MD0223	W	L	Return	0	1	04:03	0	3	30-Dec-06	4	0	299	100						5	
22	8902358393	MTV	MMB11417	B	L	Return	1699	1	04:03	1699	3	30-Dec-06	4	65.35	0	0						5	
23	8902358393	MTV	MMB11417	B	L	Return	1699	1	04:03	1699	3	30-Dec-06	4	65.35	0	0						5	
24	8902358342	MTV	MW1383	R	M	Sale	349	1	08:25	349	4	14-Nov-06	4	13.42	0	0						0	
25	8902358342	MTV	MW1383	R	M	Sale	349	1	08:25	349	4	14-Nov-06	4	13.42	0	0						0	
26	8902706056	DOCKERS	DS01907027	NA	FR	Return	99	3	07:14	297	4	04-Jan-07	4	3.81	30	9.17						3	
27	8902706056	DOCKERS	DS01907027	NA	FR	Return	99	3	07:14	297	4	04-Jan-07	4	3.81	30	9.17						3	
28	0912089920	NIKE	SX1159-103	NA	L	Return	87.5	1	07:14	87.5	4	04-Jan-07	4	3.37	12.5	12.5						3	
29	0912089920	NIKE	SX1159-103	NA	L	Return	87.5	1	07:14	87.5	4	04-Jan-07	4	3.37	12.5	12.5						3	

This is the raw data and for the TrendHub company. So there are so many features that they have given us, like starting from issuing plant code. So there might be a code for each plant, then there is a barcode, there is a brand name, style, color, size, transaction type. So we need not need all the columns for our analysis but only need certain columns. So I will just highlight which all columns we need for our analysis. Do we need the transaction type? Yes, we need the transaction type because we need to know if the person has made a sale or if he has actually returned the product.

So if his sale value is 2030, for same customer id and he has returned products worth 489, then you need to subtract the 489 from 2030 to get the total value of a particular customer. So we need this transaction type, we need this transaction price, we need the transaction quantity so that we can multiply by transaction price to get the total monetary value. Do we need the transaction time? We do not need it because date is good enough. So we have this document date which is nothing but order date. So this column is needed in order to calculate the recency value.

So this transaction type, price and transaction quantity, all these are needed for calculating the monetary value. The document date is needed for calculating the recency value. And then we have tax percentage, tax value, discount amount, discount percentage, transport all these are not needed for our analysis. So that is all. So we just select all these columns that I have highlighted in order to calculate the R, F and M values.

So if you take the count of transaction quantities for a particular customer ID, then that

is what is known as frequency. If you subtract the return values using negative sign and then add the sale and then multiply it by transaction quantity, you get the total monetary value and also recency you can get from the max of the document order date. So I have cleaned all these data.

In the next sheet, what I have done is I have assigned negative value to all the return transactions that we have and I have assigned positive value to all the sales transaction that I have. But before that what I told you is, we have to treat the outliers.

So how do we treat the outliers? I already told you that we cannot have negative values for both transaction price as well as transaction quantity. But if you select this transaction price and click filter, then you will come to know that there are negative values as well. As you can see -3, -2 and -1. So there are values that are absurd in this raw data. So it happens when you get a raw data that there are absurd values or outlier values which you have to delete after the data processing step.

TransType	TranPrice	TrnQty	Total Monetary Value	OrderDate	CustomerID	
Sale	2030	1	2030	2030	14-11-2006	11426
Sale	2030	1	2030	2030	15-11-2006	11426
Return	1699	1	1699	-1699	30-12-2006	130139
Return	1699	1	1699	-1699	30-12-2006	130139
Return	99	3	297	-297	04-01-2007	130127
Return	99	3	297	-297	04-01-2007	130127
Return	87.5	1	87.5	-87.5	04-01-2007	130127
Return	87.5	1	87.5	-87.5	04-01-2007	130127
Return	2417.58	1	2417.58	-2417.58	04-01-2007	130127
Return	2417.58	1	2417.58	-2417.58	04-01-2007	130127
Return	350	1	350	-350	04-01-2007	130127
Return	350	1	350	-350	04-01-2007	130127
Return	594	1	594	-594	13-01-2007	130152
Return	594	1	594	-594	13-01-2007	130152
Return	1575	1	1575	-1575	13-01-2007	130152
Return	1575	1	1575	-1575	13-01-2007	130152
Return	1260	1	1260	-1260	13-01-2007	130152
Return	1260	1	1260	-1260	13-01-2007	130152
Return	1199	1	1199	-1199	27-01-2007	104912
Return	1199	1	1199	-1199	27-01-2007	104912
Return	1263.74	1	1263.74	-1263.74	27-01-2007	104912
Return	1263.74	1	1263.74	-1263.74	27-01-2007	104912
Return	1393.18	1	1393.18	-1393.18	27-01-2007	104912
Return	1393.18	1	1393.18	-1393.18	27-01-2007	104912
Sale	1312.5	1	1312.5	1312.5	15-11-2006	38520
Sale	1312.5	1	1312.5	1312.5	15-11-2006	38520
Sale	1393	1	1393	1393	15-11-2006	38520
Sale	1393	1	1393	1393	15-11-2006	38520

1. columns unnecessary for analysis are deleted
 2. Negative values in transaction quantity and transaction price are filtered out since it can't be
 3. Returns are considered as negative values since it has to be subtracted if the customer returns

BUSINESS INTELLIGENCE & ANALYTICS

So even in transaction price there are lots of negative values. So all these values we have to filter out before we go forward with our analysis. So that is the first two steps that I have taken. So after data cleaning what I have got is, I have removed all the transaction price and transaction quantity which were negative. So if you now click the filter and see all are positive values, as you can see all are positive values, even transaction quantity all are positive values. So I have filtered out all the negative values. I have assigned negative sign for the rows that have transaction type as return. So sales are positive. As just see this column E, in this we have sales which are positive values

and return which are negative values. So that all the returns get subtracted for a particular customer ID from the total monetary value.

Then we have the column which is order date. Order date is nothing but the date on which the particular transaction was done and the customer ID. So we have filtered out pretty much the very outliers that were there and also we have assigned this negative to returns and for ease of calculation. So the final filtered data would look like this, in which we have the customer ID and we have here calculated the total value. The total value is nothing but the transaction price into transaction quantity. So we are multiplying quantity and price in order to get the total monetary value of the transaction.

So we have a huge data and this is a preprocess data on which we will be applying the pivot tables. So you need to have knowledge on how to use pivot tables because we are using only pivot tables for RFM calculation here. So it is pretty easy. So you can go into insert and then select pivot tables and you can select which part of transaction or which part of rows that you need for calculation. So I have selected the entire data set which has 11365 data and we are feeding into the pivot tables.

The screenshot displays the Microsoft Excel interface with the PivotTable Fields task pane open on the right. The task pane shows a list of fields to be added to the report: TransType, TranPrice, TrnQty, Total Value, DocumentDate, and CustomerID. The 'Rows' and 'Columns' sections are currently empty. A woman is visible in the bottom right corner of the screenshot, gesturing with her hands. The NPTEL logo is in the top right corner. At the bottom of the screenshot, there is a red banner with the text 'BUSINESS INTELLIGENCE & ANALYTICS'.

So this is how we get the pivot table and in this we can select one particular feature as row. So how do we want to segment this entire data, based on which value? Obviously it is the customer ID, that is the primary key even if you are using it in Excel. So that will be the primary key on which you want the customer data. For example, a customer ID called 12 might have done 575 transaction over the entire life cycle. So do we need 575 times the data or clubbed into 1? Obviously we have to club it.

So we are using the customer ID as the row label. So we just have to drag and drop it in the row label and then whatever value we want, like the total value we want, right. So we can just put it here as the sum of the total value. So the functions that the pivot table can perform can be either sum or it can obtain the max or it can obtain the count. So it can do many such functions. For example, I am here changing the sum of transaction quantity into, we need actually the count of transaction quantity, right, to know the frequency we need the count of transaction quantity.

So I am changing the sum into count for transaction quantity alone. So this column is nothing but this is the frequency. This column is nothing but monetary value. Similarly you can calculate for recency as well. But recency needs to be reverse scaled as well.

So I have already done the pivot table. So this is the pivot table which I have already done and kept. So I showed you how to do. You can practice on your own.

So we have customer ID. So these are unique values. So till now in raw data, we might have had 100 customer transactions for a particular ID. But using this pivot table we have classified this customer ID uniquely, so that you know there are no repetitions or redundancy. So we have removed the redundancy now. So even in SQL, there are commands that you can use to filter out the customer IDs and print out all these informations.

The screenshot displays an Excel PivotTable with the following data:

Cust id	TOTAL MONETARY VALUE	COUNT OF TXN	MAX OF ORDER DATE	DAYS SINCE LAST ORDER
10415	45485.91	67	01-01-2009	5421
10424			25-11-2006	6189
10458			17-11-2007	5832
10515			27-03-2008	5701
10574			07-07-2008	5599
10596	14572	18	05-07-2008	5601
10613	8771.64	6	30-01-2007	6123
10616	2398	2	28-10-2007	5852
10618	135568.5	105	16-01-2009	5406
10683	4241	4	01-11-2007	5848
10694	4056	4	25-12-2006	6159
10735	4216	6	23-04-2007	6040
10904	14400.56	11	22-01-2009	5400
10909	4666.8	6	01-06-2008	5635
10915	980	2	07-10-2007	5873
10924	6013	4	30-11-2007	5819
10929	26494.38	30	17-01-2009	5405
10935	3120	2	29-09-2007	5881
10989	1138.8	2	11-02-2008	5746
11013	1560	2	24-05-2008	5643
11082	12436.22	8	11-01-2007	6142
11100	15403.7	10	15-10-2008	5499
11113	3000	2	18-11-2007	5831
11125	2598	2	30-08-2007	5911
11134	833	2	23-09-2007	5887
11158	8078	14	08-08-2007	5933
11165	798	2	26-04-2007	6037
11184	3640	2	11-06-2007	5901

The PivotTable Fields task pane on the right shows the following fields:

- TransType
- TransPrice
- TrnQty
- Total Value
- DocumentDate
- CustomerID

Annotations in the image include:

1. Row labels are customer ID which have been chose as rows on
2. Sum of total value gives the total monetary transactions perfor
3. Count of transactions quantity gives the total number of visits t
4. Days since last order(Current date-most recent transaction)-REC

The NPTEL logo is visible in the top right corner. A woman is visible in the bottom right corner of the screen. The text 'BUSINESS INTELLIGENCE & ANALYTICS' is displayed at the bottom of the image.

So the next column is total of monetary value. How did we get this? We multiplied the

transaction price and transaction quantity and also the transaction price was assumed as negative for returns, so that we subtracted properly. So from that we are getting the sum of all the values for a particular customer ID. So 10415 customer would never be repeated again in this data. So that is a unique record that we are having. So the second column is total monetary value.

The third column is nothing but count of transaction that is we had a column that was transaction quantity so we just have taken the count of how many times the customer has done the transaction. So if we take this record then we come to know that 10415 customer has done 67 transactions over his lifetime. So that is present as a single record in this pivot table. Here we are just printing out the max of order date.

So max is also a function of pivot table. So as you can see here, max of order date that is what I have printed here so that we can, you know just take the difference of max of order date with present order date, so that we will get this column which is nothing but days since last order. So what we have got, total monetary value is nothing but monetary value. Count of transaction is nothing but frequency value and days since last order is nothing but recency value. So we have got the R, F and M now and yeah, so we have got R, F and M but what is the problem with the data? It is not scaled properly.

If you take the M value it is 45000. The days since last order is 5000. So all these are not on the similar scale. So what we have to do is, we have to arrange this into buckets. It can be 10 buckets or 5 buckets or any number of buckets that you need and here in this example, I am grouping into 10 buckets. So the recency, frequency and monetary value will have a value from 1 to 10 and not from 1 to 5. So the last sheet that I am having has the final values that we need for the RFM calculations.

So these 3 columns are nothing but the columns that we got from the pivot table and using that I have scaled the recency, frequency and monetary value in the columns which are H, I and J. Just see this column. This is the formula that I have used for scaling. So percent rank exc, this is the formula that can be used similar to ntail function in SQL. So what does the ntail function in SQL do? It will group into how many ever groups that you need the data to be grouped in.

So similar function is done by the percent rank .exc function. What it does is, we have given which column we have to scale. That is we have given this G column, that is G2 to G1539. So we have 1539 values which have to be grouped and scaled. So we have selected that and we are multiplying by 10.

Why we are multiplying by 10 is that if we do not do we will get a value from 0 to 1. So it will be like 0.1, 0.2, 0.3, etc. So we want a whole number, that is why we are

multiplying by 10 and this 10-, why have I done this 10- for recency is, can you just take a minute and you know think why I have subtracted in from 10.

I think you guessed it right. So I have done the reverse scaling here. So reverse scaling is needed for recency alone. If you see I have not done it for any other function, that any other matrix like frequency or monetary I have not done that 10 minus because I am not doing the reverse scaling. So you know why reverse scaling is needed for recency. So that is why I have subtracted the value from 10, so that we get the 10 value for the customer whose recency was the highest or who has visited the store very recently. So that is, if you can see the 5, 4, 2, 1 that is the time lapsed between the last customer order and today is 5, 4, 2, 1 days that guy gets the value as 10 and the guy who has visited even before that, gets the value lesser, that is 1.

So we have scaled the recency properly. Now we are going to scale the frequency. Frequency nothing but same formula, person rank.exc but which column are we selecting? This column, that is count of transaction column or the frequency column we are selecting and we are scaling. I have added 1, because when we do this we will get from 0 to 9 but if a value if a frequency value is 0, then even if the recency and monetary value was high, if we multiply it by 0 then the whole value becomes 0. So that is very bad kind of segmentation because a person who was not recent, like his recency is 0 but he was a good customer like frequency and monetary was very high, then we need that customer in the segmentation or if the recency was 0, if we give it as 0 itself then that customer would not even come into the important bucket. So that is why I have just added 1, so that we get the values from 1 to 10 and not from 0 to 9. So that is how I have scaled the frequency and monetary is similar to frequency because you can see that it is $1 + \text{percent rank.exc}$ and which column are we selecting? The monetary value column for calculation.

So now we have scaled frequency, recency and monetary value and now what we have to do is calculate the RFM value which is nothing but this into this into this multiply all the three values. If you want to add weights, then well and good you can add. As of now I am not adding because this is a simple RFM analysis and we are not adding any weights. So this is how we calculate the RFM values and using these RFM values the next step would be customer segmentation.

The screenshot shows an Excel spreadsheet with the following data columns: FREQUENCY, REGENCY, Recency scaled, Frequency scaled, Monetary scaled, and RFM. A yellow text box contains the following instructions:

1. Recency value has been obtained by grouping "Days since last order into 10 deciles and then scaling them from 0-9"
2. Reverse scaling done for Recency since the most recent customer needs to be given the highest R value
3. Frequency values have been obtained by grouping "Count of txn qty" into deciles and then scaling it to 0-10
4. Monetary values have been obtained by grouping and scaling "Sum of total value" column
5. RFM=R**F**M
6. Next step is customer segmentation which will be done with this preprocessed data in python

The spreadsheet shows a pivot table with the following data points:

Row	FREQUENCY	REGENCY	Recency scaled	Frequency scaled	Monetary scaled	RFM
2	67	5421	10	10	10	1000
3	24	6189	1	10	10	100
4	6	5832	5	6	5	150
5	2	5701	7	1	6	42
6	6	5599	7	6	8	336
7	18	5601	7	10	9	630
8	6	6123	2	6	8	96
9	2	5852	5	1	1	5
10	105	5406	10	10	10	1000
11	4	5848	5	3	4	60
12	4	6159	1	3	4	12
13	6	6040	4	6	4	96
14	11	5400	10	9	9	810
15	6	5635	7	6	4	168
16	2	5873	5	1	1	5
17	4	5819	5	3	6	90
18	30	5405	10	10	10	1000
19	2	5881	4	1	2	8
20	2	5746	6	1	1	6
21	2	5643	7	1	1	7
22	8	6142	2	7	9	126
23	10	5499	9	8	10	720
24	2	5831	5	1	2	10
25	2	5911	4	1	1	4
26	2	5887	4	1	1	4
27	14	5933	4	9	7	252
28	2	6037	4	1	1	4
29	2	5981	4	1	1	4

The video inset shows a woman speaking, and the bottom of the screen displays "MORE VIDEOS BUSINESS INTELLIGENCE & ANALYTICS".

So now what we have got is, the entire R,F and M values of all the customers within 1 to 10. So we have segmented these customers into 10 buckets for each recency, frequency and monetary but when we combine all these R,F and M values, what can we say about how a customer is? For example, we have, say we have a customer with R value 1, frequency is 9 and monetary is 9. What does it say about a customer? It says that the customer is not at all recent. Almost he has turned off from the business because his frequency and monetary value was very high. That means he was a very good customer but because of your improper promotional tactics or something else, some problem he has turned off to another business.

So that is what it means by this customer. Take another customer whose recency is 9, frequency is also 9 and monetary value is 1. What does this mean? That customer is always coming to a store or always visiting a website and he is very recent also. Today also he visited the website but monetary value is very less. That means he is just doing it for fun or maybe he does not have the money to do it.

So we do not know what is the problem is, but can we cluster both these customers into the same bucket? No. No, right. If we cluster both these customers into the same bucket then we will be sending similar kind of marketing mails or targeted offers to these customers, then it would be very absurd because both are very different customers who should not be clubbed into a similar bucket at all. So that is why we need a clustering algorithm, in which we feed this recency, frequency and monetary values so that it will have a prefix strategy on how to cluster these customers based on how many ever K value that you give. For example, obtaining the K value there are various methods like

elbow method etc.

So using that you can select an optimal K. K is nothing but the number of clusters that we want the customers to be grouped in. So we can select an optimal K and then we can get the different customer segments or the cluster segments on which the customers can be segmented. So if there are 4 customer segments, that means that all the strategies henceforth that we are going to follow as part of our business will be similar for a particular cluster. So cluster 1 will have similar kind of marketing mails, similar kind of strategies, promotional offers, loyalty programs etc.

But this will not be similar to the cluster 2. So the cluster wise marketing strategies would be intra-cluster marketing strategies would be similar and inter-cluster marketing strategies would be very different. So that is what the businesses do after they have done the RFM analysis.

So many big business as we saw, like the Delta airlines they have already performed RFM analysis to see what kind of customers they have and what kind of customers are in the situation of churning off or what kind of customers they have to bring in or the strangers which a stranger segment they have to bring into the business in order to increase the net worth. So there are different strategies, like that used by Delta airlines. So RFM is a very very simple technique because we can just do it in a Excel or a SQL to extract the RFM values and then, you know go forward for clustering.

So it does not analyze any psychographic or you know behavioral data of the customers other than this recency, frequency and monetary matrix. So in next session, we will be seeing how we can go about with K-means clustering with the input R, F and M values. Thank you.