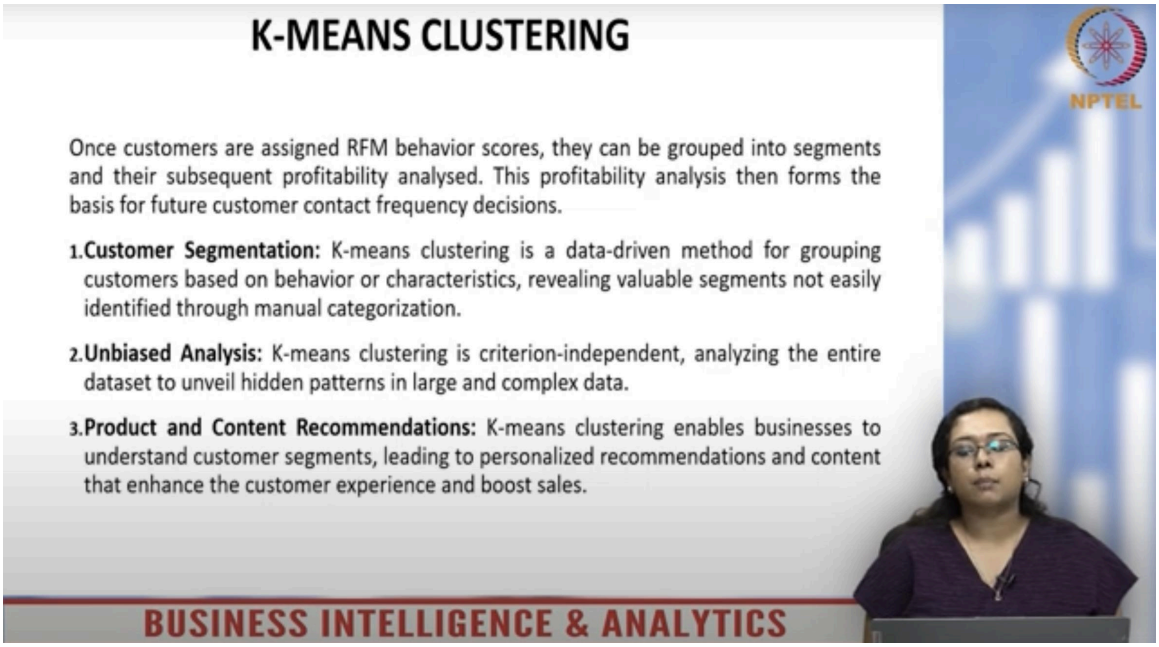
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:10
Lecture:39

RFM AND CLUSTERING

Hi, everyone and welcome to Business Intelligence and Analytics course. Today, we are going to continue what we have seen in the previous session, that is RFM analysis. Today, we are going to go forward and do the next step of RFM analysis which is clustering the customer segments with whatever R, F and M values that we have got after the RFM analysis. So if you have not gone through the RFM calculation, then I urge you to go back and read on RFM, how we do the RFM because the output, the input to the K-means clustering would be the R, F and M values of each customer ID that will be fed into the clustering algorithm which will give us a typical number of clusters of customers depending on what K value that you are choosing. So you all know that K-means clustering algorithm is a unsupervised algorithm and depending on the K value that you are going to give or the optimal K value that you can get from the ELBOW method, that we will be feeding in into the algorithm so that you will get the defined number of clusters that you need from the R, F and M values.



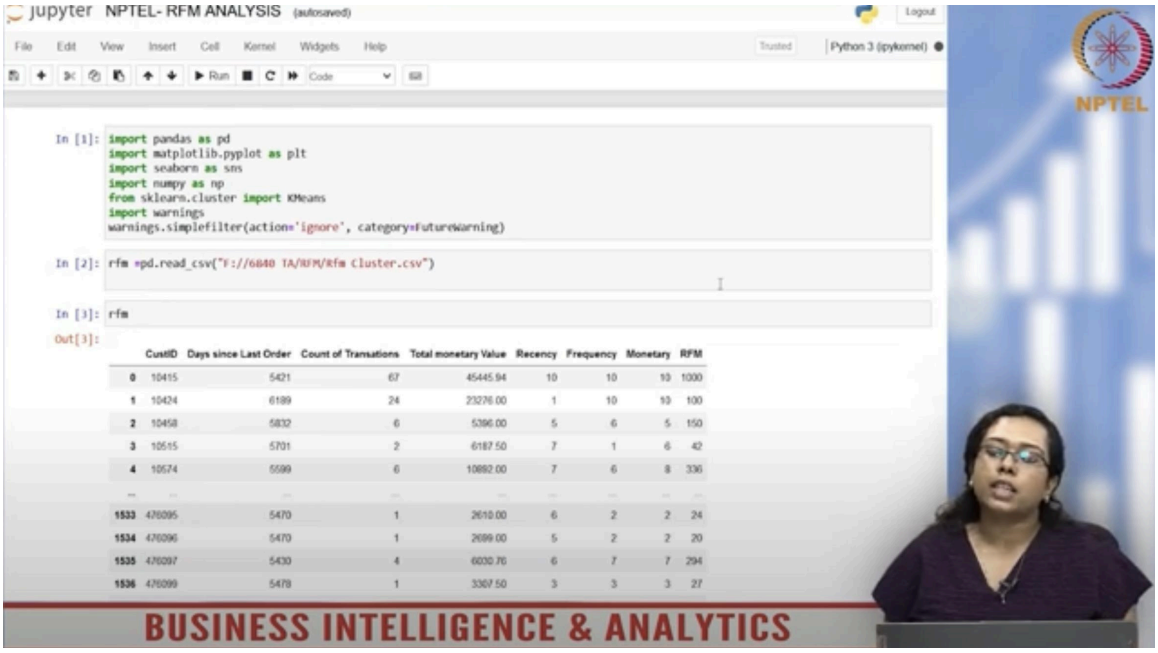
So we will go into that. So what is K-means clustering is something that you might have seen before in the previous sessions. So once the customers have got the R, F and M scores that are behavioural scores of the customers based on the past transaction data that they have done, they can be grouped into segments or different buckets based on their profitability, based on their purchase transaction history so that the companies can prioritize their customer relations or marketing strategies appropriately. So for this what we have to do is, we have to cluster these customers who we have already done RFM analysis on.

So the importance of K-means clustering is that, first obvious thing is the customer segmentation. So K-means clustering is nothing but an unsupervised data-driven method for grouping customers based on their characteristics. In here we just have three input features which are R, F and M. So it will give us valuable segments which cannot be easily otherwise observed by Excel or SQL. So we need to cluster this based on R, F and M values especially if our data is very very large then it is very difficult to do manually using Excel.

We need to have a clustering algorithm which will perform it easily for us. And unbiased analysis, K-means clustering is criterion independent and it analyses the entire data set to unveil hidden patterns in large and complex data. So here what data what we have is very large. It was about 50,000 initially, then we removed the outliers, we did the preprocessing of data post which the data became lesser and after that we grouped the customers based on customer IDs so that we have unique transaction details for each customer ID. So the data even became smaller and now we have extracted the R, F and M values which we will be feeding into the clustering algorithm. So the importance of clustering, another importance of K-means clustering is it will give us the information on which cluster a particular customer ID falls into, so that we can do the appropriate strategies or product or content recommendations based on which cluster he or she is belonging to.

For example, if you are a person who enjoys watching rom-com movies on Netflix, then Netflix will give you recommendation based on your recency, frequency and not monetary value because Netflix is, yeah maybe monetary value as well because you can take a premium subscription on Netflix as well. So based on your recency, frequency and monetary matrices it can, your product recommendation may not be similar to another person's product recommendation. Even in Netflix account one account can have many profiles. But if you go and visit those profiles, you might come to know that not, two profiles are not similar. There will be a kid's profile which will be having very different algorithms to recommend which movies or which shows you can see.

So that will be very different from an adult profile which will be having different kinds of movies and even among adult profiles, no two adults will be having the same recommendation. Say one adult prefers watching English movies, then if you go and see the recommended list it will be full of English movies, say action movies but it will not be the same for another adult. So the product and content recommendation depends on how, which cluster you belong to or which cluster the algorithm has put you into. So that is why, how and why you get the product recommendations that are coming to you. So that is why we are doing the K-means clustering and we will go into Jupyter notebook which you have to install and then, you know practice the K-means.



I am having the raw data which I have turned into, filtered and cleaned and preprocessed data with R, F and M values which I am going to feed to the K-means clustering. So we will go into the code and see how we do it. So this is the code where we do the RFM where we have we are going to do the K-means clustering on the R, F, M data that we have already got from the Excel calculation that we did in the previous session. So in this code I am using different libraries like pandas, matplotlib, seaborn, numpy, scikit-learn, etc. And the first line as you already know, what I am doing is I am importing the pandas library and giving it an alias of pd, so that we can type it easily.

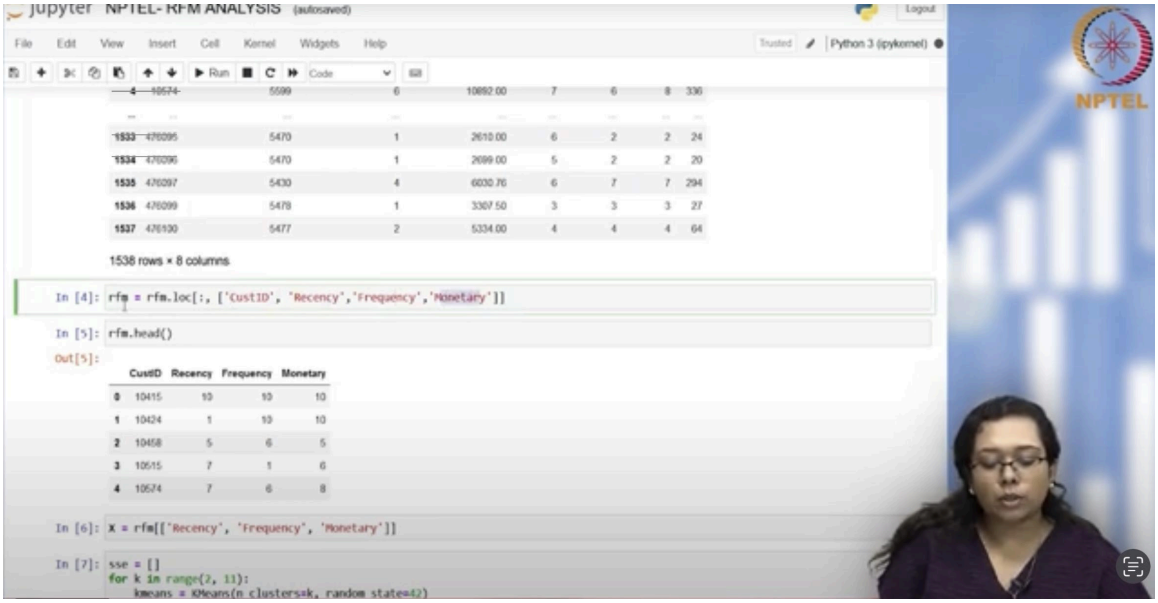
So pandas, as we all know is a very very important library, almost we use it for all the codes. So it helps us with data manipulation and analysis as well. So the next line is import matplotlib.pyplot as plt. So this is nothing but a data visualization and data plotting library that is matplotlib.

So we are downloading the pyplot module and we are renaming it as plt for ease. And the third line is import seaborn as sns. So seaborn library is nothing but a data visualization library. You will come to know why I have imported this because I have done some 3D interactive plots at the end so that we will get to know how these different clusters are clustered using these algorithms.

And the fourth line is import numpy as np which is a fundamental library for numerical operations. And the next line is from scikit-learn.cluster, we are importing the K-means. This is the important line from where we are importing the K-means algorithm and from the scikit-learn library. And next line is nothing, next two lines are just for filtering out the warnings from the scikit-learn library because when I tried coding, it was, there were many warnings, unnecessary warnings that were coming. So then, in the next line what I am doing is rfm equal to pd.read csv. So this line actually reads the csv file and the csv file's name is rfm-cluster and I have given the path name as well and it reads the file using the pandas library and it loads it to the data frame which I have named as rfm, rfm is the variable and it stores the data frame in that. So in the next line, I have just printed out the rfm for you so that you will come to know what all data is there in the rfm cluster.csv file. So it has nothing but customer ID, then days since last order, count of transaction, total monetary value, recency, frequency, monetary and rfm.

So days since last order, count of transaction and total monetary value are the unscaled r, f and m values. The scaled r, f and m values are recency, frequency and monetary and finally we have multiplied r, f and m to get the rfm value. Do we need all these columns for our K-means clustering? No. We just need these three columns which are recency, frequency and monetary and the customer ID so as to group them into clusters. So next line is rfm equals rfm.location. We are assigning the customer ID, recency, frequency and monetary into rfm because that is all we need for our analysis. So I am printing out the rfm.head, so that you will come to know that the data frame has only four columns which are customer ID, r, f and m. So I am just, in this also customer ID is not needed.

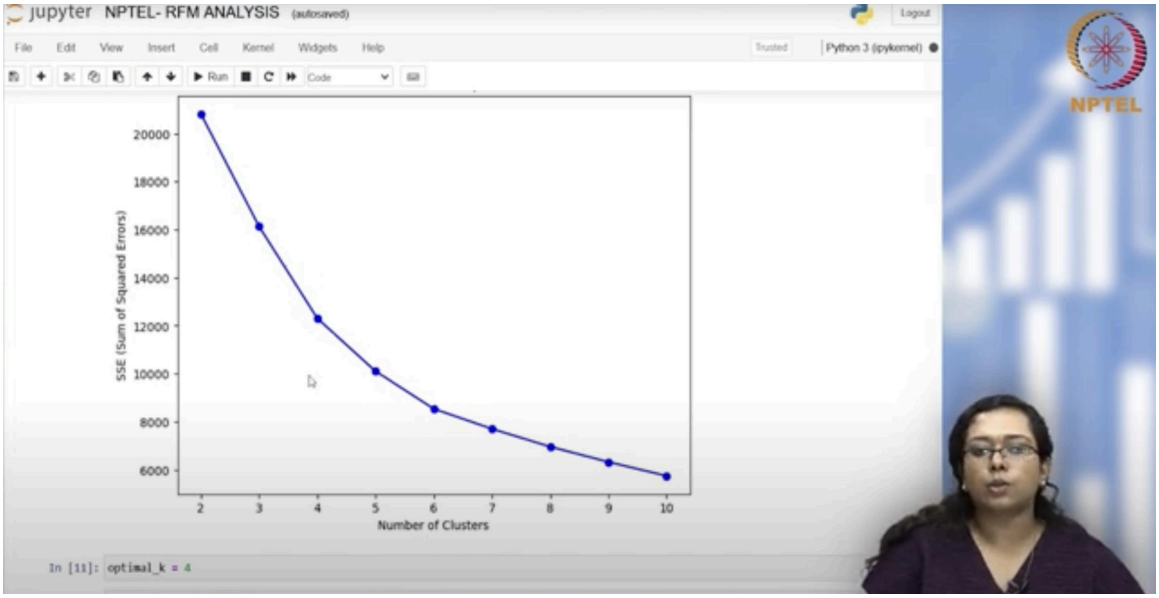
So I am just assigning x. This creates the data frame x by selecting only the recency, frequency and monetary columns from the rfm data frame because that is all we need for our analysis. Next we are going into the K-means clustering itself and for that we are initializing an empty list called SSE. So SSE is nothing but sum of squared errors and you know that the SSE has to be low in order to make a proper cluster. So we are initializing an empty list called SSE to store the sum of squared error values, for each value in the cluster K. And then what we are doing is for K in the range of 2 to 11, this is nothing but a loop that iterates over the values of K from 2 to 11 indicating that the code will perform the K-means clustering with the varying numbers of clusters from 2 to 10.



So inside the loop what we have is, what we have is we are declaring the K-means equal to K-means of number of clusters equal to 3. Since initially we have to give a cluster number, we are just assigning a random number of cluster which is 3 and this creates a clustering model with 3 clusters. So the random state we can set it to, to any number. I have just set to 42.

The next line is K-means.fit. In that x data frame is fit into the K-means algorithm and this model will partition the data into K clusters. How many we have given? That is 3 clusters based on the recency, frequency and monetary value. The next line is ssc.append within brackets K-means.inertia. So what it does is that this appends the value of the sum of squared errors for the current clustering and it will clustering model to the sum of square error list, that we have already initialized in the first line. So SSE is a measure of the model performance as you know. Low SSE means better model performance and better clustering. So that is meant by these lines and I am printing out the RFM values again for you because we will know which cluster our R, F and M belongs to. So we have got 1538 rows which is the number of records that we have and for getting the optimal K, what we have to do is, we have to use the ELBOW method.

So ELBOW method is nothing but it plots the sum of squared errors versus the number of clusters. So the sum of squared errors has to be low in order to clusterize our data. So the ELBOW is the point where it abruptly decreases. So that is the point that we have to select. So the clustering that we have done, we can take the ELBOW as 4 or 5.



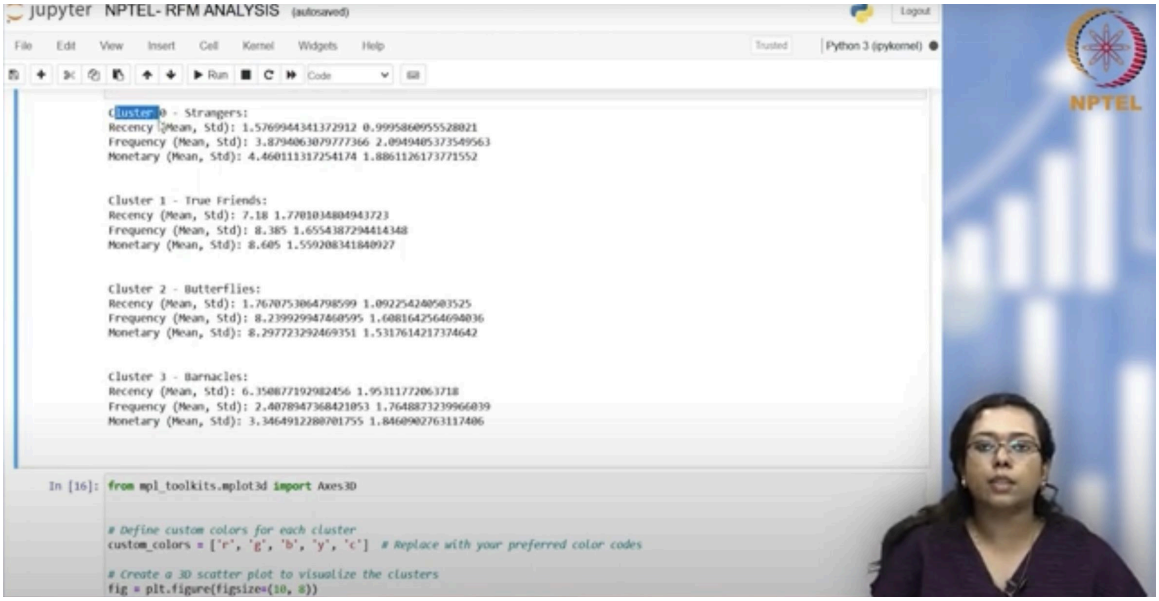
So it can be anything because that is where it abruptly changes or decreases. So I am here taking the optimal K as 4 because I want to see if we can bucket these clusters into 4 cluster segments that we already saw, that are butterflies, strangers, true friends and barnacles. So we have already seen what customers those segments are, like what are the customer statistics for each segment that belong to one of those 4 clusters. So we have seen what they are and I am just trying to see if we can bucket them into similar clusters after this, performing this algorithm. So here what we are doing is, we are changing the K.

We are changing the K value initially set as 3. Now we are setting the optimal K as 4 so that we can feed it into the K-means algorithm. Now, we are again repeating the same thing and we are also, yeah, so we are also adding a new column called clusters in this line. So in the line K-means equal to K-means within brackets number of clusters we are giving as optimal K which is 4 and then random state anything you can select and then a K-means model is created with those number of clusters and the optimal K which we have given as 4. Next is rfm within cluster equal to K-means dot fit predict of x.

In that what we are doing is, the K-means model is fitted to the data in the x data frame containing the r, f and m values and the next line is we need to print out the cluster statistics or, okay, okay. Here we are initializing the cluster stats because we want to print out the cluster statistics, because otherwise we will not be knowing what is the mean value of recency for a particular cluster or what is the, you know, mean or standard deviation value of cluster 1 for the value of recency, frequency, monetary. So we will not be knowing, if we will not print it. So I am just initializing cluster statistics so that we can print it later. Before printing out the statistics, I am just printing how many values

are there in each clusters.

So we have already fitted our data frame into K-means algorithm and we have got 4 clusters. So cluster 0 count is 539, 1 count is 200, 571 is cluster 2 count and cluster 3 has 228 values. So we have got the count already. Now we are going to define the cluster names. This I have defined already because I saw the values and the means.



So I have, you know, assigned them the 4 names or the 4 customer segments that we have already studied. So before assigning names, what are the 4 clusters that we have got? Cluster 0, cluster 0 has recency of 1, frequency of 3 and monetary value of 4. So what does it say about the customer segment? 1, 3, 4, it is not a very high value. It is actually low compared to other values that we have.

That means that they are strangers. Strangers are disloyal, unprofitable customers who do not tend to stick with the business or, you know, spend more on the business. So that is why we have clustered them as strangers or I have given the name as strangers for cluster 0 because of the cluster statistics which are recency, frequency and monetary which is coming only as 1, 3 and 4. The cluster 1 if you see, it has average value or mean value of 7, 8 and 8. That means everything is high. Recency is high, frequency is high and monetary value is high.

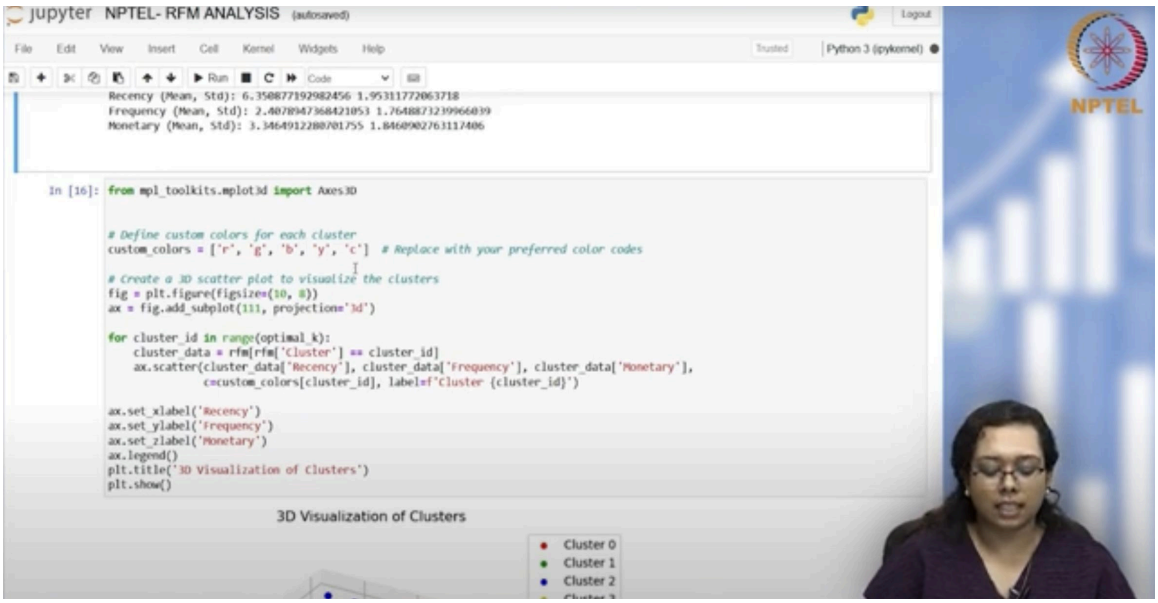
That means we had already seen that is how true friends behave and, you know, true friends as a customer segment are very loyal, very frequent as well as very profitable. So that is why I have named the cluster 1 as true friends. The cluster 2 is named as butterflies. What do butterflies tend to do? When they come and visit, they will be very

frequent and they will be having high monetary value purchase, purchases or transactions.

But they are not very loyal. They are not very recent. Okay. The recency score will be extremely low. That is what butterflies tend to do. So when I analyse the cluster 2 statistics, it was 1, 8 and 8. That means the recency value was very low. But when they do the transaction, they purchase high net worth things. So that is why the monetary value is high. That is why this cluster has been named as the butterflies.

And cluster 3 as we know, barnacles. What barnacles do is, they are very loyal. They stick to the ship, but they do not buy anything. That means that their monetary value and frequency value should be low. So once they have come, they stick forever, but they do not purchase anything. They are not very profitable. So that is why the recency is high here and the monetary value is very low.

So that is how I have clustered this into 4 clusters and by giving the optimal K as 4. And now, we will go into visualising how these clusters see, clusters you know, are, you know present. And see them in a 3D kind of scatter plot. So this has 3 dimensions, R, F and M. So a simple scatter plot will not be able to give visualisations on, you know or give proper visualisation on how they are clustered.



So since it has 3 dimensions, we have to have a 3D scatter plot and that is what I am going to do next. So the next code segment that we have in front of us provides cluster specific, you know scatter plots, that is we will have a 3D scatter plot with 4 clusters that we have gotten as output from the K-means algorithm. So we have this 3D scatter plot

that we have visualised using the matplotlib library which has 3D plotting capabilities as well. And the first line is from mpl_toolkits.mplot 3d import axes3d. So it is nothing but it imports the axes 3d module from the mptoolkits.mplot3d library enabling 3D visualisation or 3D plotting capabilities of matplotlib.

Then we are assigning random colours. So you can give colours of your choice. I have given red, green, blue, yellow and so on. And after that, I have just given the plot figure size and then, you know then we have initialised the cluster in a for loop which selects the data points in the RFM data frame that belongs to a current or particular cluster. And then we are initialising the ax.scatter which plots the scatter plot with 3 values which is recency, frequency and monetary value. And we are setting the label for the 3 axes which are recency, frequency and monetary and then I am just showing it, which is giving the title 3D visualisation of clusters and this is the plot we have got.

So this is the recency value, this is the frequency axis and the monetary is the axis behind that or the 3 dimensional, third axis that is the monetary . So if you come to see, this red guys which are the cluster 0, which we already saw as the strangers, who have less values for everything. Those are the red values or red dots that we have are cluster 0 which are strangers because they will be having less values like recency is less, frequency is less and the monetary value is also less.

The cluster 1 was true friends. Yes, cluster 1 had mean values of 7, 8 and 8 for RFM. So cluster 1 are the guys which have high values for everything that is recency, frequency and monetary.

Cluster 2 was butterflies. So butterflies are not recent at all. So cluster 2 will be having recency score as very very less. So if you can see the red and blue have recency in a similar line. So that means that red has low recency obviously, so blue also has low recency. But blue is atop red, that means that the other values that is frequency and monetary is high for blue cluster or the butterfly cluster.

The cluster number 3 or the yellow dots that you see are the fourth category that are barnacles. Barnacles have high recency. So if you see along the recency axis, the barnacles have high recency. But other two values, that is a monetary value if you see is very down and the third axis would be vertical, so it is very in the bottom of the monetary value. So that is what barnacles do.

And for better visualization or kind of a 3D interactive visualization, I have also done a 3D interactive plot, which gives a 3D visualization of clusters.

This is just for fun. So you can just see how I have done this. So I have imported the plotly.express as px and this is for 3D interactive scatter plots or graphs. So what I have done is, I have the K-means model that we have used is used to predict cluster labels for data points in the X data frame and the cluster labels that are stored in the cluster column, in the RFM data frame. That is what this line means. And the next line is for 3D scatter plot with plotly and this is the syntax for it, wherein we are feeding four things which are recency, frequency, monetary and the cluster values so that we will get a interactive plot.

And then we are just, you know this figure.update layout. This code block allows you to customize the appearance of the plot. So this is for setting the titles and axes. So that is all. That is about this line and then we are just calling the show function.

So this is the interactive plot. It is similar to the previous plot but if you see you can know the cluster statistics, if you just scroll into the cluster itself. So this has three axes. As you can see monetary, frequency and recency axis. So this yellow guys, yellow guys if you just go and scroll over it, it is falling into cluster 3.

So cluster 3 was butterflies, right. If I am not wrong cluster 3 was, sorry cluster 3 was barnacles. So we have from cluster 0 to 3 not 1 to 4. So cluster 3 was barnacles and barnacles, they have high recency but everything else is low. So we will just see how cluster 3 fares in this interactive plot.

If you go and see this yellow guy, the recency is 9. The frequency and monetary value is just 1. That means that he is very recent. He is being recent but not being of any worth to the business. So these yellow guys are like barnacles who hold the ship down. So we

can see, we can just move through all the axes and we can move through all the yellow dots and see that these barnacles are actually of no value to the company and they have to re-strategize to see how to engage these barnacles who are very loyal people, who do not leave the company at all but they do not bring anything.

So how to bring monetary value to buy these customers are something that the company has to decide on. So these people need different strategies compared to the other clusters. So if you just, you can just scroll through all the yellow guys and you can just see that all have recency, high values for recency and the other values are less. Those are the cluster 3 guys and we will go to another cluster, the orange maybe.

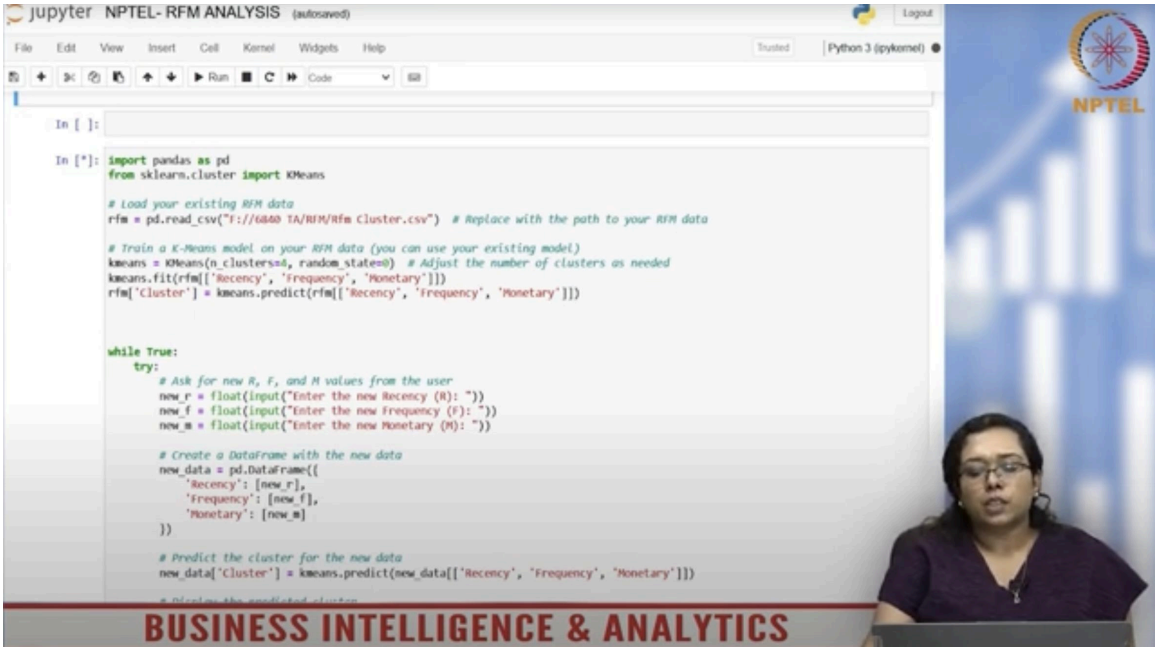
So that is cluster 2. So what was cluster 2? I guess it was butterflies. Yeah, cluster 2 was butterflies. So what do butterflies do? Butterflies are, I want you to guess, I want you to just think what butterflies does or we can even see from this interactive plot itself, you will come to know that recency is very very low. That means they are completely disloyal to the company. You cannot, you know just assume that if a person is coming today, he will come in the future. So whenever that person is coming, you have to make maximum use of that customer. That is how the butterflies behave. So the orange color guys are the butterflies here. So if you know that you know that the frequency and monetary will be high but the recency will be extremely low for the butterflies. As you can see from all the orange guys that are here, you can see the recency is very very less for these guys.

Next we will go to the cluster 0. So cluster 0 have all values which are very less. See the values for cluster 0. Those are the blue guys here. So the cluster 0, this guy has 4, 6, 4 that is pretty nice but let us go here 4, 4, 3; everything is below 5. So they are actually, if we go here it is 1, 2, 4 only. So all these blue guys are having very very less values, mostly below 5. So they are strangers. So company needs to restrategize on whether they need to spend this much on strangers, this much marketing, promotional offers and etc. to stranger category who are neither loyal, who are neither profitable or neither frequent. So that is how the company needs to strategize their marketing techniques.

So the last and final one are the violet guys who are at the top, who are nothing but they are the true friends. So we see the violet, you can see the recency is 10, frequency is 9, monetary value is 6. So the cluster 1 guys are the true friends who excel in all the 3 matrices which are recency, frequency and monetary value. So these are the guys the company should have more focus on and these are the guys where we spend less and get more or these are the 20% guys who contribute to the 80% profit. So this is the segment where the company has to put focus on and in order to prevent churn at any cost.

So that is how we can, that is the analysis that we get out of this 3D plot. So you can just play with this plot and turn it around and zoom, turn it like this and see, visualize this in

many other ways. So this is the advantage of an interactive plot. You can see from all the axes and yeah, that is the basically what the interactive plot does.



And the next one that we are going to see is, we are going to predict which, till now we saw the K-means clustering algorithm just gave us the clusters in which the different values fall into. Now if we give a new r, f and m value, can it predict which cluster it might fall into? So for that we have to use the K-means predict function. So that is what we are going to do next. So here I am just loading the existing data and this is pretty much the same thing that we have done, K-means.fit, we have done again and this is the predict function that I am calling, so that we can predict the cluster values of a new RFM value that has not been there in the data set. So here I am just initializing new r, new f and new m, which it will be getting from the user or we will be able to enter the r, f and m value and we are storing a new data frame for that, to store the r, f and m values and we are calling the predict function so that it can predict over the new r, f and m values that we are giving and we can also display the predicted cluster number. So we are just printing the location of the cluster of the new data frame that we have initialized just now. So this is nothing but the cluster statistics that are the recency matrix, the frequency matrix and the monterey matrix of the cluster and also the cluster ID.

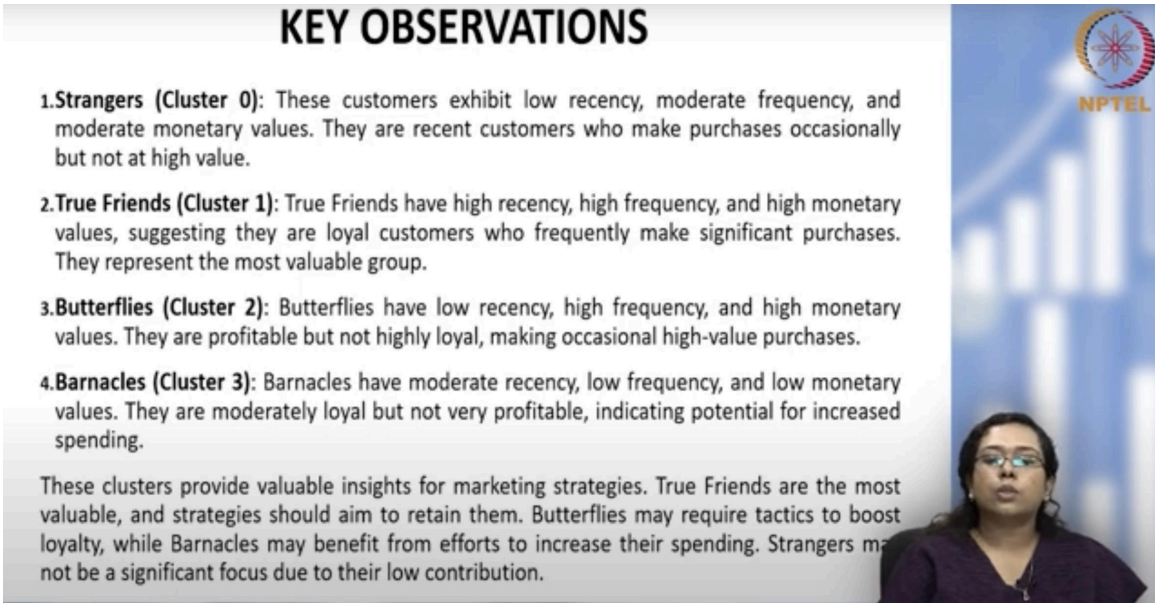
So the new predicted value will fall into either of the four clusters that we have that is cluster 0, 1, 2 and 3. So here I am just entering 9, frequency 9 and monetary 9. So you can see that the predicted cluster is cluster 1. So we know that our cluster 1 was true friends. So all was high, that is recency is high, frequency is high, monetary is high that is why they are true friends that is why the predicted cluster is coming as cluster 1.

Let us enter for something else like you know strangers maybe 1, 1 and 1. Yes it has come into cluster 0, which was strangers. Now we can maybe, you know give for butterflies whose recency is very less, frequency and monetary is high. Yes, so cluster 2 which is the butterfly.

So we have got that and the next one is, so the last cluster is barnacles cluster whose recency is very high, frequency and monetary value is very low. So that is why it has fallen into cluster 3. So we have got all the clusters, we have, you know tested for how it is predicting for the new RFM values and it is working just fine.

So that is all about the K-means analysis or K-means clustering after the RFM analysis that we have done, so that we can you know group the customers into different segments. So this is the analysis of all the clusters that we have got. We already saw that the cluster with the least values for all the R, F and M belong to cluster 0 or the strangers. Cluster 2 or the butterflies they have very, very low recency but the frequency and monetary is very, very high. Those are the butterflies. The cluster 1 are the true friends who have high values for all, that is 7, 8 and 8. Then cluster 3 are barnacles which tend to pull the business down by their existence, that is their frequency and monteray is very less but, you know they are very loyal and they are pretty recent, that is their recency is high. So the cluster 3 was barnacles.

And that is what we have got from the analysis and the key observation is that the strangers or cluster 0 is a segment where, you know they are very occasional. They are not of high value. So the business should strategize itself so as not to invest more or unnecessary on these kind of customers.



The true friends are loyal and they are very profitable. So the company should invest on, you know not losing them or not giving the opportunity to them to churn off to another business. The butterflies have low recency but when they come, they make huge purchases. So when they come, we need to give them promotional offers and, you know other, you know loyalty rewards, not loyalty rewards, rewards for, you know high value purchases so that whenever they come they make huge give huge profit to the company. So those are cluster 2 and the barnacles are of you know moderate recency or rather high recency but low frequency and monetary value. So they are not very profitable. So they have to be analyzed on how to, you know bring these barnacles and make them true friends.

So these clusters are providing very valuable insights for how the businesses should, you know strategize their marketing tactics. So we saw that they have to, you know aim to retain the true friends and you know butterflies should be having, you know, towards the butterflies, they should have tactics in order to boost their loyalty because they are not at all loyal and from barnacles we need to, you know change them into true friends so that, you know we can make them increase their spending and strangers should be the least of the priority. So hope you liked this session on RFM, you know followed by the K-Means clustering. So that is all for today. Thank you.