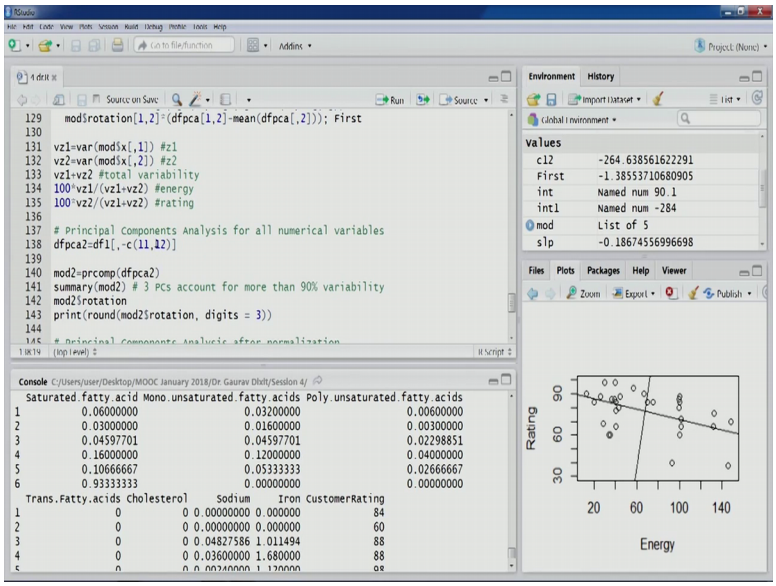
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 15
Dimension Reduction Techniques- Part III Principal Component Analysis

Welcome to the course Business Analytics and Data Mining Modelling Using R. So, in the previous lecture we were discussing the dimension reduction techniques, and specifically principal component analysis. In the previous lecture we applied principal component component analysis on breakfast cereals data base and two particular variables and had it energy and customer rating.

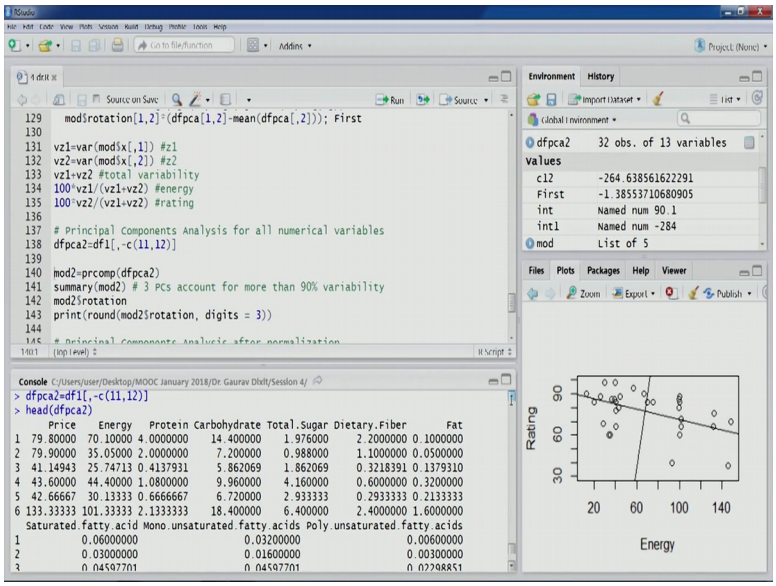
So, now today's lecture will start will applying principal component analysis, on almost all the values that are available in the data set.

(Refer Slide Time: 00:53)



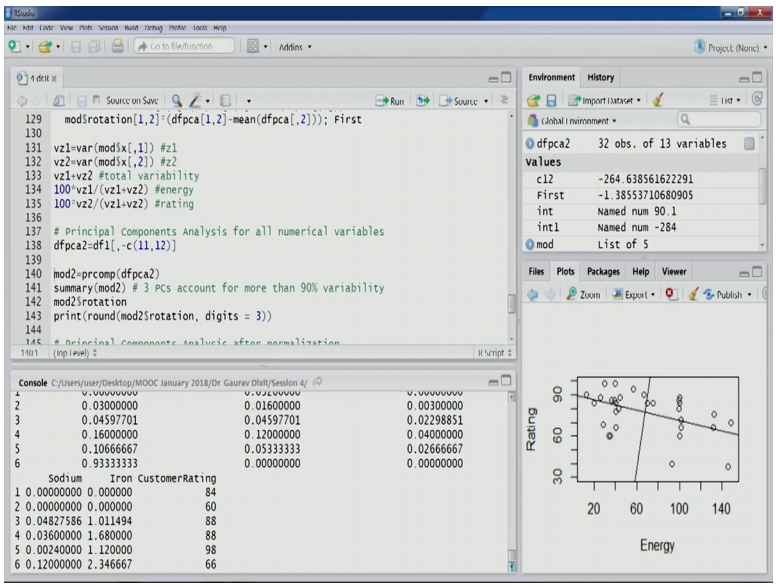
So, as we talked about in the previous lecture that will include all numerical variables except two, which are mainly having a zero values in almost all the cells, which will eliminate them ah. So, generally generally different brands try to indicate that these two you know these two particular variables stands for the fatty acid and cholesterol or 0 in their product so, that they can market them much better. So, that is why their information has been recorded, but essentially yours most of the values are 0 that is not useful for us

(Refer Slide Time: 01:30)



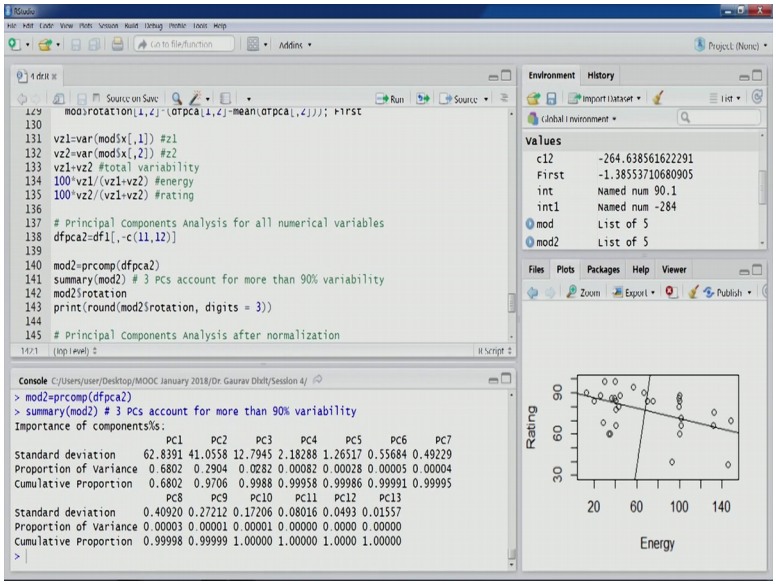
So, let us select the appropriate data frame. So, now, if you want to look at the selected variables again P C A 2 is the new data frame, here we have sub stetted ah. So, you can see the variables that all almost all the variables that were originally available in the data sets had been taken for principal component analysis had it starting on price energy, protein, carbohydrate. So, were dietary fibber fats, saturated fat fatty acids and mono unsaturated fatty acids poly unsaturated fatty acids sodium iron and last one the customer rating.

(Refer Slide Time: 02:10)



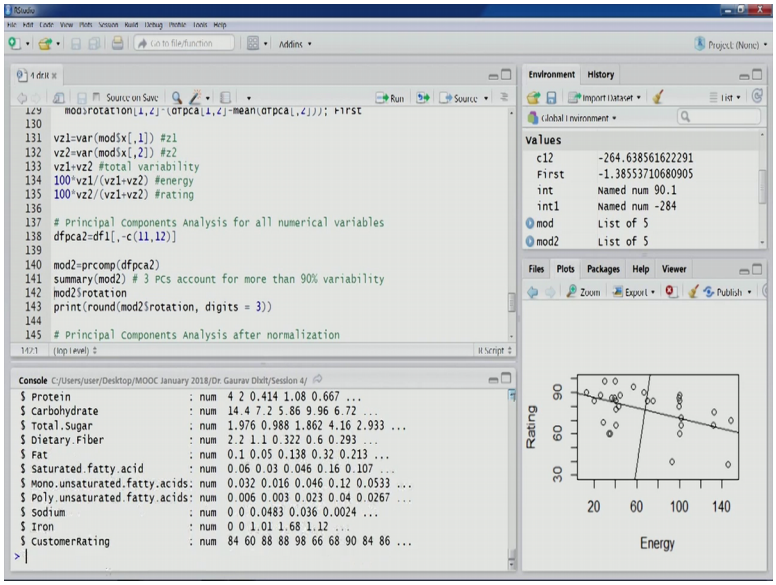
So, let us apply principal component analysis, the p r comp is the function again we are going to use it. So, let us execute this particular code, now let us look at the summary.

(Refer Slide Time: 02:30)



Now, so, you would the at 13 principal components having used. If you look at the if we look at the number of variables that we had in this new data frame on which we applied principal component analysis.

(Refer Slide Time: 02:51)



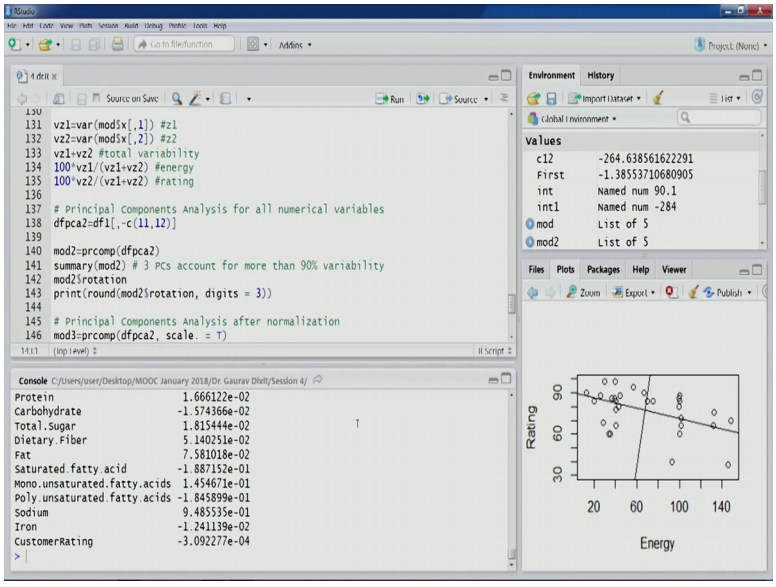
We count the number of variables, we can apply names function as well. So, 13 were there ah. So, now, 13 original variables were there, and now we have 13 principal

components. Now let us look at the different values that we got after applying principal component analysis. So, first P C 1 if you see the proportion of variance, its 68 percent and the second is prince second principal component is 29 per 29 percent. So, if you combined these two 68 and 29 its almost its its almost its almost I think 97 percent.

So, these two principal components P C 1 and P C 2 almost conti almost contributing 97 percent of the variability that was there by the original variables. So, we can eliminate other principal components. So, only these two principal components are capturing the most of the variability is therefore, the dimension can be reduced from 13 to 2 because the because of the most of the variability being captured by these two variables. Other principal component say p c; that the amount other variability that proportion are various type they cap captured is a less than 3 percent. P C is the 2.8 and then others are others are insignificant totally insignificant.

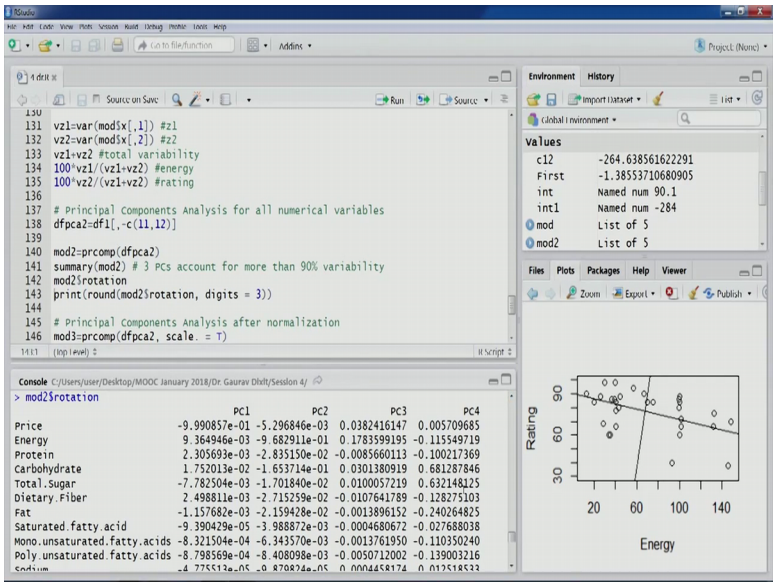
The first two we can have the first two principal components has two dimensions and therefore, we would able to reduce the dimensionality from the 13 to 2. So, let us look at other values.

(Refer Slide Time: 14:52)



So, let us look at the rotation weight weights. So, these are the weights.

(Refer Slide Time: 05:03)



Let let us look at the nicer version of this. So, we will ha will have we look at just three values, three decimal values ah. So, now, look at let us the first principal component P C 1. So, let us see which variables which original variables are contributing to P C 1 you would see that price that is minus 0.99. So, the first principal component is mainly determined by price and other original value means they are they are contributing to in in significant amounts. If we look at the second principal component, then it is you can look at the second values this is energy minus 0.968.

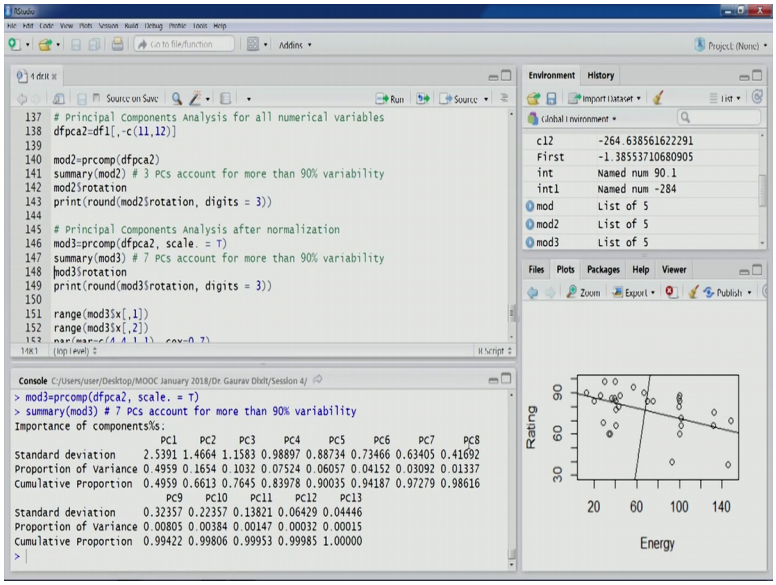
So, second principal component main contribution is coming from energy. So, P C 1 is essentially we can say price plus kind of variable and P C 2 is energy plus kind of variables. So, most of the variability being captured by price and energy; if you try to make sense of it you would see generally Indian consumers when they buy breakfast serious, they might they generally you know we might have this perception that they generally go for price and energy same is being reflected in a way here .

So, other principal components. So, proportional variance they explain was anyway quite less and then. So, they it does not make much sense to look at the contribution of original variables to these principal components. So, P C 1 , P C 2 its like price plus an energy plus. So, let us move forward now what is a problem here in this case? There is one problem that is there in this analysis that we applied. So, we look at the the variables that we are talking about its price that is in measured in a rupees, then the energy then that is

being measured in kilo calories, and protein and carbohydrate and other contents there we measured in grams and milligrams ha right. So, we look at we have having different you know measuring units. So, it is still the data that was fed to principal components analysis that was not normalized.

So, may be that was the reason we had just two principal components dominating most of the variability. So, let us apply principal component analysis after normalizing, all the numerical variables and a study. So, let us run another principal component analysis. So, again this time we are going to use the same data frame; that is d a P C A to and now we would see scaling is being down. So, scaling second argument is scaling is prove in the function now this let us execute this.

(Refer Slide Time: 08:23)



Let us look at the results, now these results are after doing normalization. So, normalization something that we have talked about in the previous lectures, we talked about that sometimes some variables because of the scales they can dom they can dominate results, they can influence the results and which might not be desirable in most of the scenarios. So, therefore, normalization is the one recommended step, before going add with the, you know going add with the building of your own model or running your model.

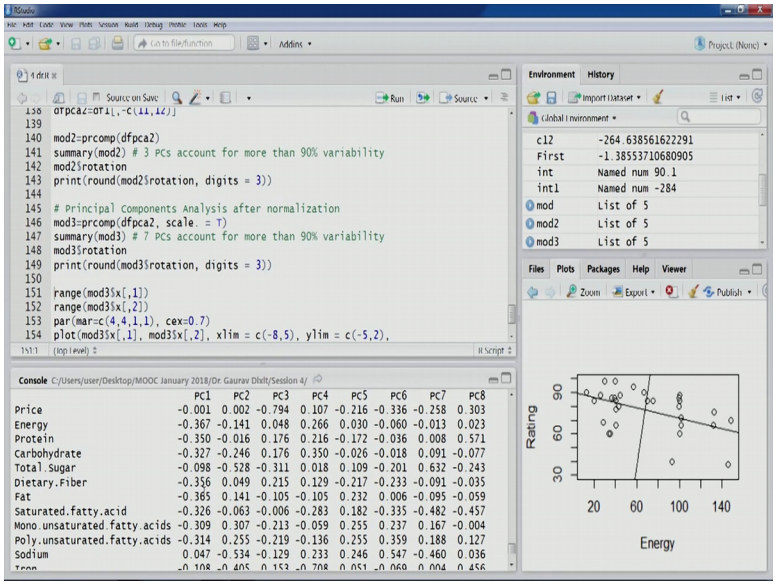
So, now this particular results that we see there after doing normalization. Now if we look at the portion of variance now we see P C 1 50 percent of variance is being captured

by P C 1 16.5 is captured by P C 2 10.3 is captured by P C 3 you can see much bigger role for P C 3, you can see even you know bigger role by P C 4 also same 7.5 percent, similarly P C 6 6 percent, we can see P C 6 4 percent P C 7 are 3 percent and after that.

So, you can see that first you know 7 first you know 7 principal components they are capturing more than 90 percent of the variability and most of the variability in the original variables. So, now, the dimensional dimensionality which we thought when we when we ran principal component analysis you know without doing normalization, we thought it was reducing from 13 to 2 that was not the actual case.

If we do normalization if we do scaling and we find out that it is actually from 13 to 7; so, we would still we requiring 7 new dimensions to capture most of the variability. Let us look at the weights of new principal components. So, let us look at the nicer version three decimal points, of to three decimal points let us start with first principal component.

(Refer Slide Time: 10:35)



In the first principal component if you see the largest contribution is coming from energy and then that is not the only dominant dominant contribution, he would see similar number protein, carbohydrate, dietary fibber, fat they and the other thinks also contributing in similar fashion right. So, this is how P C 1 is being determined. And the if you look at the P C 2 then sugar is dominating in this particular component 52 contribution coming from sugar, then we will look at another other numbers we would

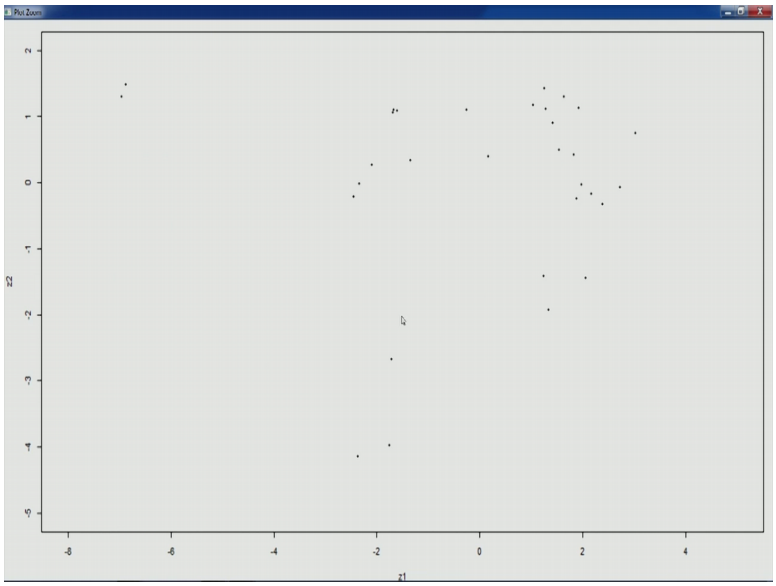
see sodium is also there even sodium is bigger than sodiums number is bigger than sugars, sodium sugar and then the iron. So, these are dominating the P C 2.

Similarly we look for look at P C 3, then the biggest biggest weight that we can see is coming from price that is minus 0.79; and then after that much smaller much smaller weights much smaller weights after that and we can see That point minus 0.31 that is again coming from sugar in principal component 3.

So, if we look at the principal component one the weights are also with minus sign right. So, therefore, you know energy and protein. So, this is a mainly signifying PC1 is mainly signifying the particular principal components which is determined by energy, protein, carbohydrate fat saturated fat fatty acids many of this contents right. P C 2 we see that is mainly determined by sugar and sodium. We will look at the P C 3 it is mainly determined by price. So, P C 3 could be called price plus P C 2 is mainly can be called sugar and sodium sugar sugar and sodium, P C 1 is may can be you know termed as health plus. So, these could be the new names for these different principal component the different new dimensions, and since we require first seven principal components. So, similarly will have to do a similar excises for other principal components as well.

Now, let us what will do? I will plot the new dimension that we have just computed. So, let us look at principal component 1 and principal component 2; let us look go back to the results that we had earlier the proportion of variance proportion of variance by P C 1 and P C 2 is 50 and 16.5. So, let us plots these two dimension and then we can compare it with the original plotting that we had done earlier. Let us look at the range of particular variables minus 6.96 23.02. So, you can see appropriately values has has been specified, let us look at the range for the second variable and you would see that minus 4.14 to 1.49, you can see the appropriately values have been specified. Let us change the margin and correct expansion through par function and let us plot now let us zoom to this particular plot.

(Refer Slide Time: 14:51)

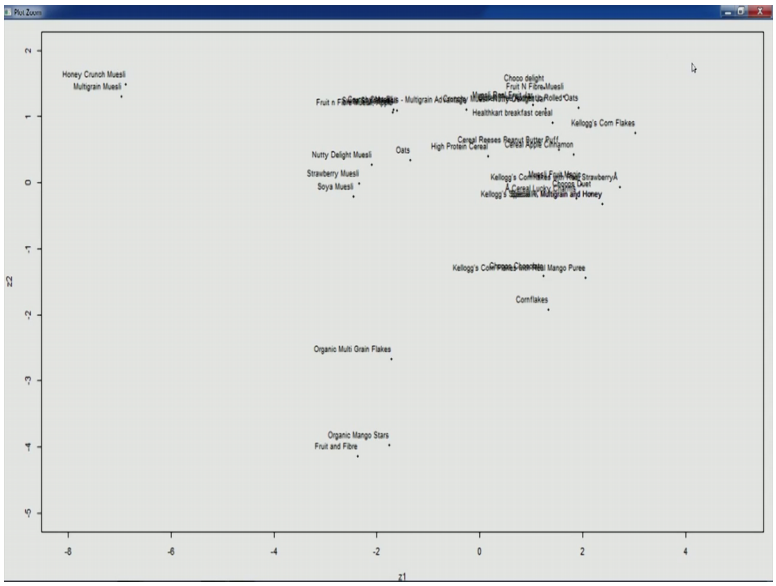


The marker point has been chose has been chosen as the smaller we have chosen a smaller marker for generating this plot, now you would see these are the these are the points for z one and z two .

So, quite different from the case that we had earlier had when be applied principal component analysis on a energy and customer rating. We saw distribution redistribution happening from 8713 to 90 and 10, and the and that was also you know without normalization. Now if we do normalize we can see this kind of scenario. Now variability that is being captured by just these two dimensions is slightly less in comparison to previous two models that we developed.

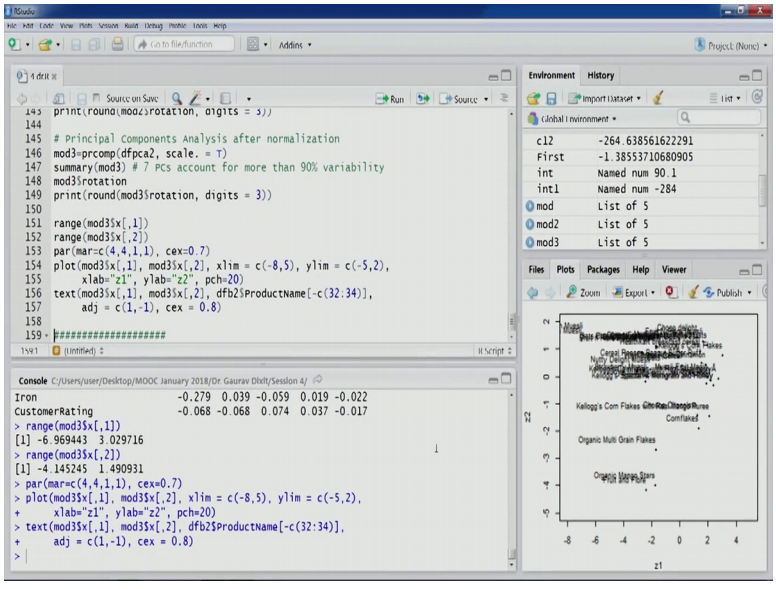
If we want to a further analyze formation, we can label all these point we can label all these point by their product name.

(Refer Slide Time: 16:02)



So, we can make a; we can analyse these a new dimensions further now. So, these are the all the points have been labelled by their product names. So, now, we want to further analyse right for example, if as we move along in the z 1 direction and will look at the weights, and that for actually contributing for this particular directions.

(Refer Slide Time: 16:25)



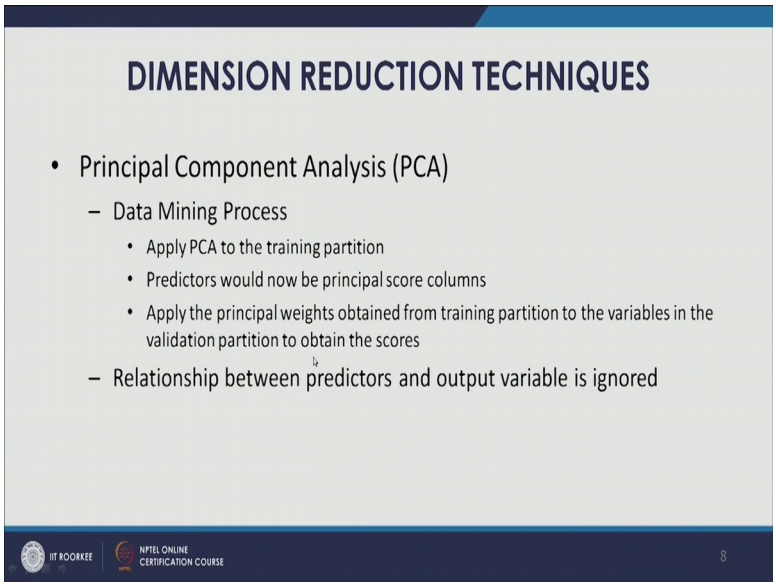
So, let us look at the P C 1. So, this was the P C 1 and it. So, as we move along to the z 1 direction from left to right, the energy content would actually decrease right. So, you would see that has move along from left to right the energy content is actually would

actually decrease and then the protein would also protein would also decrease similarly other carbohydrate fiber. So, as we move along from left to right probably less healthier options are more clubbed on the righter side of this particular plot.

So, this is how we can analyse similarly for the second directions also, we can make similar kind of analysis. For example, sugar and sodium they dominated sugar sodium and iron, they dominated the second directions second principal component. So, as we move from bottom to top and this direction get two dimensions, he would see that these decrease in these three contents, that is sugar, sodium and iron.

So, therefore, more healthy more healthy cereals would be would slightly be in the middle in mid section and then left mid section right. So, that is we are probably more healthier options are there.

(Refer Slide Time: 18:23)



So, let us go back to our discussions. Now we look at the principal component analysis; how it could be used in the data mining process? So, how we can actually applied principal component analysis in our data mining modelling and it use the dimension to lesser number. So, first step is going to be applying P C A to the training partition. So, will have training partition were validation partition and test partition. So, first step would be applied P C A to the training partition. So, now, will have new predictors, and they will be you know now we different principal is four columns, they would be new predictors.

So, the original variables we can we might will not be using further and new principal new score columns that we saw through that mod dollar x value right similarly for all the principal components, we can find new values and they are going to be new predictors. So, as we has been talking about will have new names for also new name for them; as were for example, we were talking about health plus, price plus, energy plus, kind of new variables for new principles four columns .

So, now once these we have new predictors, we are ready to build our model on training partition, but how do we evaluate our model? So, for that we would be requiring validation data set and test data set. So, the principal weights that be obtained while we applied P C A to the training partition, the same principal weights can be can be applied can be used to compute the variables new variables from the validation partition. So, the validation partition we can apply the weights that we computed from applying P C A to training partition, can we used to obtain new scores; and then these new predictors can we used to perform to actually test the model and then refined it, and then test on new partition that is test partition .

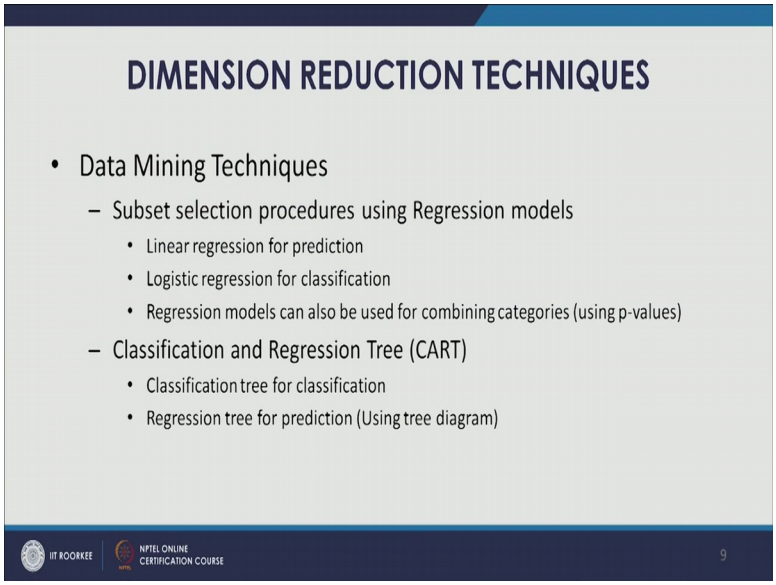
Now, we look further at P C A the, what we have been doing is in P C A is be mainly focused on numerical variables. So, we generally selected all the numerical variables and then we applied P C A on them, and then we looked an analyse the results to find out how many new predictors will have and whether the dimensions are going to come down or not.

So, those are the things that we looked at, but we look at the way P C A is done, we join generally exploit the relationship between predictors and output variables that relationship is generally ignored. So, that is one limitation of this principal component analysis. So, this particular limitation can be overcome using some other methods. So, limitation of principal component analysis that it does not include the relationship between predictors and the outcome variable, that can be overcome using some other methods. So, these are.

So, we come to our next category for dimension reduction techniques that is data mining techniques. So, some of the data mining techniques that we would be covering in more detail in the coming lectures, they can also we used to reduce the dimensions. So, first

one that we are going to discuss briefly is regression models. So, we can apply some of the subset selection procedure using a regression models.

(Refer Slide Time: 22:13)



So, for example, Linear regression for prediction task. So, for prediction task we can apply a linear regression. So, there in using the different using the significants of the coefficient that we get for different variables right, we can find out which of the important variables and therefore, we can we can get rid of the insignificant variables and also variables probably having low coefficient value. So, we can also if the we can also drops some of the values, if the coefficient numbers coefficient values is on the lower side those variables can also be drops even if that that is significant.

So, now domain knowledge also important, sometimes even though the coefficient value is on the lower side the variable might be of more importance, but if that is not the case probably we can drop those variables also. So, we can drop insignificant variable and some of the significant variables, which are carrying low value.

So, that is how selection subset selection procedures of regression models can we applied linear regressions, different subset selection method that we will be covering in coming lectures, can be run to find out the best subset, which is able to explain the model or fit the data. For classification task we can apply logistic regression, and the same process can be adopted there as well. We can use we can find out the significant relationships, and then we can also have a look at the coefficient values and thereby we can determine

which variables to drop and thereby we can reduce the dimensions. We can also look at which regression models, the subset are explaining most of the variance that we can do through multiple R square values.

Regression model can also be used for combining categories. So, we can use p values to actually find out that. |If there are you know few if there are few categories for which we have in sin which are insignificant, had there is category which is insignificant it can actually we combined with the difference category, it can be combined with the difference category and the this category can be eliminated. If we have 2 category which are having similar coefficient similar value had similar coefficient value they have similar influence on the outcome variable or output variable, and those variables on those category can also be combined. So, much can be done after analysing the results.

Another technique that can be used for dimension reduction is classification and regression tree. So, in a coming in a coming session, we will discuss classification and regression tree in more details ah, but to give you an idea that there is we in this in this under this technique, we develop a classification tree for classification task and we develop a regression tree for prediction task. So, while we build this model using different using full of variables, the large number of variables, in result is going to be a tree diagram.

So, which would be represent which would be giving us the different classification rules or prediction values and will also be incorporating the important variables that would use to build that particular tree. So, if a variable does not show up in that particular tree diagram, does not figure in the in that particular tree diagram; that means, that particular variables can directly be eliminated. So, that is how dimension can also be reduced. So, we can start from a large number of predictors and then some of them can actually be eliminated if they do not figure in the tree diagram.

Similarly, in the classification and in the classification tree as well we can combined categories if there is similar kind of if they are coming in the in the in the in the same branch, and there is possibilities of having you know similar classification rules for both of them, and probably we can combined they we can combined them. Similarly for regression tree as well if a particular variable is not coming up in the regression tree diagram, and also those variables can be eliminated from the analysis.

So, more detail on these two techniques regression model whether it is linear regression or logistic regression, and how they can we used for subset how subset selection procedures, can be developed using these models and how classification and regression discard can be used for this model for dimension reduction, we will discuss when we will discuss this when we come to the lecture we will discuss this particular techniques.

Now, one difference between principal component analysis and these models regression models and classification regression tree is that these models they in they account for the relationship between the predictors and the outcome variable. So, whether it's a regression model that, the way linear regression is model, its the relationship between outcome variables and the predictors that is actually incorporated in the modelling process. Similarly for logist logistic regression also the, predictors and the outcome variables there relationship incorporated and any subset selection for this procedures that are used later on they are based on those relationships so, that being one big difference.

Similarly, in classification and regression tree also, while we develop the tree diagram it is the it is the relationship between the variables that help us reach to the terminal nodes are leaf nodes, and do our classification on prediction. Therefore, it is the underlying relationship between those variables that is also playing its role and determining the importants of variables. So, will stop here and in the next section will, I will discuss the performance matrix so.

Thank you.