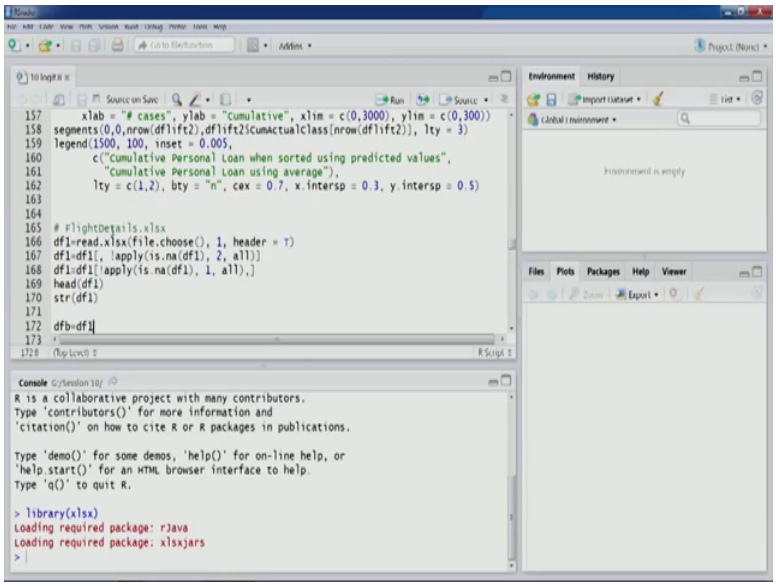
Business Analytics & Data Mining Modeling Using R

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

Lecture - 50
Logistic Regression - Part V

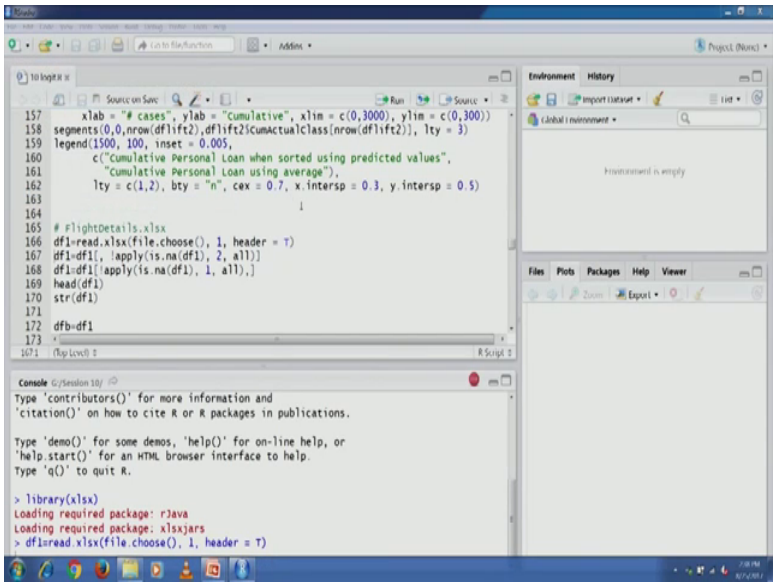
Welcome to welcome to the course Business Analytics and Data Mining Modeling Using R. So in previous few lectures we have been discussing Logistic Regression. So let us start our discussion from the same point where we left in previous lecture. So we were doing an exercise in R environment and we completed that exercise so that was using promotional offers; a data set. So what we are going to do in this particular lecture is we will use another data set and other problem and we will go through the complete analysis that is required in logistic regression.

(Refer Slide Time: 00:49)



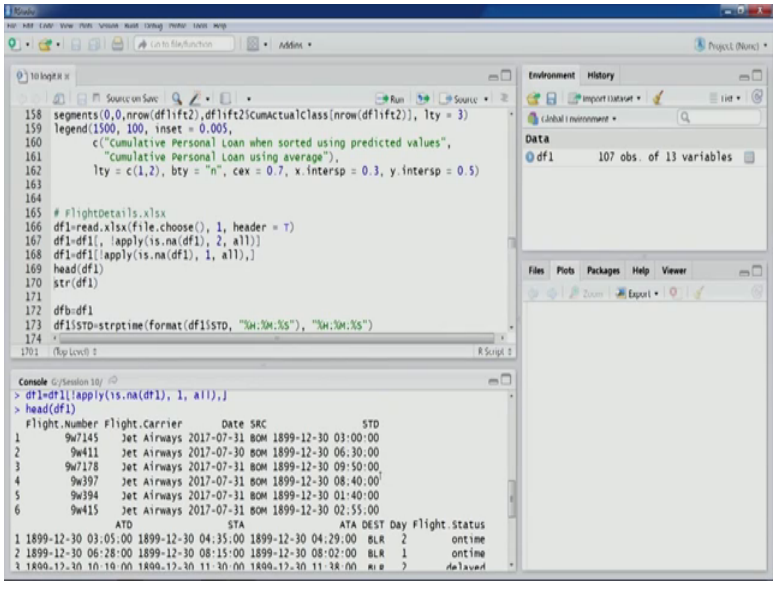
So the data set that we are going to use for this lectures exercise is on flight details. So this particular data set we have a used before as well. So let us so this is the data set; flight details dot xlsx. So, this particular data set have been used before. So let us load this, import this data set into R environment.

(Refer Slide Time: 01:26)



So we have 108 observation 13 variables; however, there are some na rows and na columns. So let us get rid of them. So let us first remove na columns then na rows and you would see 107 observation 13 variables. Let us look at the first six observations. So this particular dataset we had used before when we discuss naïve Bayes.

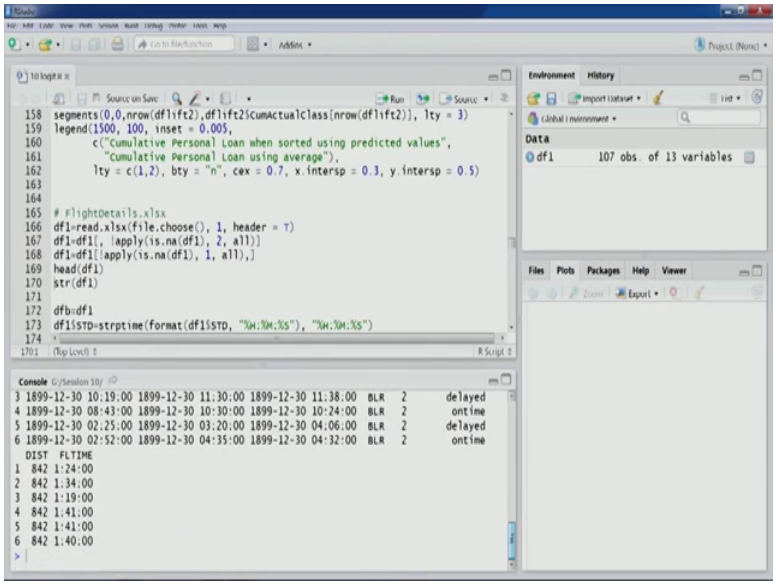
(Refer Slide Time: 01:50)



So now again we will understand; so this particular dataset most of the variables that we have they are factor variable, as you can see in first six observation. We have flight number, flight carrier, then date, then source and schedule time of departure, then we

have actual time of departure, then we have scheduled time of arrival, then actual time of arrival, then a destination, then we have day whether it was Sunday or Monday.

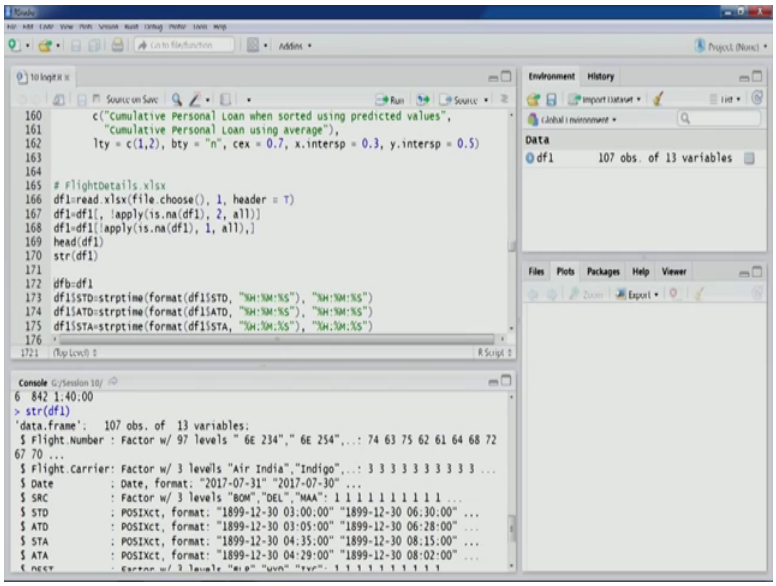
(Refer Slide Time: 02:16)



So we had information on just 2 days; data on just 2 days; then flight status and then two additional variables. So when we use this particular data set in a naïve Bayes.technique, we did not have these two variables; distance and flight time. So we have added these two variables distance between the source and destination and also flight time.

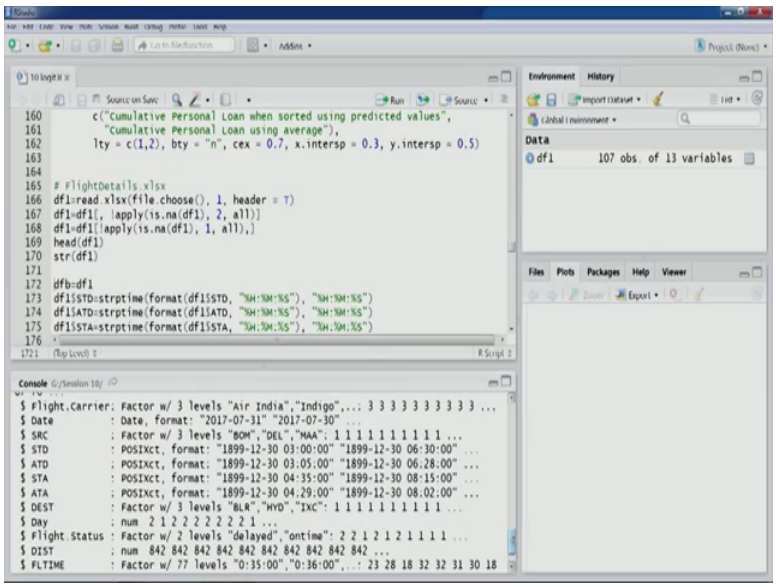
So this particular data set, so let us look at the structure and then we will discuss further. So let us look at this information. So, flight number if we see that the flight number is something that we are going not going to use in this particular analysis.

(Refer Slide Time: 03:07)



However, flight numbers for each of the you know flights that we have with us right; 97 of them out of 107 observations, some of them must be repeating. So then we have flight carrier so we have 3 carriers; Air India, Indigo and Jet Airways. So we will discuss them. Then we have date; so a data flight. We have flights on 2 days that is 30th July 2017 and 31st July 2017.

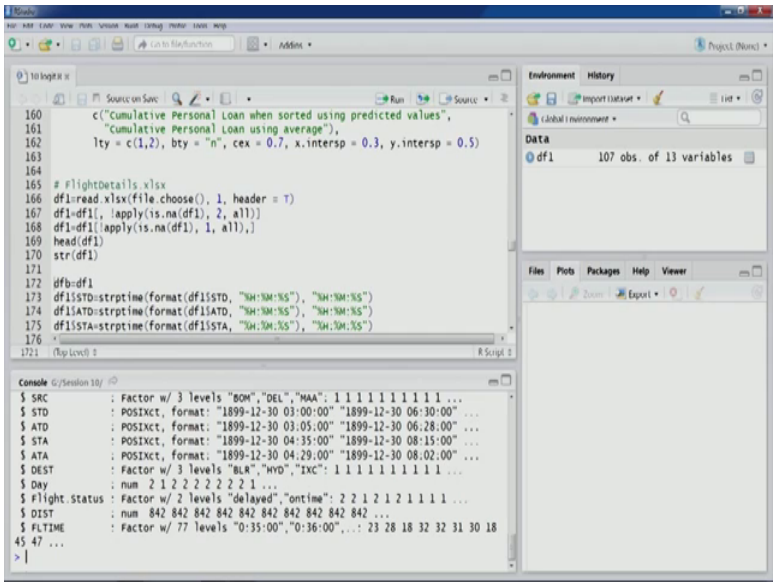
(Refer Slide Time: 03:35)



However, as we did in naïve Bayes.`we would like we would not like to use this information and date is not important for our analysis and we will consider them you

know we will not considered date and we look at the time intervals of this specific time intervals, departure intervals of flights. So the main problem remains same. So this is a classification problem that we are going to model using logistic regression modeling. So where we would like to predict the delays of flights right.

(Refer Slide Time: 04:30)



So other variables we are familiar with a day and flight status, distance, flight time. So these are the other variables flight time. As you can see this is right now it is Factor, so this has to be converted into a numeric variable distances flight status is also ok. However, we would like to change the labels, you can see this is a delayed and on time. However, since we are modeling for delayed, so we are trying to predict delays. So therefore, our reference category has to be on time and modeling has to be done with respect to delayed. So we need to change these labels for our outcome variable that is flight status. So we would like to predict the flight status whether particular flight is going to be on time or delayed; focus is on delayed. So the on time is going to be our reference category. Just like we used to have other techniques other examples, just like we used to have 1 and 0. So a focus was on always class 1; members belonging to class 1. So we always look to identify build a model which would be classify an observation into class 1.

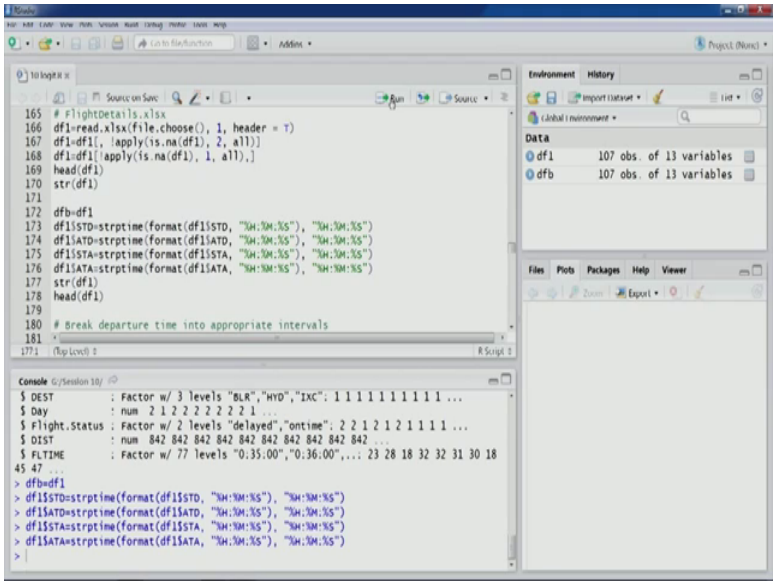
So, here in this case as I talked about we need to, we would we require to change the levels; destination is ok, day this also we would have to change. So day is factor variable,

2 days; Sunday and Monday. So right now this is numeric variable, so we will be required to change this variable as well. The schedule departure you know departure time actual time of departure and these four information only will a try to derive another variable using these four information.

So flight status has been derived already been derived using some of these variables and we will derive another variable that departure time interval using some of these variables. Other after those derivation after those variable derivation we would not be using these variables.

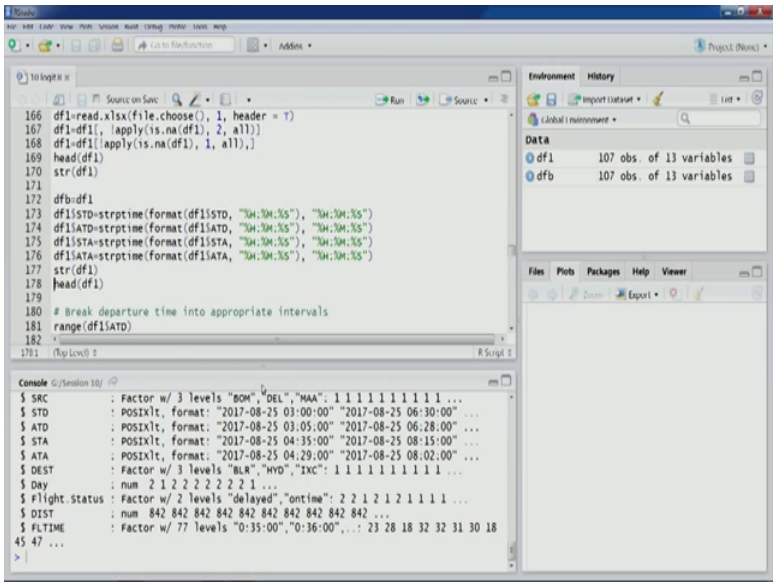
Source is appropriately mentioned factor, 3 levels date we would not be using, flight carrier we would we would be using appropriately mention, flight number again we would not be using. So let us start some of these transformation. So before we go ahead, let us take a backup of this particular data set. So we will take a backup and then we would because as we said, we are not interested in actual dates of those flights.

(Refer Slide Time: 06:55)



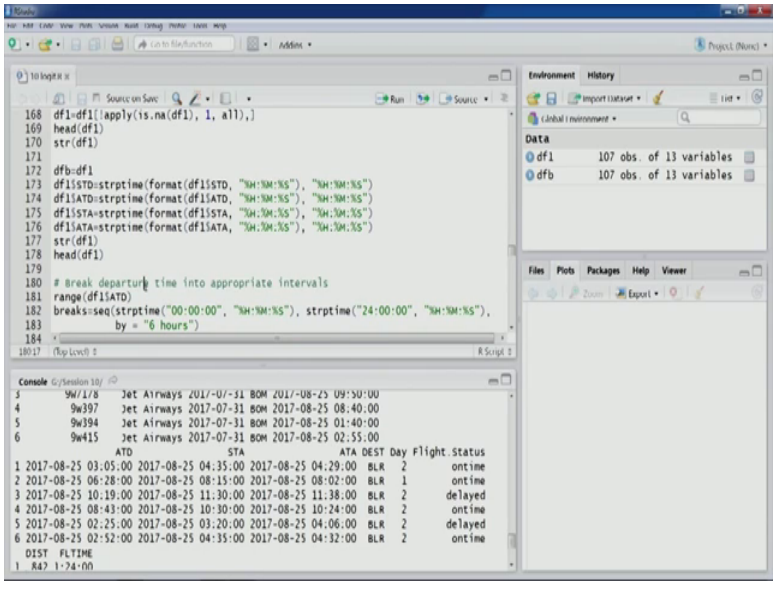
So, I will like to change the same because we need to do, we need to use these particular columns for certain variable derivations. So we would like to change these dates to, so that it is it comes out to be the same date for all the flights and therefore, various derivation that we want to perform there are no issues in that.

(Refer Slide Time: 07:23)



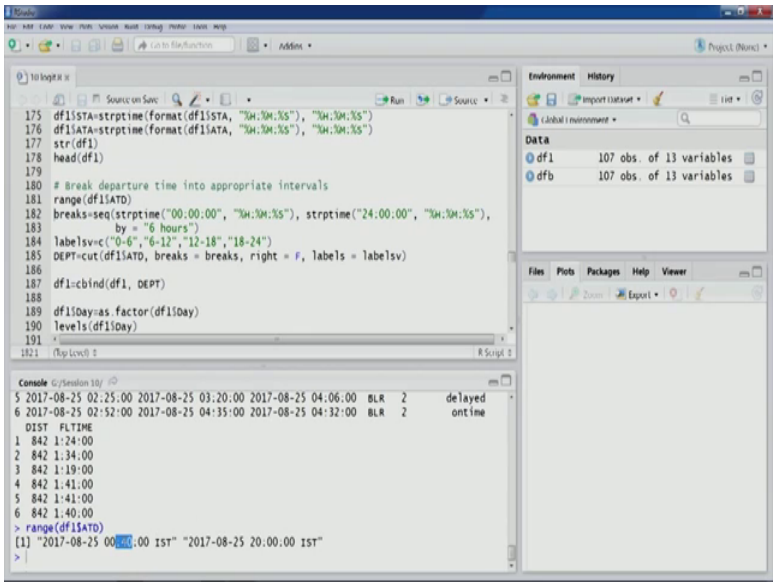
So, now you can see now in all these 4 dates, they have the dates has been changed. Earlier, it was earlier it was 1899; so this particular aspect we have already discussed during new base why 18 under 99 this particular date was coming there. Now this is what we require before we go for further variable transformation. So with this, so first six observation not much change only these four variables have been changed, dates have been change as you can see.

(Refer Slide Time: 08:00)



Now, let us first variable transformation that we are going to perform is on departure time. So, we would like to break departure time in to appropriate intervals. So for example, let us look at the range now because now actual time of departure now. It is now you know we have excluded that date information now for all the observation we are using the same date. So therefore, range can be appropriately captured for our analysis.

(Refer Slide Time: 08:30)

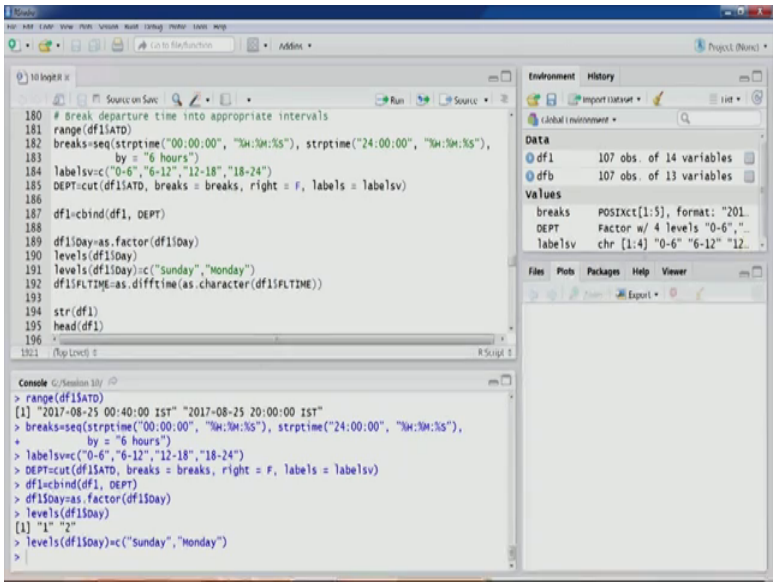


So, we have a flights ranging from flights ranging from flights ranging from 12:40 pm to 20 hours right. So, 40:00 hours 40 minutes to 20 hours right; so this is the range. Now we would like to break this time into appropriate intervals. So as we have done during naïve Bayes.we can see we would like to break this particular departure time into 4 intervals, you can see the labels also 0 to 6, 6 to 12, 12 to 18 and 18 to 24.

So, 24 hours time format; you would like to create four categories of that from 0 to 6 and 6 to 12 and 12 to 18, 18 to 12. So for this we would be requiring this particular breaks variable which would be used in the cut function as you can see in next line of code; this one that breaks is a breaks this particular variable is going to be used here and it will cut the different observation in actual time of departure using these breaks.

So let us compute this. So you would see in environment section, we will have breaks a variable and 5 values are there and let us also create labels.

(Refer Slide Time: 09:48)



So these breaks can be used to create these 4 categories. So, let us execute this. So you would see that depth this variable has been created with it is a factor with 4 levels right. So, we get it this particular variable into appropriate formats. Let us append this particular variable into the data frame. .

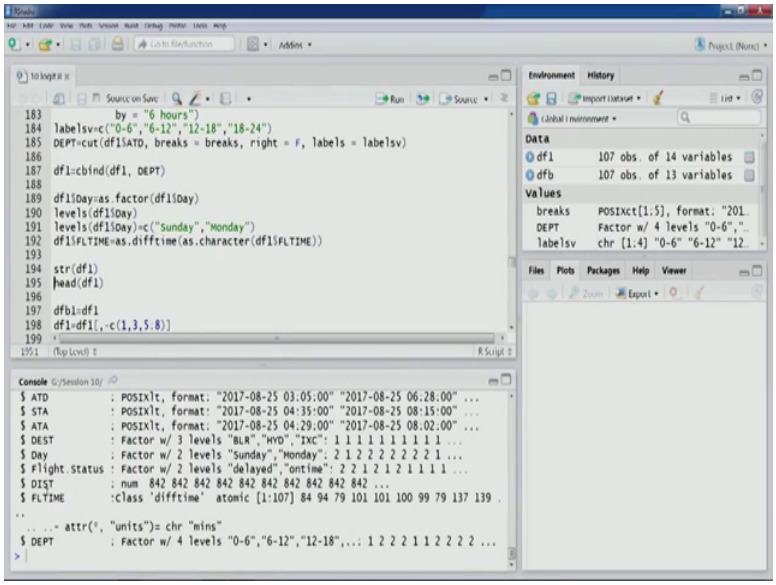
Now let us focus on other variables. So we had noticed that day was also you know day was we wanted to be a factor variable, categorical variable having a Sundays and Mondays flights on 2 day. However, this was stored as numeric. So let us change it to factor. Let us look at the level 1 and 2.

So we also would like to change these level names instead of 1 and 2 we would like to have Sunday and Monday; specifically specified. So we will do that. Once this is done we will focus on another variable that is flight time. So flight time that information we had in that time notation as hours colon minutes colon seconds. So, we would like to change it into a format that could be useful for our analysis. So, we would like to change them all those that this file flight duration flight time variable into minutes. So all those values we would like to convert into minutes.

So this is the function that can be used as dot difftime. So because these are you know time intervals, so their differences between two times. So this particular function as dot difftime can be used and you would see that we would be able to change it into minutes.

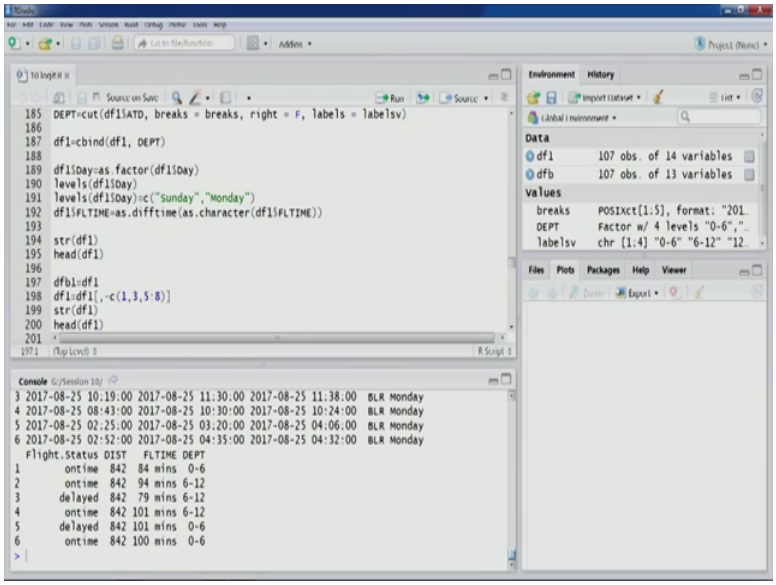
So first you would see that this particular variable has been changed to this. Let us look at this structure; so if you go to fl time you see that class difftime has been created.

(Refer Slide Time: 11:51)



So this is now atomic factor and now all these values they are minutes right. So you can see character mins here this information, so all these values that were there time intervals. Now they have been converted into minutes.

(Refer Slide Time: 12:14)

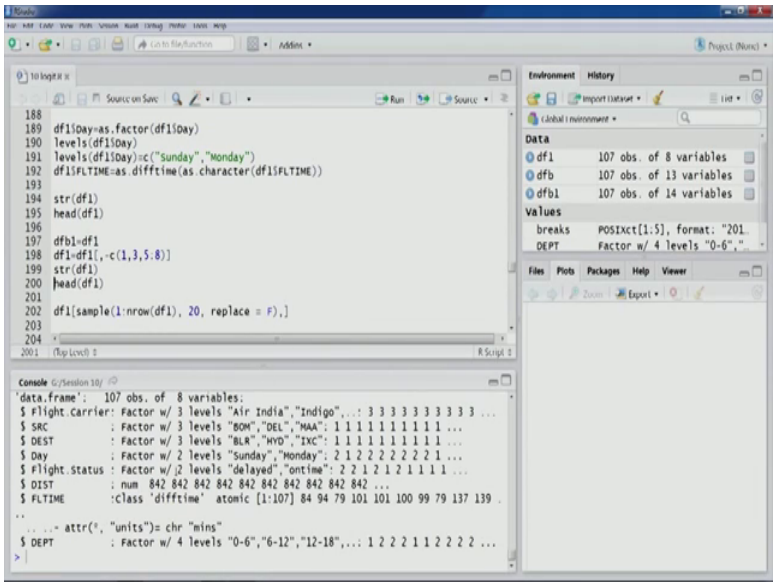


So with this almost we are almost there. So you can see that flight time; now you can see 84 minutes, 94 minutes, 79 minutes. So all the flight time duration have been

appropriately converted. We have created the departure variable, each of the flight has been correctly labeled as per it is newly created category in departure and so let us focus on some other transformations; so before that a let us take a backup. .

Now as we have talked about that some of these variables we would not be considering. So for example, variable 1 that is flight number, then date that is 3 column number 3; we would not consider, then from column number 5 to 8. So these are the actual dates scheduled time of departure, actual time of departure, scheduled time of arrival and actual time of arrival. So these also, these 4 columns also we would not be taking into model.

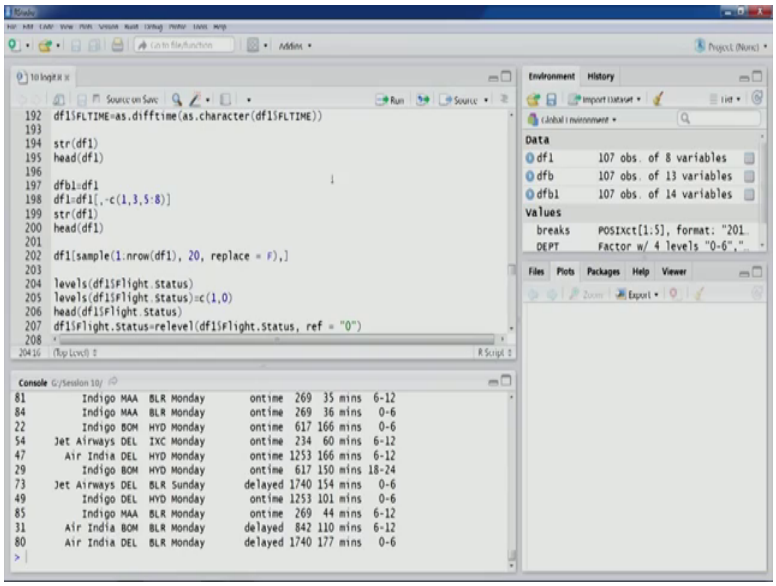
(Refer Slide Time: 13:17)



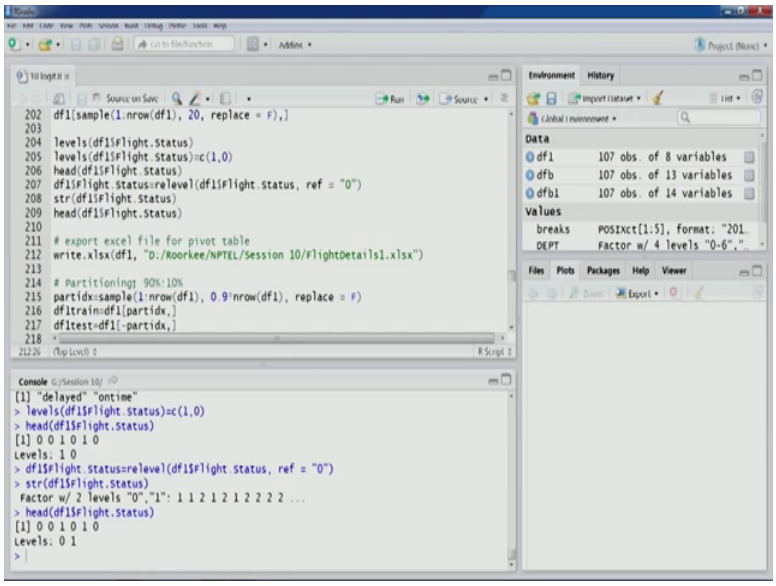
So let us get rid of these variables, these columns. Now so these are the variables that we are left with and also they are in appropriate format. So we would like to use these variables in our logistic regression model. So we now we have flight carrier, source, destination, day, flight status, our outcome variable; however, we need to change the levels as we talked about; distance and we have flight time, minutes and then we have this departure interval.

So once this is done and these are first six observations. If we want to take you know any random sample of 20; so this is sample function can be used in this fashion. So this particular command will give us 20 randomly selected rows. So we can have a look at different values for different observations randomly drawn observations.

(Refer Slide Time: 14:12)



(Refer Slide Time: 14:19)



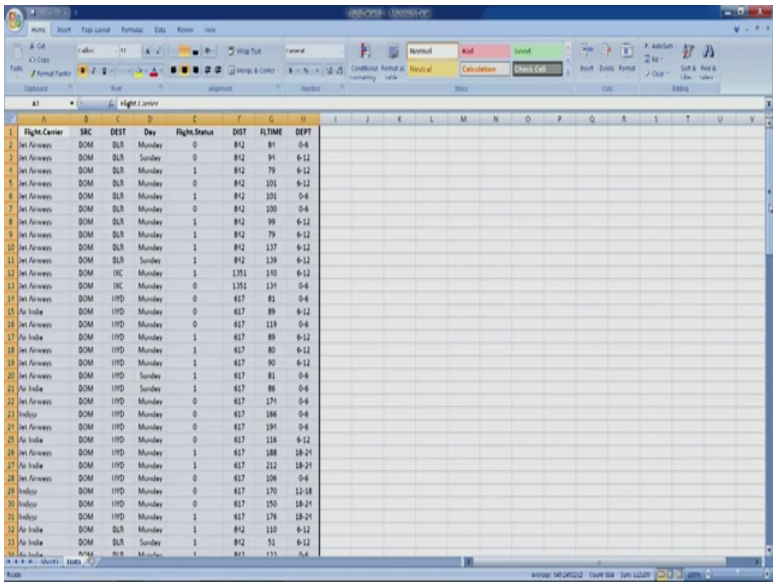
Now, let us focus on the outcome variable. So outcome variable, the levels are delayed and on time in that particular order. Now we would like to change this into numeric code because later on the different later on our code would be a much easier to write, if we have the numeric codes right because we would be creating lift curve and cumulative lift curve we would be we would require numeric code, so that we can create the cumulative actual class. So all those things for all those things we would prefer this level for outcome variable to be 1 and 0.

So let us change this. So 1, here as you can see; 1 is corresponding to delayed and 0 is corresponding to on time. Because our task is to predict delayed flights right. So therefore, one has to be assigned to delayed right. So let us execute this. Let us look at the first six observation of this outcome variable, flight status; you can see levels have changed now 1 to 0. However, if you see that ordering is 1 and 0. So 1 would become reference category and that means the rate would become reference category. So we did not want that. So we would like to change this. So re level function can be used to perform this kind of change you can see. In the re level first you need you need to pass on the factor variable and then the second argument is for the reference. So, we can select the reference category here, the other things would be appropriately changed.

So, I be execute this code and look at the structure of this particular variable. Now you can see 0 and 1; in the correct order and 1 representing now delayed and 0 representing on time flights. So now, this particular variable; let us look at first six observations also. Now this particular variable is in the desired you know state as we want it to be. Now at this moment if we want certain analysis; descriptive analysis, we can do so as we did in naïve Bayes.as well. We can once this data is ready for logistic regression model, all the key variables are there and the data frame; final data frame. Then we can write this particular file into our disk and then we can apply you know pivot table excel, base pivot tables to write some summary to for further analysis.

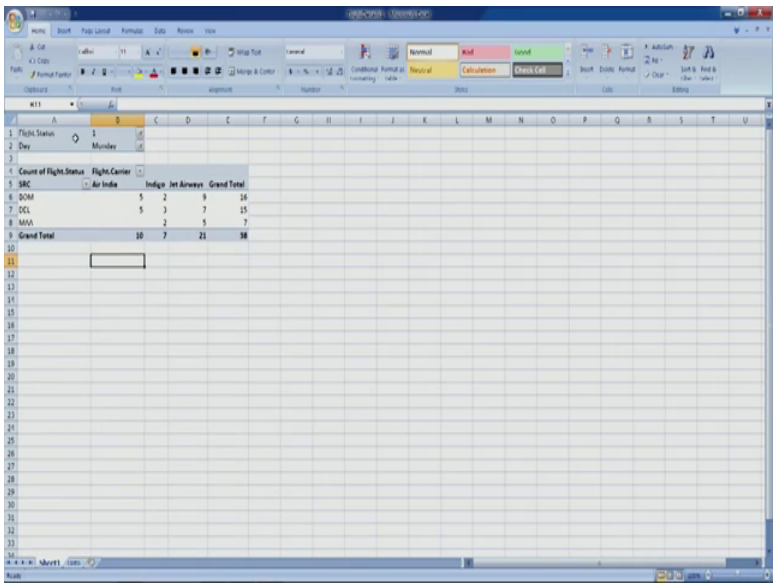
So let us, so I have already created one pivot table. So let us open this. So this was the data; pivot table this particular exercise we have already done.

(Refer Slide Time: 16:58)



So we have to just select the data and you know get it into appropriate presentable format and then select all the all the columns, all the rows and then create pivot table using the insert tab and within that this pivot table option and the pivot table would be ready for us in this fashion.

(Refer Slide Time: 17:11)



So, I have already selected few of the important variables here, in appropriate filters. So you can see in row levels I have, in row level I have SRC that is source; then in column level I have flight carrier those three flight carrier that we have; Air India, Indigo and Jet

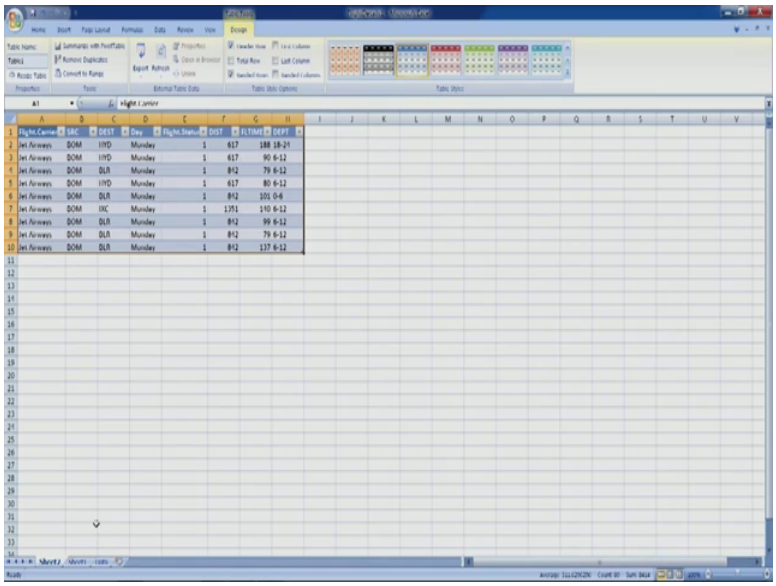
Airways and three source we have a that is BOM, DEL and MAA that is Mumbai, Delhi, Mumbai airport, Delhi airport and Madras airport. And then we have in the values areas we have account of flight status that means, how many flights are actually delayed.

So how many flights are there; count of flight. So overall, now however, in the report filter I have flight status and day as you can see here. So these are report filters point. So flight status you can see I have pre selected one. So this can be changed if we click on this filter you can see all the flights count of all the flights or delayed flight 1 representing delayed. So therefore delayed flights and 0 representing on time so that can be done. .

In day also we have a filter. So this is since this is report filter, so we are going to have a filter for all the variables. So Sunday and Monday you can see here that we can have all you know all days Sundays and Monday's together total and then either Sunday or Monday. So those numbers would be reflected. So we these this particular descriptive table will change depending on the change in levels of report filter variable.

So let us look at some of these numbers. So as we can see that when we are interested in finding the count of delayed flights, you can see the flight status is selected as 1 right here; in the filter flight status is selected as 1. Therefore, the we are interested in understanding the delayed flights from these descriptive statistics, summery statistics. So we can see that then off further we have selected Monday. So delayed flights and then Monday; so these are the numbers. So we can see when the source is Mumbai airport that is BOM, then we can see total number of delayed flights are 16 and you would see Jet Airways more number of flights of Jet ways Jet Airways are delayed.

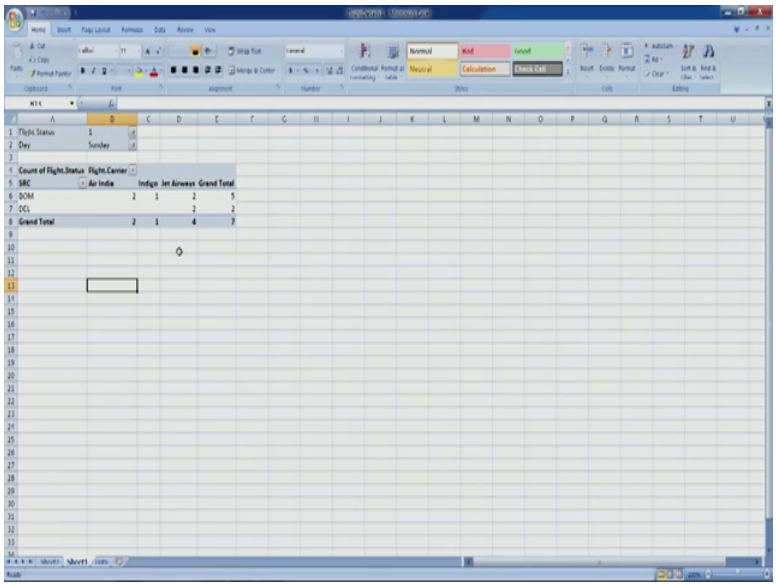
(Refer Slide Time: 19:56)



So in this table once you click you will get the specific observations as well right. So now if we look at, if we look at the second row we see that Delhi. So overall total number of delayed flights for Mumbai and airport both these airports. So these airports are too busy airports in our country and you can see number of delayed air fights are similar number is there. And Madras airport we have just 7 delayed flights; however, across 3 carriers we see that Jet Airways we have more number of flights; in terms of number we have known more number of delayed flights.

So, in terms of just the number right; in terms of just the number and this information is for Monday. So now what we can do is a because with respect to delayed flights, let us look at the what happens during Sundays. Let us select Sunday and we do; so now you would see during Sundays we did not have any flight in our data set originating from Madras airport; so that is gone.

(Refer Slide Time: 21:04)



So that is also one problem in modeling exercise as I have I have talked about in previous lectures as well; that if some of the combinations are not covered then that could be a problem when we go about predicting new observation because if that observation falls into that zone and in our training partition that was not covered; some of those combinations were not covered in our training partitions then prediction would not be possible.

So the same thing is reflected here. For example, Sunday we do not have resource destination MMA Madras. We do not have any flights from that source and if we look at the number of flights are quite few in comparison to Monday. However, if we look at again more number of delayed flights or from Mumbai and again more number of delayed flights are from Jet Airways. So, in terms of numbers total number of flights are also less and during Sundays and total number of therefore total number of delayed flights are also less.

So, if we look proportion wise then the that does not seem to be much difference between Sunday and Monday. So the same thing we should expect when we build our logistic regression model, the same thing we should be expecting in our results as well; that there in terms of you know we look proportion wise.

Then there is seem to be any difference in delayed flights, you know whether it is Sunday or Monday. The similar kind of exercise can also be performed for on time

flights. So we can go to flight status report filter, we can select 0 and that would represent the on time flights. Now we can see the Sunday numbers here and Bombay 3 flights on time, Delhi 2 flights on time and the Madras 1 flight on time. So we did have a flight on Sunday from Madras, but that was on time.

So we can change the day filter as well to a Monday to see the difference. So we expect more number of flights during Monday. So that is the case. So we see that a more number of flights; 15, 22, 19. However, if we really look at these numbers during Mondays we see that there are more number of flights, there are more number of flights from Indigo which seem to be on time. However proportional wise our number of flights that they are running on Mondays are also more; are also higher, so that could be the region.

So this kind of analysis we can do using pivot tables and that would give us some insights when we go into former technique like logistic regression, what we should be expecting. So this kind of analysis can also help us in grouping as we have been talking about in previous lecture; grouping some of the categories on this we would be able to understand that which source destination can be grouped or which days can be grouped here, we just have 2 days. So therefore, the question would be whether we should incorporate day at all? If it is, the if does not seem to be a significant predictor for delayed flights as per the descriptive stats that we saw.

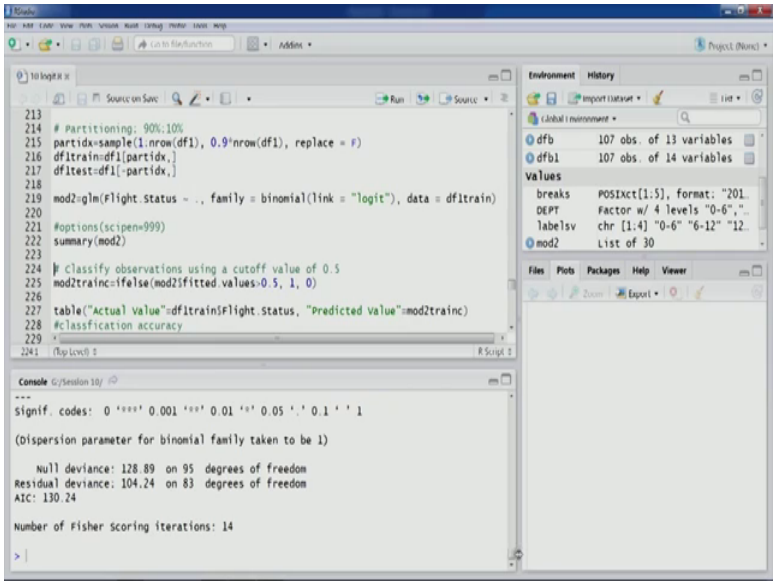
So these kind of decisions can be these kind of insights can be derived from descriptive stats. So what we will do? We will go back to R environment and we will move to our next step that is a partitioning. So because this particular data set is quite small; just 107 observations and for training partition you would like to have more number of observations; so that the model is slightly a more stable, more number of combinations because we are using a factor in many.

There are you know majority of the predictors that have they are factor they are categorical. So there you know there are going to be more combinations of values that you would like to cover in our model. So more observation we would like to have in the training partition because of this small sample size.

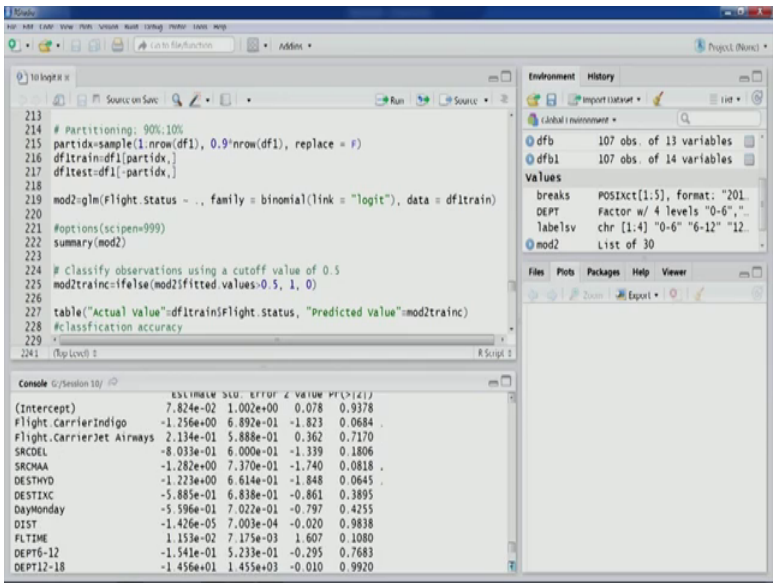
So let us do this partition partitioning; 90 percent for training partition and the remaining 10 percent for testing. You can see our partitions are created; 96 observation in the

training partition and 11 observation, remaining 11 observations in test partition. Now as we did in previous lecture, the same function glm can be used. Flight status is our outcome variable, other variables are going to be going to be used as predictors other things remain same. So let us run this code. You get the model. Let us look at the summary.

(Refer Slide Time: 26:11)



(Refer Slide Time: 26:16)



So if we look at the results of this logistic regression model that we have just created. We can see the results most of the variables because of this smaller data size, we do not

see much significance. However, we do see three variables to be significant at 90 percent confidence interval. You can see small dot here that is for 90 percent confidence interval. So, these three variables seem to be significant; first one is Flight Carrier Indigo.

So, the so the estimate is negative. So this seems to be significant at 90 percent confidence interval and so therefore what we can understand is with reference to Indigo, with reference to Air India which is the reference category for flight carrier; Indigo of flights from Indigo flights from indigo carrier. They seem to be less delayed because remember this particular modeling exercise with respect to delayed flights.

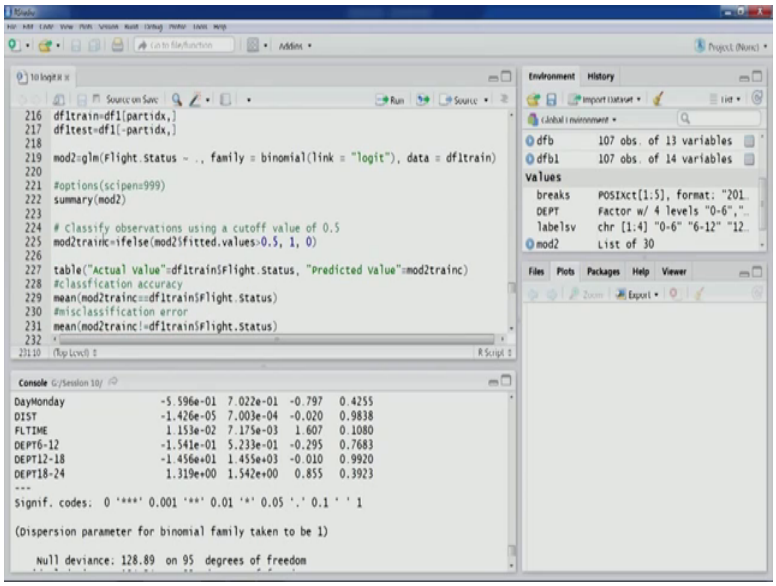
So therefore, if the coefficient is negative, so therefore we should expect that logit values are going to be that logit value there is going to be less cost, there is going to be decrease and therefore there is decrease in probabilities value and therefore there is more chance for it to be on time rather than delayed.

So, in this fashion we can interpret. However, the append be interpret these results. We should also look at first whether the variable is significant. So we have to check the significance as per our accepted level of confidence interval. We can look at the significance and once the variables are a significant then we can look at their coefficient values to understand the level of impact that they have. .

So flight from Indigo they seem to be less delayed or more on time in comparison to Air India right. The same cannot be said about Jet Airways because this is insignificant relationship. Now, other very other significant coefficient is source that is from Madras airport. So it seems that flights which originate from a Madras airport they also seem to be you know a less delayed right so more on time.

So this is also again at 90 percent confidence interval. Similarly we have another significant variable that is destination Hyderabad. So this is dummy code for Hyderabad. So airport, so we can see here also that the flights which arrive at Hyderabad; so where the destination is Hyderabad. They seem to be less delayed or more on time with respect to our reference category. Same is also that Madras airport was also with respect to the reference category.

(Refer Slide Time: 29:19)



Now other variables we can see they do not seem to be significant. For example, day Monday; so as we saw in our pivot table that a proportion wise there did not seem to be much difference on in on flights during Mondays or Sundays. So this also, in the results also logistic regression model results also this particular variable dummy variable does not come out to be significant.

So we can see that what we expected from our pivot table exercise, is the same thing is reflected here in the model. A distance this also is highly insignificant; so distance is not a the key predictor here. So this can also we so from this we can also understand which variable can be dropped and another modeling could be done. For example, distance is highly insignificant. So probably this variable can be dropped.

However, as I have pointed out in earlier lectures also, in data mining modeling our goal is prediction. So we are not even if a particular predictor is insignificant, but it is of practical importance in terms of predicting tasks, you know prediction or classification tasks or other data mining tasks; we would still like to keep it in the model.

However, in this case distance does not seem to be of much practical importance and highly insignificant. So probably this can be dropped. Flight time we can see that it was you know quite close to being significant at 90 percent confidence interval. So probably a flight time is also quite practical in terms of predicting delayed flights. So this we should anyway keep there and the and the then this next one is departure time interval.

So, with respect to the reference category that is 0 to 6 and these 3 departure interval they do not seem to be significant.

So probably at the departure time intervals also do not matter with respect to the data set that we have. So what we can do? We can look at the another modeling approach. So from this model we can further understand which variables are important; which variable are insignificant and if they are significant what is the impact that they have and whether another model with only the important variables can be done. Even if a particular variable is insignificant as we talked about it can still be kept in the model, if it is it provides some practical importance. So with this we will stop here and we will continue our discussion on this particular modeling exercise on flight delays.

Thank you.