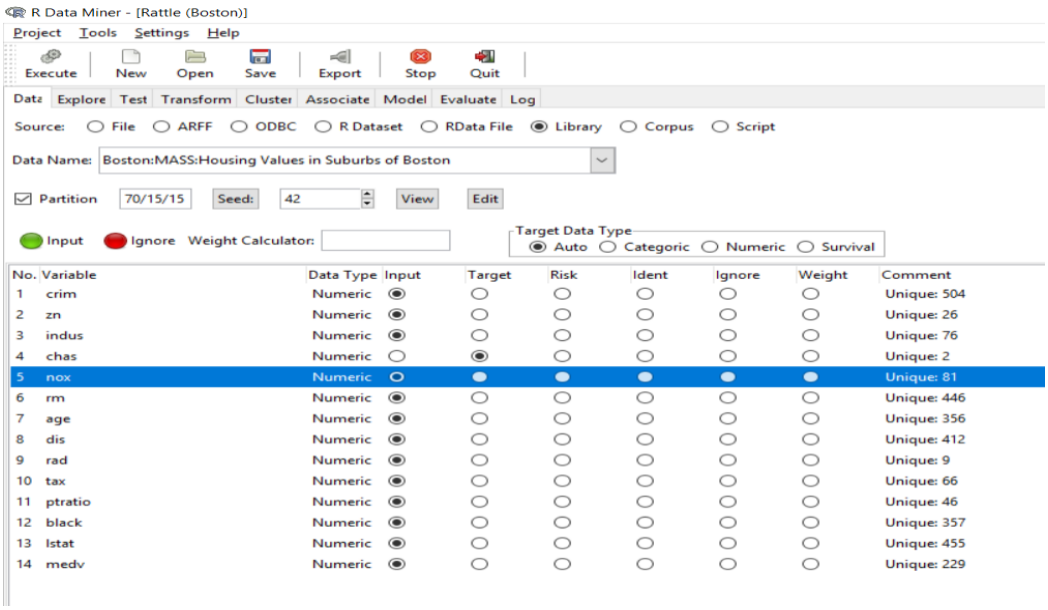
Read Data

Action 1: Select the data (Boston:MASS:Housing Values in Suburbs of Boston) from the dropdown menu called Data Name.

Action 2: Press “Execute”, which will show you the variables in the dataset.



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	crim	Numeric	◉	○	○	○	○	○	Unique: 504
2	zn	Numeric	◉	○	○	○	○	○	Unique: 26
3	indus	Numeric	◉	○	○	○	○	○	Unique: 76
4	chas	Numeric	○	◉	○	○	○	○	Unique: 2
5	nox	Numeric	○	●	●	●	●	●	Unique: 81
6	rm	Numeric	◉	○	○	○	○	○	Unique: 446
7	age	Numeric	◉	○	○	○	○	○	Unique: 356
8	dis	Numeric	◉	○	○	○	○	○	Unique: 412
9	rad	Numeric	◉	○	○	○	○	○	Unique: 9
10	tax	Numeric	◉	○	○	○	○	○	Unique: 66
11	ptratio	Numeric	◉	○	○	○	○	○	Unique: 46
12	black	Numeric	◉	○	○	○	○	○	Unique: 357
13	lstat	Numeric	◉	○	○	○	○	○	Unique: 455
14	medv	Numeric	◉	○	○	○	○	○	Unique: 229

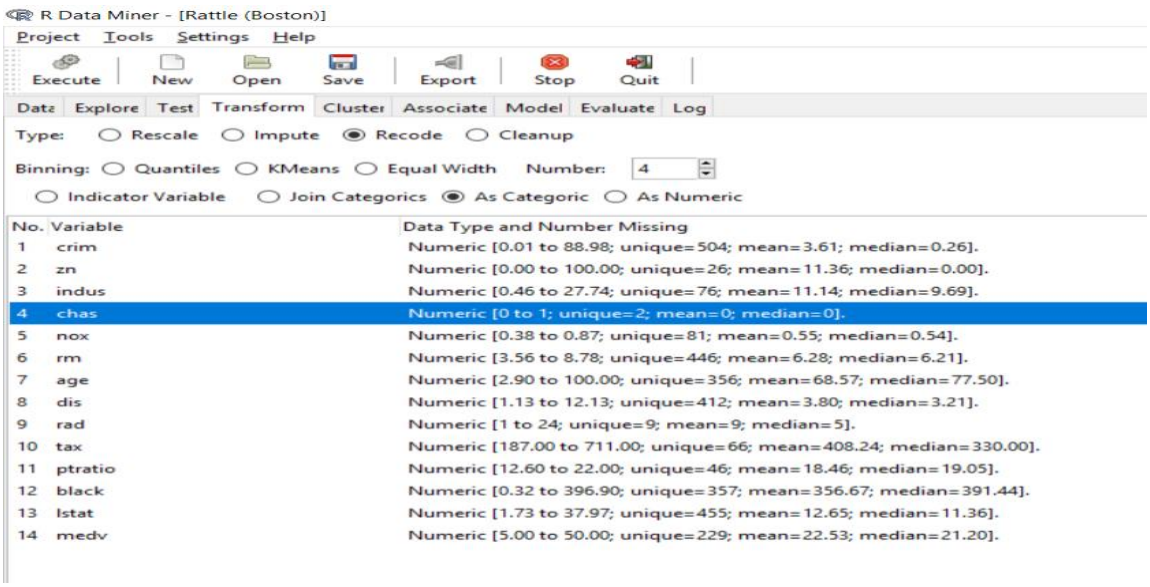
The variable CHAS is currently coded as **numeric** data. We need to convert it into **categorical** data.

Convert from Numerical to Categorical

Action 3: Go to the Transform tab.

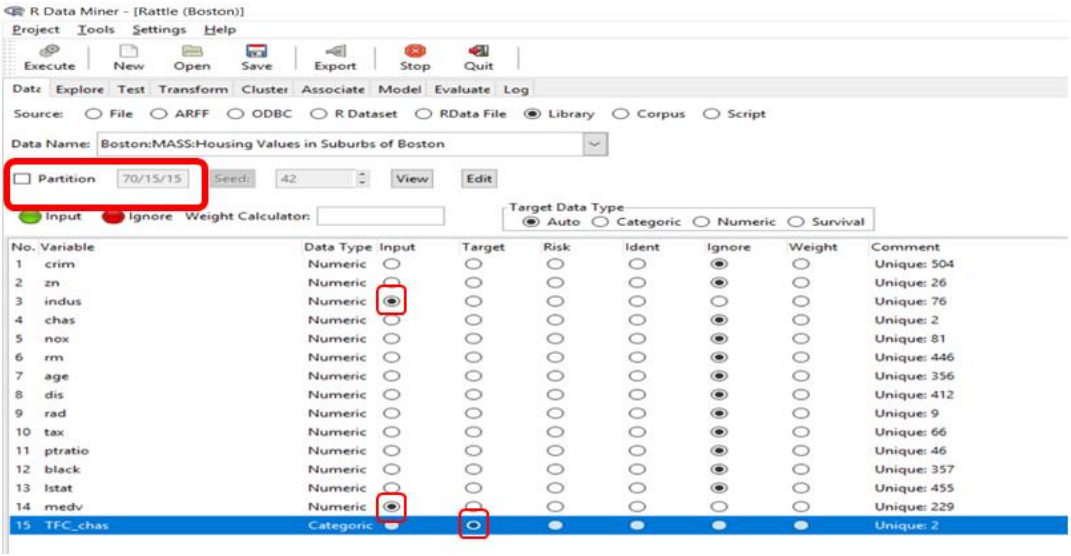
Action 4: Select Type→ Recode→ As Categoric

Action 5: Select the desired variable (CHAS in this case) and then Execute.



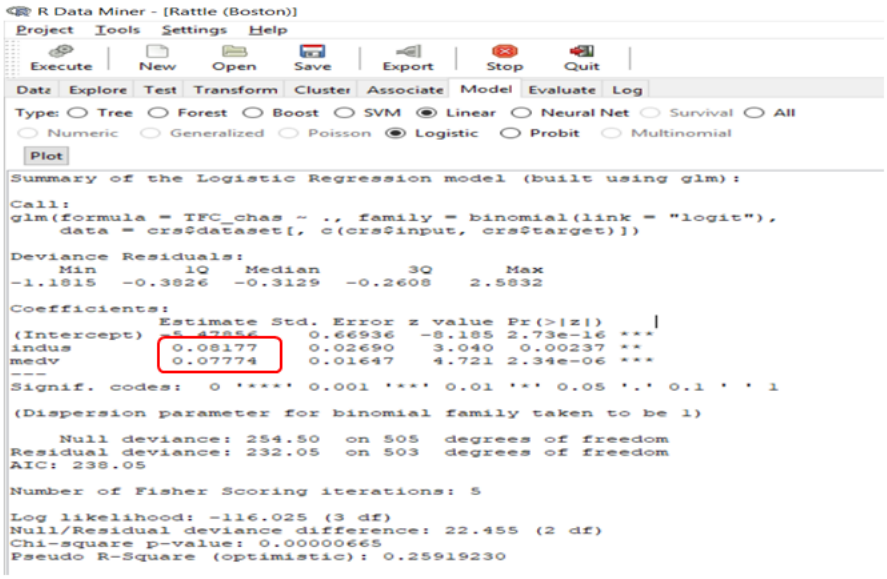
Action 6: Go to the Data tab again and uncheck the box of Partition (as specified in the assignment: “do NOT partition the data”).

Action 7: Make indus and medv as the Input variables and TFS_chas(converted into categoric variable) as the Target variable. Ignore the rest of the variables. EXECUTE



Action 8: Go to the Model tab.

Action 9: Select Type:→ Linear →Logistic



Checking Coefficients – Marked in red above

Error Matrix for the Training Data

Action 10: Go to the Evaluate tab.

Action 11: Select Type: →Error Matrix and Data:→ Training

```
Error matrix for the Linear model on Boston [**train**] (counts):

      Predicted
Actual   0 1 Error
     0 469 2   0.4
     1  33 2  94.3

Error matrix for the Linear model on Boston [**train**] (proportions):

      Predicted
Actual    0   1 Error
     0 92.7 0.4   0.4
     1  6.5 0.4  94.3

Overall error: 6.9%, Averaged class error: 47.35%
```

Deliverable 2

We have fitted a model to predict a response variable with two classes (Yes/ No).

Interpretation -
Regression Model -Both the predictor variables "INDUS" and "MEDV" predict higher probability that the tract will bound Charles river.

Error Matrix - The error rate is around 7%. However, the error rate across two classes is quite different. For positive class, the prediction error is close to 94%, while for the negative class it is merely 0.5 %

Deliverable 3 - This is an example of a possible response. The actual answer may be different and yet correct. The idea is to discuss some of the limitations of the data and/ or model.

The huge difference between the overall error rate and average class error rate is indicative of asymmetric prediction accuracy across the two classes. This may be due to the fact that there are very few observations with CHAS =1 in the data. There are only 35 Chas =1 out of 506 observations i.e., only 6.9%.

We could have achieved the an error rate of 6.9% by simply predicting Chas =0 for everyone. This would have an error rate of 6.9% only. However, what is the value of such a prediction.

Thus, we need to think about
How to improve the model?
Do we collect more data? of what type?
Would collecting more variable help?