

**Student:** Xiaomin Wu

**Week project:** week 36, Logistic Regression

### Data preparation for the training:

Dataset creation:

First of all, modify data to current/useful format, change all semicolons (;) to colons (,). I also modify the format of 'Sales Rating' to binary, change 'Good' to '1', 'Bad' to '0'.

As last time, import data to pandas data frame, set weekday and seller as X and sales rating as y, set 80% of data as training set, rest of those as test set.

Then save the original format data before dummy variables, that makes all variables change to binary.

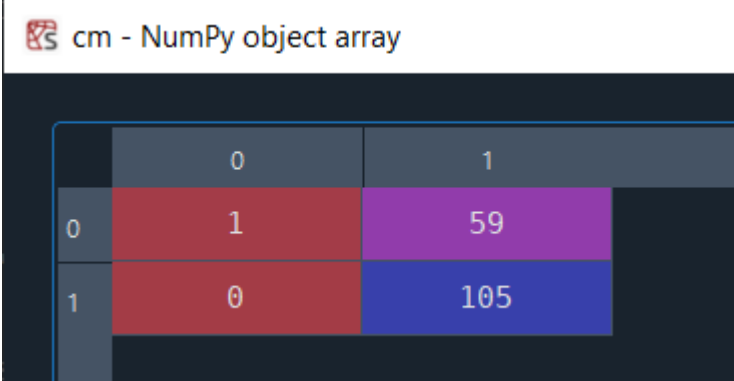
Dummy variables: In this case, I dummy all X values, keep all seller name as dummy variables, remove one weekday dummy column to avoid the 'dummy variable trap'.

Scaling: there isn't any data scaled in this case.

### Relevant metrics for the case:

Accuracy is 0.6424242424242425

The confusion matrix is:



cm - NumPy object array

	0	1
0	1	59
1	0	105

(I was tried to calculate r2 score, like linear regression, but it's negative. Then I realize it makes no sense in logistic regression.)

### Conclusions of the results:

It's not a good model for this case, since there is only 64% accuracy. As we mentioned about how many samples are necessary,  $(2 \times 10) / 0.2 = 100$ , so there are already enough numbers of data. That mean the low/unexpected accuracy is not because of leak of data. So in my opinion, the sales rating is not so strong logistic relate with both weekdays and seller.