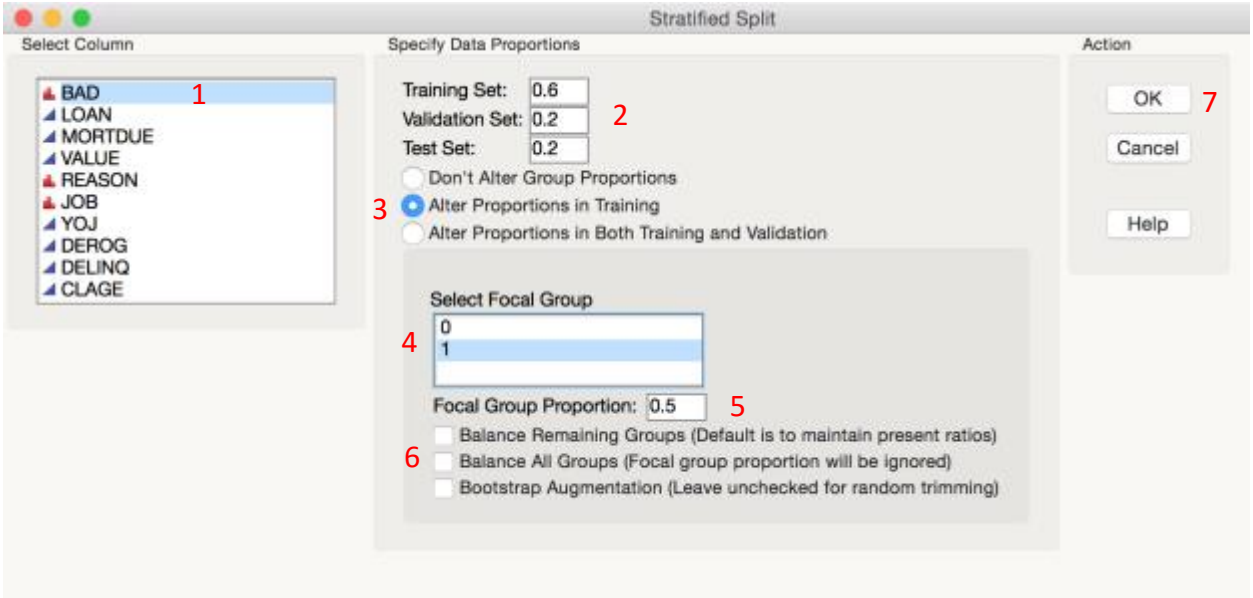
Stratified Data Partitioning (with balancing options) Add-in

This add-in can be used for oversampling when the target variable class of interest is underrepresented or rare. The add-in allows you to:

- Partition data into training, validation, and test sets, stratified according to a target variable
- Balance the training (or training and validation) set to have equal numbers of the focal and non-focal groups

Instructions for using the add-in correspond to the numbered sections below:



1. **Select the categorical target variable**. This variable can be binary, or can have multiple classes or groups

2. **Specify the data proportions for the training, validation and test sets**. Note that after balancing, the proportions in the balanced sets will not match the specified proportions (observations will be trimmed or added to achieve the balancing – see step 6 below). The proportions in the unbalanced set will match the specified proportions (relative to the original data set).

3. **Select the sets to be balanced**:
   - *Don't Alter Group Proportions*: Creates training, validation and test sets in the specified proportions, stratified by the selected target variable. Does not balance any of the sets.
   - *Alter Proportions in Training*: Balances the training set according to the *Focal Group Proportion* specified in step 5 and the options selected in step 6. The original

proportions of the focal and non-focal groups are maintained in the validation and test sets.
- *Alter Proportions in Both Training and Validation*: Balances the training and validation sets, and maintains the original proportions of the focal and non-focal groups in the test set.

4. **Select Focal Group**: Select the value of the target variable of interest (generally the rare or under-represented group).

5. **Focal Group Proportion**: The training (or training and validation) sets will be balanced. By default, 50% of the observations will be from the focal group and 50% from the non-focal group(s).

6. **Additional balancing options**: The first two options relate to how to handle balancing when there are more than 2 groups (see the end of this document for an example contrasting these two options). The third option relates to how the balancing is accomplished (by dropping observations or by adding bootstrapped observations).

    - *Balance remaining groups*: Each non-focal group will have the same number of observations.
    - *Balance all groups*: The number of rows in all groups will be the same. The proportion of focal rows to the count of all other rows is p : (1-p), where p is the proportion specified in step 5.
    - *Bootstrap Augmentation (or Leave Unchecked for Random Trimming)*: This option specifies how the balancing is accomplished (recall that observations are trimmed or added to achieve balancing).
        - *Random Trimming (Default)*: Data are first partitioned according to the proportions specified (in step 2). Then, observations in the training (or training and validation) set are trimmed randomly so that the number of focal rows represents the specified proportion (step 5) of the total number of rows.
        - *Bootstrap Augmentation*: The observations are bootstrapped (and added to the data table) so the number of rows in the focal group in the training (or training and validation) set achieve the specified ratio of focal to non-focal rows (from step 5).

7. **Click OK to run**. The script produces the following:
    - A *new data table* that has been balanced according to the specifications selected.
    - *Distribution output of the target variable (Y) by validation* – this allows you to see the balancing of the target variable across the training (or training and validation) set
    - *Distribution output of the validation set (Y) by the target variable* – this allows you to see the partitioning of the data for each of the classes of the target variable. Note

that after balancing, observations will either be trimmed or added (to the training or the training and validation sets). As a result, the proportions for the balanced sets will not equal the proportions specified in step 2.

Example: *Balance remaining groups* vs *Balance all groups*

In this example we use a 5-nomial setting with 50% focal group representation to illustrate the difference in the methods.

	group	Overall Counts	0.6 * count
1	a	76	45.6
2	b	87	52.2
3	c	144	86.4
4	d	90	54
5	e	134	80.4

In each instance, assume that "C" is selected as the focal group, a focal group proportion of 0.5 has been set, 60% of the data is to be reserved for training, and we are only altering the training group.

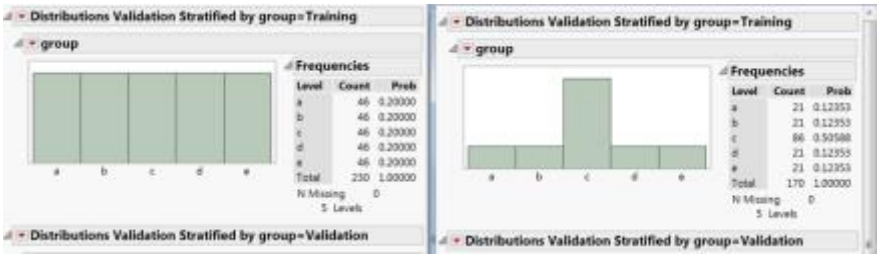
Balance all Groups (results are shown on the left):

Data is deleted to achieve equal proportions across all groups. The constraint (i.e., group of minimal size) is group "A", which has 76 rows overall, 46 (60%) of which will be devoted to training. This forces ALL groups to have 46 rows devoted to training… so random samples of size 46 are taken from each group.

Balance remaining groups (results are shown on the right):

Data is deleted to achieve balance across all NON-focal groups, AFTER taking into consideration the focal group's proportion, without deleting focal group data (unless necessary to achieve the balance).

From the table, we see that the focal group, "C", has 144 rows, 86 (60%) of which are to be devoted to Training. Since this must, as specified, represent 50% of the training data, then 86 elements must represent the other 50% of the data, which is to be split evenly among the 4 non-focal groups. Rounding, this results in 21 elements per group. (86 rows / 4 = 21.5 rows per group, but we cannot use 22 per group, because that would be 88 rows total, forcing us to augment the training data, which this method does not do… this method only deletes rows.)



In a dichotomous setting, there will be no difference between these two options if 50% is chosen for the focal group representation. By extension, in a k-nomial setting, if 100(1/k)% is chosen for the focal percentage, there will be no difference in the results.