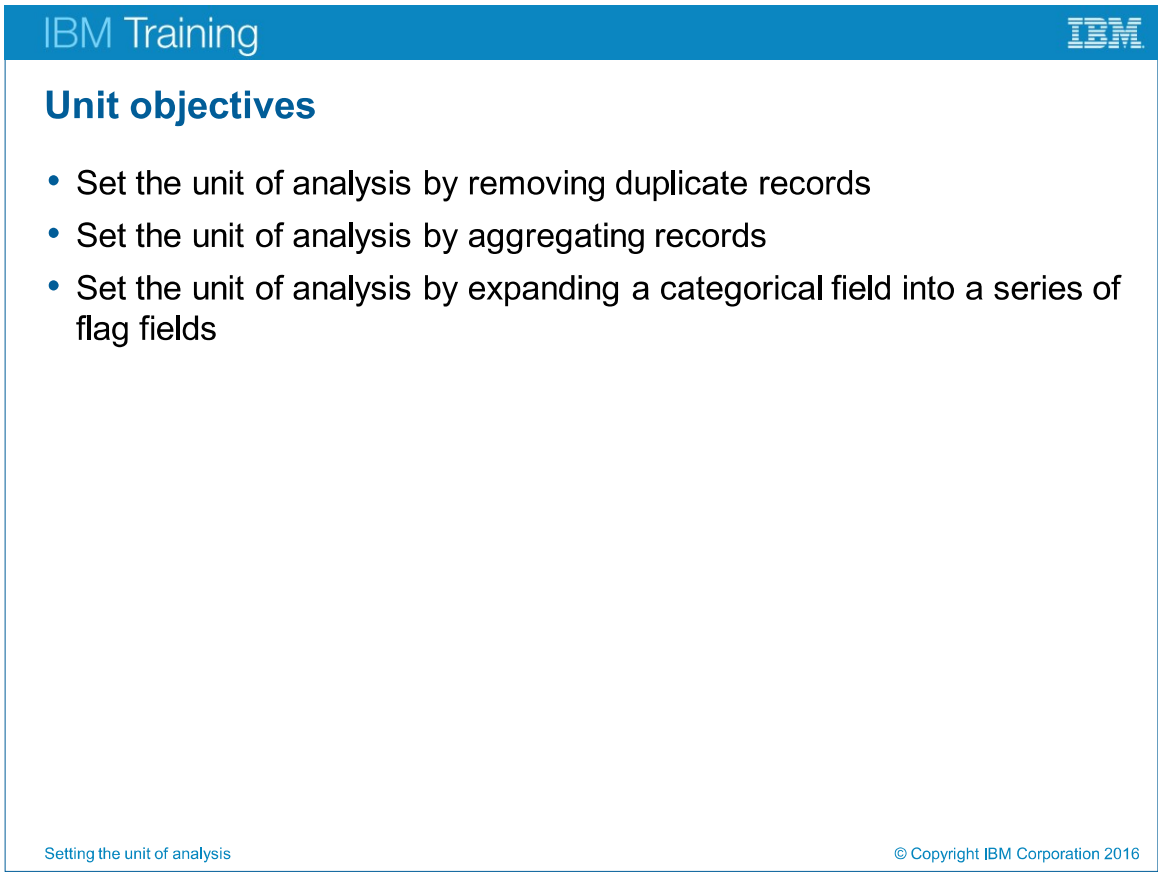
IBM Training

Setting the unit of analysis

IBM SPSS Modeler (v18)



Unit objectives

After importing and exploring the data, the next task is to set the unit of analysis, one of the tasks in the Data Preparation stage in the CRISP-DM process model. This unit presents three methods how you can set the unit of analysis.

Before reviewing this unit you should be familiar with:

- CRISP-DM
- IBM SPSS Modeler streams, nodes and palettes
- methods to collect initial data
- methods to explore the data

IBM Training

Identify the required unit of analysis

ID	AGE	CHURNED
1	21	F
2	43	F
3	56	F
4	38	F
5	32	T
5	32	T

Data errors

ID	PRODUCT	YEAR	REVENUES
1	A	2012	100
2	B	2014	50
2	C	2011	200
3	B	2011	50
3	C	2007	200
3	D	2005	10

Transactional data

Setting the unit of analysis

Identify the required unit of analysis

A dataset may not have the required unit of analysis. For example, when your dataset has a field age and you want to compute the mean age, your dataset should have only one record per person.

When the same record appears more than once it may be that your dataset is erroneous and in this case you should remove duplicate records. Another situation where you can expect multiple records is when information is collected at different moments in time. For example, a customer may have purchased products at different times, and each purchase is a record. This is referred to as transactional data, which is especially occurring in databases. Another example of information that is collected at different points in time is when a patient visits a hospital several times and each visit is a record.

When you have a transactional dataset as shown on this slide, a business question such as the frequency of co-occurrence of two products cannot be answered in IBM SPSS Modeler. To answer that question IBM SPSS Modeler requires a data structure where the products make up the fields, not the records.

IBM Training

Methods to create datasets with the required unit of analysis: Distinct

Distinct

ID	PRODUCT	YEAR	REV
1	A	2012	100
2	B	2014	50
2	C	2011	200
3	B	2011	50
3	C	2007	200
3	D	2005	10

ID	PRODUCT	YEAR	REV
1	A	2012	100
2	B	2014	50
3	B	2011	50

Setting the unit of analysis

Methods to create datasets with the required unit of analysis: Distinct

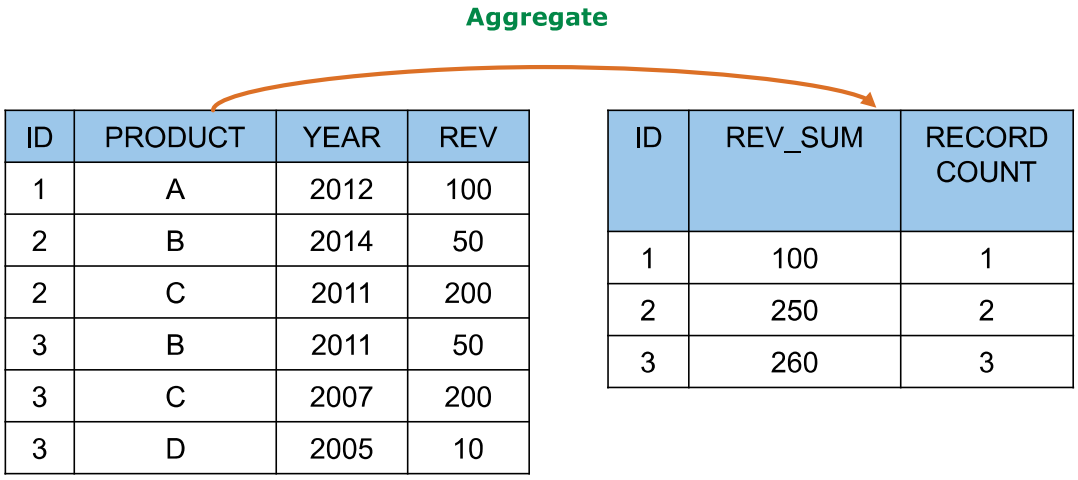
You can create a dataset with one record per customer in three ways. This slide shows the option to keep one of the records in the group, named Distinct in IBM SPSS Modeler.

In this example the group of records is defined by ID and the first record of each ID is retained. Because the data are sorted descending by year, you will retain the most recent record.

There are different choices in how you can use Distinct to create the record, which will be presented later in this unit.

IBM Training

Methods to create datasets with the required unit of analysis: Aggregate



ID	PRODUCT	YEAR	REV
1	A	2012	100
2	B	2014	50
2	C	2011	200
3	B	2011	50
3	C	2007	200
3	D	2005	10

ID	REV_SUM	RECORD COUNT
1	100	1
2	250	2
3	260	3

Setting the unit of analysis

Methods to create datasets with the required unit of analysis: Aggregate

Another method is to summarize the information over the records in the group. This option is called Aggregate.

In this example, the group of records is defined by ID, and the sum of REV is computed per ID. Also, a field is created that keeps track of the number of records in the source file.

IBM Training

Methods to create datasets with the required unit of analysis: SetToFlag

SetToFlag

ID	PRODUCT	YEAR	REV
1	A	2012	100
2	B	2014	50
2	C	2011	200
3	B	2011	50
3	C	2007	200
3	D	2005	10

ID	A	B	C	D
1	T	F	F	F
2	F	T	T	F
3	F	T	T	T

Setting the unit of analysis

Methods to create datasets with the required unit of analysis: SetToFlag

The last method is useful to transform a nominal field into a series of flag fields, so that the categories make up the columns of the dataset instead of the rows. This operation is called SetToFlag.

In this example ID defines a group of records, and the nominal field PRODUCT with categories A, B, C and D is transformed into a new dataset with one record per ID, with the fields A, B, C and D flagging if one has purchased the particular product.

The method that you will use (Distinct, Aggregate, or SetToFlag) will be determined by your analysis requirements. For example, if you want to keep track of the products purchased, then SetToFlag is the preferred method. You can also choose more than one method and combine the datasets later.

Note: IBM SPSS Modeler also provides an option to restructure the dataset from records into fields and retaining the values of the continuous field. In the example above you would create revenues_A, revenues_B, and revenues_C. Restructuring data in this way is presented in the *Advanced Data Preparation Using IBM SPSS Modeler* course.

IBM Training

Distinct records

- A group of records is defined by key fields.
- Only one record of the group will be retained.
- Use the Distinct node (Record Ops). 

- The Distinct node has more capabilities than only removing duplicates.

Setting the unit of analysis

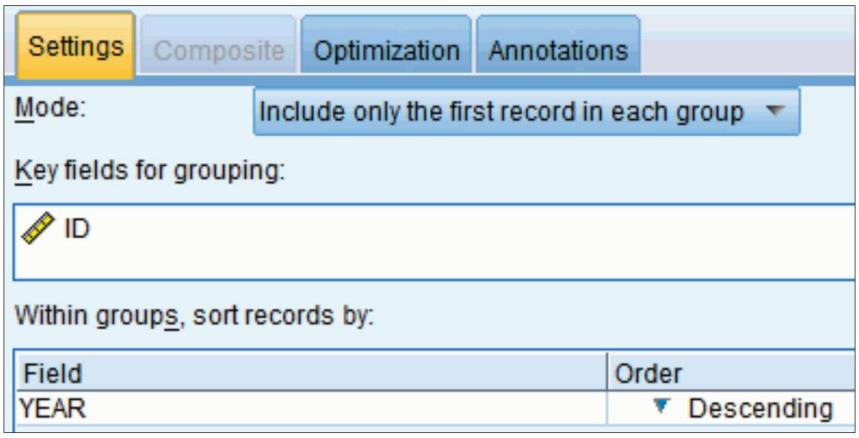
Distinct records

The Distinct node, located in the Record Ops palette, removes duplicate records. Duplicate records are defined by key fields and records with the same values on the key fields are treated as duplicate records. When you define all fields as key you can identify identical records and the data can be cleansed. To deal with transactional data, you should only specify a field such as CUSTOMER_ID as key.

The Distinct node not only lets you remove duplicates, it also gives you more flexibility in how you want to create a composite record; this is detailed in the next slides.

IBM Training

Explore the Distinct dialog box: Define the unit of analysis



Setting the unit of analysis

Explore the Distinct dialog box: Define the unit of analysis

The Distinct dialog box has three main tabs: Settings, and Composite and Optimization. The first two tabs are presented in this unit. The Optimization tab sets options to improve performance; refer to the online Help for more information.

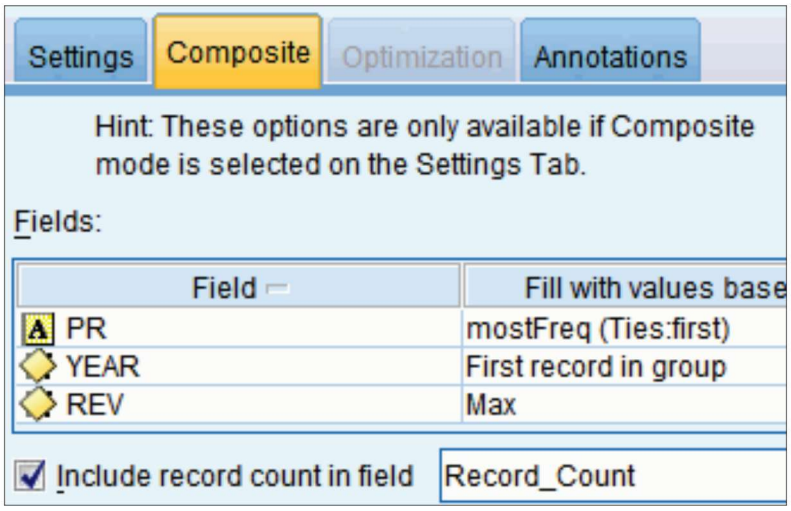
On the Settings tab, under Key fields for grouping, specify the field(s) that defines (define) the unit of analysis. You can sort the records within each group, ascending or descending, to ensure that a particular record is the first record in the group. Sorting records in the Distinct node itself makes a separate Sort node upstream from the Distinct node redundant, and is recommended.

Mode controls how records are formed in the output dataset. IBM SPSS Modeler provides three options:

- Include only the first record in each group. Retains the first record of a group of records. Use this option if you want to remove duplicate records.
- Discard only the first record in each group either include the first record or exclude the first record. Removes the first record in the group and retains the rest. Use this option to examine the duplicate records.
- Create a composite record for each group, which will make the Composite tab available. Refer to the next slide for more information.

IBM Training

Explore the Distinct dialog box: Create composite records



Setting the unit of analysis

Explore the Distinct dialog box: Create composite records

On the Composite tab you specify how the new record must be built from the source records. For example, you can have the first record's value of field X, the last record's value of field Y, and the maximum value of field Z.

This slide shows an example of creating a composite record for a transactional dataset. For a categorical field named PRODUCT, the most frequent product will be output for each group of records. For example, when a customer has 5 records with product A for the first record, A for the second, B for the third, C for the fourth, and D for the fifth, the value that is output will be A. When there are more products with the same frequency of occurrence, for example A A B B C, then the first product will be output, A in this example.

The Composite tab also provides an option to create an extra field, named Record_Count by default, which returns the number of input records that were grouped to form an output record.

IBM Training

Aggregate records

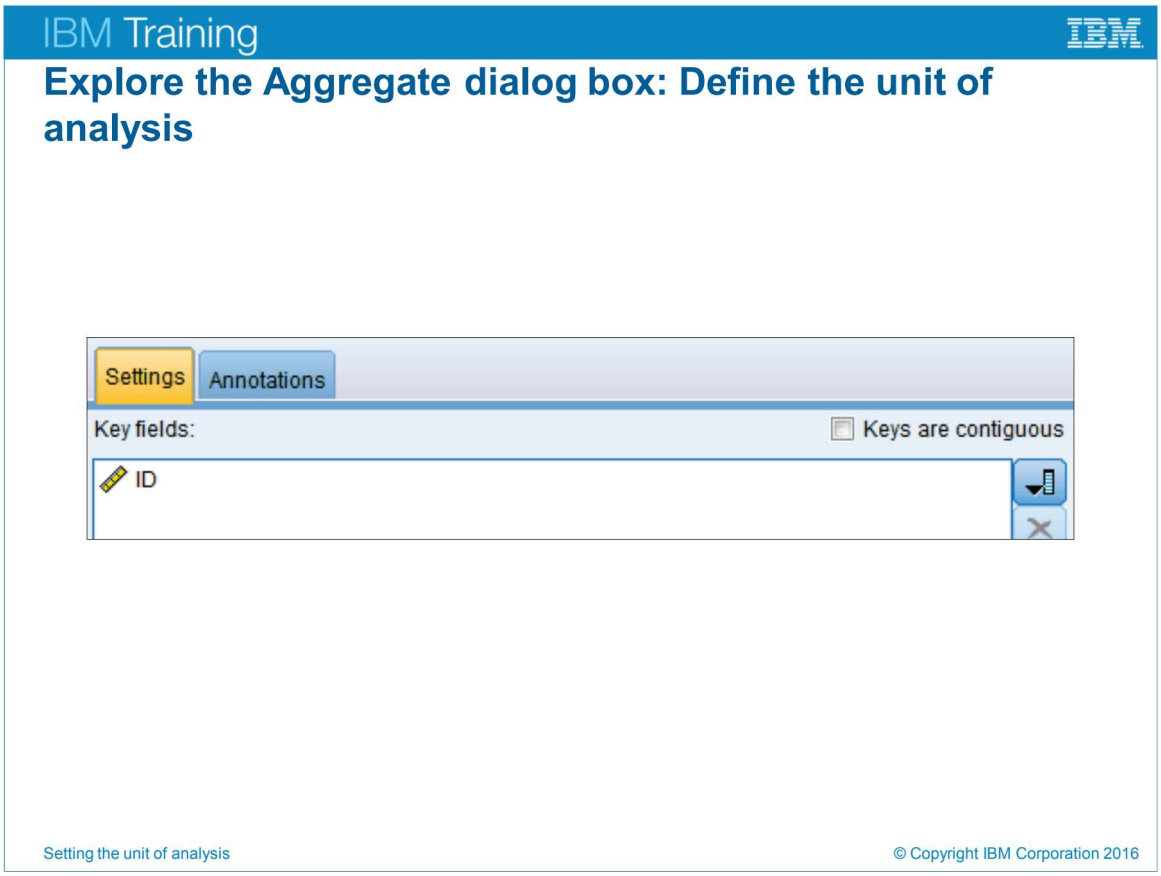
- A group of records is defined by key fields.
- Aggregation will summarize information over all records in each group.
- Use the Aggregate node (Record Ops). 

Setting the unit of analysis

Aggregate records

You can deal with multiple records per customer by using the Distinct node. Another option is to aggregate information over all the records that a customer has. For example, you can have the mean value of a certain field, computed over all the customer's records. The Distinct node mimics this functionality by creating composite records and offering aggregate statistics such as minimum, mean and maximum, but the Aggregate node provides more statistics.

The Aggregate node, located in the Record Ops palette, replaces a group of input records with one aggregated output record. After passing data through the Aggregate node, the overall file structure has changed because the record definition is altered.



Explore the Aggregate dialog box: Define the unit of analysis

As in the Distinct node a group of records is defined by key fields. A key field such as ID will group the records of a customer into one record.

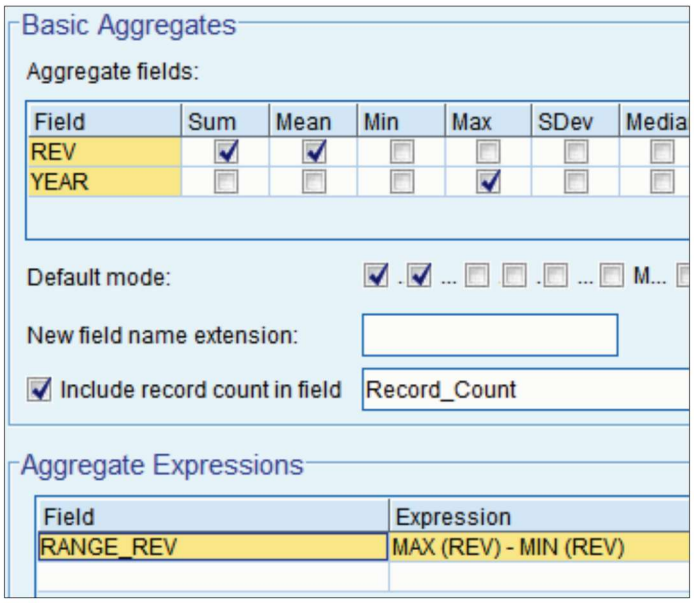
When no key field is specified, the aggregation will be over all the records in the dataset, and thus will result in one record.

If you want to retain a field value that is constant for all records in an aggregate group, such as gender, add the field to the list of key fields.

To improve performance you can enable the Keys are contiguous option if the data are already sorted on the key fields.

IBM Training

Explore the Aggregate dialog box: Define the new fields



Setting the unit of analysis

Explore the Aggregate dialog box: Define the new fields

Under Basic Aggregates, select the fields that you want to aggregate values for and select the statistic(s). You can choose Sum, Mean, Min, Max, SDev (standard deviation), Median, Count (the number of records having a non-$null$ value), Variance, 1st and 3rd Quartile (25th and 75th percentile). Enable the Include record count in field option to create a field that stores the number of records aggregated to form each output record.

Aggregate Expressions operate on the group of records as defined by the key field(s). For example, although the range (the difference between maximum and minimum value) is not available as one of the statistics, it can be created by using the built-in aggregate functions MAX en MIN. When you imported from a database, you can also use the functions supported by your database.

Note: User-defined blanks are included in the computation of the statistics. For example, you declared 999 as blank value for AGE. Requesting the Max statistic for AGE will return 999 although this value is declared as blank. Also the Mean statistic will be affected by this blank value. Therefore, nullify blank values upstream from the Aggregate node.

IBM Training

Transform a field into a series of flag fields

- Transform a categorical field into a series of flag fields.
- Especially relevant in the case of transactional data.
- Use the SetToFlag node (Field Ops). 

Setting the unit of analysis

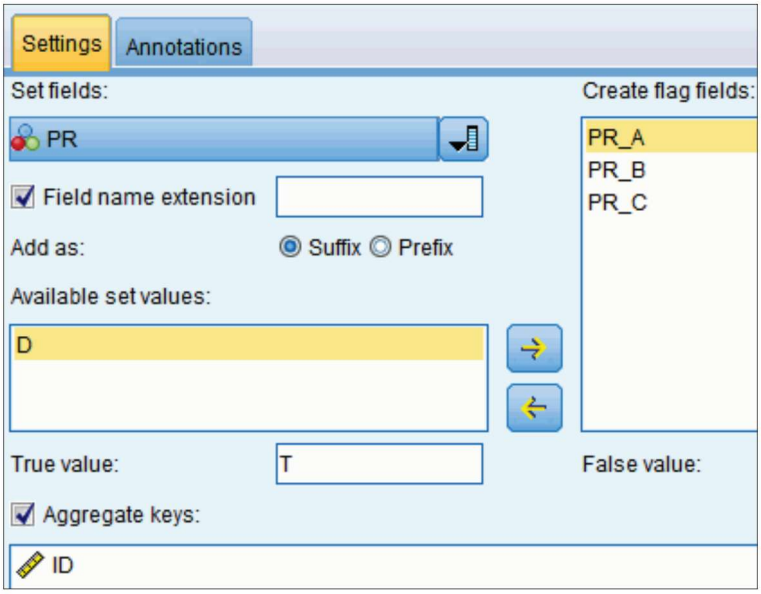
Transform a field into a series of flag fields

It may be necessary to convert information held in a categorical field into a collection of flag fields. This is especially true when the data are of a transactional nature.

The SetToFlag node, located in the Field Ops palette, expands a nominal field into a series of flag fields, with the option to aggregate the data.

IBM Training

Set a field to flag fields: Define the unit of analysis and define new fields



Setting the unit of analysis

Set a field to flag fields: Define the unit of analysis and define new fields

On the Settings tab, under Set fields, select the categorical field that you want to expand in flags. The area under Available set values will be populated with the categories of the selected categorical field, provided that the field is instantiated. If no values are available, add a Type node upstream from the SetToFlag node and instantiate the field. Move the categories that you want to create flag fields for to the Create flag field area. Optionally, extend the field name for the new flag fields, either as suffix or prefix.

By default, the true value and the false value will be T and F, respectively. You can change these values if you want.

Enable the option Aggregate keys and select the appropriate key field(s) to change the unit of analysis. If you do not specify an aggregate key field, the unit of analysis will not change and you will have as many records downstream from the SetToFlag node as you had upstream from the SetToFlag node.