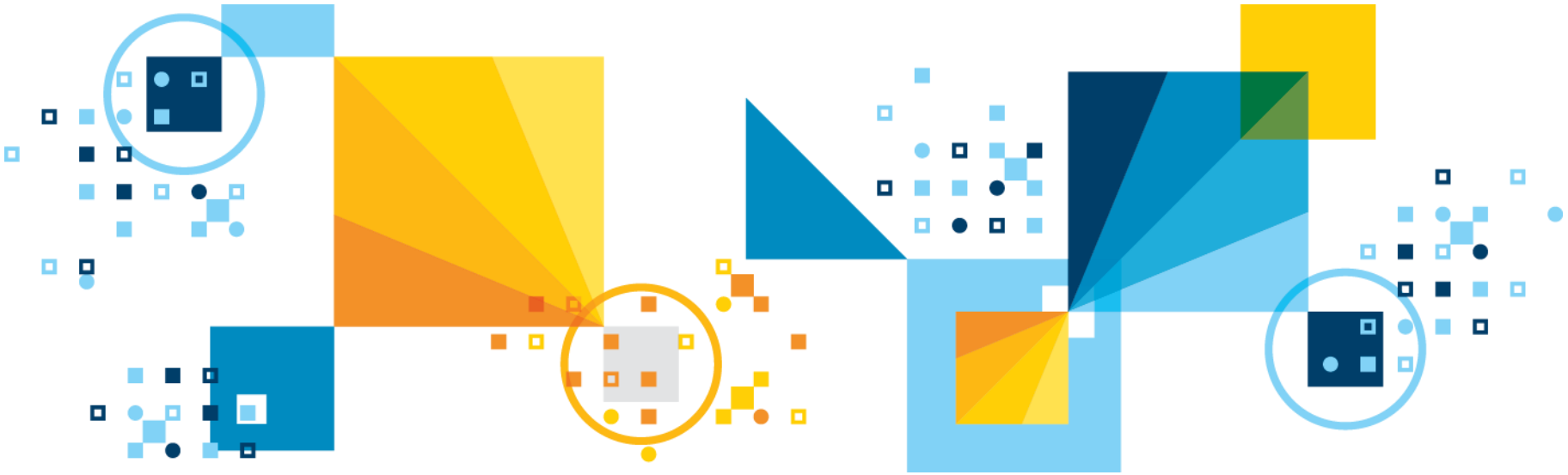
Predictive Modeling Fundamentals I
Lesson 2



Setting the Stage….

Why this is important to know…

1. Fundamental introduction to Data Mining and its application to business problems
2. Ability to utilize software tools for advanced analytics

After this session, you will be able to…

1. Understand the first steps of CRISP-DM methodology
2. Understand data preparation and preprocessing
3. Perform data preparation with SPSS Modeler

Speaking to you today…

Armand Ruiz
Product Manager

Mikhail Lakirovich
Product Marketing Manager

Agenda

- 1. Business Understanding
  - The Titanic Use Case
- 2. Data Understanding
  - Tools for Data Exploration in IBM SPSS Modeler
- 3. Data Preparation
  - Major Tasks in Data preprocessing
  - Tools for Data Preparation in IBM SPSS Modeler

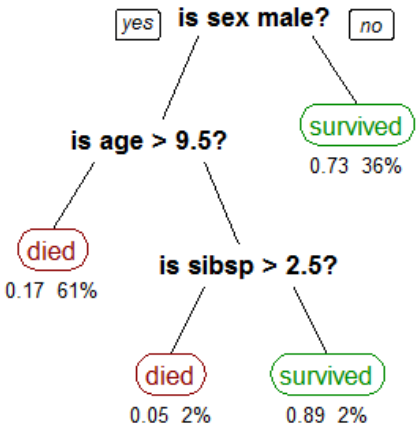
- Lab 2: Understand the data available and prepare it for modelling
- Solution

1. Business Understanding: Initial phase focuses on understanding the project objectives and requirements

Use Case: Sinking of the Titanic User Case

- 1502 out of 2224 passengers and crew died
- Not enough lifejackets for all passengers and crew
- Some groups were more likely to survive than others, such as women, children, and the upper-class

- **Challenge:** Analysis if a passenger is likely to survive using Data Mining



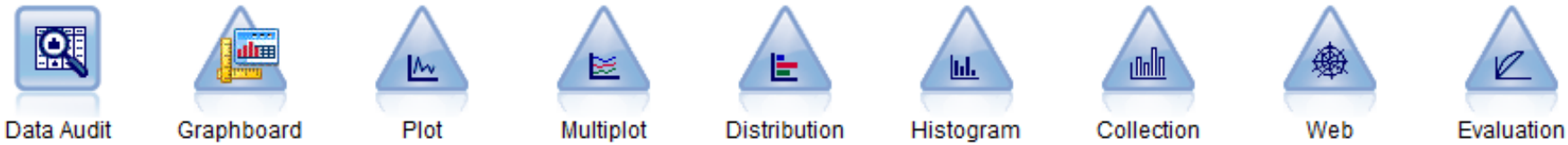
2. Data Understanding: Initial data collection and getting familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

kaggle

Train

Test

Tools for Data Exploration for IBM SPSS Modeler



DEMO!!

3. Data Preparation: Construction the final dataset (data that will be fed into the model) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Why Data Preprocessing?

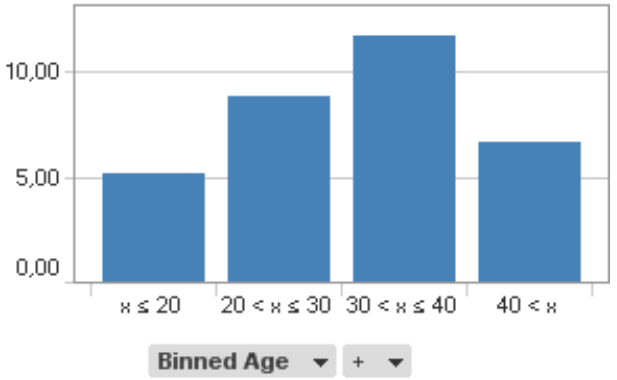
- Data in the real world is dirty
  - incomplete: lacking attribute values or lacking certain attributes of interest
    - Gender=“”
  - noisy: containing errors or outliers
    - Age=“-25”
  - inconsistent: containing discrepancies in codes or names
    - Discrepancy between duplicate ratings
    - Rating on ‘1-5’ scale vs on ‘A-E’ scale

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Feature Selection
  - Feature Extraction

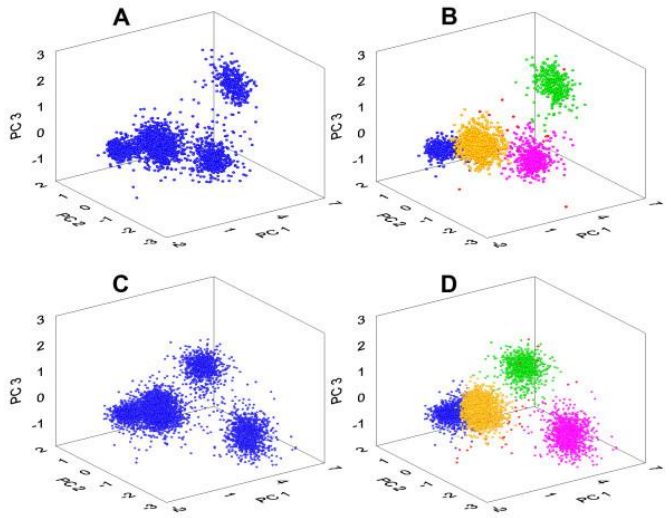
How to Handle Noisy Data?

- Binning method
  - first sort data and partition into (equi-depth) bins



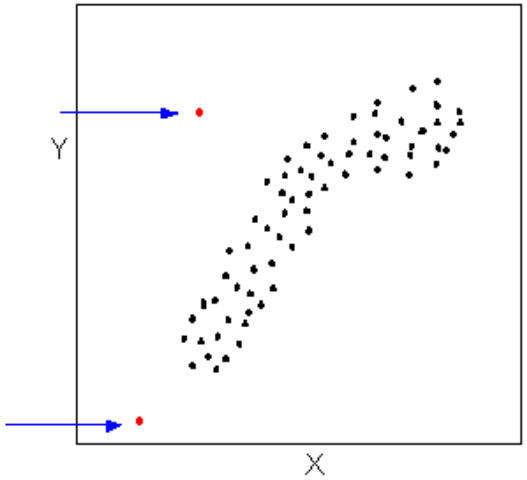
- Clustering
  - detect and remove outliers

- Expert knowledge
  - model detects potential outliers and they are validated by subject matter expert



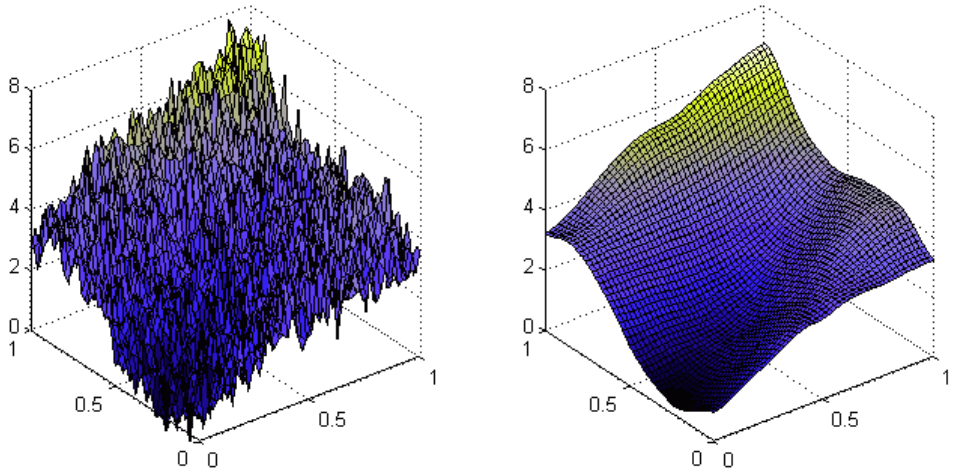
Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

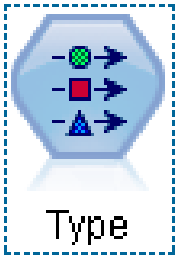


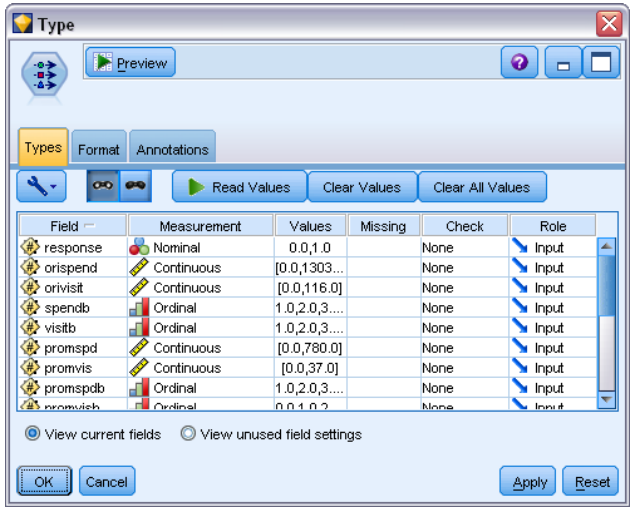
Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction: create new attributes



Tools for Data Preparation for IBM SPSS Modeler: Type Node





-Data type declarations

-Setting field directions

-Declarations of user defined blanks

-Checking of values (Range and Set values)

Continuous: Used to describe numeric values, such as a range of 0–100 or 0.75–1.25. A continuous value can be an integer, real number, or date/time.

Categorical: Used for string values when an exact number of distinct values is unknown.

Flag: Used for data with two distinct values, such as Yes and No or 1 and 2.

Nominal: Used to describe data with multiple distinct values, each treated as a member of a set, such as small/medium/large. Nominal data can have any storage—numeric, string, or date/time.

Ordinal: Used to describe data with multiple distinct values that have an inherent order. For example, salary categories or satisfaction rankings can be typed as ordinal data.

Typeless: Used for data that does not conform to any of the above types, for fields with a single value, or for nominal data where the set has more members than the defined maximum.

Field Direction in the Type Node

INPUT: The field will be used as an input or predictor to a modeling technique

TARGET: The field will be the output or target for a modeling technique

BOTH: Direction suitable for the association modeling nodes. Allows the field to be an input and an output in an association rule. All other modeling techniques will ignore the field

NONE: The field will not be used in modeling.

PARTITION: indicates a field used to partition the data into separate samples for training, testing and (optional) validation purposes.

Direction used in Modeling

Missing Data: What does it matter?

It is necessary to carefully examine our data to determine how we will handle our data and its quality *before model building*

The higher the quality of data in data mining, the more accurate the prediction

Missing data that is not specified can change the variable type or affect the range of values…therefore…

Files with missing data can produce *misleading results* if it is not identified prior to analysis

Types of Missing Data

For numeric variables:

If a record has no value (e.g. didn't reveal income). Replaced automatically using system missing: $NULL$

For string variables:

A blank space is a valid entry for a string variable. Will have to 'tell' Modeler to pick up these values. Two forms:

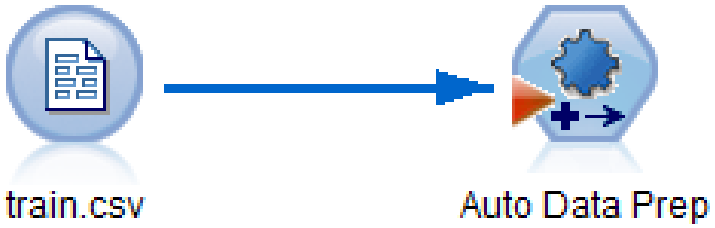
- White Space: ' '
- Empty Strings: ' ' … empty strings are a subset of white spaces

User-specified missings are called: BLANK VALUES

- Missing values in the variable CHILDREN are coded as 99

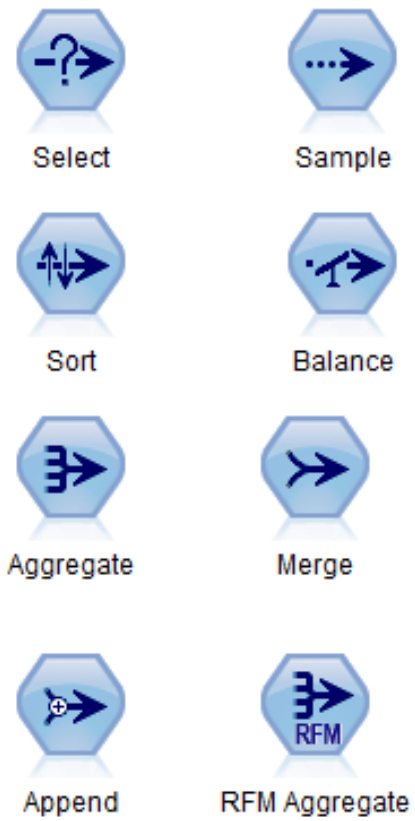
DEMO!!

Auto Data Preparation node: Handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques.

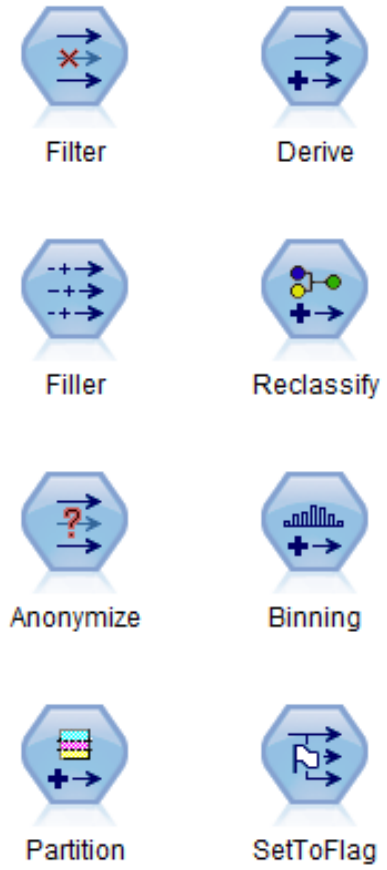


DEMO!!

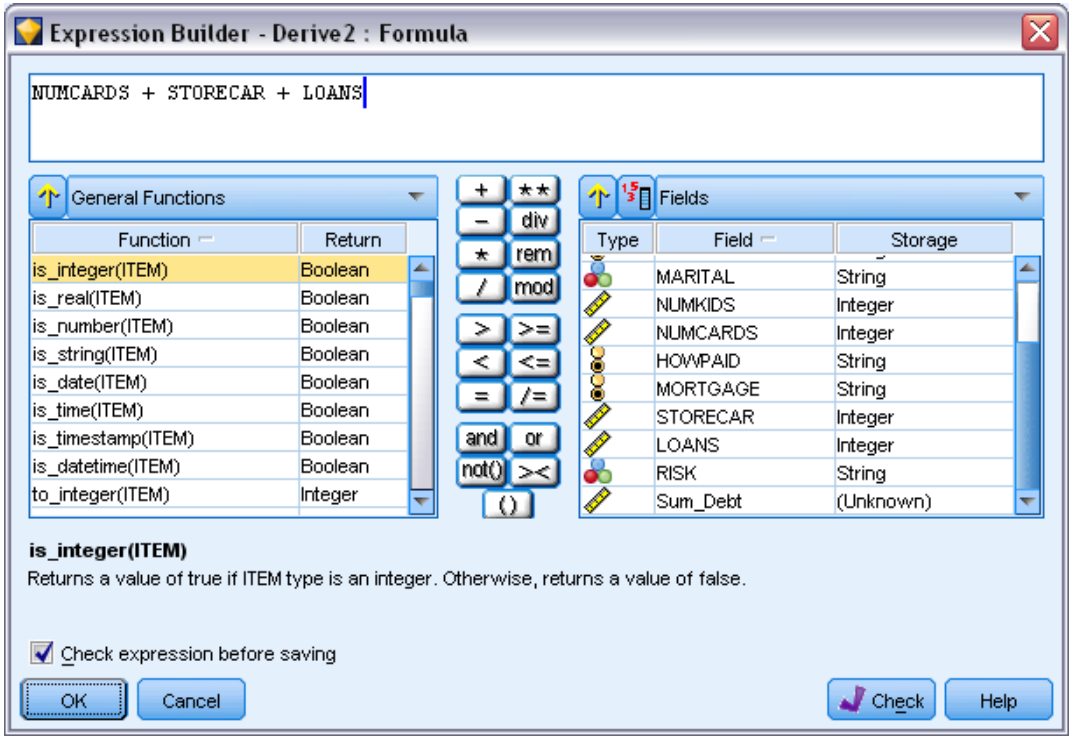
Record Operations



Field Operations



CLEM language: The Expression Builder



CLEM = Control Language Expression Manipulation

DEMO!!

Lab 2:

- Load Titanic Data
- Explore the Data
- Prepare the Data for Modelling