AGENDA

1. Data Preparation (Data Quality)
2. Data Preparation (CLEM Language)
3. Data Preparation (Feature extraction)

Data Science Academy

I

Data Preparation (Data Quality)

Data Science Academy

Data Quality

- Data are rarely perfect. In fact, most data contain coding errors, missing values, or other types of inconsistencies that make analysis tricky at times. One way to avoid potential pitfalls is to conduct a thorough quality analysis of available data before modeling.
  - Missing data include values that are blank or coded as a non-response (such as $null$, ?, or 999).
  - Data errors are usually typographical errors made in entering the data.
  - Measurement errors include data that are entered correctly but are based on an incorrect measurement scheme.
  - Coding inconsistencies typically involve nonstandard units of measurement or value inconsistencies, such as the use of both M and male for gender.
  - Bad metadata include mismatches between the apparent meaning of a field and the meaning stated in a field name or definition.

Bütün hüquqlar qorunur.

Writing a Data Quality Report

- Have you identified missing attributes and blank fields? If so, is there meaning behind such missing values?
- Are there spelling inconsistencies that may cause problems in later merges or transformations?
- Have you explored deviations to determine whether they are "noise" or phenomena worth analyzing further?
- Have you conducted a plausibility check for values? Take notes on any apparent conflicts (such as teenagers with high income levels).
- Have you considered excluding data that has no impact on your hypotheses?
- Are the data stored in flat files? If so, are the delimiters consistent among files? Does each record contain the same number of fields?

Bütün hüquqlar qorunur.

What is an outlier?

- The outliers – or extreme values – can represent a danger for the analysis, because they directly affect mean and standard deviation. That's why we should detect, and in some cases remove the outliers before running such tests.
  - An outlier is an unusually small or unusually large value in a data set.
  - It might be an incorrectly recorded data value.
  - It might be a data value that was incorrectly included in the data set.
  - It might be a correctly recorded data value that belongs in the data set!

Bütün hüquqlar qorunur.

Identification of Outliers

- There are two methods for identifying the outliers:
  - A **numerical method**, based on the standardized values

    A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
  - A **graphical method**, bases on the boxplot chart

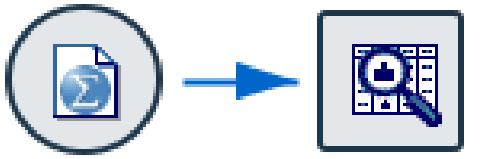
Bütün hüquqlar qorunur.

Controlling outliers

1. Import *“customer_dbase.sav”*;
2. Use **Data Audit** node;
3. Open **Quality** tab;
4. In the **Action** column, specify handling for outliers and extremes for specific fields as desired:
   - **Coerce** – Replaces outliers and extreme values with the nearest value that would not be considered extreme.
   - **Discard** – Discards records with outlying or extreme values for the specified field.
   - **Nullify** – Replaces outliers and extremes with the null or system-missing value.

Complete fields (%): 88.37% Complete records (%): 11.56%

Field	Measurement	Outliers	Extremes	Action	Impute Missing
region	Nominal	--	--	--	Never
townsize	Ordinal	--	--	--	Never
gender	Nominal	--	--	--	Never
age	Continuous	0	0	None	Never
agecat	Ordinal	--	--	--	Never
birthmonth	Nominal	--	--	--	Never
ed	Continuous	0	0	None	Never
edcat	Ordinal	--	--	--	Never
jobcat	Nominal	--	--	--	Never
union	Nominal	--	--	--	Never
employ	Ordinal	--	--	--	Never
empcat	Ordinal	--	--	--	Never
retire	Nominal	--	--	--	Never
income	Continuous	63	23	None	Never
lninc	Continuous	15	0	None	Never
inccat	Ordinal	--	--	Coerce	Never
debtinc	Continuous	52	1	Discard	Never
creddebt	Continuous	50	23	Nullify	Never
lncreddebt	Continuous	27	1	Coerce outliers / discard extremes	Never
othdebt	Continuous	65	28	Coerce outliers / nullify extremes	Never
lnothdebt	Continuous	24	0		Never
default	Nominal	--	--	--	Never

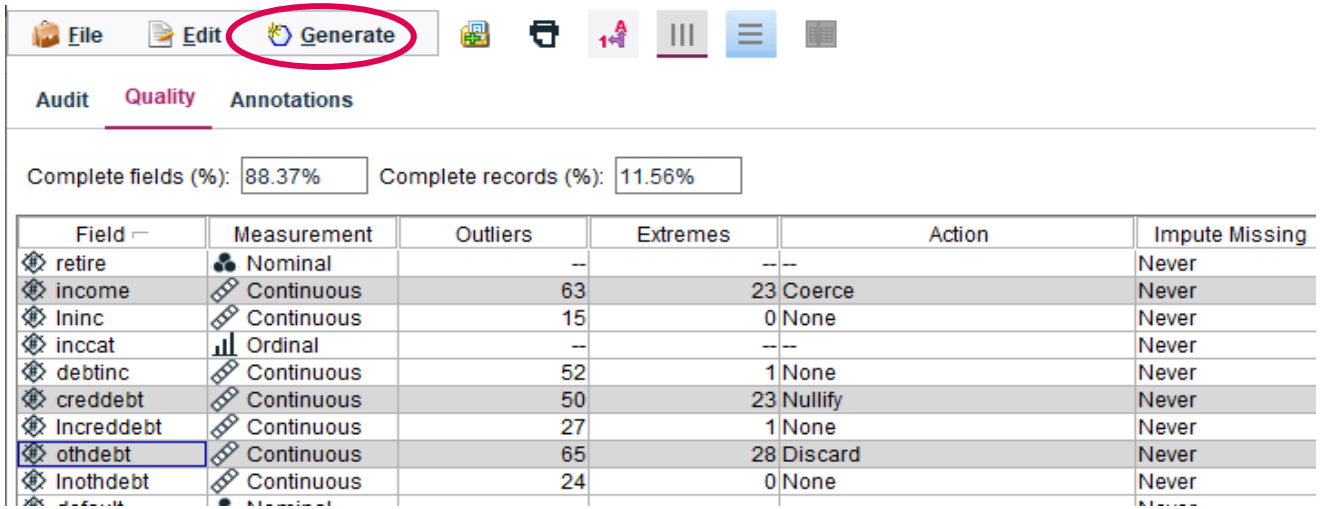


customer_dbase.sav 129 Fields

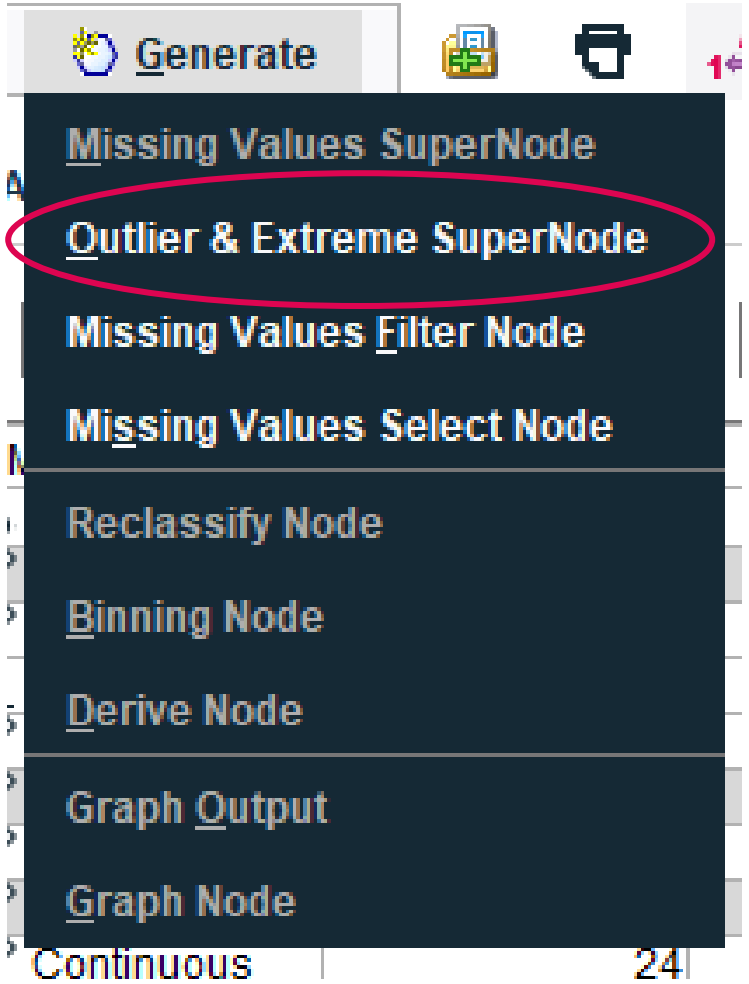
Bütün hüquqlar qorunur.

Removing outliers and extremes

1. Select **Coerce** in *income* row, **Nullify** in *creddebt* row, **Discard** in *otherdebt* raw.
2. Select *income* raw, *creddebt* raw and *otherdebt* raw.
3. Select **Outliers & Extreme Supernode** from **Generate** tab;



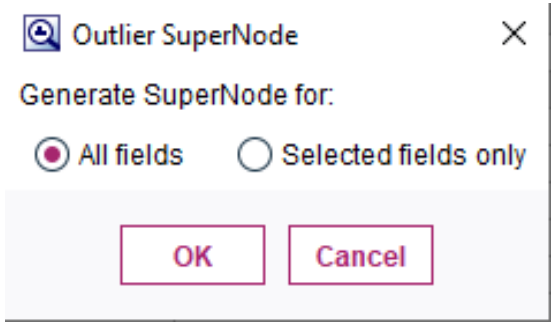
Field	Measurement	Outliers	Extremes	Action	Impute Missing
retire	Nominal	--	--	--	Never
income	Continuous	63	23	Coerce	Never
lninc	Continuous	15	0	None	Never
inccat	Ordinal	--	--	--	Never
debtinc	Continuous	52	1	None	Never
creddebt	Continuous	50	23	Nullify	Never
lncreddebt	Continuous	27	1	None	Never
othdebt	Continuous	65	28	Discard	Never
lnothdebt	Continuous	24	0	None	Never

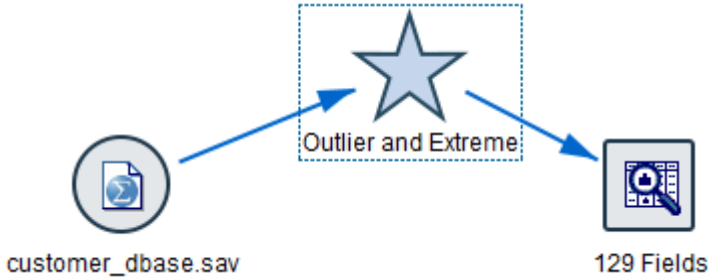


Bütün hüquqlar qorunur.

Removing outliers and extremes

4. Select **All fields** or **Selected fields only**;
5. Click **Ok**;
6. Connect **SuperNode** with **Source** node;
7. Add **Data Audit** node again and connect to **SuperNode** in order to compare the number of outliers in quality tab.





Bütün hüquqlar qorunur.

What is missing value?

Missing data, also known as missing values, is where some of the observations in a data set are blank. There are 4 types of missing data in SPSS Modeler.

Types of missing data	Definition
$Null$ value	Applies only to numeric fields. This is a cell that is empty or has an illegal value
White space	Applies only to string fields. This is a cell that is empty or has spaces
Empty string	Applies only to string values. This is a cell that is empty. Empty string is a subset of white space.
Blank value	This is predefined code, and it applies to any type of field.

Bütün hüquqlar qorunur.

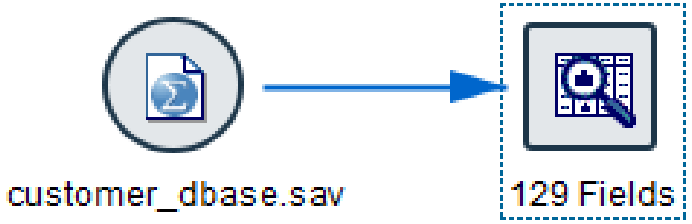
Handling Missing Values

- There are 3 ways to handle missing value:
  - Remove fields
  - Remove cases
  - Impute missing values

Bütün hüquqlar qorunur.

Impute Missing Values

1. Add statistics file
2. Import *“customer_dbase.sav”*;
3. Add **Type** node;
4. Edit **Type** node and click on **Read Values**;
5. Add **Data Audit** node from **Output** palette;
6. Click **Run**;
7. Open **Quality** tab of output;
8. Select **Specify**.



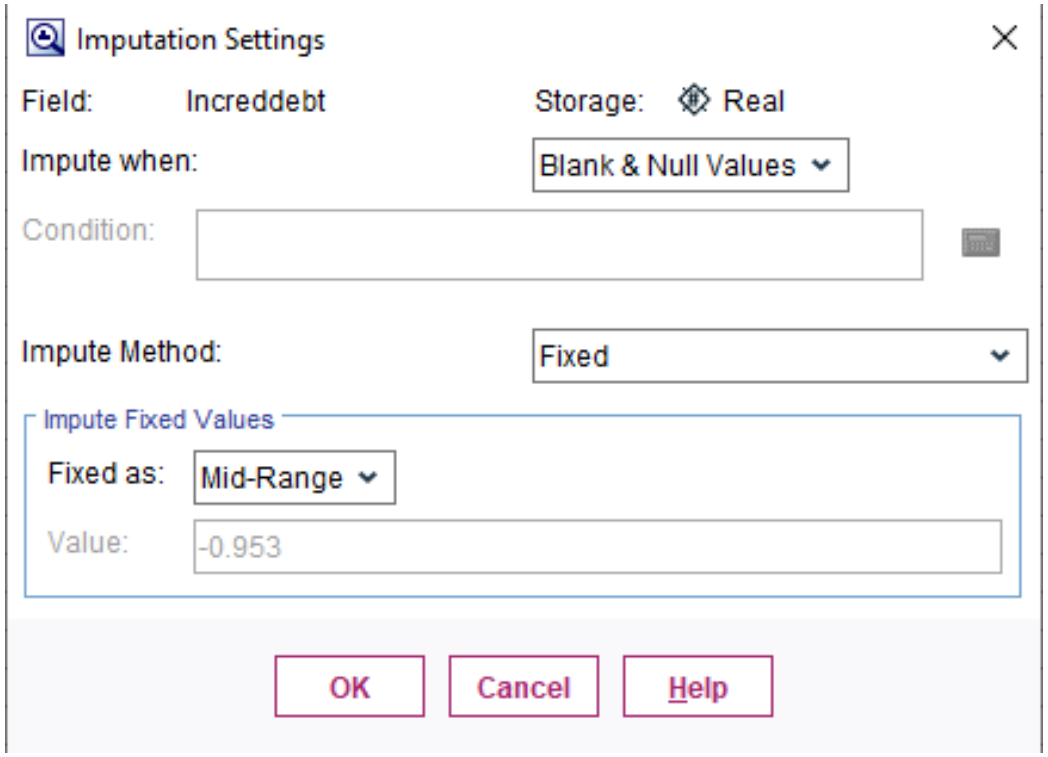
Impute Missing	Method
Never	Fixed
Never	Fixed
Never	Fixed
Blank Values	Fixed
Null Values	Fixed
Blank & Null Value	Fixed
Condition...	Fixed
Specify...	Fixed
Never	Fixed
Never	Fixed
Never	Fixed

Bütün hüquqlar qorunur.

Specify Imputation method

- In this dialog box, you can specify which imputation method you want to use, and once you have chosen a method, you can then further specify details about the imputation. There are several imputation methods:

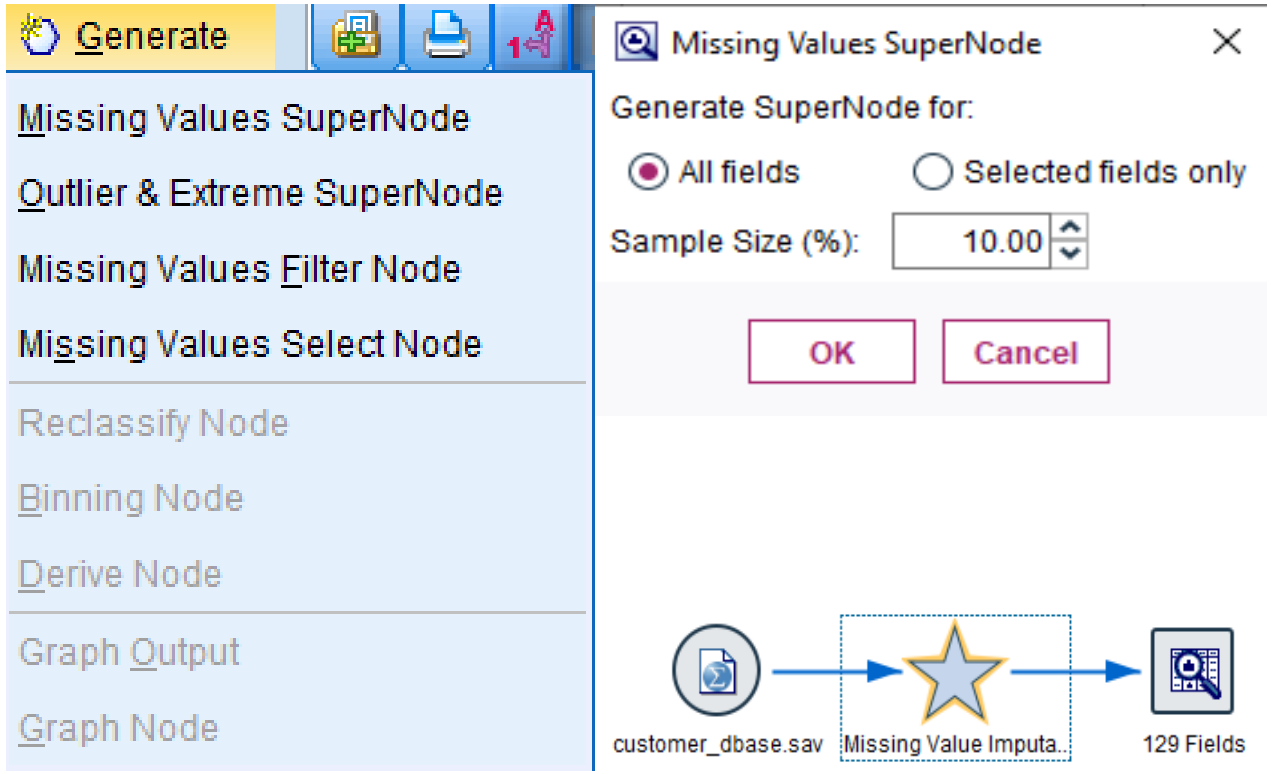
  - Fixed uses the same value for all cases. This fixed value can be a constant, the mode, the mean, or a median (the options will vary depending on the measurement level of the field).
  - Random uses a random value based on a normal or uniform distribution. This allows for there to be variation for the field with imputed values.
  - Expression allows you to create your own equation to specify missing values.
  - Algorithm uses a value predicted by a C&R Tree model.



Bütün hüquqlar qorunur.

Impute Missing Values

9. Select **Missing Values SuperNode** from **Generate** menu;
10. Click **Ok**;
11. Connect Supernode to Source node.



Bütün hüquqlar qorunur.

II

Data Preparation (CLEM Language)

Data Science Academy

What is CLEM Language?

- CLEM is the **Control Language for Expression Manipulation**, which is used to build expressions within SPSS Modeler streams

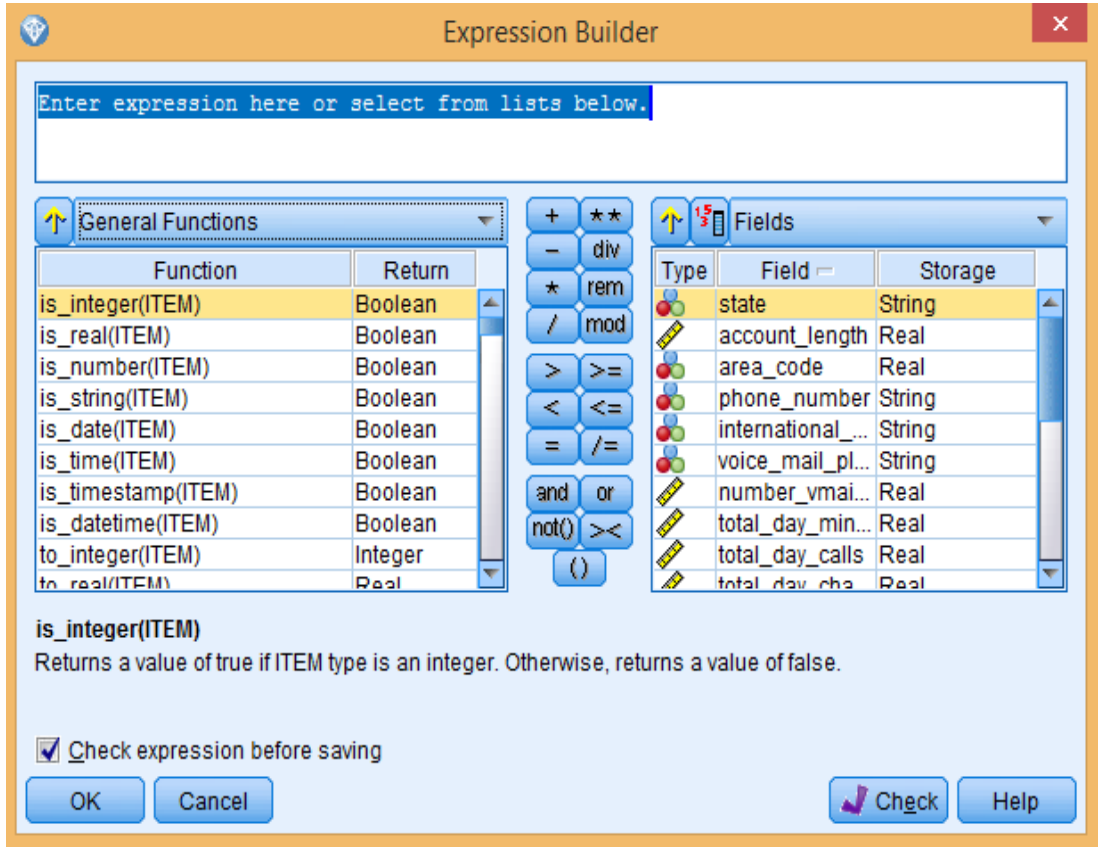
- CLEM is actually used in a number of SPSS nodes

Bütün hüquqlar qorunur.

CLEM Expression

CLEM expressions

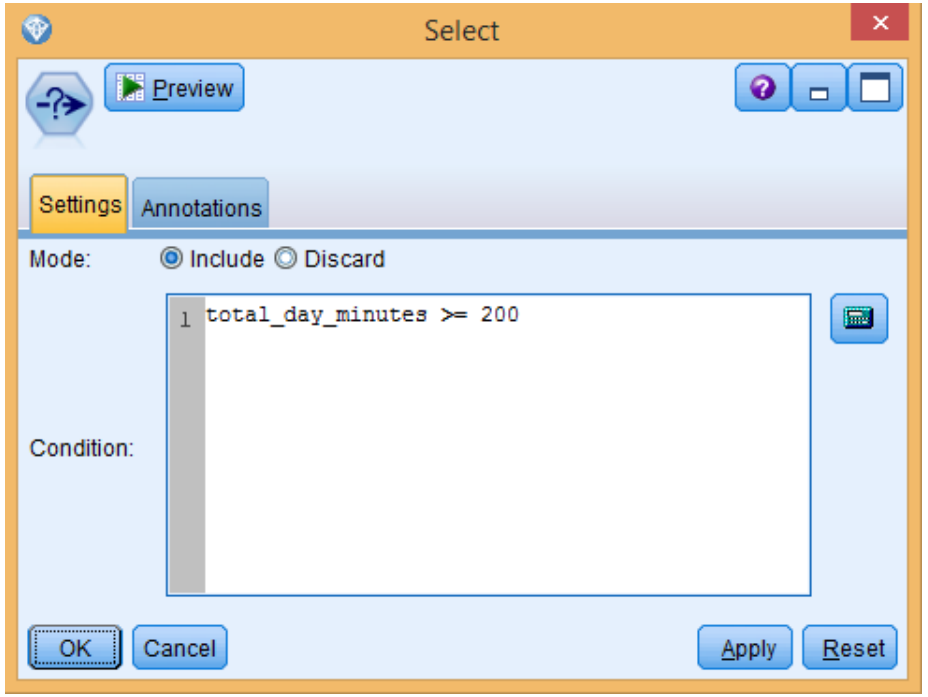
- Values
- Functions
- Operators
- Field names



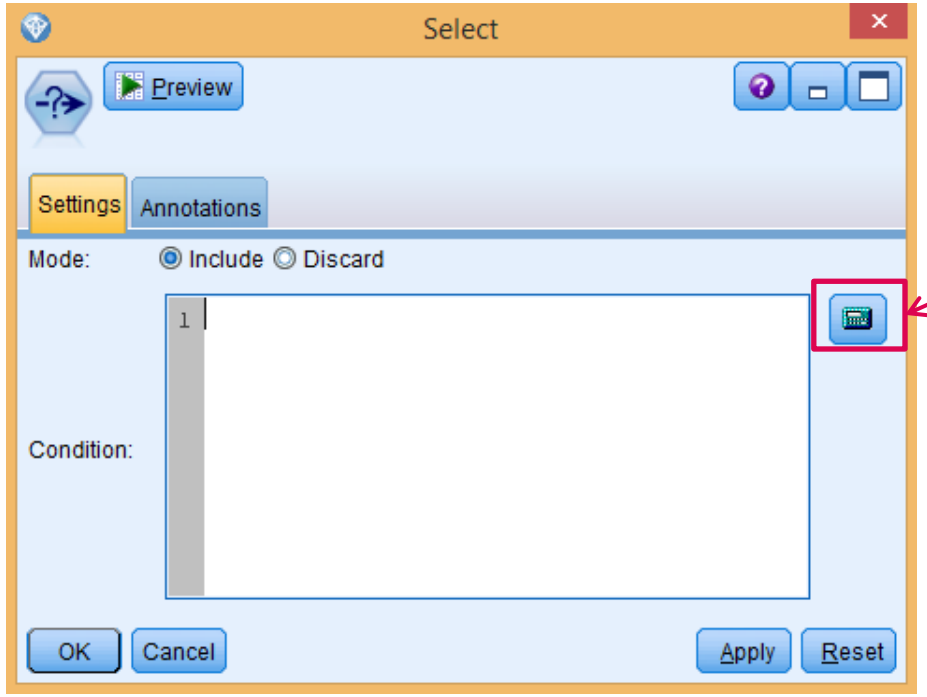
Bütün hüquqlar qorunur.

Using Clem

- In nodes like “Select” and “Derive”, expressions and conditions can be entered directly (by typing them in).
- By using the **SPSS Expression Builder.** The Expression Builder is invoked by clicking the Launch expression builder button in the respective dialog (for example, in a Select node you click the Launch expression builder button to launch the Expression builder).



or



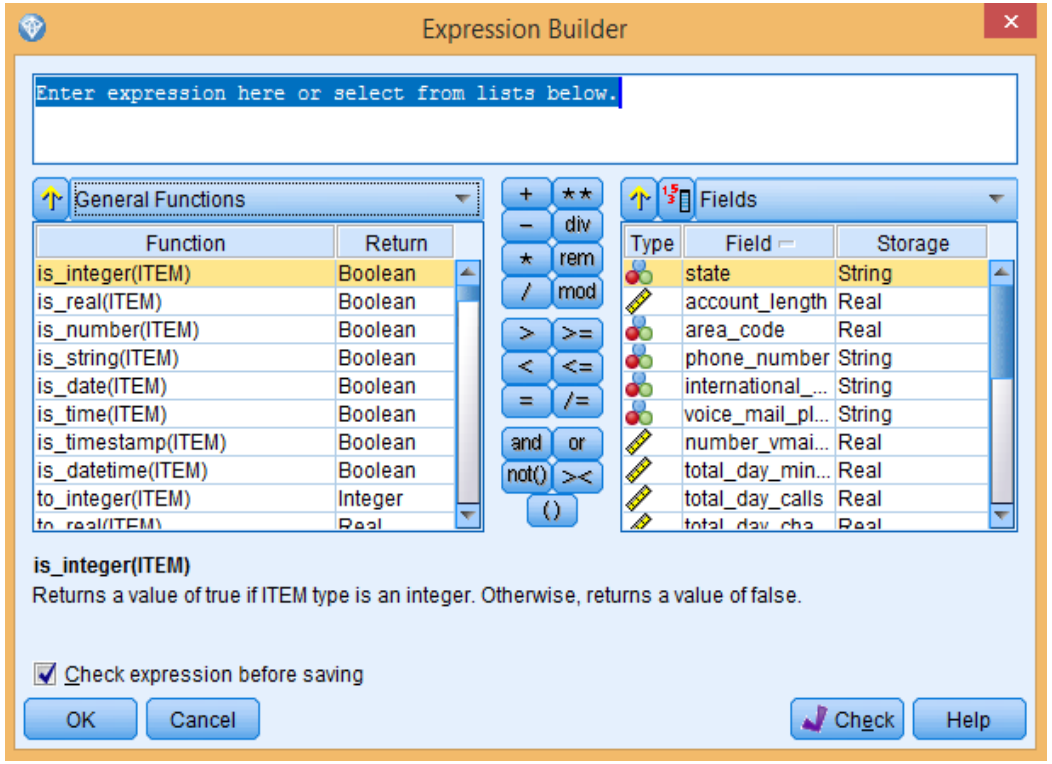
Expression Builder

Bütün hüquqlar qorunur.

Expression Builder

The Expression builder provides powerful tool to construct expressions and in addition, the Expression Builder can check the validity of your expression in real-time

- Across the top of the Expression Builder dialog is where you can enter your expression directly.
- **“General Functions”** is a list of all available SPSS functions to the expression builder
- The middle of the dialog shows the expression operators for selection
- Under **“Fields”** you can view and select a field (from your data) to be used in the expression
- When you clock on a function, its description is displayed under the Function list.
- You can use the **“Check”** button to test the syntax of your expression
- A value can be picked from a list of values (provided the field has been instantiated) by clicking the fields Value button or right clicking on the field itself



Bütün hüquqlar qorunur.

Text selection with CLEM

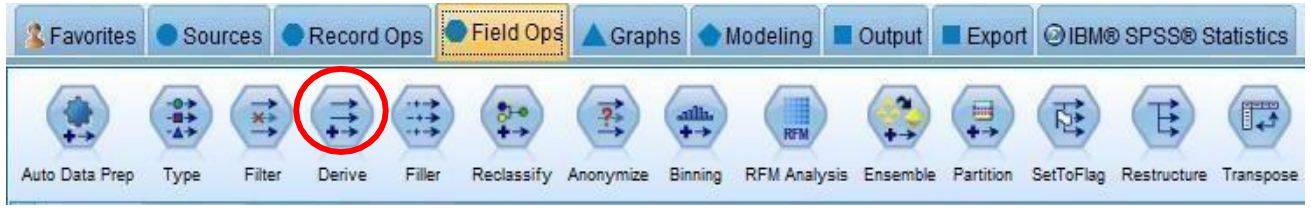
Formula	Description
startstring(LENGTH, STRING)	Extracts the first N characters from the specified string. If the string length is less than or equal to the specified length, then it is unchanged
endstring(LENGTH, STRING)	Extracts the last N characters from the specified string. If the string length is less than or equal to the specified length, then it is unchanged
substring(N,LEN, STRING)	Returns a string SUBSTRING, which consists of the LEN characters of the string STRING, starting from the character at subscript N.

Bütün hüquqlar qorunur.

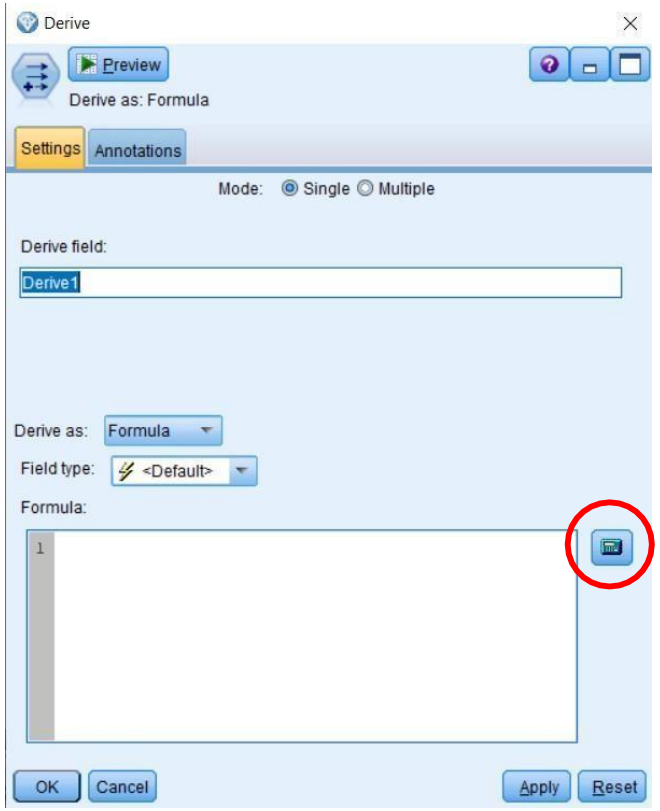
Text selection with CLEM

- Open *"customer_dbase.sav"*
- Double click the **Derive** node from **Field Ops** tab;
- Double click the **Derive** node on the canvas;
- Click button

1



2



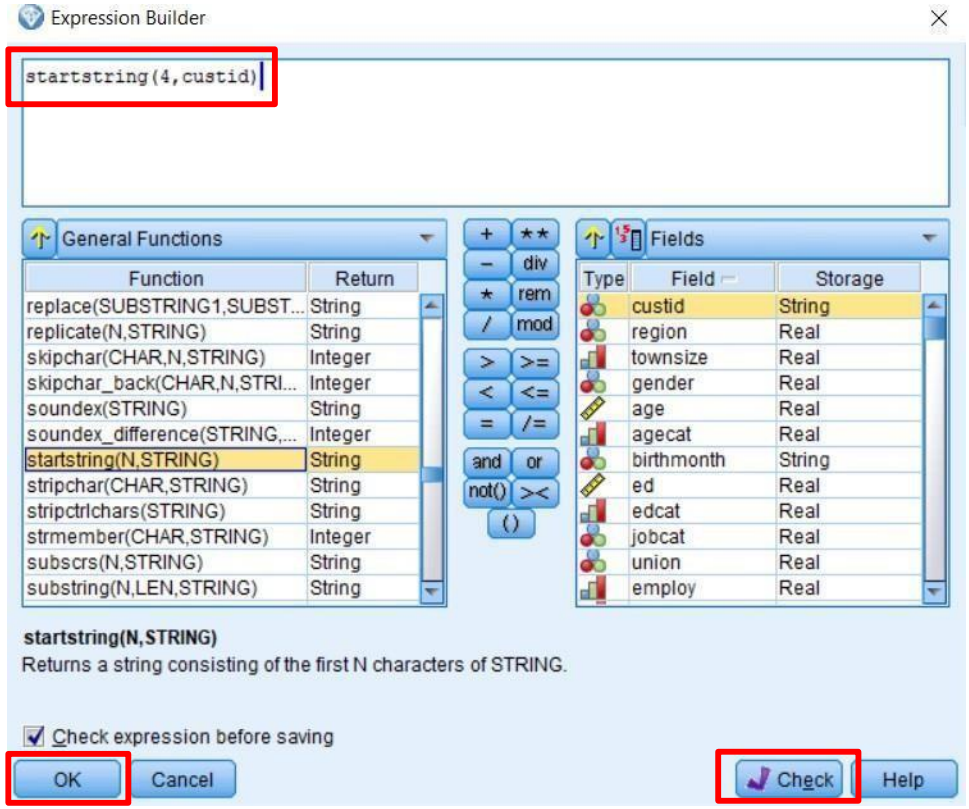
Bütün hüquqlar qorunur.

Text selection with CLEM

Task 1: Selecting the first 4 characters from the customer ID

3

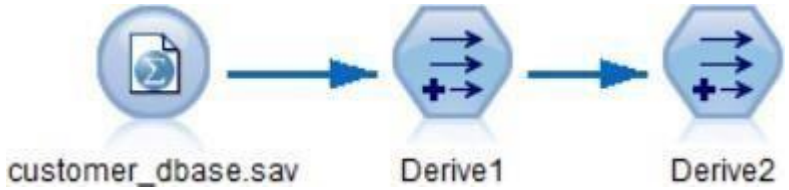
- Select **Starstring** function from **General Functions** section;
- Write "4" in the first part of the expression and select "custid" for the second part of expression;
- Click **Check** and click **OK**;
- Click **OK**.



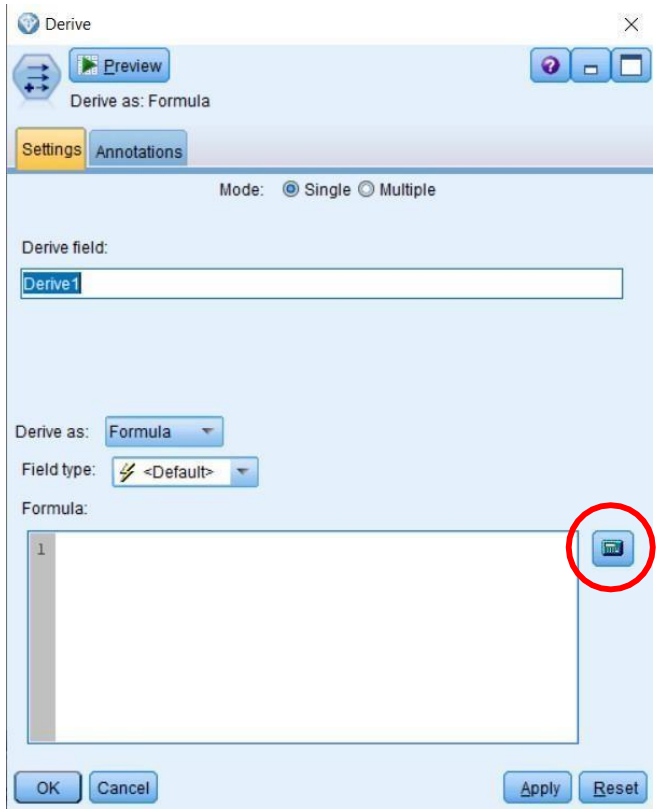
Bütün hüquqlar qorunur.

Text selection with CLEM

- Add the second Derivenode;
- Double click the **Derive** node on the canvas;
- Click button



1



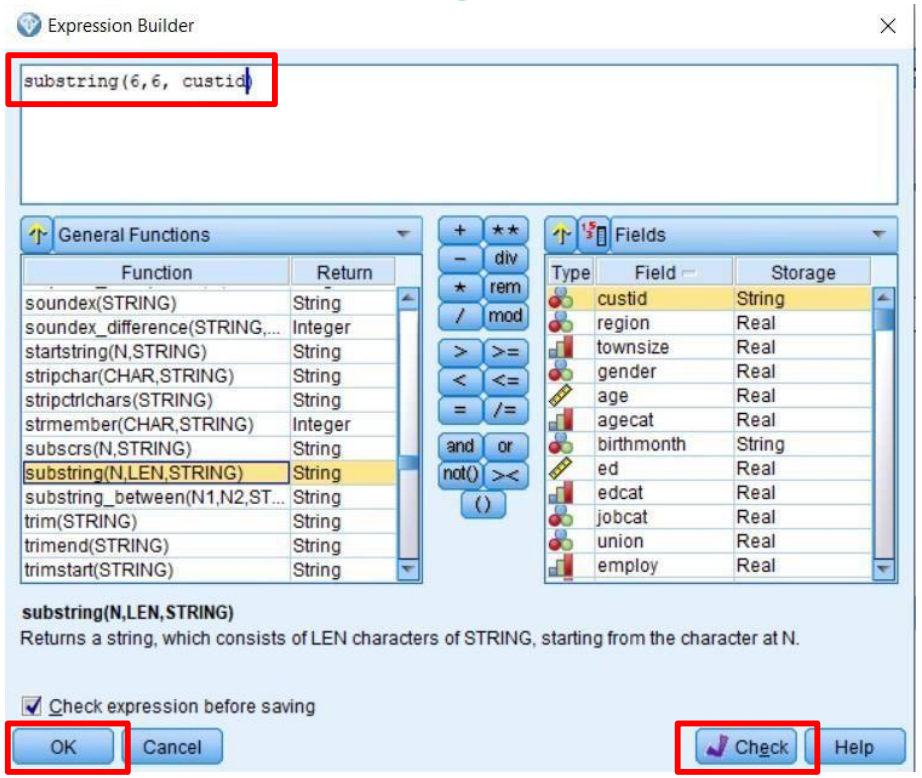
Bütün hüquqlar qorunur.

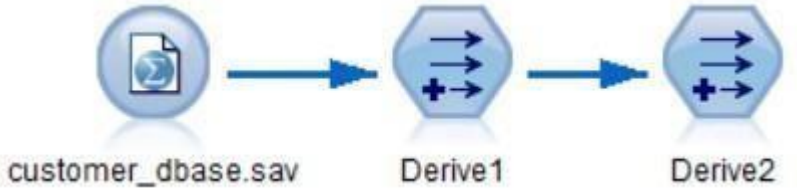
Text selection with CLEM

Task 2: Selecting the middle 6 characters from the customer ID.

2

- Select **Substring** function from **General Functions** section;
- Write "6" in the first part of expression;
- Write "6" in the second part of expression and select "custid" for the third part of expression;
- Click **Check** and click **Ok;**
- Click **OK.**



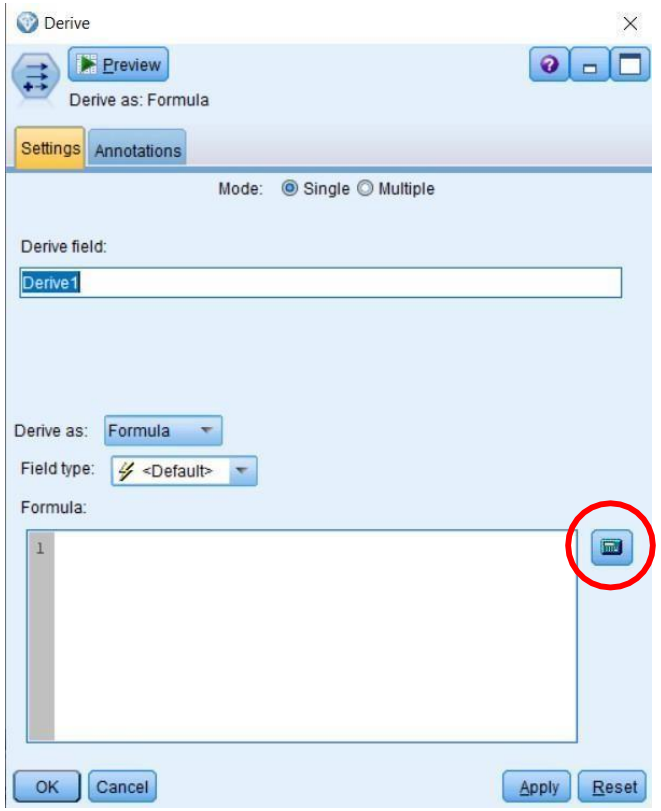


Bütün hüquqlar qorunur.

Text selection with CLEM

- Add the third Derive node;
- Double click the **Derive** node on the canvas;
- Click button

1

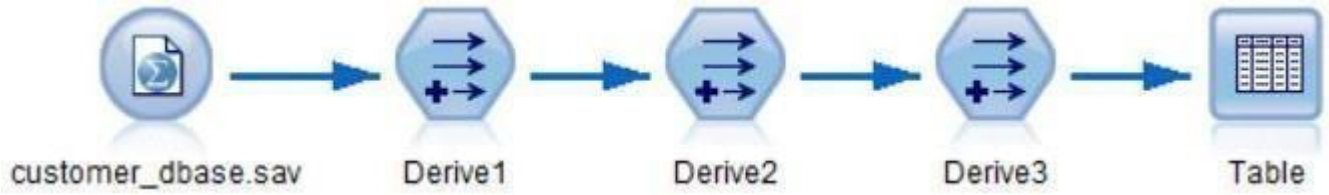


Bütün hüquqlar qorunur.

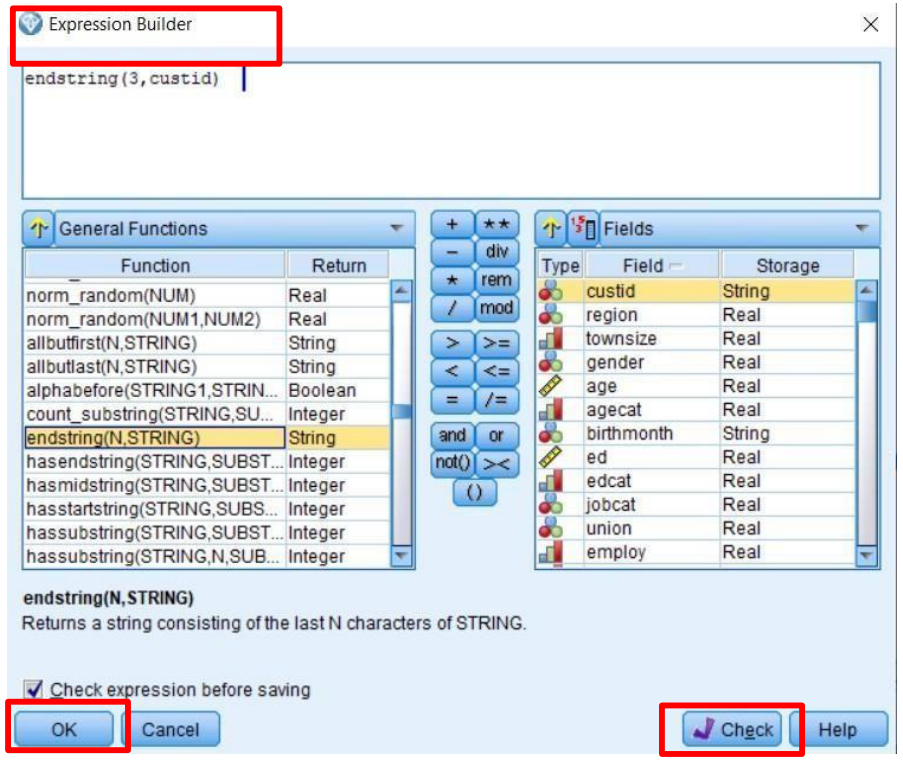
Text selection with CLEM

Task 3: Selecting the last 3 characters from the customer ID.

- Select **Endstring** function from **General Functions** section;
- Write "3" in the first part of expression and select "custid" for the second part of expression;
- Click **Check** and click **Ok;**
- Click **OK.**
- Add **Table** node from **Output** palette;
- Click **Run** ( ) button from the toolbar



2



Bütün hüquqlar qorunur.

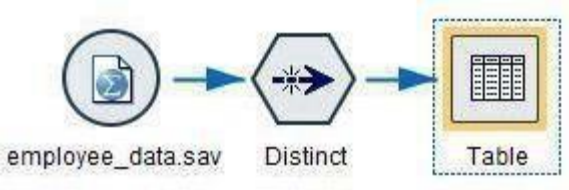
Working with Date Column

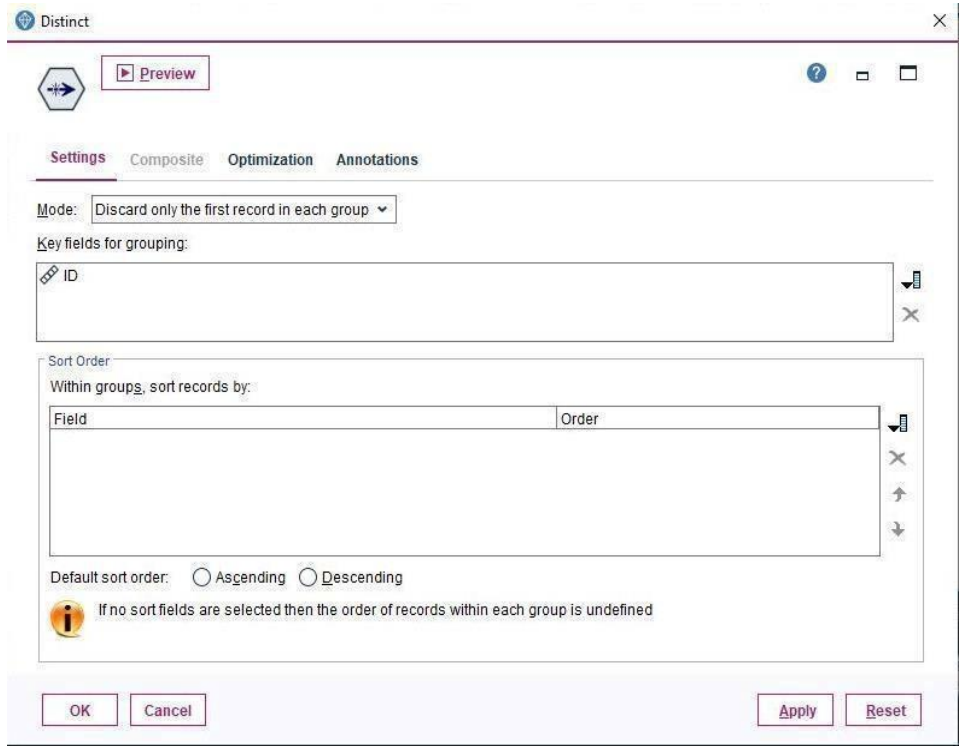
Formula	Description
datetime_year(DATE)	Returns the year from a DATE or timestamp. The result is an integer such as 2002
datetime_month(DATE)	Returns the month from a DATE or timestamp. The result is an integer in a range 1 to 12
datetime_day(DATE)	Returns the day from a given DATE or timestamp. The result is an integer in the range 1 to 31

Bütün hüquqlar qorunur.

Exploring Duplicates

- Open “Employee data.sav”;
- Select **Distinct** node from **Record Ops** palette;
- Open **Mode** to “Discard only the record in each group”;
- Select “id” for **Key fields for grouping** field;
- Connect Table node;

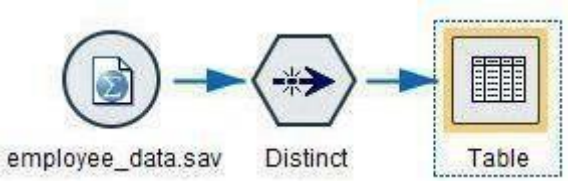


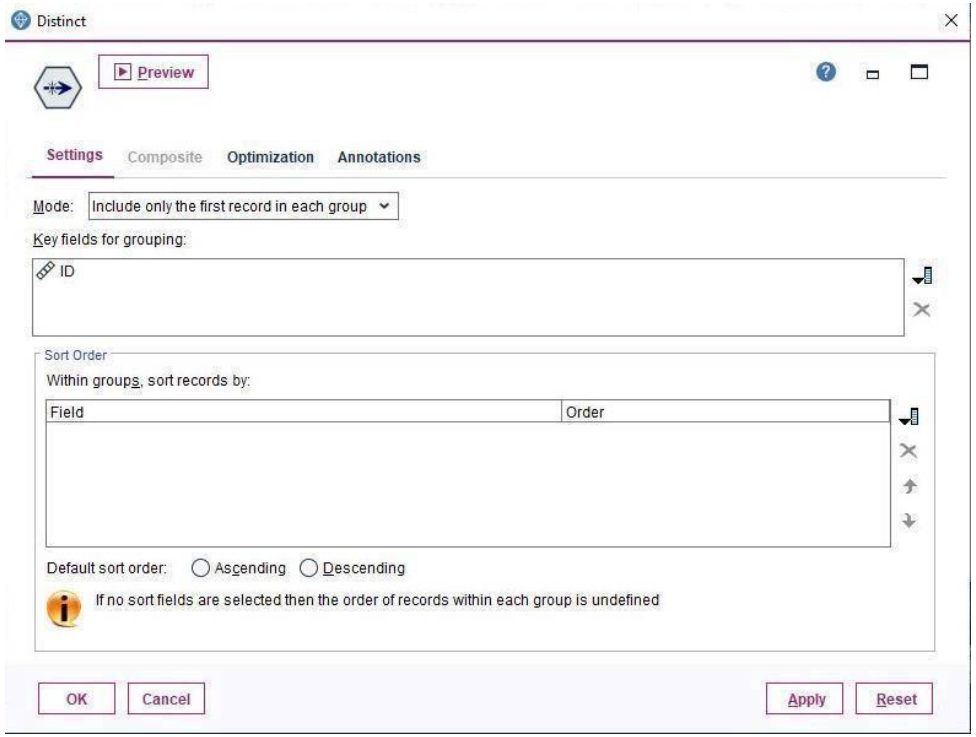


Bütün hüquqlar qorunur.

Deleting Duplicates

- Open “Employee data.sav”;
- Select **Distinct** node from **Record Ops** palette;
- Open **Edit menu** of **Distinct** node;
- Change **Mode** to “Include only first record in each group”;
- Select “id” for **Key fields for grouping** field;
- Connect Table node

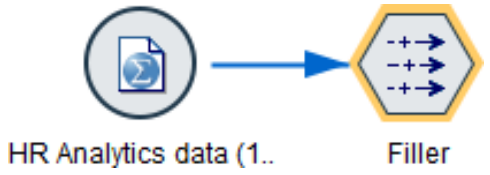


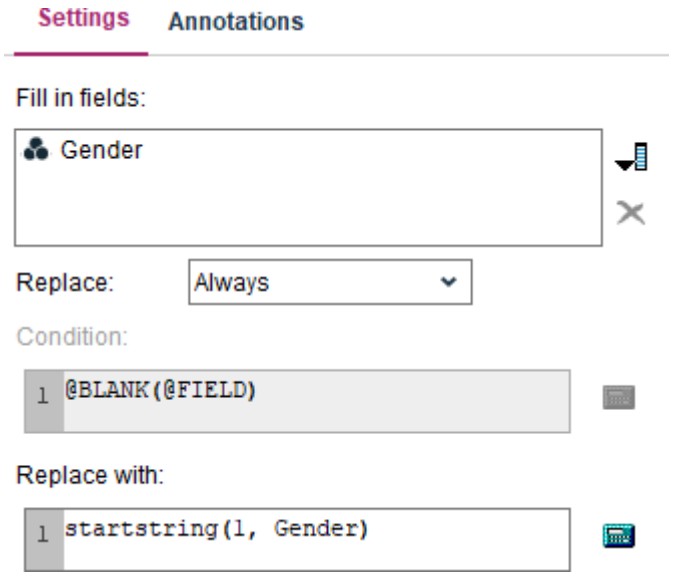


Bütün hüquqlar qorunur.

Replacing Values

- Open “HR Analytics data.sav”;
- Select **Filter** node from **Field Ops** palette;
- Open **Edit menu** of Filler node;
- Write “Gender” in **fill in fields;**
- Select **Always** in **Replace** section;
- Write “startstring(1,Gender)” in **Replace with** section;
- Click Preview





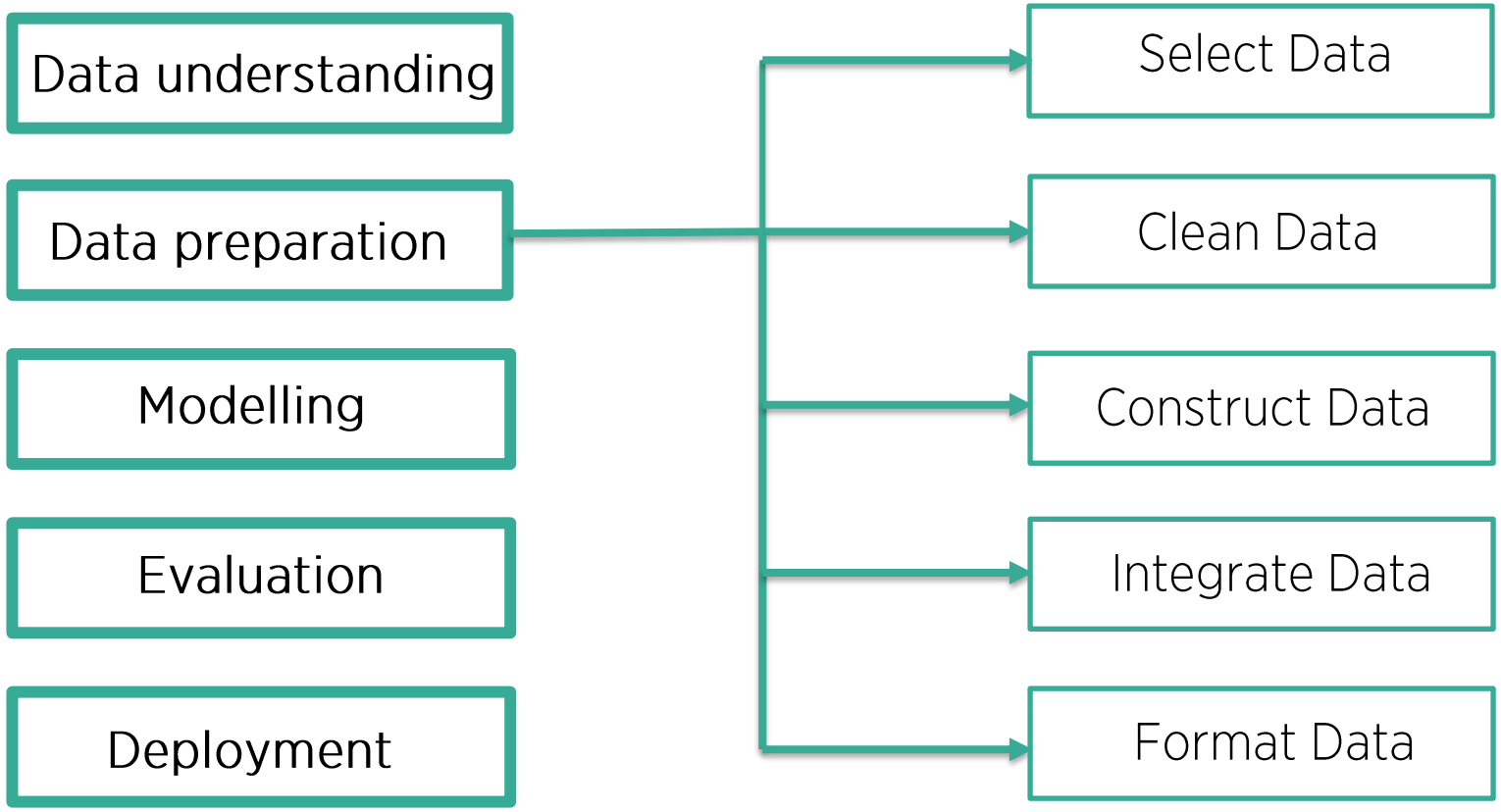
Bütün hüquqlar qorunur.

III

Data Preparation (Feature extraction)

Data Science Academy

Data Preparation



Bütün hüquqlar qorunur.

Data Preparation

- Data preparation is one of the most important and often time-consuming aspects of data mining
- In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort
- Devoting adequate energy to the earlier business understanding and data understanding phases can minimize this overhead, but you still need to expand a good amount of effort preparing and packaging the data for mining

Bütün hüquqlar qorunur.

Selecting Data

Based upon the initial data collection conducted in the previous CRISP-DM phase, you are ready to begin selecting the data relevant to your data mining goals. Generally, there are two ways to select data:

- Selecting items (rows) involves making decisions such as which accounts, products, or customers to include.
- Selecting attributes or characteristics (column involves making decisions about the use of characteristics such as transaction amount or household income

Bütün hüquqlar qorunur.

Including or Excluding Data

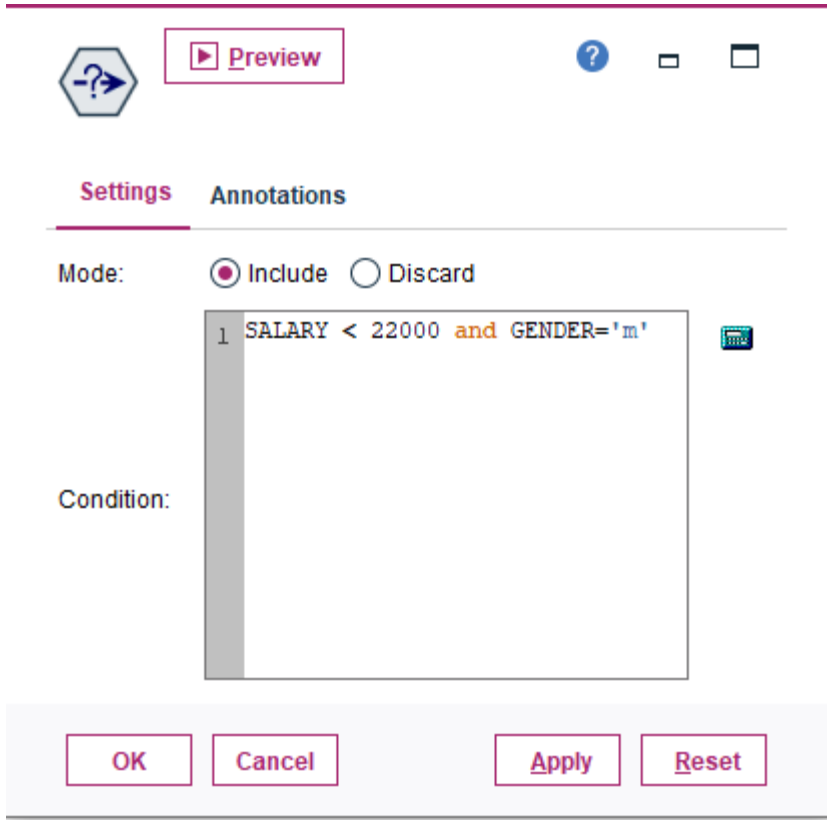
Questions to Consider

- Is a given attribute relevant to you data mining goals?
- Does the quality of a particular data set or attribute preclude the validity of your results?
- Can you salvage such data?
- Are there any constraints on using particular fields such as gender or race?

Bütün hüquqlar qorunur.

Filtering out records using Select node

- Open “Employee_data.sav”;
- Connect **Select** node from **Record Ops** palette;
- Open **Edit menu** of Select node;
- Change **Mode** to “Include”;
- Write “Salary < 22000 and GENDER=‘m’” to **Condition** field;
- Click **Preview**;



Bütün hüquqlar qorunur.

Cleaning Data

Cleaning your data involves taking a closer look at the problems in the data that you've chosen to include for analysis. There are several ways to clean data using the Record and Field Operation nodes in IBM SPSS Modeler.

Data Problem	Possible Solution
Missing data	Exclude rows or characteristics. Or, fill blanks with an estimated value
Data errors	Use logic to manually discover errors and replace. Or, exclude characteristics.
Coding inconsistencies	Decide upon a single coding scheme, then convert and replace values.
Missing or bad metadata	Manually examine suspect field and track down correct meaning

Bütün hüquqlar qorunur.

Cleaning Data

- Raise the data quality to the level required by the selected analysis techniques.
- This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modelling

Bütün hüquqlar qorunur.

Constructing New Data

It is frequently the case that you'll need to construct new data. For example, it may be useful to create a new column flagging the purchase of an extended warranty for each transaction. This new field, purchased_warranty, can easily be generated using a Set to Flag node in IBM SPSS Modeler.

There are two ways to construct new data:

- Deriving attributes (columns or characteristics)
- Generating records (rows)

Bütün hüquqlar qorunur.

Deriving Attributes

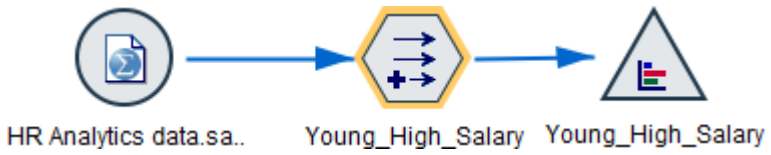
Task List

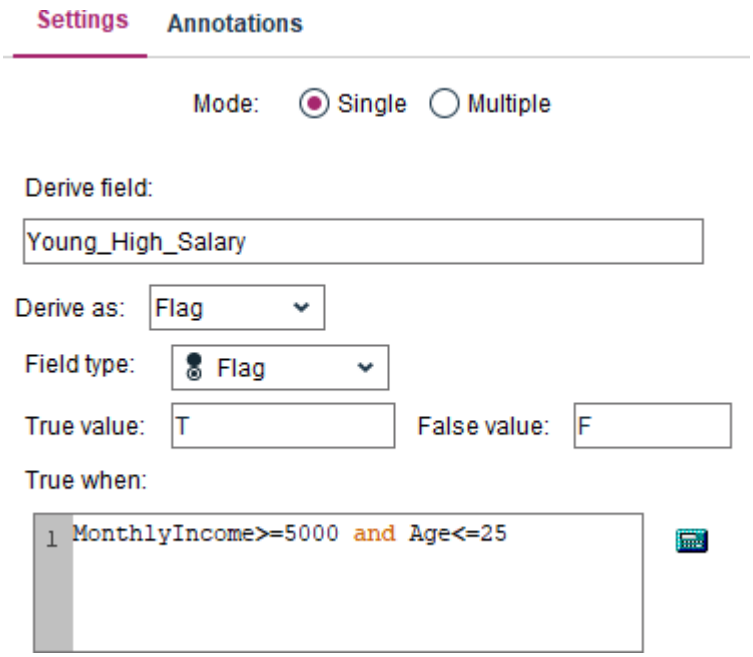
- Consider the data requirements for modeling when deriving attributes. Does the modeling algorithm expect a particular type of data, such as numeric? If so, perform the necessary transformations.
- Do the data need to be normalized before modeling?
- Can missing attributes be constructed using aggregation, averaging, or induction?
- Based upon your background knowledge, are there important facts (such as length of time spent as the Web site) that can be derived from existing fields?

Bütün hüquqlar qorunur.

Deriving new flag variables

- Open “Hr Analytics data.sav”;
- Select **Derive** node from **Field Ops** palette;
- Open **Edit menu** of **Derive** node;
- Write “Young_High_Salary” in **Derive field**;
- Select **Flag** in **Derive as** section;
- Write “MonthlyIncome>5000 and Age<25” in **True when** section;
- Add distribution plot to display frequencies of new variable

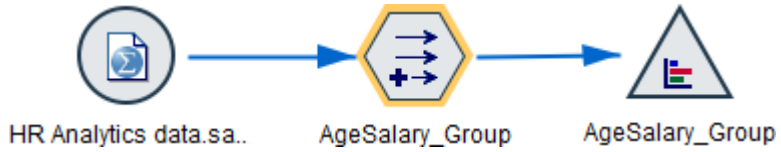


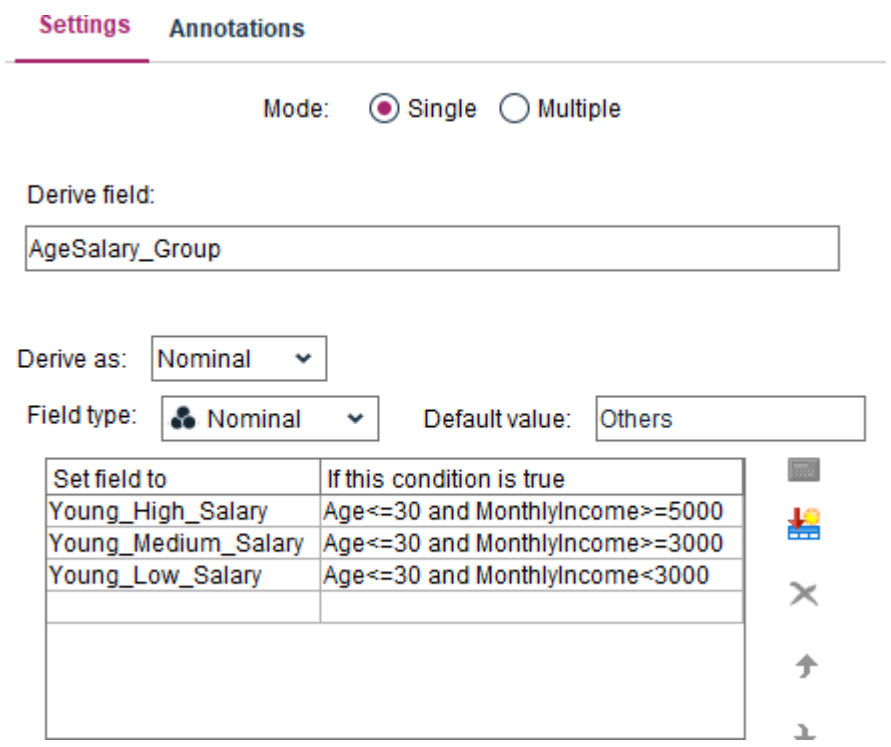


Bütün hüquqlar qorunur.

Deriving new nominal variables

- Open “Hr Analytics data.sav”;
- Select **Derive** node from **Field Ops** palette;
- Open **Edit menu** of **Derive** node;
- **Derive field:** “AgeSalary_Group”;
- **Derive as:** Nominal
- **Default value:** Others
- **Set field to:** Young_High_Salary, **If this condition is true:** MonthlyIncome>5000 and Age<25.....continue as shown.
- Add distribution plot to display frequencies of new variable

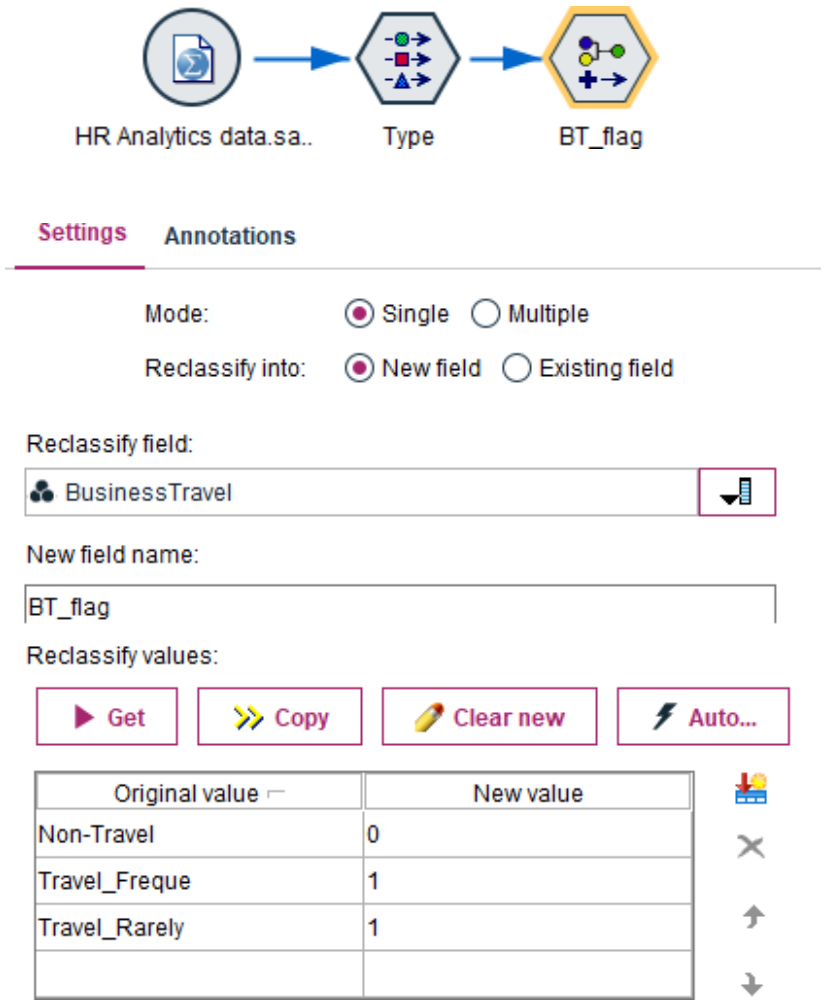




Bütün hüquqlar qorunur.

Recording nominal variables

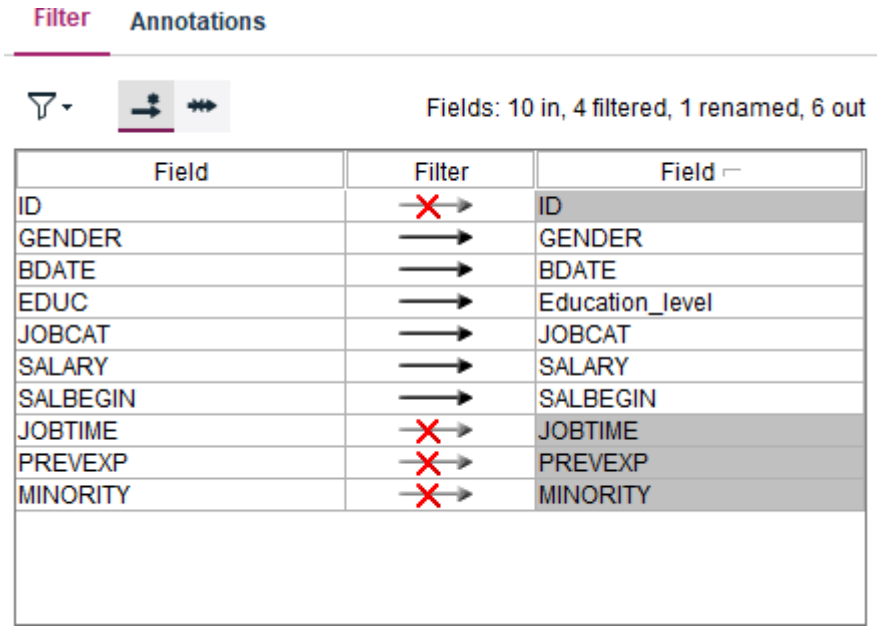
- Open “Hr Analytics data.sav”;
- Select **Reclassify** node from **Fields Ops** palette;
- Open **Edit menu** of **Reclassify** node;
- **Reclassify field:** “Business Travel”;
- **New field name:** BT_flag
- **Reclassify values:** Get
- Non-Travel -> 0
- Travel_Freque -> 1
- Travel_Rarely ->1



Bütün hüquqlar qorunur.

Filtering out fields using Filter node

- Open “Employee_data.sav”;
- Connect **Filter** node from **Field Ops** palatte;
- Open **Edit menu** of Filter node;
- Click on **arrow** button on *ID, JOBTIME, PREVEXP, MINORITY* row.
- Rename *EDUC* in **Field** as “Education_level”;
- Click Preview

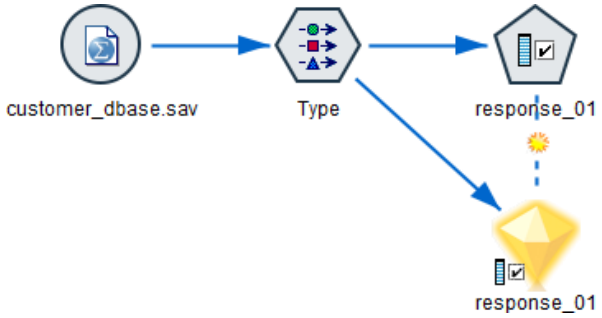


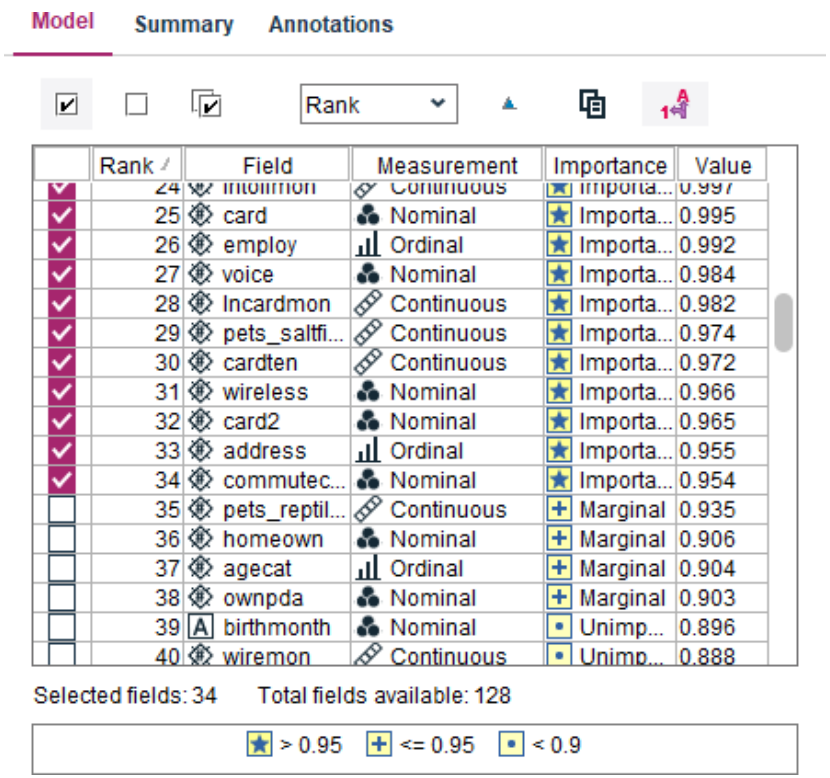
Field	Filter	Field
ID	✗	ID
GENDER	→	GENDER
BDATE	→	BDATE
EDUC	→	Education_level
JOBCAT	→	JOBCAT
SALARY	→	SALARY
SALBEGIN	→	SALBEGIN
JOBTIME	✗	JOBTIME
PREVEXP	✗	PREVEXP
MINORITY	✗	MINORITY

Bütün hüquqlar qorunur.

Variable Screening using Feature selection node

- Open “customer_dbase”;
- Connect **Type node** and Click **read the values** in edit menu;
- Change types if necessary;
- Connect **Feature selection** node from **Modeling** palette;
- Run the Stream;
- Double click to **model nugget;**





Rank	Field	Measurement	Importance	Value
24	intolmon	Continuous	Importa...	0.997
25	card	Nominal	Importa...	0.995
26	employ	Ordinal	Importa...	0.992
27	voice	Nominal	Importa...	0.984
28	lncardmon	Continuous	Importa...	0.982
29	pets_saltfi...	Continuous	Importa...	0.974
30	cardten	Continuous	Importa...	0.972
31	wireless	Nominal	Importa...	0.966
32	card2	Nominal	Importa...	0.965
33	address	Ordinal	Importa...	0.955
34	commutec...	Nominal	Importa...	0.954
35	pets_reptil...	Continuous	Marginal	0.935
36	homeown	Nominal	Marginal	0.906
37	agecat	Ordinal	Marginal	0.904
38	ownpda	Nominal	Marginal	0.903
39	birthmonth	Nominal	Unimp...	0.896
40	wiremon	Continuous	Unimp...	0.888

Bütün hüquqlar qorunur.

Thank you!

Data Science Academy