Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

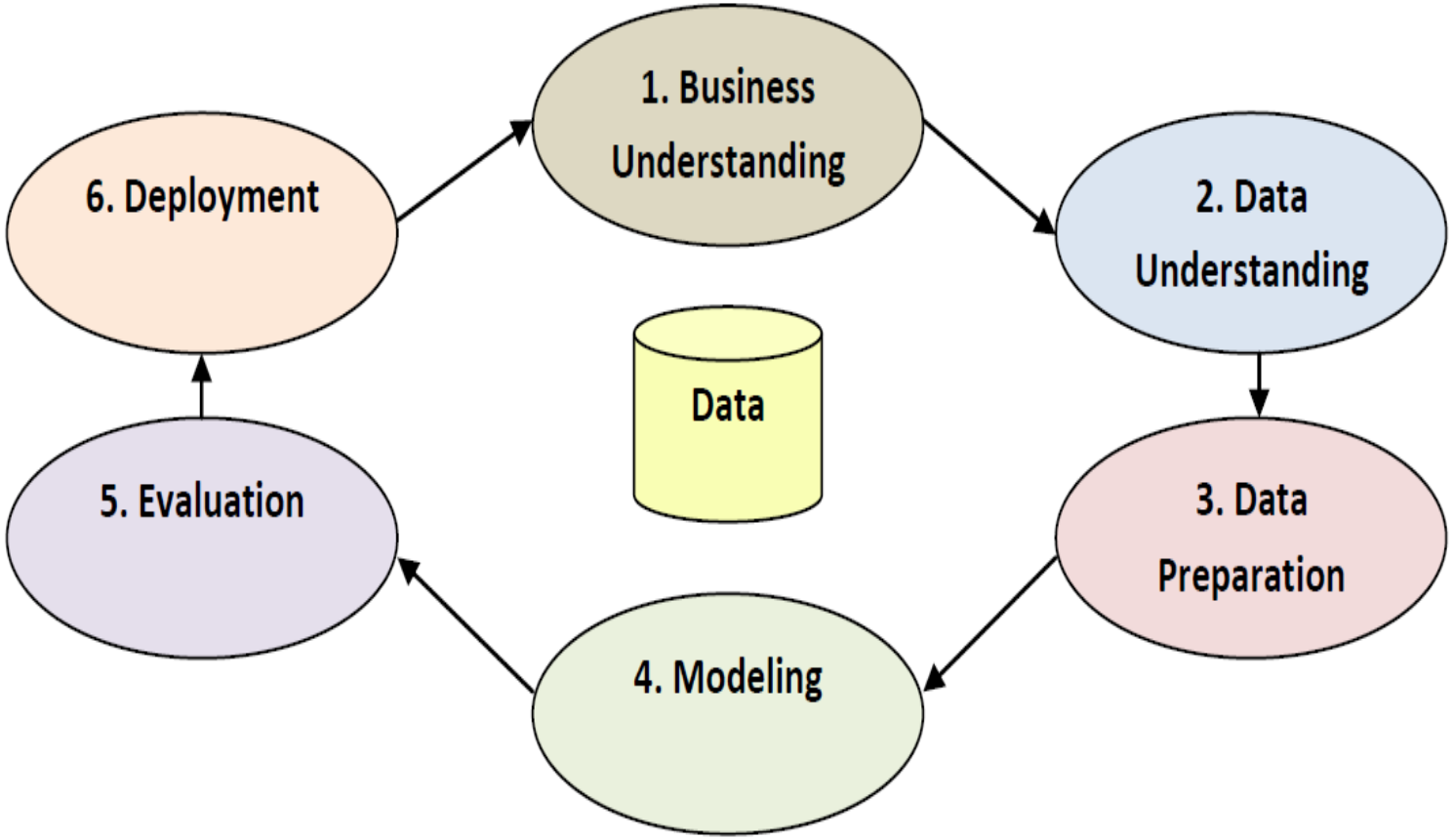
Παραδείγματα Προετοιμασίας Δεδομένων

Ενδεικτική Βιβλιογραφία

M. North, Data Mining for the Masses, 2012, ISBN: 978-0615684376

This book is licensed under a Creative Commons Attribution 3.0 License

CRISP-DM Conceptual Model
Cross-industry standard process for data mining



CRISP-DM Step 1:
Business (Organizational) Understanding

- Η κατανόηση συνδέεται με την απάντηση σε ερωτήσεις όπως:
- Πώς μπορούμε να αυξήσουμε το περιθώριο κέρδους ανά μονάδα προϊόντος;
- Πώς μπορούμε να προβλέψουμε και να διορθώσουμε ατέλειες κατασκευής έτσι ώστε να αποφύγουμε την αποστολή ενός ελαττωματικού προϊόντος;

Παράδειγμα.

- Έστω ότι εργάζεστε για εταιρεία που προμηθεύει πετρέλαιο θέρμανσης για οικιακή χρήση. Τα κύρια ερωτήματα μπορούν να διατυπωθούν ως εξής:

- Ποιοι παράγοντες σχετίζονται με τη χρήση του πετρελαίου θέρμανσης;
- Πώς η εταιρεία θα μπορούσε να χρησιμοποιήσει τη γνώση αυτών των παραγόντων για την καλύτερη διαχείριση των αποθεμάτων της, καθώς και την πρόβλεψη της ζήτησης;

CRISP-DM Step 2: Data Understanding

- Από πού προέρχονται τα δεδομένα που θα αναλύσουμε; Από ποιόν συλλέγονται; Χρησιμοποιήθηκε μια τυποποιημένη μέθοδος συλλογής (collection); Τι σημαίνουν οι διάφορες στήλες και οι γραμμές των δεδομένων; Υπάρχουν ακρωνύμια ή συντομογραφίες που είναι άγνωστα ή ασαφή;

Τα δεδομένα-παρατηρήσεις συλλέγονται από τα στοιχεία πελατών και πωλήσεων.

Σε ορολογία του εργαλείου Rapid Miner οι παρατηρήσεις λέγονται observations ή examples.

Μόνωση [1..10]	Θερμοκρασία Περιβάλλοντος (°F)	Κάτοικοι	Μέγεθος Οικίας [1..8]	Πετρέλαιο θέρμανσης	Ταυτότητα νοικοκυριού
6	74	4	8	1300	1
10	43	3	6	2000	2
3	81	5	8	1500	3
κ.τ.λ.					

CRISP-DM Step 3: Data Preparation

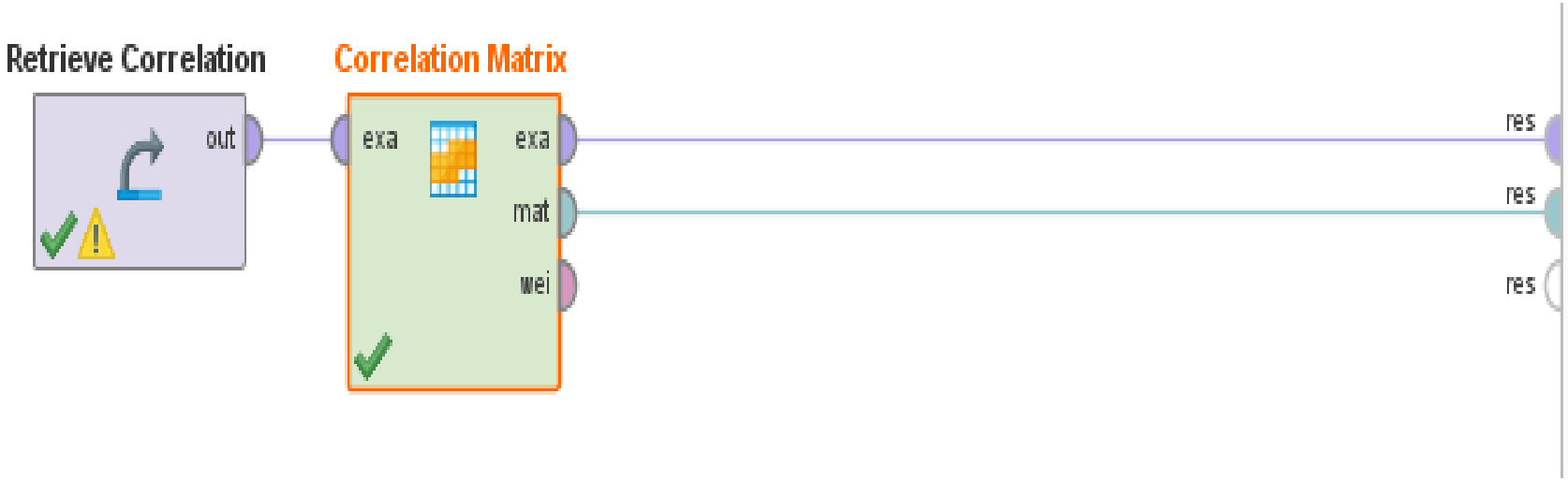
- Η Προετοιμασία των δεδομένων (Data Preparation) περιλαμβάνει μια σειρά από δραστηριότητες. Μπορεί να ενώνει δύο ή περισσότερα σύνολα δεδομένων, να περιορίζει σύνολα δεδομένων μόνον σε εκείνες τις μεταβλητές που έχουν ενδιαφέρον σε μια συγκεκριμένη περίπτωση εξόρυξης δεδομένων, να καθαρίζει δεδομένα από «ακραίες» παρατηρήσεις, να συμπληρώνει – διαχειρίζεται ελλιπή δεδομένα, να μορφοποιεί εκ νέου δεδομένα για λόγους συνέπειας κ.λπ.
- Στο παράδειγμά μας βλέπουμε ελλιπή δεδομένα και λανθασμένα

- Ελλιπή και εσφαλμένα δεδομένα

Μόνωση [1..10]	Θερμοκρασία Περιβάλλοντος (°F)	Κάτοικοι	Μέγεθος Οικίας [1..8]	Πετρέλαιο θέρμανσης	Ταυτότητα νοικοκυριού
6	74	4	8	1300	1
10	43	3	6	2000	2
3	81	?	10	1500	3
κ.τ.λ.					

CRISP-DM Step 4: Modeling

- Απλουστεύοντας λίγο, ένα μοντέλο στην εξόρυξη δεδομένων είναι μια ηλεκτρονική αναπαράσταση παρατηρήσεων – μετρήσεων (observations) του πραγματικού κόσμου. Τα μοντέλα προκύπτουν από την εφαρμογή αλγορίθμων που «αναλαμβάνουν» την αναζήτηση, τον εντοπισμό, και την εμφάνιση προτύπων ή μηνυμάτων στα δεδομένα.
- Υπάρχουν δύο βασικά είδη μοντέλων εξόρυξης: εκείνα που ταξινομούν (classify) και εκείνα που προβλέπουν (predict).


Retrieve Correlation
out
Correlation Matrix
exa
exa
mat
wei
res
res
res

CRISP-DM Step 5: Evaluation

Attributes	Insulation	Temperature	Heating Oil	Num Occu...	Avg Age	Home Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num Occup	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg Age	0.643	-0.673	0.848	-0.048	1	0.307
Home Size	0.201	-0.214	0.381	-0.023	0.307	1

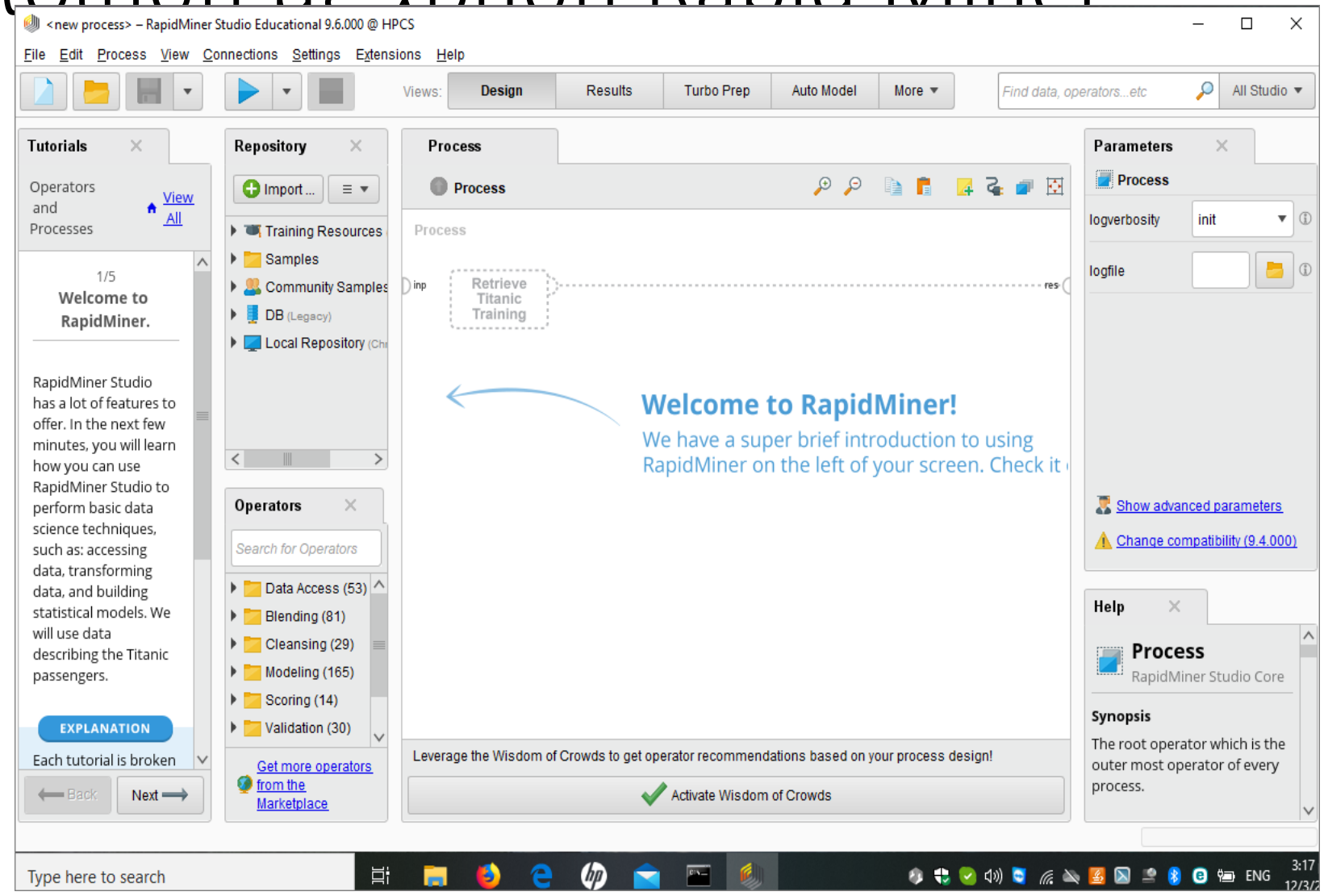
CRISP-DM Step 6: Deployment

Ενέργειες βασιζόμενες σε ότι μάθαμε από το μοντέλο.

Η μόνωση συσχετίζεται με τα άλλα χαρακτηριστικά. Ευκαιρία να ξεκινήσουμε νέες δραστηριότητες ή συνεργασίες με άλλες εταιρείες που θα εστιάζουν σε μονώσεις.

Θυμηθείτε την κυκλική, επαναληπτική φύση του CRISP-DM.

Υλοποίηση με χρήση Rapid Miner



παράδειγμα: DM and RapidMiner

- Θα μελετήσουμε μια εταιρεία που προμηθεύει πετρέλαιο θέρμανσης για το σπίτι σε πανεθνική κλίμακα. Ένας περιφερειακός διευθυντής πωλήσεων αισθάνεται την ανάγκη να κατανοήσει τα είδη των συμπεριφορών και άλλους παράγοντες που μπορεί να επηρεάσουν τη ζήτηση για πετρέλαιο θέρμανσης στην εγχώρια αγορά.
- Προβληματίζεται κυρίως λόγω της μεταβλητότητας των τιμών της αγοράς πετρελαίου θέρμανσης, σε συνδυασμό με τη μεγάλη μεταβλητότητα στο μέγεθος των παραγγελιών για το σπίτι του πετρελαίου θέρμανσης.

  (M. North, Data Mining for the Masses)

DM and RapidMiner

Το κύριο ερώτημα μπορεί να διατυπωθεί ως εξής:

- Ποιοι παράγοντες σχετίζονται με τη χρήση του πετρελαίου θέρμανσης, και πώς η εταιρεία θα μπορούσε να χρησιμοποιήσει τη γνώση αυτών των παραγόντων για την καλύτερη διαχείριση των αποθεμάτων της, καθώς και την πρόβλεψη της ζήτησης;

DM and RapidMiner

- Υπάρχουν πολλοί παράγοντες που επηρεάζουν την κατανάλωση πετρελαίου θέρμανσης. Πιστεύουμε ότι η διερεύνηση της σχέσης μεταξύ ορισμένων από τους παράγοντες αυτούς, θα βοηθήσει την εταιρεία να παρακολουθεί καλύτερα και να ανταποκριθεί στη ζήτηση του πετρελαίου θέρμανσης.
- Στην έρευνά μας έχει επιλεγεί η στατιστική συσχέτιση (correlation) ως ένας απλός τρόπος μοντελοποίησης της σχέσης μεταξύ των παραγόντων που η εταιρεία επιθυμεί να διερευνήσει.

DM and RapidMiner

- Η συσχέτιση - correlation είναι ένα στατιστικό μέτρο που περιγράφει πόσο ισχυρές είναι οι σχέσεις μεταξύ των ιδιοτήτων σε ένα σύνολο δεδομένων ("**Correlation** is a statistical measure of how strong the relationships are between attributes in a data set").
- Ακολουθεί η δημιουργία ενός πίνακα συσχέτισης έξι χαρακτηριστικών. Τα δεδομένα που χρησιμοποιούνται αντλούνται κυρίως από το σύστημα τιμολόγησης πελατών (billing) της εταιρείας.

DM and RapidMiner

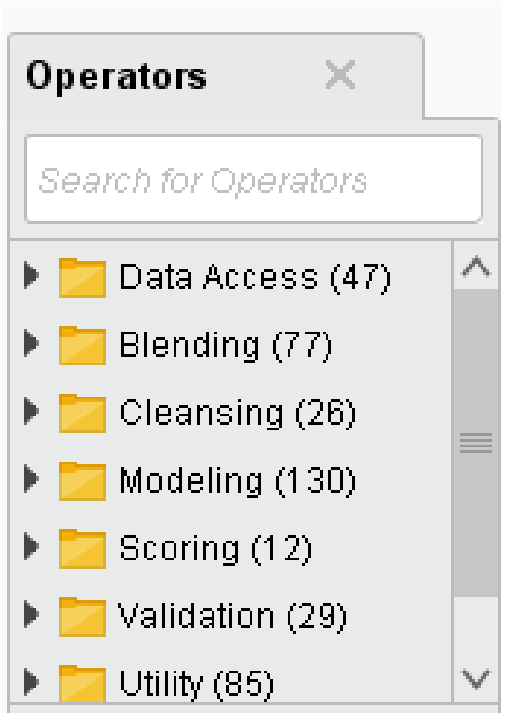
1) Insulation: Είναι μια εκτίμηση που κυμαίνονται από ένα έως δέκα και αναφέρεται στη μόνωση κάθε σπιτιού. Ένα σπίτι με βαθμολογία ένα δεν είναι καλά μονωμένο, ενώ ένα σπίτι με μια βαθμολογία των δέκα βαθμών έχει εξαιρετική μόνωση.
2) Temperature: Μέση εξωτερική θερμοκρασία περιβάλλοντος σε κάθε σπίτι για το προηγούμενο έτος. Μετράται σε βαθμούς Fahrenheit.
3) Heating_Oil: Συνολικός αριθμός μονάδων λίτρων πετρελαίου θέρμανσης που αγοράστηκαν από τον ιδιοκτήτη του σπιτιού το προηγούμενο έτος.
4) Num_Occupants: Συνολικός αριθμός των ενοίκων σε κάθε σπίτι.
5) Avg_Age: Μέση ηλικία ενοίκων.
6) Home_Size: Βαθμολογία, σε μια κλίμακα από ένα έως οκτώ του που αναφέρεται στο συνολικό μέγεθος του σπιτιού. Όσο μεγαλύτερος ο αριθμός τόσο μεγαλύτερο το σπίτι.

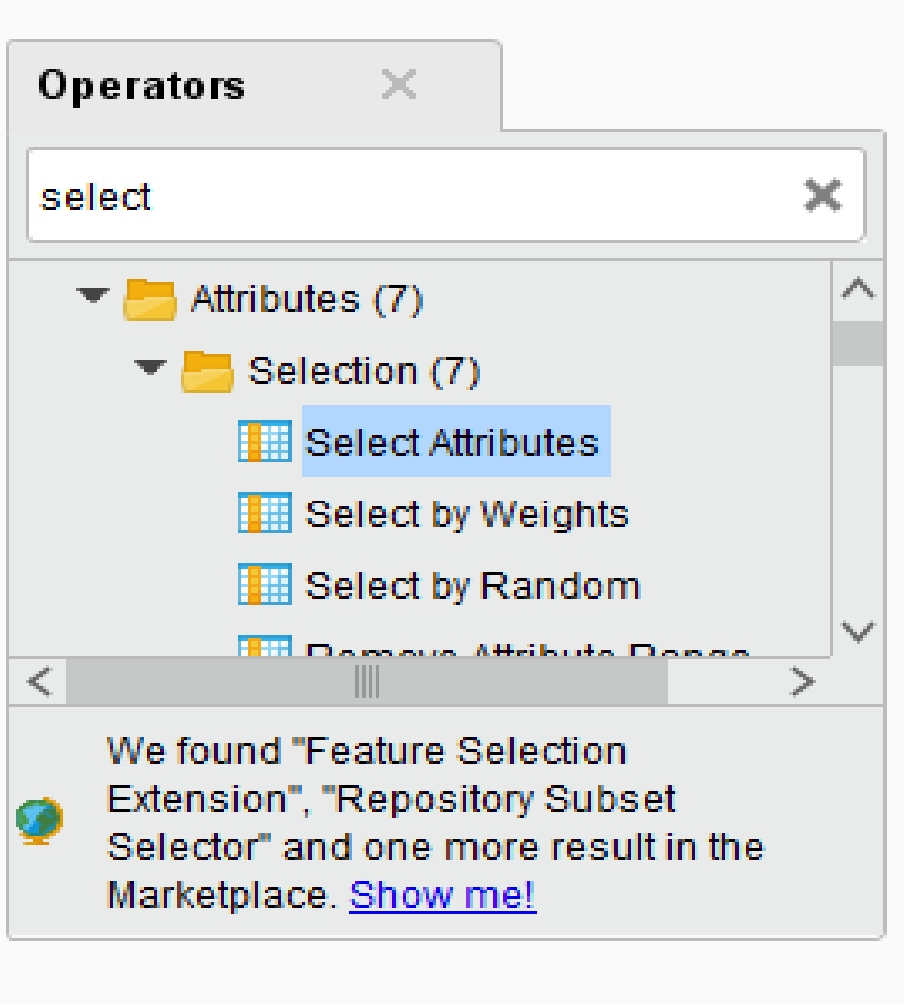
$$\frac{(F-32)*5}{9} = C$$

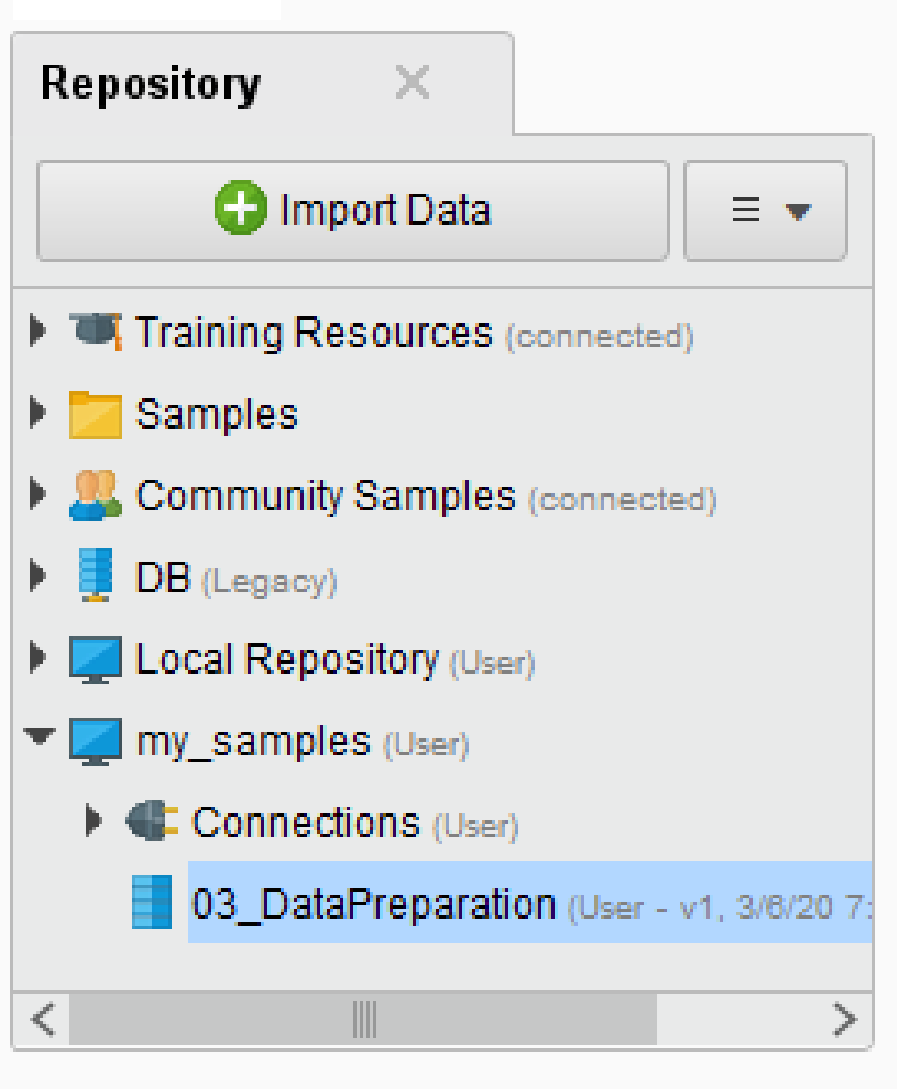
https://sites.google.com/site/dataminingforthemasses/

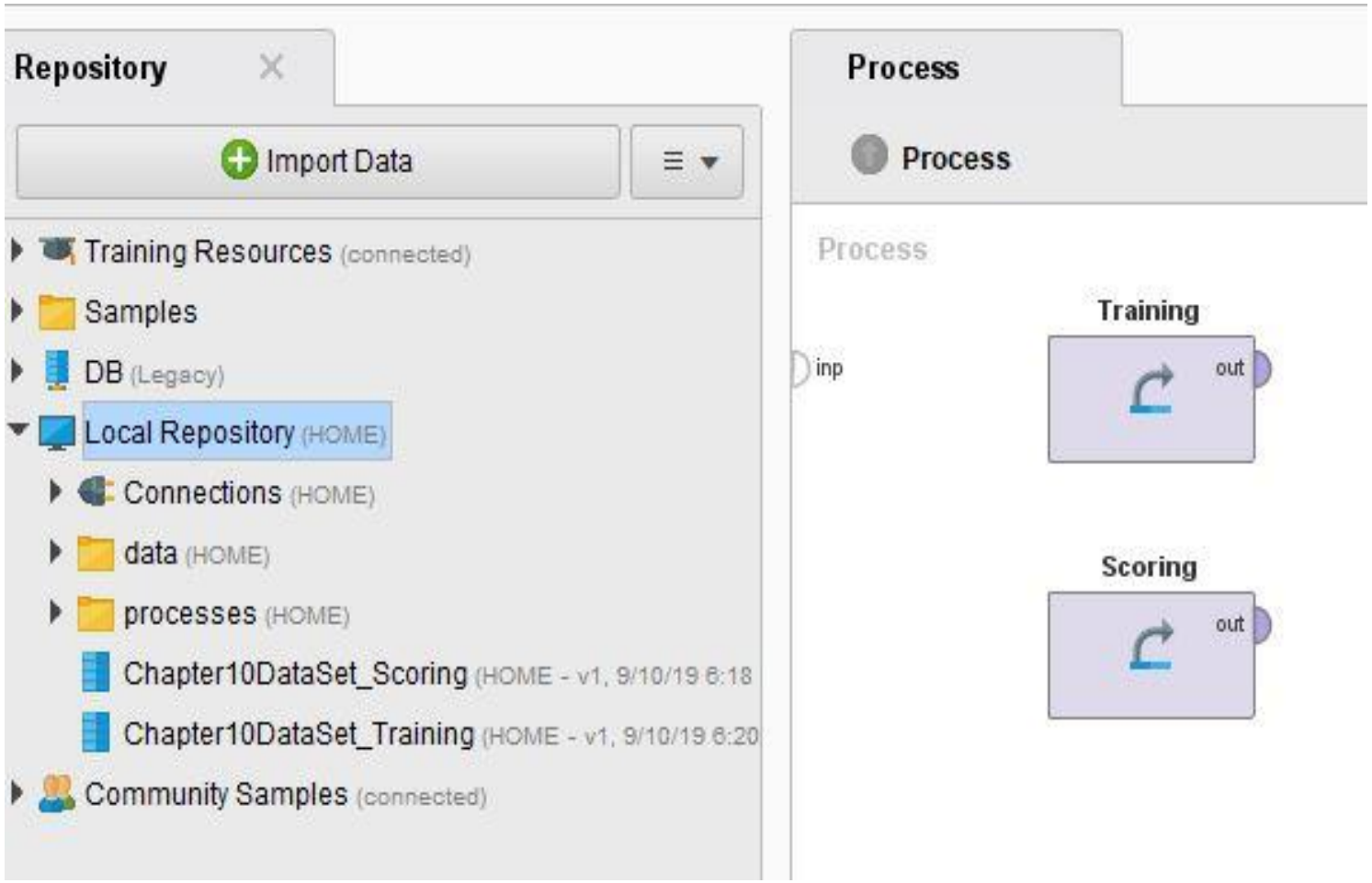
Insulation,Temperature,Heating_Oil,Num_Occupants,Avg_Age,Home_Size
6,74,132,4,23.8,4
10,43,263,4,56.7,4
3,81,145,2,28.0,6
9,50,196,4,45.1,3
2,80,131,5,20.8,2
5,76,129,3,21.5,3
5,72,131,4,23.5,3
6,88,161,2,38.2,6
5,77,184,3,42.5,3
10,42,225,3,51.1,1
6,90,178,2,42.1,2
3,83,121,1,19.8,2
10,43,186,5,45.1,6
8,59,206,2,50.1,8

Ορολογία εργαλείου Rapid Miner

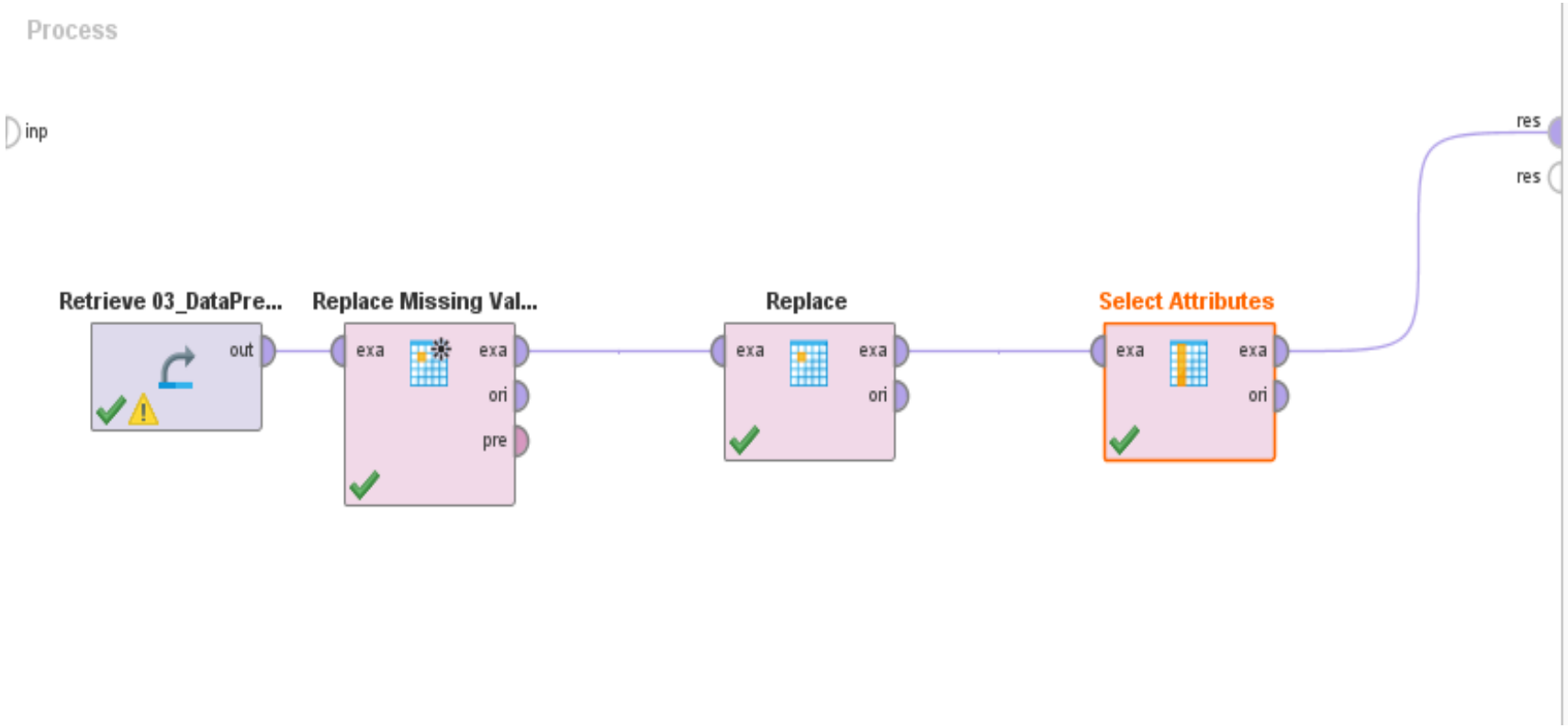




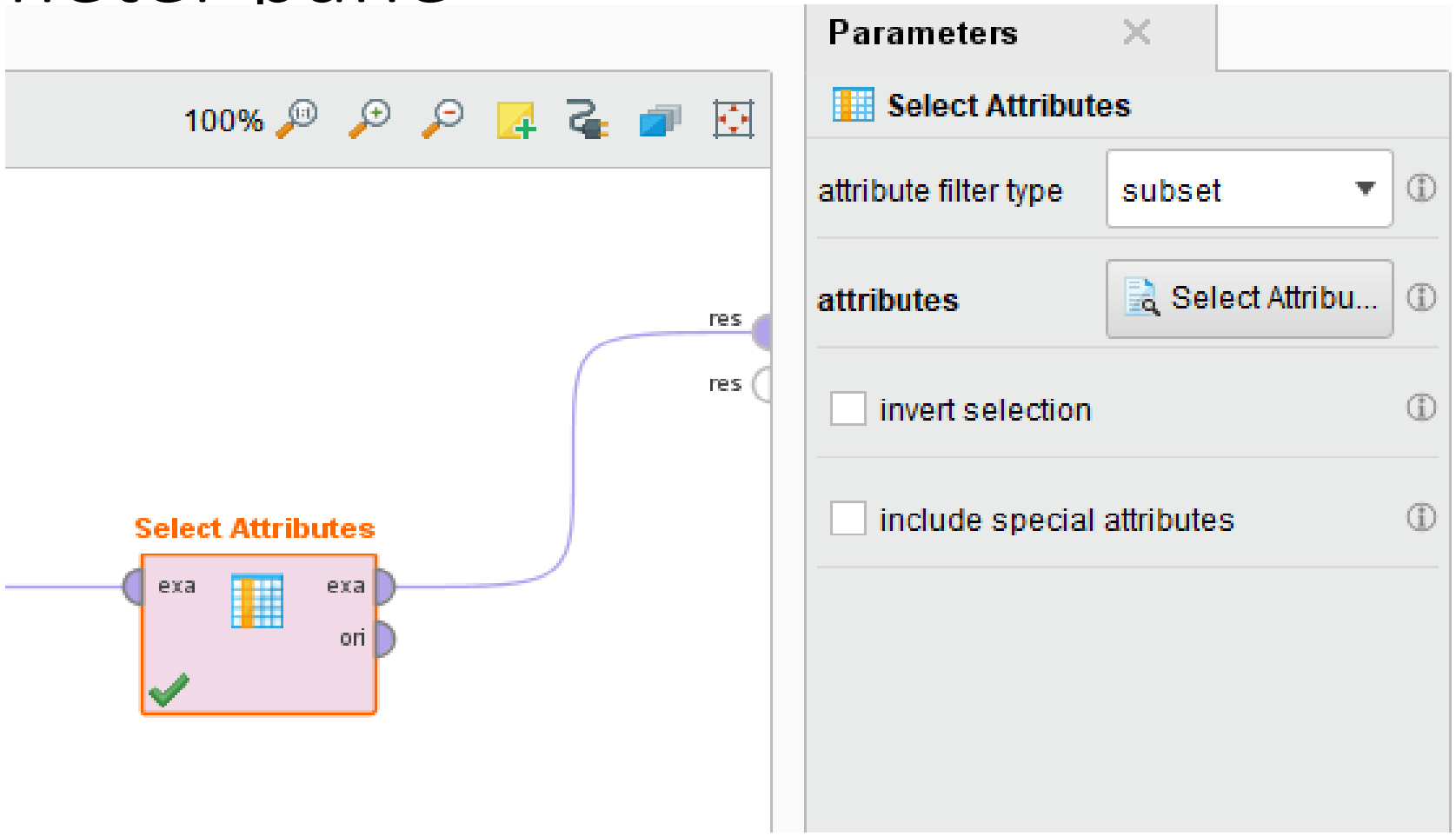

Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
DB (Legacy)
Local Repository (User)
my_samples (User)
Connections (User)
03_DataPreparation (User - v1, 3/6/20 7:


Repository
Import Data
Training Resources (connected)
Samples
DB (Legacy)
Local Repository (HOME)
Connections (HOME)
data (HOME)
processes (HOME)
Chapter10DataSet_Scoring (HOME - v1, 9/10/19 6:18
Chapter10DataSet_Training (HOME - v1, 9/10/19 6:20
Community Samples (connected)
Process
Process
Process
inp
Training
out
Scoring
out

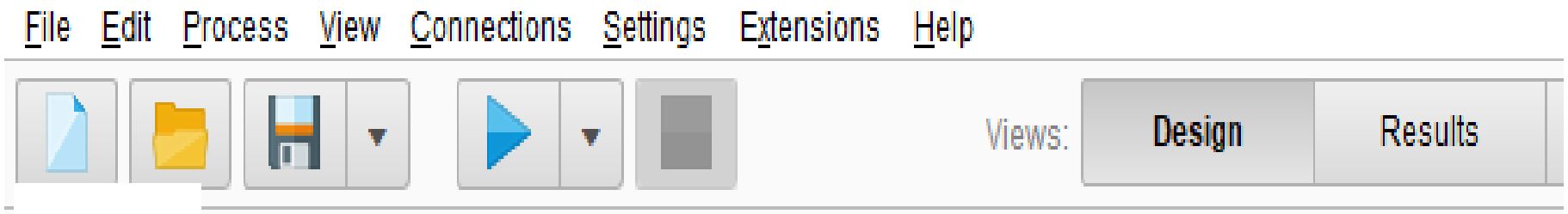
Παράδειγμα μοντέλου (Rapid Miner process)



Parameter pane



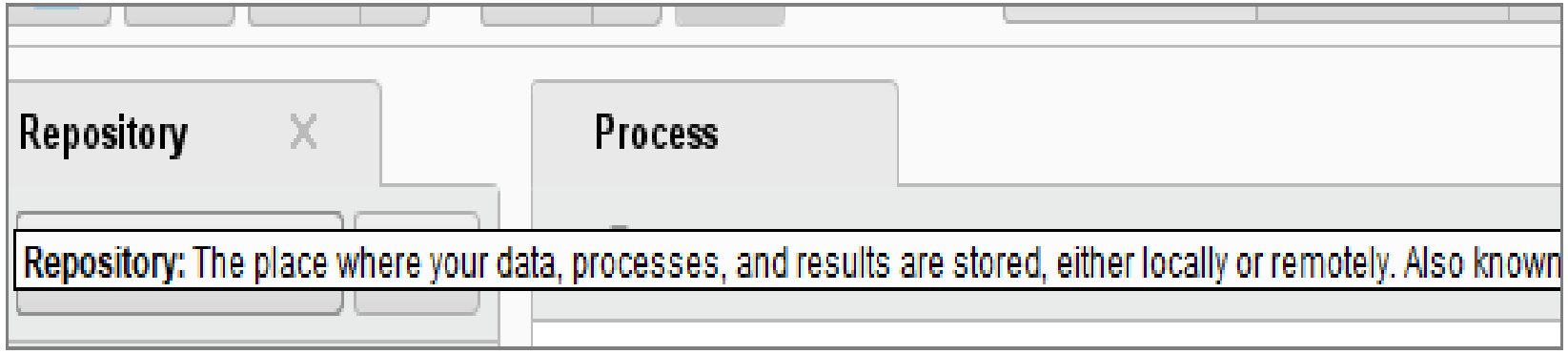
Βασικό μενού



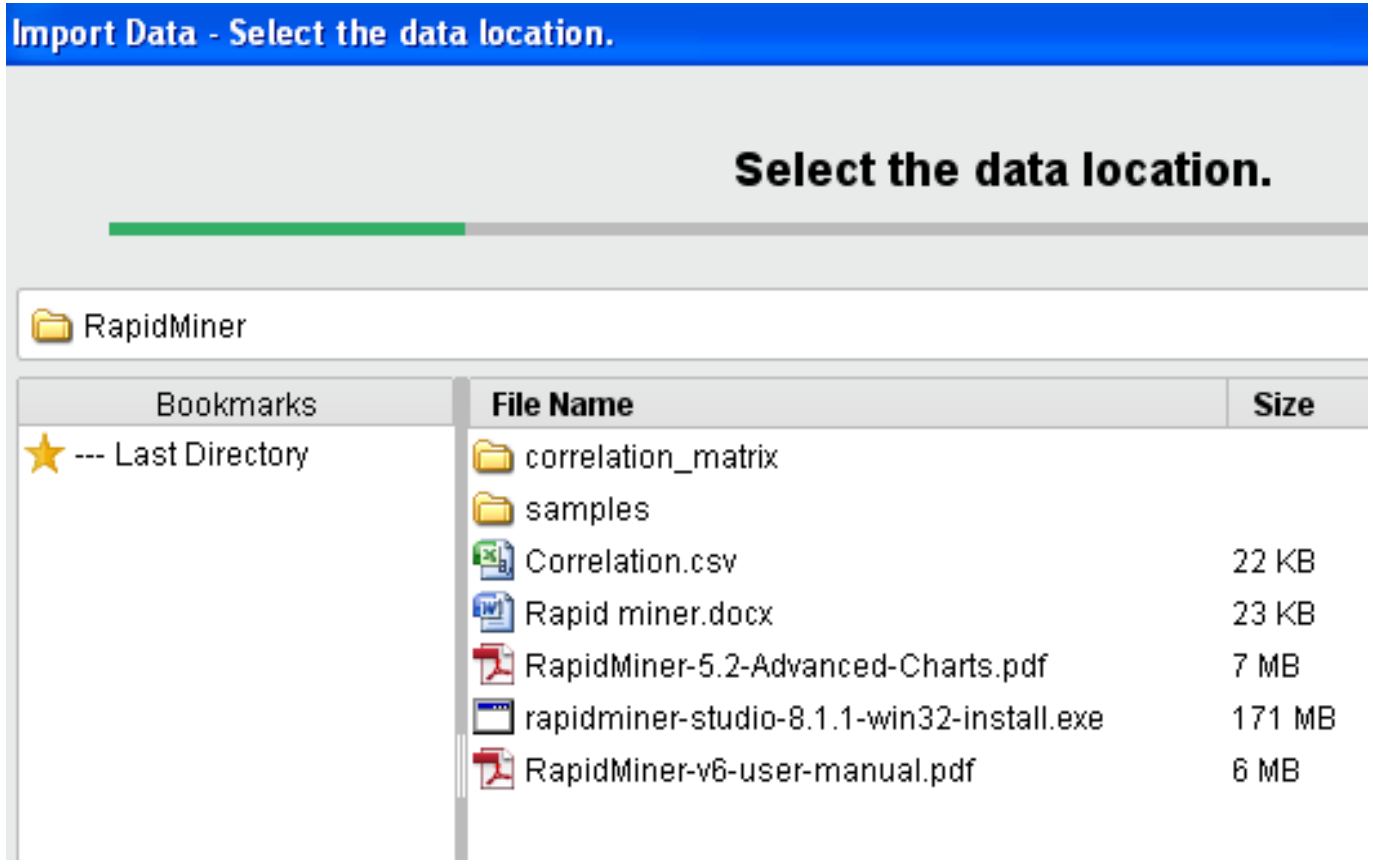
Στατιστική συσχέτιση - correlation

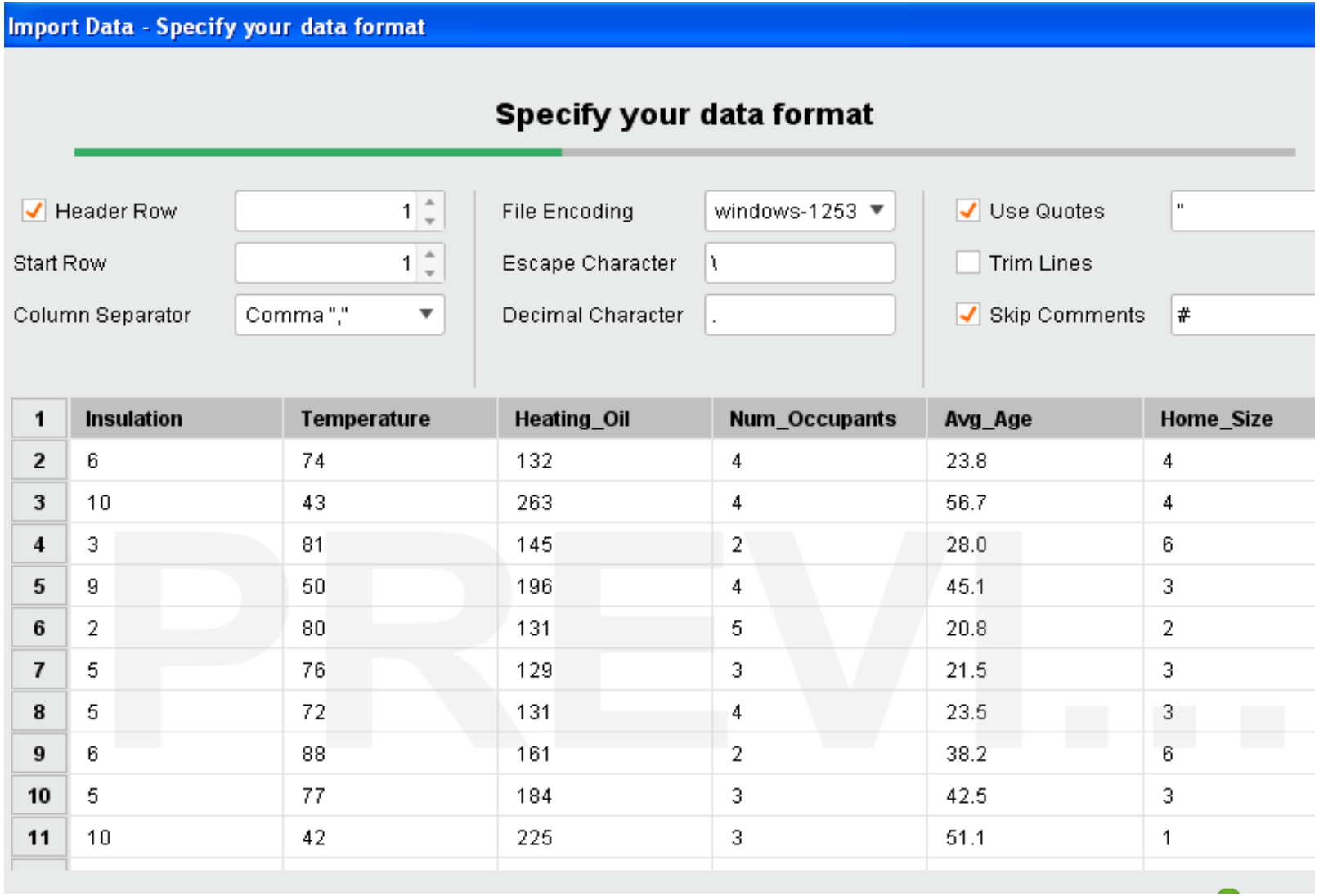
Επιλέξτε ένα νέο έργο στο εργαλείο RapidMiner.

Επιλέξτε την προσθήκη δεδομένων.

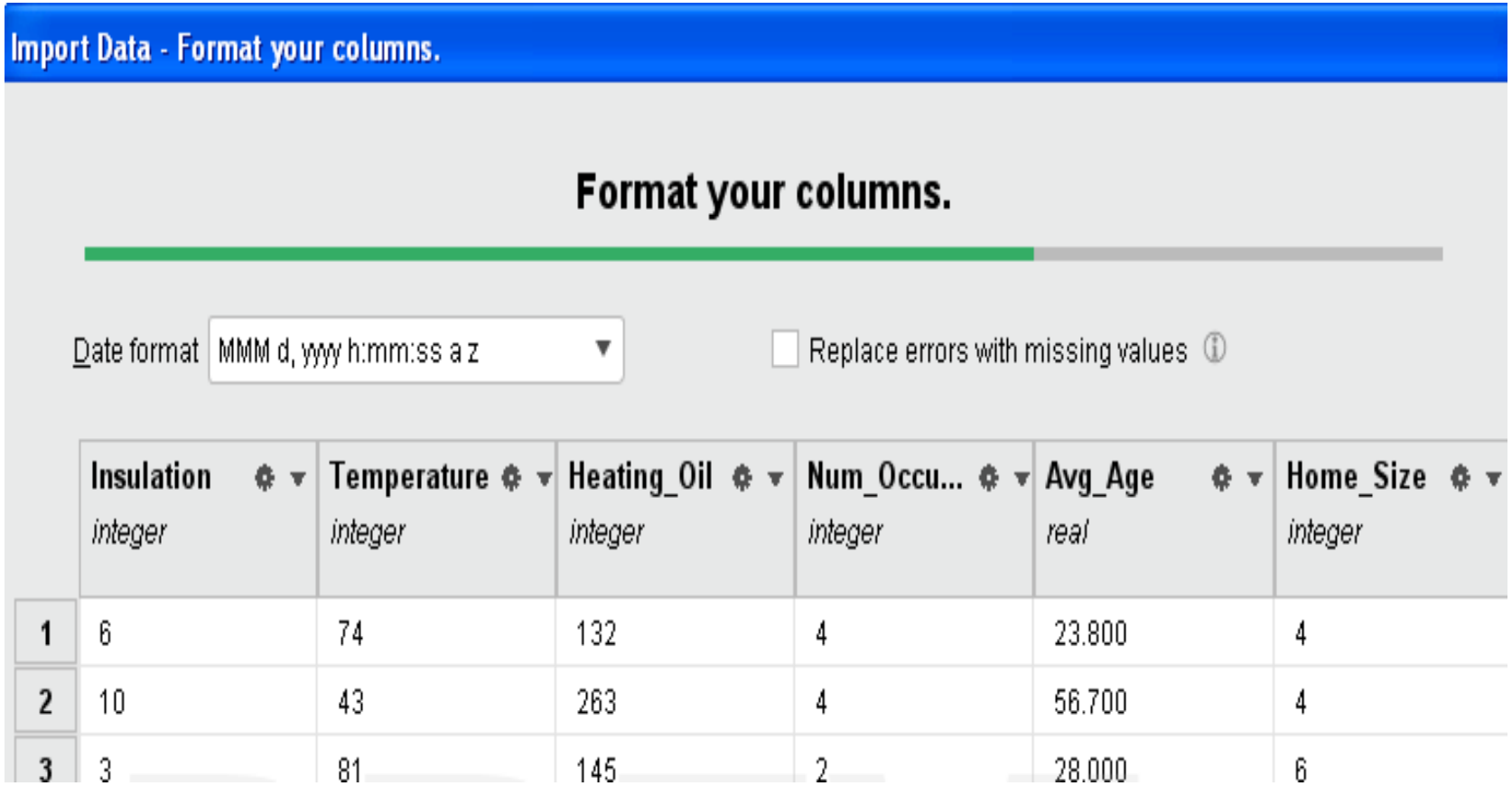



Import Data - Where is your data?
Where is your data?
My Computer
Database

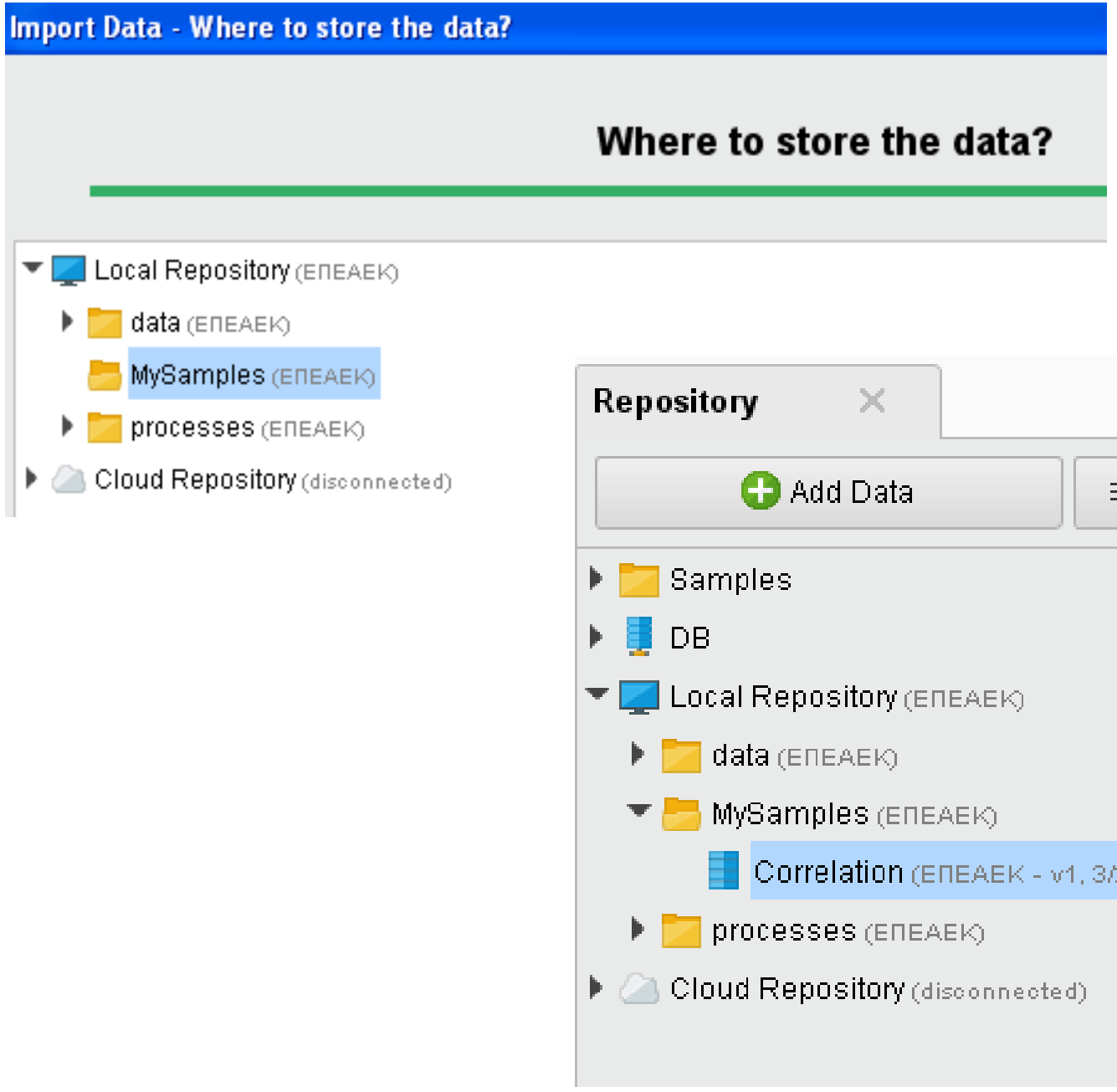

Import Data - Select the data location.
Select the data location.
RapidMiner
Bookmarks
--- Last Directory
File Name
Size
correlation_matrix
samples
Correlation.csv
22 KB
Rapid miner.docx
23 KB
RapidMiner-5.2-Advanced-Charts.pdf
7 MB
rapidminer-studio-8.1.1-win32-install.exe
171 MB
RapidMiner-v6-user-manual.pdf
6 MB



1	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
2	6	74	132	4	23.8	4
3	10	43	263	4	56.7	4
4	3	81	145	2	28.0	6
5	9	50	196	4	45.1	3
6	2	80	131	5	20.8	2
7	5	76	129	3	21.5	3
8	5	72	131	4	23.5	3
9	6	88	161	2	38.2	6
10	5	77	184	3	42.5	3
11	10	42	225	3	51.1	1



	Insulation *integer*	Temperature *integer*	Heating_Oil *integer*	Num_Occu... *integer*	Avg_Age *real*	Home_Size *integer*
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28.000	6


Import Data - Where to store the data?
Where to store the data?
Local Repository (ΕΠΕΑΕΚ)
data (ΕΠΕΑΕΚ)
MySamples (ΕΠΕΑΕΚ)
processes (ΕΠΕΑΕΚ)
Cloud Repository (disconnected)
Repository
Add Data
Samples
DB
Local Repository (ΕΠΕΑΕΚ)
data (ΕΠΕΑΕΚ)
MySamples (ΕΠΕΑΕΚ)
Correlation (ΕΠΕΑΕΚ - v1, 3/26/18
processes (ΕΠΕΑΕΚ)
Cloud Repository (disconnected)

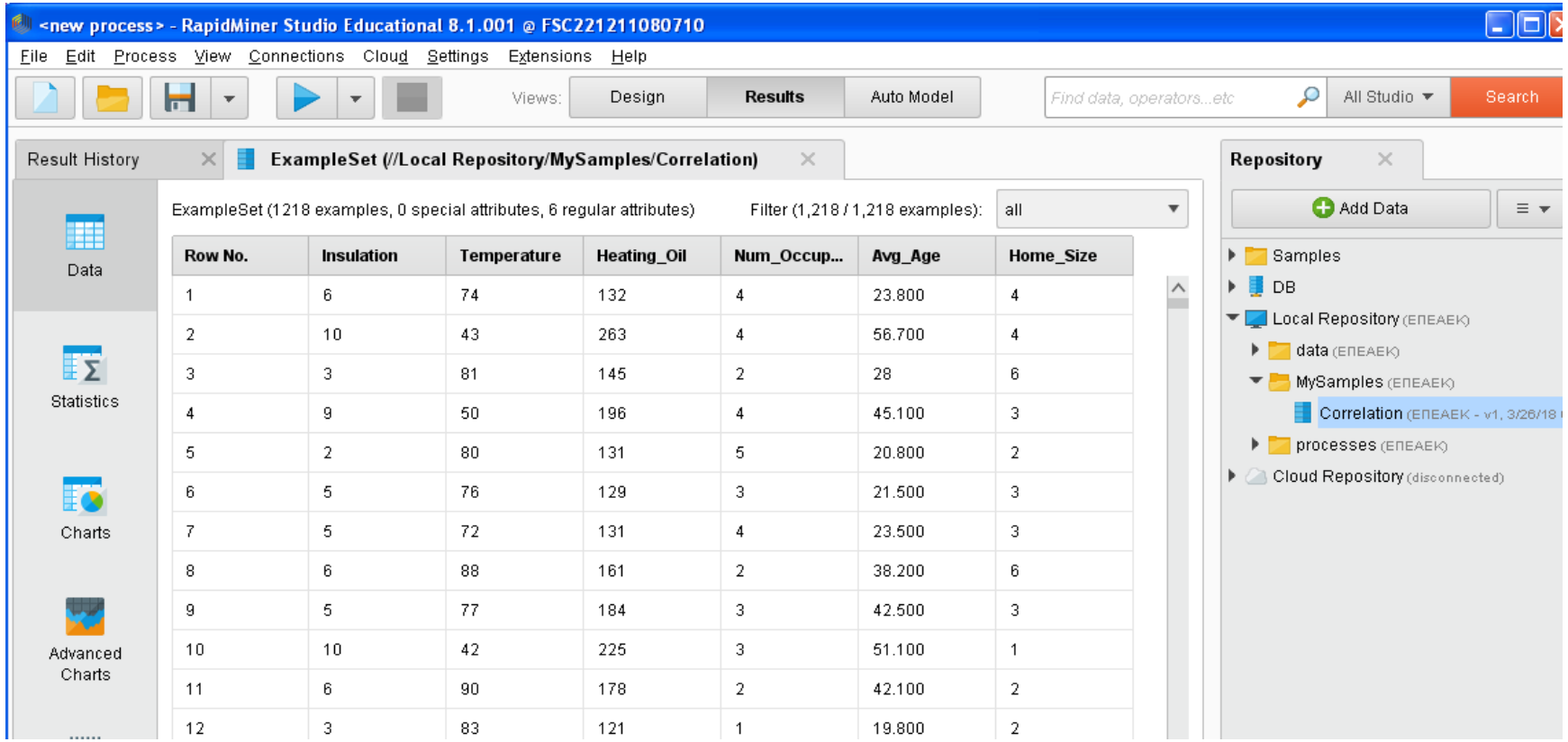
Να πως φαίνεται το Dataset

ExampleSet (//Local Repository/MySamples/Correlation)

ExampleSet (1218 examples, 0 special attributes, 6 regular attributes) Filter (1,218 / 1,218 examples): all

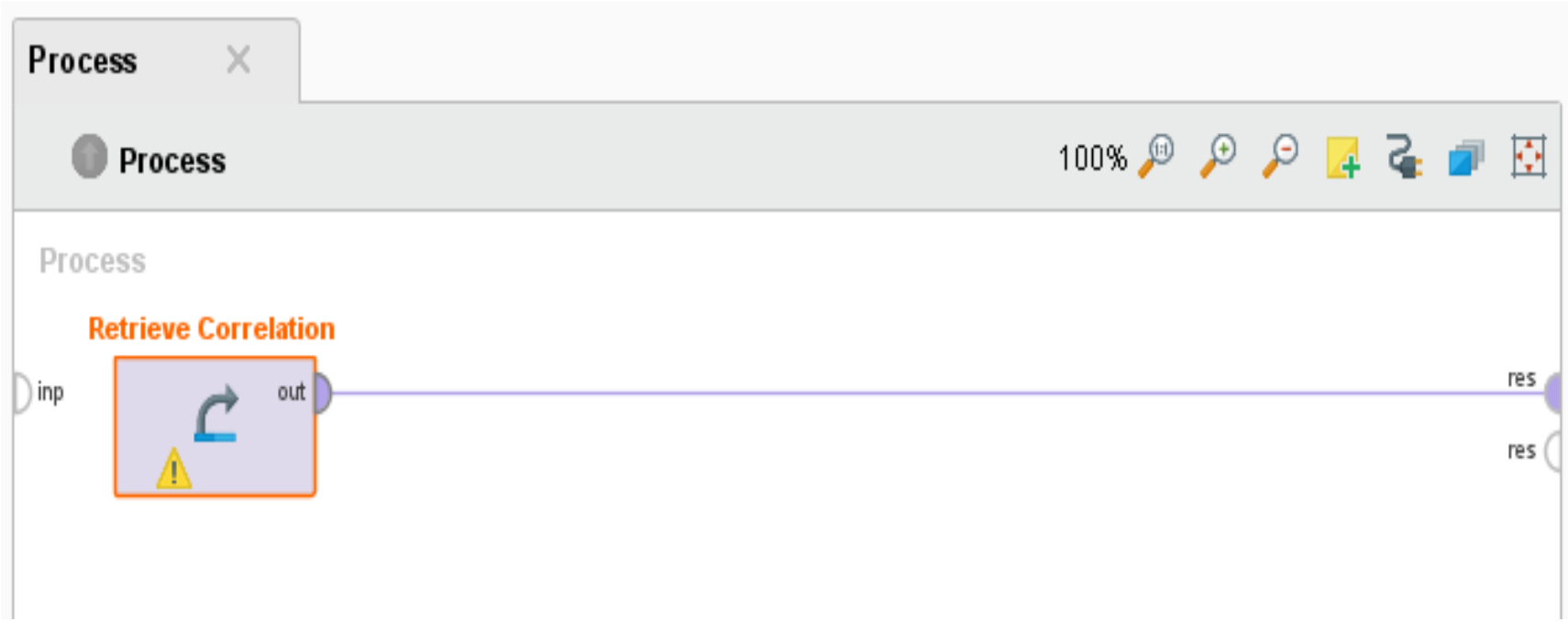
Row No.	Insulation	Temperature	Heating_Oil	Num_Occup...	Avg_Age	Home_Size
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28	6
4	9	50	196	4	45.100	3
5	2	80	131	5	20.800	2
6	5	76	129	3	21.500	3
7	5	72	121	4	22.500	2

Οι δύο views. Είμαστε σε view Results.

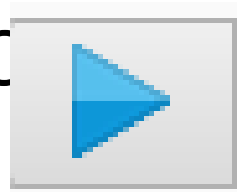


Επιλέγουμε DESIGN view και μεταφέρουμε με drag and drop το Data set στην περιοχή σχεδίασης του μοντέλου μας.

Συνδέουμε το process Retrieve Correlation (βλέπε port out στο process) μέσω spline με την έξοδο (res). Η γραμμή σύνδεσης (spline) ορίζεται με drag and drop ή με κλικ στο port out στο process και κλικ στην έξοδο (res)



Εκτελέστε με χρήση του εικονιδίο



Show the data in a table n in Turbo Prep Auto Model Filter (1,218 / 1,218 examples): all

Data

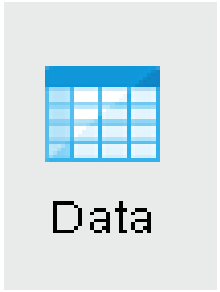
Statistics

Visualizations

Annotations

Row No.	Insulation_R...	Outdoor_Te...	Num_Occup...	Home_Age	Home_Size	Heating_Oil...
1	6	74	4	12	3356	132
2	10	43	4	45	2415	263
3	3	81	2	16	5899	145
4	9	50	4	34	3005	196
5	2	80	5	9	1164	131
6	5	76	3	10	2935	129
7	5	72	4	12	1678	131
8	6	88	2	27	4761	161
9	5	77	3	31	2149	184
10	10	42	3	40	574	225

Μπορούμε να επιλέξουμε Data View ή Statistics

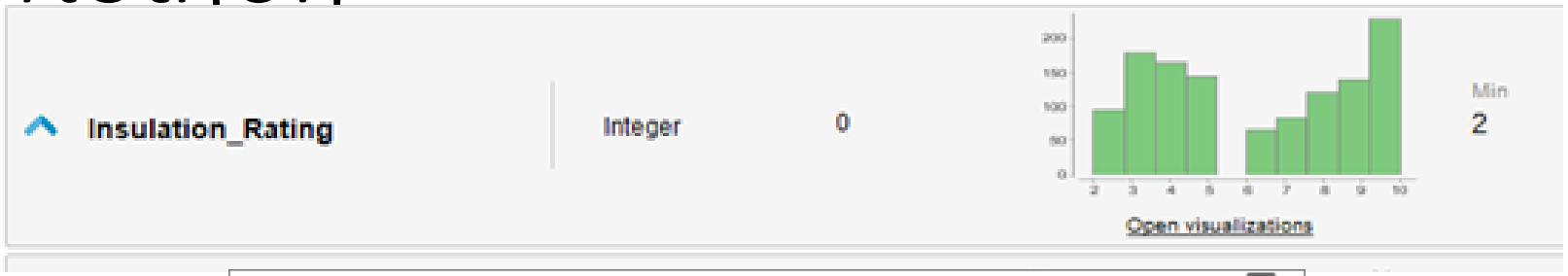


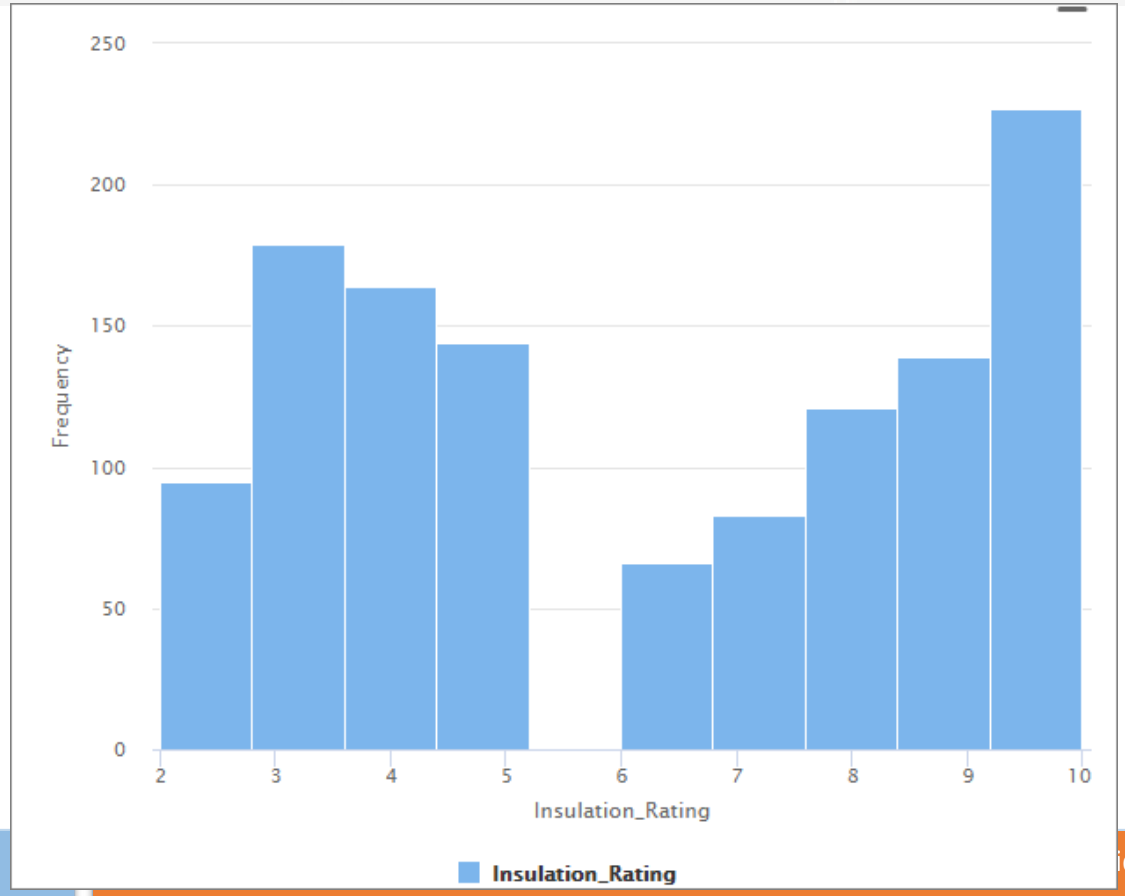
ExampleSet (1218 examples, 0 special attributes, 6 regular attributes) Filter (1,218 / 1,218 examples): all

Row No.	Insulation	Temperature	Heating_Oil	Num_Occup...	Avg_Age	Home_Size
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28	6

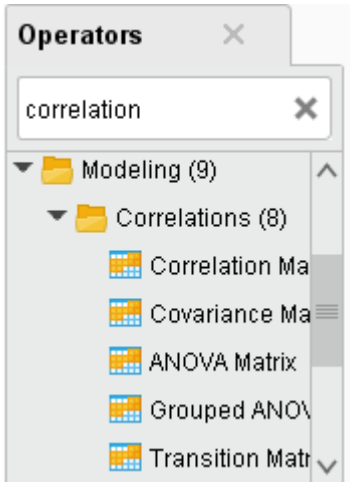
Name	Type	Missing	Filter (6 / 6 attributes): Search for Attributes	
Insulation	Integer	0	Min 2	Max 10
Temperature	Integer	0	Min 38	Max 90
Heating_Oil	Integer	0	Min 114	Max 301
Num_Occupants	Integer	0	Min 1	Max 10
Avg_Age	Real	0	Min 15.100	Max 72.200
Home_Size	Integer	0	Min 1	Max 8

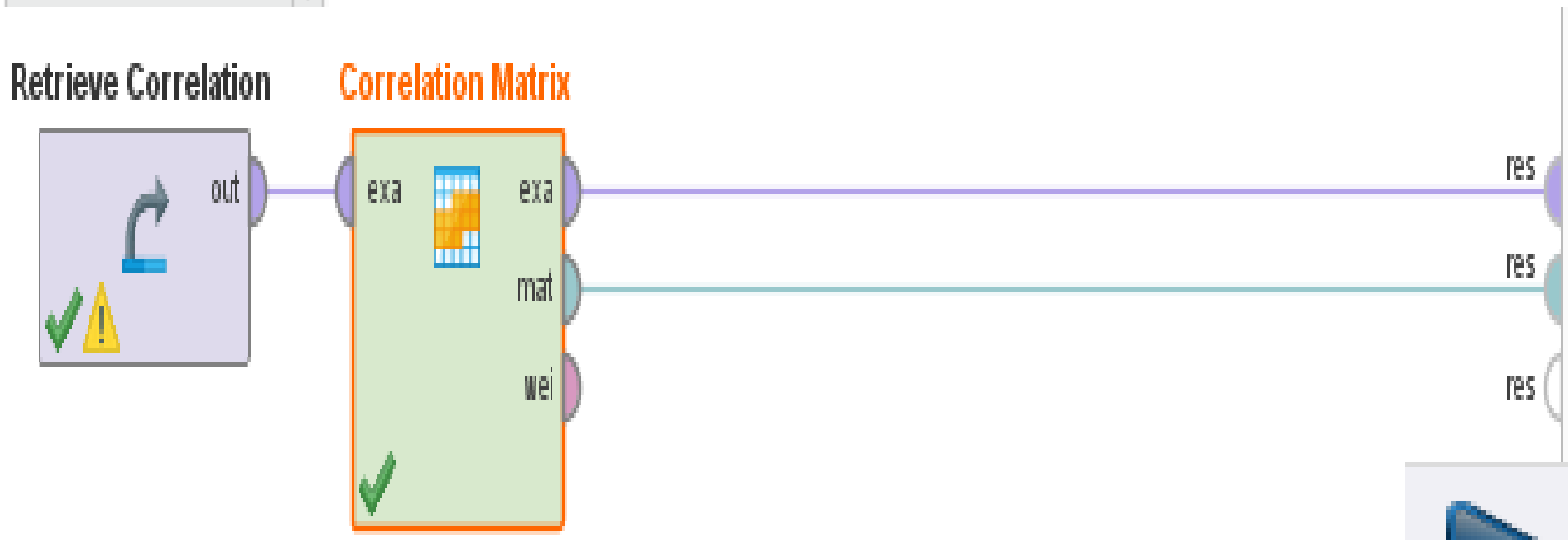
οπτικοποίηση





Αναζητούμε τελεστή (operator) Correlation και από τη λίστα των σχετικών τελεστών επιλέγουμε Correlation Matrix.





Τα αποτελέσματα σε Correlation Matrix

Correlation Matrix (Correlation Matrix) | ExampleSet (Correlation Matrix)

Attributes	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num_Occup...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1

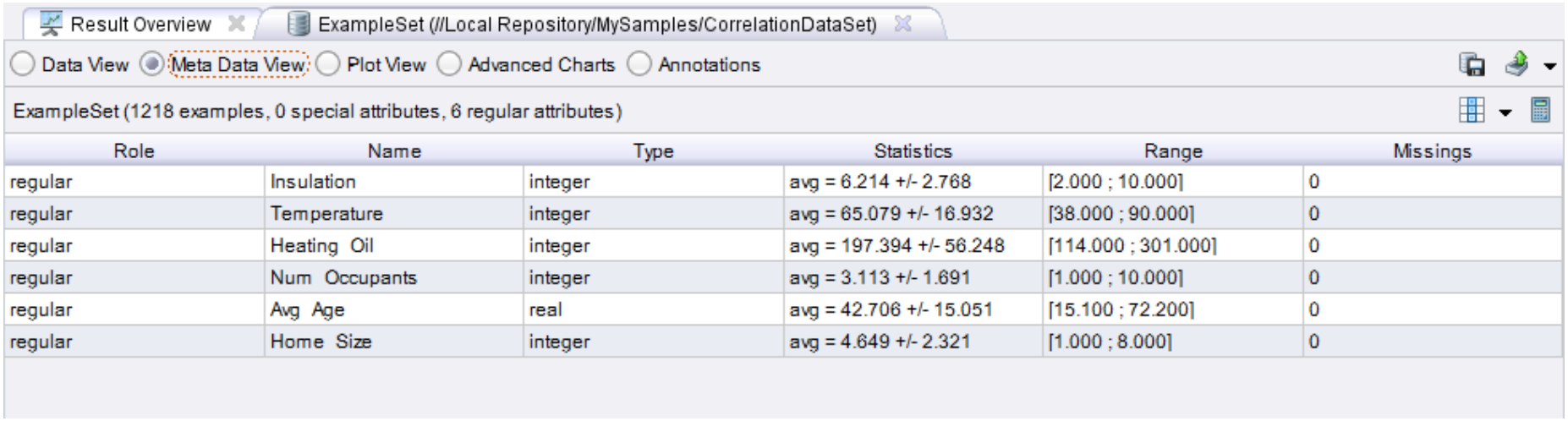
Result Overview | ExampleSet (//Local Repository/MySamples/CorrelationDataSet)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (1218 examples, 0 special attributes, 6 regular attributes)

Row No.	Insulation	Temperature	Heating Oil	Num Occu...	Avg Age	Home Size
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28	6
4	9	50	196	4	45.100	3
5	2	80	131	5	20.800	2
6	5	76	129	3	21.500	3
7	5	72	131	4	23.500	3
8	6	88	161	2	38.200	6
9	5	77	184	3	42.500	3
10	10	42	225	3	51.100	1
11	6	90	178	2	42.100	2
12	3	83	121	1	19.800	2
13	10	43	186	5	45.100	6
14	8	59	206	2	50.100	8
15	4	86	179	5	41.400	6
16	4	80	156	3	32.800	3
17	4	78	135	4	22.800	5
18	4	76	186	1	50.500	4
19	10	47	282	2	62	6
20	10	55	193	4	47.100	8

Log



Role	Name	Type	Statistics	Range	Missings
regular	Insulation	integer	avg = 6.214 +/- 2.768	[2.000 ; 10.000]	0
regular	Temperature	integer	avg = 65.079 +/- 16.932	[38.000 ; 90.000]	0
regular	Heating_Oil	integer	avg = 197.394 +/- 56.248	[114.000 ; 301.000]	0
regular	Num_Occupants	integer	avg = 3.113 +/- 1.691	[1.000 ; 10.000]	0
regular	Avg_Age	real	avg = 42.706 +/- 15.051	[15.100 ; 72.200]	0
regular	Home_Size	integer	avg = 4.649 +/- 2.321	[1.000 ; 8.000]	0

- Οι συντελεστές συσχέτισης (correlation coefficients) μεταξύ 0 και 1 αντιπροσωπεύουν θετικές συσχετίσεις (positive correlations) και οι συντελεστές μεταξύ 0 και -1 αρνητικές (negative correlations).

- positive correlations: Όσον η τιμή ενός χαρακτηριστικού (attribute) αυξάνεται, η τιμή του άλλου χαρακτηριστικού αυξάνεται επίσης

- negative correlations: Όσον η τιμή ενός χαρακτηριστικού μειώνεται, η τιμή του άλλου χαρακτηριστικού μειώνεται επίσης

Εξετάζουμε τη σχέση (relationship) μεταξύ των ιδιοτήτων (attributes) Heating_Oil και Insulation.

Όσο υψηλότερη είναι η εξωτερική θερμοκρασία τόσο μικρότερη είναι η ανάγκης για μόνωση

ΤΕΛΟΣ