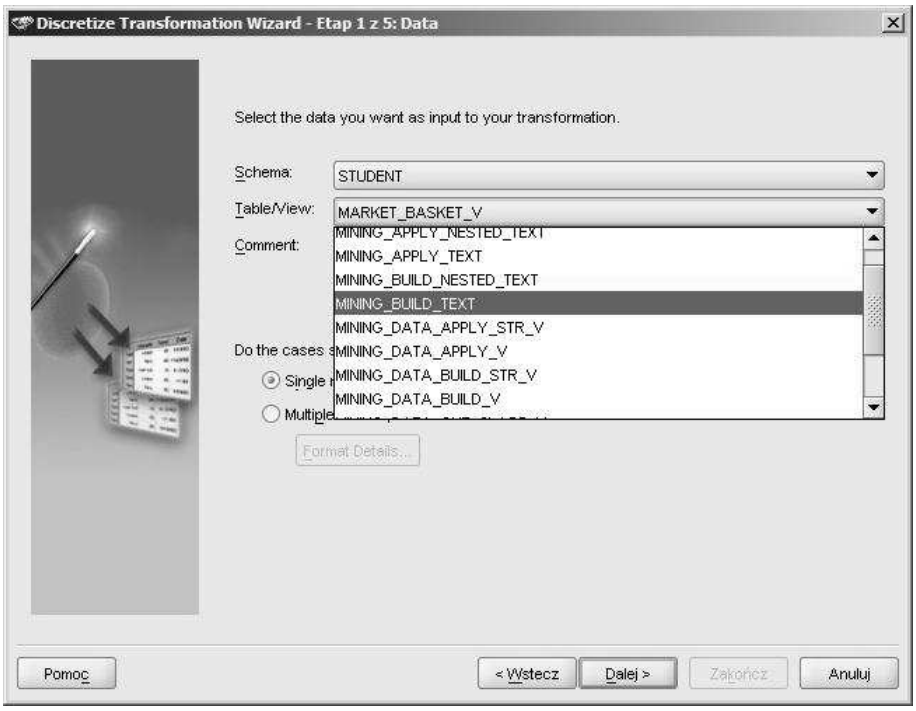
Laboratorium 1

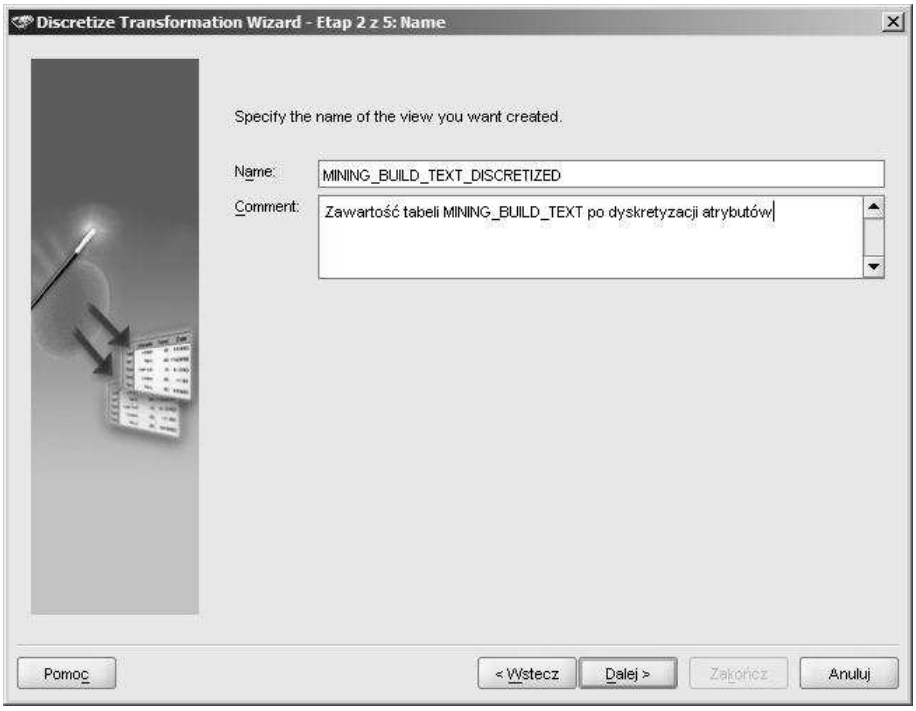
Przygotowanie danych do eksploracji.

1. Uruchom narzędzie Oracle Data Miner i połącz się z serwerem bazy danych.
2. Z menu głównego wybierz Data→Transform→Discretize.
3. Po wyświetleniu ekranu powitalnego wybierz przycisk Dalej>

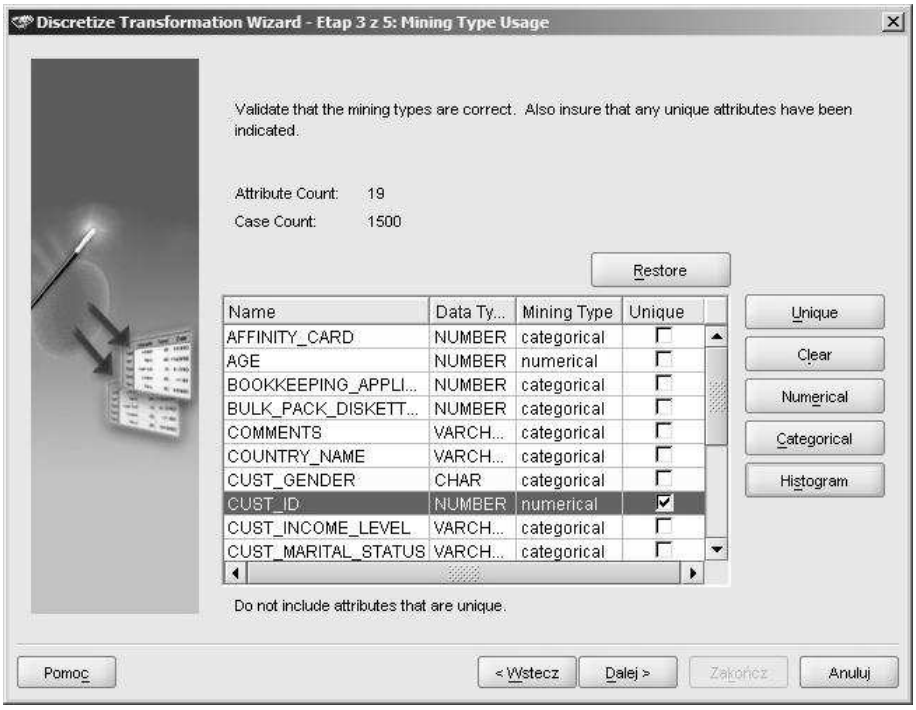
4. Wybierz schemat `STUDENT` i tabelę `MINING_BUILD_TEXT`, upewnij się, że zaznaczona jest opcja Single record per case. Kliknij przycisk Dalej>.



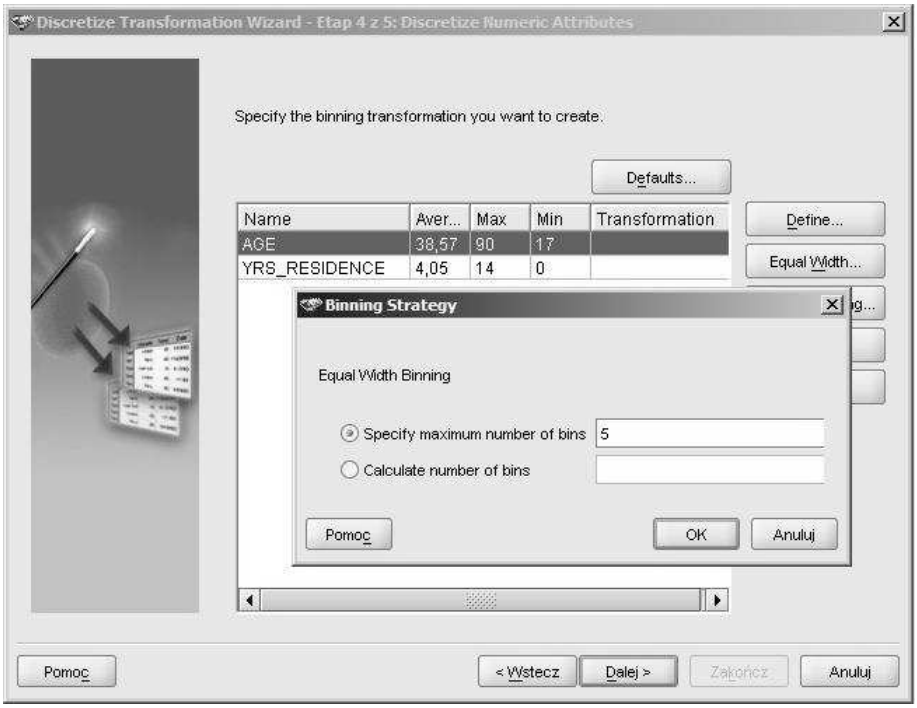
5. Podaj nazwę perspektywy wynikowej (`MINING_BUILD_TEXT_DISCRETIZED`) oraz opis słowny perspektywy. Kliknij przycisk Dalej>.



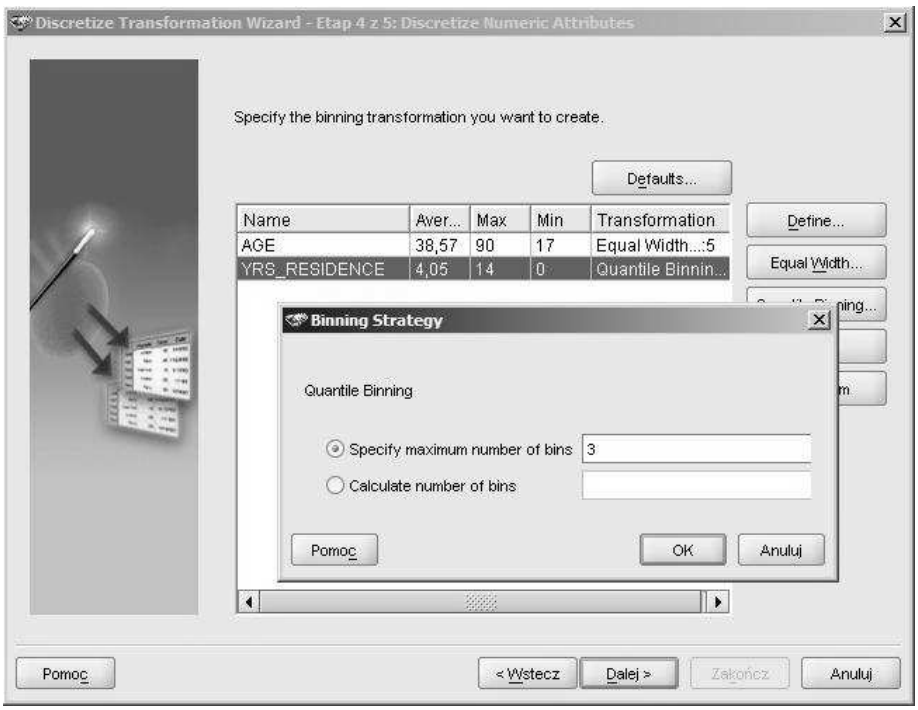
6. Upewnij się, że atrybut `CUST_ID` został zaznaczony jako unikalny (nie będzie brany pod uwagę w procesie eksploracji). Zwróć uwagę, że niektóre atrybuty typu `NUMBER` zostały zidentyfikowane jako atrybuty kategoryczne – dotyczy to atrybutów o niewielkiej liczbie różnych wartości. Kliknij przycisk Dalej>.



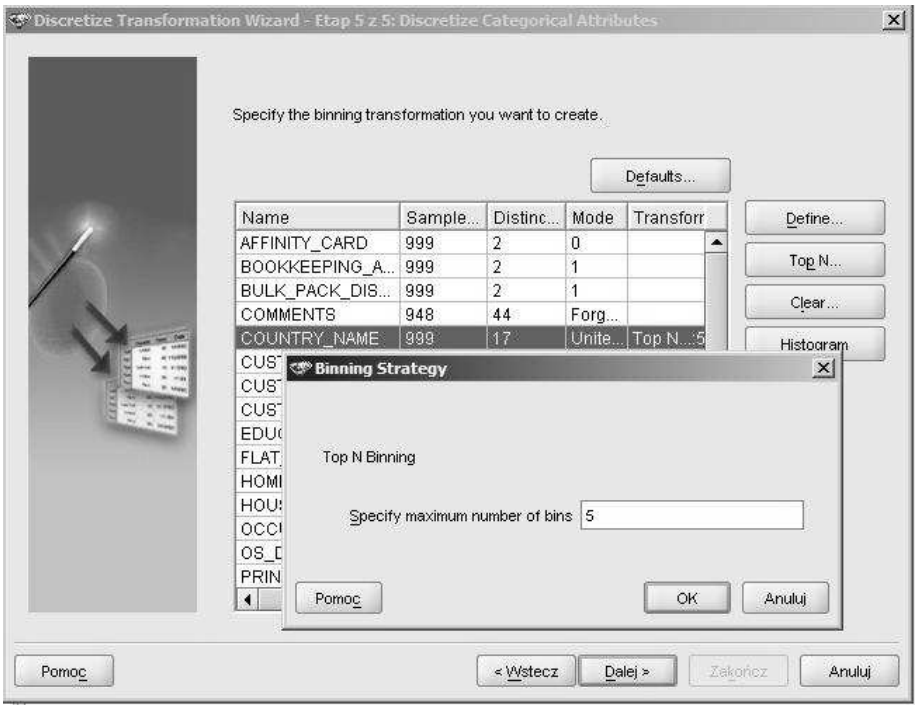
7. Zaznacz atrybut AGE. Zwróć uwagę na rozpiętość wartości: minimalnej, średniej i maksymalnej. Kliknij przycisk Equal Width. Wybierz opcję Specify maximum number of bins i wpisz wartość 5. Atrybut AGE zostanie podzielony na 5 przedziałów o równej szerokości. Kliknij przycisk OK.



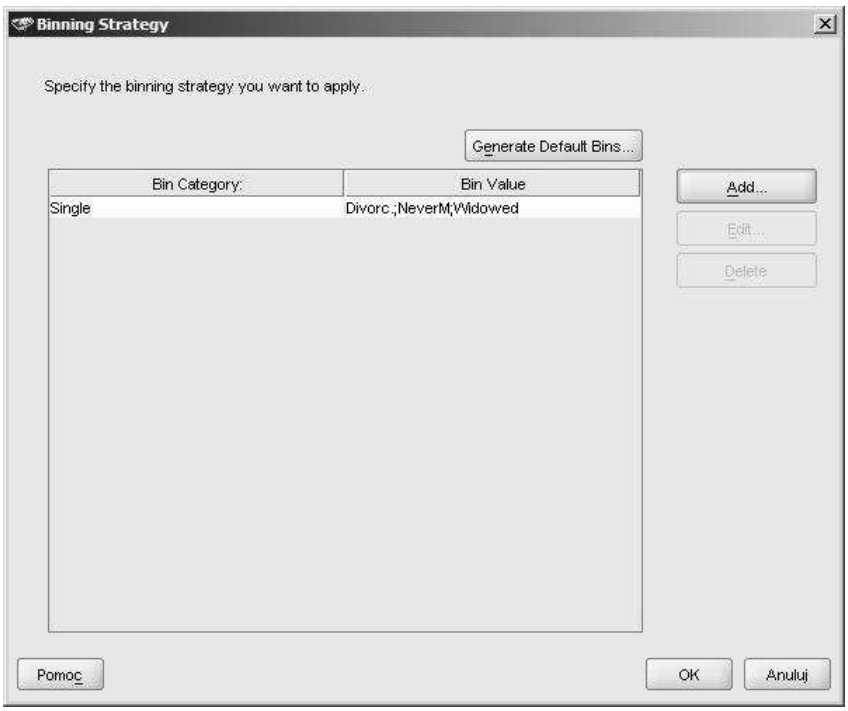
8. Następnie, zaznacz atrybut YRS_RESIDENCE i kliknij przycisk Quantile Binning. Wybierz opcję Specify maximum number of bins i wpisz wartość 5. Atrybut YRS_RESIDENCE zostanie podzielony na 5 równolicznych grup. Aby zakończyć, kliknij przycisk OK. Kliknij przycisk Dalej>.



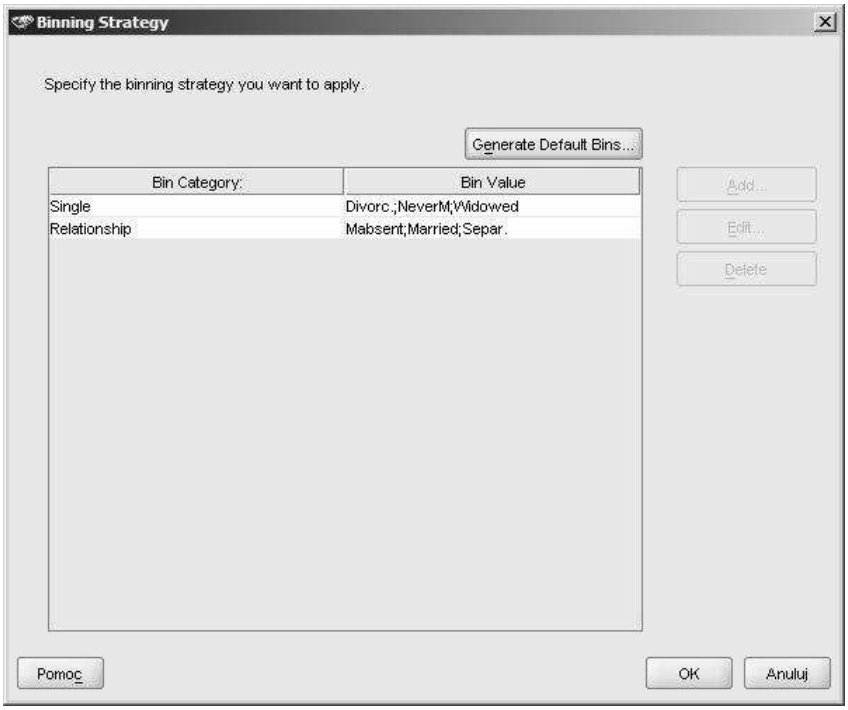
9. Obejrzyj listę atrybutów kategorycznych, zwróć uwagę na atrybuty o dużej liczbie wartości. Zaznacz atrybut COUNTRY_NAME. Kliknij przycisk Top N. W pole Specify maximum number of bins wpisz wartość 5. Kliknij przycisk OK. W perspektywie wynikowej pozostanie 5 najczęściej pojawiających się nazw krajów, a wszystkie pozostałe kraje zostaną umieszczone w zbiorczej kategorii Others.



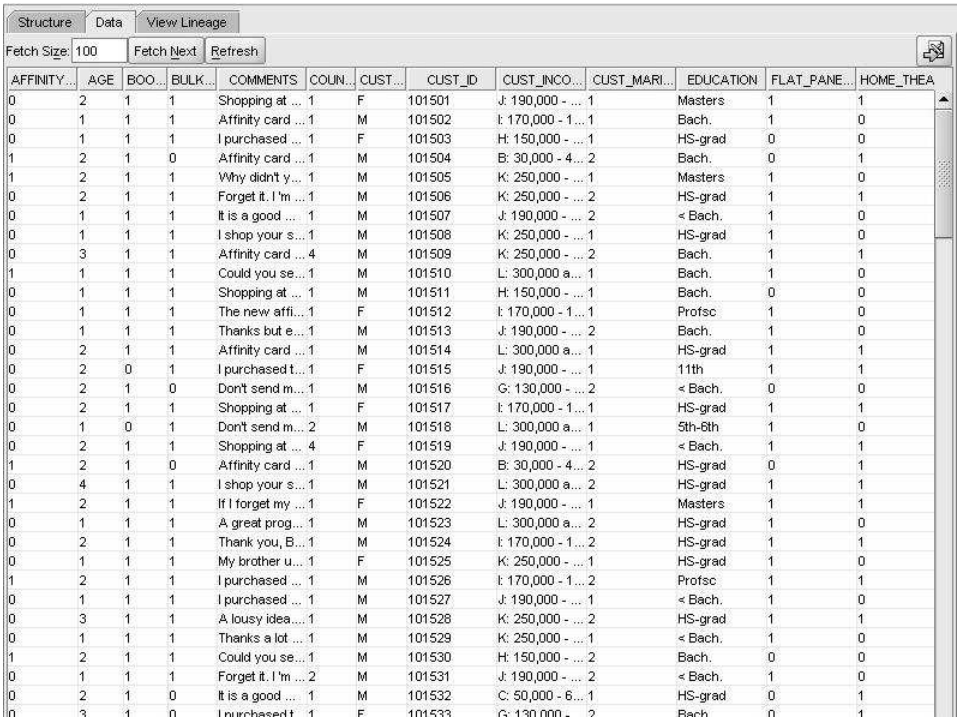
10. Zaznacz atrybut CUST_MARITAL_STATUS i kliknij przycisk Define. Kliknij przycisk Add. W pole Bin Category wpisz **Single**. W polu All Distinct Values zaznacz wartość Divorced i kliknij przycisk Następnie zaznacz wartość NeverM i ponownie kliknij przycisk. Na koniec zaznacz wartość Widowed i kliknij przycisk. Kliknij przycisk OK. W tym momencie ekran komputera powinien wyglądać następująco.



11. W analogiczny sposób przygotuj kategorię **Relationship** i włącz do niej wartości Mabsetn, Married, Separ. Po zakończeniu ekran komputera powinien wyglądać następująco.

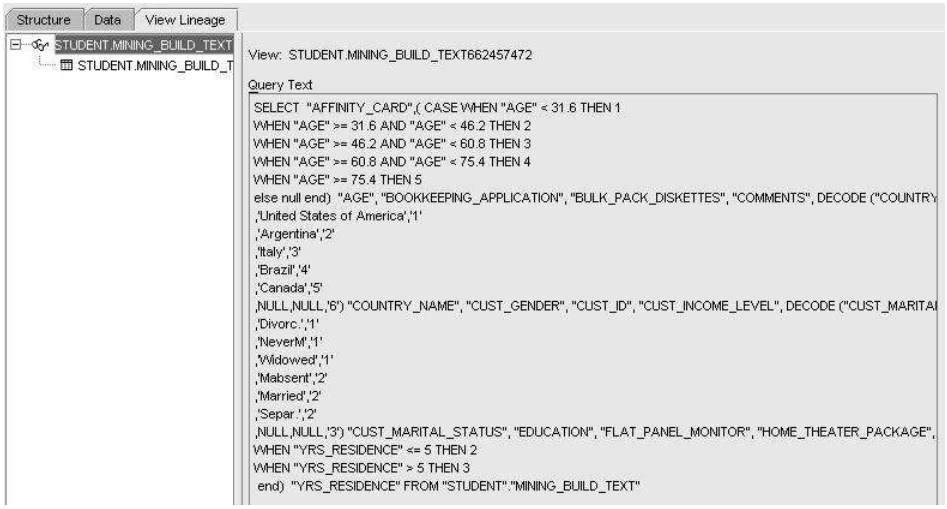


12. Kliknij przycisk Dalej>. Kliknij przycisk Zakończ. Rozwiń drzewo obiektów po lewej stronie ekranu i przejdź do student@miner→Data Sources→STUDENT→Views. Zaznacz perspektywę `MINING_BUILD_TEXT_DISCRETIZED`. W głównym oknie przejdź do prawego panelu i kliknij na zakładkę Data. Zwróć uwagę na wartości w atrybutach AGE, `COUNTRY_NAME`, `CUST_MARITAL_STATUS` i `YRS_RESIDENCE`.



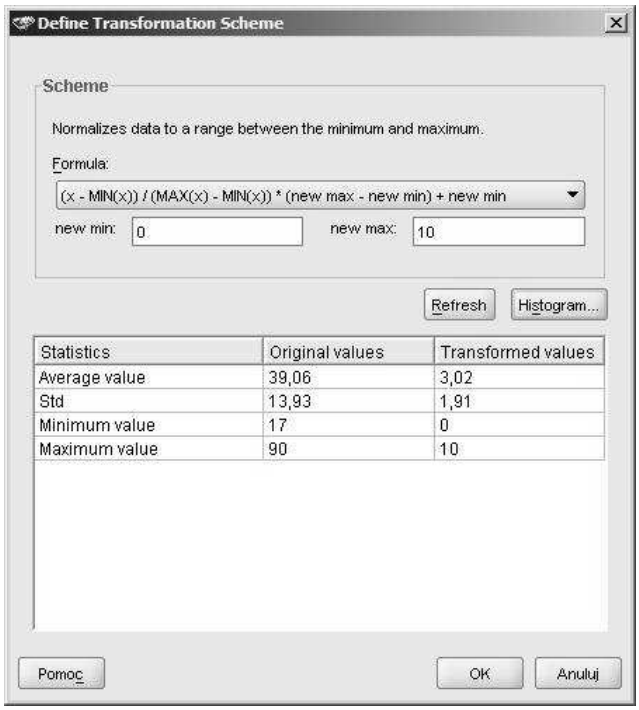
AFFINITY...	AGE	BOO...	BULK...	COMMENTS	COUN...	CUST...	CUST_ID	CUST_INCO...	CUST_MARI...	EDUCATION	FLAT_PANE...	HOME_THEA
0	2	1	1	Shopping at ...	1	F	101501	J: 190,000 - ...	1	Masters	1	1
0	1	1	1	Affinity card ...	1	M	101502	I: 170,000 - 1...	1	Bach.	1	0
0	1	1	1	I purchased ...	1	F	101503	H: 150,000 - ...	1	HS-grad	0	0
1	2	1	0	Affinity card ...	1	M	101504	B: 30,000 - 4...	2	Bach.	0	1
1	2	1	1	Why didn't y...	1	M	101505	K: 250,000 - ...	1	Masters	1	0
0	2	1	1	Forget it. I 'm ...	1	M	101506	K: 250,000 - ...	2	HS-grad	1	1
0	1	1	1	It is a good ...	1	M	101507	J: 190,000 - ...	2	< Bach.	1	0
0	1	1	1	I shop your s...	1	M	101508	K: 250,000 - ...	1	HS-grad	1	0
0	3	1	1	Affinity card ...	4	M	101509	K: 250,000 - ...	2	Bach.	1	1
1	1	1	1	Could you se...	1	M	101510	L: 300,000 a...	1	Bach.	1	0
0	1	1	1	Shopping at ...	1	M	101511	H: 150,000 - ...	1	Bach.	0	0
0	1	1	1	The new affi...	1	F	101512	I: 170,000 - 1...	1	Profsc	1	0
0	1	1	1	Thanks but e...	1	M	101513	J: 190,000 - ...	2	Bach.	1	0
0	2	1	1	Affinity card ...	1	M	101514	L: 300,000 a...	1	HS-grad	1	1
0	2	0	1	I purchased t...	1	F	101515	J: 190,000 - ...	1	11th	1	1
0	2	1	0	Don't send m...	1	M	101516	G: 130,000 - ...	2	< Bach.	0	0
0	2	1	1	Shopping at ...	1	F	101517	I: 170,000 - 1...	1	HS-grad	1	1
0	1	0	1	Don't send m...	2	M	101518	L: 300,000 a...	1	5th-6th	1	0
0	2	1	1	Shopping at ...	4	F	101519	J: 190,000 - ...	1	< Bach.	1	1
1	2	1	0	Affinity card ...	1	M	101520	B: 30,000 - 4...	2	HS-grad	0	1
0	4	1	1	I shop your s...	1	M	101521	L: 300,000 a...	2	HS-grad	1	1
1	2	1	1	If I forget my ...	1	F	101522	J: 190,000 - ...	1	Masters	1	1
0	1	1	1	A great prog...	1	M	101523	L: 300,000 a...	2	HS-grad	1	0
0	2	1	1	Thank you, B...	1	M	101524	I: 170,000 - 1...	2	HS-grad	1	1
0	1	1	1	My brother u...	1	F	101525	K: 250,000 - ...	1	HS-grad	1	0
1	2	1	1	I purchased ...	1	M	101526	I: 170,000 - 1...	2	Profsc	1	1
0	1	1	1	I purchased ...	1	M	101527	J: 190,000 - ...	1	< Bach.	1	0
0	3	1	1	A lousy idea....	1	M	101528	K: 250,000 - ...	2	HS-grad	1	1
0	1	1	1	Thanks a lot ...	1	M	101529	K: 250,000 - ...	1	< Bach.	1	0
1	2	1	1	Could you se...	1	M	101530	H: 150,000 - ...	2	Bach.	0	0
0	1	1	1	Forget it. I 'm ...	2	M	101531	J: 190,000 - ...	2	< Bach.	1	0
0	2	1	0	It is a good ...	1	M	101532	C: 50,000 - 6...	1	HS-grad	0	1
0	3	1	0	I purchased t...	1	F	101533	G: 130,000 - ...	2	Bach.	0	1

13. Kliknij na zakładkę View Lineage. Przeanalizuj kod perspektywy wynikowej realizującej poszczególne kroki dyskretyzacji.

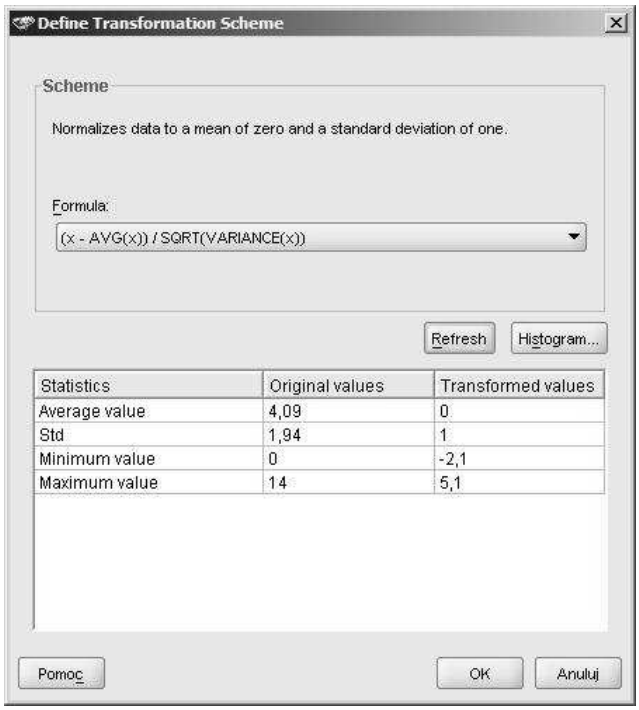


14. Z menu głównego wybierz Data→Transform→Normalize. Na ekranie powitalnym kliknij przycisk Dalej>. Wybierz schemat STUDENT. Wybierz ponownie tabelę MINING_BUILD_TEXT. Kliknij przycisk Dalej>. Podaj nazwę perspektywy wynikowej (MINING_BUILD_TEXT_NORMALIZED) i krótki opis zawartości perspektywy (np. zawartość tabeli MINING_BUILD_TEXT po normalizacji). Kliknij przycisk Dalej>.

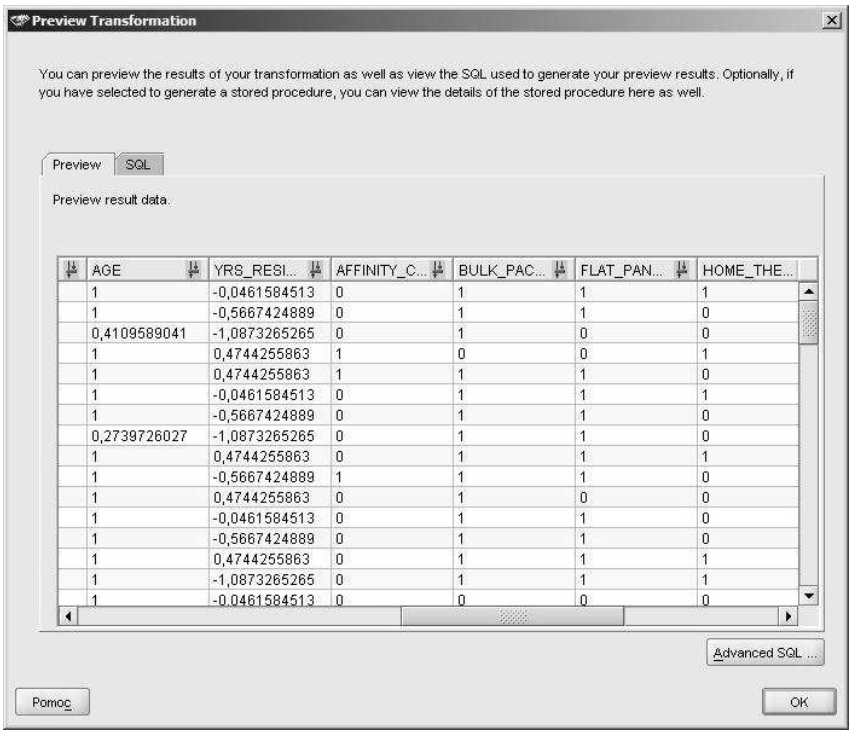
15. Zaznacz atrybut AGE. Przeanalizuj statystyki związane z atrybutem. Kliknij przycisk Define. Z listy dostępnych transformacji wybierz transformację MinMax (pierwsza pozycja na liście). Jako nową wartość minimalną wpisz **0** a jako nową wartość maksymalną wpisz **10**. Kliknij przycisk Refresh. Kliknij przycisk OK.



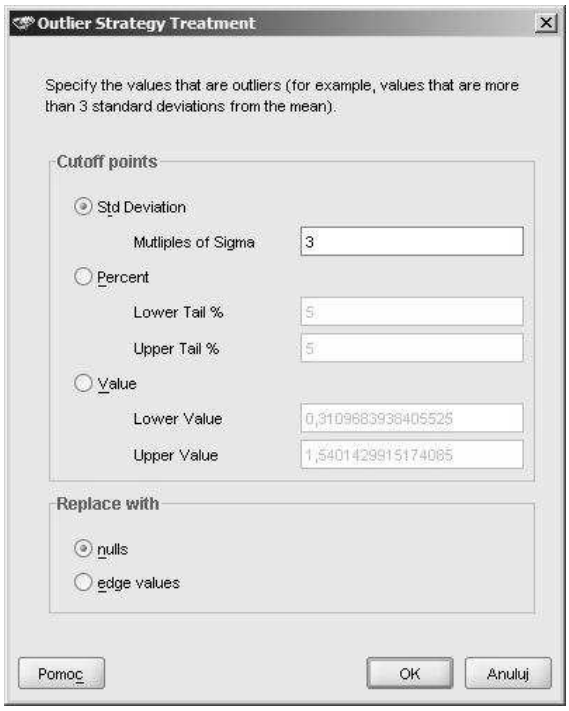
16. Zaznacz atrybut YRS_RESIDENCE i kliknij przycisk Define. Z listy dostępnych transformacji wybierz normalizację przez odchylenia standardowe (druga pozycja na liście). Kliknij przycisk Refresh. Kliknij przycisk OK. Kliknij przycisk Dalej>.



17. Kliknij przycisk Preview Transform. Zwróć uwagę na wartości atrybutów AGE i YRS_RESIDENCE. Kliknij na zakładce SQL i obejrzyj kod perspektywy dokonującej normalizacji atrybutów numerycznych. Kliknij przycisk OK. Kliknij przycisk Zakończ.



18. Z menu głównego wybierz Data→Transform→Outlier Treatment. Na ekranie powitalnym kliknij przycisk Dalej>. Wybierz schemat STUDENT. Wybierz tabelę MINING_BUILD_TEXT_NORMALIZED. Kliknij przycisk Dalej>. Podaj nazwę perspektywy wynikowej (MINING_BUILD_TEXT_NOOUTLIERS) i opis zawartości perspektywy (zawartość perspektywy MINING_BUILD_TEXT_NORMALIZED po usunięciu osobliwości). Kliknij przycisk Dalej>.
19. Upewnij się, że atrybut CUST_ID jest zaznaczony jako unikalny. Sprawdź, czy poszczególne atrybuty zostały poprawnie zaklasyfikowane jako kategoryczne lub numeryczne. Kliknij przycisk Dalej>.
20. Zaznacz atrybut AGE. Kliknij przycisk Define. Wybierz wielokrotność odchylenia standardowego jako preferowaną metodę identyfikacji osobliwości, jako wartość graniczną wpisz **3** (dane odległe o więcej niż 3 wartości odchylenia standardowego od średniej zostaną uznane za osobliwości). Upewnij się, że u dołu okna zaznaczona jest wartość Replace with nulls. Kliknij przycisk OK.



21. Zaznacz atrybut YRS_RESIDENCE. Kliknij przycisk Define. Wybierz procent wartości granicznych jako preferowaną metodę identyfikacji osobliwości, jako wartość dolnego i górnego odcięcia wpisz **5%** (po 5% najniższych i najwyższych wartości zostanie uznanych za osobliwości). Upewnij się, że u dołu okna zaznaczona jest wartość Replace with edge values. Kliknij przycisk OK. Kliknij przycisk Dalej>.
22. Kliknij przycisk Preview Transform i znajdź wiersze, w których znaleziono osobliwości w atrybucie AGE. Czy możesz zidentyfikować wiersze, w których osobliwości wystąpiły w atrybucie YRS_RESIDENCE? Kliknij na zakładce SQL i obejrzyj kod perspektywy dokonującej identyfikacji osobliwości. Czy potrafisz dostrzec poważną wadę wykorzystywanego narzędzia?
23. Połącz się z bazą danych wykorzystując iSQLPlus. Wykonaj skrypt preparation.sql. Po każdym kroku przeanalizuj uzyskane wyniki (komentarz jest umieszczony wewnątrz skryptu).