

WYDZIAŁ  
MATEMATYKI  
I FIZYKI STOSOWANEJ  
POLITECHNIKI RZESZOWSKIEJ

**Wielowymiarowa analiza danych**  
Projekt

Konrad Olszewski 169826

Krystian Pupiec 169833

Rzeszów, 29 maja 2024

# Spis treści

<b>1. Cel projektu</b>	<b>3</b>
<b>2. Opis danych</b>	<b>3</b>
<b>3. ETL</b>	<b>4</b>
3.1. ETL 1	4
3.2. ETL 2	5
3.3. ETL 3	6
3.4. ETL 4	7
3.5. Execute SQL Task, finalny przepływ danych	8
<b>4. Kostka OLAP</b>	<b>9</b>
4.1. Hierarchie	12
<b>5. Scenariusze - KPI</b>	<b>13</b>
5.1. Scenariusz 1	13
5.2. Scenariusz 2	14
5.3. Scenariusz 3	15
5.4. Scenariusz 4	16
5.5. Scenariusz 5	17
<b>6. Data mining</b>	<b>18</b>
6.1. Drzewo decyzyjne	19
6.2. Sieć neuronowa	22
6.3. Klastrowanie	24
<b>7. Wizualizacja danych - Power BI</b>	<b>30</b>
7.1. Wizualizacje	31
<b>8. Podsumowanie</b>	<b>33</b>

## 1. Cel projektu

Celem niniejszego projektu jest przeprowadzenie wielowymiarowej analizy danych na zestawie danych składającym się z 700,000 wierszy. Projekt obejmuje cały proces od przygotowania danych, przez ich przetwarzanie, aż po analizę i wizualizację wyników.

## 2. Opis danych

Zebrane dane przedstawiają informacje kredytowe klientów pewnego banku. Dane te zostały sztucznie wygenerowane przy pomocy skryptu utworzonego w języku programowania - python. W skład tych danych wchodzi takie argumenty jak: *id\_klienta*, *plec*, *wiek*, *miasto*, *typ\_kredytu*, *kwota\_kredytu*, *oprocentowanie*, *okres\_kredytu*, *data\_zawarcia*, *status*, *wykształcenie*, *zawod*.

	A	B	C	D	E	F	G	H	I	J	K	L
1	id_klienta,plec,wiek,miasto,typ_kredytu,kwota_kredytu,oprocentowanie,okres_kredytu,data_zawarcia,status,wykształcenie,zawo											
2	1,K,47,Bydgo	_szcz	konsumpcyjny,	471358,2.98,59,15-03-2019,zalegly,wyzsze,prawnik								
3	b,M,57,Wroclaw,	konsolidacyjny,	256603,6.92,31,03-06-2021,aktywny,wyzsze,dzialalnosc_wlasna									
4	3,K,82,Gdy nia,	inwestycyjny,	525801,12.68,122,04-11-2021,splacony,wyzsze,tynkarz									
5	4,K,87,Czestochowa,	konsumpcyjny,	387963,1.76,108,09-07-2021,splacony,wyzsze,zolnierz									
6	g,M,31,Warszawa,	hipoteczny,	300296,1.94,12,05-11-2011,zalegly,wyzsze,sprzedawca									

Rysunek 2.1: Widok utworzonych danych

Dodatkowo na potrzeby późniejszych procesów ETL odłączona została ostatnia kolumna *zawod* od głównego pliku danych a także do bazy danych została wprowadzona tabela z poprawnymi nazwami miast w języku polskim.

Przedstawione dane zawierają informacje z okresu 01-01-2000r do 31-12-2022r. Dane zostały utworzone zachowując przy tym rozkład normalny. Oprócz tego zostały one podzielone na dwa oddzielne pliki tekstowe *.csv*.

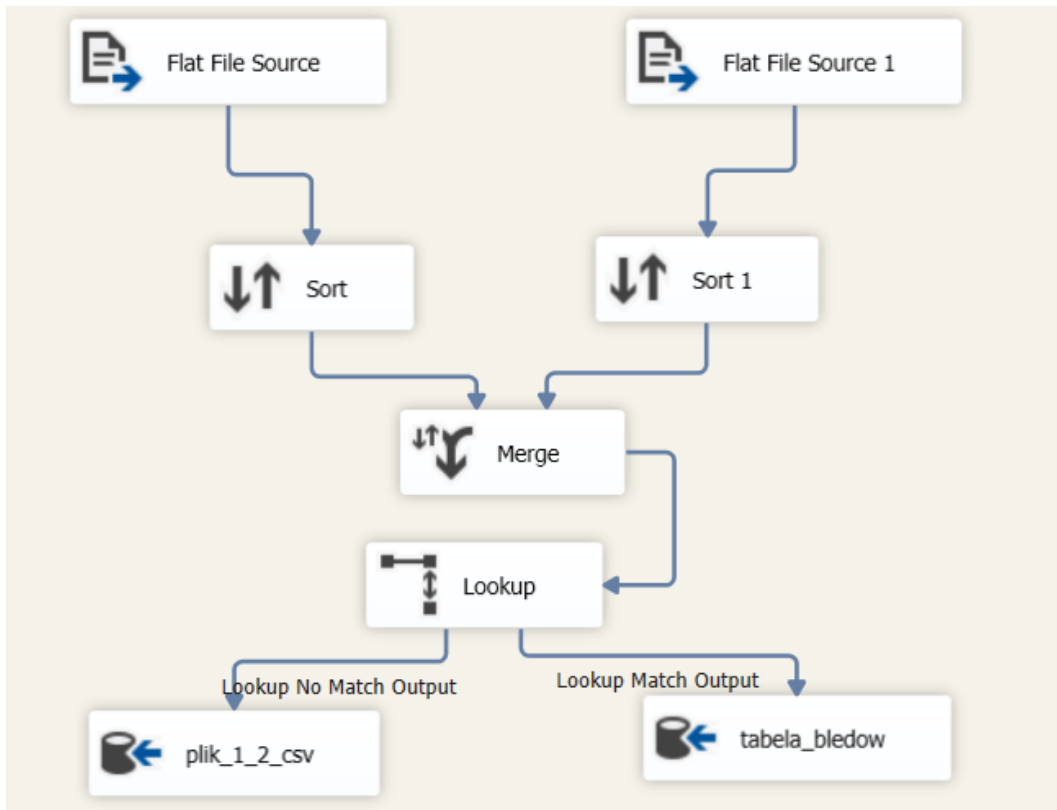
### 3. ETL

W tym etapie przedstawione zostaną utworzone na potrzeby projektu 4 przepływy ETL mające na celu wczytanie danych z odpowiednich źródeł, oczyszczenie danych z błędów, przyłączenie im odpowiednich nowych kolumn, poprawienie pisowni oraz danych zawierających niechcane, błędne wartości.

#### 3.1. ETL 1

Poniższy przepływ ETL wykorzystuje takie procesy i obiekty jak *Flat File Source*, *Sort*, *Merge*, *Lookup*, *OLE DB Destination*. Odpowiadają one odpowiednio za:

- **Flat File Source** - jest to obiekt pełniący funkcję pobierania danych z źródła w tym wypadku pliku tekstowego w formacie *.csv*,
- **Sort** - odpowiednie posortowanie danych na podstawie każdej kolumny w sposób rosnący,
- **Merge** - proces ten odpowiada za złączenie posortowanych danych w 1 plik,
- **Lookup** - operacja *Lookup* w tym przypadku odpowiada za sprawdzenie czy wprowadzane dane istnieją już w docelowej tabeli w bazie danych na podstawie wartości *id\_klient*. W momencie znalezienia pasujących danych wiersze te zostają przeniesione do tabeli z błędami natomiast nowe dane zostają wpisane do tabeli głównej (proces ten analogicznie zostanie wykorzystany w kolejnych przepływach danych ETL),
- **OLE DB Destination** - jest to obiekt pełniący funkcję połączenia z bazą danych oraz miejsca docelowego przepływających przez ETL danych. W tym przykładzie został on wykorzystany do połączenia z tabelą główną jak i błędów.



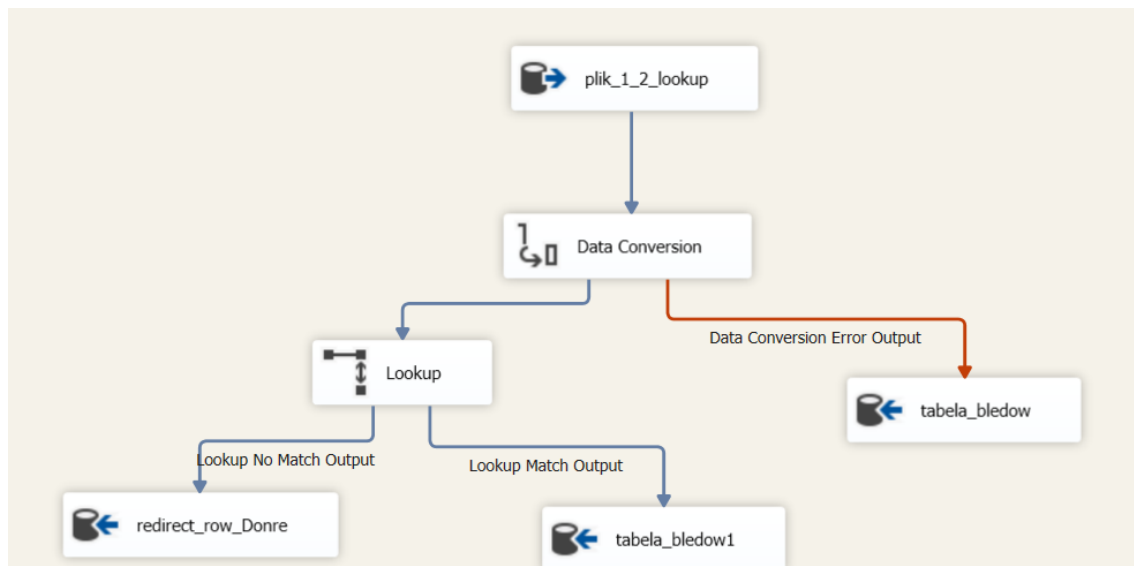
Rysunek 3.2: Widok przepływu nr 1

### 3.2. ETL 2

Poniższy ETL utworzony został przy użyciu funkcji *redirect row* oraz *data conversion*.

**Data conversion** - używany jest do sprawdzenia poprawności typu przesyłanych danych. Sprawdza on takie kolumny jak *id\_klienta*, *data\_zawarcia*, *wiek*, *kwota\_kredytu*, *okres\_kredytu*. W momencie znalezienia błędnych danych przy użyciu zawartej w tej operacji funkcji *redirect row* zostają one przesłane do tabeli błędów. Proces ten uznany jest również jako tzw. "obsługa błędów".

Dodatkowo jak w poprzednim przykładzie wykorzystana została operacja *Lookup* mająca za zadanie nie wprowadzania tych samych danych do tabeli głównej.

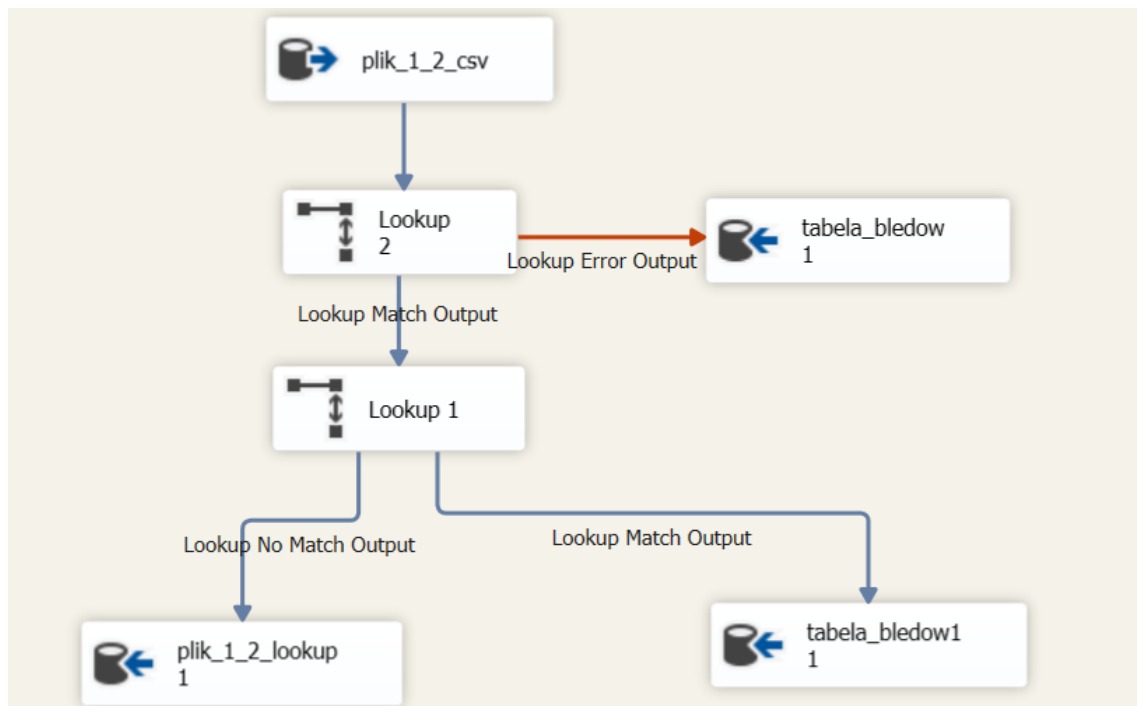


Rysunek 3.3: Widok przepływu nr 2

### 3.3. ETL 3

W przepływie danych nr 3 przedstawiona została operacja *Lookup* z dwóch opcji.

- **Lookup 2** - wykorzystuje zawartą w bazie danych tabelę z zawodami by w odpowiedni sposób przypisać do klientów zawody poprzez ich id jako nowo dołączona kolumna do głównej tabeli z danymi. Jeżeli w tym procesie pojawiają się błędy automatycznie zostają one przesyłane do tabeli błędów z odpowiednimi informacjami o nich.
- **Lookup 1** - analogicznie jak w poprzednich przykładach sprawdza występowanie wprowadzanych danych w docelowej tabeli głównej.



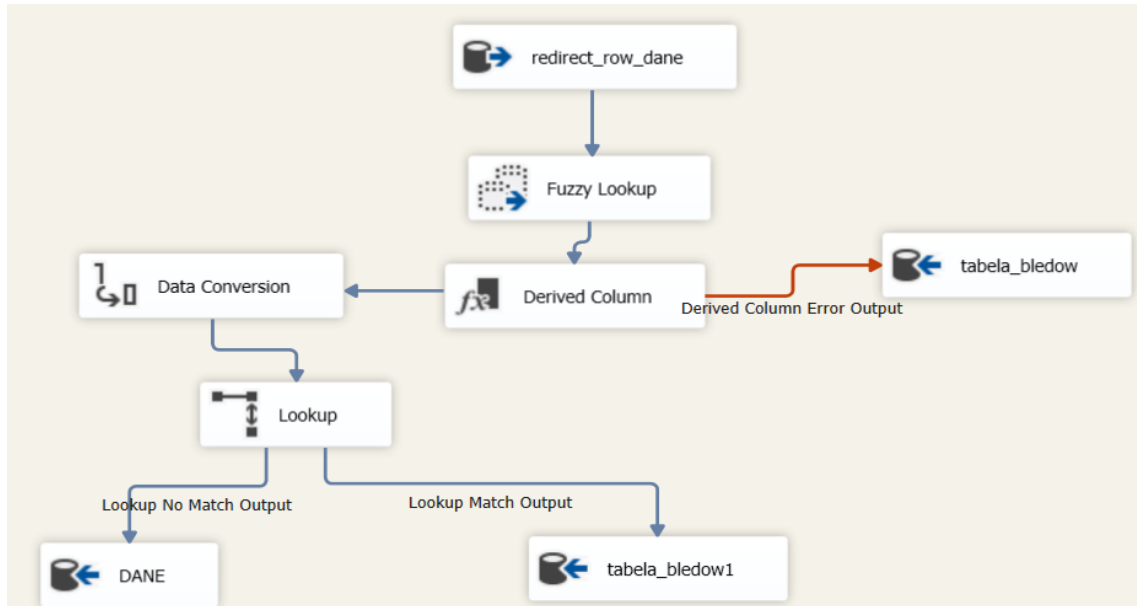
Rysunek 3.4: Widok przepływu nr 3

### 3.4. ETL 4

ETL 4 jest to najbardziej rozbudowany spośród przedstawionych wcześniej przepływów danych. Wykorzystuje on takie operacje jak *Fuzzy Lookup*, *Derived Column*, *Data Conversion* jak i *Lookup*.

- **Fuzzy Lookup** - odpowiada za zamianę znaków na polski w *nazwa\_miasta*. Porównuje pisownię w wspomnianej kolumnie z tabelą z poprawnymi miastami gdzie następnie na podstawie podobieństwa zamienia nazwy miast na odpowiadające tym w tabeli *poprawne\_miasta*,
- **Derived Column** - element ten został dodany ze względu iż przy użyciu operacji *Fuzzy Lookup* w przypadku miasta *Lodz* nie wykrywane były odpowiednie wiersze w głównej tabeli co przełożyło się na przypisanie wartości dla tego miasta jako *NULL*. W tym celu użyty element *Derived Column* zmienia wartości *NULL* na nazwę miasta *Łódź*. Dodatkowo operacja ta została wyposażona w obsługę błędów,

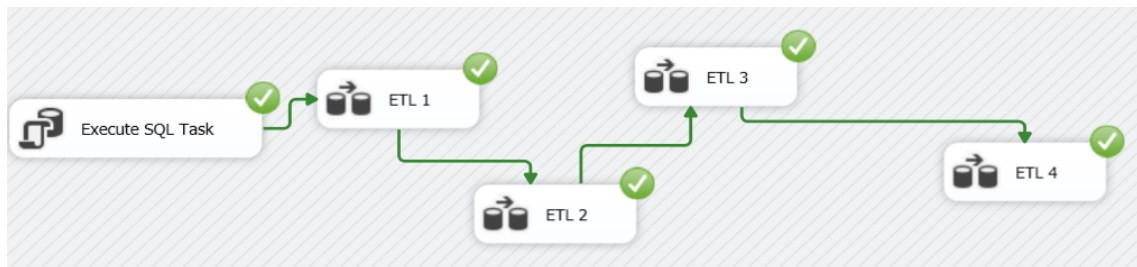
- **Data Conversion** - zmienia typ kolumny *zawod* na *string*,
- **Lookup** - tak jak w poprzednich przykładach sprawdza występowanie odpowiednich wierszy w tabeli.



Rysunek 3.5: Widok przepływu nr 4

### 3.5. Execute SQL Task, finalny przepływ danych

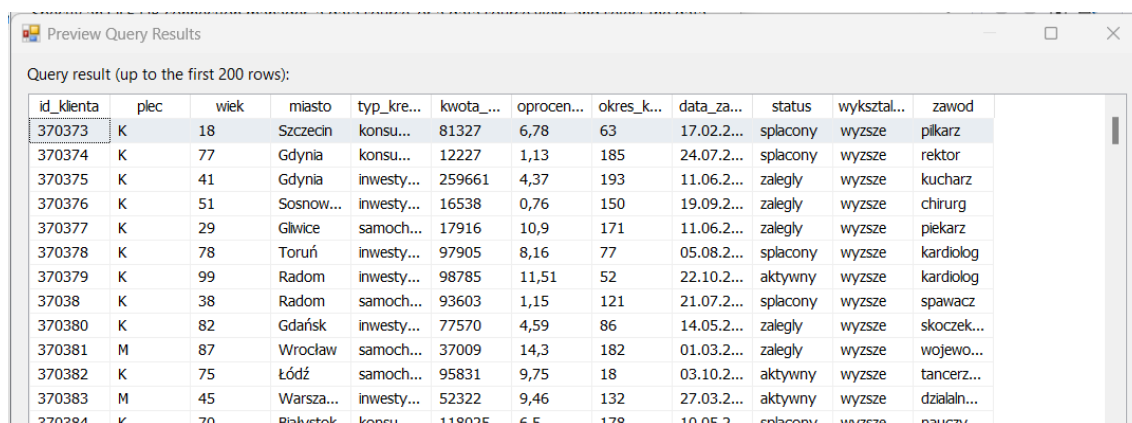
Wszystkie przedstawione przepływy uzupełnia dodatkowy element *Execute SQL Task*, który opróżnia tabelę błędów przed każdym uruchomieniem procesu przepływu danych.



Rysunek 3.6: Poprawnie działający przepływ danych



Poniżej ukazany został podgląd oczyszczonych danych po procesach ETL.



Query result (up to the first 200 rows):

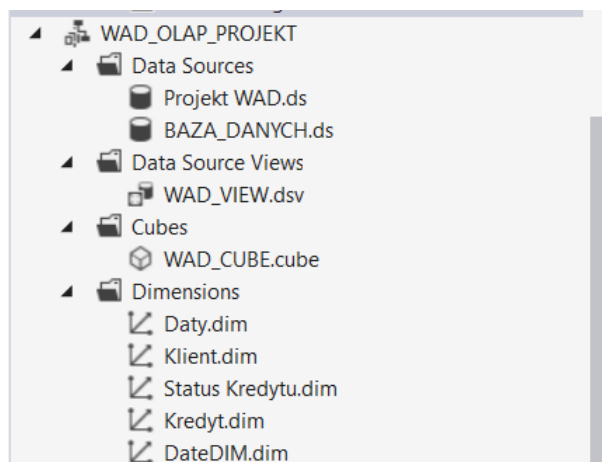
id_klienta	plec	wiek	miasto	typ_kre...	kwota_...	oprocen...	okres_k...	data_z...	status	wyksztal...	zawod
370373	K	18	Szczecin	konsu...	81327	6,78	63	17.02.2...	splacony	wyzsze	piłkarz
370374	K	77	Gdynia	konsu...	12227	1,13	185	24.07.2...	splacony	wyzsze	rektor
370375	K	41	Gdynia	inwesty...	259661	4,37	193	11.06.2...	zalegly	wyzsze	kucharz
370376	K	51	Sosnow...	inwesty...	16538	0,76	150	19.09.2...	zalegly	wyzsze	chirurg
370377	K	29	Głwice	samoch...	17916	10,9	171	11.06.2...	zalegly	wyzsze	piekarz
370378	K	78	Toruń	inwesty...	97905	8,16	77	05.08.2...	splacony	wyzsze	kardiolog
370379	K	99	Radom	inwesty...	98785	11,51	52	22.10.2...	aktywny	wyzsze	kardiolog
37038	K	38	Radom	samoch...	93603	1,15	121	21.07.2...	splacony	wyzsze	spawacz
370380	K	82	Gdańsk	inwesty...	77570	4,59	86	14.05.2...	zalegly	wyzsze	skoczek...
370381	M	87	Wrocław	samoch...	37009	14,3	182	01.03.2...	zalegly	wyzsze	wojewo...
370382	K	75	Łódź	samoch...	95831	9,75	18	03.10.2...	aktywny	wyzsze	tancerz...
370383	M	45	Warsza...	inwesty...	52322	9,46	132	27.03.2...	aktywny	wyzsze	działaln...
370384	K	70	Białystok	konsu...	118025	6,5	178	10.05.2...	splacony	wyzsze	naucz...

Rysunek 3.7: Finalne dane

## 4. Kostka OLAP

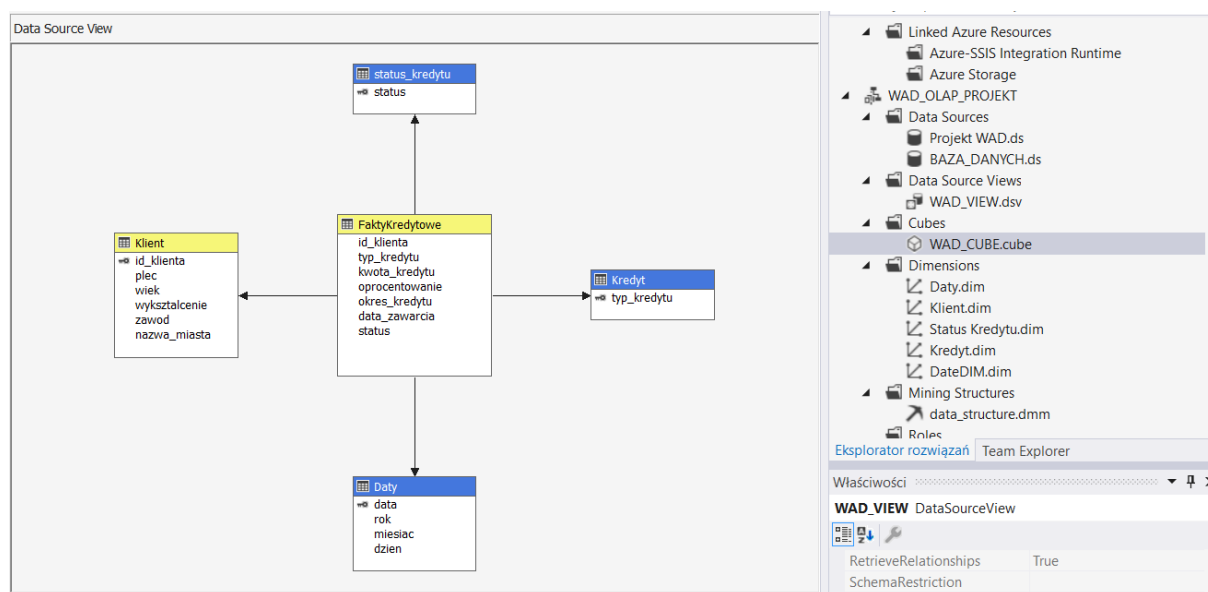
Po przygotowaniu danych, przetworzone dane zostały użyte do utworzenia tabeli faktów oraz tablic wymiarów. Dzięki temu możliwe było zbudowanie kostki OLAP (Online Analytical Processing), która umożliwia szybkie i efektywne wykonywanie zapytań analitycznych. W tym celu został utworzony nowy projekt w Visual Studio przy pomocy dodatku SSAS, który pozwala na tworzenie wielowymiarowej analizy danych jak i data miningu.

Na potrzeby tworzenia kostki wybrane zostało źródło danych, z których następnie został utworzony ich widok jak i kostka OLAP.



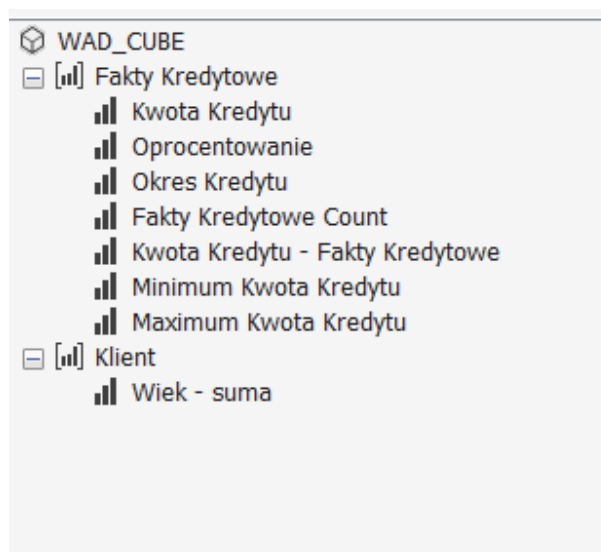
Rysunek 4.8: Elementy kostki

Tak utworzone elementy prezentują poniższą kostkę OLAP.



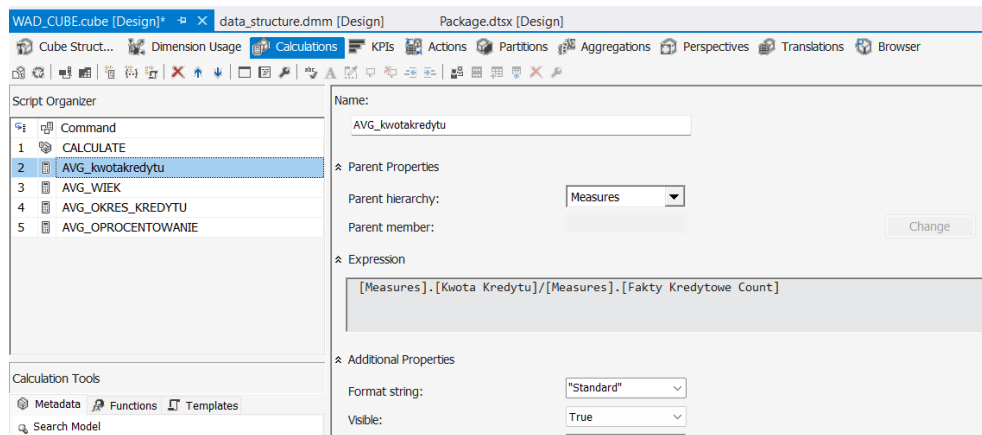
Rysunek 4.9: Kostka OLAP

W powyższej kostce utworzone zostały dodatkowe miary przy użyciu funkcji tworzenia reprezentujące sumy odpowiednich kolumn a także średnią określoną przez czas. Do ostatniej wspomnianej mierze wymagane było pomimo posiadania wymiaru czasu zawierającego wszystkie daty z okresu 01-01-2000 do 31-12-2022 dodanie nowego wymiaru *DateDIM*. Od tego momentu wspomniany wymiar pełnił funkcję wymiaru czasu.

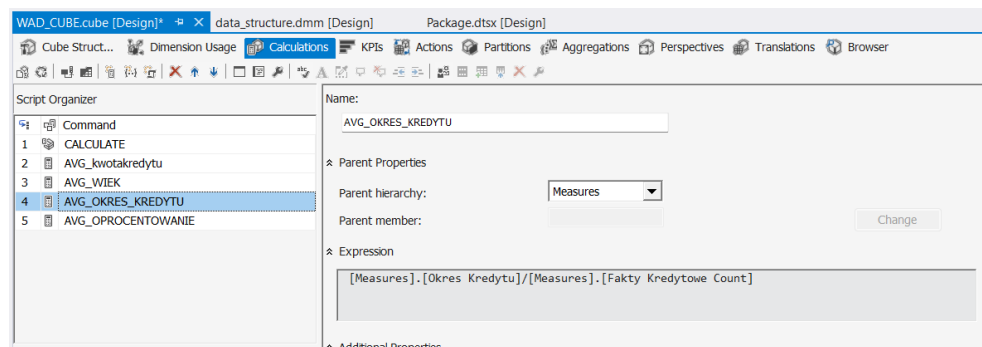


Rysunek 4.10: Utworzone miary przy użyciu funkcji tworzenia

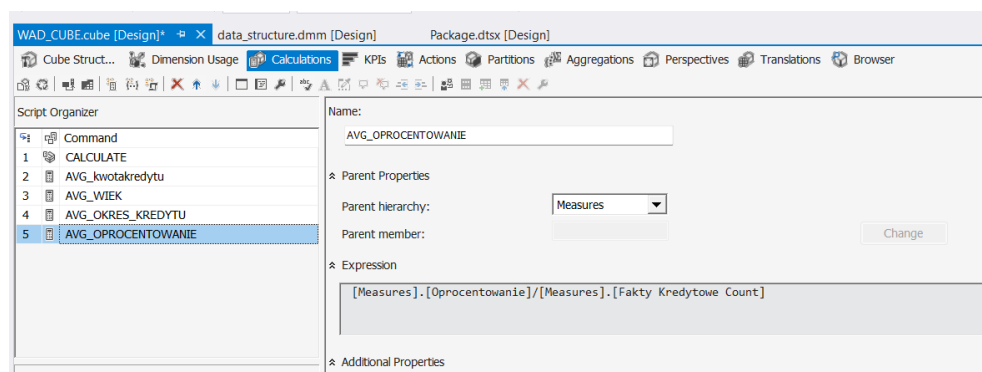
Na podstawie powyższych miar przy użyciu języka DAX w zakładce *CALCULATIONS* utworzone zostały dodatkowe miary reprezentujące średnią wartość odpowiednich argumentów takich jak *kwota\_kredytu*, *wiek*, *okres\_kredytu*, *oprocentowanie*.



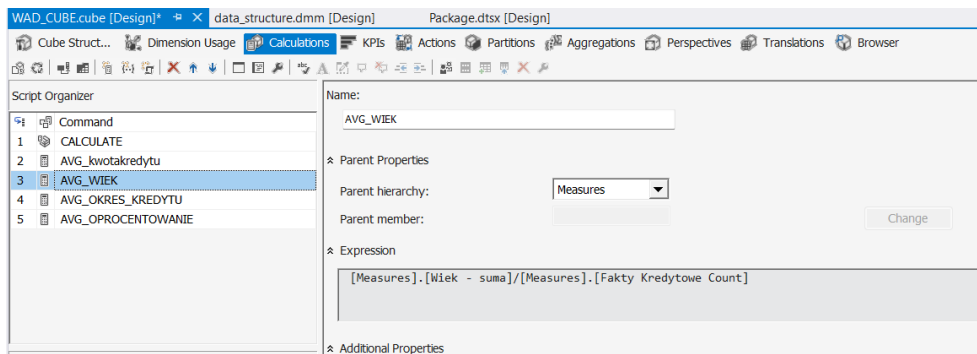
Rysunek 4.11: Miara - średnia kwota kredytu



Rysunek 4.12: Miara - średni okres kredytu



Rysunek 4.13: Miara - średnie oprocentowanie

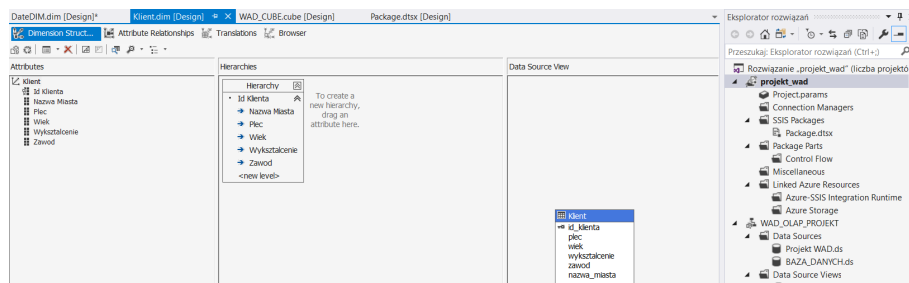


Rysunek 4.14: Miara - średni wiek klienta

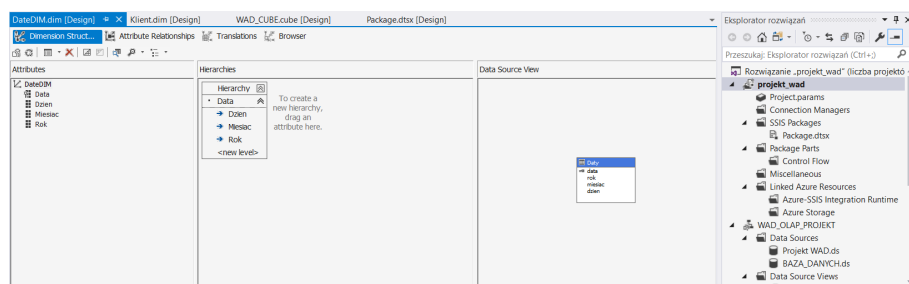
Dzięki tak utworzonym miarom możliwe było zaprojektowanie nowych scenariuszy z określonym KPI (Key Performance Indicator), które zostaną zaprezentowane w kolejnym etapie projektu.

## 4.1. Hierarchie

Dodatkowo w utworzonej kostce OLAP stworzone zostały dwie dla wymiaru czasu a także klienta.



Rysunek 4.15: Hierarchia - Klient



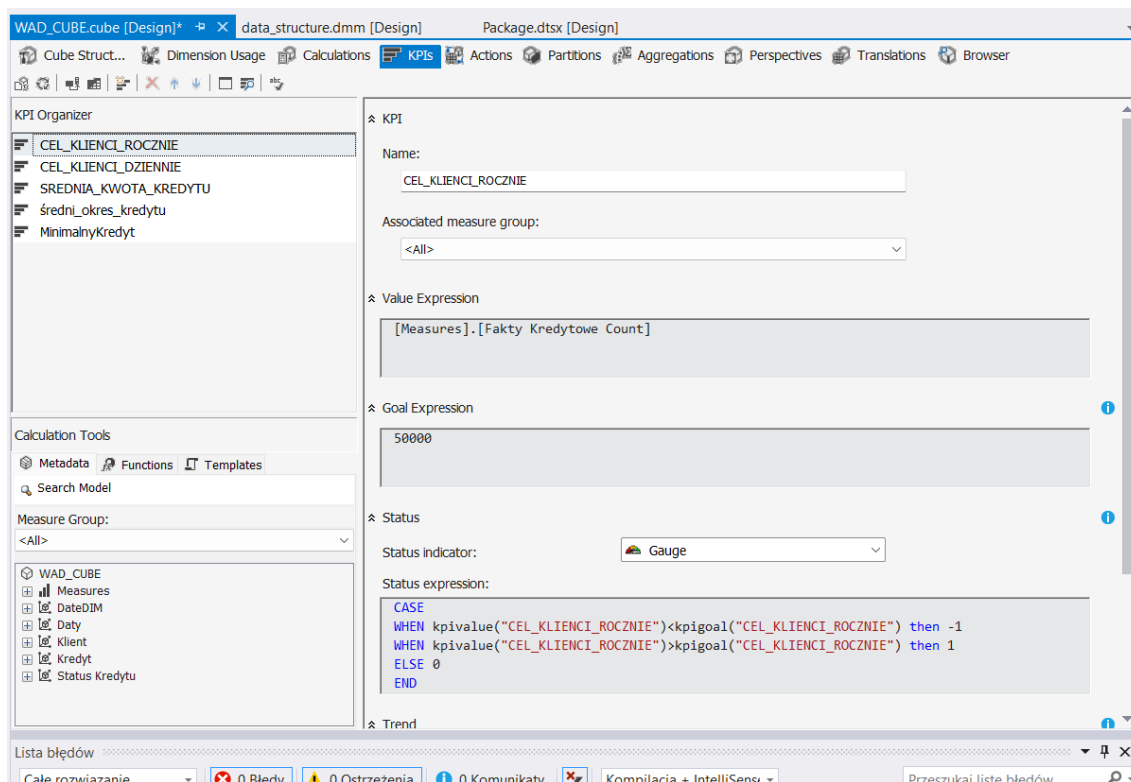
Rysunek 4.16: Hierarchia - DateDIM

## 5. Scenariusze - KPI

W tym punkcie przedstawione zostaną utworzone scenariusze wraz z odpowiednimi do nich KPI. Wszystkie wskaźniki zostały utworzone w modelu kostki w zakładce *KPI*.

### 5.1. Scenariusz 1

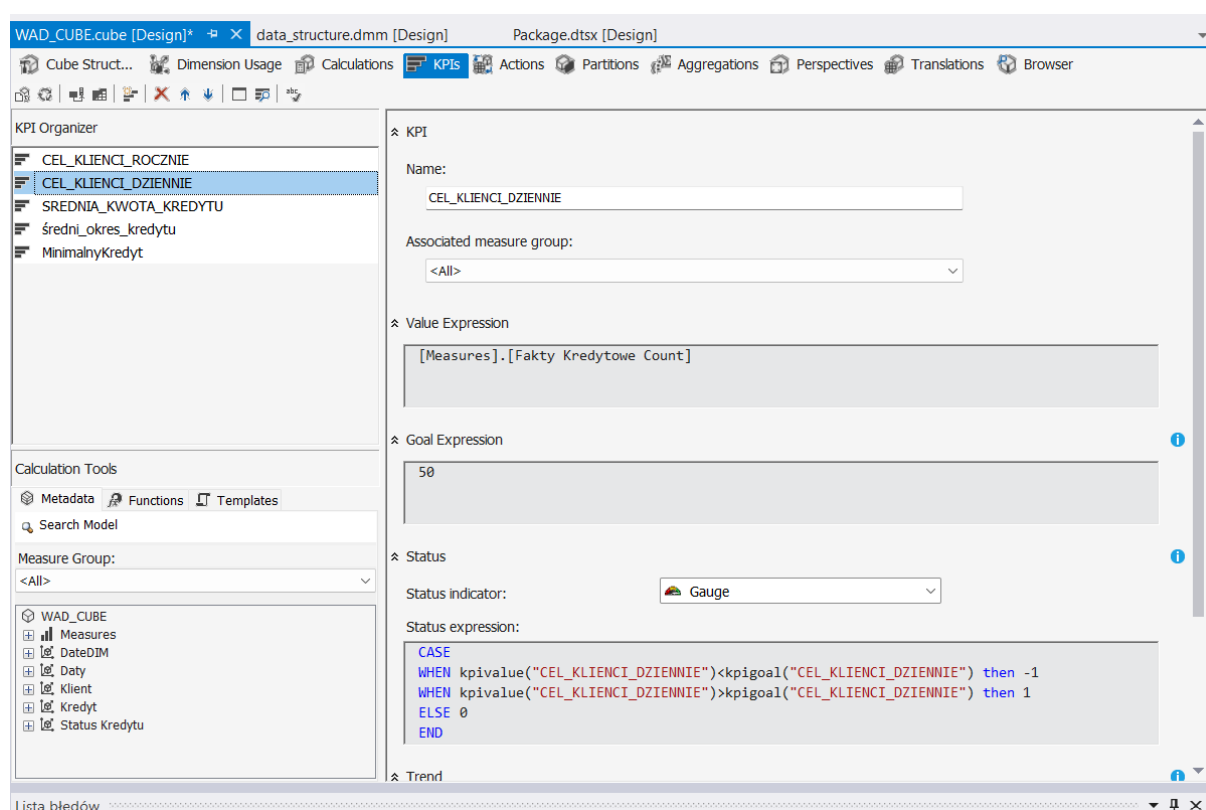
Pierwszym utworzony scenariusz prezentuje minimalny cel ilości rocznych klientów, którym bank udziela kredytu. W tym celu utworzony został wskaźnik przedstawiający tą wartość. Cel został ustawiony na 50000. Jako wartość analizowana wykrozystano *[Measures].[Fakty Kredytowe Count]*. Dodatkowo zostały dodane odpowiednie reakcje na wartość poziomu KPI. Wskaźnik ten może pomóc bankowi udzielającemu kredyty sprawdzić czy przyciąga on odpowiednią ilość klientów. Analizując tą wartość może wprowadzać nowe sposoby reklamy czy zachęty klientów do wybrania właśnie ich usług a także podniesienia innych standardów wspomagającym temu procesowi takich jak na przykład obsługa klienta.



Rysunek 5.17: KPI 1 - roczna ilość klientów

## 5.2. Scenariusz 2

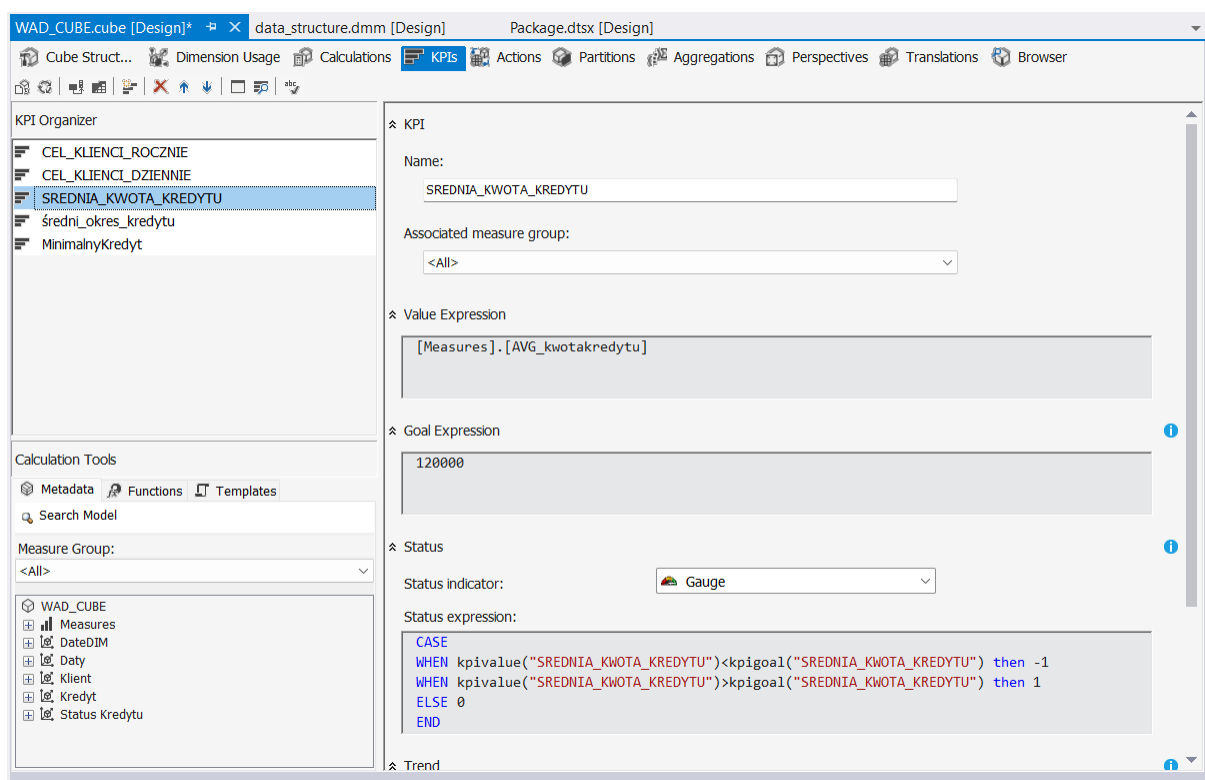
Scenariusz nr 2 podobnie jak w poprzednim dotyczy ilości klientów korzystających z usług banku lecz w tym przypadku dotyczy to wartości dziennej. Cel posiada wartość 50 a analizowana miara została analogicznie wykorzystana jak w poprzednim przykładzie. KPI dotyczący minimalnej liczby obsłużonych klientów dziennie może motywować pracowników do optymalizacji swoich działań, co może prowadzić do skrócenia czasu obsługi klienta i poprawy ogólnej efektywności operacyjnej banku. Monitorowanie liczby obsłużonych klientów dziennie pozwala bankowi lepiej zrozumieć przepływ klientów i dostosować alokację zasobów, takich jak personel i infrastruktura, aby lepiej sprostać zapotrzebowaniu w różnych oddziałach i w różnych porach dnia.



Rysunek 5.18: KPI 2 - dzienna ilość klientów

### 5.3. Scenariusz 3

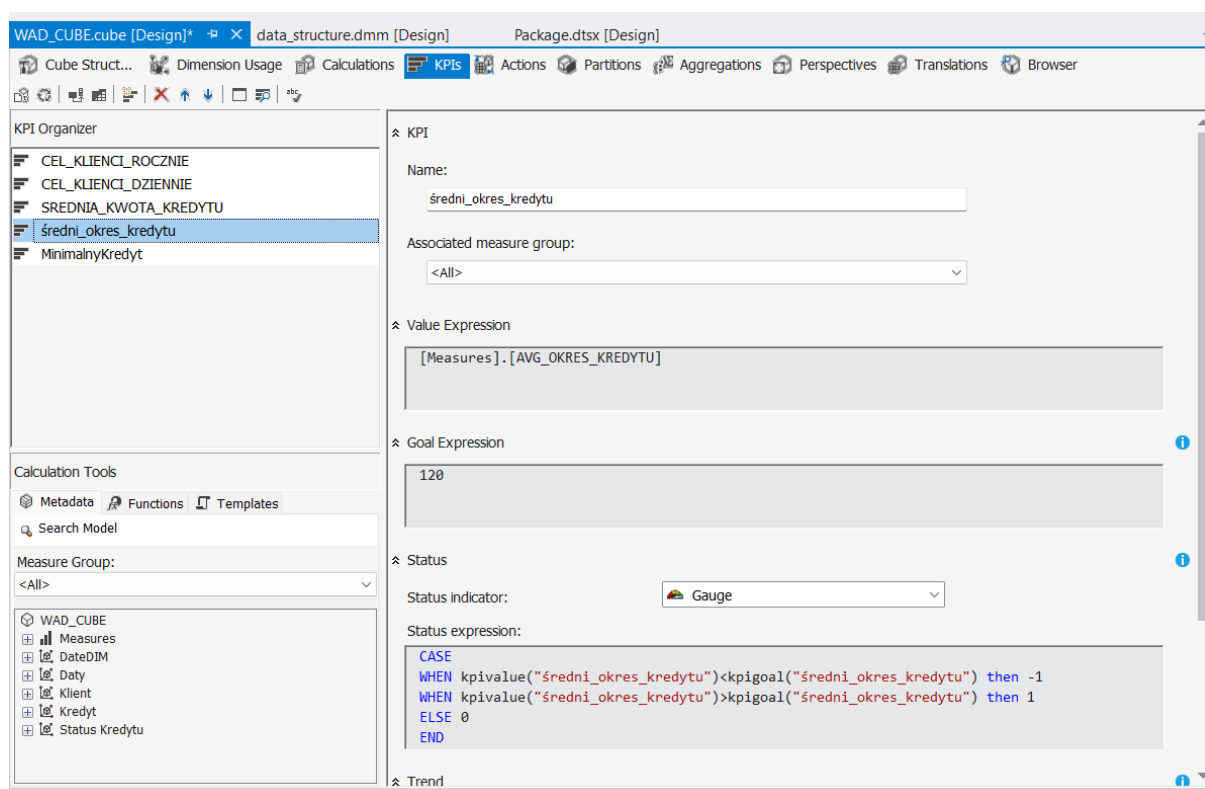
Scenariusz nr 3 przedstawia średnią udzielaną kwotę kredytu. Celem banku jest by ta od początku jego istnienia miała wartość minimalnie 120000. Do analizy tego wskaźnika wykorzystano miarę  $[Measures].[AVG\_kwotakredytu]$ . Wyższa średnia kwota udzielanego kredytu bezpośrednio przekłada się na większe przychody z odsetek i opłat. Kredyty o wyższych kwotach generują większe dochody dla banku w dłuższym okresie, co jest korzystne dla jego wyników finansowych. Konsekwentne udzielanie kredytów o wyższej wartości może pomóc bankowi w budowaniu reputacji jako instytucji finansowej zdolnej do obsługi większych i bardziej skomplikowanych potrzeb finansowych klientów. Może to również przyciągnąć klientów poszukujących większych kredytów. Klienci poszukujący wyższych kwot kredytów oczekują profesjonalnej obsługi i doradztwa. Skupienie się na takim KPI może motywować bank do podnoszenia standardów obsługi, co zwiększa satysfakcję klientów i ich lojalność.



Rysunek 5.19: KPI 3 - średnia kwota udzielanego kredytu

## 5.4. Scenariusz 4

W tym przykładzie wzięty zostanie pod uwagę średni okres udzielanego kredytu. Jego cel określony jest jako minimum 120 miesięcy a analizowaną do tego miarą jest `[Measures].[AVG_OKRES_KREDYTU]`. Kredyty o dłuższym okresie spłaty zapewniają bankowi stabilny i przewidywalny przepływ dochodów z odsetek przez dłuższy czas. Długoterminowe kredyty mogą generować regularne przychody, co przyczynia się do stabilności finansowej banku. Dłuższy okres spłaty może pozwolić bankowi oferować klientom bardziej korzystne warunki kredytowania, takie jak niższe miesięczne raty. To może przyciągnąć więcej klientów i zwiększyć konkurencyjność banku na rynku. Dłuższe okresy spłaty mogą być mniej ryzykowne dla banku, pod warunkiem, że są udzielane kredytobiorcom z dobrą zdolnością kredytową. Rozłożenie spłat na dłuższy okres może zmniejszyć miesięczne obciążenia finansowe klientów.

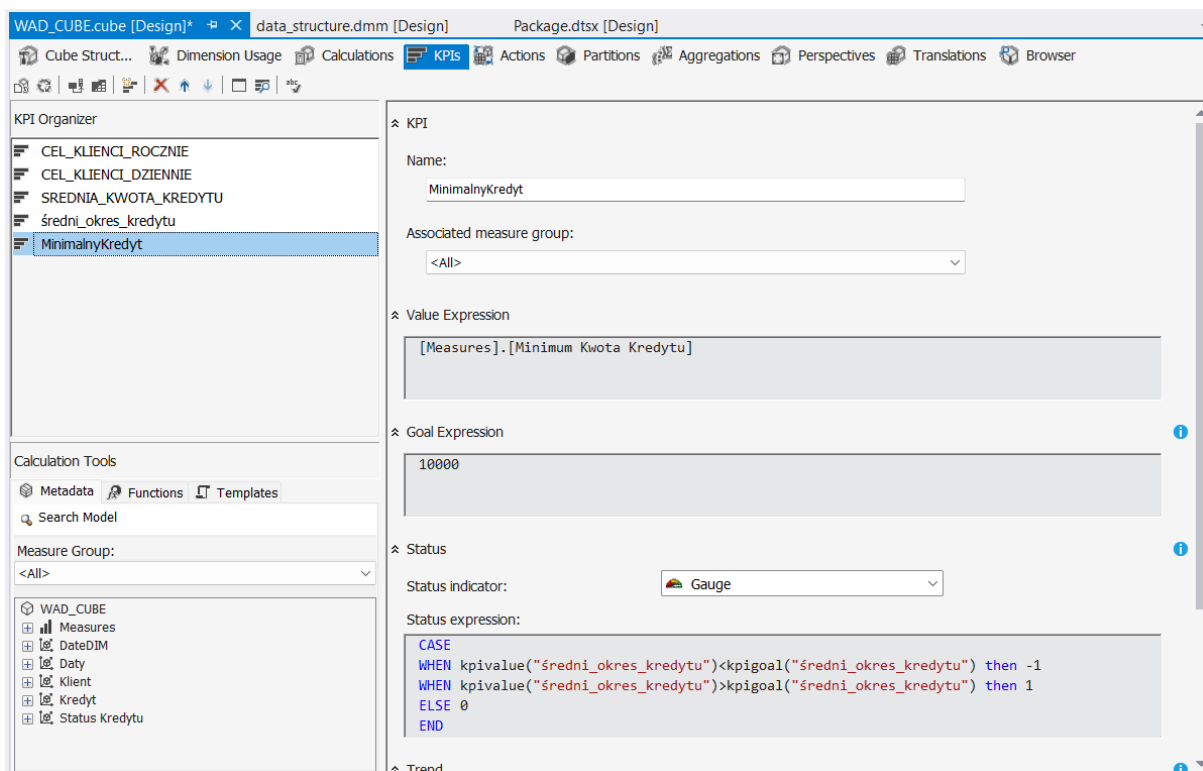


Rysunek 5.20: KPI 4 - średni okres udzielanego kredytu



## 5.5. Scenariusz 5

Ostatni scenariusz przedstawia minimalną wartość udzielanego kredytu by ta wynosiła minimalnie 10000. Do analizy wykorzystano *[Measures].[Minimum Kwota Kredytu]*. Udzielanie większych kredytów może generować bankowi większe zyski ze względu na wyższe odsetki i prowizje. Obsługa mniejszej liczby transakcji o mniejszej wartości może być mniej opłacalna ze względu na koszty operacyjne. Koncentracja na większych kredytach może ułatwić zarządzanie i zmniejszyć koszty administracyjne. Większe kredyty mogą być bardziej rentowne dla banku, ponieważ mogą one być skierowane do bardziej stabilnych klientów lub projektów, co zmniejsza ryzyko niewypłacalności. Klienci otrzymujący większe kredyty mogą być bardziej skłonni do korzystania z innych usług bankowych, co może zwiększyć lojalność i wartość życia klienta dla banku.

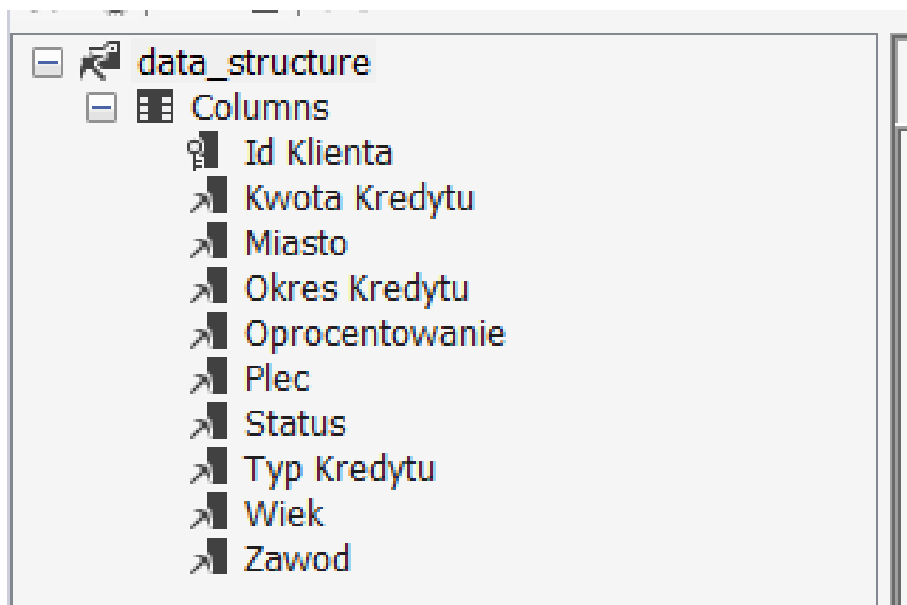


Rysunek 5.21: KPI 5 - minimalna kwota udzielanego kredytu

## 6. Data mining

W tym punkcie wykorzystana zostanie operacja Data mining czyli proces odkrywania wzorców, trendów i relacji w dużych zbiorach danych. W Visual Studio daje możliwość używania różnych narzędzi i bibliotek do przeprowadzania analizy danych i wykonywania operacji związanych z Data Mining. W projekcie skupiona została uwaga na takich opcjach jak: *Drzewo decyzyjne*, *Sieć neuronowa*, *Klastrowanie*.

Głównym celem wykorzystania modeli Data mining będzie predykcja kwoty udzielanego kredytu na podstawie wielu parametrów przy użyciu klucza jako *id\_klienta*. Wybrane elementy zostały przedstawione poniżej.



Rysunek 6.22: Wybrane parametry do predykcji

Structure	data_structure	Neural_network	cluster
	Microsoft_Decision_Trees	Microsoft_Neural_Network	Microsoft_Clustering
Id Klienta	Key	Key	Key
Kwota Kredytu	PredictOnly	PredictOnly	PredictOnly
Miasto	Input	Input	Input
Okres Kredytu	Input	Input	Input
Oprocentowanie	Input	Input	Input
Plec	Input	Input	Input
Status	Input	Input	Input
Typ Kredytu	Input	Input	Input
Wiek	Input	Input	Input
Zawod	Input	Input	Input

Rysunek 6.23: Parametry wchodzące w skład modeli

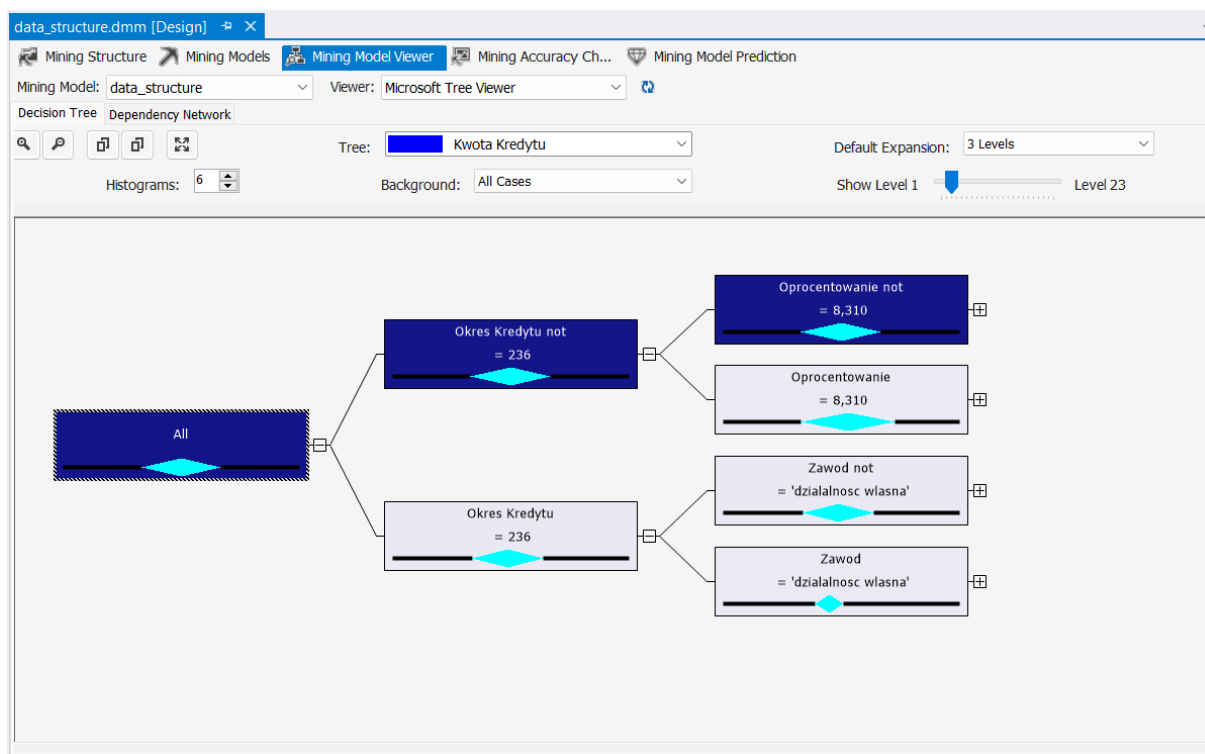
W kolejnych etapach zaprezentowane zostaną widoki jak i wyniki uzyskane poprzez wymienione modele.

## 6.1. Drzewo decyzyjne

Pierwszym wybranym modelem do predykcji kwoty kredytu na podstawie wybranych parametrów jest drzewo decyzyjne. Drzewo decyzyjne to model predykcyjny w analizie danych, który używa struktury drzewa do reprezentowania i podejmowania decyzji na podstawie zestawu warunków. Każdy wierzchołek drzewa reprezentuje test na jednej ze zmiennych, a każda gałąź wychodząca z wierzchołka reprezentuje możliwy wynik tego testu. Model ten jest często wykorzystywany do klasyfikacji i prognozowania, ponieważ umożliwia interpretację procesu decyzyjnego.

### Wizualizacja

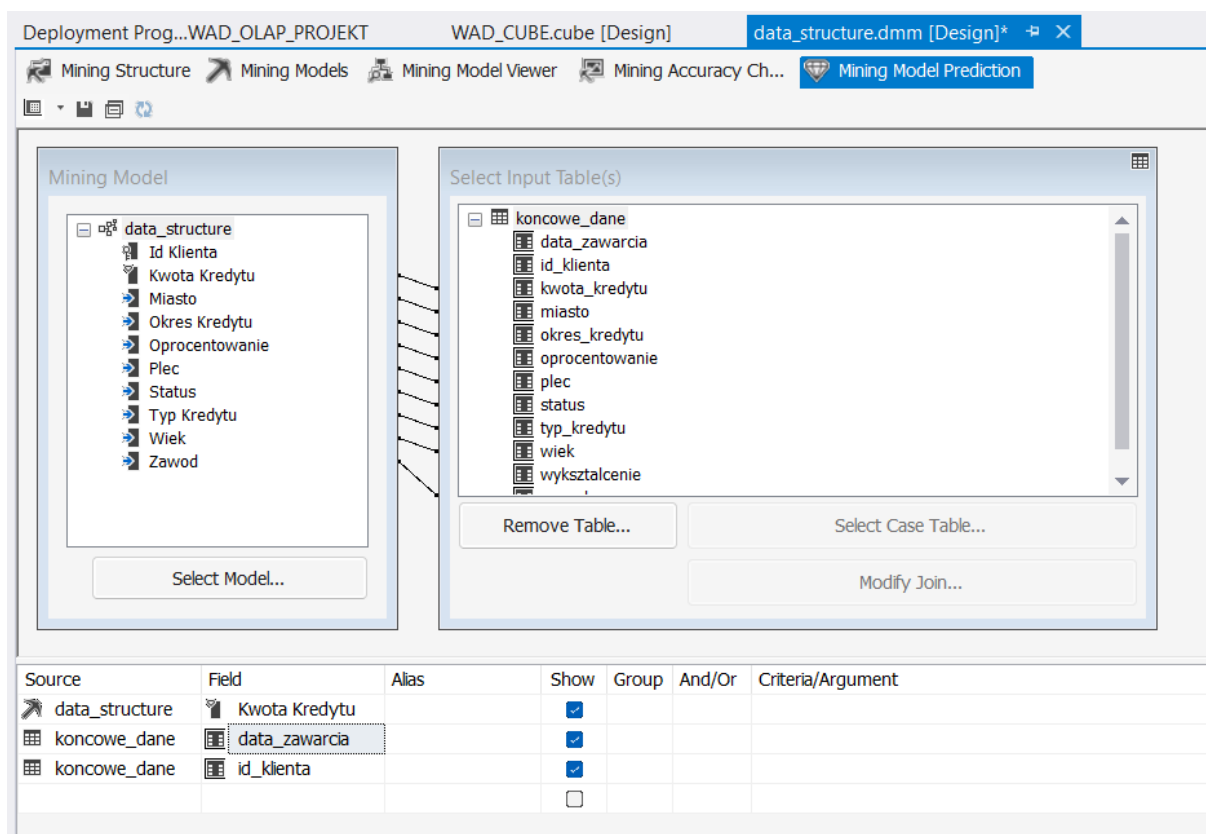
Poniżej przedstawiona została wizualizacja utworzonego drzewa decyzyjnego. Ukazane ono zostało tylko w formacie 3 poziomów ze względu na jego wielkość. Poprzez wciśnięcie znaków ”+” można rozszerzyć jego widok.



Rysunek 6.24: Utworzone drzewo decyzyjne

## Wyniki predykcji

Po utworzeniu powyższego drzewa została wykonana predykcja dla kwoty kredytu udzielanego klientom. Poniżej zostają przedstawione poniżej wraz z odpowiednim doborem zmiennych, który zostanie wykorzystany w ten sam sposób dla każdego modelu.



Rysunek 6.25: Tworzenie predykcji

WAD_VIEW.dsv [Design]*		Deployment Prog...WAD_OLAP_PROJEKT	WAD_CUBE.cube [Design]*	data_structure.dmm [Design]*
Mining Structure	Mining Models	Mining Model Viewer	Mining Accuracy Ch...	Mining Model Prediction
Kwota Kred...	data_zawarcia	id_klienta		
144722,68...	17.02.2014...	370373		
144722,68...	24.07.2020...	370374		
144722,68...	11.06.2021...	370375		
144722,68...	19.09.2013...	370376		
144722,68...	11.06.2015...	370377		
144722,68...	05.08.2006...	370378		
144722,68...	22.10.2005...	370379		
144722,68...	21.07.2020...	37038		
144722,68...	14.05.2003...	370380		
144722,68...	01.03.2021...	370381		
144722,68...	03.10.2018...	370382		
144722,68...	27.03.2018...	370383		
144722,68...	10.05.2015...	370384		
144722,68...	22.05.2018...	370385		
144722,68...	31.08.2008...	370386		
144722,68...	01.01.2011...	370387		
144722,68...	26.02.2004...	370388		
144722,68...	18.01.2002...	370389		
144722,68...	23.06.2001...	37039		
144722,68...	25.03.2010...	370390		
144722,68...	21.04.2015...	370391		
144722,68...	24.06.2005...	370392		
144722,68...	22.12.2016...	370393		
144722,68...	30.09.2019...	370394		
144722,68...	22.02.2010...	370395		

Query run completed with 700000 rows fetched

Rysunek 6.26: Wyniki predykcji

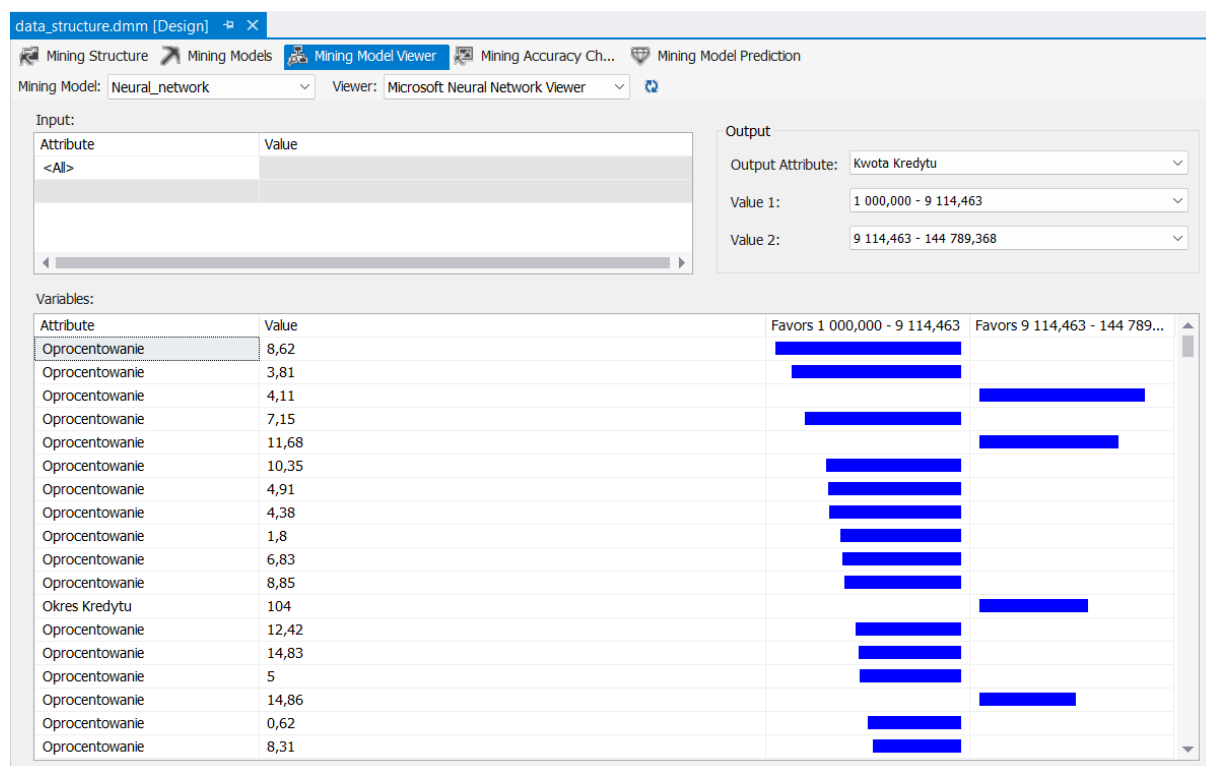
Jak można zauważyć wyniki drzewa decyzyjnego w większości przypadków przedstawiały te same wartości. Analizując dogłębniej można było zauważyć, że są one różne dla innych wierszy.

## 6.2. Sieć neuronowa

Kolejnym wykorzystanym modelem jest sieć neuronowa. Sieć neuronowa to model obliczeniowy inspirowany strukturą mózgu, składający się z połączonych sztucznych neuronów, które przetwarzają i przekształcają dane wejściowe, generując odpowiedź na wyjściu. Jest to rodzaj algorytmu uczenia maszynowego, który może być używany do klasyfikacji, regresji, rozpoznawania wzorców i innych zadań, zdolny do uczenia się na podstawie doświadczenia poprzez dostosowywanie wagi połączeń między neuronami. Sieci neuronowe znalazły zastosowanie w wielu dziedzinach, takich jak przetwarzanie obrazów, przetwarzanie języka naturalnego, rozpoznawanie mowy i wiele innych. Dla tego przypadku analogicznie jak w poprzednim zostały wykonane te same operacje.

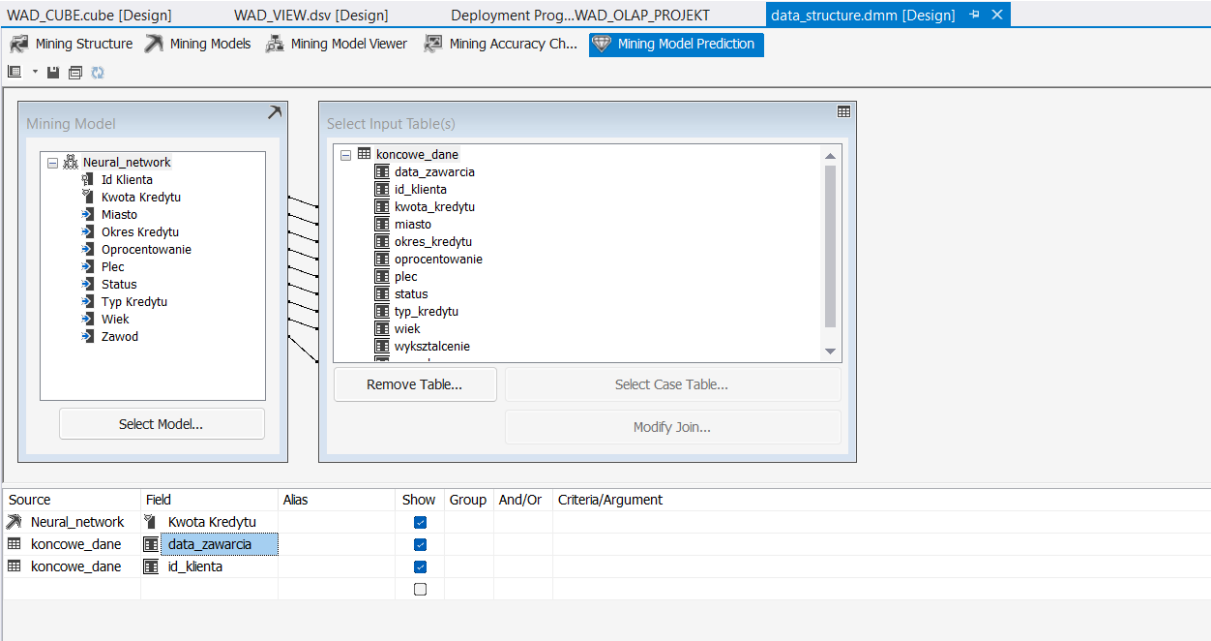
### Wizualizacja

Poniżej przedstawiony został obraz wyliczonych parametrów dla sieci neuronowej w Visual Studio.



Rysunek 6.27: Utworzona sieć neuronowa

# Wyniki predykcji



Rysunek 6.28: Tworzenie predykcji

WAD\_CUBE.cube [Design]WAD\_VIEW.dsv [Design]Deployment Prog...WAD\_OLAP\_PROJEKTdata\_structure.dmm [Design]

Mining StructureMining ModelsMining Model ViewerMining Accuracy Ch...Mining Model Prediction

Kwota Kredytu	data_zawarcia	id_klienta
139554,666645...	17.02.2014 00:00:00	370373
169111,886454...	24.07.2020 00:00:00	370374
129501,210120...	11.06.2021 00:00:00	370375
129033,504492...	19.09.2013 00:00:00	370376
127556,393820...	11.06.2015 00:00:00	370377
116298,675897...	05.08.2006 00:00:00	370378
156180,948250...	22.10.2005 00:00:00	370379
170635,05496122	21.07.2020 00:00:00	37038
124590,440868...	14.05.2003 00:00:00	370380
154900,667331...	01.03.2021 00:00:00	370381
188552,879153...	03.10.2018 00:00:00	370382
164812,99584847	27.03.2018 00:00:00	370383
155327,286453...	10.05.2015 00:00:00	370384
162525,847328...	22.05.2018 00:00:00	370385
139495,495929...	31.08.2008 00:00:00	370386
118955,479567...	01.01.2011 00:00:00	370387
167589,060449...	26.02.2004 00:00:00	370388
221049,254770...	18.01.2002 00:00:00	370389
134204,141864...	23.06.2001 00:00:00	37039
185641,402952...	25.03.2010 00:00:00	370390
162426,44803719	21.04.2015 00:00:00	370391
128248,156646...	24.06.2005 00:00:00	370392
169954,011113...	22.12.2016 00:00:00	370393
146604,08242575	30.09.2019 00:00:00	370394
118542,250506...	22.02.2010 00:00:00	370395

Query run completed with 700000 rows fetched

Rysunek 6.29: Wyniki predykcji

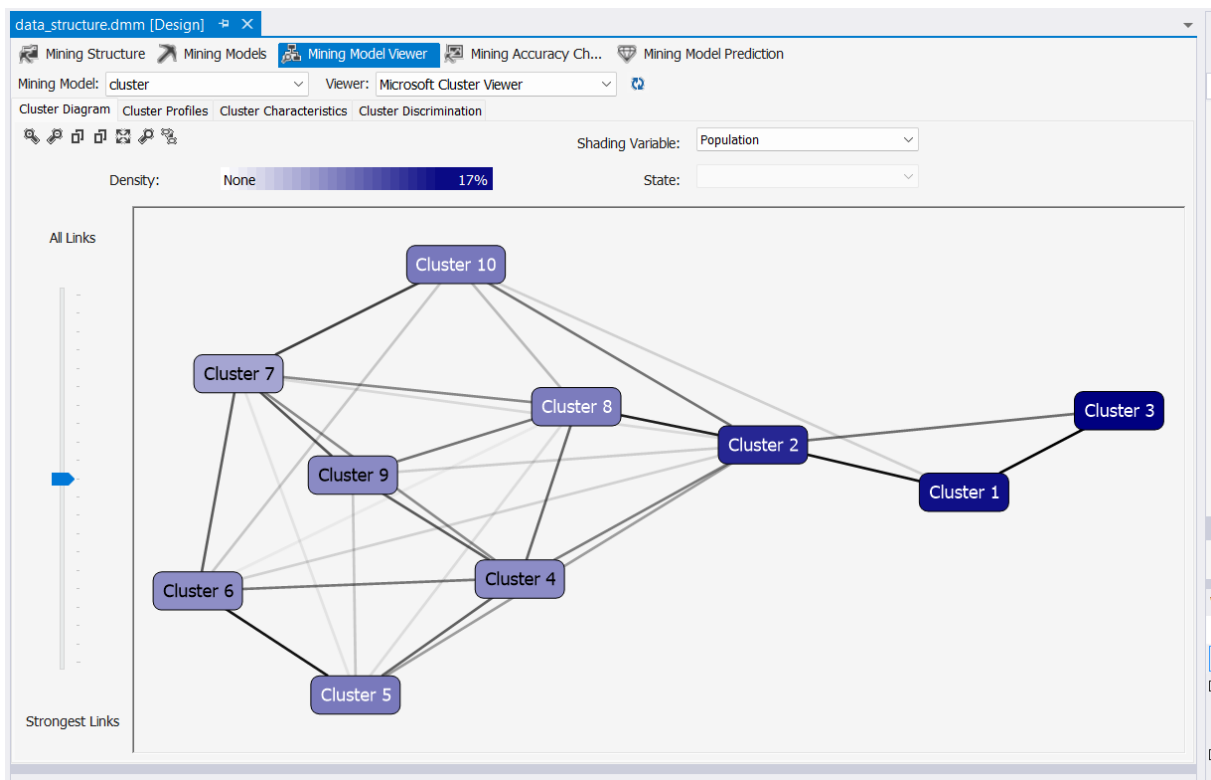
Jak można zauważyć wyniki predykcji dla sieci neuronowej w dużej mierze różnią się od wyników dla drzewa decyzyjnego. Przedstawiają one różne wartości.

### 6.3. Klastrowanie

Kolejnym wykorzystanym modelem są klastry. Klastrowanie to technika analizy danych używana do grupowania zbioru danych na podstawie podobieństwa między ich punktami. Celem jest znalezienie naturalnych grup (klastrów) w danych, gdzie obiekty wewnątrz klastra są bardziej podobne do siebie niż do obiektów w innych klastrach. Istnieje wiele różnych metod klastrowania, takich jak k-means, hierarchiczne klastrowanie czy klastrowanie gęstościowe, które mogą być stosowane w zależności od struktury danych i oczekiwanych wyników. Klastrowanie znajduje zastosowanie w analizie danych, uczeniu nienadzorowanym, segmentacji rynku czy wykrywaniu anomalii.

#### Wizualizacja

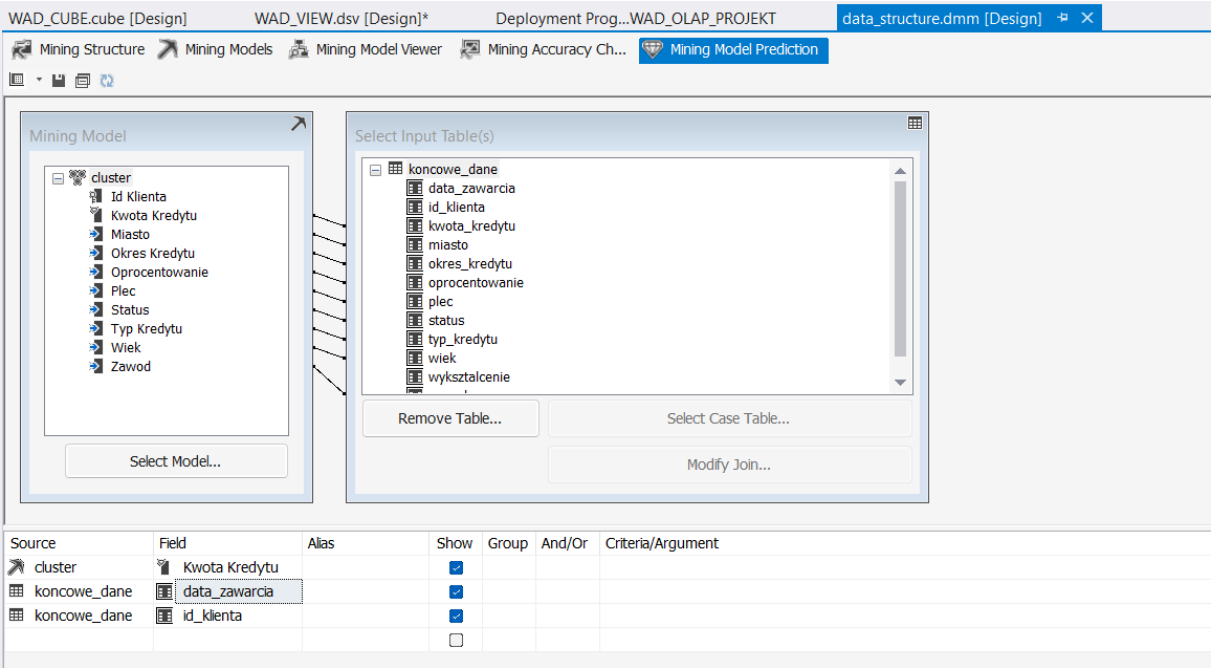
Poniżej zaprezentowane zostały utworzone klastry dla modelu w Visual Studio.



Rysunek 6.30: Utworzone klastry



# Wyniki predykcji



Rysunek 6.31: Tworzenie predykcji

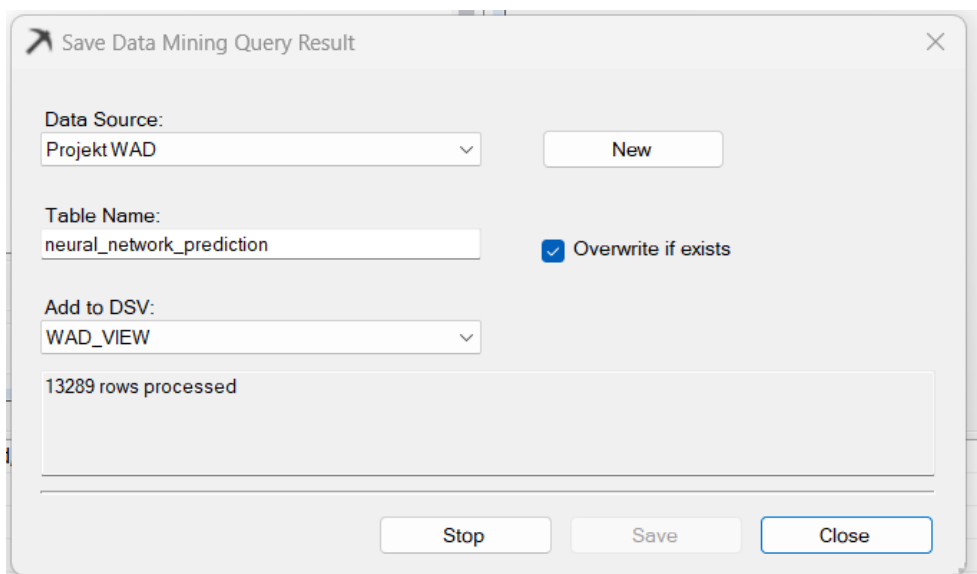
Kwota Kredytu	data_zawarcia	id_klienta
144864,99337891	17.02.2014 00:00:00	370373
144847,382196732	24.07.2020 00:00:00	370374
144824,635213987	11.06.2021 00:00:00	370375
144412,820007191	19.09.2013 00:00:00	370376
144723,599069384	11.06.2015 00:00:00	370377
145264,729809652	05.08.2006 00:00:00	370378
145015,496656195	22.10.2005 00:00:00	370379
145442,065024098	21.07.2020 00:00:00	37038
144539,113499085	14.05.2003 00:00:00	370380
144353,629947039	01.03.2021 00:00:00	370381
144575,522648085	03.10.2018 00:00:00	370382
145399,314814623	27.03.2018 00:00:00	370383
144668,559945018	10.05.2015 00:00:00	370384
144635,136456379	22.05.2018 00:00:00	370385
144394,725981445	31.08.2008 00:00:00	370386
144397,78336477	01.01.2011 00:00:00	370387
144234,717623555	26.02.2004 00:00:00	370388
144504,708253485	18.01.2002 00:00:00	370389
145622,109924241	23.06.2001 00:00:00	37039
145247,024836821	25.03.2010 00:00:00	370390
144593,885788007	21.04.2015 00:00:00	370391
144907,744733778	24.06.2005 00:00:00	370392
145247,690856288	22.12.2016 00:00:00	370393
144233,969019753	30.09.2019 00:00:00	370394
144896,116359837	23.03.2019 00:00:00	370395

Query run completed with 700000 rows fetched

Rysunek 6.32: Wyniki predykcji

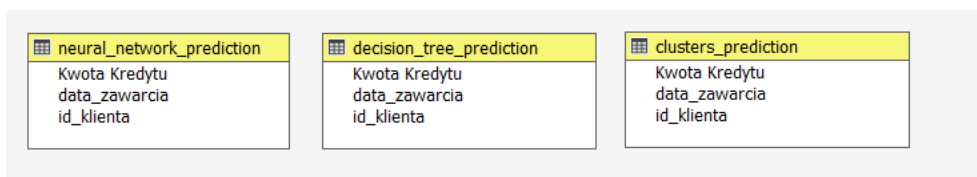
Uzyskane wartości różnią się od pozostałych lecz są one zbliżone do wyników otrzymanych z drzewa decyzyjnego.

Wizualizacja danych otrzymanych przy użyciu powyższych modelu Data Miningu zostanie ukazana w sekcji poświęconej Power BI, gdzie zostanie porównana z średnią kwotą roczną udzielanego kredytu. W tym celu wykonany został zapis danych do odpowiednich tabeli o nazwach *neural\_network\_prediction*, *clusters\_prediction*, *decision\_tree\_prediction*.

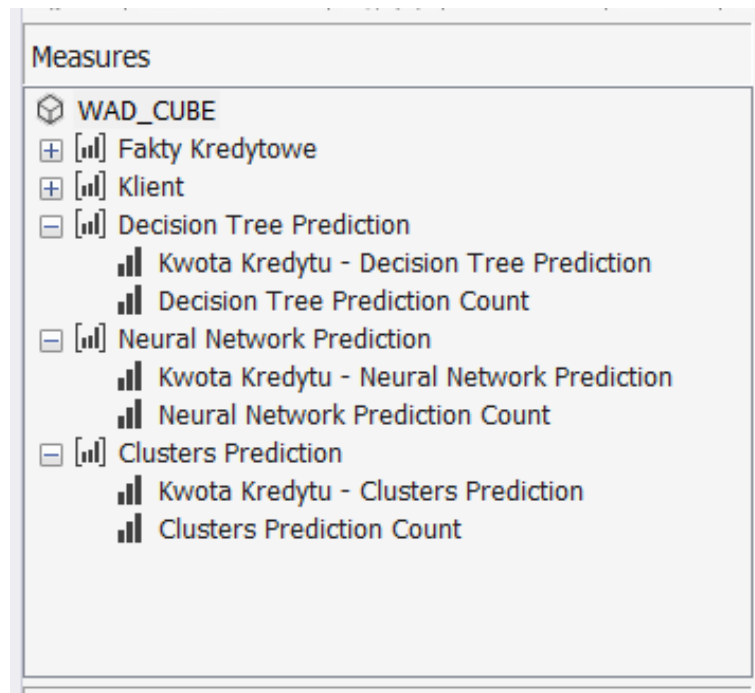


Rysunek 6.33: Przykład zapisu

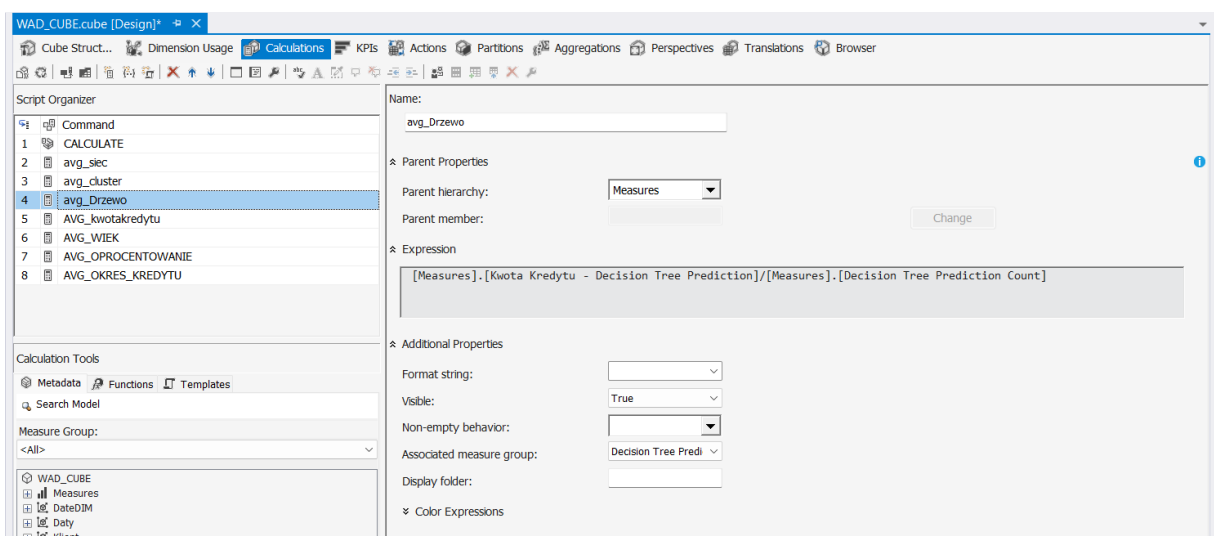
W celu wyliczenia średniej przewidywanej wartości dla danego roku zostały utworzone nowe miary zwracające sumę oraz liczbę wierszy danej tabeli. Dodatkowo w zakładce *CALCULATIONS* bazując na poprzednich przykładach utworzona została miara zawierająca średnią wartość kredytu. By wszystkie miary działały poprawnie tabele zostały połączone z tabelą wymiaru czasu poprzez odpowiednie relacje.



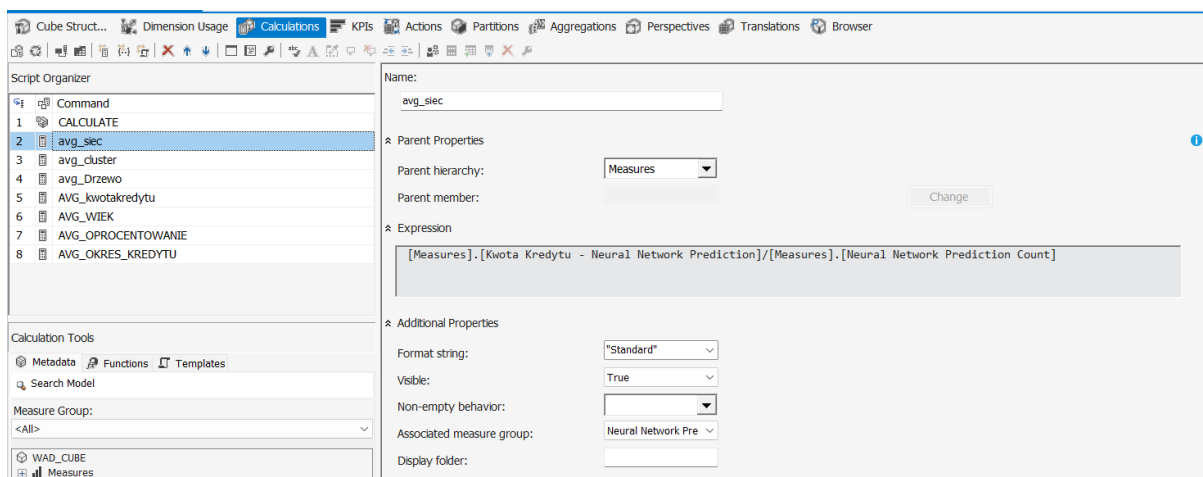
Rysunek 6.34: Tabele znajdujące się w kostce



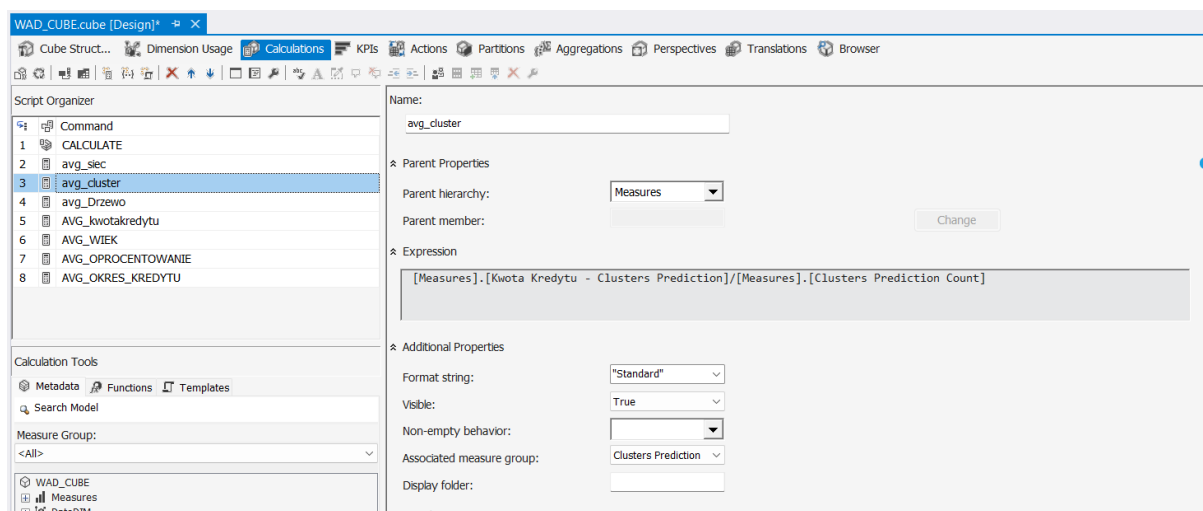
Rysunek 6.35: Miary utworzone dla wyników Data Mining



Rysunek 6.36: Miara średnia wartość kredytu - Drzewo decyzyjne

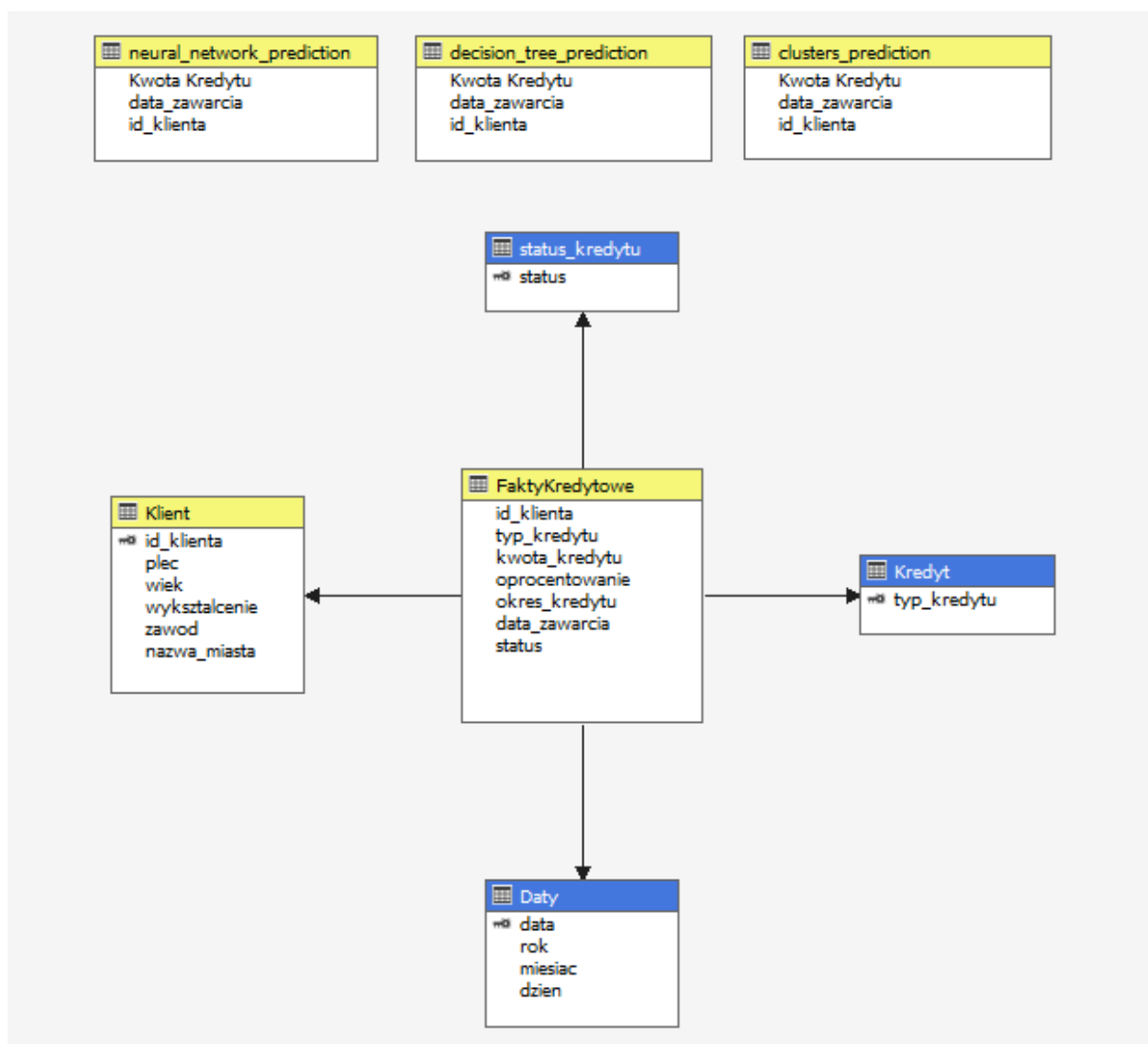


Rysunek 6.37: Miara średnia wartość kredytu - Sieć neuronowa



Rysunek 6.38: Miara średnia wartość kredytu - Klasyfikacja

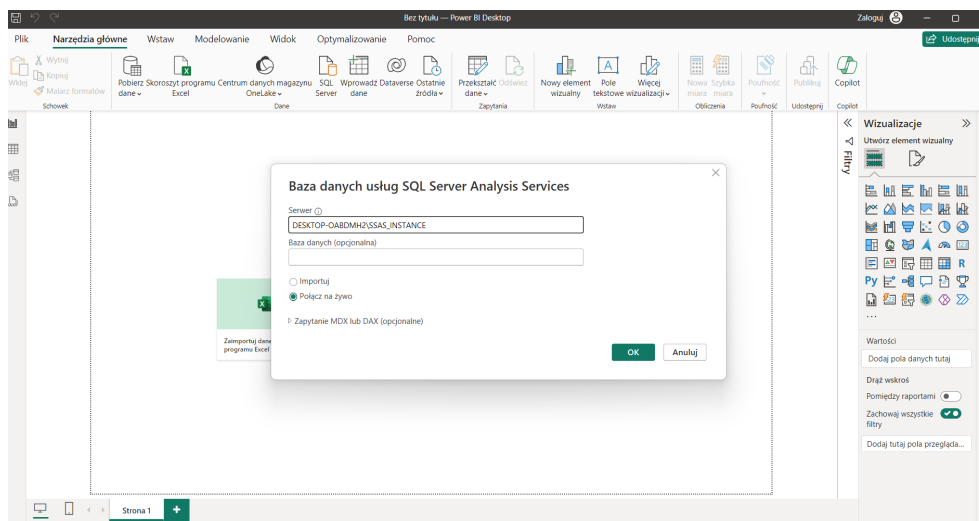
Finalnie utworzona kostka OLAP gotowa do analizy danych wygląda następująco.



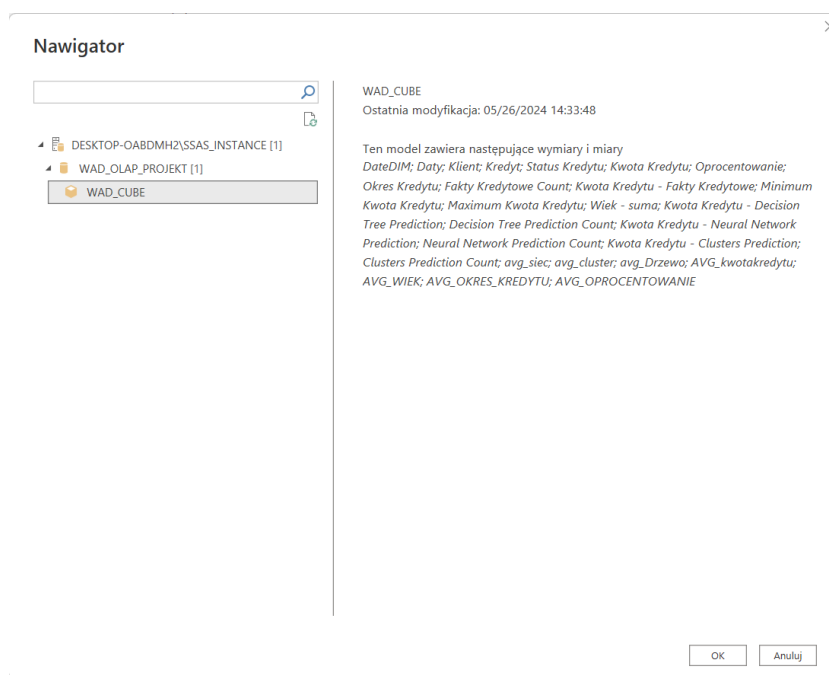
Rysunek 6.39: Finalna kostka OLAP

## 7. Wizualizacja danych - Power BI

W celu wizualizacji zebranych danych wykorzystane zostanie oprogramowanie typu open-source czyli Power BI Desktop. W tym celu na wstępie zostanie utworzone połączenie z usługą SSAS, gdzie zawarta jest utworzona kostka OLAP.



Rysunek 7.40: Podłączanie się do SSAS w oprogramowaniu Power BI

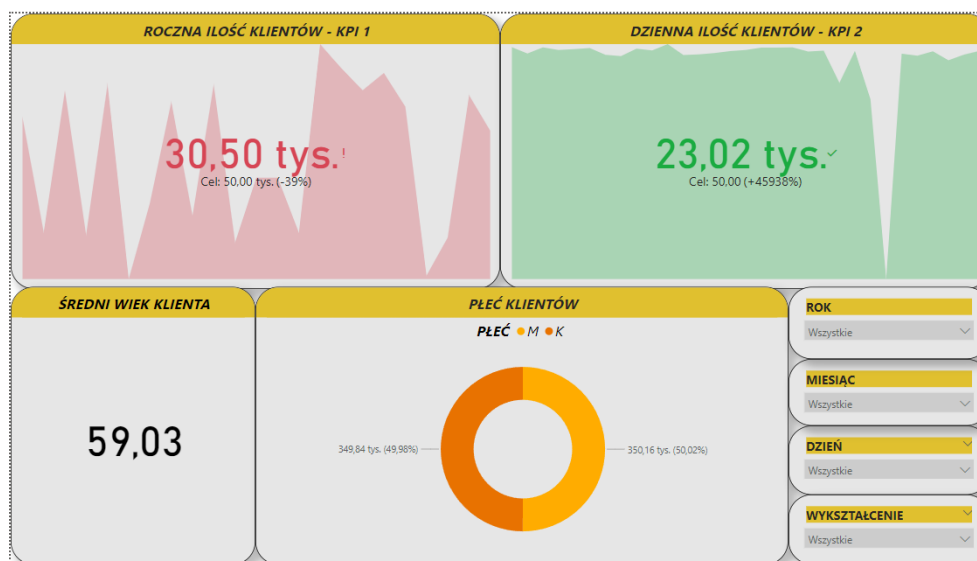


Rysunek 7.41: Wybór utworzonej kostki z SSAS

## 7.1. Wizualizacje

### KPI cz. 1 + informacje o klientach

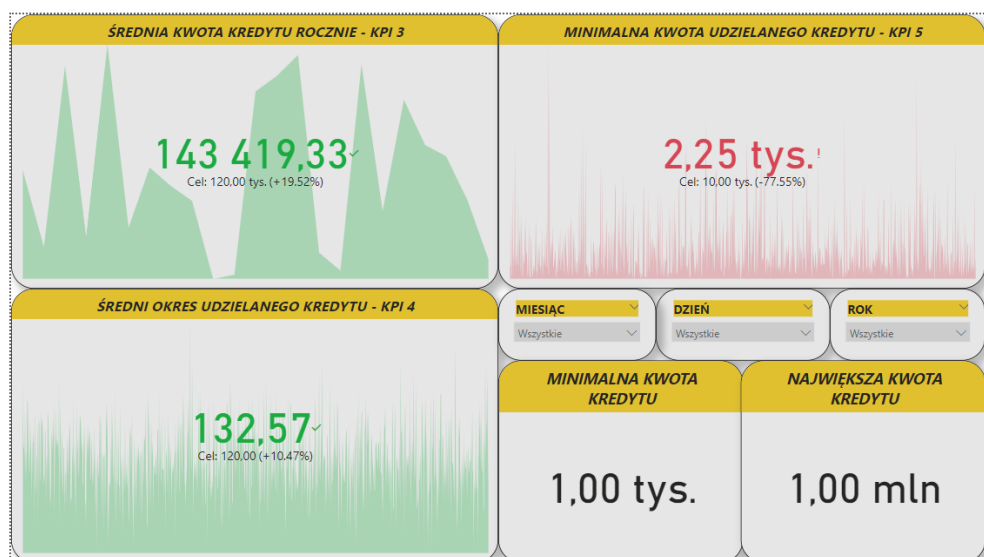
W tym raporcie utworzone zostały 2 wskaźniki (Scenariusz 1, Scenariusz 2) przedstawiające cel ilości klientów rocznie jak i dziennie. Dodatkowo w wartościach jak i na wykresach dodane zostały informacje mówiące o średnim wieku klientów a także o jakiej płci są klienci banku. Oprócz tego dodane zostały 4 fragmentatory pozwalające na filtrowanie odpowiednio przedstawionych danych. Niektóre fragmentatory nie wpływają na wskaźniki na przykład zmiana miesiąca czy dni na roczny KPI ilości klientów.



Rysunek 7.42: Raport 1

### KPI cz. 2 + informacje kredytowe

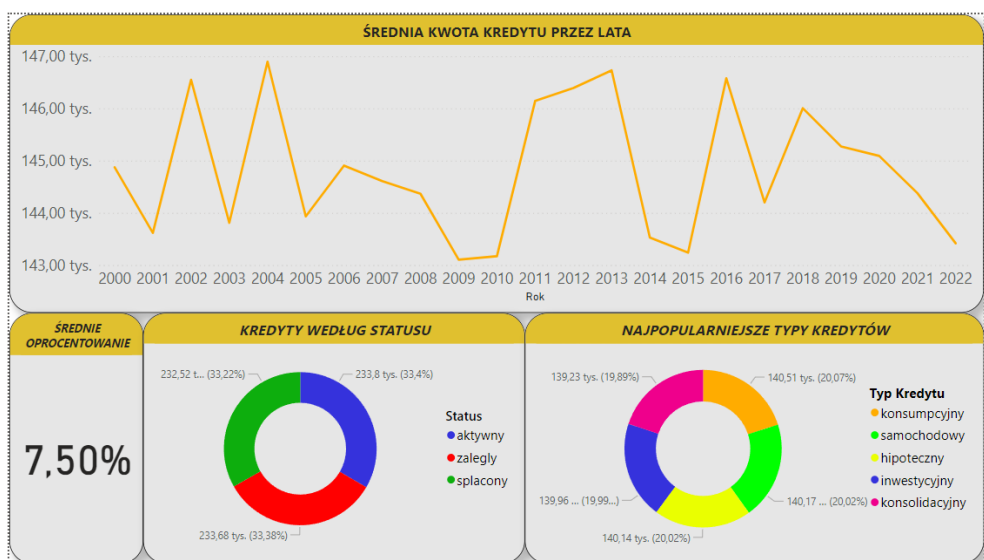
Poniższy raport przedstawia pozostałe 3 KPI (Scenariusz 3, Scenariusz 4, Scenariusz 5). Dotyczą one średniej kwoty kredytu rocznie, średniego okresu udzielanego kredytu w miesiącach a także minimalnej kwoty udzielanego kredytu. Dodatkowo umieszczone zostały wartości przedstawiające najmniejszą i największą kwotę udzielonego kredytu. Oprócz wizualizacji dodane zostały fragmentatory, dzięki którym można odpowiednio filtrować dane. Raport również został zabezpieczony by nie wpływały one na każdy rodzaj wizualizacji.



Rysunek 7.43: Raport 2

## Informacje kredytowe

Trzeci raport przedstawia wyłącznie informacje kredytowe. Zawiera on wykres liniowy przebiegu średniej kwoty kredytu w zakresie dat 01-01-2000 do 31-12-2022. Posiada również wartość prezentującą średnie oprocentowanie dla udzielonego kredytu, wykres kołowy przedstawiający procentowo ilość kredytów o danym statusie a także w ten sam sposób przedstawione informacje o najpopularniejszych typach kredytów.

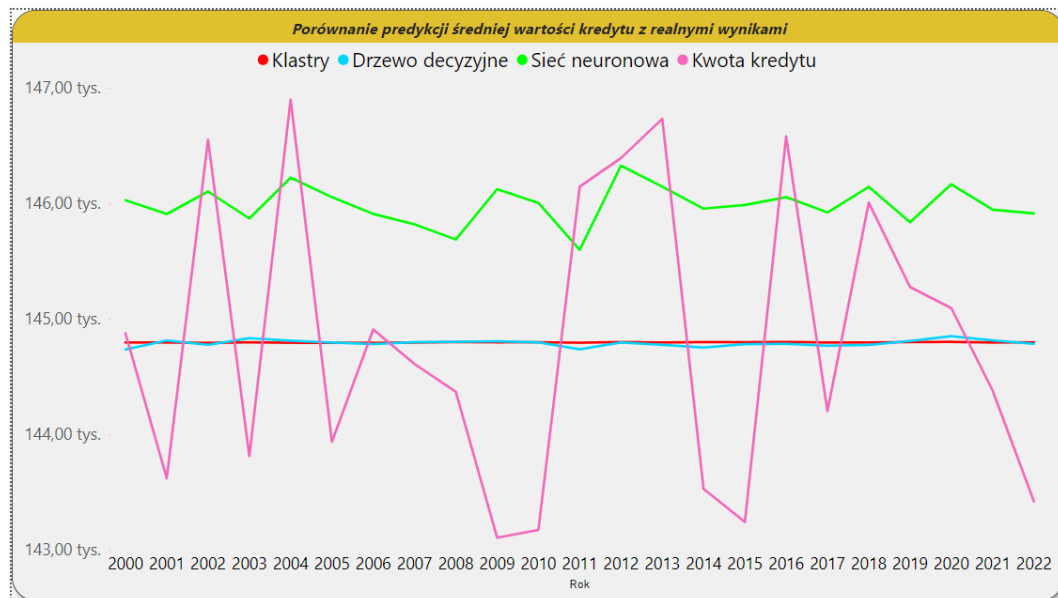


Rysunek 7.44: Raport 3



## Raport Data Mining

Ostatni czwarty przedstawia predykcje średniej kwoty kredytu według danych lat. Porównany odpowiednio został każdy z utworzonych modeli Data Miningu z rzeczywistą średnią.



Rysunek 7.45: Raport 4

## 8. Podsumowanie

Projekt ten obejmuje pełen cykl przetwarzania danych, od ich przygotowania przez proces ETL, po zaawansowaną analizę wielowymiarową i wizualizację wyników. Dzięki zastosowanym narzędziom i technikom, możliwe było przeprowadzenie kompleksowej analizy danych, co dostarczyło wartościowych wniosków. Przez realizację tego projektu, nabyte zostało praktyczne doświadczenie w pracy z dużymi zbiorami danych, wykorzystaniu narzędzi ETL, budowie hurtowni danych oraz tworzeniu zaawansowanych wizualizacji, co jest kluczowe w dzisiejszym świecie analizy danych.