

# **Sprawozdanie - Hurtownia na podstawie danych sklepu internetowego**

Hurtownie Danych  
Prowadzący: dr inż. Dariusz Gąsior

Bączyński Konrad, 211147  
Charewicz Jan, 212825

## Spis treści

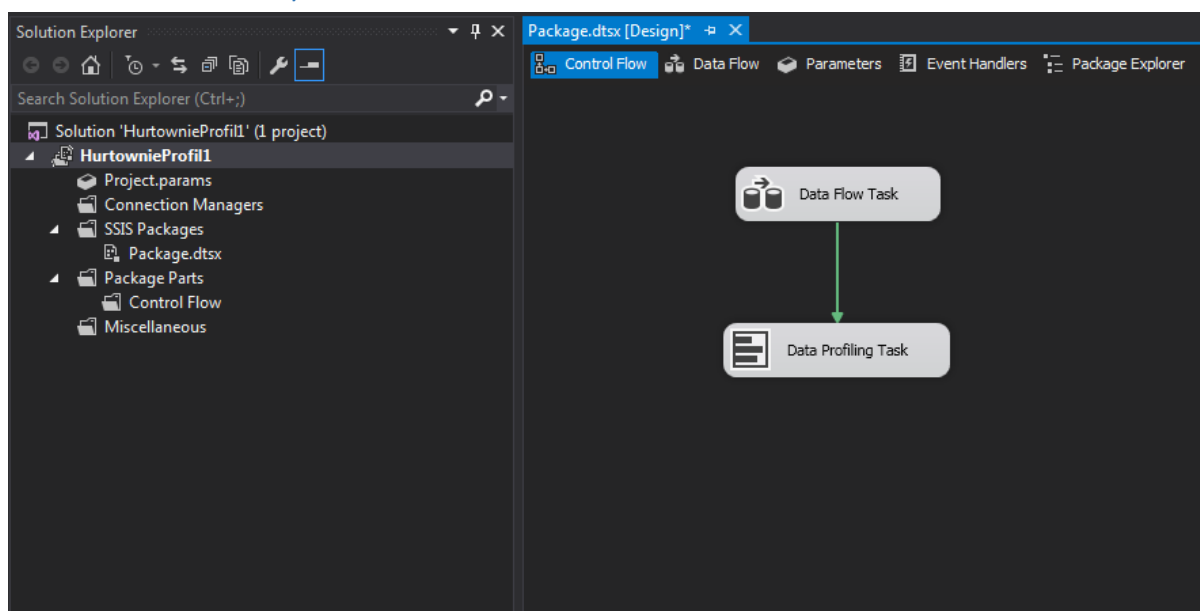
1. Wybór i opis źródeł .....	3
2. Jakość danych .....	4
3. Projekt bazy analitycznej.....	5
4. Proces ETL-owy ( <i>Extract, Transform and Load</i> ).....	6
5. Kostka.....	8
6. Raporty .....	9
• Raport 1 – Liczba klientów w perspektywie poszczególnych krajów i miesięcy.....	9
• Raport 2 – Całkowita miesięczna sprzedaż w roku 2004 z podziałem na miesiące. ....	10
• Raport 3 – Udział krajów w rozkładzie liczby sprzedanych produktów różnych kategorii.	10
7. Podsumowanie .....	11

## 1. Wybór i opis źródeł

Do naszego projektu, wybraliśmy znalezione w internecie źródło danych, będące sztandarową reprezentacją sklepu internetowego - w tym przypadku - Sklepu Dell'a na terenie Stanów Zjednoczonych. Znalezione plik .sql - na różnych forach, był po wielokrotnie podawany jako nawet dobry materiał na hurtownie danych. W pliku znaleziono zarówno tworzenie tabel, jak i liczne polecenia INSERT (...). Baza zawierała generowane (nierzadko o rozkładzie zbliżonym do rozkładu równomiernego) informacje na temat klientów, zamówień, pozycji na zamówieniach, produktach.

Motywację do wyboru źródła odnaleźliśmy w rekomendacjach tego źródła na forach, a także w bardzo przyjaznym formacie pliku. Niestety, jako, że język użyty w pliku, okazał się PostgreSQL'em, musieliśmy dokonać małych transformacji formatowania źródła, a by lepiej się pracowało w następnych etapach dane umieszczono (przy użyciu tak przeformatowanego i akceptowalnego przez MS pliku .sql) w bazie MS SQL SERVER i to ją wykorzystywaliśmy dalej jako źródło.

## 2. Jakość danych



Rysunek 1 Data Profiling Task w Visual Studio 2015.

W celu wykonania oceny jakości danych, tworzy się profile na podstawie danych źródłowych, całość eksportowana jest do pliku .xml które to parsuje Data Profile Viewer (rys. poniżej). Całe zadanie opiewało na odpowiednią konfigurację zadań i połączeń w Projekcie Analysis Services.

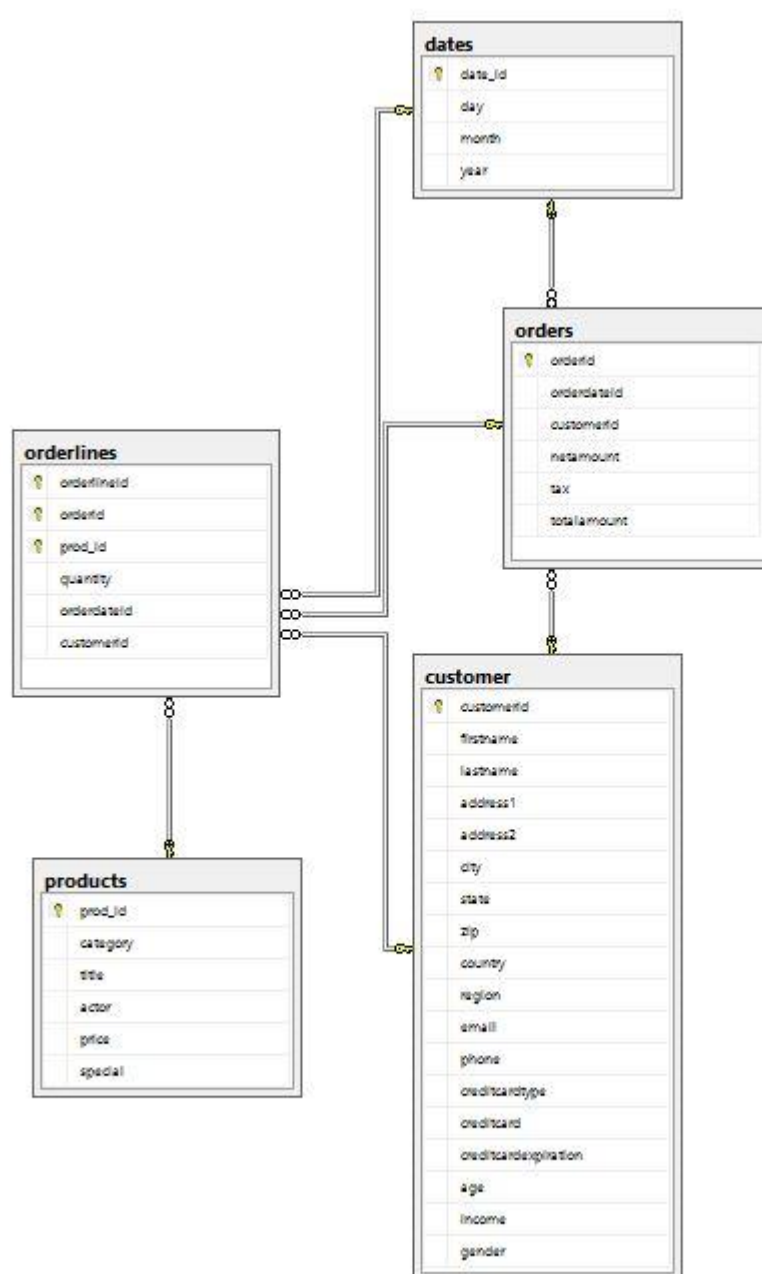
The screenshot shows the Data Profile Viewer application. The left pane displays a tree view of data sources and tables. The right pane shows a table of Column Null Ratio Profiles for the [dbo].[orders] table. The table has columns for Column, Null Count, and Null Percentage. The data shows zero nulls for all columns.

Column	Null Count	Null Percentage
customerid	0	0.0000 %
netamount	0	0.0000 %
orderdate	0	0.0000 %
orderid	0	0.0000 %
tax	0	0.0000 %
totalamount	0	0.0000 %

Rysunek 2 Data Profile Viewer

Jak pokazuje [Rysunek 1] źródło faktycznie okazało się bardzo w porządku, wskaźniki pól o wartości NULL w całej bazie wynosiły okrągłe 0. Z profilowania już można było się całkiem sporo dowiedzieć, np. to, że minimalny wiek klientów wynosił 18, pochodzili oni w około połowie z jednego kraju i częściej robili mniejsze zamówienia.

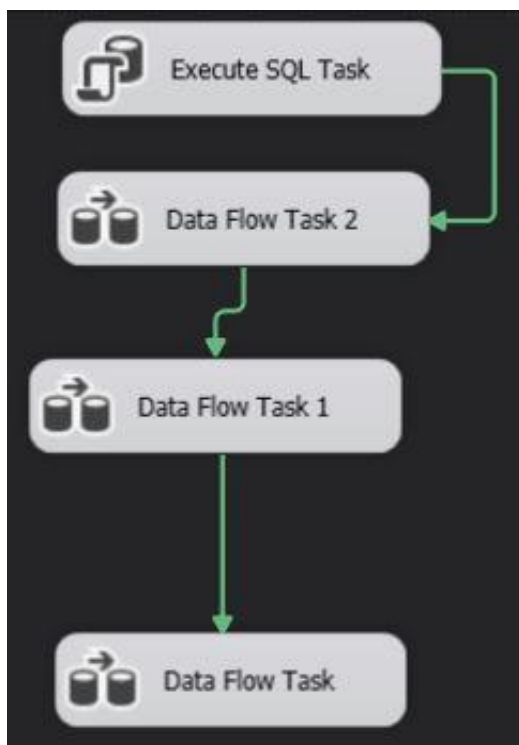
### 3. Projekt bazy analitycznej



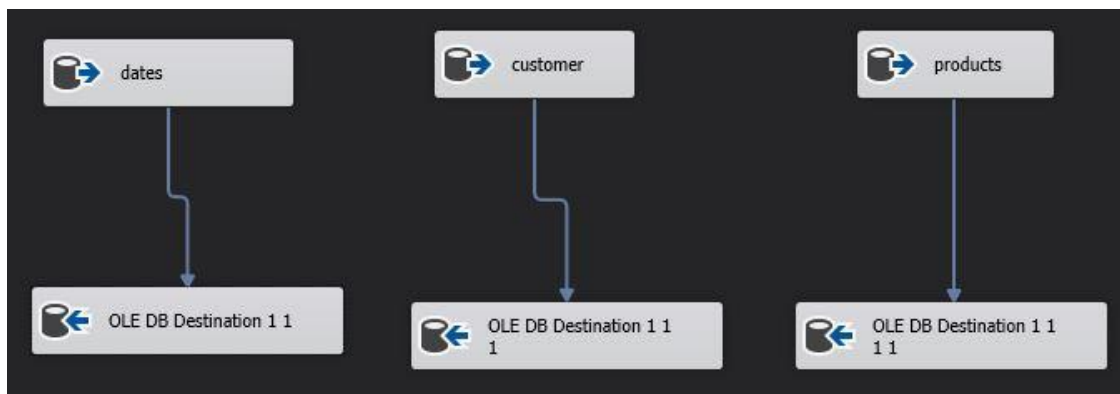
Rysunek 3 Diagram tabel bazy analitycznej.

Powyższy diagram [Rysunek 3] przedstawia zaproponowany po konsultacji z prowadzącym schemat docelowej bazy analitycznej. W niektórych miejscach należało więc zastosować spłaszczenia, by ze schematu z dwoma potencjalnymi tabelami faktów, skupić się na pozycjach na zamówieniu jako głównym wpisie w hurtowni.

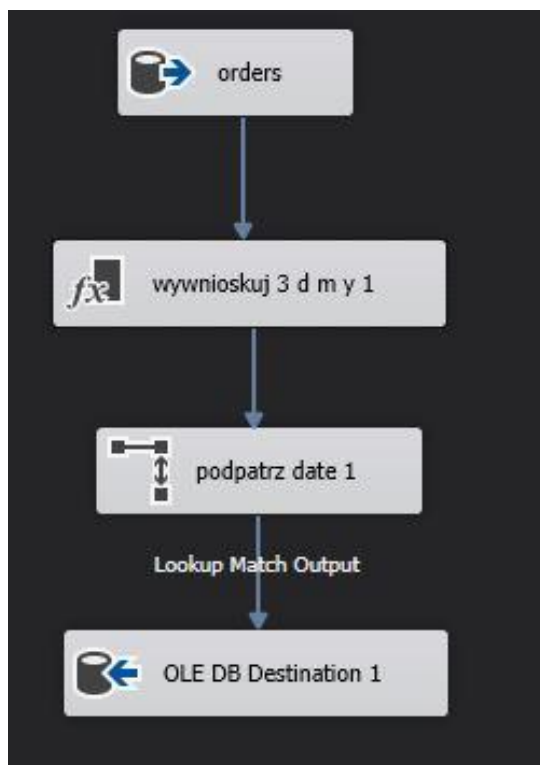
#### 4. Proces ETL-owy (*Extract, Transform and Load*)



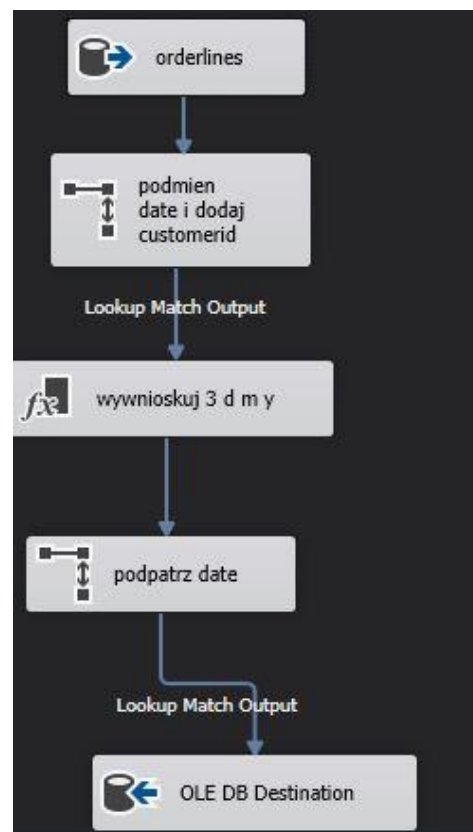
Rysunek 4 Control flow.



Rysunek 5 Data Flow Task 2.



Rysunek 6 Data Flow Task 1

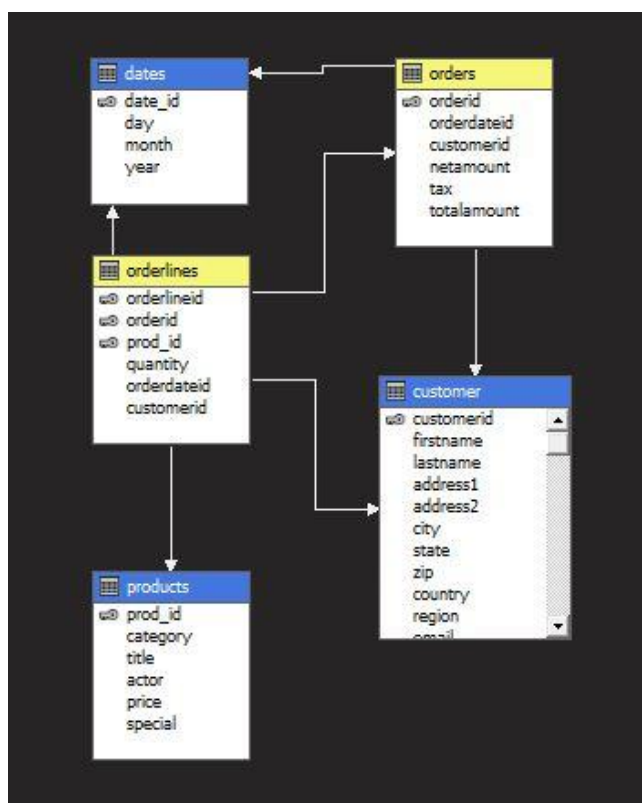


Rysunek 7 Data Flow Task.

Powyższe rysunki [4-7] przedstawiają realizację procesu ETL. Dzięki niemu uzyskujemy bazę wypełnioną przetworzonymi już częściowo danymi. Na przykład dodajemy dodatkową kolumnę do tabeli i wnioskujemy ją na podstawie innej kolumny w innej tabeli. Ważnym elementem tego procesu jest przygotowanie najpierw tabel pobocznych, potencjalnie wykorzystywanych w blockach typu *lookup*, a dopiero w następnym etapie procesowanie tabeli faktów.

## 5. Kostka

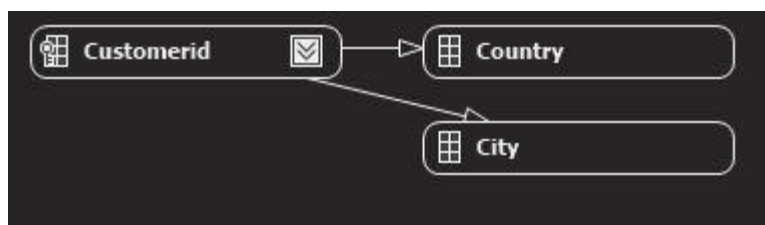
Zamieszczone obok rysunki przedstawiają proces tworzenia i konfiguracji kostki. [Rysunek 8] przedstawia strukturę kostki. Kostka pozwala na szybką analizę danych. [Rysunek 9] przedstawia wykorzystane wymiary do budowy kostki oraz dwie tabele faktów. [Rysunek 10] przedstawia wymiar klienta i relacje między jego atrybutami.



Rysunek 8 Struktura kostki.

Measure Groups		
Dimensions	Orderlines	Orders
Customer	Customerid	Customerid
Products	Prod Id	
Orders	Orderid	Orderid
Dates (Orderline D...	Date Id	Date Id
Dates (Order Data)	Orders	Orderline Data

Rysunek 9 Wykorzystane wymiary.



Rysunek 10 Relacje między atrybutami Klienta.

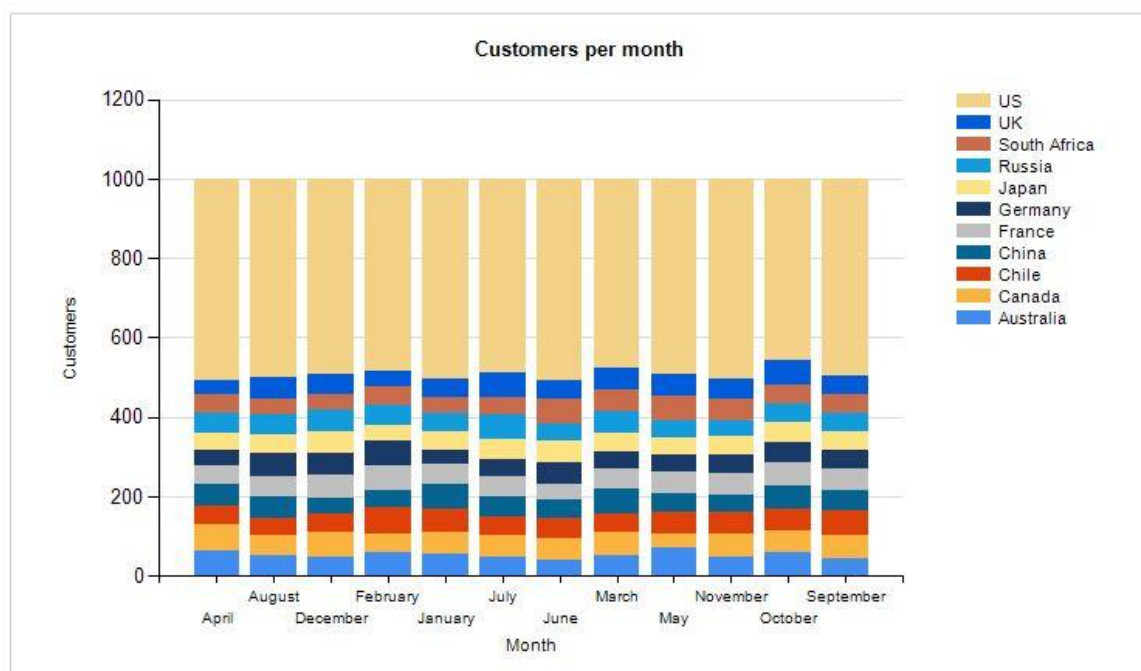


## 6. Raporty

- Raport 1 – Liczba klientów w perspektywie poszczególnych krajów i miesięcy.

Country	Month	Number of customers
Australia	January	56
	February	59
	March	53
	April	62
	May	69
	June	40
	July	48
	August	50
	September	45
	October	59
	November	49
	December	49
Canada	January	52
	February	47
	March	58
	April	67
	May	38
	June	54
	July	54
	August	50
	September	56
	October	54

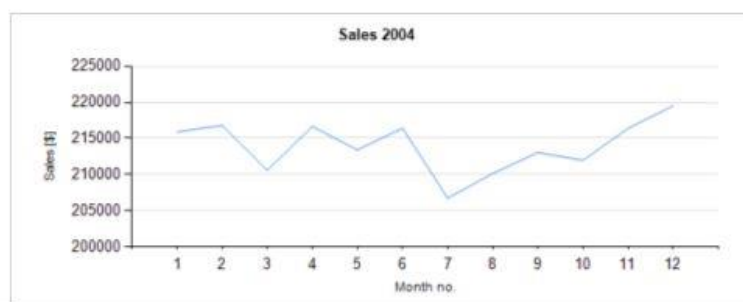
Tabela 1 Liczba klientów w poszczególnych krajach z podziałem na miesiące.



Rysunek 11 Liczba klientów w poszczególnych krajach z podziałem na miesiące.

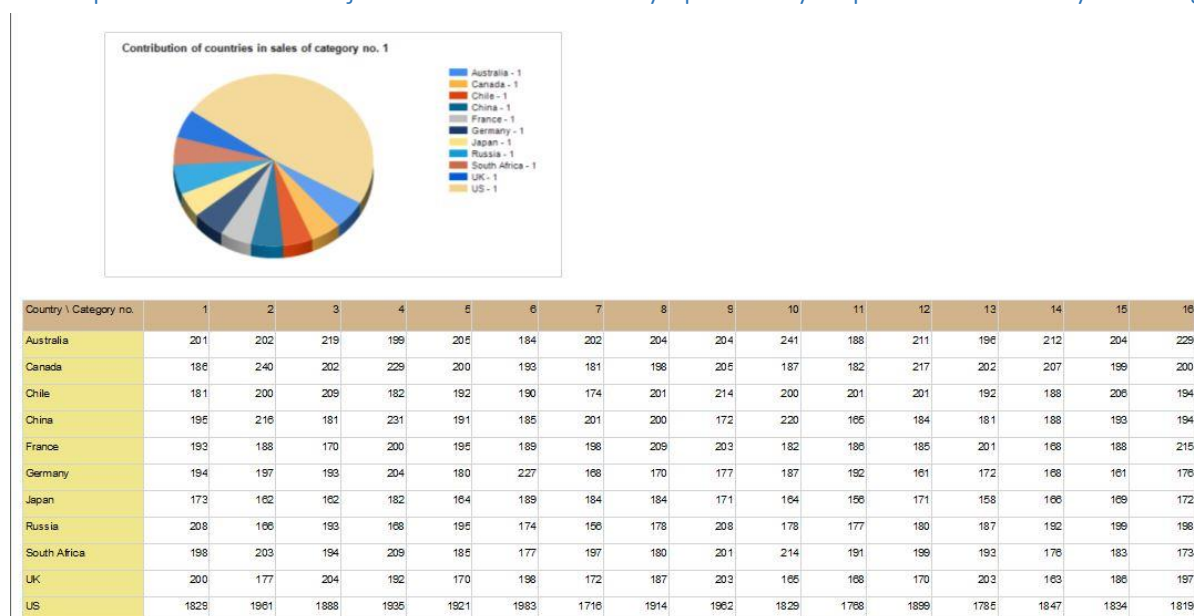
- Raport 2 – Całkowita miesięczna sprzedaż w roku 2004 z podziałem na miesiące.

Monthly sales		
No.	Month	Total sales
1	January	\$215898.76
2	February	\$216792.09
3	March	\$210564.40
4	April	\$216642.01
5	May	\$213395.19
6	June	\$216412.59
7	July	\$206707.38
8	August	\$210115.92
9	September	\$213034.71
10	October	\$211952.17
11	November	\$216373.17
12	December	\$219498.40



Rysunek 12 Całkowita miesięczna sprzedaż w poszczególnych miesiącach.

- Raport 3 – Udział krajów w rozkładzie liczby sprzedanych produktów różnych kategorii.



Rysunek 13 Liczba produktów sprzedanych w poszczególnych krajach w danych kategoriach.

Powyższe rysunki przedstawiają raporty wygenerowane dzięki analizie danych. Dzięki nim możemy zorientować się jak wielu klientów miał nasz sklep w każdym z krajów na przestrzeni całego roku. Możemy również dowiedzieć się jaka była całkowita miesięczna sprzedaż. Oraz jak rozkładała się sprzedaż w poszczególnych krajach z podziałem na kategorie.

## 7. Podsumowanie

Hurtownie danych to dość specyficzna baza danych, pozwalająca dokładniej przyjrzeć się danemu wycinkowi rzeczywistości. Wszystkie wymienione w sprawozdaniu etapy są niezbędne, by uzyskać zdolność wnioskowania. W przypadku wybranego źródła danych, już przy profilowaniu danych uzyskaliśmy obraz danych o dość równomiernym, poniekąd „sztucznym”, czy „nienaturalnym” rozkładzie. Nie obniżyło to wartości edukacyjnych wyniesionych z całego procesu budowania i pracy z hurtownią danych. Przygotowane raporty są, naszym zdaniem, bardzo trafne jeśli chodzi o tematykę wybranego źródła, gdyż przy rzeczywistych danych, można by mówić o tym jak ów sklep się rozwijał, czy można zauważyć progres w poziomie sprzedaży, scharakteryzować w którym kraju warto by pomyśleć nad otwarciem nowego oddziału, czy odpowiedzieć na pytanie produkty których kategorii jak się sprzedają (nawet w poszczególnych krajach), a znając ich charakterystykę szukać argumentacji z czego może to wynikać.