**HURTOWNIE DANYCH**

Laboratorium 2

Maciej Kopiński 254578

**Zad. 1. Ekstrakcja danych**

1. Utworzyć tabelę Sprzedaz zawierającą dane dotyczące liczby sprzedanych produktów przez pracowników firmy AdventureWorks w poszczególnych latach zgodnie z podanym schematem:

Sprzedaz(pracID, prodID, "Nazwa produktu", Rok, Liczba).

Wykonać to samo zapytanie przy użyciu polecenia pivot:

* 1. w kolejnych kolumnach wyświetlić lata;
  2. w kolejnych kolumnach wyświetlić 5 najlepszych produktów.

1. Utworzyć zestawienie, które dla poszczególnych miesięcy i lat przedstawi informację o liczbie różnych klientów. Przygotuj zapytanie z i bez użycia polecenia pivot.
2. Utworzyć zestawienie zawierające w wierszach imiona i nazwiska sprzedawców,   
   a w kolumnach kolejne lata. Wartością będzie liczba obsłużonych transakcji.

Wyświetlić tylko tych sprzedawców, którzy pracowali przez wszystkie 4 lata.

1. Zdefiniować zapytanie wyznaczające sumę kwot sprzedaży towarów oraz liczbę różnych produktów w zamówieniach w poszczególnych latach, miesiącach, dniach.
2. Wykorzystując polecenie CASE przygotować podsumowania do zestawienia z poprzedniego zadania tak, aby sumowane były kwoty zamówień oraz obliczana liczba różnych produktów dla poszczególnych miesięcy i dni tygodnia.
3. Przygotować zestawienie, w którym dla wybranych klientów przygotujemy kartę lojalnościową:
   1. srebrną, jeśli klient wykonał co najmniej 5 transakcji w sklepie;
   2. złotą, jeśli klient wykonał co najmniej 2 transakcje, której kwota przekraczała 150% średniej wartości zamówień w bazie (łączna kwota przekraczała 300% i każda z transakcji przekraczała 150%);
   3. platynową, jeśli klient co roku spełniał warunki otrzymania karty złotej.

Schemat wynikowej tabeli powinien wyglądać następująco:

KartyLojalnosciowe(Imie, Nazwisko, Liczba transakcji, Łączna kwota transakcji, Kolor karty)

**Zad. 2. Analiza danych**

1. Przedstaw wyniki zadania 1 w postaci tabel i wykresów przestawnych w programie MS Excel. Zinterpretuj wyniki.
2. Przygotuj 5 dodatkowych tabel/wykresów, które pokażą ciekawe zależności w bazie AdventureWorks przy użyciu narzędzia Power BI lub Tableau.

**Zad. 3. Ocena jakości danych – profilowanie danych**

1. Przeanalizować, scharakteryzować i ocenić dane znajdujące się w plikach dane\_1.txt oraz dane\_2.csv wykorzystując profilowanie danych pakietu Visual Studio: projekt *Integration Services Project*. Z menu bocznego *SSIS Toolbox* należy:
   1. wybrać bloczek *Data Profiling Task* (przeciągnąć na kanwę projektu),
   2. określić dane w sekcji *Destination*,
   3. skonfigurować *Quick Profile*,
   4. uruchomić pakiet (Run).
2. Aby obejrzeć wynik należy użyć *Data Profile Viewer* lub ponownie edytując bloczek *Data Profiling Task* użyć opcji *Open Profile Viewer*.

**Rozwiązania i wnioski:**

**Ad.1**

**1)**

WITH Sprzedaz AS

(

SELECT Sales.SalesPerson.BusinessEntityID, Sales.SalesOrderDetail.ProductID, Production.Product.Name, YEAR(Sales.SalesOrderHeader.OrderDate) AS Rok, SUM(Sales.SalesOrderDetail.OrderQty) AS Liczba

FROM Sales.SalesPerson

RIGHT JOIN Sales.SalesOrderHeader ON Sales.SalesOrderHeader.SalesPersonID = Sales.SalesPerson.BusinessEntityID

JOIN Sales.SalesOrderDetail ON Sales.SalesOrderDetail.SalesOrderID = Sales.SalesOrderHeader.SalesOrderID

JOIN Production.Product ON Production.Product.ProductID = Sales.SalesOrderDetail.ProductID

GROUP BY Sales.SalesPerson.BusinessEntityID, Sales.SalesOrderDetail.ProductID, Production.Product.Name, YEAR(Sales.SalesOrderHeader.OrderDate)

)

SELECT \*

FROM Sprzedaz;

**1.a)**

SELECT PracownikID, ProduktID, "Nazwa produktu", [2011], [2012], [2013], [2014]

FROM (

SELECT sp.BusinessEntityID AS PracownikID, sod.ProductID AS ProduktID, Production.Product.Name AS "Nazwa produktu", YEAR(soh.OrderDate) AS Rok, sod.OrderQty AS "Liczba produktów"

FROM Sales.SalesPerson sp

RIGHT JOIN Sales.SalesOrderHeader soh ON soh.SalesPersonID = sp.BusinessEntityID

JOIN Sales.SalesOrderDetail sod ON sod.SalesOrderID = soh.SalesOrderID

JOIN Production.Product ON Production.Product.ProductID = sod.ProductID

) AS sumamry

PIVOT(

SUM("Liczba produktów")

FOR Rok IN ([2011], [2012], [2013], [2014])

) AS P

ORDER BY 1;

**1.b)**

SELECT pracID, Rok, [712], [870], [711], [715], [708]

FROM(

SELECT ssp.BusinessEntityID AS pracID, pp.ProductID AS prodID,

YEAR(ssoh.OrderDate) AS Rok, ssod.OrderQty AS Liczba

FROM Sales.SalesPerson ssp

RIGHT JOIN Sales.SalesOrderHeader ssoh ON ssp.BusinessEntityID = ssoh.SalesPersonID

JOIN Sales.SalesOrderDetail ssod ON ssod.SalesOrderID = ssoh.SalesOrderID

JOIN Production.Product pp ON pp.ProductID = ssod.ProductID

WHERE pp.ProductID IN (SELECT TOP 5 ProductID FROM Sales.SalesOrderDetail GROUP BY ProductID ORDER BY SUM(OrderQty) DESC)

) x

PIVOT(

SUM(Liczba)

FOR

prodID in ([712], [870], [711], [715], [708])

) y

ORDER BY 1, 2;

**2)**

SELECT Rok, [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]

FROM (

SELECT DISTINCT YEAR(s.OrderDate) AS Rok, MONTH(s.OrderDate) AS Miesiac, s.CustomerId AS "Roczni klienci"

FROM Sales.SalesOrderHeader s) AS summary

PIVOT(

COUNT("Roczni klienci")

FOR Miesiac IN ([1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12])) AS P

ORDER BY 1;

**3)**

SELECT "Imie i nazwisko", [2011], [2012], [2013], [2014]

FROM (

SELECT p.FirstName + ' ' + p.LastName AS "Imie i nazwisko", YEAR(Sales.SalesOrderHeader.OrderDate) AS Rok, Sales.SalesOrderHeader.SalesOrderID AS "Liczba transakcji"

FROM Person.Person p

JOIN Sales.SalesPerson ON Sales.SalesPerson.BusinessEntityID = p.BusinessEntityID

JOIN Sales.SalesOrderHeader ON Sales.SalesOrderHeader.SalesPersonID = Sales.SalesPerson.BusinessEntityID) AS summary

PIVOT(

COUNT("Liczba transakcji")

FOR Rok IN ([2011], [2012], [2013], [2014])) AS P;

**4)**

SELECT YEAR(Sales.SalesOrderHeader.OrderDate) AS Rok, MONTH(Sales.SalesOrderHeader.OrderDate) AS Miesiąc, DAY(Sales.SalesOrderHeader.OrderDate) AS Dzień,

SUM(Sales.SalesOrderHeader.TotalDue) AS Suma, COUNT(DISTINCT Sales.SalesOrderDetail.ProductID) AS "Liczba różnych produktów"

FROM Sales.SalesOrderHeader

JOIN Sales.SalesOrderDetail ON Sales.SalesOrderDetail.SalesOrderID = Sales.SalesOrderHeader.SalesOrderID

GROUP BY YEAR(Sales.SalesOrderHeader.OrderDate), MONTH(Sales.SalesOrderHeader.OrderDate), DAY(Sales.SalesOrderHeader.OrderDate)

ORDER BY 1, 2, 3;

**5)**

SELECT CASE

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 1 THEN 'Styczeń'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 2 THEN 'Luty'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 3 THEN 'Marzec'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 4 THEN 'Kwiecień'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 5 THEN 'Maj'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 6 THEN 'Czerwiec'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 7 THEN 'Lipiec'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 8 THEN 'Sierpień'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 9 THEN 'Wrzesień'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 10 THEN 'Październik'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 11 THEN 'Listopad'

WHEN MONTH(Sales.SalesOrderHeader.OrderDate) = 12 THEN 'Grudzień'

ELSE NULL

END Miesiąc,

CASE

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 1 THEN 'Poniedziałek'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 2 THEN 'Wtorek'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 3 THEN 'Środa'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 4 THEN 'Czwartek'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 5 THEN 'Piątek'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 6 THEN 'Sobota'

WHEN DATEPART(dw, Sales.SalesOrderHeader.OrderDate) = 7 THEN 'Niedziela'

ELSE NULL

END Dzień,

SUM(Sales.SalesOrderHeader.TotalDue) AS Suma, COUNT(DISTINCT Sales.SalesOrderDetail.ProductID) AS "Liczba różnych produktów"

FROM Sales.SalesOrderHeader

JOIN Sales.SalesOrderDetail ON Sales.SalesOrderDetail.SalesOrderID = Sales.SalesOrderHeader.SalesOrderID

GROUP BY MONTH(Sales.SalesOrderHeader.OrderDate), DATEPART(dw, Sales.SalesOrderHeader.OrderDate) ORDER BY MONTH(Sales.SalesOrderHeader.OrderDate),

DATEPART(dw, Sales.SalesOrderHeader.OrderDate);

**6)**

SELECT COUNT(\*) FROM (

SELECT p.FirstName, p.Lastname, COUNT(ssoh.SalesOrderID) AS "Liczba transakcji", SUM(ssoh.TotalDue) AS "Łączna kowata transakcji",

CASE

WHEN (SELECT COUNT(soh.TotalDue)

FROM Sales.Customer c

JOIN Sales.SalesOrderHeader soh ON soh.CustomerID = c.CustomerID

WHERE soh.TotalDue > 1.5\*(SELECT AVG(soh.TotalDue) FROM Sales.SalesOrderHeader soh) AND soh.CustomerID = cus.CustomerID) >= 2 THEN (

CASE

WHEN (SELECT COUNT(DISTINCT Y)

FROM (

SELECT CustomerID, YEAR(sohh.OrderDate) Y, COUNT(sohh.TotalDue) C

FROM Sales.SalesOrderHeader sohh

WHERE sohh.TotalDue > 1.5\*(SELECT AVG(soh.TotalDue) FROM Sales.SalesOrderHeader soh)

GROUP BY CustomerID, YEAR(sohh.OrderDate)

HAVING COUNT(sohh.TotalDue) >= 2

) B ) >= (SELECT COUNT(DISTINCT soh.OrderDate) FROM Sales.SalesOrderHeader soh WHERE soh.CustomerID = cus.CustomerID) THEN 'Platynowa'

ELSE 'Złota'

END

)

WHEN COUNT(ssoh.SalesOrderID) >= 5 THEN 'Srebrna'

END Karta

FROM Person.Person P

JOIN Sales.Customer cus ON cus.PersonID = p.BusinessEntityID

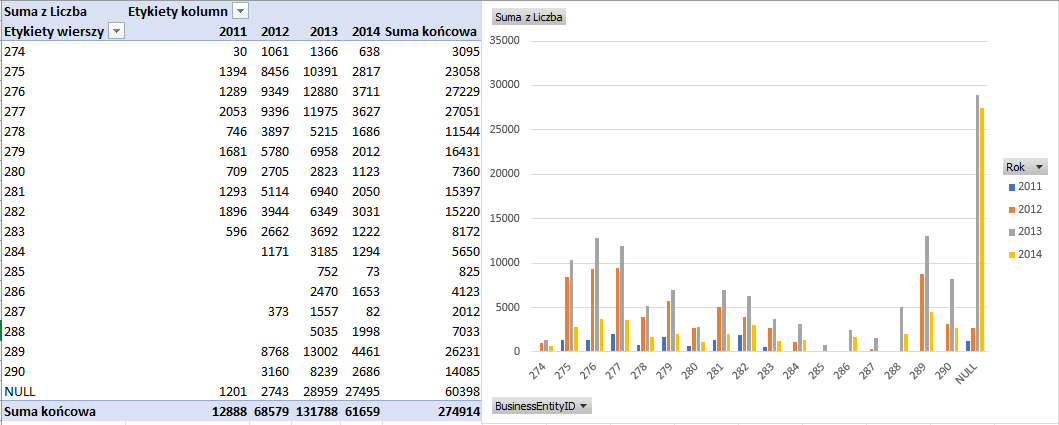
JOIN Sales.SalesOrderHeader ssoh ON ssoh.CustomerID = cus.CustomerID

GROUP BY p.FirstName, p.LastName, cus.CustomerID, YEAR(ssoh.OrderDate)) A

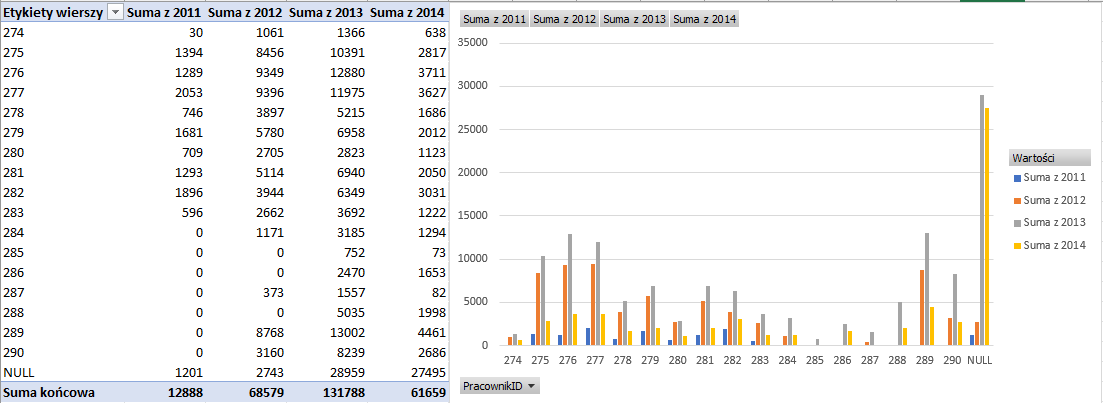
WHERE A.Karta = 'Platynowa';

**Ad.2 – Wnioski do wykresów powstałych z wyników kwerend z zadania 1**

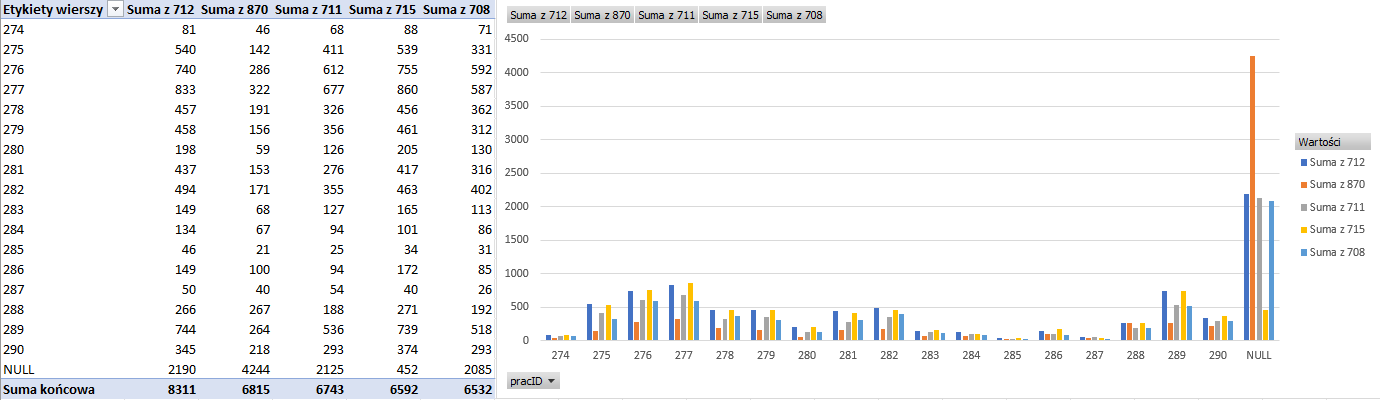
**1)**

****

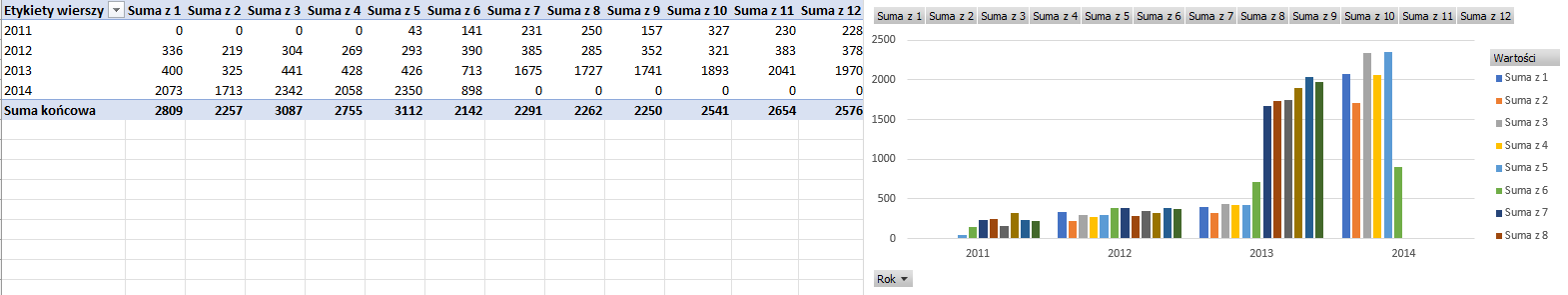
**1a)**

****

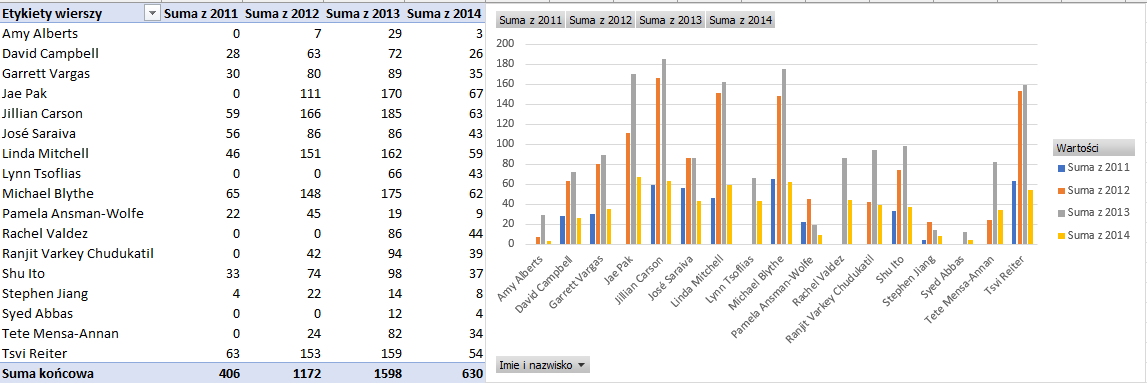
**1b)**

****

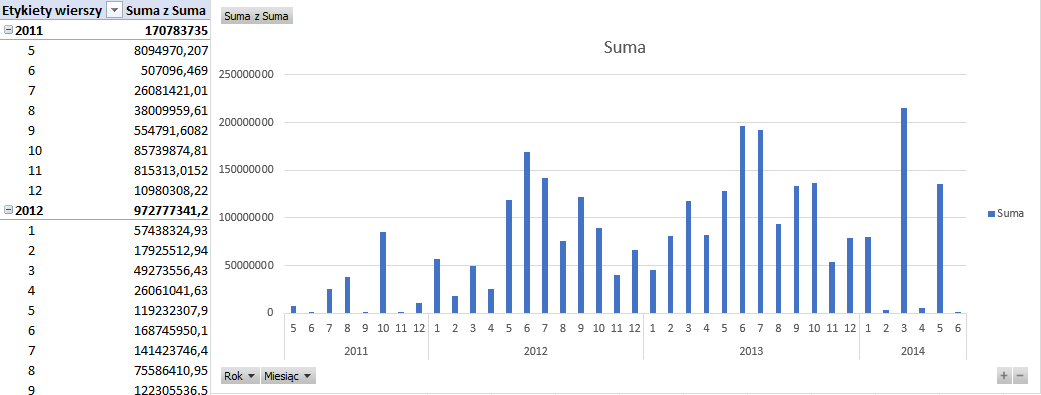
**2)**

****

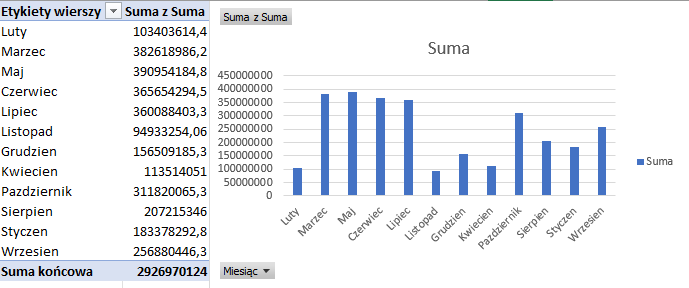
**3)**

****

**4)**

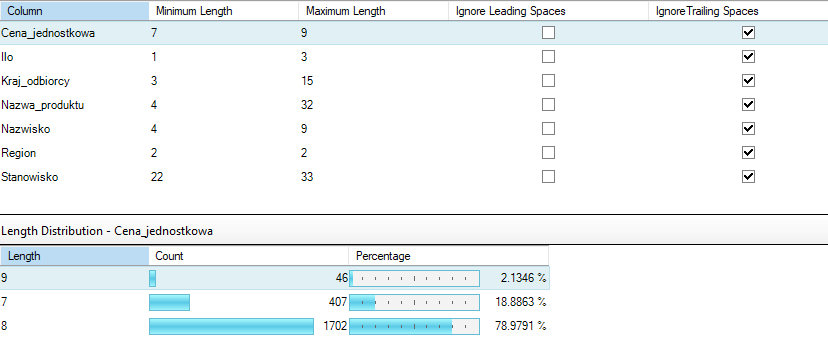
****

**5)**

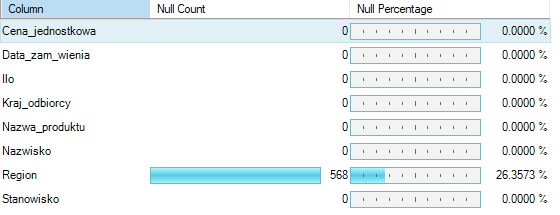
****

**Ad.3**

Zrzuty ekranu oraz wnioski z wyników dotyczących pliku dane\_1.txt:



Dane z pliku dane\_1.txt odnoszą się najprawdopodobniej do jakiegoś rodzaju restauracji. Jeśli chodzi o powyższe wyniki, to najciekawsza jest różnica w długości ceny jednostkowej. Zamówienia zaczynają się od kilku do nawet kilkuset złotych.



Wiele zamówień nie posiada określonego regionu (prawie 30%).

Obraz zawierający tekst

Opis wygenerowany automatycznie

Restauracja rozpoczęła rejestrowanie zamówień w bazie danych w roku 1996 i działała przez prawie 2 lata.

Obraz zawierający stół

Opis wygenerowany automatycznie

Z powyższego zrzutu ekranu możemy uzyskać najwięcej informacji – przykładowo, że dostawy odbywały się do aż 21 różnych krajów, lub że restauracja oferowała aż 77 różnych produktów. Inną ciekawą informacją jest również to, że w ostatnim zarejestrowanym dniu w bazie wystąpiło najwięcej zamówień.

Zrzuty ekranu z wyników dotyczących pliku dane\_2.csv:

Obraz zawierający stół

Opis wygenerowany automatycznie

Dane z pliku dane\_2.txt odnoszą się do pewnego rodzaju bazy pomiarów składu powietrza w Londynie. Widzimy, że wszystkie z pomiarów są bardzo dokładne.

Obraz zawierający stół

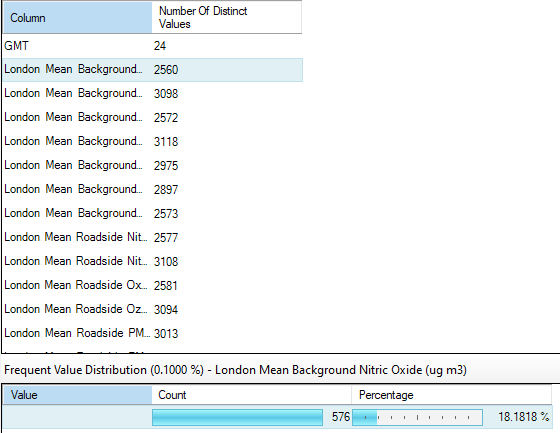
Opis wygenerowany automatycznie

Wiele pomiarów nie było wykonywane od początku istnienia bazy. Przeważająca większość ma również wartości o zbliżonych wartościach (długość 18 znaków). Nie licząc długości 0, można przypuszczać, że pomiary o niewielkich wartościach wynikają z błędów urządzeń pomiarowych.

Obraz zawierający stół

Opis wygenerowany automatycznie

Nie ma pól o pustych wartościach.



Powyższy zrzut ekranu potwierdza, że większość z wartości pomiarowych znacząco różni się od siebie (duża dokładność pomiarowa).

**Wnioski:**

**Wnioski do zadań 1 oraz 2 (Wnioski do zadania nr 3 znajdują się pod zrzutami ekranu zamieszczonymi wyżej):**

1. **oraz 1a)**

Dwa powyższe zrzuty ekranu prezentują to samo zestawienie, na którym warto zauważyć, że w 2013 roku następuje gwałtowny wzrost transakcji bez określonego sprzedawcy – sprzedaż internetowa.

**1b)**

Widzimy, że najpopularniejsze produkty były kupowane w dużych ilościach bez pomocy sprzedawców.

**2)**

Widać tutaj dwukrotny wzrost klientów w połowie 2013 roku. Biorąc pod uwagę poprzednie wykresy można przypuszczać, że właśnie wtedy do sklepu została wprowadzona sprzedaż internetowa. Po jej wprowadzaniu sprzedaż utrzymywała się na podobnym, wysokim poziomie.

**3)**

U prawie wszystkich sprzedawców możemy zaobserwować wzrost sprzedanych produktów wraz z biegiem lat. W przypadku 2014 roku trzeba brać pod uwagę, że dysponujemy danymi z jedynie połowy roku.

**4)**

Z roku na rok widać stopniowy wzrost łącznej kwoty zamówień. Na wykresie nie zamieszczałem liczby produktów, ponieważ jest ona nieporównywalnie mniejsza od sumy kwot zamówień.

**5)**

Na powyższym wykresie widać, że łączna kwota zamówień osiąga najwyższe wartości w ciepłych miesiącach, a najniższe – w najchłodniejszych.

**Zadanie z zajęć:**

SELECT Dzień, [S], [M], [L]

FROM (

SELECT DISTINCT DATEPART(dw, soh.OrderDate) AS Dzień,

CASE

WHEN pro.Size = '38' THEN 'S'

WHEN pro.Size = '40' THEN 'M'

WHEN pro.Size = '46' THEN 'L' END Rozmiar, soh.CustomerID AS "Liczba różnych klientów"

FROM Sales.SalesOrderHeader soh

JOIN Sales.SalesOrderDetail sod ON sod.SalesOrderID = soh.SalesOrderID

JOIN Production.Product pro ON pro.ProductID = sod.ProductID

) AS summary

PIVOT(

COUNT("Liczba różnych klientów")

FOR Rozmiar IN ([S], [M], [L])) AS P

ORDER BY 1;

**Obraz zawierający stół

Opis wygenerowany automatycznie**

**Wnioski:**

Widać, że klienci chętniej kupują przedmioty o mniejszych rozmiarach, lecz dzień tygodnia nie wpływa na ilość tych zakupów.