

Hurtownie danych

Lista 3 – Podstawy SQL: funkcje grupujące i okienkowe



Aleksander stepaniuk

nr. indeksu: 272644

Politechnika Wrocławska, Informatyka Stosowana

**Rozwiązania:**

Zadanie 1.

1. Kolejno: Rollup, Cube, Grouping sets

A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

2. *Przygotować zestawienie przedstawiające łączną kwotę zniżek z podziałem na kategorie…*

A screenshot of a computer

AI-generated content may be incorrect.A screen shot of a computer code

AI-generated content may be incorrect.

Zad2.

1. *Dla kategorii ‘Bikes’ przygotuj zestawienie prezentujące procentowy udział kwot*…

a) Bikes

A screenshot of a computer

AI-generated content may be incorrect.

b) Components

A screenshot of a computer

AI-generated content may be incorrect.

c) Clothing

A screenshot of a computer

AI-generated content may be incorrect.

d) Accessories

A screenshot of a computer

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

2. *Przygotuj zestawienie dla sprzedawców z podziałem na lata i miesiące prezentujące liczbę obsłużonych przez nich zamówień w ciągu roku, narastająco, sumarycznie w poprzecnim i obecnym miesiącu. (Wykorzystaj funkcje okna)*

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

3. *Przygotuj ranking klientów w zależności od liczby zakupionych produktów. Porównaj rozwiązania uzyskane przez funkcje rank i dense\_rank…*

A screenshot of a computer

AI-generated content may be incorrect.A screen shot of a computer

AI-generated content may be incorrect.

4. *Przygotuj ranking produktów w zależności od średniej liczby sprzedanych sztuk. Wyróżnij 3 (prawie równoliczne)* grupy produktów…

A screenshot of a computer screen

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

Zad3.

*Ocena jakości danych (profilowanie danych) zostało wykonanie w jęzuku Python za pomocą bibliotek pandas i ydata\_profiling.*

A screen shot of a computer program

AI-generated content may be incorrect.

Wszystkie rekordy mają id transakcji, ale tylko blisko 2/3 posiada dane w kolumnie Location.

A screenshot of a computer

AI-generated content may be incorrect.

Ilość wystąpień słów ‘Error’ i ‘Unknown’ w różnych kolumnach wydaje się być zbliżona (około 300-600 rekordów na 10002 czyli jakieś 3%-6%) , tylko kolumna ID nie posiada takich rekordów.

A screenshot of a computer error

AI-generated content may be incorrect.

Najwięcej brakujących wartości (pustych pól) widać w kolumnach metody płatności i lokalizacji. Jest ich tam o cały rząd wielkości więcej niż w pozostałych kolumnach.

A screenshot of a computer screen

AI-generated content may be incorrect.

Wygenerowany raport za pomocą ydata\_profiling:A screenshot of a computer

AI-generated content may be incorrect.

Możemy na tej stronie html interaktywnie przeglądać nasze dane, tu kilka najważniejszych wykresów:

ID:

A screenshot of a computer

AI-generated content may be incorrect.

Widać że ID nie jest unikalne, i niektóre pojedyncze rekordy występują więcej niż jeden raz.

Kolumna Item:

A white and grey horizontal lines

AI-generated content may be incorrect.

Widać że ilość produktów jest podobna w poszczególnych kategoriach, z wyjątkiem brakujących danych, które to kategorią nie są, ale widać że jest to łącznie około 1000 rekordów (~10%), które nie mają podanej wartości w ‘Item’.

Kolumna Quantity:

A white and grey striped background

AI-generated content may be incorrect.

Widać, że ilość oscyluje w granicach 1-5, pojedyncze wartości -2 i 100 które prawdopodobnie zostały wprowadzone do danych ręcznie lub są wynikiem błędu, więc można je kategoryzować razem z resztą błędów.

Kolumna Price per Unit:

A white and grey rectangular object

AI-generated content may be incorrect.

Widać że jest prawie dwa razy więcej wartości 3.0 i 4.0 niż pozostałych wartości takich jak 2.0, 5.0, 1.0, 1.5. Pozostałe wartości typu ERROR, UNKNOWN i (Missing) czyli puste pole stanowią zdecydowaną mniejszość.

Kolumna Payment method:

A white and grey horizontal lines

AI-generated content may be incorrect.

Widać że większość danych to Digital Wallet, Credit Card lub Cash, jest też standardowo trochę pustych rekordów z wartościami ERROR lub UNKNOWN ale również pojedyncze literówki w Digital Walle (brakujące ‘t’ na końcu) oraz CreditCard pisane razem. Widać też że ponad 25% rekordów jest brakujących (puste pola)

Mapa korelacji:

A screenshot of a computer

AI-generated content may be incorrect.

Widać największą korelację pomiędzy Price per unit, a Item oraz Total Spent i polami Item, Price Per Unit, Quantity. Reszta pól jest bez żadnej korelacji.

Duplikaty:

A screenshot of a computer

AI-generated content may be incorrect.

Widać także że jeden rekord występuje aż 3 razy w dokłanie tej samej formie.

**Wnioski:**

Wykorzystanie funkcji grupujących (rollup, cube, grouping sets) okazało się przydatne do uzyskania dodatkowych informacji o danych, takich jak sumy całkowite, sumy częściowe czy rankingu i podsumowania na różnych poziomach szczegółów w relatywnie prosty sposób. Wyniki takiego zapytania są właściwie gotowe do użycia w raportach i analizach.

Największe zniżki dotyczą produktów z kategorii 'Bikes'. Łączna kwota zniżek dla tej kategorii wynosi ponad 0,5 miliona. Widać, że niektóre produkty nigdy nie potrzebowały zniżki, co może sugerować ich wysoką jakość lub popularność wśród klientów. Inne produkty były przeceniane tylko w niektórych latach lub miesiącach, co może być związane z sezonowością sprzedaży lub innymi czynnikami rynkowymi.

Najwięcej zniżek zostało udzielonych w 2012 i 2013 roku (około 200tys w 2012 i 300tys w 2013) - może to sugerować, że w tych latach sklep dynamicznie się rozwijał i wprowadzał nowe promocje, aby przyciągnąć klientów (pik w sprzedaży 2013 rok). We wszystkich kategoriach widać, że najwięcej promocji zostało nałożone w 2013 roku, jednak dla rowerów i komponentów 2012 rok jest podobnie obfity w zniżki, co sugeruje, że sklep najpierw nakładał zniżki na rowery i komponenty a z czasem, rozszerzył to również o pozostałe kategorie.

Najlepszym pracownikiem był Jillian Carson, który w ciągu 2013 roku obsłużył 185 transakcji (w każdym miesiącu regularnie radził sobie dobrze, najgorszym był Amy Alberts, który obsłużył zaledwie 3 transakcje.

Najlepszym klientem był Reuben D'Sa który zakupił aż 2737 produktów. Najgorszych klientów było ponad 3500, czyli tych, którzy kupili tylko jeden produkt.

Najlepszym produktem był Full-Finger Gloves L (9 sztuk średnio sprzedanych). Na drugim miejscu ten sam produkt w rozmiarze M.

Pandas Profiling jest doskonałym narzędziem, aby automatycznie wygenerować ciekawe statystyki dotyczące zbioru danych. Udało się z łatwością zobaczyć ile jest unikalnych wartości, brakujących rekordów i inne ciekawe statystyki.