3. sprawozdanie z laboratorium Hurtownie Danych

Mikołaj Kubś, 272662

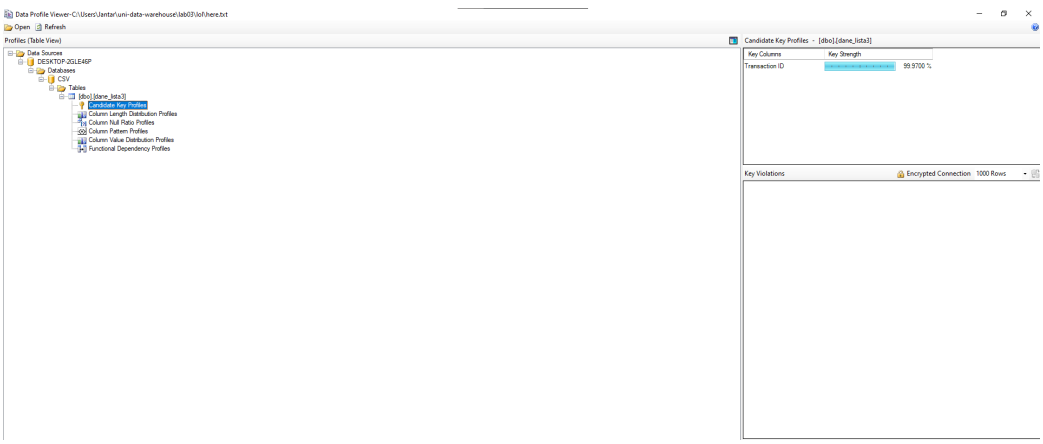
25 marca 2025

1 Zadanie 1 - funkcje grupujące

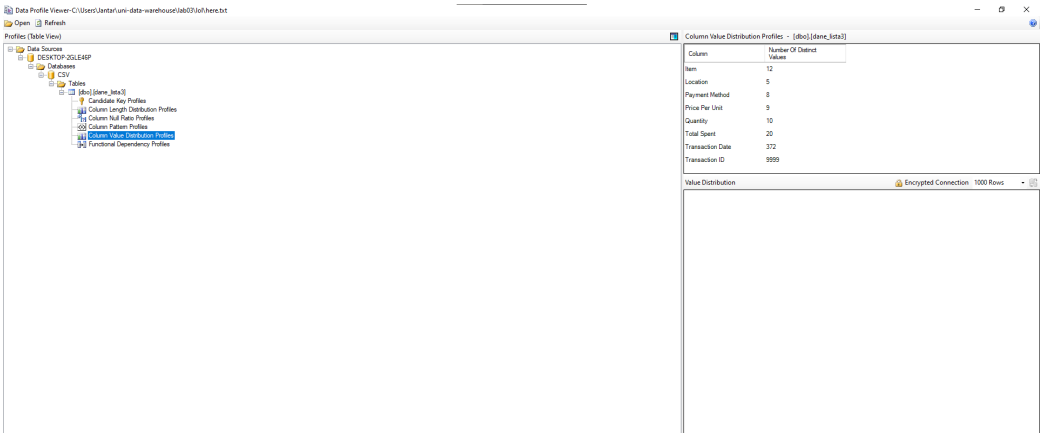
2 Zadanie 2 - funkcje okienkowe

3 Zadanie 3 - profilowanie danych

Po próbach analizy w SSIS uzyskano tylko część rezultatów.



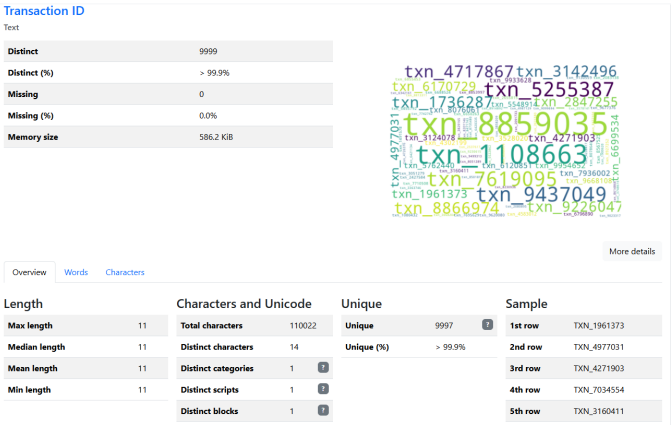
Rysunek 1: Profil kolumny kandydującej



Rysunek 2: Liczba unikatowych wartości w kolumnach

Z powodu problemów technicznych resztę analizy przeprowadzono używając *Pythona* oraz bibliotek *Pandas* i *ydata_profiling*.

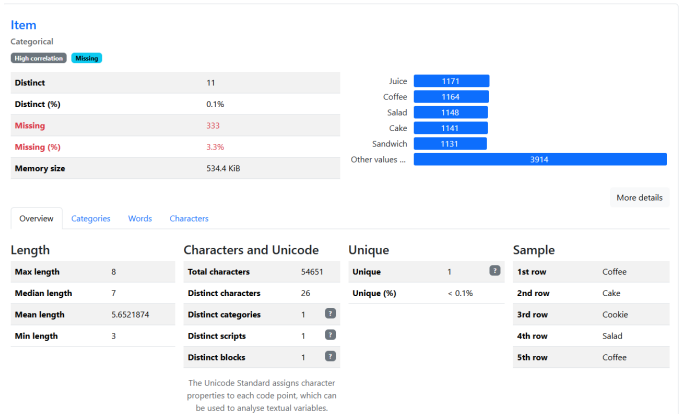
3.1 Transaction ID



Rysunek 3: Profil kolumny *Transaction ID*

Tak jak też wywnioskowano wcześniej, jest to jedyna kolumna nadająca się na klucz kandydujący. Są w niej tylko 3 duplikaty. Zawsze ma też tą samą długość - 11 znaków.

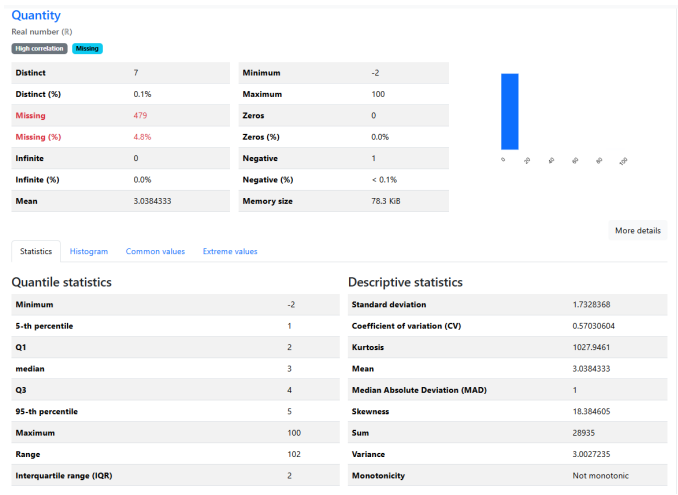
3.2 Item



Rysunek 4: Profil kolumny *Item*

Kolumna jest kategoryczna i przypisuje transakcjom kategorię kupionego towaru. W 3.3% wierszy brakuje wartości. Dodatkowo w 3.4% to *UNKNOWN*.

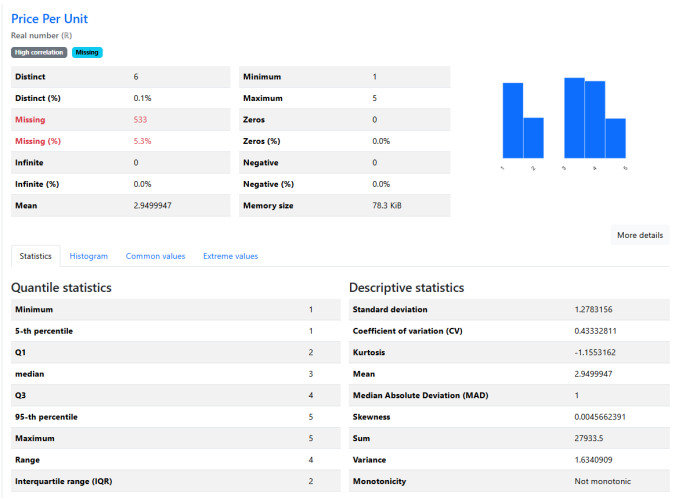
3.3 Quantity



Rysunek 5: Profil kolumny *Quantity*

W kolumnie występują anomalnie. Zdarzyła się raz wartość ujemna: -2. Zdarzyła się raz wartość 100, znacznie przekraczająca poza zakres innych (1-5), ale być może dozwolona. Brakuje danych w 4.8% wierszy. Wartości standardowe, czyli od 1 do 5 włącznie, występują z podobną częstością.

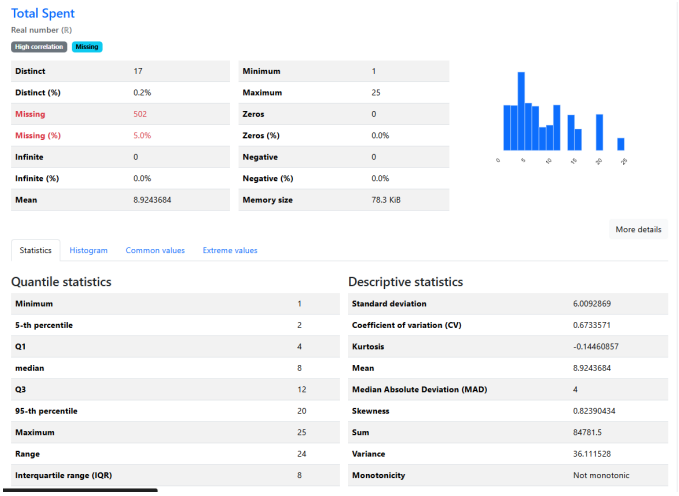
3.4 Price Per Unit



Rysunek 6: Profil kolumny *Price Per Unit*

Kolumna w większości składa się z liczb całkowitych, ale 11.3% wartości to jedyne z wartością po przecinku. Wszystkie wynoszą dokładnie 1,5. Brakuje wartości dla 5.3%. Standardowe i jedyne poza brakującymi wartości to liczby całkowite od 1-5 oraz 1,5. Najczęściej występuje wartość 3.

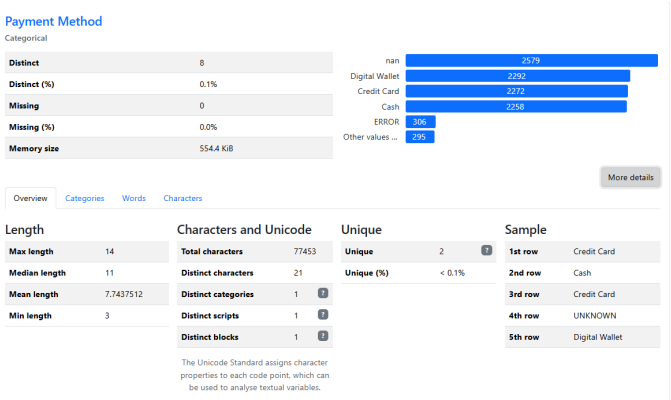
3.5 Total Spent



Rysunek 7: Profil kolumny *Total Spent*

Brakuje 5.02% wartości. Przedział to liczby od 1 do 25 (ale tylko 17 z nich jest w danych). Większość wartości to liczby całkowite, ale występują też wielokrotności 1,5, sugerując, że cena 1,5 w Price Per Unit nie jest żadną anomalią. Najczęściej występuje wartość 6. Większość wartości jest bliżej początku rozkładu niż końca.

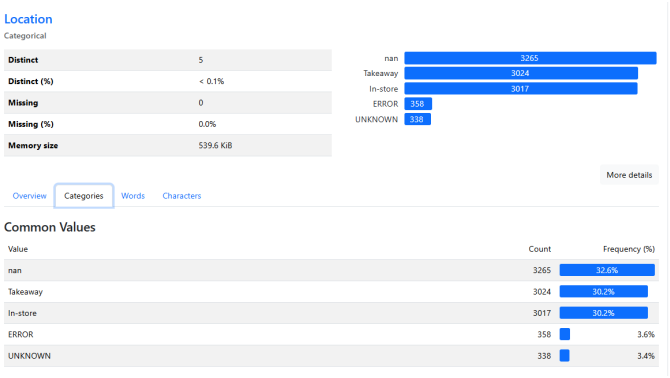
3.6 Payment Method



Rysunek 8: Profil kolumny *Payment Method*

Niepoprawne to 31,78% całości - połaczenie *nan*, *ERROR* i *UNKNOWN*. Oczywiście przy odczytywaniu karty mogą zdarzać się błędy, ale wtedy transakcje raczej powinny być odrzucane i nie znajdować się w bazie danych. Dodatkowo 2 razy wystąpiły literówki - Digital Walle i CreditCard zamiast Digital Wallet i Credit Card (występujących znacznie częściej). Poprawne kategorie to Digital Wallet, Credit Card i Cash.

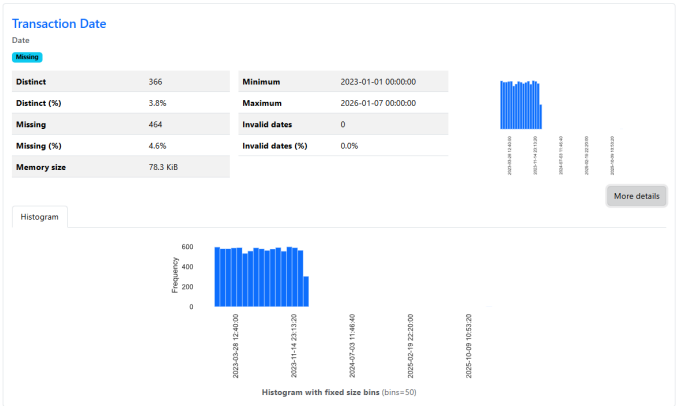
3.7 Location



Rysunek 9: Profil kolumny *Location*

Podobnie jak poprzednio, wielu danych brakuje. Nan, ERROR i UNKNOWN występują z częstością 39,61%. Poza tym, 2 poprawne kategorie to Takeaway lub In-store.

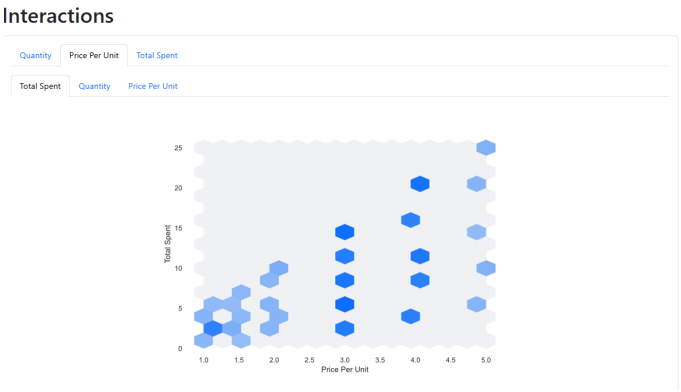
3.8 Transaction Date



Rysunek 10: Profil kolumny *Transaction Date*

Przez jedną literówkę w danych zakres wydaje się być od 2023-01-01 do 2026-01-07. Rzeczywiście jednak dane są od 2023-01-01 do 2023-12-31, a wystąpiła literówka w 2026 i rok powinien zostać zmieniony na 2023. Poza tym rozkład jest dość równomierny, jedynie w grudniu jest wyraźny spadek. Brakuje 4,6% wartości.

3.9 Przykład interakcji



Rysunek 11: Przykład interakcji między Price Per Unit a Total Spent

Jak widać, występuje pozytywna korelacja miedzy Price Per Unit a Total Spent.