

# SPRAWOZDANIE

Zajęcia: Eksploracja i wizualizacja danych

Prowadzący: prof. dr hab. inż. Vasyl Martsenyuk

Laboratorium nr 4 Data: 24.11.2021 Temat: Użycie KNIME w celu wizualizacji Wariant: 1	Piotr Rybka Informatyka II stopień, niestacjonarne, III semestr, gr. 1
--	---

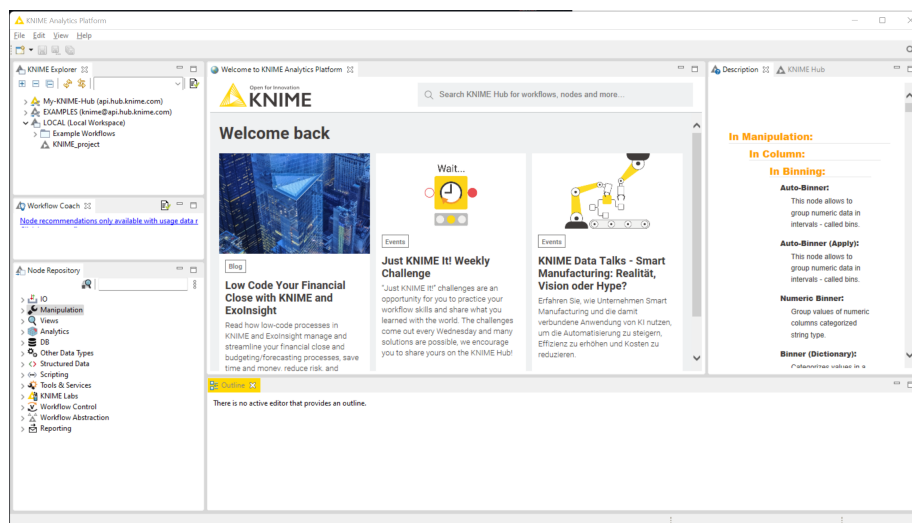
Repozytorium z kodem programu:

[https://github.com/prybka82-student/eksploracja\\_i\\_wizualizacja\\_danych/tree/zadanie04](https://github.com/prybka82-student/eksploracja_i_wizualizacja_danych/tree/zadanie04)

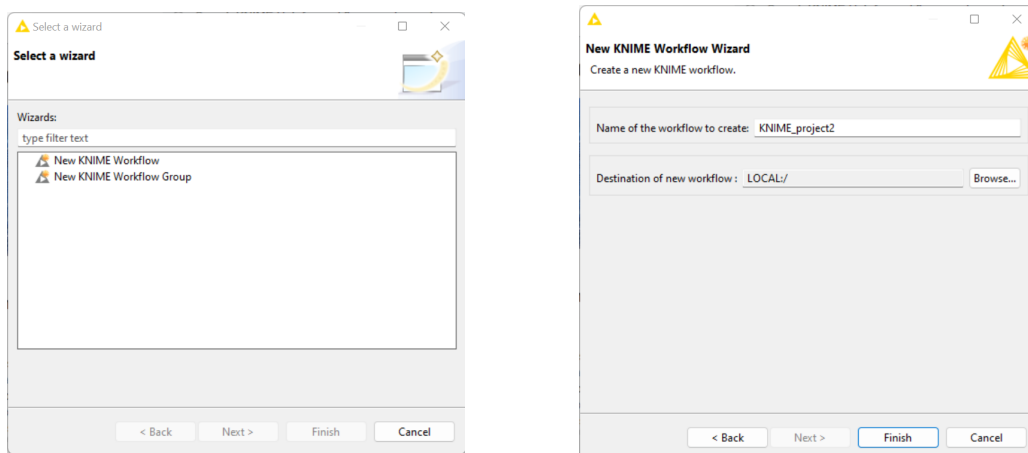
# Polecenie

Na podstawie danych ze zbioru [1] wypróbować podstawowe metody biblioteki „pyspark”.

**Zadanie 1.** Pograć program „Knime” ze strony <https://www.knime.com/downloads> i zainstalować



Rysunek 1: Interfejs zainstalowanego programu „Knime”



Rysunek 2: Tworzenie nowego projektu



Rysunek 3: Węzeł wczytujący pliki .csv

## Zadanie 2. Utworzyć nowy projekt i wczytać dane z pliku .csv

Tworzenie projektu (tzw. „workflow”) – 2.

Dodanie do przepływu (ang. *workflow*) węzła CSV Reader pozwalającego wczytywać pliki .csv – 3.

Dostępne ustawienia deserializacji pliku .csv (rys. 4):

- separator wierszy (*Row delimiter*),
- separator kolumn (*Column delimiter*),
- ograniczniki napisów (*Quote char*),

- automatyczne rozpoznawanie nagłówków (*Has column header*).

Wybór kolumn i ustawianie typów danych – 5.

Załadowanie danych przez wybranie komendy „Execute” – 6.

### Zadanie 3. Wydzielić podgrupy danych numerycznych przy użyciu węzła Numeric Binner

Dodanie do przepływu węzła Numeric Binner – 7.

Ustawienie przedziałów wartości na podstawie danych z kolumny „years”. Ustawione przedziały (rys. 8):

Nazwa	Dolne ograniczenie	Górne ograniczenie
Lata 1990-1992	– inf	1992
Lata 1992-1994	1992	1994
Lata 1994-1996	1994	1996
Lata 1996-1998	1996	1998
Lata 1998-	1998	inf

Uruchomienie węzła – analogicznie jak poprzednio (rys. 9).

### Zadanie 4. Korzystając z danych podzielonych na podgrupy, wygenerować wykres rozrzutu

Połączenie węzła Scatter Plot z uprzednio dodanymi do przepływu – rys. 10.

Ustawienia węzła Scatter Plot, generowanie i wyświetlanie wykresu – rys. 11.

Uzyskany wykres – rys. 12.

### Zadanie 5. Wygenerować histogram

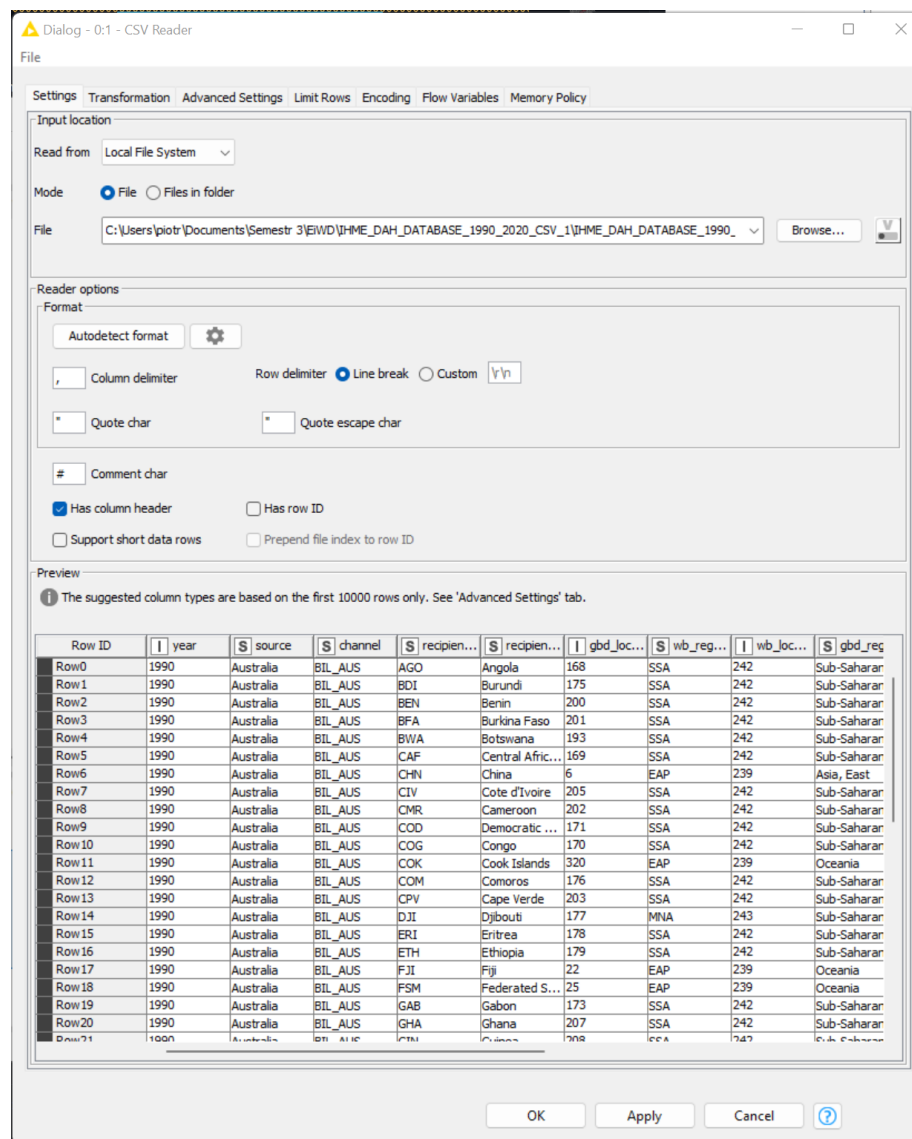
Dodanie węzła generującego histogram do węzła CSV Reader i wskazanie, jakie dane zawierać będzie histogram (w przykładzie – kolumny „year”, „gbd\_location\_id”, „gbd\_region\_id”).

Wygenerowany histogram i dodatkowe ustawienia: **binning column** – kolumna, względem której dzielone są wartości, **Aggregation method** – zastosowana miara (w przykładzie – **row count** – liczba wierszy) – rys. 14

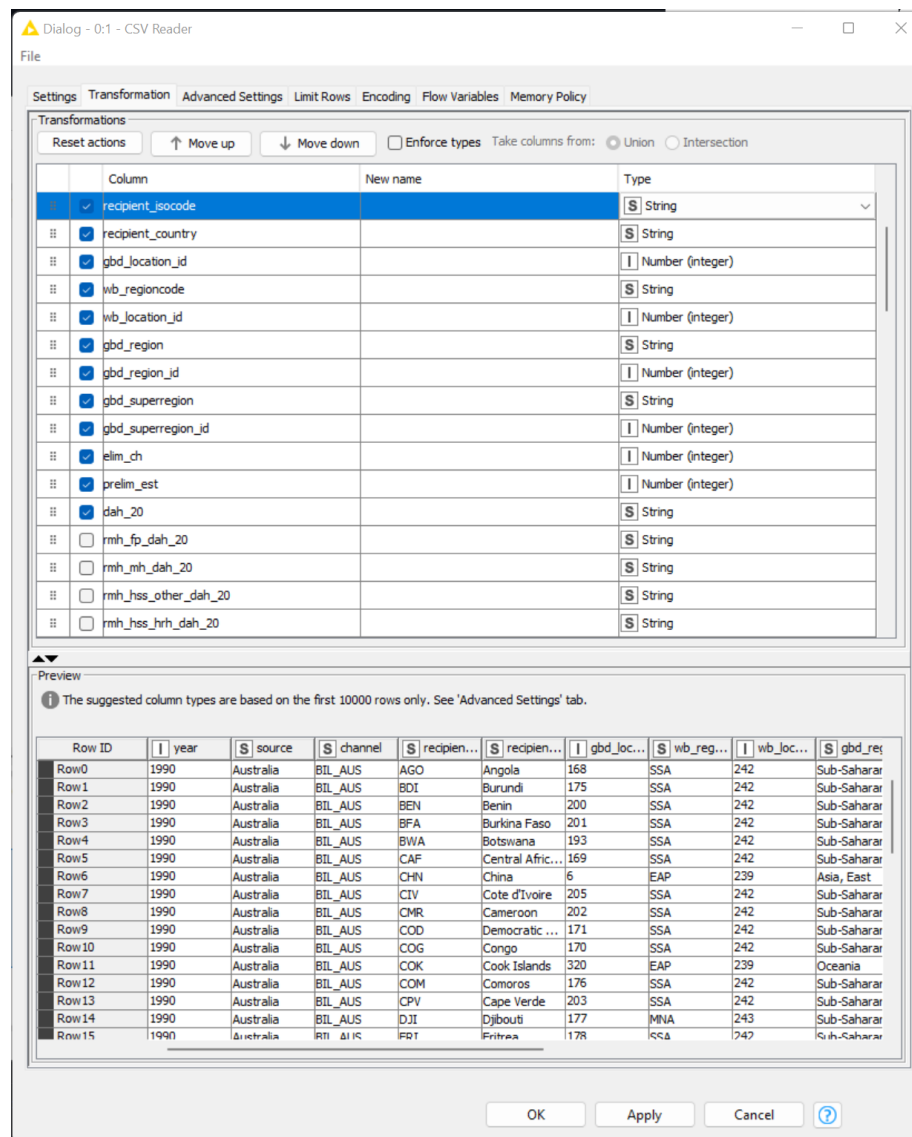
## **Zadanie 6. Wygenerować macierz wykresów rozrzutu**

Węzeł generujący macierz wykresów rozrzutu podłączony do źródła danych – 15

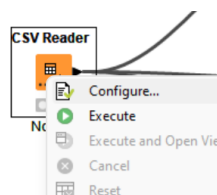
Wygenerowany wykres i jego ustawienia – wybór danych do każdego wykresu w macierzy (w przykładzie kolumny „year”, „source”, „channel”) – rys. 16.



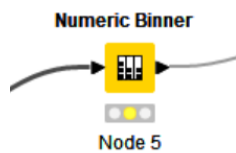
Rysunek 4: Parametry deserializacji pliku .csv



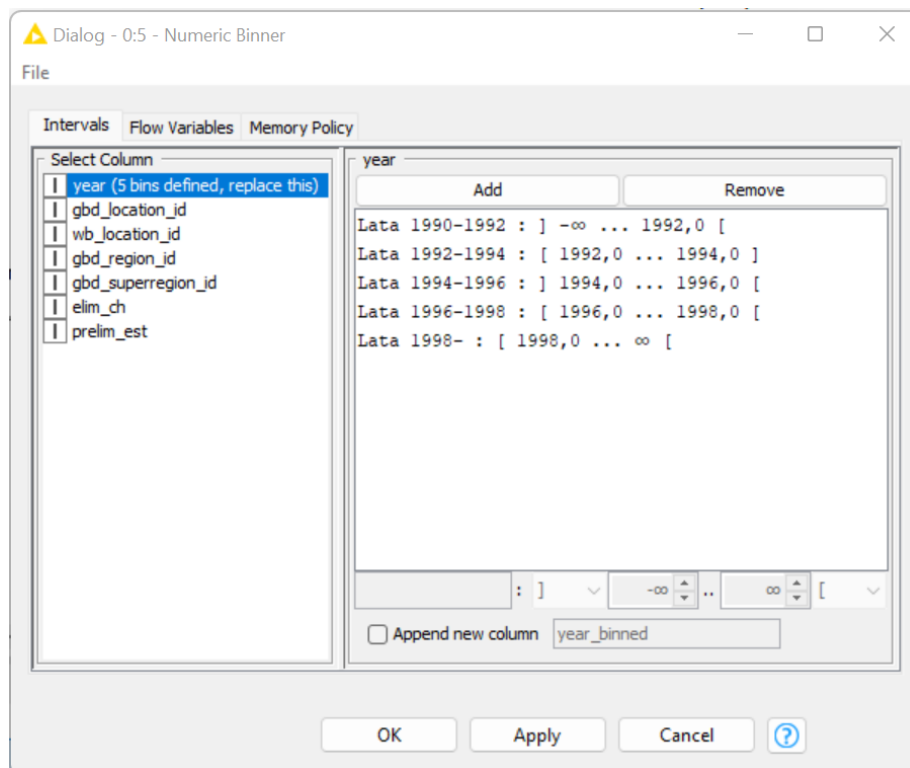
Rysunek 5: Okno wybierania kolumn i typu danych



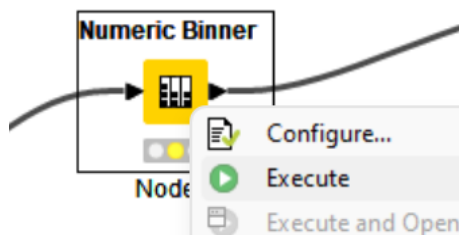
Rysunek 6: Uruchomienie węzła CSV reader



Rysunek 7: Węzeł Numeric Binner

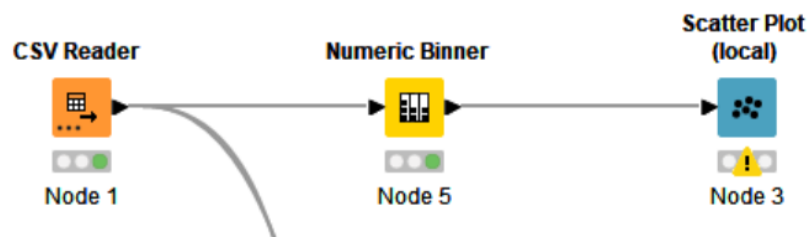


Rysunek 8: Węzeł Numeric Binner

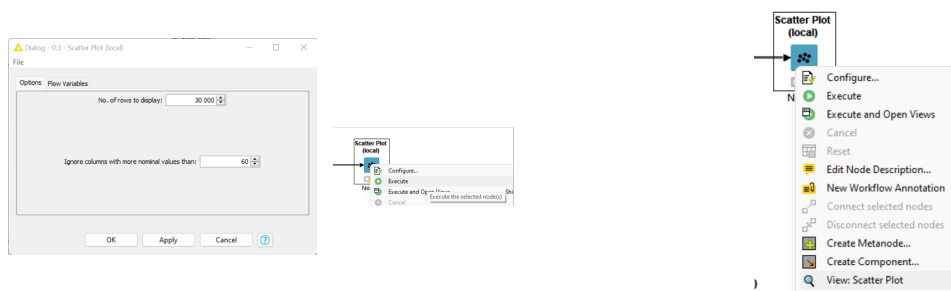


Rysunek 9: Węzeł Uruchomienie węzła Numeric Binner

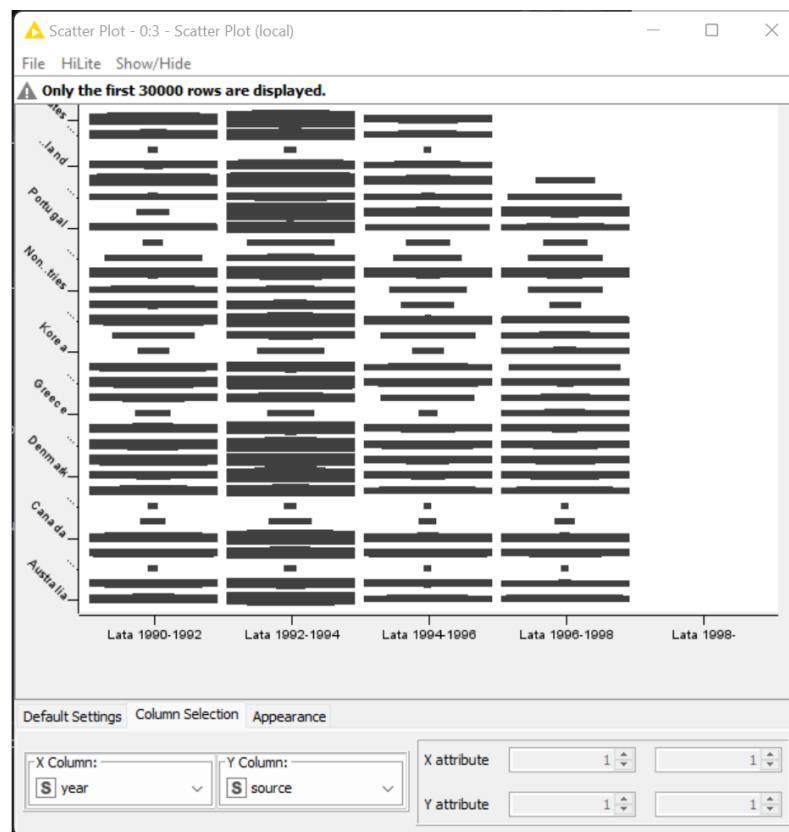




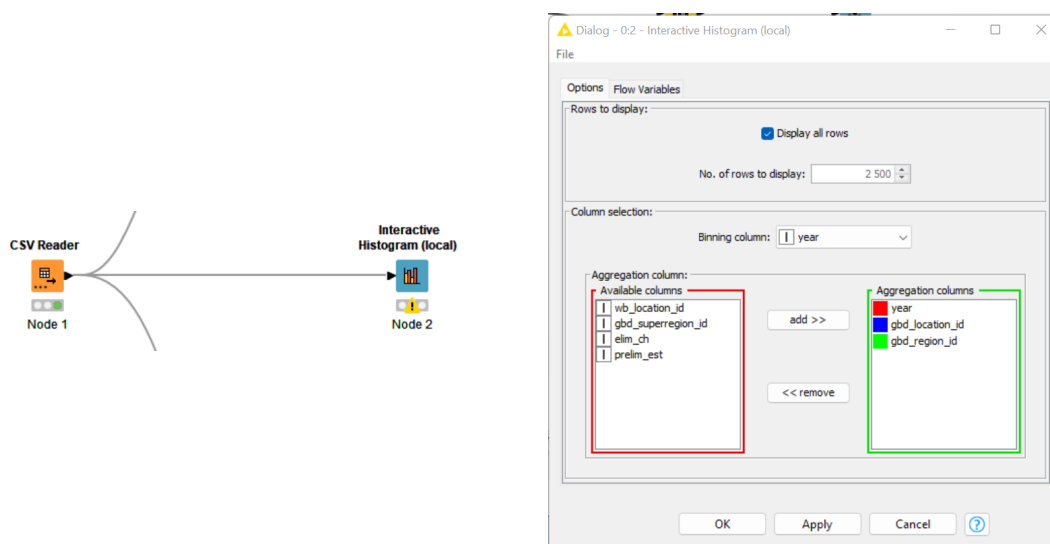
Rysunek 10: Połączenie węzłów CSV Reader, Numeric Binner, Scatter Plot



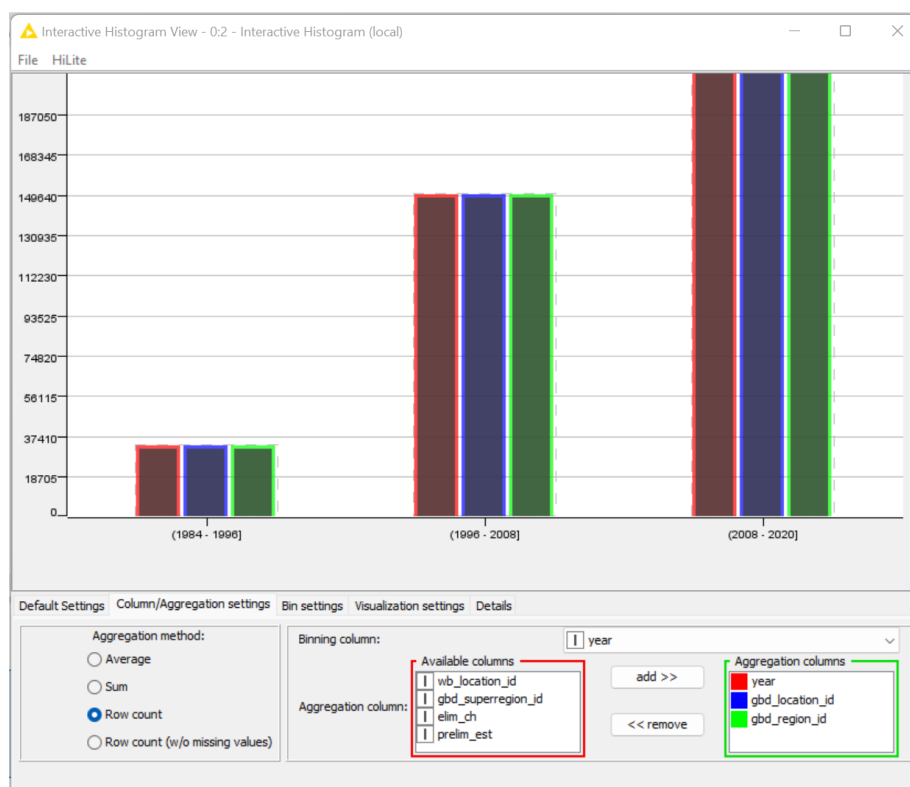
Rysunek 11: Ustawienia węzła Scatter Plot, uruchomienie i wyświetlanie wykresu



Rysunek 12: Wykres rozrzutu na podstawie danych z kolumny „year” (podzielonych na 5 przedziałów) i „source”



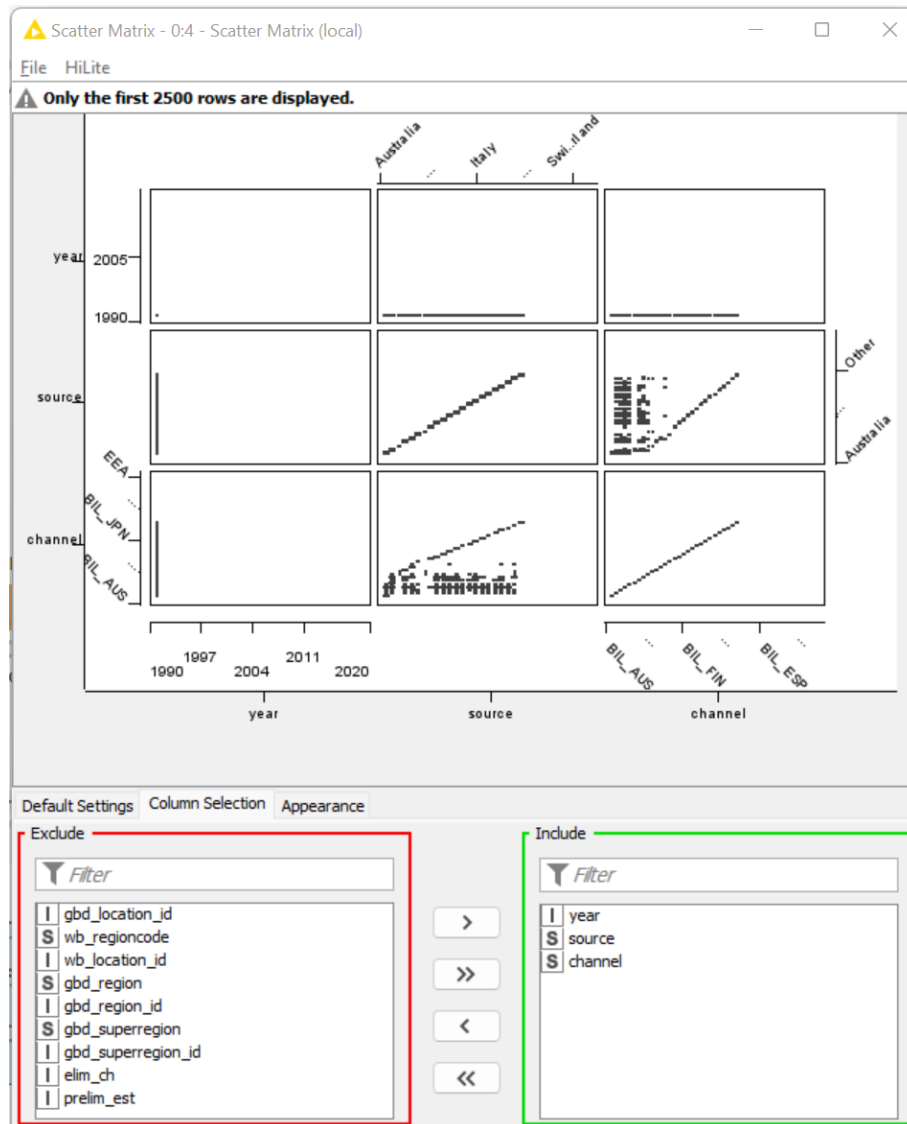
Rysunek 13: Węzeł histogramu i jego ustawienia



Rysunek 14: Wygenerowany histogram i dodatkowe ustawienia wykresu



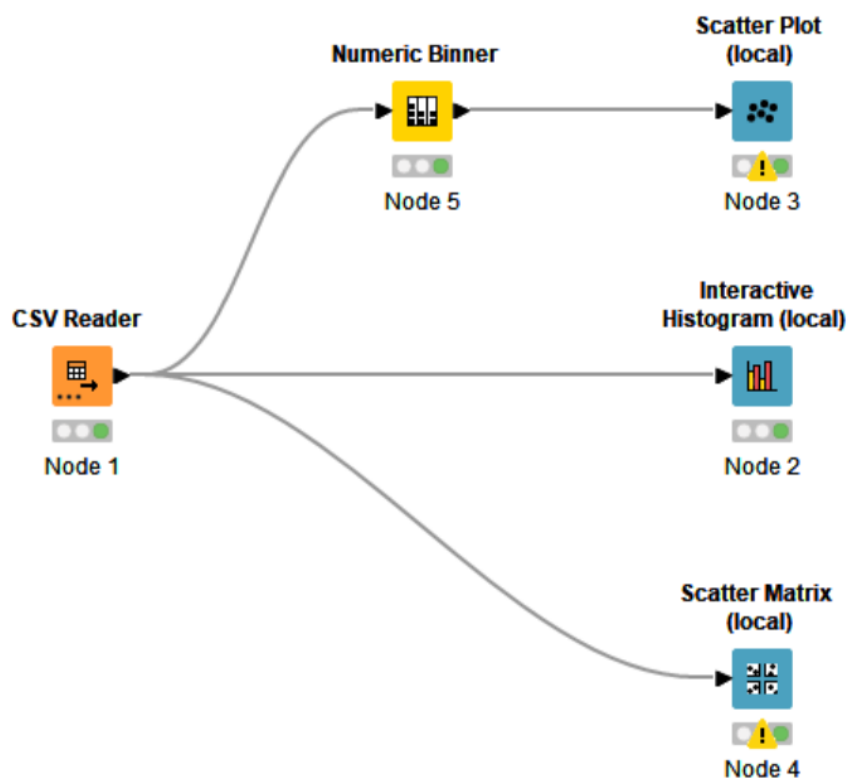
Rysunek 15: Połączenie węzła **Scatter Matrix** ze źródłem danych (plik **.csv**)



Rysunek 16: Wygenerowana macierz wykresów rozrzutu i jej ustawienia

# Wnioski

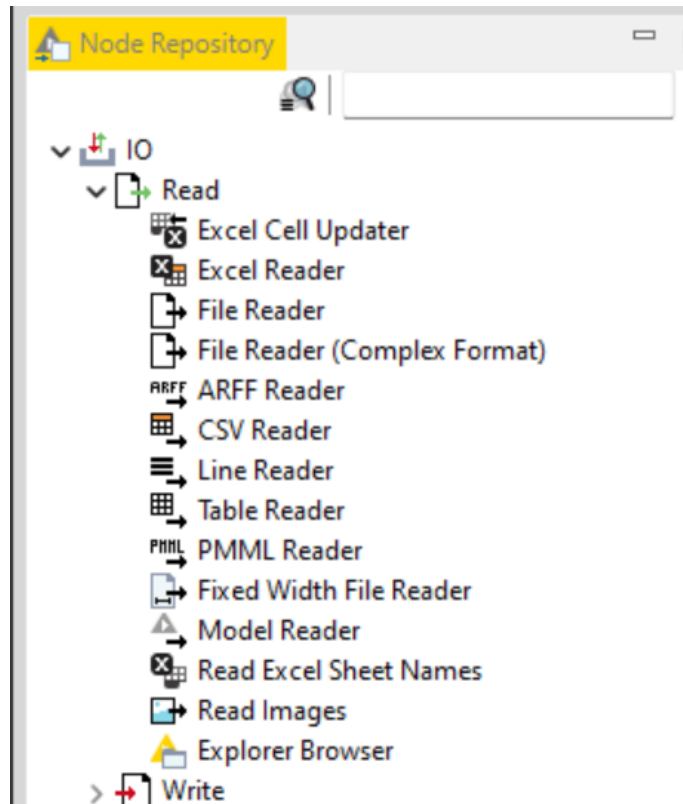
W programie KNIME przetwarzanie danych polega na ułożeniu odpowiednich węzłów (ang. *nodes*) w tzw. przepływie – rys. 17.



Rysunek 17: Układ węzłów w przepływie

Istnieje kilka rodzajów węzłów, m.in.

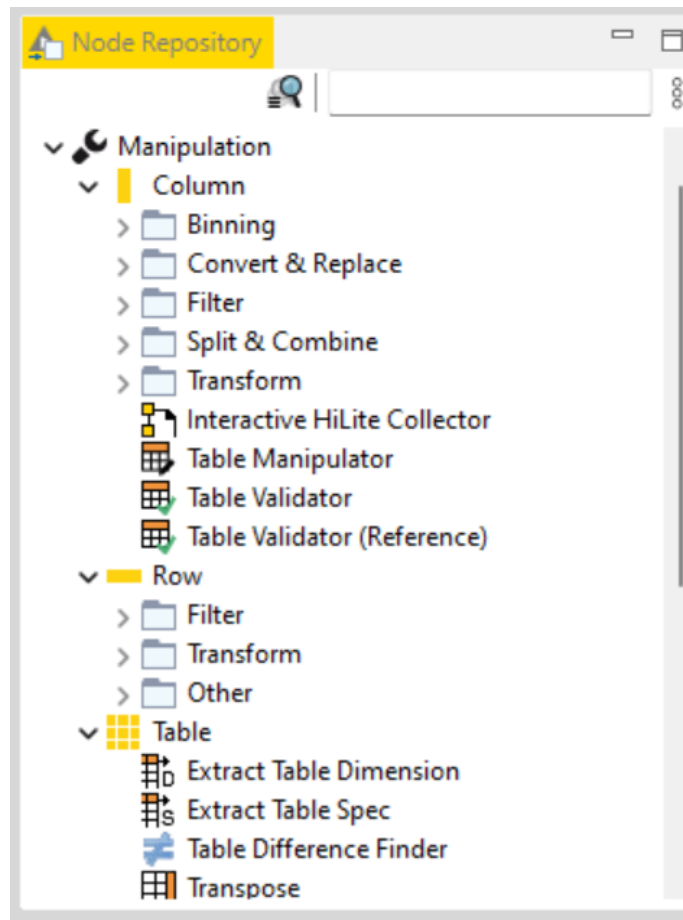
- węzły pobierające dane, np. **CSV Reader** – rys. 18;
- węzły przetwarzające dane, np. **Numeric Binner** – rys. 19;
- węzły wizualizujące dane, np. **Scatter Plot** – rys. 20;



Rysunek 18: Węzły odczytujące dostępne w programie KNIME

Podwójne kliknięcie w każdy węzeł prowadzi do jego ustawień. W przypadku węzłów wizualizujących najważniejsze ustawienia dostępne są po wyświetleniu danej wizualizacji.

Dane przechowywane w węźle nie są od razu dostępne po jego dodaniu do przepływu. Można je w każdej chwili usunąć opcją **Reset**. Wykorzystanie węzła lub węzłów pobierających z niego dane wymaga zawsze załadowania danych opcją **Execute**.

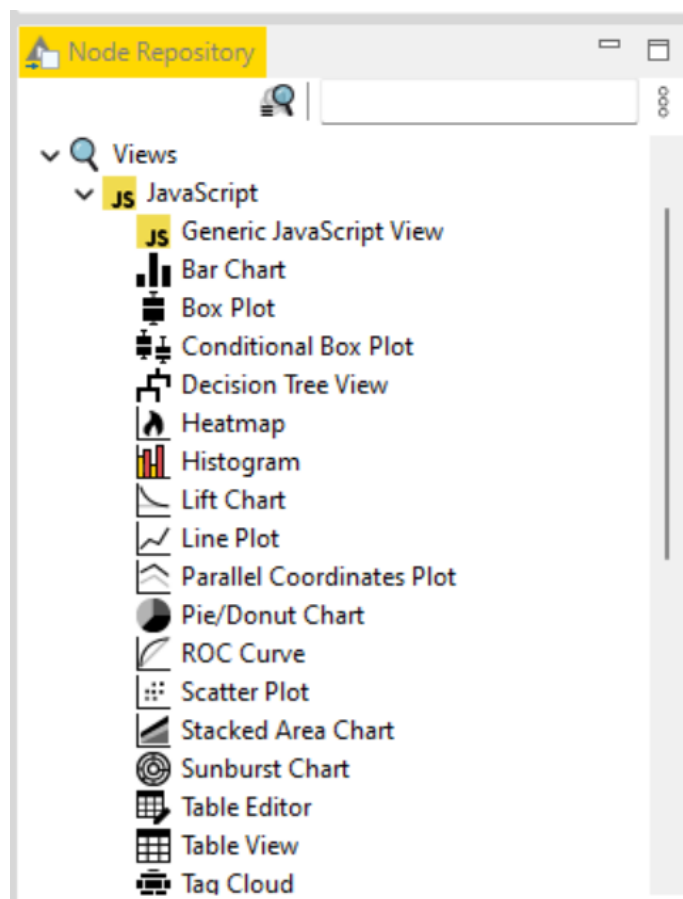


Rysunek 19: Węzły przetwarzające dane w programie KNIME

## Strony internetowe

- [1] *Development Assistance for Health Database 1990-2020*. URL: <http://ghdx.healthdata.org/record/ihme-data/development-assistance-health-database-1990-2020> (term. wiz. 21.10.2021).





Rysunek 20: Węzły wizualizujące dane w programie KNIME