**Instytut Informatyki  
Wydział Nauk Ścisłych i Technicznych  
Uniwersytet Rzeszowski**

**Przedmiot:**

**Hurtownie danych**

**Dokumentacja projektu:**

***DataFusion***

**Wykonali: Dawid Olko i Piotr Smoła**

**Prowadzący: mgr inż. Adam Szczur**

**Rzeszów 2025**

Spis treści

[1. Temat i cel projektu 3](#_Toc198666736)

[2. Techniczne aspekty projektu 4](#_Toc198666737)

[2.1. Funkcjonalności aplikacji 4](#_Toc198666738)

[2.2. Wykorzystane technologie 6](#_Toc198666739)

[2.3. Projekt GUI 7](#_Toc198666740)

[3. Wygląd i użytkowanie aplikacji 12](#_Toc198666741)

[3.1. Wymagania do uruchomienia aplikacji 12](#_Toc198666742)

[3.2. Obsługa aplikacji 12](#_Toc198666743)

[3.2.1. Wczytanie zbioru danych 12](#_Toc198666744)

[3.2.2. Analiza statystyczna danych 13](#_Toc198666745)

[3.2.3. Ekstrakcja podtablic i zastępowanie wartości 13](#_Toc198666746)

[3.2.4. Skalowanie i wizualizacja danych 14](#_Toc198666747)

[3.2.5. Czyszczenie i transformacja danych 15](#_Toc198666748)

[3.2.6. Zapisywanie danych 16](#_Toc198666749)

[4. Eksperymenty na danych 17](#_Toc198666750)

[4.1. Wykorzystane zbiory danych 17](#_Toc198666751)

[4.2. Przebieg eksperymentów i wyniki 20](#_Toc198666752)

[4.3. Analiza uzyskanych wyników i wnioski 24](#_Toc198666753)

[5. Literatura 26](#_Toc198666754)

# Temat i cel projektu

**Temat projektu:**

Projekt dotyczy stworzenia aplikacji opartej na graficznym interfejsie użytkownika (GUI), która ma na celu wspomaganie procesów analizy i obróbki danych. Program wykorzystuje dane z dwóch publicznych zbiorów: **UCI Adult Dataset** (dane o zarobkach) oraz **Chronic Kidney Disease Dataset** (dane o chorobach nerek). Aplikacja umożliwia użytkownikowi wykonanie różnych operacji na danych, takich jak obliczanie miar statystycznych, przekształcanie danych (np. kodowanie zmiennych kategorycznych, standaryzacja), a także umożliwia usuwanie, zamienianie i filtrowanie danych.

**Cel projektu:**

Celem projektu jest zaprojektowanie oraz wykonanie aplikacji do analizy danych, która pozwala na:

* **Ładowanie danych** z wybranych zestawów danych (UCI Adult Dataset oraz Chronic Kidney Disease Dataset).
* **Obliczanie miar statystycznych** (średnia, mediana, odchylenie standardowe, korelacja itp.) zarówno dla danych numerycznych, jak i kategorycznych.
* **Podstawowe operacje na danych**:
  + Zastępowanie wartości w kolumnach danych (np. zmiana wartości w kolumnie z <=50K na >50K).
  + Usuwanie lub zachowanie wybranych wierszy oraz kolumn.
  + Wykonywanie operacji standaryzacji (np. za pomocą StandardScaler, MinMaxScaler).
* **Reprezentacja danych** w postaci wykresów, takich jak wykresy słupkowe, punktowe, kołowe, oraz opcjonalnie mapa cieplna.
* **Obsługa danych brakujących**: umożliwienie ich wypełniania średnią, medianą lub usuwania.

# Techniczne aspekty projektu

## Funkcjonalności aplikacji

Poniżej znajdują się wszystkie wymagania (wariant dla projektu 2‑osobowego) wraz z krótkim opisem sposobu, w jaki zostanie zrealizowany w aplikacji **DataFusion**.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Funkcjonalność** | **Status** | **Uszczegółowienie implementacji** |
| 1 | **Odczyt danych z CSV** | ✅ | load\_dataset() obsługuje *Adult* i *Chronic Kidney Disease* oraz niestandardowe pliki CSV z automatyczną detekcją nagłówków i typów danych. |
| 2 | **Miary statystyczne (numeryczne i kategoryczne)** | ✅ | Numeryczne: min, max, mean, median, std, variance, mode, skewness, kurtosis. Kategoryczne: słownik liczebności każdej kategorii + mode. Obliczenia realizuje compute\_statistics(). |
| 3 | **Korelacje atrybutów (dwie metody)** | ✅ | compute\_correlation() generuje macierze **Pearson** i **Spearman**, prezentowane w modalnym oknie po kliknięciu przycisku "Correlation". |
| 4 | **Ekstrakcja podtablic (dwa warianty)** | ✅ | extract\_subtable() pozwala usunąć wskazane wiersze/kolumny albo zachować tylko podane (flaga keep). Obsługuje pojedyncze indeksy oraz zakresy indeksów podane w formie "start-koniec". |
| 5 | **Zastępowanie wartości** | ✅ | replace\_values() zamienia wskazaną wartość w kolumnie; replace\_all\_values() podmienia wszystkie wartości w kolumnie. Funkcja automatycznie obsługuje różne typy danych, w tym numeryczne i kategoryczne. |
| 6 | **Skalowanie i standaryzacja** | ✅ | scale\_columns() korzysta z **StandardScaler** lub **MinMaxScaler** z biblioteki scikit-learn do normalizacji wybranych kolumn numerycznych. |
| 7 | **Obsługa braków danych** | ✅ | handle\_missing\_values() umożliwia usunięcie wierszy z brakami (*remove*) albo wypełnienie ich średnią, medianą lub modą dla kolumn numerycznych. |
| 8 | **Usuwanie duplikatów** | ✅ | Przycisk w zakładce *Data Cleaning & Transformation* wywołuje remove\_duplicates(), który eliminuje powtarzające się wiersze z zestawu danych. |
| 9 | **Kodowanie symboliczne (trzy metody)** | ✅ | • **One‑Hot** (one\_hot\_encoding() – generuje kolumny binarne dla każdej kategorii),<br>• **Binary** (binary\_encoding() – kody numeryczne dla kategorii),<br>• **Target** (target\_encoding() – średnia zmiennej celu dla każdej kategorii). |
| 10 | **Wizualizacje danych (≥ 4 rodzaje)** | ✅ | Dostępne: Histogram (z wykrywaniem typów danych), Boxplot (z oznaczeniem kwartyli), Bar Chart, Line Plot (z wykrywaniem outlierów), Pie Chart (z opcją łączenia małych kategorii). Wykresy generuje generate\_plot() z konfigurowalnymi opcjami (etykiety wartości, siatka, tytuł wykresu, motyw kolorystyczny). |
| 12 | **Intuicyjne GUI** | ✅ | PySimpleGUI, pięć zakładek: Creators & Info, Data & Stats, Replacement & Subtable, Scaling & Visualisation, Data Cleaning & Transformation. Okno aplikacji ma stały rozmiar i wycentrowaną pozycję, z ikoną aplikacji. |
| 13 | **Dokumentacja projektu** | ✅\* | Niniejszy opis stanowi część rozdziału 2.1 dokumentacji tworzonej wg wzoru *dokumentacjaHD\_2.docx*. |

\* Ostateczna kompozycja dokumentu przygotowana po zakończeniu testów.

## Wykorzystane technologie

**Python 3.x** Jest głównym językiem programowania wykorzystanym w projekcie DataFusion. Wybór tego języka umożliwił szybką implementację złożonych funkcjonalności dzięki jego czytelnej składni oraz bogatemu ekosystemowi bibliotek do analizy danych. Python oferuje również doskonałe wsparcie dla operacji na różnych typach danych, co jest kluczowe w aplikacji przetwarzającej dane z różnych źródeł.

**PySimpleGUI** Biblioteka odpowiedzialna za budowę intuicyjnego interfejsu graficznego aplikacji. Zastosowanie PySimpleGUI pozwoliło na utworzenie przejrzystego, modułowego układu z wykorzystaniem systemu zakładek, tabel, przycisków i pól tekstowych. Biblioteka ta umożliwiła również implementację zaawansowanych elementów interfejsu, takich jak modalne okna dialogowe do wyświetlania wyników operacji oraz interaktywne elementy formularzy do konfiguracji parametrów transformacji danych.

**Pandas** Stanowi fundament operacji na danych w aplikacji DataFusion. Wykorzystano struktury DataFrame do przechowywania, filtrowania i transformacji danych tabelarycznych. Pandas dostarcza zaawansowanych funkcji do wczytywania plików CSV z automatyczną detekcją typów danych, obsługi wartości brakujących oraz przeprowadzania złożonych operacji na kolumnach i wierszach. Biblioteka ta jest również wykorzystywana do obliczania statystyk opisowych i agregacji danych.

**NumPy** Zapewnia wydajne struktury danych i operacje na tablicach wielowymiarowych, które są wykorzystywane w obliczeniach statystycznych i transformacjach numerycznych. NumPy umożliwia szybkie przetwarzanie dużych zbiorów danych dzięki zoptymalizowanym algorytmom napisanym w C. W aplikacji używamy tej biblioteki przede wszystkim do operacji skalowania, normalizacji oraz przy obliczaniu macierzy korelacji.

**SciPy (stats)** Dostarcza zaawansowanych funkcji statystycznych wykorzystywanych w aplikacji, takich jak obliczanie skośności (skew) i kurtozy (kurtosis) dla zmiennych numerycznych. Te miary są kluczowe dla zrozumienia rozkładów danych i identyfikacji potencjalnych anomalii. SciPy uzupełnia funkcjonalność NumPy i Pandas w zakresie zaawansowanej analizy statystycznej.

**Matplotlib** + **FigureCanvasTkAgg** Stanowi podstawę modułu wizualizacji w aplikacji DataFusion. Matplotlib umożliwia tworzenie różnorodnych typów wykresów (histogramy, wykresy pudełkowe, wykresy słupkowe, liniowe i kołowe) z rozbudowanymi opcjami konfiguracyjnymi. Klasa FigureCanvasTkAgg zapewnia integrację wykresów Matplotlib z interfejsem tkinter używanym przez PySimpleGUI, co pozwala na osadzanie interaktywnych wizualizacji bezpośrednio w oknach aplikacji.

**Scikit-learn** Dostarcza narzędzi do przetwarzania danych wykorzystywanych w modułach transformacji i skalowania. W aplikacji używamy przede wszystkim klas StandardScaler i MinMaxScaler do normalizacji danych numerycznych. Te transformacje są kluczowe dla przygotowania danych do dalszej analizy, szczególnie gdy zmienne mają różne jednostki i zakresy wartości. Biblioteka zapewnia spójny interfejs do różnych metod przetwarzania wstępnego.

**PIL (Pillow)** Biblioteka do przetwarzania obrazów, wykorzystywana w aplikacji do obsługi ikon i logo. PIL umożliwia wczytywanie, modyfikowanie i wyświetlanie obrazów w różnych formatach, co pozwala na dopasowanie elementów graficznych do interfejsu użytkownika. W projekcie DataFusion, Pillow jest używana do ustawienia ikony aplikacji i manipulacji obrazami wyświetlanymi w zakładce informacyjnej.

## Projekt GUI

**A screenshot of a computer

Description automatically generated**

**Obraz 1: Okno obsługi wartości brakujących**

**Tytuł**: Obsługa brakujących wartości w DataFusion

Okno przedstawia wyniki zastosowania strategii 'remove' (usuwanie) do obsługi brakujących wartości w zbiorze danych. Widoczna jest tabela prezentująca wyczyszczony zestaw danych, z którego usunięto wiersze zawierające wartości NULL. Tabela zawiera kolumny takie jak "id", "age", "bp", "sg", "al" i "su" z odpowiadającymi im wartościami. Implementacja funkcji handle\_missing\_values() umożliwia usuwanie niekompletnych wierszy, co jest jedną z kluczowych metod.

A screenshot of a computer

Description automatically generated

**Obraz 2: Rezultat ekstrakcji podtablicy**

**Tytuł**: Wyodrębnianie podtablic z zestawu Adult Dataset

Zrzut ekranu prezentuje okno modalne z wynikami ekstrakcji podtablicy z zestawu danych Adult Dataset. Widocznych jest kilka pierwszych wierszy (z informacją, że pokazano 100 z 32559 wierszy) zawierających dane o osobach z kolumnami: workclass, fnlwgt, education, education-num, marital-status i occupation. Dane są prezentowane w formie tabeli z możliwością przewijania. Ta funkcjonalność realizuje wymaganie ekstrakcji podtablic, umożliwiając użytkownikowi pracę na wybranym podzbiorze danych.

A screenshot of a computer

Description automatically generated

**Obraz 3: Analiza korelacji**

**Tytuł**: Macierze korelacji Pearsona i Spearmana

Okno przedstawia wyniki analizy korelacji przeprowadzonej za pomocą funkcji compute\_correlation(). Na górze widoczna jest macierz korelacji Pearsona, a na dole macierz korelacji Spearmana dla zmiennych numerycznych: age, fnlwgt, education-num, capital-gain, capital-loss i hours-per-week. Macierze korelacji są przedstawione w formie tabel z wartościami współczynników zaokrąglonymi do trzech miejsc po przecinku. Wartości 1.0 na przekątnej reprezentują korelację zmiennej z samą sobą. Ten element realizuje wymaganie dotyczące analizy korelacji przy użyciu dwóch różnych metod.

A screenshot of a computer

Description automatically generated

**Obraz 4: Wyniki skalowania**

**Tytuł**: Standaryzacja zmiennej age za pomocą StandardScaler

Ekran prezentuje wyniki zastosowania StandardScaler do kolumny "age". Widoczna jest tabela zawierająca znormalizowane wartości, gdzie oryginalne wartości wieku zostały przekształcone do skali o średniej 0 i odchyleniu standardowym 1. Wartości po skalowaniu mieszczą się w przybliżonym zakresie od -1.62 do 2.62, co jest charakterystyczne dla standaryzacji. Ta funkcjonalność jest realizowana przez funkcję scale\_columns() wykorzystującą bibliotekę scikit-learn.

**A screenshot of a computer program

Description automatically generated**

**Obraz 5: Interfejs skalowania i wizualizacji**

**Tytuł**: Zakładka Scaling & Visualization w aplikacji DataFusion

Zrzut ekranu przedstawia zakładkę "Scaling & Visualization", która zawiera dwie główne sekcje: "Data Scaling" (u góry) oraz "Data Visualization" (u dołu). W części dotyczącej skalowania widoczne są opcje wyboru kolumn do skalowania (aktualnie wybrana kolumna "age"), wybór metody skalowania (StandardScaler lub MinMaxScaler) oraz przyciski do wykonania operacji. W sekcji wizualizacji dostępne są opcje wyboru kolumny do wykresu, typu wykresu (Histogram, Boxplot, Bar Chart, Line Plot, Pie Chart) oraz różne opcje formatowania, jak pokazywanie etykiet wartości, siatki czy wybór motywu kolorystycznego.

A screenshot of a computer

Description automatically generated

**Obraz 6: Histogram zmiennej age**

**Tytuł**: Wizualizacja rozkładu znormalizowanej zmiennej age

Okno prezentuje histogram dla znormalizowanej kolumny "age". Na wykresie widoczny jest rozkład wartości zmiennej po standaryzacji, z wartościami na osi X w zakresie od -3 do 2. Każdy słupek histogramu jest opatrzony etykietą z liczalnością (np. 80, 75, 44). W prawym górnym rogu znajduje się informacja o wartościach min., max. i średniej. Wykres został wygenerowany za pomocą funkcji generate\_plot() z użyciem biblioteki matplotlib, co stanowi realizację wymagania dotyczącego wizualizacji danych.

A screenshot of a computer

Description automatically generated

**Obraz 7: Statystyki zestawu danych Adult Dataset**

**Tytuł**: Analiza statystyczna zestawu Adult Dataset

Ekran przedstawia zakładkę "Data & Stats" z wczytanym zbiorem Adult Dataset (32561 wierszy, 15 kolumn). W górnej części widoczne są opcje ładowania danych. Poniżej znajdują się panele ze statystykami dla kolumn numerycznych i kategorycznych. Dla zmiennych numerycznych (age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week) pokazane są miary takie jak Min, Max, Mean, Median, Std, Mode, Variance, Skewness i Kurtosis. Dla zmiennych kategorycznych (workclass, education, marital-status itp.) wyświetlone są wartości liczebności każdej kategorii oraz moda. Ta funkcjonalność realizuje wymaganie dotyczące obliczania miar statystycznych.

A screenshot of a computer

Description automatically generated

**Obraz 8: Ekran powitalny aplikacji**

**Tytuł**: Strona powitalna aplikacji DataFusion

Zrzut ekranu przedstawia zakładkę "Creators & Info" - stronę powitalną aplikacji DataFusion. Na środku ekranu widoczne jest logo aplikacji przedstawiające trójwymiarową strukturę danych z nazwą "DataFusion - Data Warehouse & Analysis". Poniżej znajdują się informacje o twórcach projektu: Dawidzie Olko i Piotrze Smole. U dołu umieszczony jest przycisk "Go to data" umożliwiający przejście do głównej części aplikacji. Ten ekran stanowi element intuicyjnego interfejsu użytkownika.

A screenshot of a computer

Description automatically generated

**Obraz 9: Zakładka Replacement & Subtable**

**Tytuł**: Interfejs ekstrakcji podtablic i zastępowania wartości

Ekran pokazuje zakładkę "Replacement & Subtable" z dwiema głównymi sekcjami: "Subtable Extraction" (u góry) oraz "Value Replacement" (u dołu). W sekcji ekstrakcji podtablic użytkownik może wybrać między usuwaniem a zachowywaniem wierszy/kolumn, wprowadzić indeksy wierszy (w przykładzie "2-3") oraz wybrać kolumny do operacji (wybrana "age"). W sekcji zastępowania wartości dostępne są opcje wyboru kolumny, pobierania jej unikalnych wartości, wprowadzania nowej wartości oraz przyciski do wykonania operacji zastępowania. Ta zakładka realizuje dwa wymagania: ekstrakcję podtablic oraz zastępowanie wartości.

A screenshot of a computer

Description automatically generated

**Obraz 10: Zakładka Data Cleaning & Transformation**

**Tytuł**: Interfejs czyszczenia i transformacji danych

Zrzut ekranu przedstawia zakładkę "Data Cleaning & Transformation" z trzema głównymi sekcjami: "Handling Missing Values" (obsługa wartości brakujących), "Duplicates" (obsługa duplikatów) oraz "Encoding" (kodowanie zmiennych kategorycznych). W sekcji obsługi wartości brakujących wybrana jest strategia "remove". Sekcja duplikatów zawiera przyciski do usuwania duplikatów i zapisywania wyników. W sekcji kodowania dostępne są opcje wyboru kolumny do kodowania, wybór metody kodowania (One-Hot, Binary, Target) oraz, w przypadku kodowania target, wybór kolumny docelowej. Ta zakładka realizuje wymagania dotyczące obsługi braków danych, usuwania duplikatów oraz kodowania symbolicznego.

Przedstawione zrzuty ekranu dokumentują realizację wszystkich kluczowych funkcjonalności aplikacji DataFusion, zgodnie z wymaganiami projektu, prezentując zarówno interfejs użytkownika, jak i przykładowe wyniki działania poszczególnych funkcji przetwarzania danych.

# Wygląd i użytkowanie aplikacji

## Wymagania do uruchomienia aplikacji

**Wymagania sprzętowe:**

* Komputer z procesorem minimum 1.5 GHz
* Minimum 4 GB pamięci RAM (zalecane 8 GB dla większych zbiorów danych)
* 100 MB wolnego miejsca na dysku twardym
* Rozdzielczość ekranu minimum 1024x768 pikseli

**Wymagania programowe:**

* System operacyjny: Windows 10/11, macOS 10.14 lub nowszy, Linux z aktualnym środowiskiem graficznym
* Python w wersji 3.6 lub nowszej
* Zainstalowane biblioteki Python:
  + PySimpleGUI (interfejs graficzny)
  + pandas (przetwarzanie danych)
  + numpy (operacje numeryczne)
  + matplotlib (wizualizacja danych)
  + scikit-learn (skalowanie i preprocessing)
  + scipy (funkcje statystyczne)
  + Pillow (obsługa grafiki)

**Instalacja wymaganych bibliotek:**

pip install PySimpleGUI pandas numpy matplotlib scikit-learn scipy pillow

## Obsługa aplikacji

### 3.2.1. Wczytanie zbioru danych

Aplikacja DataFusion umożliwia wczytanie danych z predefiniowanych zestawów lub własnego pliku CSV.

**Krok 1:** Uruchom aplikację poprzez wykonanie pliku main.py w terminalu lub wierszu poleceń:

python main.py

**Krok 2:** Po uruchomieniu aplikacji wybierz zakładkę "Data & Stats" z górnego menu (domyślnie otwiera się ekran powitalny "Creators & Info").

**Krok 3:** W sekcji "Load Data" wybierz źródło danych:

* Aby wczytać predefiniowany zestaw danych:
  1. Zaznacz opcję "Predefined Dataset"
  2. Z rozwijanej listy wybierz jedną z opcji ("Adult Dataset" lub "Kidney Disease Dataset")
  3. Kliknij przycisk "Load Data"
* Aby wczytać własny plik CSV:
  1. Zaznacz opcję "Custom CSV File"
  2. Kliknij przycisk "Browse..." i wybierz plik CSV z dysku
  3. Kliknij przycisk "Load Data"

**Krok 4:** Po załadowaniu danych w polu tekstowym pod sekcją "Load Data" pojawi się informacja o wczytanym zbiorze danych, zawierająca liczbę wierszy, kolumn oraz nazwy kolumn.

**Wskazówki:**

* Aplikacja automatycznie wykrywa nagłówki w plikach CSV
* Wartości brakujące oznaczone jako "?" są automatycznie konwertowane na NaN
* W przypadku plików bez nagłówków, kolumny będą nazwane automatycznie (col\_a, col\_b, itd.)
* Aby powrócić do oryginalnego zestawu danych po wykonaniu modyfikacji, użyj przycisku "Restore Original Data"

### 3.2.2. Analiza statystyczna danych

**Krok 1:** Po wczytaniu danych, w zakładce "Data & Stats" kliknij przycisk "Compute Stats".

**Krok 2:** W dolnej części zakładki pojawią się dwie tabele:

* Górna tabela zawiera statystyki dla zmiennych numerycznych:
  + Min - wartość minimalna
  + Max - wartość maksymalna
  + Mean - średnia arytmetyczna
  + Median - mediana (wartość środkowa)
  + Std - odchylenie standardowe
  + Mode - wartość najczęstsza
  + Variance - wariancja
  + Skewness - skośność (miara asymetrii rozkładu)
  + Kurtosis - kurtoza (miara "ogonów" rozkładu)
* Dolna tabela zawiera statystyki dla zmiennych kategorycznych:
  + Value Counts - liczności poszczególnych kategorii
  + Mode - wartość najczęstsza

**Krok 3:** Aby przeglądać dane w tabelach, użyj pasków przewijania.

**Krok 4:** Aby przeprowadzić analizę korelacji zmiennych numerycznych:

1. Kliknij przycisk "Correlation"
2. W nowym oknie zostaną wyświetlone dwie macierze korelacji:
   * Macierz korelacji Pearsona (liniowa zależność między zmiennymi)
   * Macierz korelacji Spearmana (monotoniczna zależność między zmiennymi)

**Wskazówki:**

* Wartości korelacji bliskie 1 oznaczają silną dodatnią korelację
* Wartości bliskie -1 oznaczają silną ujemną korelację
* Wartości bliskie 0 wskazują na brak korelacji
* Przekątna macierzy korelacji zawsze zawiera wartości 1 (korelacja zmiennej z samą sobą)

### 3.2.3. Ekstrakcja podtablic i zastępowanie wartości

**Krok 1:** Przejdź do zakładki "Replacement & Subtable".

**Ekstrakcja podtablic:**

**Krok 2:** W sekcji "Subtable Extraction" wybierz jedną z opcji:

* "Remove Rows/Columns" - aby usunąć wskazane wiersze/kolumny
* "Keep Rows/Columns" - aby zachować tylko wskazane wiersze/kolumny

**Krok 3:** W polu "Rows" wprowadź indeksy wierszy, które chcesz uwzględnić w operacji. Możesz użyć:

* Pojedynczych indeksów oddzielonych przecinkami, np. "0, 2, 5"
* Zakresów indeksów, np. "2-10"
* Kombinacji powyższych, np. "0, 2-5, 10, 15-20"

**Krok 4:** Wybierz kolumny do uwzględnienia w operacji:

1. Z rozwijanej listy "Columns" wybierz kolumnę
2. Kliknij "Add Column" aby dodać ją do listy wybranych kolumn
3. Powtórz dla wszystkich potrzebnych kolumn
4. Aby usunąć wszystkie wybrane kolumny, kliknij "Clear Columns"

**Krok 5:** Kliknij "Extract Subtable" aby wykonać operację.

**Krok 6:** W nowym oknie zostanie wyświetlona wynikowa podtablica. Możesz ją zamknąć klikając "Close".

**Krok 7:** Aby zapisać wynikową podtablicę do pliku CSV, kliknij "Save Subtable".

**Zastępowanie wartości:**

**Krok 8:** W sekcji "Value Replacement":

1. Z rozwijanej listy "Select Column" wybierz kolumnę, w której chcesz zastąpić wartości
2. Kliknij "Get Values" aby załadować unikalne wartości występujące w tej kolumnie
3. Z rozwijanej listy "Current Value" wybierz wartość do zastąpienia
4. W polu "New Value" wprowadź nową wartość
5. Kliknij "Replace Values in Column" aby wykonać operację
6. W nowym oknie zostanie wyświetlona tabela po zastąpieniu wartości

**Krok 9:** Aby zastąpić wszystkie wartości w kolumnie jedną wartością:

1. Z rozwijanej listy "Replace All in Column" wybierz kolumnę
2. W polu "New Value for All" wprowadź nową wartość
3. Kliknij "Replace All Values"

**Wskazówki:**

* Po wykonaniu operacji ekstrakcji podtablicy lub zastąpienia wartości, oryginalna tabela w pamięci aplikacji zostaje zaktualizowana
* Aby powrócić do oryginalnego zestawu danych, wróć do zakładki "Data & Stats" i kliknij "Restore Original Data"
* Operacje zastępowania wartości uwzględniają typ danych kolumny (numeryczny/kategoryczny)

### 3.2.4. Skalowanie i wizualizacja danych

**Krok 1:** Przejdź do zakładki "Scaling & Visualization".

**Skalowanie danych:**

**Krok 2:** W sekcji "Data Scaling":

1. Z rozwijanej listy "Columns to Scale" wybierz kolumnę do skalowania
2. Kliknij "Add" aby dodać ją do listy wybranych kolumn
3. Wybierz metodę skalowania:
   * "StandardScaler" - standaryzacja (średnia 0, odchylenie standardowe 1)
   * "MinMaxScaler" - normalizacja do zakresu [0, 1]
4. Kliknij "Apply Scaling" aby wykonać operację
5. W nowym oknie zostanie wyświetlona tabela po skalowaniu

**Krok 3:** Aby zapisać dane po skalowaniu, kliknij "Save Scaled Data".

**Krok 4:** Aby przywrócić oryginalne wartości, kliknij "Restore Original".

**Wizualizacja danych:**

**Krok 5:** W sekcji "Data Visualization":

1. Z rozwijanej listy "Select Column to Plot" wybierz kolumnę do wizualizacji
2. Wybierz typ wykresu:
   * "Histogram" - dla zmiennych numerycznych i kategorycznych
   * "Boxplot" - dla zmiennych numerycznych
   * "Bar Chart" - dla zmiennych kategorycznych i numerycznych
   * "Line Plot" - dla zmiennych numerycznych
   * "Pie Chart" - dla zmiennych kategorycznych
3. Skonfiguruj opcje wykresu:
   * Zaznacz "Show Value Labels" aby wyświetlić etykiety wartości
   * Zaznacz "Show Grid Lines" aby wyświetlić linie siatki
   * Ustaw rozmiar etykiet za pomocą suwaka "Label Size"
   * Wprowadź tytuł wykresu w polu "Chart Title"
   * Wybierz motyw kolorystyczny z listy "Color Theme"
4. Kliknij "Generate Plot" aby utworzyć wykres

**Krok 6:** Wykres zostanie wyświetlony w nowym oknie.

**Krok 7:** Aby zapisać wykres jako obraz, kliknij "Save Plot as Image" w oknie wykresu.

**Wskazówki:**

* Skalowanie można zastosować tylko do zmiennych numerycznych
* Dla zmiennych kategorycznych odpowiednie są wykresy typu Bar Chart i Pie Chart
* Histogram i Boxplot najlepiej przedstawiają rozkład zmiennych numerycznych
* Line Plot jest przydatny do wykrywania trendów i anomalii w danych
* Na wykresach typu Boxplot oznaczone są wartości odstające (outliers)

### 3.2.5. Czyszczenie i transformacja danych

**Krok 1:** Przejdź do zakładki "Data Cleaning & Transformation".

**Obsługa wartości brakujących:**

**Krok 2:** W sekcji "Handling Missing Values":

1. Z rozwijanej listy wybierz strategię obsługi wartości brakujących:
   * "remove" - usunięcie wierszy z wartościami brakującymi
   * "fill\_mean" - wypełnienie braków średnią wartością kolumny
   * "fill\_median" - wypełnienie braków medianą kolumny
   * "fill\_mode" - wypełnienie braków najczęstszą wartością kolumny
2. Kliknij "Apply Missing Values Handling" aby wykonać operację
3. W nowym oknie zostanie wyświetlona tabela po zastosowaniu wybranej strategii

**Krok 3:** Aby zapisać dane po obsłudze wartości brakujących, kliknij "Save Cleaned Data".

**Usuwanie duplikatów:**

**Krok 4:** W sekcji "Duplicates":

1. Kliknij "Remove Duplicates" aby usunąć zduplikowane wiersze
2. W nowym oknie zostanie wyświetlona tabela po usunięciu duplikatów
3. Aby zapisać dane po usunięciu duplikatów, kliknij "Save After Duplicate Removal"

**Kodowanie zmiennych kategorycznych:**

**Krok 5:** W sekcji "Encoding":

1. Z rozwijanej listy "Select Column to Encode" wybierz kolumnę kategoryczną do zakodowania
2. Wybierz metodę kodowania:
   * "One-Hot Encoding" - tworzenie kolumn binarnych dla każdej kategorii
   * "Binary Encoding" - kodowanie na wartości numeryczne
   * "Target Encoding" - zastąpienie kategorii średnią wartością zmiennej celu
3. Jeśli wybierzesz "Target Encoding", z rozwijanej listy "Select Target Column" wybierz zmienną celu
4. Kliknij "Apply Encoding" aby wykonać operację
5. W nowym oknie zostanie wyświetlona tabela po kodowaniu

**Krok 6:** Aby zapisać dane po kodowaniu, kliknij "Save Encoded Data".

**Wskazówki:**

* One-Hot Encoding zwiększa liczbę kolumn, tworząc nową kolumnę dla każdej unikalnej wartości
* Target Encoding jest przydatny w przygotowaniu danych do uczenia maszynowego
* Usuwanie duplikatów może znacznie zmniejszyć rozmiar zbioru danych, jeśli zawiera on wiele powtórzeń
* Strategie wypełniania braków danych (fill\_mean, fill\_median, fill\_mode) działają tylko na kolumnach numerycznych
* Po wykonaniu operacji kodowania, oryginalna kolumna kategoryczna pozostaje w zbiorze danych

### 3.2.6. Zapisywanie danych

**Krok 1:** Po wykonaniu operacji na danych, możesz zapisać wyniki do pliku CSV.

**Krok 2:** Kliknij odpowiedni przycisk zapisu, w zależności od wykonanej operacji:

* "Save Current Data" w zakładce "Data & Stats"
* "Save Subtable" w zakładce "Replacement & Subtable"
* "Save After Replacement" w zakładce "Replacement & Subtable"
* "Save Scaled Data" w zakładce "Scaling & Visualization"
* "Save Cleaned Data" w zakładce "Data Cleaning & Transformation"
* "Save After Duplicate Removal" w zakładce "Data Cleaning & Transformation"
* "Save Encoded Data" w zakładce "Data Cleaning & Transformation"

**Krok 3:** W oknie opcji zapisu:

1. Zaznacz opcję "Include Row Indices" jeśli chcesz dołączyć indeksy wierszy
2. Zaznacz opcję "Include Column Headers" jeśli chcesz dołączyć nagłówki kolumn
3. Kliknij "Save" aby kontynuować

**Krok 4:** W oknie dialogowym wyboru pliku:

1. Wybierz lokalizację zapisu
2. Wprowadź nazwę pliku
3. Kliknij "Save" aby zapisać plik

**Wskazówki:**

* Plik zostanie zapisany w formacie CSV
* Jeśli nazwa pliku nie kończy się rozszerzeniem .csv, zostanie ono dodane automatycznie
* Po udanym zapisie pojawi się komunikat potwierdzający
* Domyślnie, nagłówki kolumn są dołączane do pliku, a indeksy wierszy nie są dołączane

# Eksperymenty na danych

## Wykorzystane zbiory danych

**a) Adult Census Income – opis przygotował Dawid Olko**

* **Pochodzenie** – UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Adult>), oryginalnie zestaw powstał na bazie spisu powszechnego USA z 1994 r., zebrany przez U.S. Census Bureau.
* **Tematyka / problem** – predykcja rocznego dochodu gospodarstwa domowego (<=50K lub >50K) na podstawie danych demograficznych i zawodowych.
* **Rozmiar** – 48 842 rekordy (32 561 w zbiorze treningowym adult.data, 16 281 w zbiorze testowym adult.test), 14 atrybutów wejściowych + 1 cel (dochód).
* **Format danych** – pliki CSV z separatorem przecinkowym, zawierające wartości atrybutów bez nagłówków. Każdy wiersz odpowiada jednemu rekordowi.

**Struktura cech:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nazwa atrybutu** | **Opis** | **Typ danych** | **Wartości/Zakres** |
| age | Wiek w latach | Numeryczny (int) | 17-90 |
| workclass | Klasa zatrudnienia | Kategoryczny | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt | Waga finalna (reprezentacja w populacji) | Numeryczny | 12285-1490400 |
| education | Poziom wykształcenia | Kategoryczny | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| education-num | Liczba lat edukacji | Numeryczny (int) | 1-16 |
| marital-status | Stan cywilny | Kategoryczny | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| occupation | Zawód | Kategoryczny | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| relationship | Relacja rodzinna | Kategoryczny | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| race | Rasa | Kategoryczny | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | Płeć | Kategoryczny | Female, Male |
| capital-gain | Zysk kapitałowy | Numeryczny | 0-99999 |
| capital-loss | Strata kapitałowa | Numeryczny | 0-4356 |
| hours-per-week | Godziny pracy tygodniowo | Numeryczny (int) | 1-99 |
| native-country | Kraj pochodzenia | Kategoryczny | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |
| income | Dochód roczny (zmienna celu) | Kategoryczny | <=50K, >50K |

* **Problematyka jakości danych** – wartość ? oznacza brak (ok. 7% przypadków w kolumnach workclass, occupation i native-country), umiarkowana nierównowaga klas (~76% w kategorii <= 50K).
* **Zastosowanie w projekcie** – obliczanie statystyk, wizualizacja rozkładów, klasyfikacja przy użyciu Logistic Regression oraz segmentacja K-Means.

**b) Chronic Kidney Disease – opis przygotował Piotr Smoła**

* **Pochodzenie** – UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease>); dane kliniczne zebrane w 2015 roku w Apollo Hospital w Indiach przez zespół badawczy pod kierownictwem Dr. P. Soundarapandian (opublikowane przez Tavakkoli et al.).
* **Tematyka / problem** – diagnostyka i wczesne wykrywanie przewlekłej choroby nerek (klasyfikacja: ckd vs notckd).
* **Rozmiar** – 400 rekordów, 24 atrybuty + etykieta classification.
* **Format danych** – pojedynczy plik CSV z separatorem przecinkowym, zawierający nagłówki kolumn oraz wartości atrybutów dla każdego pacjenta.

**Struktura cech:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nazwa atrybutu** | **Opis** | **Typ danych** | **Wartości/Zakres** | **% braków danych** |
| age | Wiek w latach | Numeryczny | 2-90 | 9% |
| bp | Ciśnienie krwi (mm/Hg) | Numeryczny | 50-180 | 12% |
| sg | Ciężar właściwy moczu | Kategoryczny | 1.005, 1.010, 1.015, 1.020, 1.025 | 11% |
| al | Albumina w moczu | Kategoryczny | 0, 1, 2, 3, 4, 5 | 11% |
| su | Cukier w moczu | Kategoryczny | 0, 1, 2, 3, 4, 5 | 11% |
| rbc | Czerwone krwinki w moczu | Kategoryczny | normal, abnormal | 19% |
| pc | Komórki ropne w moczu | Kategoryczny | normal, abnormal | 16% |
| pcc | Skupiska komórek ropnych | Kategoryczny | present, notpresent | 12% |
| ba | Bakterie w moczu | Kategoryczny | present, notpresent | 12% |
| bgr | Poziom cukru we krwi (mmoL/L) | Numeryczny | 22-490 | 12% |
| bu | Mocznik we krwi (mmoL/L) | Numeryczny | 1.5-391 | 12% |
| sc | Kreatynina w surowicy (mmoL/L) | Numeryczny | 0.4-76 | 12% |
| sod | Sód w surowicy (mEq/L) | Numeryczny | 4.5-163 | 19% |
| pot | Potas w surowicy (mEq/L) | Numeryczny | 2.5-47 | 19% |
| hemo | Hemoglobina (g/dL) | Numeryczny | 3.1-17.8 | 12% |
| pcv | Hematokryt | Numeryczny | 9-54 | 19% |
| wc | Liczba białych krwinek (x1000/mm³) | Numeryczny | 2200-26400 | 19% |
| rc | Liczba czerwonych krwinek (miliony/mm³) | Numeryczny | 2.1-8 | 19% |
| htn | Nadciśnienie tętnicze | Kategoryczny | yes, no | 12% |
| dm | Cukrzyca | Kategoryczny | yes, no | 12% |
| cad | Choroba wieńcowa | Kategoryczny | yes, no | 12% |
| appet | Apetyt | Kategoryczny | good, poor | 12% |
| pe | Obrzęki obwodowe | Kategoryczny | yes, no | 12% |
| ane | Anemia | Kategoryczny | yes, no | 12% |
| classification | Etykieta klasy (zmienna celu) | Kategoryczny | ckd, notckd | 0% |

* **Problematyka jakości danych** – braki dochodzące do 19% w kolumnach rbc, sod, pot, pcv, wc i rc. Umiarkowana nierównowaga klas (250 przypadków ckd, 150 przypadków notckd).
* **Zastosowanie w projekcie** – testowanie procedur czyszczenia danych (usuwanie braków, imputacja), skalowanie zmiennych i eksperymenty z wizualizacją (heatmapy, boxploty).

## Przebieg eksperymentów i wyniki

**Eksperyment 1: Analiza wpływu standaryzacji na rozkład zmiennej age w zbiorze Adult Dataset (Dawid Olko)**

**Cel eksperymentu:**

Zbadanie wpływu standaryzacji na rozkład zmiennej age w zbiorze danych Adult Dataset oraz ocena jej przydatności do dalszego modelowania.

**Metodologia:**

1. Wczytanie zbioru danych Adult Dataset
2. Obliczenie podstawowych statystyk dla zmiennej age przed transformacją
3. Wizualizacja rozkładu zmiennej age za pomocą histogramu
4. Zastosowanie StandardScaler do zmiennej age
5. Obliczenie statystyk dla znormalizowanej zmiennej
6. Wizualizacja rozkładu znormalizowanej zmiennej
7. Porównanie wyników

**Przebieg eksperymentu:**

1. Wczytałem zbiór Adult Dataset używając opcji "Predefined Dataset" w zakładce "Data & Stats".
2. Obliczyłem podstawowe statystyki dla nieprzetworzonej zmiennej age klikając przycisk "Compute Stats". Otrzymałem następujące wartości:

|  |  |
| --- | --- |
| **Statystyka** | **Wartość dla age** |
| Min | 17.0 |
| Max | 90.0 |
| Mean | 38.58 |
| Median | 37.0 |
| Std | 13.64 |
| Variance | 186.06 |
| Skewness | 0.56 |
| Kurtosis | -0.17 |

1. Wygenerowałem histogram dla zmiennej age, który pokazał rozkład z lekką asymetrią prawostronną.
2. Przeszedłem do zakładki "Scaling & Visualization" i zastosowałem StandardScaler do zmiennej age.
3. Obliczyłem statystyki dla znormalizowanej zmiennej:

|  |  |
| --- | --- |
| **Statystyka** | **Wartość dla znormalizowanej age** |
| Min | -1.58 |
| Max | 3.77 |
| Mean | 0.00 |
| Median | -0.12 |
| Std | 1.00 |
| Variance | 1.00 |
| Skewness | 0.56 |
| Kurtosis | -0.17 |

1. Wygenerowałem histogram dla znormalizowanej zmiennej age.

**Wyniki:**

Porównanie rozkładów przed i po standaryzacji pokazało następujące zależności:

1. Średnia zmiennej została przesunięta z 38.58 do 0.0
2. Odchylenie standardowe zostało zmienione z 13.64 do 1.0
3. Kształt rozkładu (asymetria i kurtoza) pozostał niezmieniony
4. Znormalizowane wartości zawierają się w przybliżonym zakresie [-2, 4]

Histogram dla znormalizowanej zmiennej age wykazał ten sam kształt rozkładu co pierwotna zmienna, ale z wartościami na osi X zmienionymi zgodnie z przekształceniem standaryzacji.

**Wnioski:**

Standaryzacja zmiennej age nie zmieniła charakterystyki jej rozkładu, ale umożliwiła wyrażenie wartości w jednostkach odchylenia standardowego od średniej. Jest to korzystne dla algorytmów uczenia maszynowego, które są wrażliwe na skalę danych, takich jak algorytmy oparte na odległościach (np. k-najbliższych sąsiadów) czy metody wykorzystujące gradientowe metody optymalizacji.

Zmienne znormalizowane ułatwiają również interpretację obserwacji odstających - wartości powyżej 2 lub poniżej -2 są potencjalnymi outlierami, co odpowiada wartościom age poniżej ok. 11 lat i powyżej ok. 66 lat w oryginalnej skali.

**Eksperyment 2: Analiza wpływu metody obsługi wartości brakujących na statystyki zmiennych w zbiorze Chronic Kidney Disease (Piotr Smoła)**

**Cel eksperymentu:**

Zbadanie wpływu różnych strategii obsługi wartości brakujących na statystyki zmiennych numerycznych w zbiorze Chronic Kidney Disease oraz porównanie skuteczności tych metod.

**Metodologia:**

1. Wczytanie zbioru danych Chronic Kidney Disease
2. Obliczenie podstawowych statystyk dla zmiennych numerycznych przed przetwarzaniem
3. Zastosowanie trzech strategii obsługi wartości brakujących:
   * Usunięcie wierszy z brakującymi wartościami (remove)
   * Wypełnienie braków średnią (fill\_mean)
   * Wypełnienie braków medianą (fill\_median)
4. Obliczenie statystyk po zastosowaniu każdej ze strategii
5. Porównanie wyników

**Przebieg eksperymentu:**

1. Wczytałem zbiór Chronic Kidney Disease używając opcji "Predefined Dataset" w zakładce "Data & Stats".
2. Obliczyłem podstawowe statystyki dla oryginalnego zbioru danych, skupiając się na trzech zmiennych z największą liczbą braków: 'hemo', 'pcv' i 'rc'.
3. Przeszedłem do zakładki "Data Cleaning & Transformation" i zastosowałem trzy różne strategie obsługi wartości brakujących.
4. Po każdej strategii obliczałem statystyki dla wybranych zmiennych.

**Wyniki:**

**Statystyki przed przetwarzaniem:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Zmienna** | **Liczba obserwacji** | **Min** | **Max** | **Mean** | **Median** | **Std** |
| hemo | 352 | 3.1 | 17.8 | 12.53 | 12.8 | 2.67 |
| pcv | 324 | 9.0 | 54.0 | 38.88 | 40.0 | 7.18 |
| rc | 324 | 2.1 | 8.0 | 4.65 | 4.7 | 0.87 |

**Statystyki po usunięciu wierszy z brakującymi wartościami (remove):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Zmienna** | **Liczba obserwacji** | **Min** | **Max** | **Mean** | **Median** | **Std** |
| hemo | 248 | 3.1 | 17.8 | 12.31 | 12.45 | 2.83 |
| pcv | 248 | 9.0 | 54.0 | 37.95 | 38.0 | 7.58 |
| rc | 248 | 2.1 | 8.0 | 4.58 | 4.6 | 0.92 |

**Statystyki po wypełnieniu braków średnią (fill\_mean):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Zmienna** | **Liczba obserwacji** | **Min** | **Max** | **Mean** | **Median** | **Std** |
| hemo | 400 | 3.1 | 17.8 | 12.53 | 12.8 | 2.50 |
| pcv | 400 | 9.0 | 54.0 | 38.88 | 40.0 | 6.47 |
| rc | 400 | 2.1 | 8.0 | 4.65 | 4.7 | 0.78 |

**Statystyki po wypełnieniu braków medianą (fill\_median):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Zmienna** | **Liczba obserwacji** | **Min** | **Max** | **Mean** | **Median** | **Std** |
| hemo | 400 | 3.1 | 17.8 | 12.57 | 12.8 | 2.51 |
| pcv | 400 | 9.0 | 54.0 | 39.10 | 40.0 | 6.46 |
| rc | 400 | 2.1 | 8.0 | 4.66 | 4.7 | 0.78 |

**Wnioski:**

1. **Strategia usuwania wierszy z brakującymi wartościami (remove)** zmniejszyła liczbę obserwacji o 38% (z 400 do 248), co jest znaczącą redukcją danych. Zaobserwowałem niewielkie zmiany w średnich wartościach zmiennych (np. średnia 'hemo' spadła z 12.53 do 12.31), co może sugerować, że wartości brakujące nie są całkowicie losowe (MCAR), a raczej zależą od wartości samych zmiennych (MAR).
2. **Strategia wypełniania braków średnią (fill\_mean)** zachowała wszystkie obserwacje i nie zmieniła średnich wartości zmiennych, co jest oczywiste, gdyż brakujące wartości zostały zastąpione właśnie średnią. Zauważalny jest jednak spadek odchylenia standardowego dla wszystkich trzech zmiennych, co wskazuje na sztuczne zmniejszenie zmienności danych.
3. **Strategia wypełniania braków medianą (fill\_median)** również zachowała wszystkie obserwacje, ale spowodowała niewielkie zmiany w średnich wartościach (np. średnia 'pcv' wzrosła z 38.88 do 39.10). Podobnie jak w przypadku strategii fill\_mean, odchylenie standardowe uległo zmniejszeniu.

Podsumowując, każda z metod obsługi wartości brakujących ma swoje zalety i wady. Usuwanie wierszy pozwala zachować naturalną zmienność danych, ale kosztem utraty informacji. Imputacja średnią lub medianą zachowuje liczbę obserwacji, ale zmniejsza zmienność i może wprowadzać sztuczne wzorce w danych. Wybór metody powinien zależeć od specyfiki problemu, mechanizmu powstawania braków oraz planowanych dalszych analiz.

## Analiza uzyskanych wyników i wnioski

**Analiza wyników eksperymentu z Adult Dataset (Dawid Olko)**

Przeprowadzony eksperyment standaryzacji zmiennej 'age' w zbiorze Adult Dataset dostarczył interesujących obserwacji dotyczących transformacji danych i jej wpływu na rozkład zmiennej.

**Kluczowe obserwacje:**

1. **Zmiana skali przy zachowaniu kształtu rozkładu**: Standaryzacja skutecznie przekształciła zmienną 'age' do rozkładu o średniej 0 i odchyleniu standardowym 1, zachowując przy tym oryginalny kształt rozkładu. Miary asymetrii (skewness) i spłaszczenia (kurtosis) pozostały niezmienione, co potwierdza, że standaryzacja jest transformacją liniową, która nie wpływa na kształt rozkładu.
2. **Identyfikacja obserwacji odstających**: Po standaryzacji wyraźnie widoczne stały się obserwacje odstające. Wartości poniżej -2 i powyżej 2 na znormalizowanej skali odpowiadały odpowiednio osobom bardzo młodym (poniżej ~11 lat) i starszym (powyżej ~66 lat). Standaryzacja ułatwiła ich identyfikację bez konieczności szczegółowej znajomości oryginalnej skali zmiennej.
3. **Interpretacja w kontekście populacji**: Znormalizowane wartości można interpretować w odniesieniu do populacji - wartość 0 oznacza wiek dokładnie na poziomie średniej (38.58 lat), wartość 1 oznacza wiek o jedno odchylenie standardowe powyżej średniej (ok. 52.22 lat), itd. Ta interpretacja jest intuicyjna i ułatwia porównywanie różnych obserwacji.

**Wnioski:**

1. **Przydatność w modelowaniu**: Standaryzacja zmiennej 'age' będzie szczególnie przydatna w modelach wrażliwych na skalę danych, takich jak regresja logistyczna, SVM czy sieci neuronowe. Zmienne o jednolitej skali zazwyczaj prowadzą do szybszej zbieżności algorytmów optymalizacji i bardziej stabilnych wyników.
2. **Porównanie z innymi zmiennymi**: Dzięki standaryzacji można bezpośrednio porównywać względny wpływ wieku z innymi znormalizowanymi zmiennymi w modelu, co byłoby trudne przy oryginalnych skalach z powodu różnych jednostek i zakresów wartości.
3. **Zalecenia dla dalszych analiz**: Dla pełnego przygotowania danych do modelowania, należałoby zastosować podobną standaryzację do pozostałych zmiennych numerycznych w zbiorze, takich jak 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss' i 'hours-per-week'. Dla zmiennych kategorycznych odpowiednie byłoby zastosowanie kodowania one-hot lub innej metody kodowania.

Podsumowując, standaryzacja okazała się skuteczną metodą transformacji zmiennej 'age', przygotowując ją do dalszego wykorzystania w modelach predykcyjnych przy jednoczesnym zachowaniu jej statystycznych właściwości. Eksperyment ten pokazał praktyczne zastosowanie funkcji skalowania zaimplementowanej w aplikacji DataFusion oraz jej użyteczność w przygotowaniu danych do analizy.

**Analiza wyników eksperymentu z Chronic Kidney Disease (Piotr Smoła)**

Przeprowadzony eksperyment porównujący różne strategie obsługi wartości brakujących w zbiorze Chronic Kidney Disease dostarczył cennych informacji na temat wpływu tych metod na statystyki opisowe i strukturę danych.

**Kluczowe obserwacje:**

1. **Wpływ na liczebność próby**: Strategia usuwania wierszy z brakującymi wartościami (remove) spowodowała utratę 38% obserwacji (z 400 do 248), co jest znaczącą redukcją w i tak niewielkim zbiorze danych. To może prowadzić do utraty istotnych informacji i potencjalnie wpływać na reprezentatywność próby.
2. **Zmiany w statystykach opisowych**:
   * Przy strategii **remove** zauważyłem niewielkie, ale zauważalne zmiany w średnich wartościach zmiennych (np. średnia 'hemo' spadła z 12.53 do 12.31), co sugeruje, że braki danych nie są całkowicie losowe.
   * Strategia **fill\_mean** zachowała średnie wartości (co jest oczywiste, gdyż brakujące wartości zostały zastąpione średnią), ale zmniejszyła odchylenie standardowe dla wszystkich badanych zmiennych.
   * Strategia **fill\_median** spowodowała niewielkie zmiany w średnich i także zmniejszyła zmienność danych, choć w nieco innym stopniu niż fill\_mean.
3. **Mechanizm powstawania braków**: Obserwowane zmiany w statystykach po usunięciu wierszy z brakującymi wartościami sugerują, że braki w danych nie są całkowicie losowe (MCAR - Missing Completely At Random), a raczej zależą od wartości samych zmiennych (MAR - Missing At Random) lub nawet od brakujących wartości (MNAR - Missing Not At Random). To ważna obserwacja, która powinna wpływać na wybór metody imputacji.

**Wnioski:**

1. **Kompromis między kompletnością a naturalną zmiennością**: Każda z badanych metod przedstawia kompromis - usuwanie wierszy zachowuje naturalną zmienność kosztem utraty danych, podczas gdy metody imputacji zachowują wszystkie obserwacje, ale mogą zniekształcać rozkład i zmniejszać zmienność.
2. **Zastosowanie w diagnostyce medycznej**: W kontekście danych medycznych, takich jak Chronic Kidney Disease, utrata 38% obserwacji może być problematyczna, szczególnie jeśli braki nie są losowe. Z drugiej strony, sztuczne wprowadzanie wartości może prowadzić do błędnych diagnoz, jeśli model będzie trenowany na takich danych.
3. **Zalecenia dla dalszych analiz**:
   * Dla tego konkretnego zbioru danych rekomendowałbym zastosowanie bardziej zaawansowanych technik imputacji, takich jak KNN lub algorytmy oparte na drzewach decyzyjnych, które mogą lepiej zachować relacje między zmiennymi.
   * Warto również rozważyć wykorzystanie samego faktu braku danych jako dodatkowej informacji w modelowaniu, tworząc zmienne binarne oznaczające obecność lub brak wartości dla danej cechy.
   * Przed zastosowaniem modeli predykcyjnych należałoby przeprowadzić analizę wrażliwości, sprawdzając jak różne metody imputacji wpływają na wyniki klasyfikacji.
4. **Uwagi dotyczące implementacji w DataFusion**: Aplikacja DataFusion skutecznie zaimplementowała różne strategie obsługi wartości brakujących, umożliwiając ich porównanie. Warto rozważyć rozszerzenie funkcjonalności o bardziej zaawansowane metody imputacji w przyszłych wersjach.

Podsumowując, eksperyment pokazał, że wybór metody obsługi wartości brakujących ma znaczący wpływ na charakterystykę danych i powinien być dokonywany świadomie, z uwzględnieniem specyfiki problemu, mechanizmu powstawania braków oraz planowanych dalszych analiz. W przypadku danych medycznych, gdzie każda obserwacja może być krytyczna, a jednocześnie kompletność i poprawność danych są kluczowe, odpowiednia strategia obsługi wartości brakujących może mieć istotny wpływ na jakość modeli diagnostycznych.

# Literatura

Rząsa, W. (2025). *Inteligentne metody eksploracji baz danych: Pojęcia podstawowe*. Materiały wykładowe, Uniwersytet Rzeszowski.

Rząsa, W. (2025). *Eksploracja tekstu*. Materiały wykładowe, Uniwersytet Rzeszowski.

Hand, D., Mannila, H., Smith, P. (2005). *Eksploracja danych*. WNT, Warszawa.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., Kurhan, L. A. (2007). *Data Mining: A Knowledge Discovery Approach*. Springer.

Hwang, C. L., Yoon, K. (1981). *Multiple Attribute Decision Making: Methods and Applications*. Springer-Verlag, Berlin.