Odkrywanie ukrytych struktur i wizualizacja danych wielowymiarowych

Analiza zbioru Titanic metodą MDS

Dominika Szulc, Wiktoria Jarząb

2025-04-03

Spis treści

1 Cel analizy

W poniższym raporcie zbadamy wielowymiarową strukturę danych dotyczących pasażerów Titanica oraz zidentyfikujemy naturalne skupiska (grupy) pasażerów. Analiza zostanie przeprowadzona w sposób nienadzorowany (bez wykorzystania zmiennej informującej o ich ocaleniu), aby sprawdzić, czy istnieją ukryte wzorce, od których zależało przetrwanie pasażerów.

W tym celu zastosujemy metodę skalowania wielowymiarowego (MDS) do redukcji wymiarowości danych ze zbioru `titanic_train`. Pozwoli to na odwzorowanie wielowymiarowej przestrzeni cech na płaszczyźnie (2D) przy zachowaniu oryginalnych podobieństw (odległości) między pasażerami.

2 Dane `titanic_train`

2.1 Przygotowanie danych

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
```

```
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Dane `titanic_train` z pakietu `titanic` po wczytaniu wymagają pewnych modyfikacji - nie wszystkie typy zmiennych są poprawne, dlatego zamieniamy:

- `Sex` na typ `factor`,
- `Survivied` na typ `factor` z nowymi nazwami poziomów:
  - `Yes` dla 1,
  - `No` dla 0,
- `Embarked` na typ `factor`,
- `Pclass` na typ `factor`.

Tak przygotowane dane przypisujemy do zmiennej `titanic`. To na tym zbiorze będziemy wykonywać dalsze operacje.

2.2 Informacje o danych

Dane `titanic` mają 891 przypadków i 12 cech.

Tablica 1: Opis danych titanic

Nazwa Kolumny	Typ Danych	Opis Danych
PassengerId	ciągłe	ID pasażera
Survived	jakościowa	czy osoba przeżyła katastrofę
Pclass	jakościowa	numer klasy: 1, 2, 3
Name	character	imię i nazwisko pasażera
Sex	jakościowa	płeć
Age	ciągłe	wiek
SibSp	ciągłe	liczba rodzieństwa lub małżonek/ka na pokładzie
Parch	ciągłe	liczba rodziców lub dzieci na pokładzie
Ticket	character	numer biletu
Fare	ciągłe	opłata za pasażera
Cabin	character	numer kabiny
Embarked	jakościowa	port okrętowania

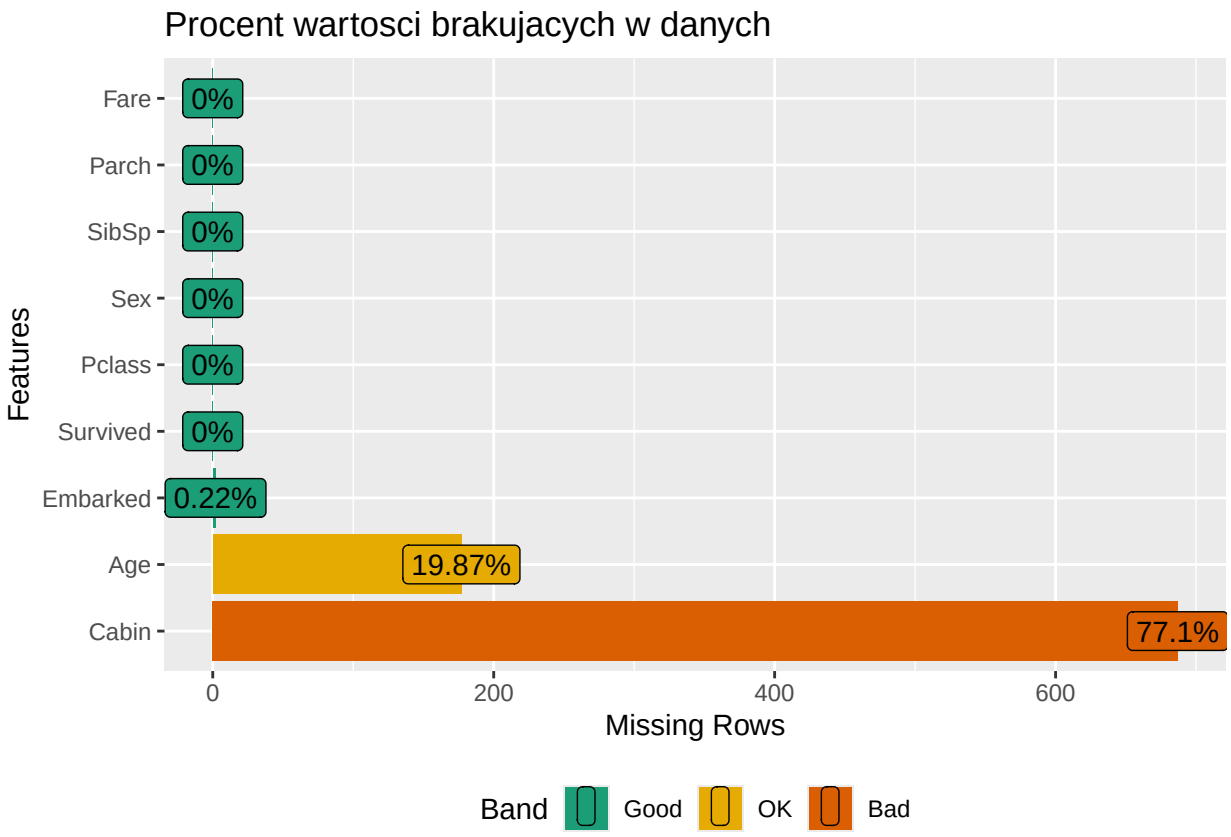
W tabeli **??** został przedstawiony opis poszczególnych zmiennych. Widać, że cechy `PassengerID`, `Ticket` oraz `Name` są danymi identyfikacyjnymi, zatem można je usunąć, bo nie będą przydatne w dalszej analizie.

2.2.1 Wartości brakujące

Nie wszystkie braki danych zostały poprawnie wczytane. Niektóre zapisano jako puste wiersze w kolumnie, dlatego zamieniamy je na standardowe kodowanie `NA`.

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## i The deprecated feature was likely used in the DataExplorer package.
##   Please report the issue at
##   <https://github.com/boxuancui/DataExplorer/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```

```
## generated.
```



Rysunek 1: Wykres przedstawiający procentowy udział wartości brakujących w danych titanic

Po przeanalizowaniu wykresu 1 zdecydowałyśmy się na usunięcie danych dotyczących kabiny - `Cabin`.

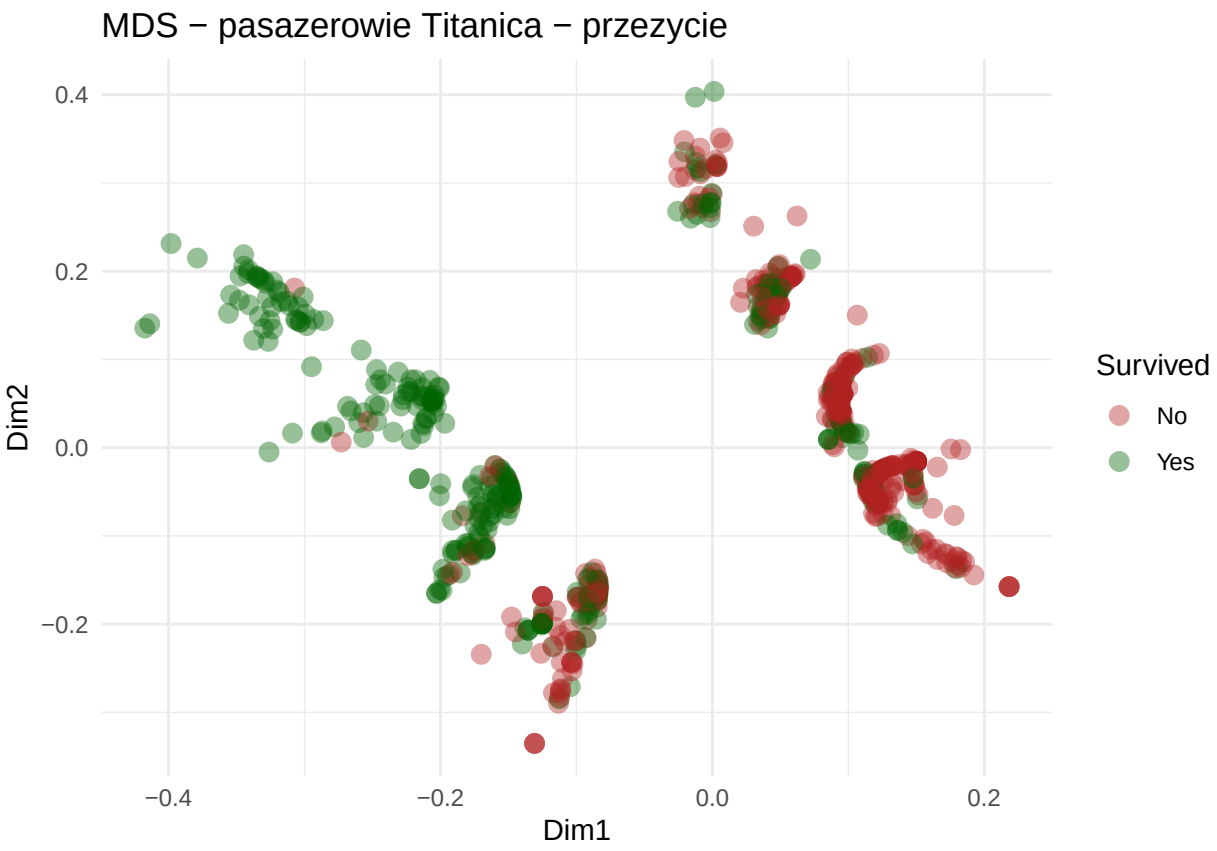
Ponadto warto zwrócić uwagę na zmienną `Age`, gdzie również znajduje się dużo wartości brakujących. W tym przypadku zdecydowałyśmy się jednak na zastąpienie ich, stosując metodę k-najbliższych sąsiadów.

```
##     SibSp     Parch      Fare     SibSp     Parch      Fare
##    0.0000    0.0000    0.0000    8.0000    6.0000  512.3292
```

3.2 Wizualizacja

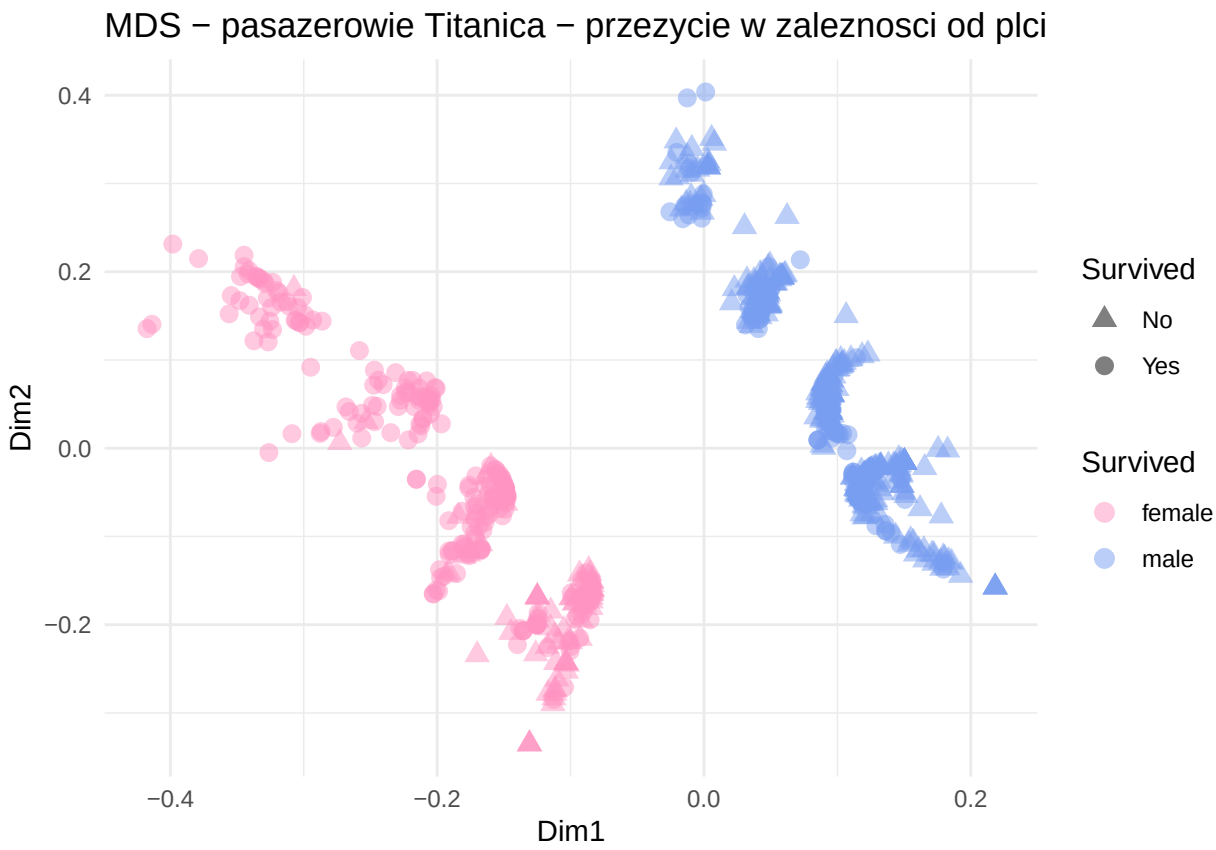
Korzystając z wyznaczonych danych `mds.tabela`, przyjrzymy się przeżyciu pasażerów rejsu. Zbadamy to w zależność od płci oraz klasy pasażerskiej. Otrzymane wyniki porównamy z oficjalnymi danymi dotyczącymi katastrofy.

3.2.1 Przeżycie



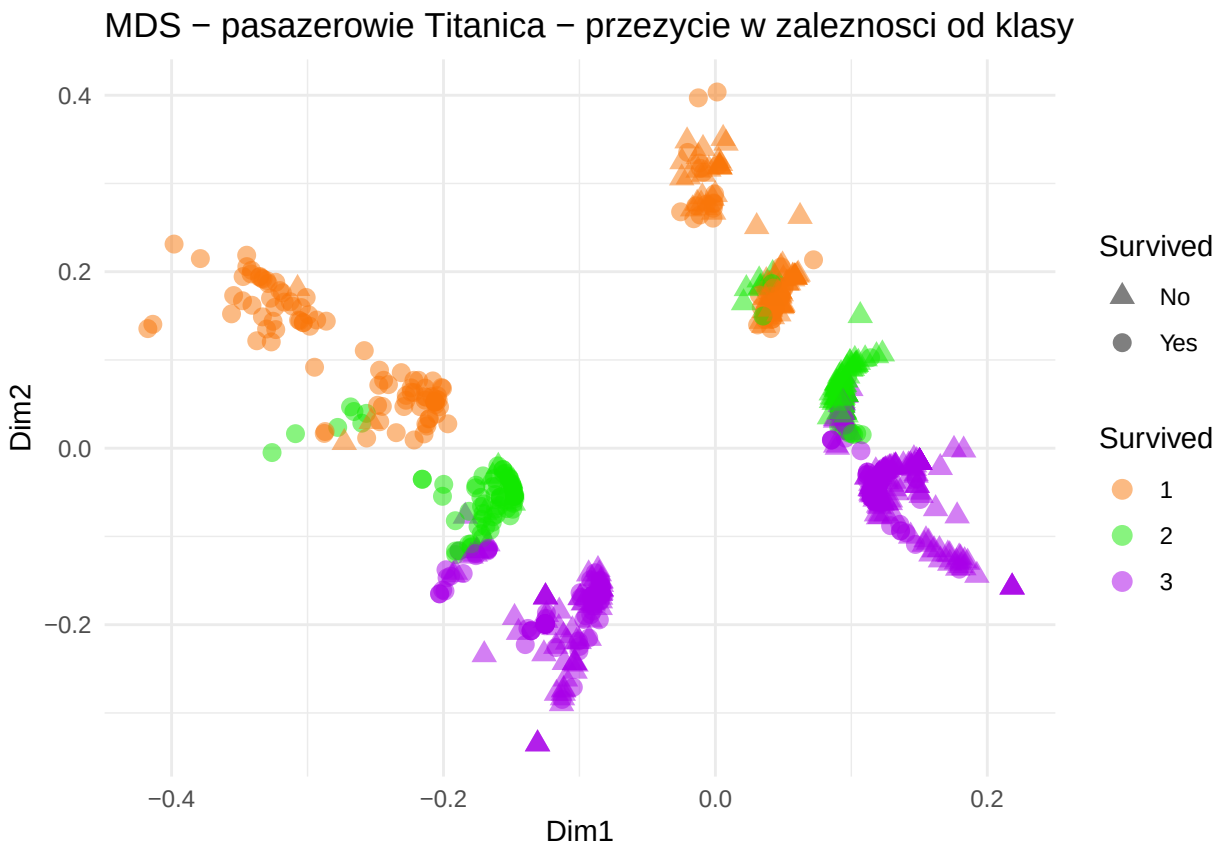
Rysunek 3: Wykres przedstawiający przeżycie pasażerów dla danych MDS

Na wykresie 3 mamy dwa skupiska punktów w dużej odległości od siebie. Każda z nich jest podzielona na mniejsze podgrupy. Brak obserwacji odstających. Ponadto mniej więcej połowa pasażerów zginęła w katastrofie. Jest to podobne do oficjalnych danych.



Rysunek 4: Wykres przedstawiający przeżycie pasażerów dla danych MDS w zależności od płci

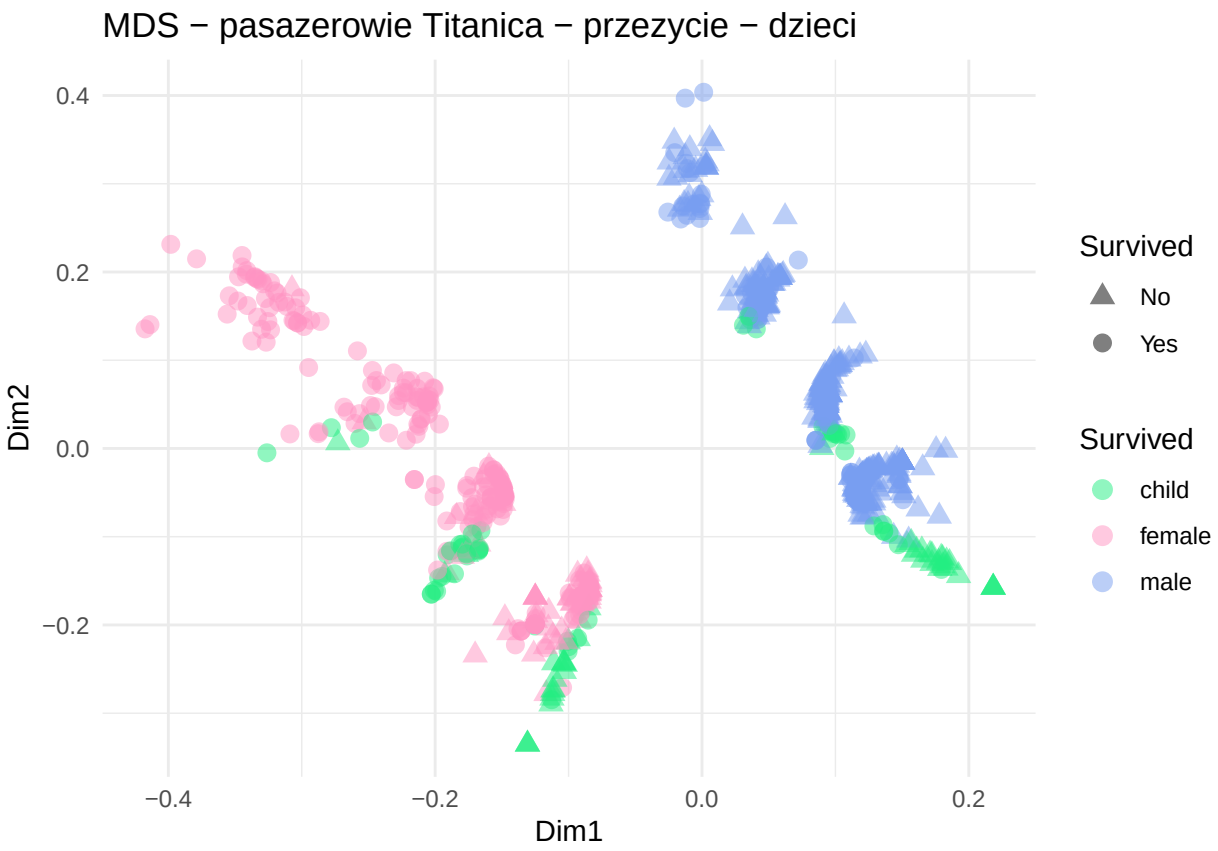
3.2.1.1 Płeć Zatem podział na dwa skupiska to podział ze względu na płeć (wykres 4), więc podgrupy to może być podział na klasy albo wiek. Zginęła zdecydowana większość mężczyzn (mniej więcej 75%), co jest zgodne z oficjalnymi danymi.



Rysunek 5: Wykres przedstawiający przeżycie pasażerów dla danych MDS w zależności od klasy pasażerskiej

3.2.1.2 Klasa Zatem podgrupy mogą być przybliżeniem podziału na klasy (wykres 5). Widać, że zginęli niemal wszyscy mężczyźni podróżujących trzecią klasą. Najwięcej przeżyło kobiet z pierwszej i drugiej klasy. Jest to zgodne z oficjalnymi danymi.

3.2.1.3 Dzieci Sprawdzimy, jaka część podróżujących to dzieci oraz ile z nich zginęło w katastrofie.



Rysunek 6: Wykres przedstawiający przeżycie pasażerów dla danych MDS dla dzieci (Age < 16)

Zdecydowana większość dzieci (wykres 6) podróżowała w trzeciej klasie. Przeżyły praktycznie wszystkie dziewczynki, w przeciwieństwie do chłopców. Śmierć poniosły głównie dzieci z trzeciej klasy. Jest to zgodne z oficjalnymi danymi.

4 Wnioski

Przeżycie pasażerów Titanica zależało od ich płci oraz klasy, którą podróżowali.

Większość kobiet przeżyła (wykres 4). Śmierć poniosły głównie pasażerki trzeciej klasy i niektóre z drugiej klasy (wykres 5).

Zginęła zdecydowana większość mężczyzn (wykres 4). Spośród wszystkich klas przeżyli jedynie nieliczni (wykres 5).

Wiek nie miał dużego wpływu na przeżycie. Około połowa wszystkich podróżujących dzieci straciła życie w katastrofie (wykres 6).