SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

|  |  |
| --- | --- |
| Laboratorium Nr 3  Data 25.10.2025  Temat: "Przygotowanie danych dla modelu predykcji"  Wariant 1 | Artur Rolak  Informatyka  II stopień, stacjonarne,  1 semestr, gr. 1b |

<https://github.com/rlNkoo/Studia-Magisterskie/tree/main/Nauka%20o%20Danych%20L>

Polecenie:

CEL ZADANIA

Przygotowanie i analiza zestawu danych do treningu modelu ML, z naciskiem na identyfikację i naprawę problemów w danych rzeczywistych.

DATASET

Wykorzystać dowolny dataset znaleziony na Kaggle poprzez kagglehub. **Wymagania minimalne:**

* Minimum 500 obserwacji (wierszy)
* Minimum 4 zmienne (kolumny)
* Dataset musi zawierać zarówno zmienne numeryczne jak i kategoryczne
* W sprawozdaniu podać: nazwę datasetu, źródło, liczbę wierszy i kolumn

WYMAGANIA ZADANIA

**CZĘŚĆ 1: Eksploracja i Raportowanie**

1.1. Analiza struktury danych:

* Identyfikacja typów zmiennych (numeryczne/kategoryczne/datetime)
* Statystyki opisowe dla każdej kolumny
* Macierz korelacji dla zmiennych numerycznych
* Analiza: jak zmienne wpływają na główną kolumnę (target) i na siebie nawzajem?
* Identyfikacja potencjalnych problemów multikolinearności

1.2. Analiza brakujących danych i duplikatów:

* Wizualizacja brakujących danych (heatmapa, bar plot)
* Klasyfikacja braków: MCAR/MAR/MNAR z uzasadnieniem
* Utworzenie DataFrame z procentem braków per kolumna
* Analiza duplikatów: procentowy stosunek ilości danych oryginalnych do duplikatów

**CZĘŚĆ 2: Czyszczenie i Naprawa**

2.1. Obsługa wartości brakujących:

* Imputacja median/mean dla zmiennych numerycznych
* Imputacja modą dla kategorycznych
* Usunięcie kolumn z >60% braków
* Uzasadnienie wyboru każdej strategii w komentarzach
* Porównanie różnych metod imputacji dla wybranej kolumny

2.2. Wykrywanie i obsługa outlierów:

* Zastosowanie metody IQR dla zmiennych numerycznych
* Opcjonalnie: metoda Z-score lub Isolation Forest
* Wizualizacja przed/po usunięciu outlierów (box plot, scatter plot)
* Uzasadniona decyzja: capping vs usunięcie
* Analiza wpływu outlierów na statystyki opisowe

2.3. Feature Engineering - utworzyć minimum 1 nową kolumnę:

* kolumny powinny logicznie wynikać z istniejących danych
* Przykłady: ratio dwóch kolumn, binning zmiennych ciągłych, agregacje, interakcje
* Uzasadnienie, dlaczego nowe kolumny mogą być wartościowe
* Analiza korelacji nowych kolumny z targetem

**CZĘŚĆ 3: Walidacja i Wizualizacja**

3.1. Kontrola jakości:

* Sprawdzenie i usunięcie duplikatów (z uzasadnieniem metody)
* Walidacja zakresów wartości (czy są sensowne dla domeny?)
* Test normalności rozkładu dla kluczowych zmiennych numerycznych (Shapiro-Wilk, Q-Q plot)
* Sprawdzenie spójności danych między kolumnami

3.2. Wizualizacje - wykonać minimum 4 wykresy:

* Distribution plot dla najważniejszych zmiennych (przed/po transformacji)
* Correlation heatmap (przed/po feature engineering)
* Box plot lub violin plot dla wykrywania outlierów
* Scatter plot lub pair plot dla relacji między zmiennymi
* **Wszystkie wykresy umieścić w sprawozdaniu z opisem wniosków**

1. Opis programu opracowanego (kody źródłowe, rzuty ekranu)

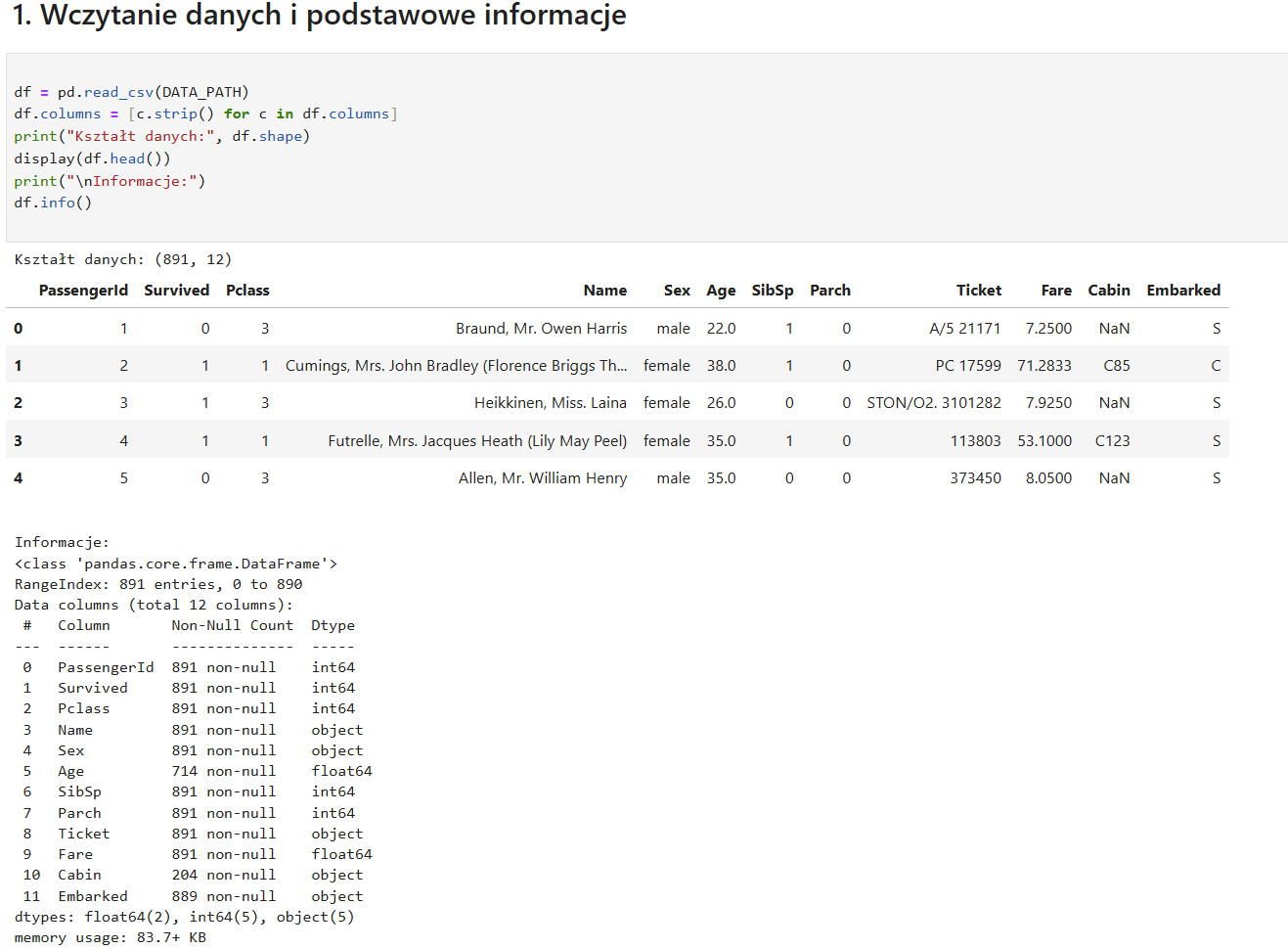
Celem opracowanego programu było **przygotowanie i analiza rzeczywistego zbioru danych** do uczenia modelu Machine Learning (ML) poprzez wykonanie eksploracji, czyszczenia i wstępnej walidacji danych.  
Program został zaimplementowany w języku **Python** w środowisku **Jupyter Notebook**, z wykorzystaniem bibliotek pandas, numpy oraz matplotlib.  
Zrealizowano pełny pipeline:

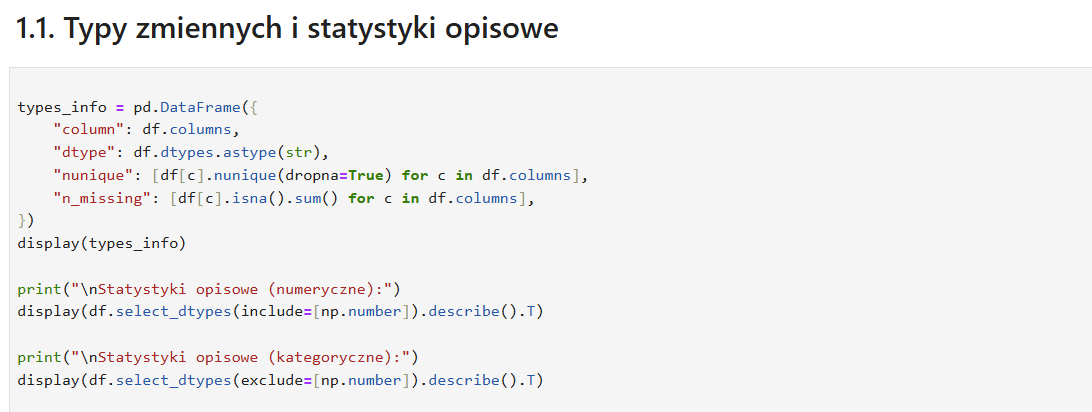
* analiza typów i statystyk,
* wykrywanie i imputacja braków,
* identyfikacja outlierów metodą IQR,
* tworzenie nowych cech (feature engineering),
* walidacja danych i wizualizacje.

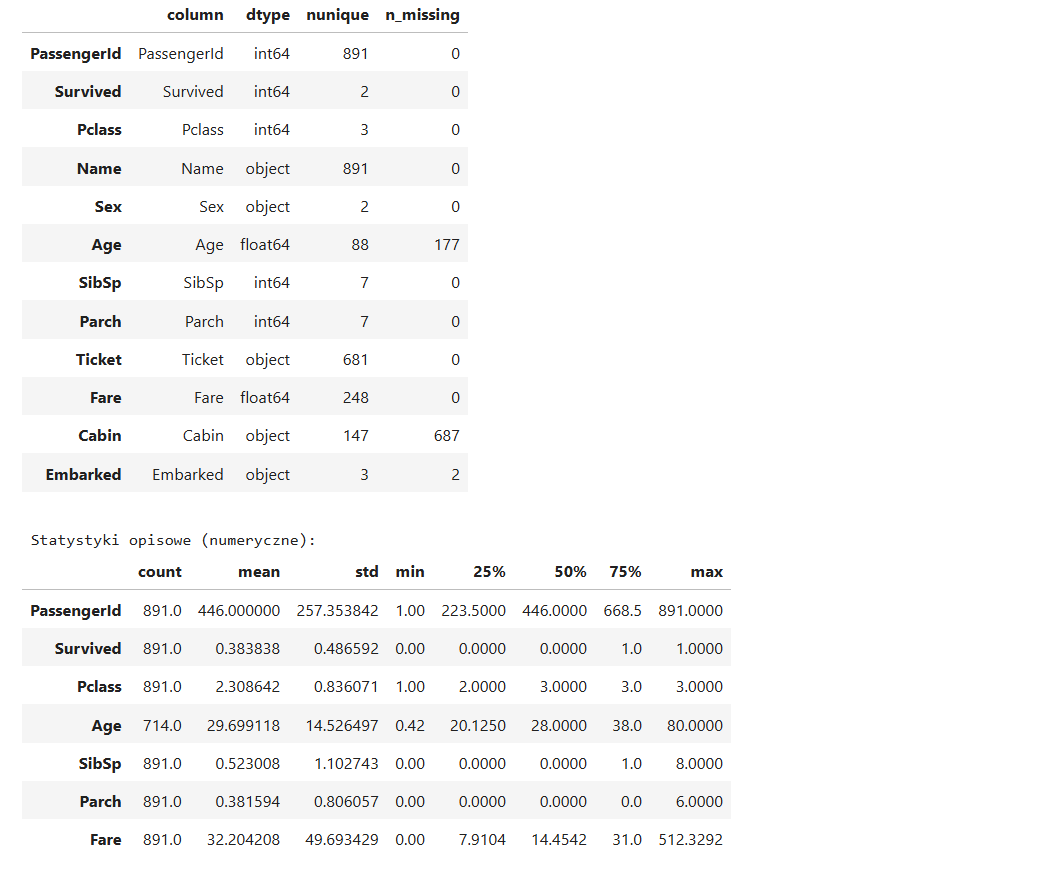
**Zbiór danych:** *Titanic Dataset*  
**Źródło:** Kaggle – Titanic: Machine Learning from Disaster  
**Liczba wierszy:** 891  
**Liczba kolumn (oryginalnie):** 12

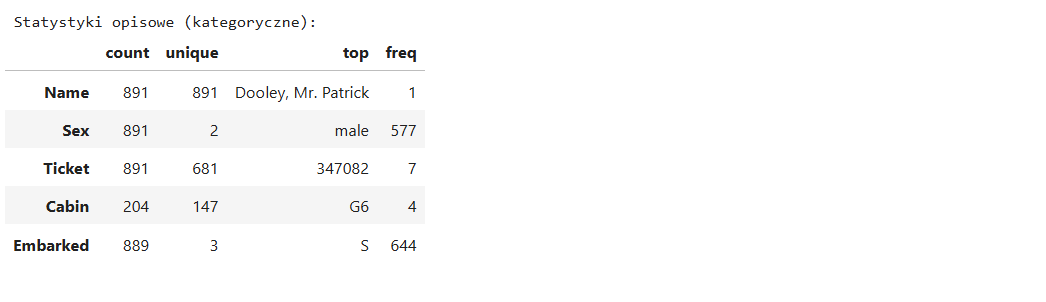
Zbiór zawiera informacje o pasażerach statku Titanic, takie jak wiek, płeć, klasa podróży, cena biletu, liczba członków rodziny na pokładzie oraz informację, czy dana osoba przeżyła katastrofę (Survived).



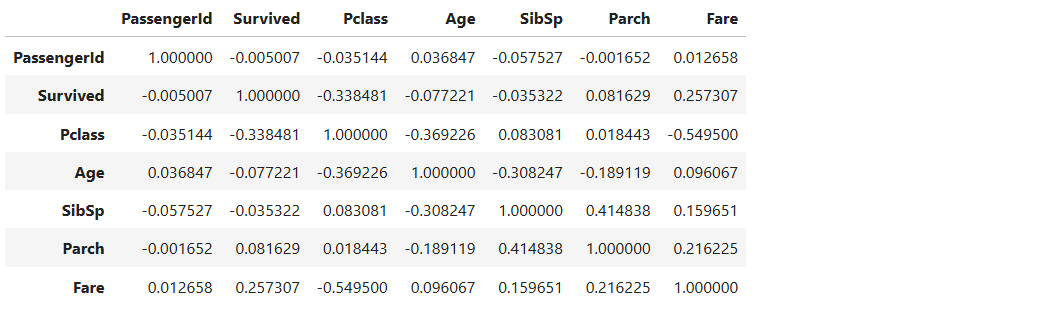


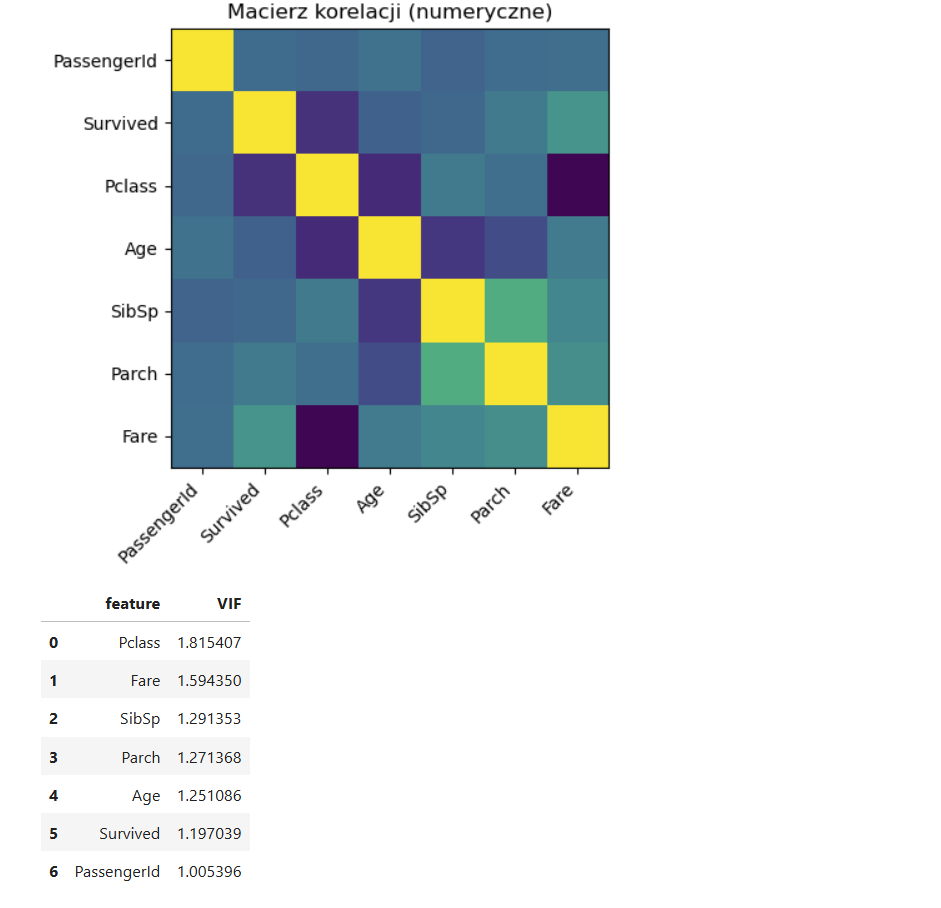






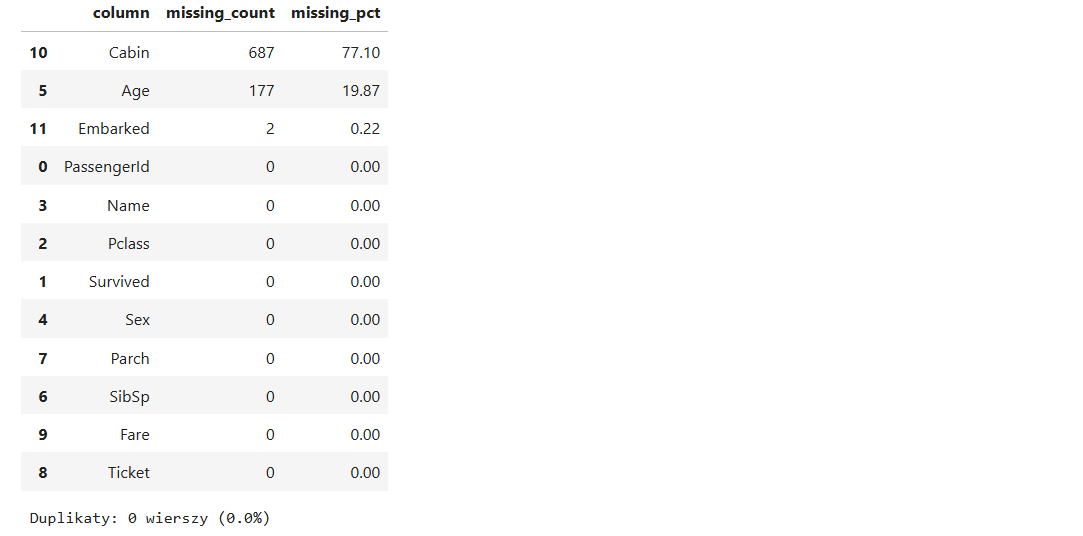


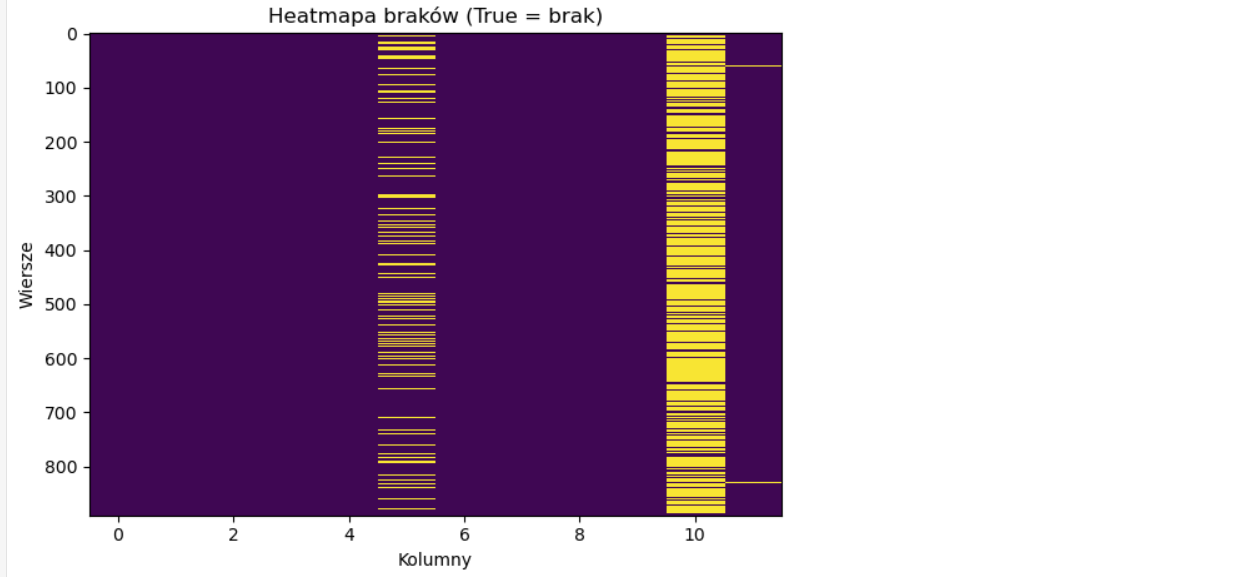


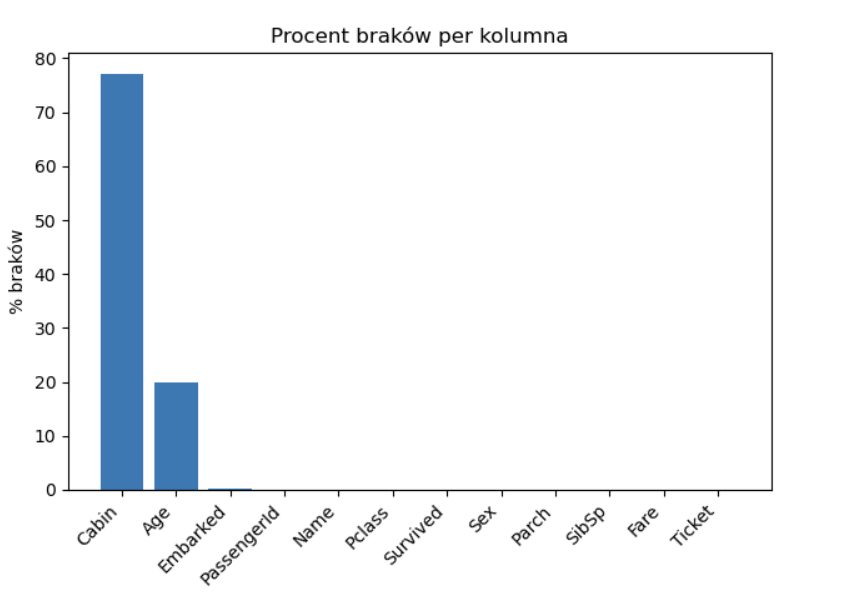


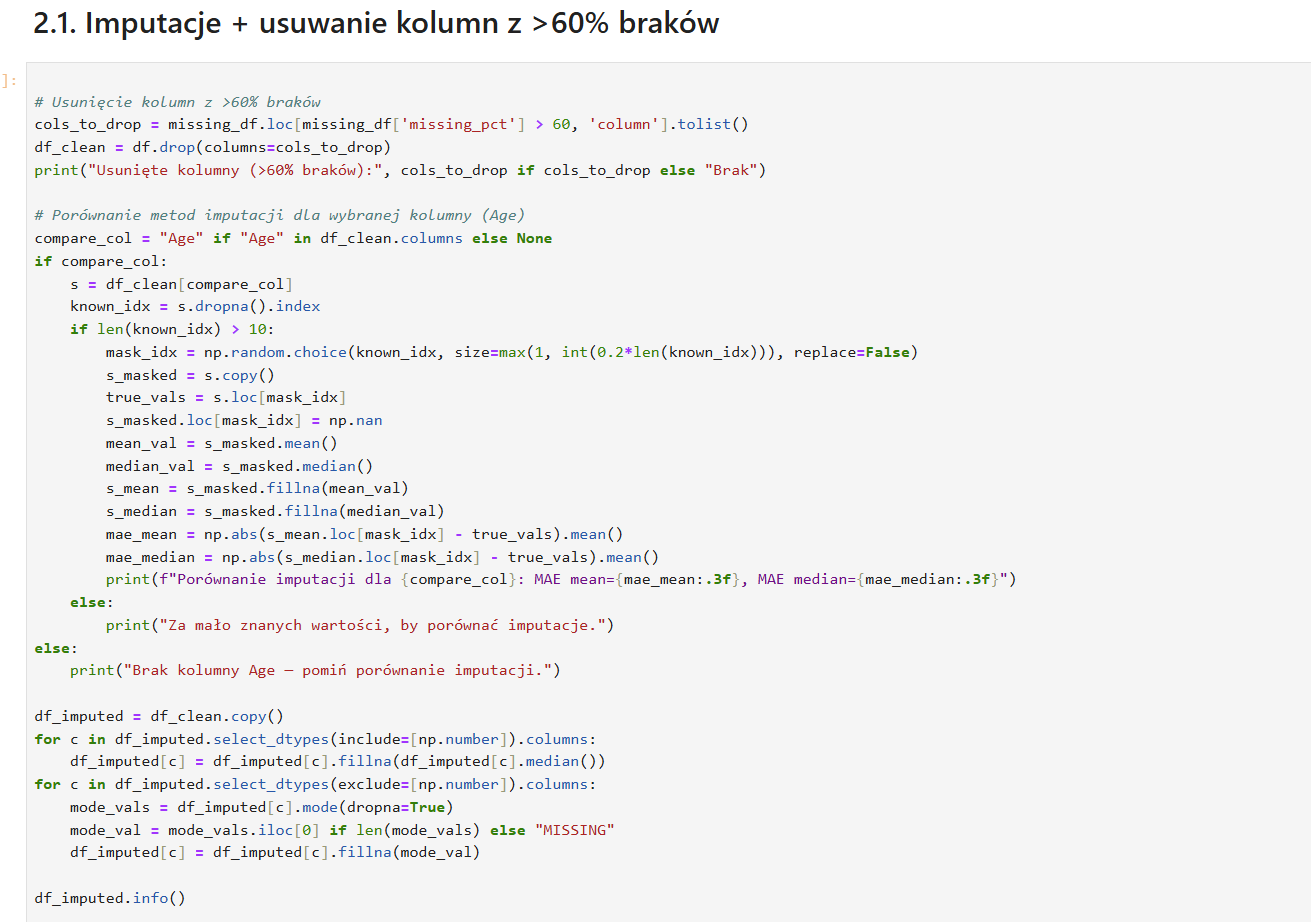


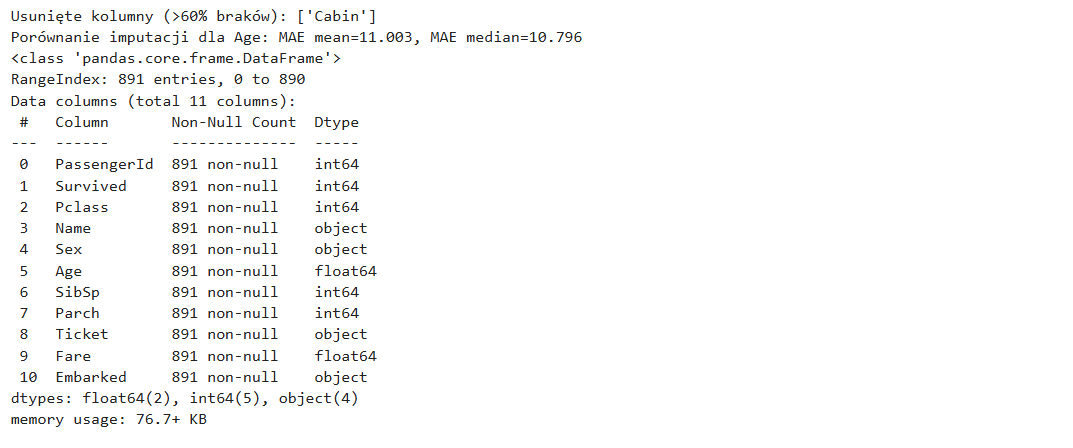


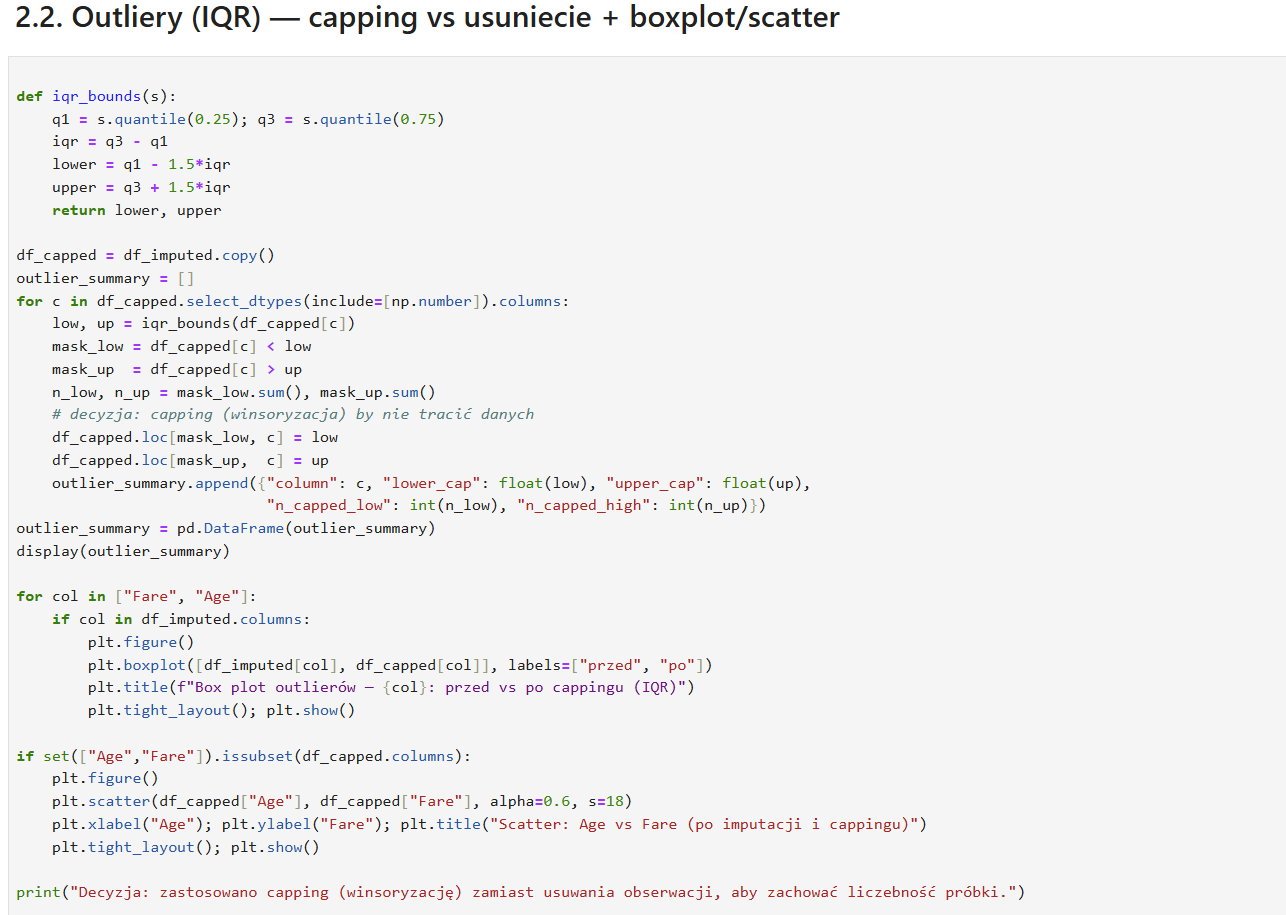


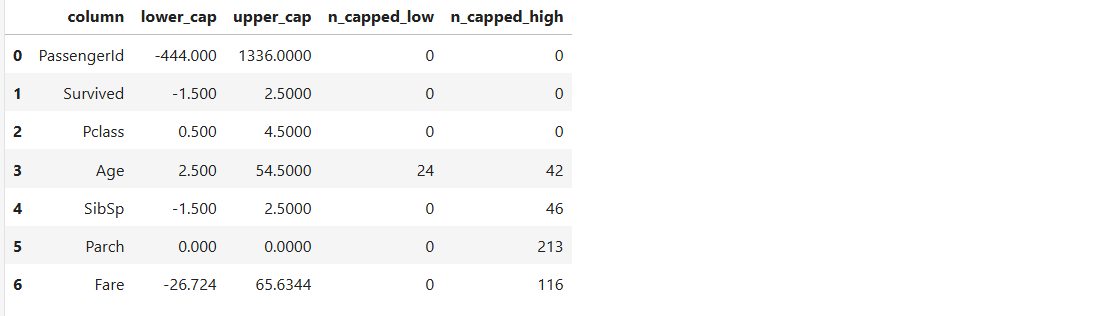


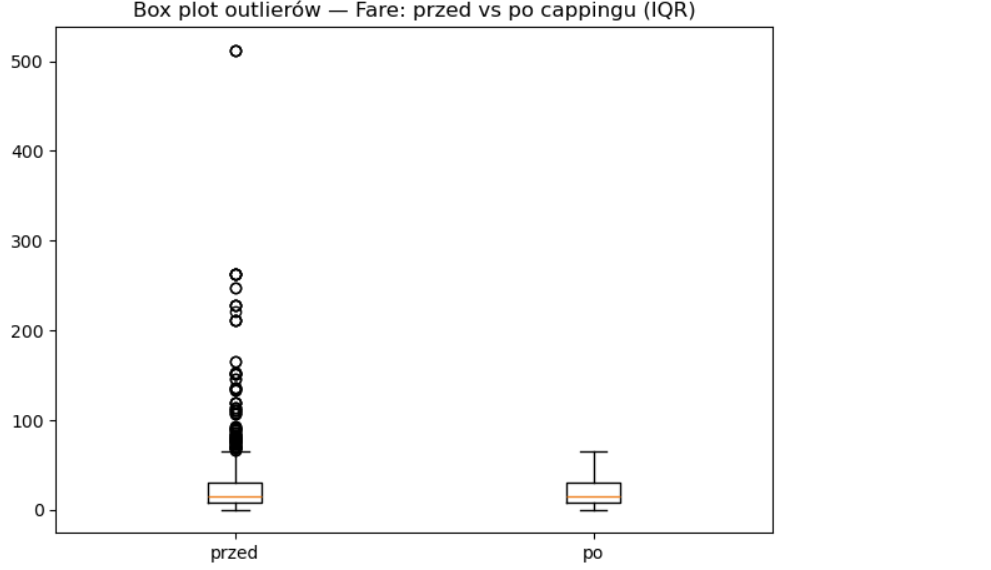






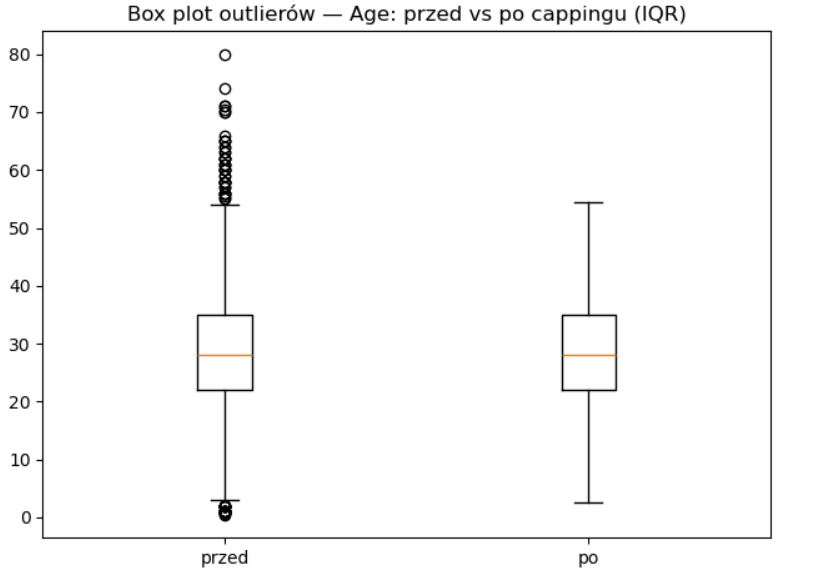


****

****

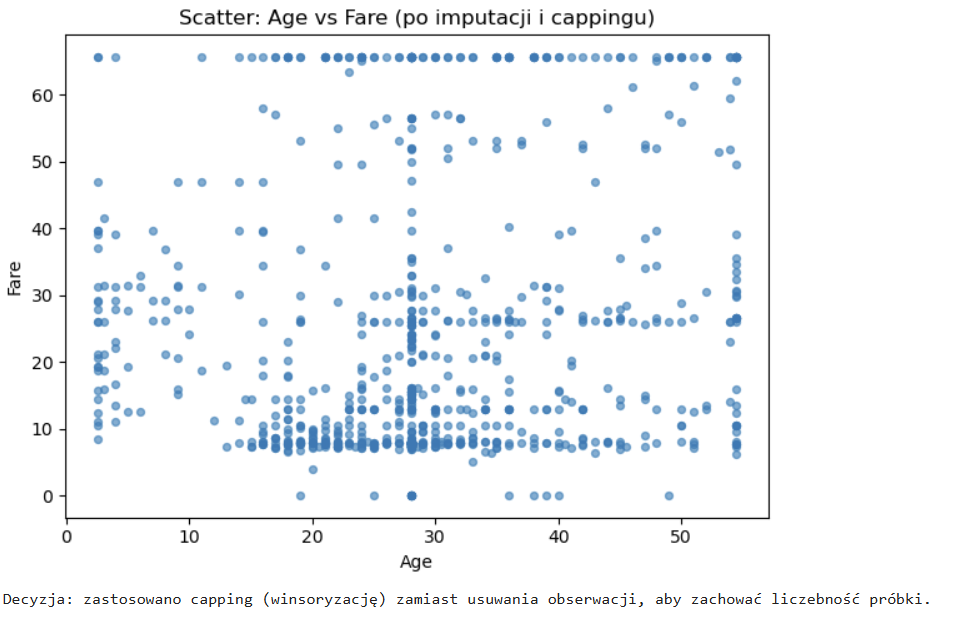
**Opis:**  
Wykres pudełkowy przedstawia rozkład zmiennej Fare (cena biletu) przed i po zastosowaniu metody IQR do ograniczenia wartości odstających.

**Wnioski:**  
Przed cappingiem występowały liczne wartości ekstremalne (powyżej 400). Po zastosowaniu winsoryzacji wartości zostały ograniczone do górnej granicy, dzięki czemu rozkład stał się bardziej równomierny. Nie utracono przy tym żadnych obserwacji.



**Opis:**  
Boxplot prezentuje rozkład wieku (Age) pasażerów przed i po ograniczeniu outlierów.

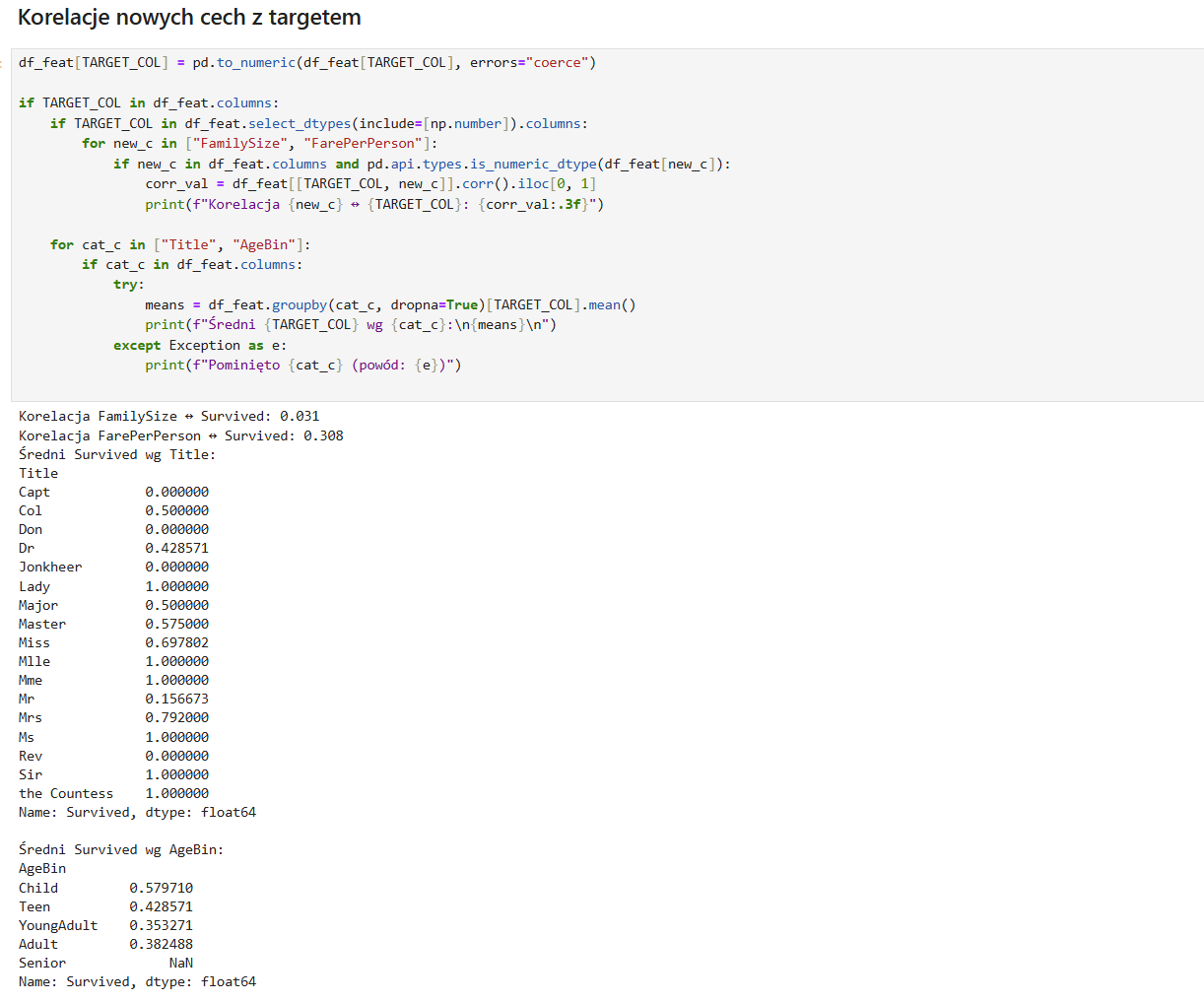
**Wnioski:**  
Wartości odstające (osoby bardzo młode i starsze) zostały przycięte do granic IQR, co zmniejszyło rozrzut i poprawiło czytelność danych.  
Rozkład wieku pozostał naturalny — większość pasażerów ma od 20 do 40 lat.

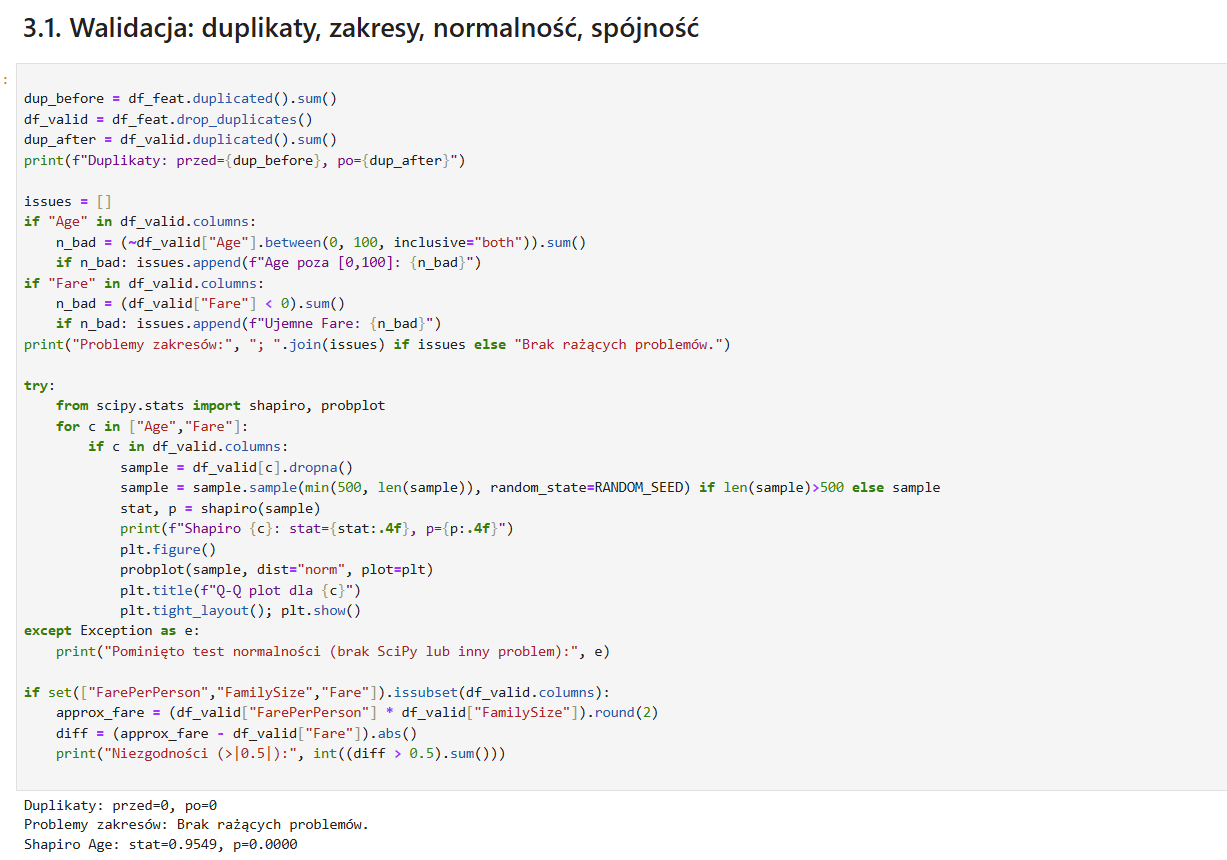


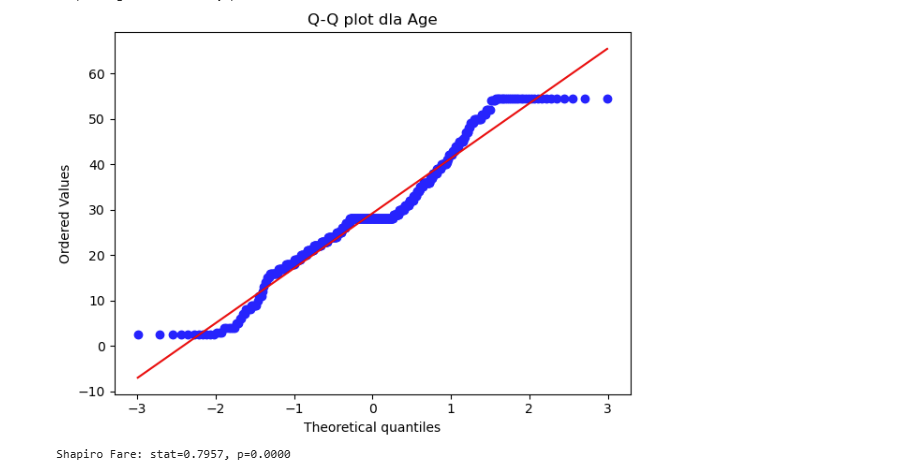
**Opis:**  
Boxplot prezentuje rozkład wieku (Age) pasażerów przed i po ograniczeniu outlierów.

**Wnioski:**  
Wartości odstające (osoby bardzo młode i starsze) zostały przycięte do granic IQR, co zmniejszyło rozrzut i poprawiło czytelność danych.  
Rozkład wieku pozostał naturalny — większość pasażerów ma od 20 do 40 lat.



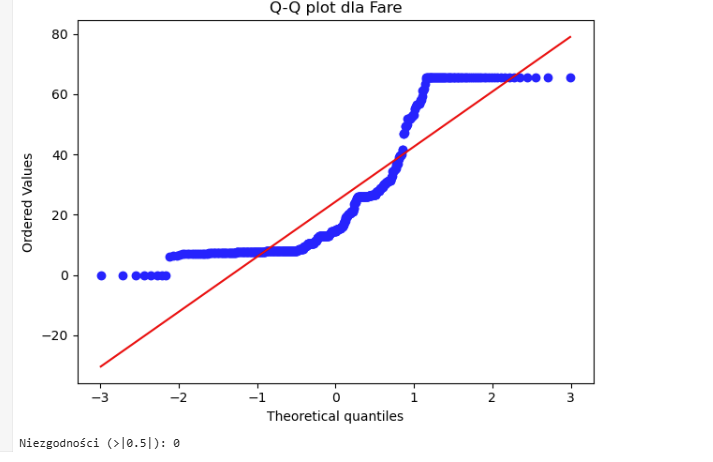






**Opis:**  
Wykres Q–Q porównuje rozkład wieku z teoretycznym rozkładem normalnym.

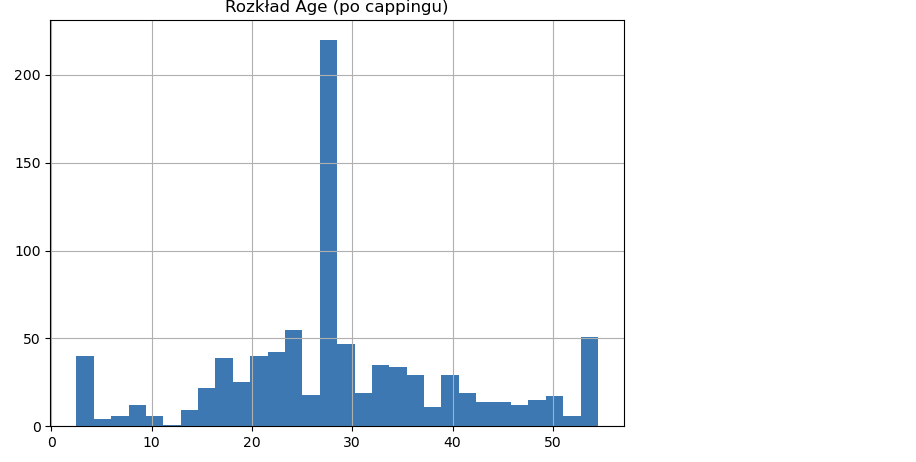
**Wnioski:**  
Wartości układają się wzdłuż linii tylko w części środkowej, natomiast końce wykresu odchylają się od niej, co potwierdza nienormalny rozkład.  
Wynik testu Shapiro–Wilka (p < 0.05) potwierdza, że Age nie ma rozkładu normalnego (lekko prawoskośny).



**Opis:**  
Wykres Q–Q sprawdza zgodność rozkładu cen biletów z rozkładem normalnym.

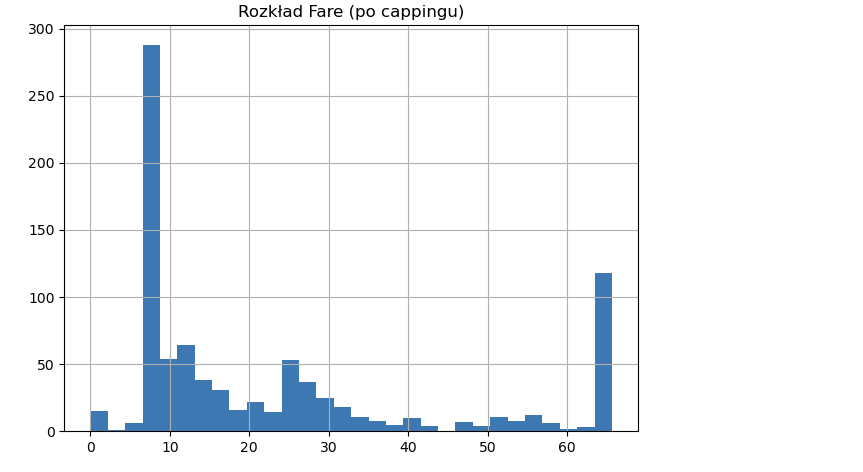
**Wnioski:**  
Rozkład jest silnie skośny w prawo — tylko część danych pokrywa się z linią teoretyczną.  
Większość pasażerów zapłaciła niską cenę, a nieliczni — bardzo wysoką. Potwierdza to wynik testu Shapiro (p < 0.05).





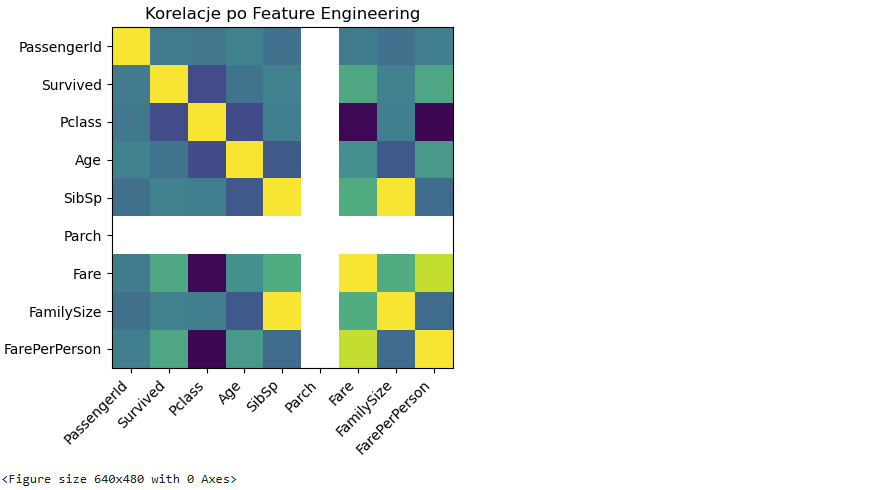
**Opis:**  
Histogram prezentuje rozkład wieku pasażerów po imputacji i cappingu.

**Wnioski:**  
Dominującą grupą są osoby w wieku około 30 lat. Widoczne mniejsze grupy dzieci i starszych.  
Wysoki słupek przy 30 latach wynika z imputacji medianą, która zastąpiła brakujące wartości.



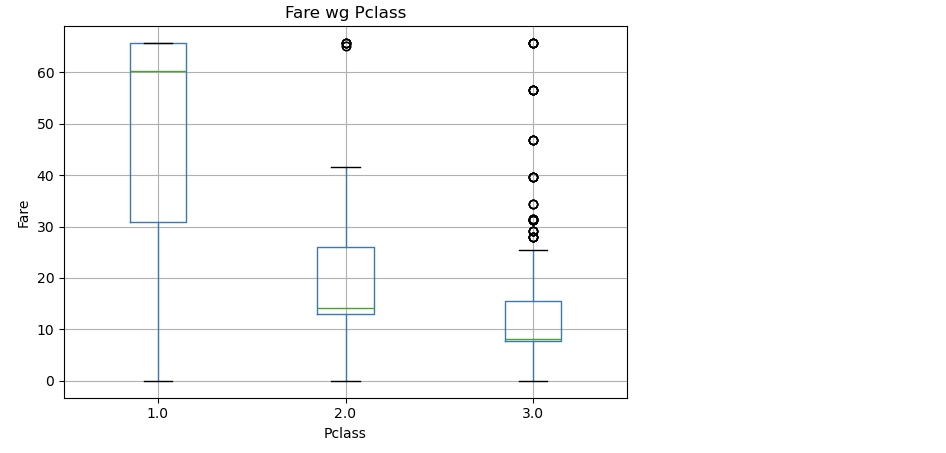
**Opis:**  
Histogram przedstawia rozkład ceny biletu po oczyszczeniu danych.

**Wnioski:**  
Rozkład jest wyraźnie skośny — większość pasażerów zapłaciła niewielkie kwoty (do ok. 50), a pojedyncze przypadki przekraczały 60.  
Outliery zostały skutecznie ograniczone, co poprawia stabilność dalszej analizy.



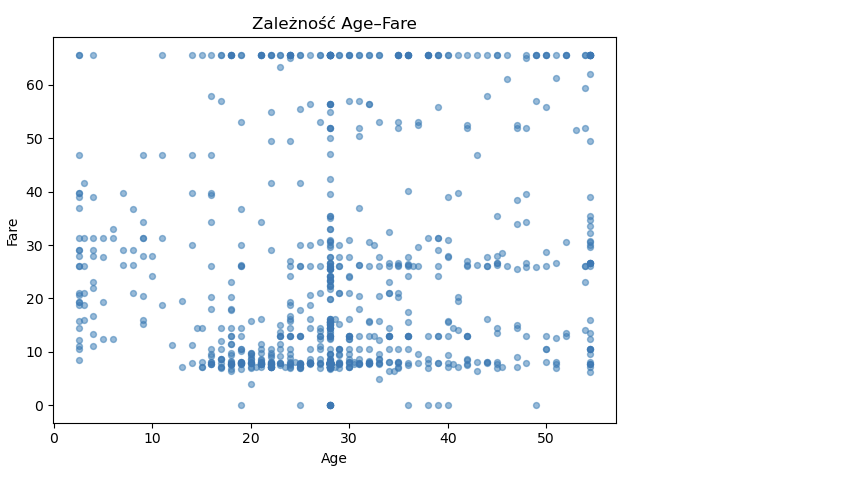
**Opis:**  
Mapa korelacji pokazuje zależności pomiędzy zmiennymi numerycznymi po dodaniu nowych cech (FamilySize, FarePerPerson).

**Wnioski:**  
Najsilniejsze korelacje obserwuje się między FamilySize, SibSp i Parch (co jest logiczne — tworzą rodzinne relacje).  
Fare jest skorelowane z Survived – droższe bilety częściej należały do pasażerów, którzy przeżyli (wyższa klasa podróży).



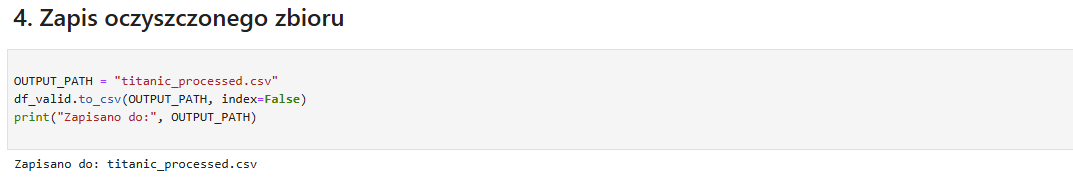
**Opis:**  
Wykres pudełkowy przedstawia ceny biletów w poszczególnych klasach pasażerskich.

**Wnioski:**  
Pasażerowie 1. klasy płacili znacznie więcej niż ci z klas 2 i 3.  
Rozrzut cen w klasie 3 jest większy, co odzwierciedla różne standardy miejsc w tej klasie.  
Zależność ta potwierdza silny wpływ klasy podróży na status pasażera i jego szanse przeżycia.



**Opis:**  
Wykres punktowy przedstawia zależność pomiędzy wiekiem pasażerów (Age) a ceną biletu (Fare) po przeprowadzeniu imputacji braków oraz cappingu wartości odstających. Każdy punkt odpowiada jednemu pasażerowi.

**Wnioski:**  
Nie widać wyraźnej zależności pomiędzy wiekiem a ceną biletu. Pasażerowie w podobnym wieku płacili bardzo różne kwoty, co sugeruje, że cena zależała głównie od klasy podróży (Pclass) i pozycji społecznej, a nie od wieku. Rozkład punktów jest równomierny, bez zauważalnego trendu rosnącego lub malejącego.



1. Wnioski

Przeprowadzona analiza zbioru *Titanic Dataset* pozwoliła kompleksowo przygotować dane do dalszego modelowania predykcyjnego. W procesie wstępnego przetwarzania zidentyfikowano i uzupełniono braki danych, głównie w kolumnach Age i Embarked. Wartości brakujące zastąpiono medianą (dla zmiennych numerycznych) i modą (dla kategorycznych), co pozwoliło zachować pełną liczebność obserwacji. Wartości odstające w zmiennych Age i Fare zostały ograniczone metodą IQR (capping), aby zminimalizować ich wpływ na analizę bez usuwania rekordów.

Analiza wykazała, że dane nie mają rozkładu normalnego — zmienna Fare jest silnie skośna w prawo, a Age lekko prawoskośna. Najsilniejsze zależności korelacyjne zaobserwowano między zmiennymi rodzinnymi (SibSp, Parch, FamilySize) oraz pomiędzy Fare i Pclass. Zmienna Survived wykazuje dodatnią zależność z Fare i ujemną z Pclass, co oznacza, że pasażerowie z wyższych klas mieli większe szanse przeżycia.

W ramach feature engineeringu utworzono nowe cechy: FamilySize, IsAlone oraz FarePerPerson, które lepiej opisują relacje rodzinne i kontekst ekonomiczny podróży. Dodatkowe walidacje potwierdziły spójność danych — brak duplikatów, błędnych zakresów i logicznych niespójności.

Podsumowując, uzyskany zbiór danych po oczyszczeniu jest kompletny, pozbawiony ekstremalnych błędów i gotowy do budowy modelu klasyfikacyjnego przewidującego przeżycie pasażerów Titanica.