

Analiza danych pasażerów Titanica



Cel analizy

Które cechy pasażerów — płeć, klasa biletu, wiek i port wejścia na pokład — były najsilniej związane z przeżyciem katastrofy?

Zbiór danych zawiera informacje o pasażerach RMS Titanic, który zatonął 15 kwietnia 1912 roku po zderzeniu z górą lodową. Dane obejmują takie atrybuty jak klasa podróży, wiek, płeć, liczba rodzeństwa/małżonków na pokładzie, liczba rodziców/dzieci na pokładzie, cena biletu oraz miejsce zaokrętowania.

Zbiór zawiera także informację o tym, czy pasażer przeżył katastrofę.

Titanic przewoził ponad 2,200 osób, z czego ponad 1,500 zginęło, co czyni tę katastrofę jedną z najbardziej tragicznych w historii morskiej.

Kolumny:

- **pclass** - Klasa biletu
- **survived** - Czy pasażer przeżył katastrofę
- **name** - Imię i nazwisko pasażera
- **sex** - Płeć pasażera
- **age** - Wiek pasażera
- **sibsp** - Liczba rodzeństwa/małżonków na pokładzie

- **parch** - Liczba rodziców/dzieci na pokładzie
- **ticket** - Numer biletu
- **fare** - Cena biletu
- **cabin** - Numer kabiny
- **embarked** - Port, w którym pasażer wszedł na pokład (C = Cherbourg, Q = Queenstown, S = Southampton)
- **boat** - Numer łodzi ratunkowej
- **body** - Numer ciała (jeśli pasażer nie przeżył i ciało zostało odnalezione)
- **home.dest** - Miejsce docelowe

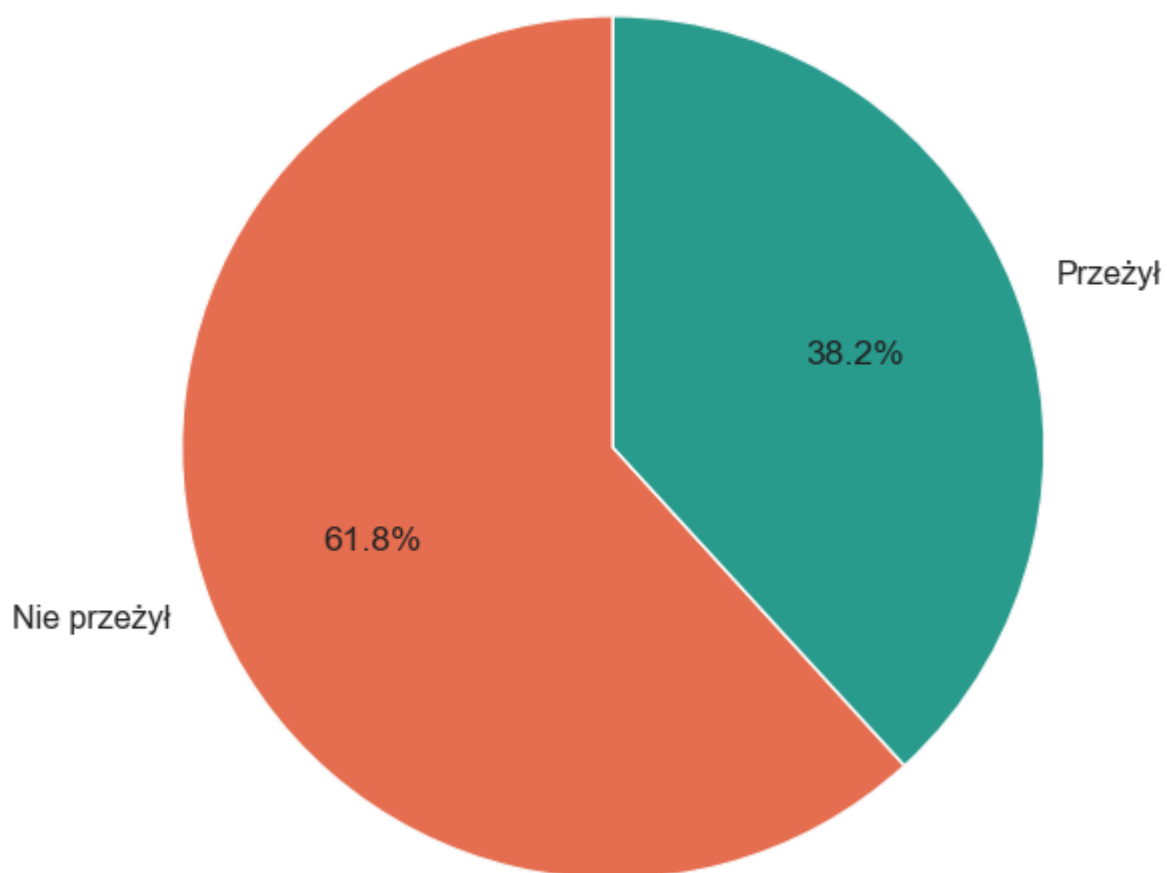
	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat
499	2.0	0.0	McCrie, Mr. James Matthew	male	30.0	0.0	0.0	233478	13.0000	NaN	S	Na
513	2.0	1.0	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1.0	0.0	237736	30.0708	NaN	C	Na
832	3.0	0.0	Goodwin, Mrs. Frederick (Augusta Tyler)	female	43.0	1.0	6.0	CA 2144	46.9000	NaN	S	Na
637	3.0	0.0	Aronsson, Mr. Ernst Axel Algot	male	24.0	0.0	0.0	349911	7.7750	NaN	S	Na
170	1.0	1.0	Ismay, Mr. Joseph Bruce	male	49.0	0.0	0.0	112058	0.0000	B52 B54 B56	S	

Zbiór zawiera **1309 rekordów** oraz **14 kolumn**.

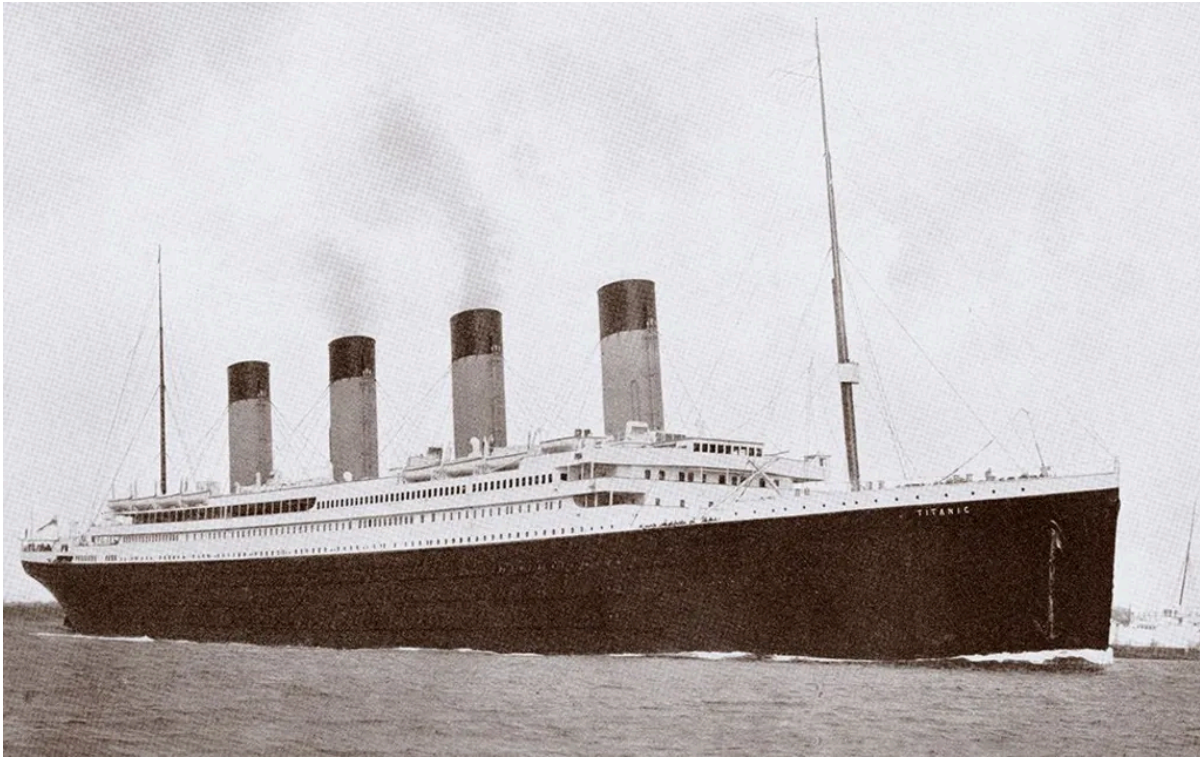
Dane obejmują **2 płcie**, **3 klasy biletów**

oraz informację, czy pasażer przeżył katastrofę.

Udział pasażerów, którzy przeżyli katastrofę



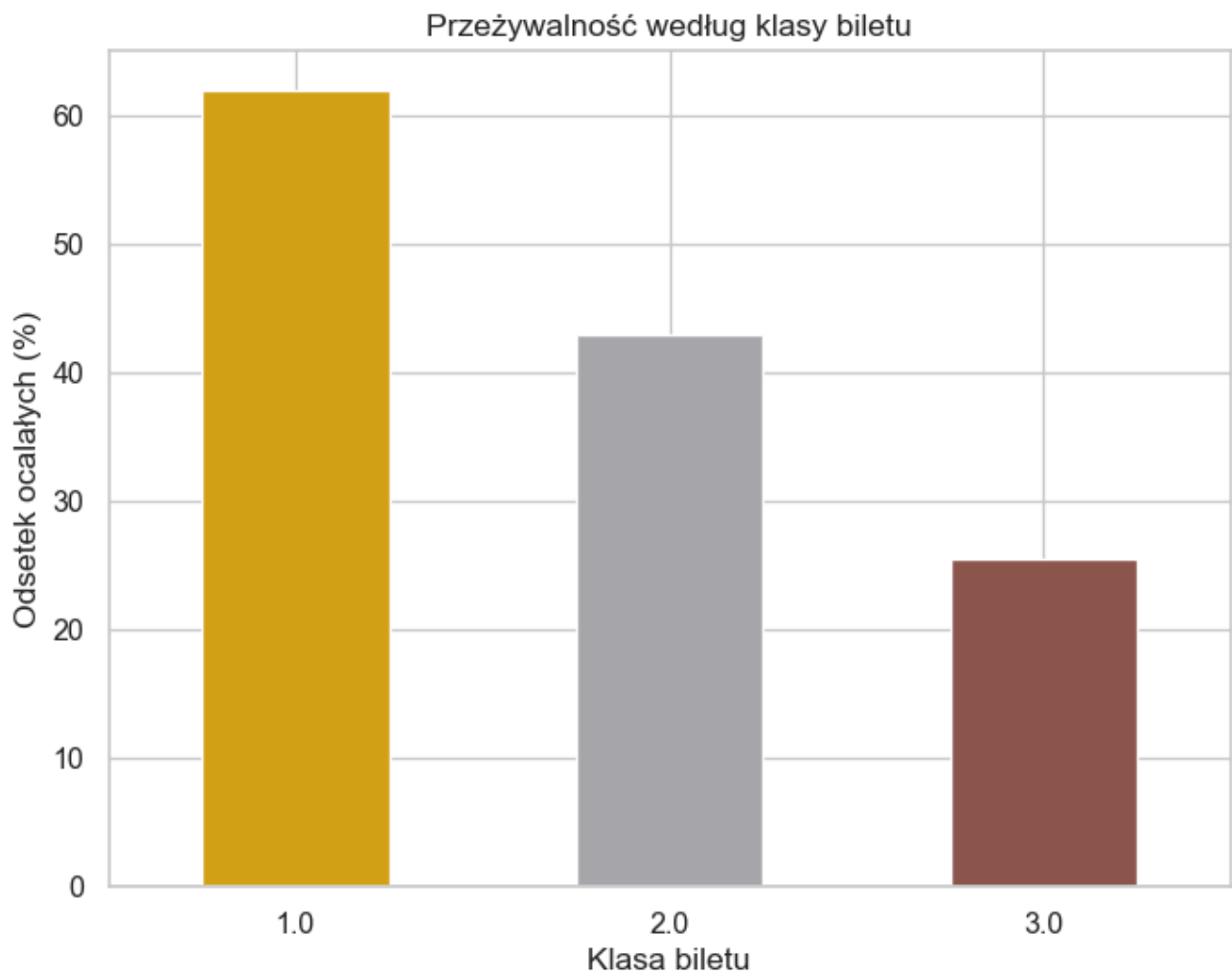
Krok 1: Ogólny przegląd danych



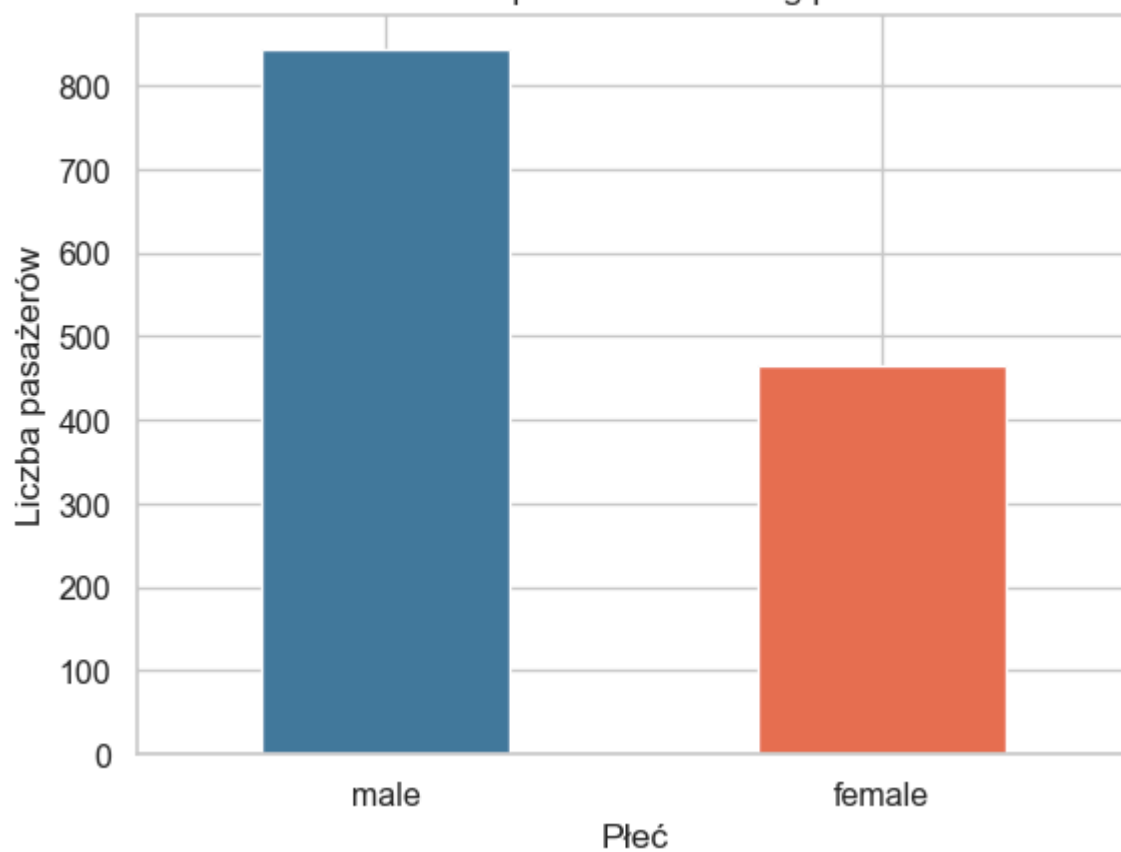
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1309 entries, 0 to 1308
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  -
0   pclass      1309 non-null   float64
1   survived    1309 non-null   float64
2   name        1309 non-null   object
3   sex         1309 non-null   object
4   age         1046 non-null   float64
5   sibsp       1309 non-null   float64
6   parch       1309 non-null   float64
7   ticket      1309 non-null   object
8   fare        1308 non-null   float64
9   cabin       295 non-null    object
10  embarked    1307 non-null   object
11  boat        486 non-null    object
12  body        121 non-null    float64
13  home.dest   745 non-null    object
dtypes: float64(7), object(7)
memory usage: 153.4+ KB
```

```
pclass      3
survived     2
name        1307
sex          2
age         98
sibsp        7
parch        8
ticket      929
fare        281
cabin       186
embarked     3
boat        27
body        121
home.dest    369
dtype: int64
```

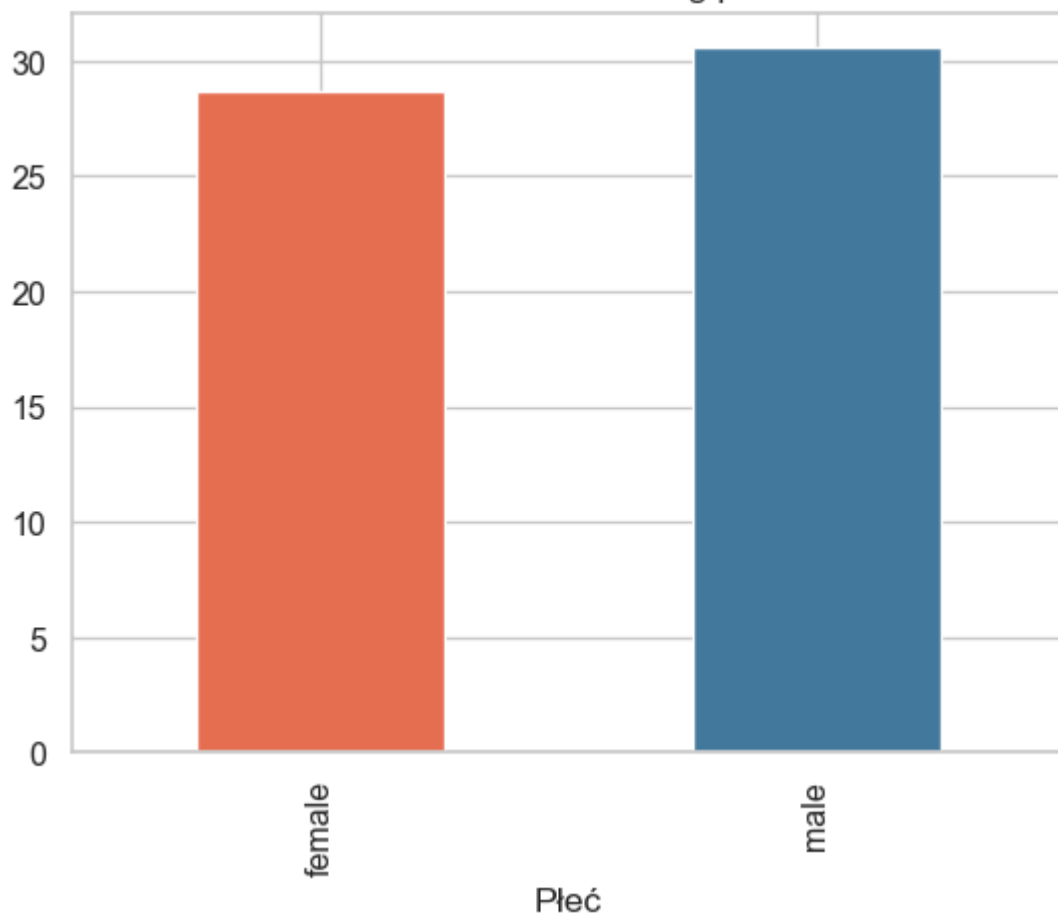
0



Liczba pasażerów według płci



Średnia wieku według płci



	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1.0	yes	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	S
1	1.0	yes	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	S
2	1.0	no	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	S
3	1.0	no	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	S
4	1.0	no	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	S

Spostrzeżenia

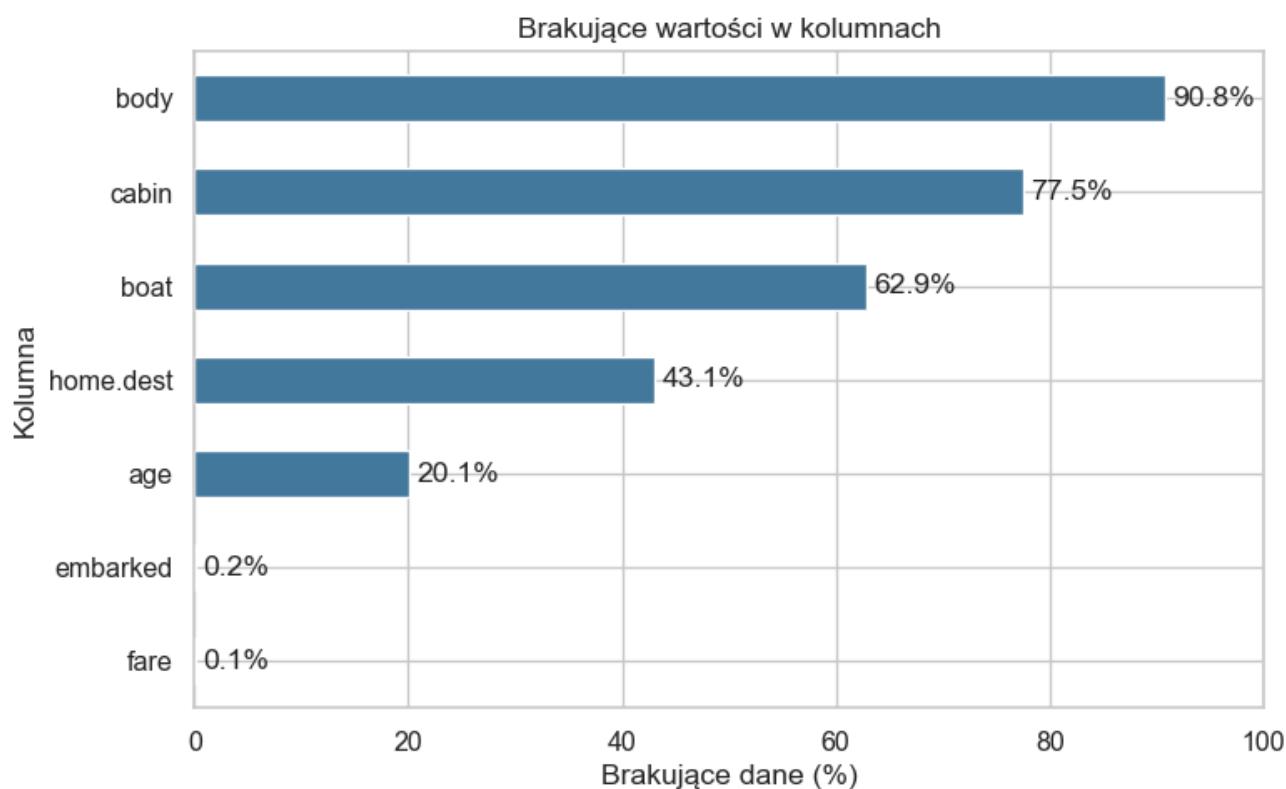
- Zbiór zawiera **1309 rekordów i 14 kolumn**.
- Dane dotyczą pasażerów trzech klas biletowych.
- W zbiorze jest **843 mężczyzn i 466 kobiet**.
- Katastrofę przeżyło **500 osób (38,2%)**, a **809 osób (61,8%)** nie przeżyło.
- Średni wiek pasażerów wynosi około **30 lat**.

Krok 2: Brakujące wartości

```
body          90.76
cabin         77.46
boat          62.87
home.dest     43.09
age           20.09
embarked      0.15
fare          0.08
dtype: float64
```

sex	pclass	survived	
male	3.0	no	128
female	3.0	yes	34
		no	30
male	1.0	no	20
	3.0	yes	16
female	1.0	yes	11
male	2.0	no	11
	1.0	yes	8
female	2.0	yes	2
male	2.0	yes	2
female	2.0	no	1

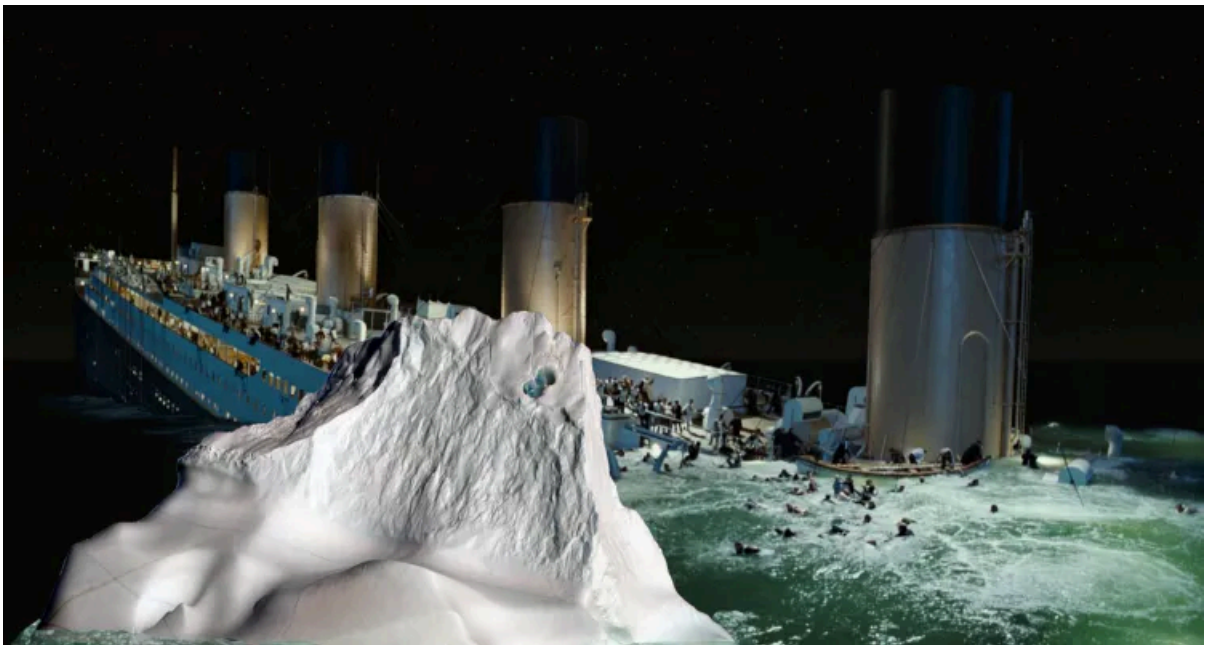
dtype: int64



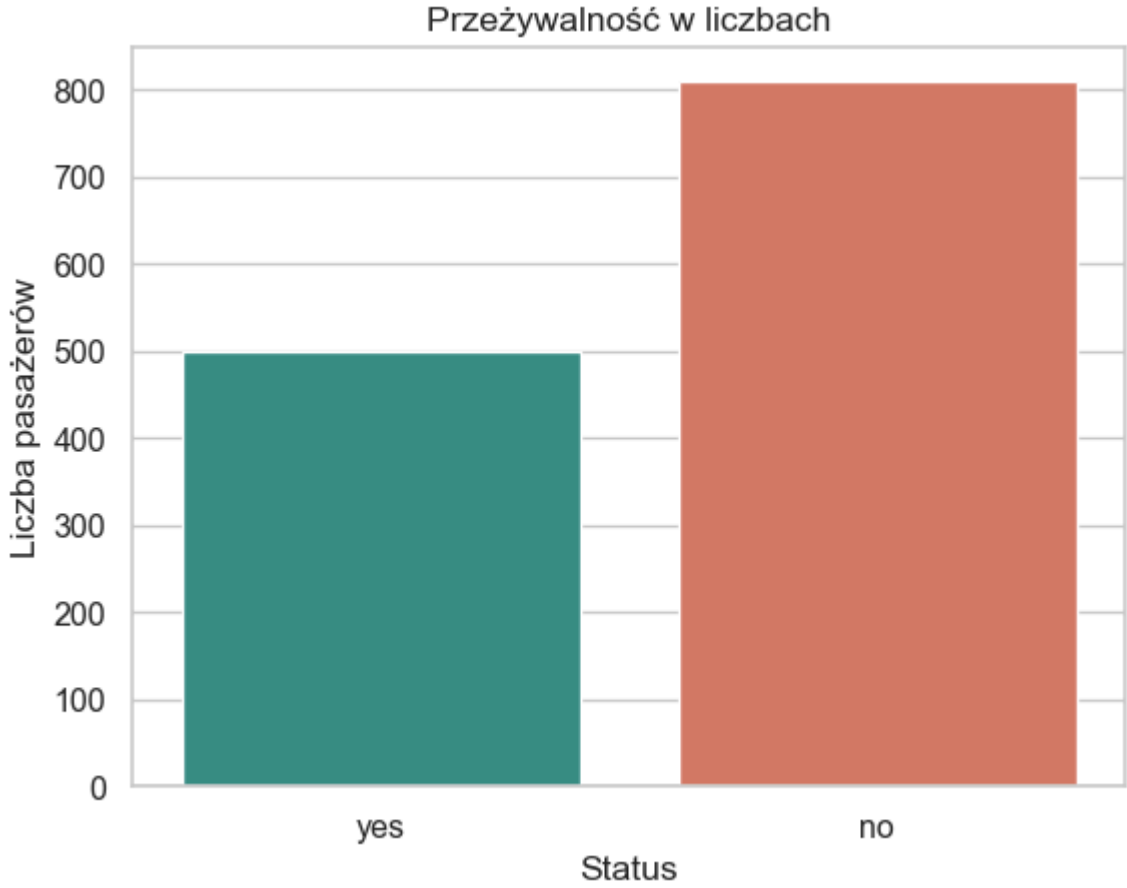
Spostrzeżenia

- Najwięcej braków występuje w kolumnach `body` (90,8%), `cabin` (77,5%) i `boat` (62,9%).
- Kolumna `age`, ważna dla dalszej analizy, ma braki dla około 20% pasażerów.
- Braki w `fare` i `embarked` są marginalne — po około 0,2%.
- Kolumny `body` oraz `boat` dotyczą zdarzeń po katastrofie, dlatego nie będą używane do analizy czynników związanych z przeżyciem.
- W dalszej części analizy brakujący wiek uzupełnimy medianą obliczoną osobno dla grup płci i klas biletowych.

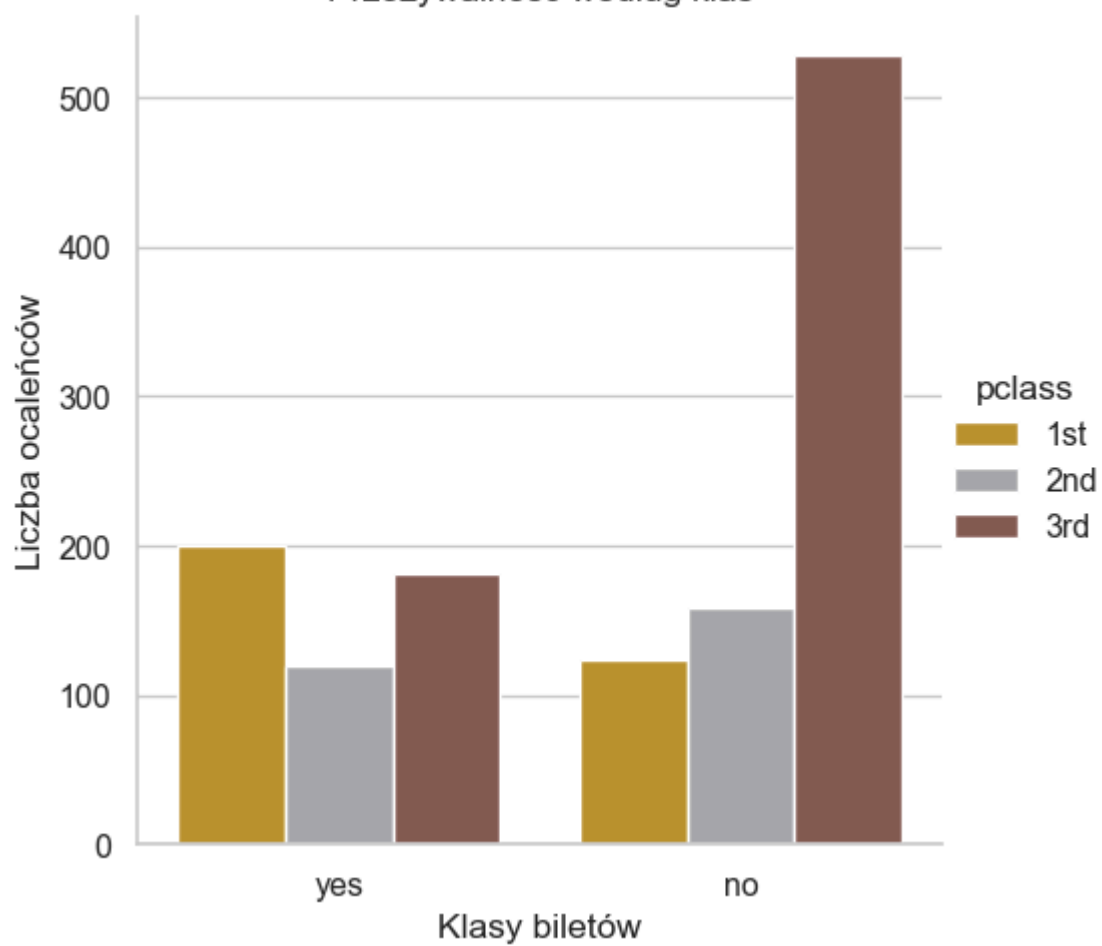
Krok 3: Analiza pojedynczych kolumn:



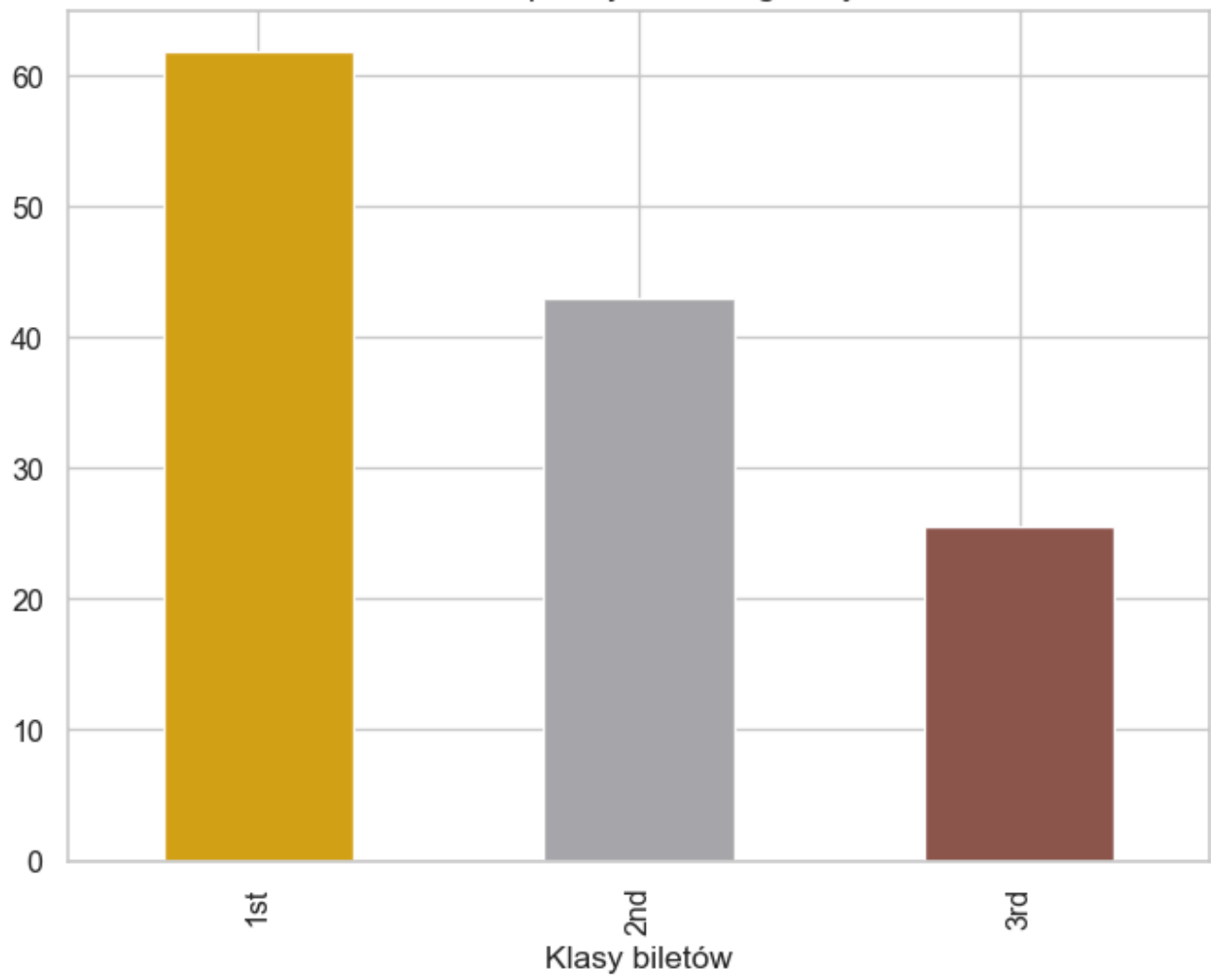
1. Przeżywalność

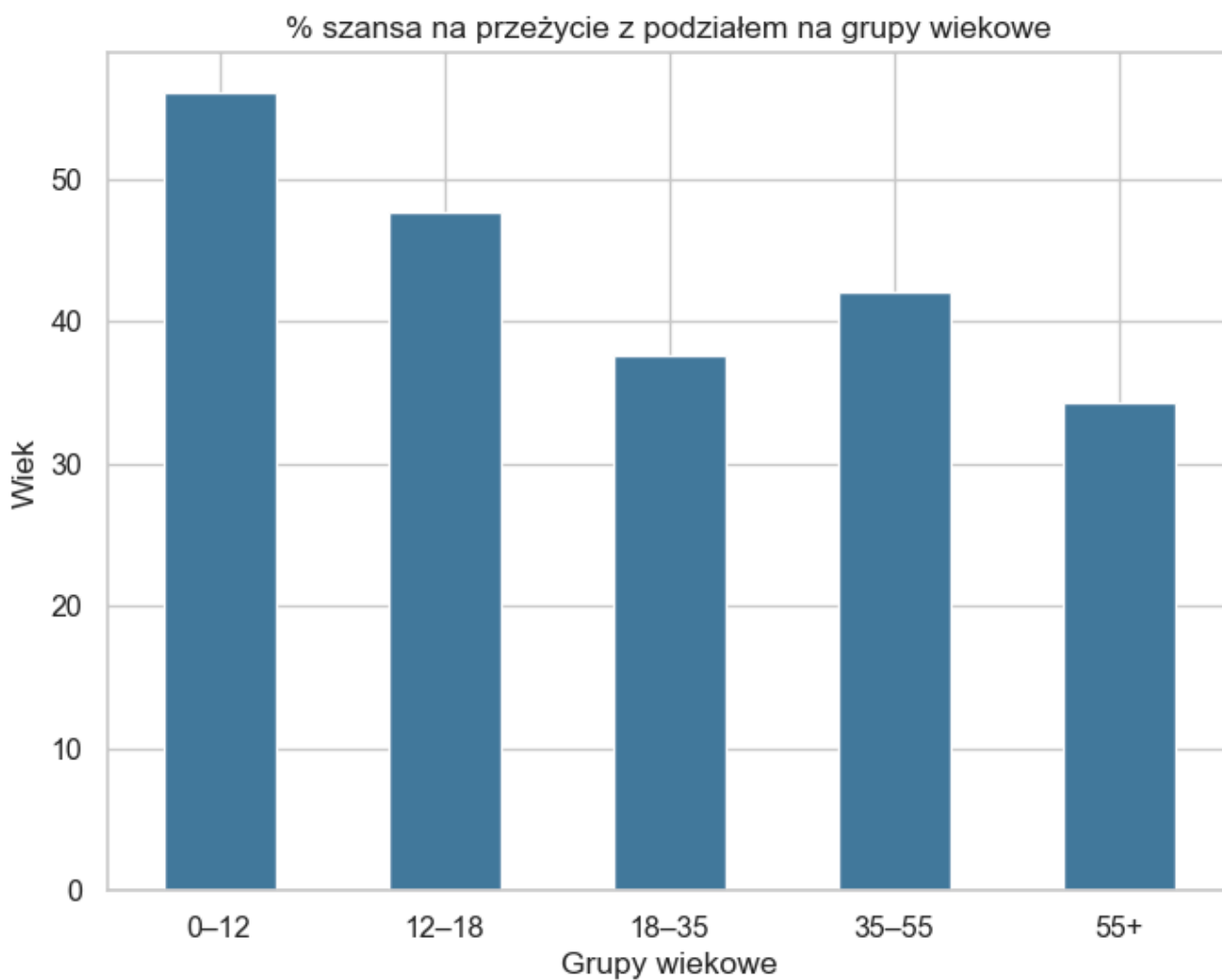
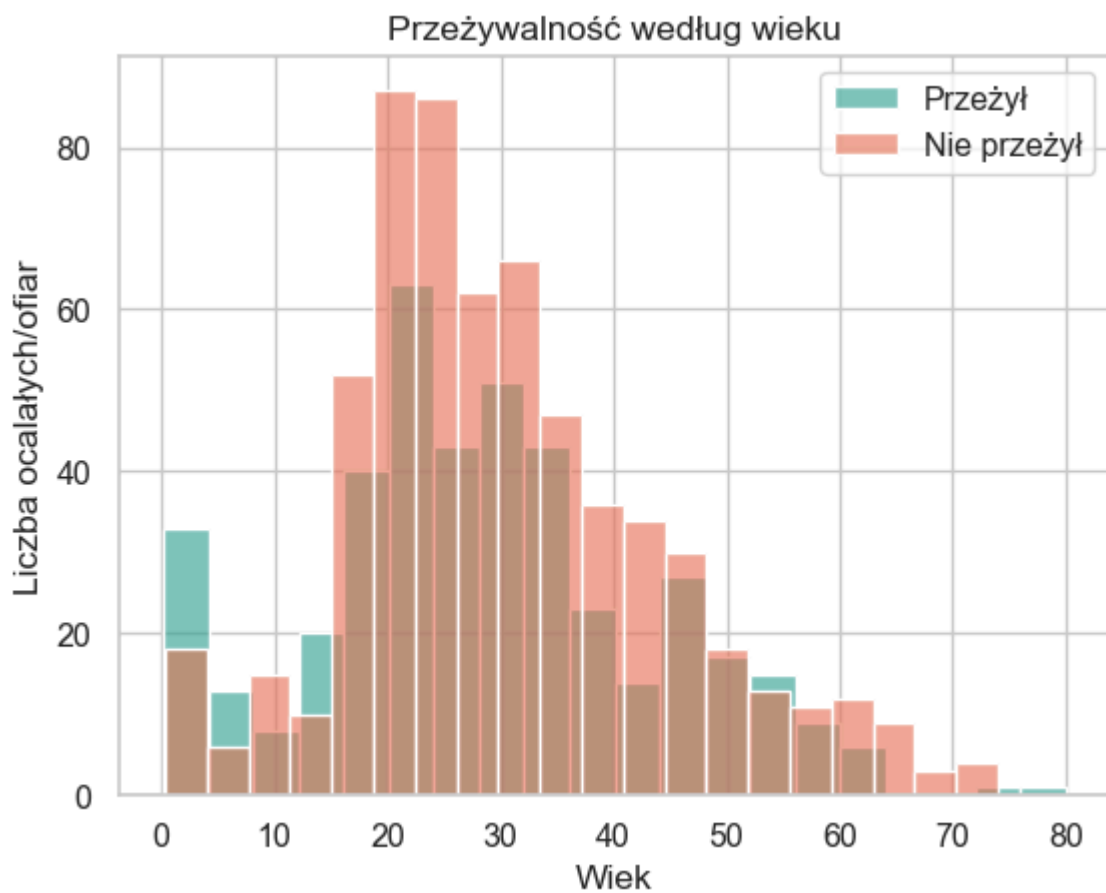


Przeżywalność według klas

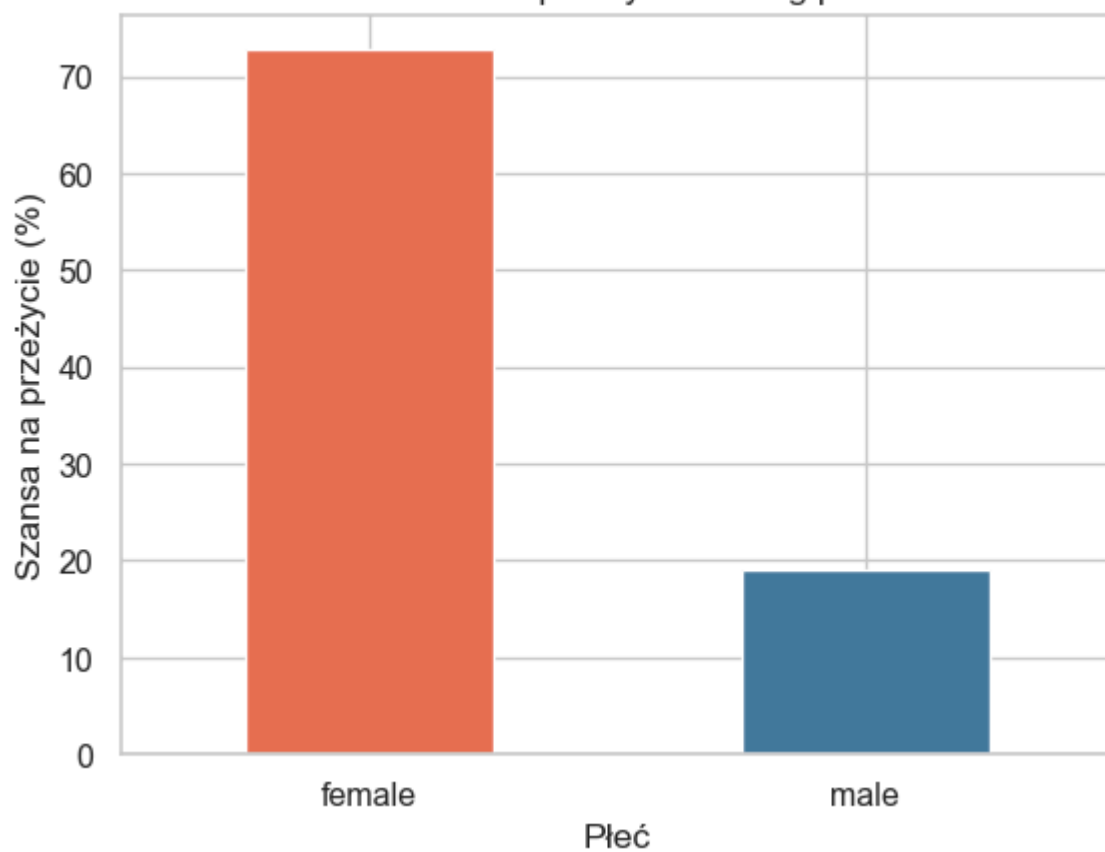


% szansa na przeżycie według klasy biletu

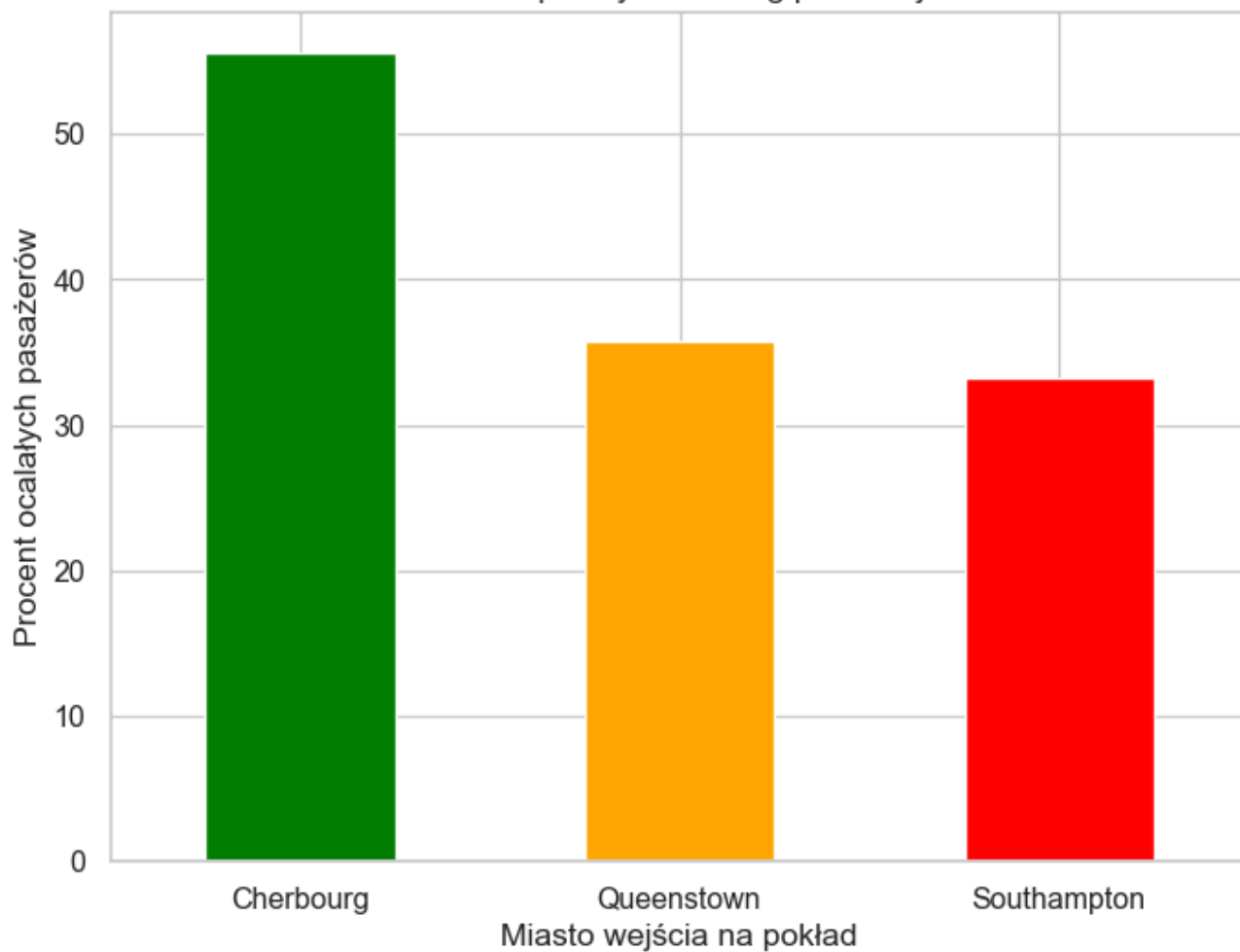




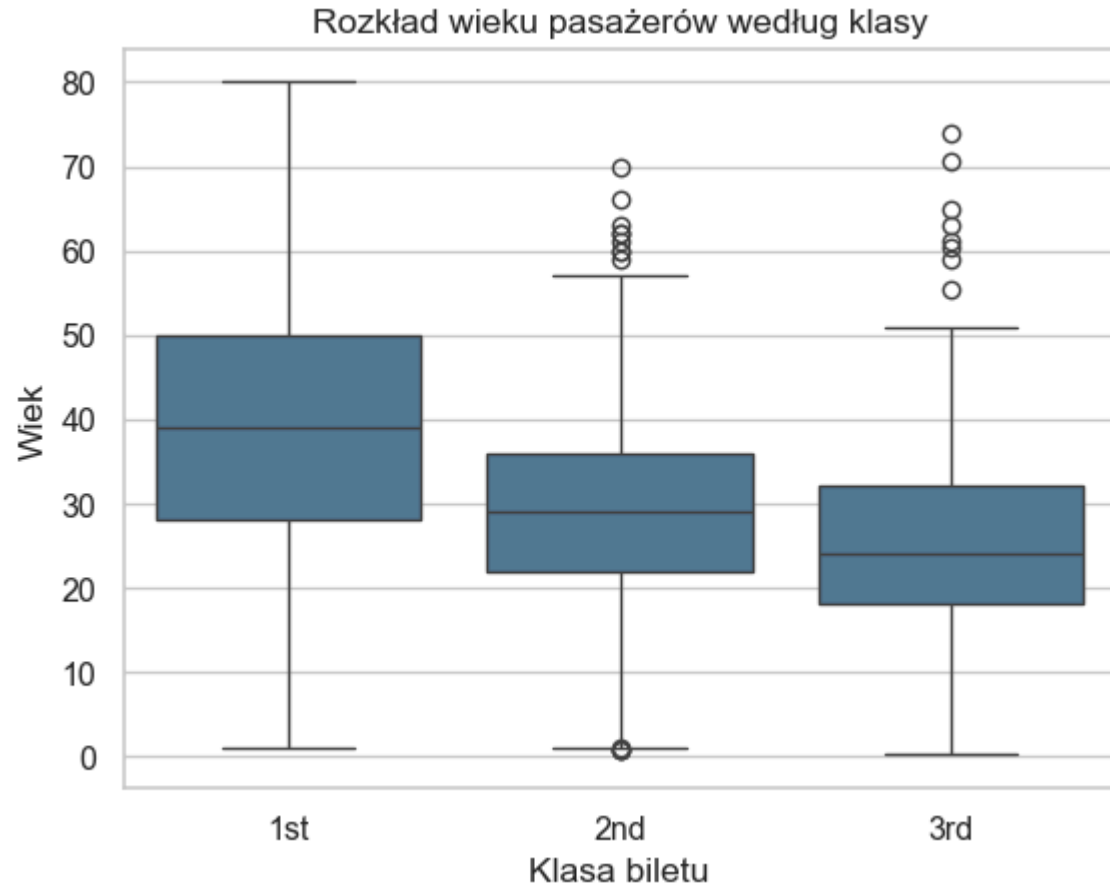
Szansa na przeżycie według płci

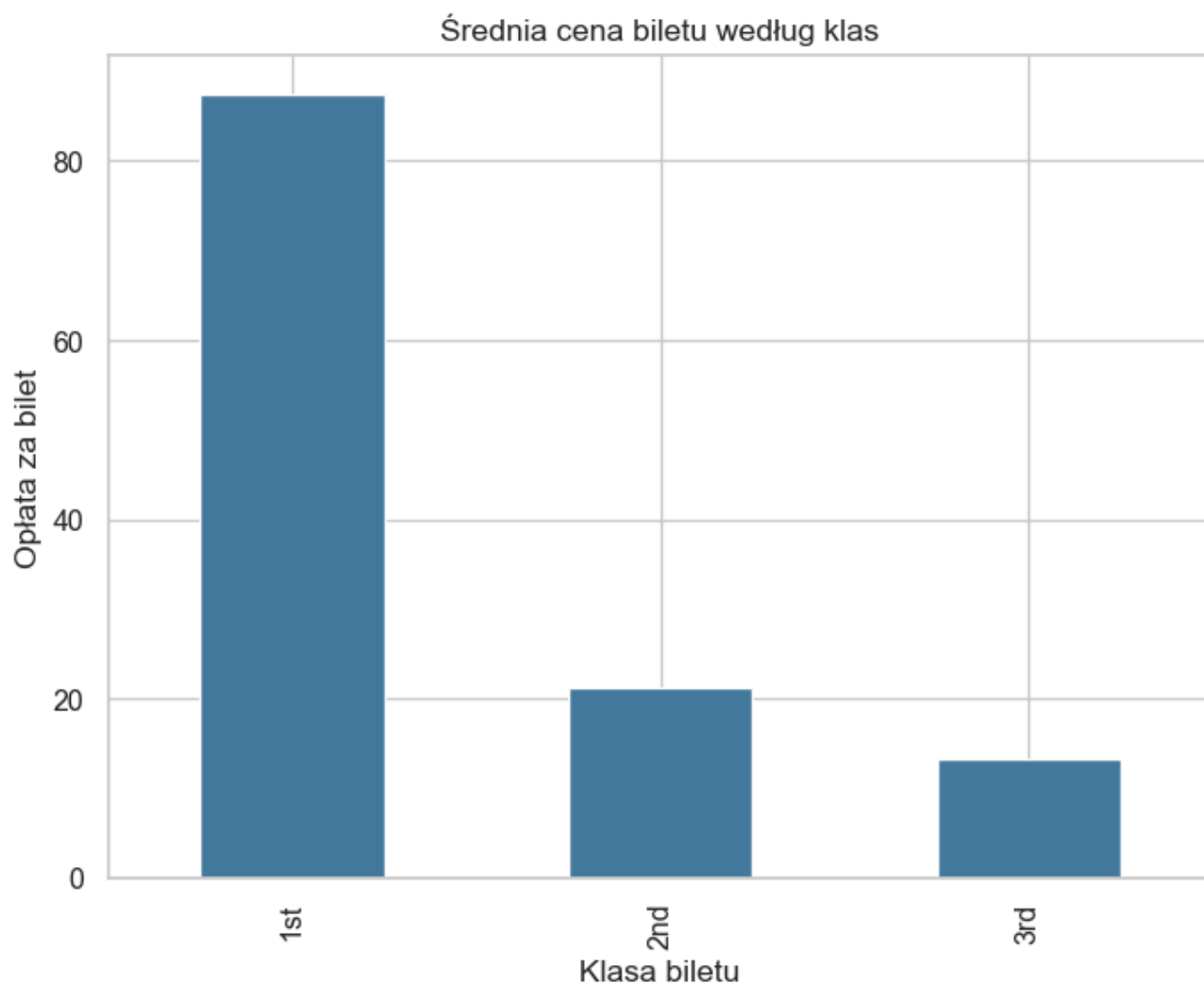


Szansa na przeżycie według portu wejścia



2. Podział na klasy





Ciekawostka historyczna: rozmieszczenie ocalałych w łodziach ratunkowych



Wykres pokazuje skład pasażerów przypisanych do poszczególnych łodzi ratunkowych. Jest to analiza uzupełniająca — numer łodzi był znany dopiero po katastrofie, dlatego nie używamy go do oceny czynników wpływających na przeżycie.

Spostrzeżenia

Przeżywalność

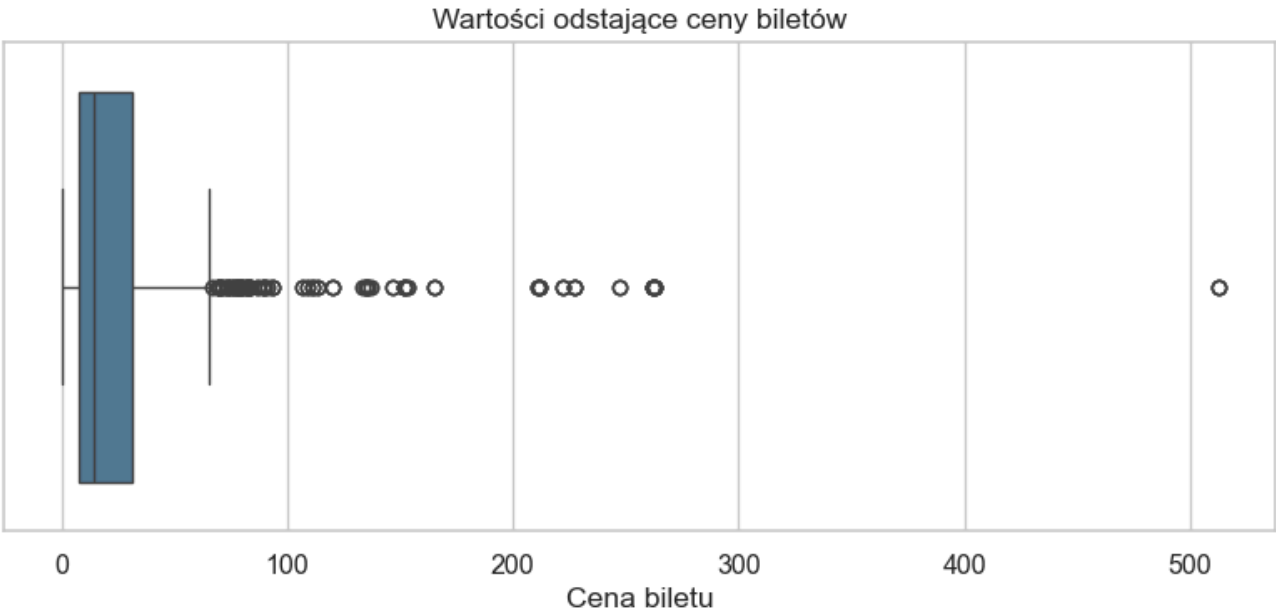
- Katastrofę przeżyło 38,2% pasażerów.
- Najwyższą przeżywalność odnotowano w klasie pierwszej (61,9%), a najniższą w klasie trzeciej (25,5%).
- Kobiety przeżywały zdecydowanie częściej niż mężczyźni.
- Najwyższy odsetek ocalałych wystąpił wśród dzieci do 12 lat (56,0%).
- Wśród portów wejścia na pokład najwyższą przeżywalność mieli pasażerowie z Cherbourga (55,6%). Ten wynik może być częściowo związany z innym składem pasażerów, np. większym udziałem osób z pierwszej klasy.

Podział na klasy

- We wszystkich klasach mężczyźni stanowili większość; największy ich udział wystąpił w klasie trzeciej (69,5%).
- Pasażerowie pierwszej klasy byli przeciętnie starsi niż pasażerowie klas drugiej i trzeciej.
- Średnia cena biletu malała wraz z klasą: 87,5 w klasie pierwszej, 21,2 w drugiej i 13,3 w trzeciej.

Krok 4: Trasformacje

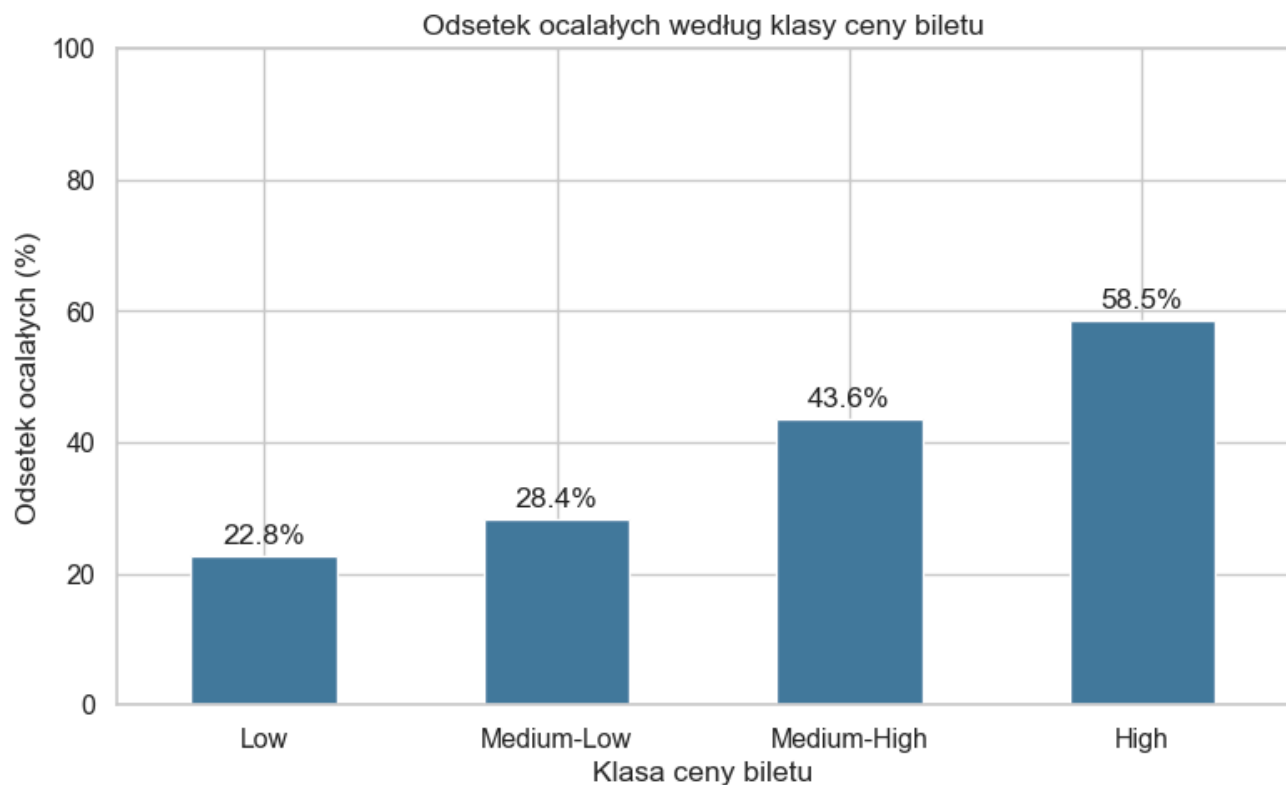
Liczba brakujących wartości w kolumnie age: 0



Liczba wykrytych wartości odstających: 171

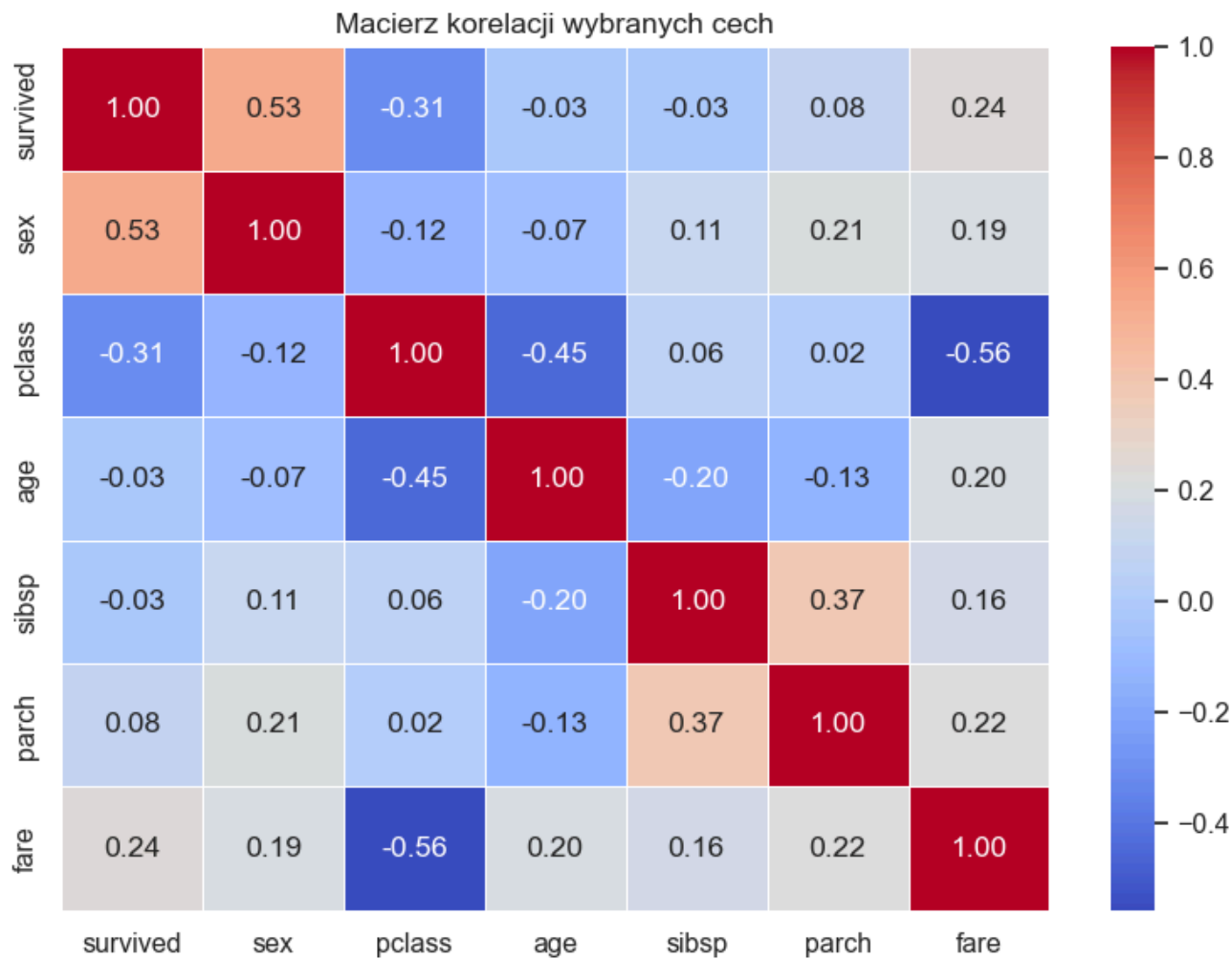
Low 337
Medium-High 328
High 323
Medium-Low 320
Name: fare_class, dtype: int64

fare_class	survived	
	no	yes
Low	77.2	22.8
Medium-Low	71.6	28.4
Medium-High	56.4	43.6
High	41.5	58.5



Krok 5: Relacje

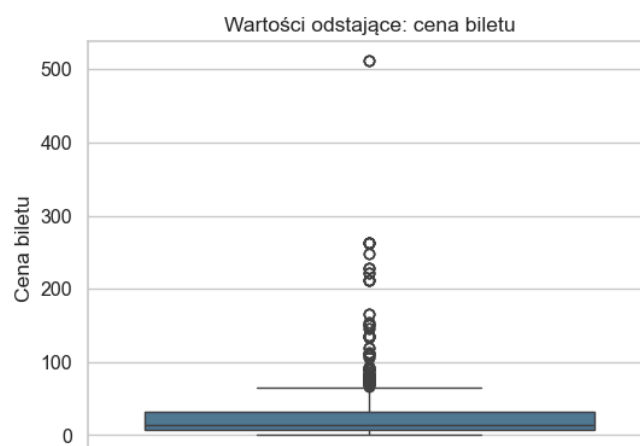
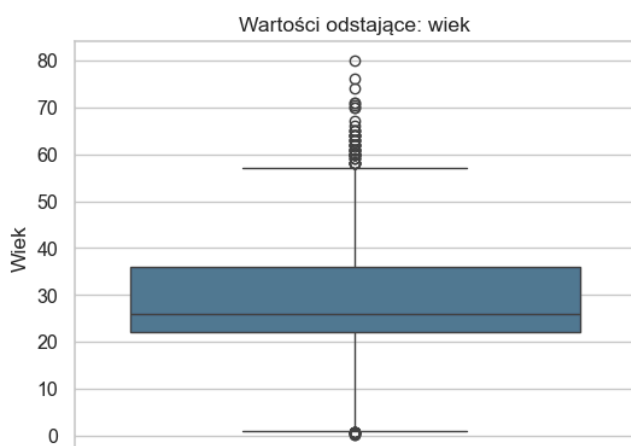
	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	b
224	1	0	Partner, Mr. Austen	0	45.5	0.0	0.0	113043	28.5000	C124	Southampton	N
1259	3	0	Turcin, Mr. Stjepan	0	36.0	0.0	0.0	349247	7.8958	NaN	Southampton	N
930	3	0	Kiernan, Mr. Philip	0	25.0	1.0	0.0	367229	7.7500	NaN	Queenstown	N
196	1	1	Marechal, Mr. Pierre	0	42.0	0.0	0.0	11774	29.7000	C47	Cherbourg	N
1269	3	0	Vande Velde, Mr. Johannes Joseph	0	33.0	0.0	0.0	345780	9.5000	NaN	Southampton	N



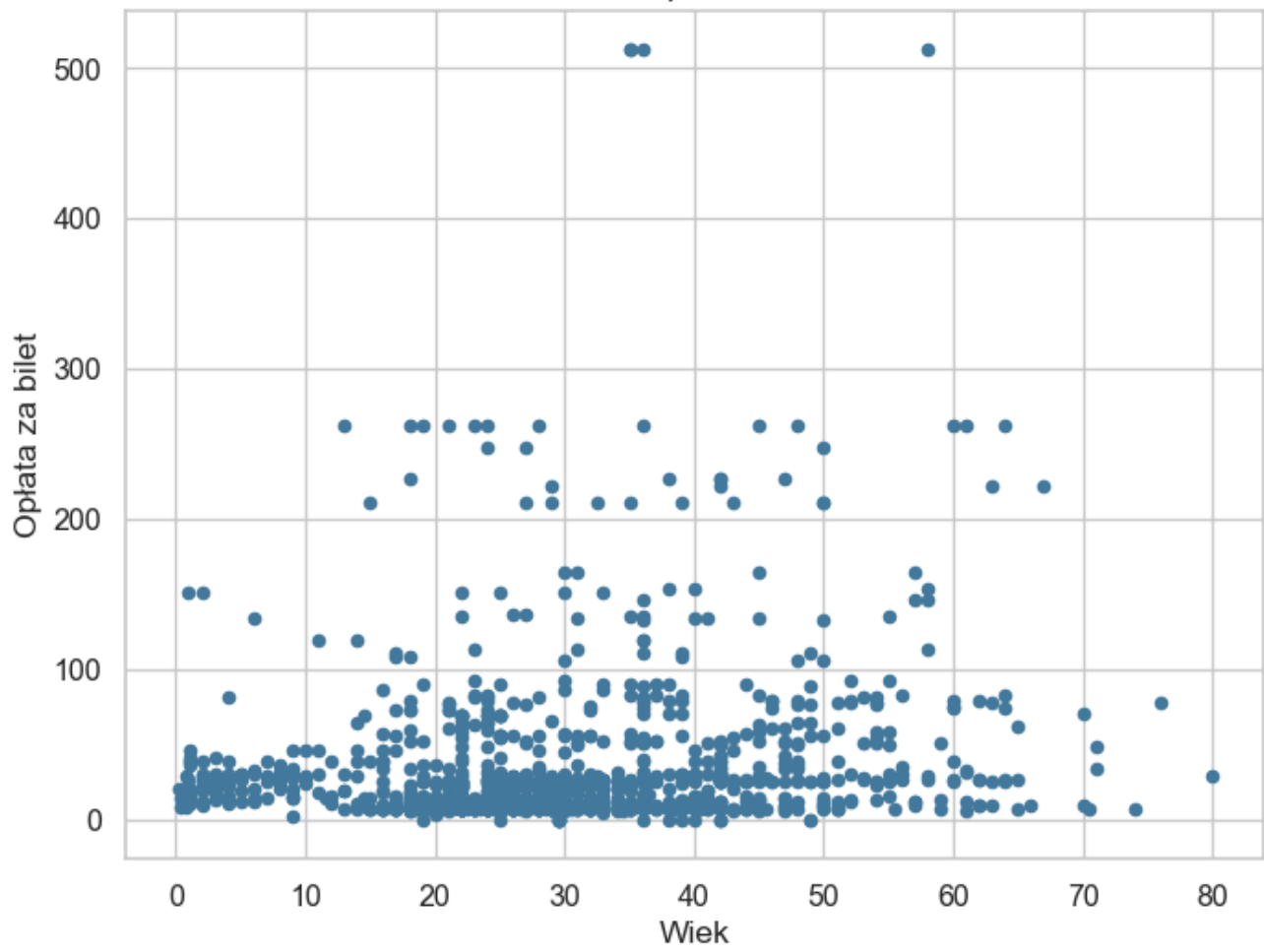
Spostrzeżenia

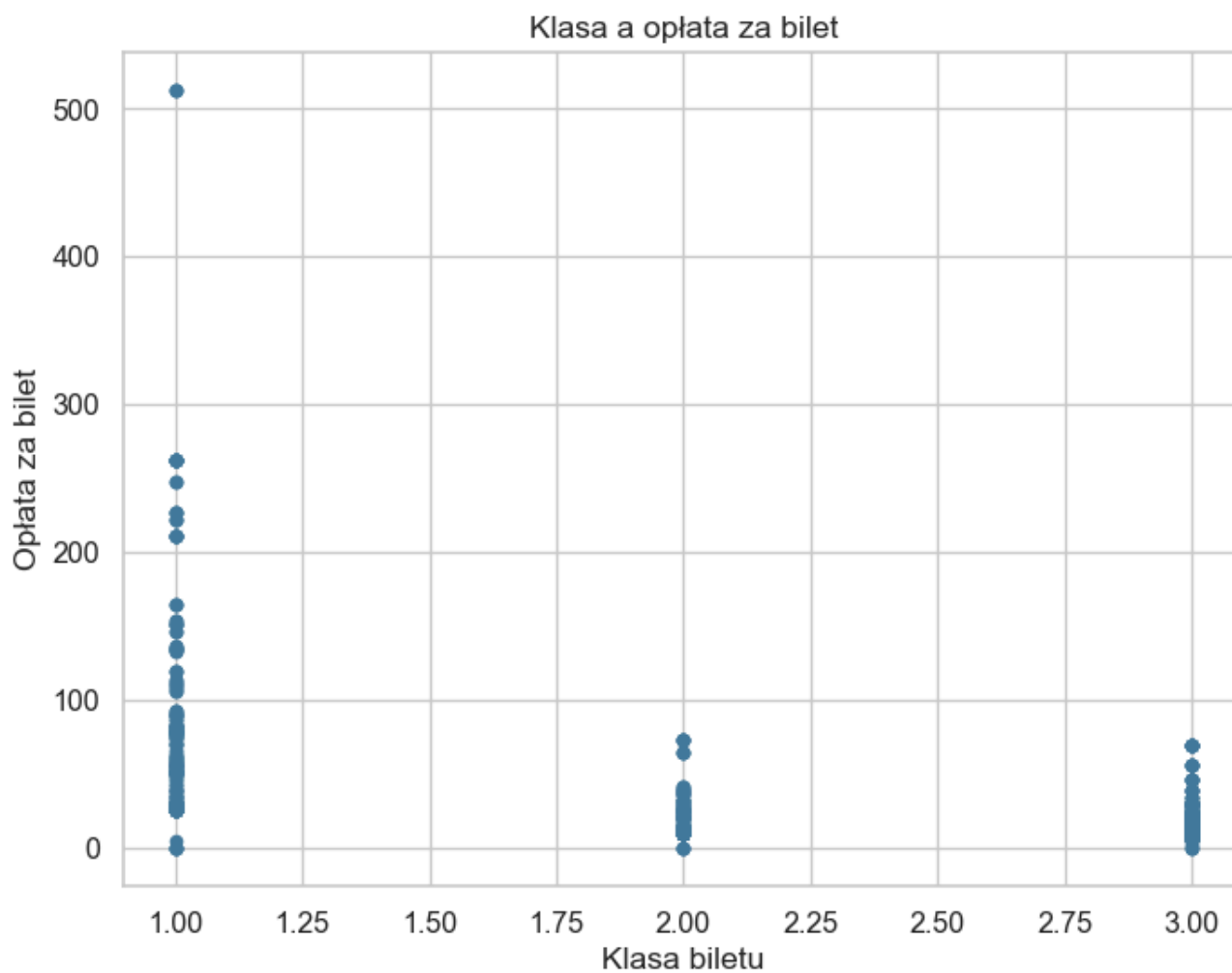
- Najsilniejsza zależność występuje między płcią a przeżyciem ($r = 0,53$). Ponieważ `female = 1`, oznacza to, że kobiety częściej przeżywały katastrofę.
- Klasa biletu i cena biletu są umiarkowanie ujemnie skorelowane ($r = -0,56$): niższy numer klasy oznaczał wyższy standard podróży i przeciętnie wyższą cenę biletu.
- Pozostałe zależności są słabe lub umiarkowane.

Krok 6: Wartości odstające



Wiek a opłata za bilet





Spostrzeżenia

- Wiek zawiera pojedyncze wartości odstające, lecz mieszczą się one w realistycznym zakresie od 0 do 80 lat.
- Najwięcej wartości odstających występuje w cenie biletu. Są to przede wszystkim bardzo drogie bilety, głównie w klasie pierwszej.
- Nie usunięto tych obserwacji, ponieważ przedstawiają rzeczywiste różnice w cenie i standardzie podróży.

Podsumowanie końcowe



1. Zakres danych i ich jakość

- Analiza obejmuje 1309 pasażerów i 14 pierwotnych kolumn.
- Katastrofę przeżyło 500 osób, czyli 38,2% pasażerów w zbiorze.
- Najwięcej braków dotyczy kolumn `body`, `cabin`, `boat` oraz `home.dest`.
- Brakujące wartości wieku uzupełniono medianą wyznaczoną w grupach płci i klas biletowych.

2. Czynniki związane z przeżyciem

- Płeć była najsilniej związana z przeżyciem: korelacja między płcią a przeżyciem wyniosła $r = 0,53$. Kobiety przeżywały znacznie częściej niż mężczyźni.
- Klasa biletu była wyraźnie powiązana z przeżywalnością: w pierwszej klasie przeżyło 61,9% pasażerów, a w trzeciej 25,5%.
- Najwyższy odsetek ocalałych wystąpił wśród dzieci do 12 lat — 56,0%.
- Wśród portów wejścia na pokład najwyższą przeżywalność odnotowano w Cherbourgu — 55,6%. Wynik może być związany z różnym składem pasażerów w portach, np. udziałem osób z pierwszej klasy.

3. Klasa podróży i cena biletu

- W każdej klasie było więcej mężczyzn niż kobiet; największy udział mężczyzn wystąpił w klasie trzeciej.
- Pasażerowie pierwszej klasy byli przeciętnie starsi niż pasażerowie klas drugiej i trzeciej.
- Średnia cena biletu wyniosła 87,5 w klasie pierwszej, 21,2 w drugiej i 13,3 w trzeciej.
- Cena biletu i klasa są umiarkowanie ujemnie skorelowane ($r = -0,56$), ponieważ niższy numer klasy oznaczał wyższy standard podróży.
- Przeżywalność rosła wraz z klasą ceny biletu: od 22,8% dla najtańszej grupy do 58,5% dla najdroższej.

4. Relacje i wartości odstające

- Poza płcią, klasą biletu i ceną nie występują silne zależności liniowe między analizowanymi cechami.
- Najwięcej wartości odstających występuje w cenie biletu; dotyczą one przede wszystkim bardzo drogich biletów w klasie pierwszej.
- Wartości odstających nie usunięto, ponieważ opisują rzeczywiste różnice w kosztach i standardzie podróży.

5. Analiza uzupełniająca: informacje po katastrofie

- Dane o łodziach ratunkowych pokazują skład pasażerów przypisanych do poszczególnych łodzi według klasy podróży.
- Informacja o kabinie jest najczęściej dostępna dla pasażerów pierwszej klasy, lecz kolumna `cabin` zawiera liczne braki.
- Informacja o odnalezieniu ciała również jest niepełna i dotyczy zdarzeń po katastrofie.
- Dane o łodziach, kabinach i ciałach są wartościami historycznymi, ale nie są używane do oceny czynników związanych z przeżyciem, ponieważ opisują skutki katastrofy lub mają znaczną liczbę braków.

Wniosek

Największe różnice w przeżywalności wiążą się z płcią, klasą biletu, wiekiem oraz ceną biletu. Analiza pokazuje zależności w zbiorze danych, ale sama nie dowodzi bezpośrednich przyczyn przeżycia.

0

0