Biblioteki i przygotowanie danych

In [3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import arff
```

In [131…
```
titanic_arff = arff.load(open("Zbiór danych Titanic.arff", 'r'))
print(titanic_arff.keys())

attribiutes = titanic_arff["attributes"]
data = titanic_arff["data"]

df = pd.DataFrame(data, columns=[x[0] for x in attribiutes])

df.head(5)
```

```
dict_keys(['description', 'relation', 'attributes', 'data'])
```

Out[131…

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
0	1.0	1	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5
1	1.0	1	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26
2	1.0	0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26
3	1.0	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26
4	1.0	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26

1. Liczebność poszczególnych etykiet dla danych zmiennych jakościowych

In [39]:
```
print('Liczba etykiet zmiennej pclass: {}'.format(len(df.pclass.unique())))
print('Liczba etykiet zmiennej survived: {}'.format(len(df.survived.unique())))
```

```
print('Liczba etykiet zmiennej name: {}'.format(len(df.name.unique())))
print('Liczba etykiet zmiennej sex: {}'.format(len(df.sex.unique())))
print('Liczba etykiet zmiennej ticket: {}'.format(len(df.ticket.unique())))
print('Liczba etykiet zmiennej cabin: {}'.format(len(df.cabin.unique())))
print('Liczba etykiet zmiennej embarked: {}'.format(len(df.embarked.unique())))
print('Liczba etykiet zmiennej boat: {}'.format(len(df.boat.unique())))
print('Liczba etykiet zmiennej home.dest: {}'.format(len(df["home.dest"].unique(
```

```
Liczba etykiet zmiennej pclass: 3
Liczba etykiet zmiennej survived: 2
Liczba etykiet zmiennej name: 1307
Liczba etykiet zmiennej sex: 2
Liczba etykiet zmiennej ticket: 929
Liczba etykiet zmiennej cabin: 187
Liczba etykiet zmiennej embarked: 4
Liczba etykiet zmiennej boat: 28
Liczba etykiet zmiennej home.dest: 370
```

2. Liczba wszystkich pasażerów.

In [43]:
```
print('Liczba wszystkich pasażerów: {}'.format(len(df)))
```

```
Liczba wszystkich pasażerów: 1309
```

3. Komentarz do wyników otrzymanych w punkcie 1 i 2.

Na podstawie liczby unikalnych etykiet można podzielić zmienne jakościowe ze względu na kardynalność - moc zbioru.

Zmienne o małej kardynalności:

- pclass: 3 klasy
- survived: 2 wartości (przeżyl/ nie przeżył)
- sex: 2 wartości plci
- embarked: 4 porty
- boat: 28 łodzi

Zmienne o dużej kardynalności:

- name: 1307 unikalnych imion i nazwisk (niemal wszyscy mają inne)
- ticket: 929 różnych numerów biletów
- cabin: 187 różnych oznaczeń kabin
- home.dest: 370 miejsc docelowych

Liczba wszystkich pasażerów: 1309. A widzimy, że kolumna name ma 1307 unikalnych etykiet, co oznacza, że niemal każda osoba w zbiorze jest unikalna.

==WNIOSKI==

- Zmienne o małej kardynalności są prostsze do zakodowania i analizy.
- Zmienne o dużej liczbie etykiet mogą być trudniejsze do przetworzenia lub modelowania – szczególnie jeśli są jakościowe. W analizie warto rozważyć ich redukcję lub transformację, aby ułatwić dalsze kroki.

4. Ile unikalnych etykiet ma zmienna mówiąca o kabinie danego pasażera?

In [96]:
```
unique_cabins = df['cabin'].unique()
print('Liczba unikalnych kabin:', len(unique_cabins))
print('Wartości (postać NumPy array):', np.array(unique_cabins))
```

```
Liczba unikalnych kabin: 187
Wartości (postać NumPy array): ['B5' 'C22 C26' 'E12' 'D7' 'A36' 'C101' None 'C62
C64' 'B35' 'A23'
 'B58 B60' 'D15' 'C6' 'D35' 'C148' 'C97' 'B49' 'C99' 'C52' 'T' 'A31' 'C7'
 'C103' 'D22' 'E33' 'A21' 'B10' 'B4' 'E40' 'B38' 'E24' 'B51 B53 B55'
 'B96 B98' 'C46' 'E31' 'E8' 'B61' 'B77' 'A9' 'C89' 'A14' 'E58' 'E49' 'E52'
 'E45' 'B22' 'B26' 'C85' 'E17' 'B71' 'B20' 'A34' 'C86' 'A16' 'A20' 'A18'
 'C54' 'C45' 'D20' 'A29' 'C95' 'E25' 'C111' 'C23 C25 C27' 'E36' 'D34'
 'D40' 'B39' 'B41' 'B102' 'C123' 'E63' 'C130' 'B86' 'C92' 'A5' 'C51' 'B42'
 'C91' 'C125' 'D10 D12' 'B82 B84' 'E50' 'D33' 'C83' 'B94' 'D49' 'D45'
 'B69' 'B11' 'E46' 'C39' 'B18' 'D11' 'C93' 'B28' 'C49' 'B52 B54 B56' 'E60'
 'C132' 'B37' 'D21' 'D19' 'C124' 'D17' 'B101' 'D28' 'D6' 'D9' 'B80' 'C106'
 'B79' 'C47' 'D30' 'C90' 'E38' 'C78' 'C30' 'C118' 'D36' 'D48' 'D47' 'C105'
 'B36' 'B30' 'D43' 'B24' 'C2' 'C65' 'B73' 'C104' 'C110' 'C50' 'B3' 'A24'
 'A32' 'A11' 'A10' 'B57 B59 B63 B66' 'C28' 'E44' 'A26' 'A6' 'A7' 'C31'
 'A19' 'B45' 'E34' 'B78' 'B50' 'C87' 'C116' 'C55 C57' 'D50' 'E68' 'E67'
 'C126' 'C68' 'C70' 'C53' 'B19' 'D46' 'D37' 'D26' 'C32' 'C80' 'C82' 'C128'
 'E39 E41' 'D' 'F4' 'D56' 'F33' 'E101' 'E77' 'F2' 'D38' 'F' 'F G63'
 'F E57' 'F E46' 'F G73' 'E121' 'F E69' 'E10' 'G6' 'F38']
```

Jak widzimy występuje etykieta "None" - to znaczy, że niektóre osoby nie miały przypisanej do siebie żadnej kabiny albo informacja o tym została zgubiona. Dlatego też w kolejnym zadaniu będzie pojawiać się literka N, jako pierwsza litera - sugerująca brak przypisanej kabiny

5. Zastąpienie obecnych etykiet w formacie LL11 do etykiet zawierających tylko pierwszą literę.

In [119…
```
df['CabinReduced'] = df['cabin'].astype(str).str[0]
df[['cabin', 'CabinReduced']].head(20)
```

Out[119…

	cabin	CabinReduced
0	B5	B
1	C22 C26	C
2	C22 C26	C
3	C22 C26	C
4	C22 C26	C
5	E12	E
6	D7	D
7	A36	A
8	C101	C
9	None	N
10	C62 C64	C
11	C62 C64	C
12	B35	B
13	None	N
14	A23	A
15	None	N
16	B58 B60	B
17	B58 B60	B
18	D15	D
19	C6	C

6. Liczba etykiet dla zmiennych z pkt 5.

In [121…

```
original_cardinality = len(df['cabin'].unique())
reduced_cardinality = len(df['CabinReduced'].unique())

print('Liczba etykiet przed redukcją:', original_cardinality)
print('Liczba etykiet po redukcji:', reduced_cardinality)

reduction_percent = 100 * (original_cardinality - reduced_cardinality) / origina
print('Procent redukcji kardynalności: {:.2f}%'.format(reduction_percent))
```

```
Liczba etykiet przed redukcją: 187
Liczba etykiet po redukcji: 9
Procent redukcji kardynalności: 95.19%
```

7. Dlaczego dokonuję redukcji akurat tej zmiennej? Jak to wpływa na przyszłe analizy, czy powoduje jakieś neghatywne skutki?

Zmienna cabin zawiera bardzo wiele unikalnych wartości (np. D15, C62, A36.), co utrudniałoby ich kodowanie oraz zwiększałoby wymiarowość danych.

Redukując zmienną cabin do pierwszej litery (np. C101 → C), uzyskujemy mniej kategorii, które często odpowiadają zapewne **pokładom statku**. Dzięki temu:

- zmniejszamy kardynalność,
- zachowujemy istotną informację o położeniu pasażera,
- ułatwiamy analizę i modelowanie

W takiej sytuacji tracimy precyzję (różnicę między np. C101 a C64), ale w zamian otrzymujemy bardziej stabilną zmienną do analizy. W większości przypadków korzyści przeważają nad negatywami.