# Titanic_Missing_Values

April 5, 2025

Titanic - Missing Values

Dominik Sakłaski, 415120

### 1. Załadowanie bibliotek

[1]:
```
import pandas as pd
import numpy as np
```

### 2. Wczytanie danych i wstępna inspekcja 20 pierwszych wierszy

[2]:
```
path = 'data_titanic.txt'
columns = [
    "pclass",
    "survived",
    "name",
    "sex",
    "age",
    "sibsp",
    "parch",
    "ticket",
    "fare",
    "cabin",
    "embarked",
    "boat",
    "body",
    "home.dest"
]

data = pd.read_csv(path, header=None, names=columns, skiprows=17)

data.index = range(1, len(data) + 1)

print("20 pierwszych wierszy danych: \n")
print(data.head(20))

data.to_csv('data_titanic_processed.csv', index=False)
```

```
20 pierwszych wierszy danych:

    pclass  survived                                             name  \
```

| | | | |
|---|---|---|---|
| 1 | 1 | 1 | Allen, Miss. Elisabeth Walton |
| 2 | 1 | 1 | Allison, Master. Hudson Trevor |
| 3 | 1 | 0 | Allison, Miss. Helen Loraine |
| 4 | 1 | 0 | Allison, Mr. Hudson Joshua Creighton |
| 5 | 1 | 0 | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) |
| 6 | 1 | 1 | Anderson, Mr. Harry |
| 7 | 1 | 1 | Andrews, Miss. Kornelia Theodosia |
| 8 | 1 | 0 | Andrews, Mr. Thomas Jr |
| 9 | 1 | 1 | Appleton, Mrs. Edward Dale (Charlotte Lamson) |
| 10 | 1 | 0 | Artagaveytia, Mr. Ramon |
| 11 | 1 | 0 | Astor, Col. John Jacob |
| 12 | 1 | 1 | Astor, Mrs. John Jacob (Madeleine Talmadge Force) |
| 13 | 1 | 1 | Aubart, Mme. Leontine Pauline |
| 14 | 1 | 1 | Barber, Miss. Ellen 'Nellie' |
| 15 | 1 | 1 | Barkworth, Mr. Algernon Henry Wilson |
| 16 | 1 | 0 | Baumann, Mr. John D |
| 17 | 1 | 0 | Baxter, Mr. Quigg Edmond |
| 18 | 1 | 1 | Baxter, Mrs. James (Helene DeLaudeniere Chaput) |
| 19 | 1 | 1 | Bazzani, Miss. Albina |
| 20 | 1 | 0 | Beattie, Mr. Thomson |

| | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | \ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | female | 29 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | |
| 2 | male | 0.9167 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | 11 | |
| 3 | female | 2 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | |
| 4 | male | 30 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | |
| 5 | female | 25 | 1 | 2 | 113781 | 151.55 | C22 C26 | S | ? | |
| 6 | male | 48 | 0 | 0 | 19952 | 26.55 | E12 | S | 3 | |
| 7 | female | 63 | 1 | 0 | 13502 | 77.9583 | D7 | S | 10 | |
| 8 | male | 39 | 0 | 0 | 112050 | 0 | A36 | S | ? | |
| 9 | female | 53 | 2 | 0 | 11769 | 51.4792 | C101 | S | D | |
| 10 | male | 71 | 0 | 0 | PC 17609 | 49.5042 | ? | C | ? | |
| 11 | male | 47 | 1 | 0 | PC 17757 | 227.525 | C62 C64 | C | ? | |
| 12 | female | 18 | 1 | 0 | PC 17757 | 227.525 | C62 C64 | C | 4 | |
| 13 | female | 24 | 0 | 0 | PC 17477 | 69.3 | B35 | C | 9 | |
| 14 | female | 26 | 0 | 0 | 19877 | 78.85 | ? | S | 6 | |
| 15 | male | 80 | 0 | 0 | 27042 | 30 | A23 | S | B | |
| 16 | male | ? | 0 | 0 | PC 17318 | 25.925 | ? | S | ? | |
| 17 | male | 24 | 0 | 1 | PC 17558 | 247.5208 | B58 B60 | C | ? | |
| 18 | female | 50 | 0 | 1 | PC 17558 | 247.5208 | B58 B60 | C | 6 | |
| 19 | female | 32 | 0 | 0 | 11813 | 76.2917 | D15 | C | 8 | |
| 20 | male | 36 | 0 | 0 | 13050 | 75.2417 | C6 | C | A | |

| | body | home.dest |
|---|---|---|
| 1 | ? | St Louis, MO |
| 2 | ? | Montreal, PQ / Chesterville, ON |
| 3 | ? | Montreal, PQ / Chesterville, ON |
| 4 | 135 | Montreal, PQ / Chesterville, ON |

```
5     ?  Montreal, PQ / Chesterville, ON
6     ?                      New York, NY
7     ?                        Hudson, NY
8     ?                       Belfast, NI
9     ?               Bayside, Queens, NY
10   22               Montevideo, Uruguay
11  124                      New York, NY
12    ?                      New York, NY
13    ?                     Paris, France
14    ?                                 ?
15    ?                     Hessle, Yorks
16    ?                      New York, NY
17    ?                      Montreal, PQ
18    ?                      Montreal, PQ
19    ?                                 ?
20    ?                      Winnipeg, MN
```

[3]:
```
print("Nazwy kolumn w danych: \n")
print(list(data.columns))
```

```
Nazwy kolumn w danych:

['pclass', 'survived', 'name', 'sex', 'age', 'sibsp', 'parch', 'ticket', 'fare',
'cabin', 'embarked', 'boat', 'body', 'home.dest']
```

**Dane zawierają informacje o pasażerach Titanica. Wyróżanimy 14 cech w zbiorze:**

- **pclass**: jest to klasa, w której podróżował pasażer. Wartości { 1, 2, 3, } odpowiadają pierwszej, drugiej i trzeciej klasie odpowiednio - ta kolumna odzwierciedla status społeczno-ekonomiczny pasażerów.
- **survived**: w tej kolumnie została zawarta informacja czy pasażer przeżył katastrofę (1 = przeżył, 0 = nie przeżył).
- **name**: ta cecha zawiera pełne imię i nazwisko pasażera.
- **sex**: ta kolumna określa płeć pasażera, male - mężczyzna lub female - kobieta.
- **age**: jest to wiek pasażera w latach ( występują również wartości np. 0.9167 które określają prawdopodobnie ułamki roczne niemowląt).
- **sibsp**: kolumna zawiera liczbę małżonków, lub rodzeństwa na pokładzie
- **parch**: kolumna zawiera liczba rodziców, lub dzieci na pokładzie.
- **ticket**: jest to numer biletu pasażera.
- **fare**: zawiera cenę biletu.
- **cabin**: kolumna zawiera numery kabin w której mieszkał pasażer.
- **embarked**: zawiera nazwe portu, w którym pasażer wsiadł na statek(C=Cherbourg, Q=Queenstown, S=Southampton).
- **boat**: cecha zawierająca numer łodzi ratunkowej, którą pasażer opuścił statek.
- **body**: cecha zawierająca numer identyfikacyjny ciała, jeśli pasażer zginął i jego ciało zostało odnalezione.
- **home.dest**: kolumna zawiera miejsce docelowe podróży.

Wiele rekordów zawiera brakujące wartości oznaczone symbolem “?”, które należy zmodyfikować.

Szczególnie jest to widoczne w kolumnach cabin, boat, body i home.dest. Z tego powodu przed użyciem funkcji pd.isnull().sum() oraz pd.isnull().mean, należy zamienić "?" na standardową reprezentację brakujących danych NaN przy użyciu funkcji replace().

### 3. Zmiana oznaczenia brakujących danych w zbiorze

[4]:
```
data.replace('?', np.nan, inplace=True)
print("20 początkowych wierszy po zamianie '?' na NaN: \n")
print(data.head(20))
```

```
20 początkowych wierszy po zamianie '?' na NaN:

    pclass  survived                                             name  \
1        1         1                    Allen, Miss. Elisabeth Walton
2        1         1                   Allison, Master. Hudson Trevor
3        1         0                     Allison, Miss. Helen Loraine
4        1         0             Allison, Mr. Hudson Joshua Creighton
5        1         0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
6        1         1                                 Anderson, Mr. Harry
7        1         1                Andrews, Miss. Kornelia Theodosia
8        1         0                             Andrews, Mr. Thomas Jr
9        1         1    Appleton, Mrs. Edward Dale (Charlotte Lamson)
10       1         0                           Artagaveytia, Mr. Ramon
11       1         0                            Astor, Col. John Jacob
12       1         1  Astor, Mrs. John Jacob (Madeleine Talmadge Force)
13       1         1                     Aubart, Mme. Leontine Pauline
14       1         1                      Barber, Miss. Ellen 'Nellie'
15       1         1            Barkworth, Mr. Algernon Henry Wilson
16       1         0                             Baumann, Mr. John D
17       1         0                         Baxter, Mr. Quigg Edmond
18       1         1  Baxter, Mrs. James (Helene DeLaudeniere Chaput)
19       1         1                           Bazzani, Miss. Albina
20       1         0                            Beattie, Mr. Thomson

       sex     age  sibsp  parch    ticket      fare    cabin embarked boat  \
1   female      29      0      0     24160  211.3375       B5        S    2
2     male  0.9167      1      2    113781    151.55  C22 C26        S   11
3   female       2      1      2    113781    151.55  C22 C26        S  NaN
4     male      30      1      2    113781    151.55  C22 C26        S  NaN
5   female      25      1      2    113781    151.55  C22 C26        S  NaN
6     male      48      0      0     19952     26.55      E12        S    3
7   female      63      1      0     13502   77.9583       D7        S   10
8     male      39      0      0    112050         0      A36        S  NaN
9   female      53      2      0     11769   51.4792     C101        S    D
10    male      71      0      0  PC 17609   49.5042      NaN        C  NaN
11    male      47      1      0  PC 17757   227.525  C62 C64        C  NaN
12  female      18      1      0  PC 17757   227.525  C62 C64        C    4
13  female      24      0      0  PC 17477      69.3      B35        C    9
14  female      26      0      0     19877     78.85      NaN        S    6
```

```
15    male      80      0      0     27042        30      A23        S    B
16    male     NaN      0      0  PC 17318    25.925      NaN        S  NaN
17    male      24      0      1  PC 17558  247.5208  B58 B60        C  NaN
18  female      50      0      1  PC 17558  247.5208  B58 B60        C    6
19  female      32      0      0     11813   76.2917      D15        C    8
20    male      36      0      0     13050   75.2417       C6        C    A
```

```
   body                        home.dest
1   NaN                    St Louis, MO
2   NaN  Montreal, PQ / Chesterville, ON
3   NaN  Montreal, PQ / Chesterville, ON
4   135  Montreal, PQ / Chesterville, ON
5   NaN  Montreal, PQ / Chesterville, ON
6   NaN                    New York, NY
7   NaN                      Hudson, NY
8   NaN                     Belfast, NI
9   NaN              Bayside, Queens, NY
10   22              Montevideo, Uruguay
11  124                    New York, NY
12  NaN                    New York, NY
13  NaN                   Paris, France
14  NaN                             NaN
15  NaN                   Hessle, Yorks
16  NaN                    New York, NY
17  NaN                    Montreal, PQ
18  NaN                    Montreal, PQ
19  NaN                             NaN
20  NaN                    Winnipeg, MN
```

### 4. Analiza udziału wartości brakujących w danych

[5]:
```
values_NaN = data.isnull().sum()
print("Liczba brakujących wartości dla każdej kolumny:\n")
print(values_NaN)

values_NaN_percent = data.isnull().mean() * 100
#dzięki pomożeniu *100 otrzymujemy wartość procentową
print("\nProcent brakujących wartości w każdej kolumnie:\n")
print(values_NaN_percent)
```

```
Liczba brakujących wartości dla każdej kolumny:

pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
```

```
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64

Procent brakujących wartości w każdej kolumnie:

pclass        0.000000
survived      0.000000
name          0.000000
sex           0.000000
age          20.091673
sibsp         0.000000
parch         0.000000
ticket        0.000000
fare          0.076394
cabin        77.463713
embarked      0.152788
boat         62.872422
body         90.756303
home.dest    43.086325
dtype: float64
```

[6]:
```python
# wybór kolumn, które posiadają wartości NaN
columns_with_nan = data.columns[data.isnull().any()].tolist()
print("Kolumny zawierające wartości NaN:", columns_with_nan)
```

```
Kolumny zawierające wartości NaN: ['age', 'fare', 'cabin', 'embarked', 'boat',
'body', 'home.dest']
```

[7]:
```python
# mapowanie po kolumnach z NaN względem 'survived' i 'pclass'
for column in columns_with_nan:
    col_nan = f'{column}Null'
    data[col_nan] = np.where(data[column].isnull(), 1, 0)

    # obliczanie udziału % NaN dla 'survived'
    mean_by_survived = data.groupby('survived')[col_nan].mean() * 100
    print(f"Udział % NaN w kolumnie '{column}' względem 'survived':\n"
        f"{mean_by_survived}\n")

    # obliczanie udziału % NaN dla 'pclass'
    mean_by_pclass = data.groupby('pclass')[col_nan].mean() * 100
    print(f"Udział % NaN w kolumnie '{column}' względem 'pclass':\n"
        f"{mean_by_pclass}\n")
```

```
Udział % NaN w kolumnie 'age' względem 'survived':
survived
0    23.485785
1    14.600000
Name: ageNull, dtype: float64

Udział % NaN w kolumnie 'age' względem 'pclass':
pclass
1    12.074303
2     5.776173
3    29.337094
Name: ageNull, dtype: float64

Udział % NaN w kolumnie 'fare' względem 'survived':
survived
0    0.123609
1    0.000000
Name: fareNull, dtype: float64

Udział % NaN w kolumnie 'fare' względem 'pclass':
pclass
1    0.000000
2    0.000000
3    0.141044
Name: fareNull, dtype: float64

Udział % NaN w kolumnie 'cabin' względem 'survived':
survived
0    87.391842
1    61.400000
Name: cabinNull, dtype: float64

Udział % NaN w kolumnie 'cabin' względem 'pclass':
pclass
1    20.743034
2    91.696751
3    97.743300
Name: cabinNull, dtype: float64

Udział % NaN w kolumnie 'embarked' względem 'survived':
survived
0    0.0
1    0.4
Name: embarkedNull, dtype: float64

Udział % NaN w kolumnie 'embarked' względem 'pclass':
pclass
1    0.619195
```

```
2    0.000000
3    0.000000
Name: embarkedNull, dtype: float64

Udział % NaN w kolumnie 'boat' względem 'survived':
survived
0    98.887515
1     4.600000
Name: boatNull, dtype: float64

Udział % NaN w kolumnie 'boat' względem 'pclass':
pclass
1    37.770898
2    59.566787
3    75.599436
Name: boatNull, dtype: float64

Udział % NaN w kolumnie 'body' względem 'survived':
survived
0     85.043263
1    100.000000
Name: bodyNull, dtype: float64

Udział % NaN w kolumnie 'body' względem 'pclass':
pclass
1    89.164087
2    88.808664
3    92.242595
Name: bodyNull, dtype: float64

Udział % NaN w kolumnie 'home.dest' względem 'survived':
survived
0    50.803461
1    30.600000
Name: home.destNull, dtype: float64

Udział % NaN w kolumnie 'home.dest' względem 'pclass':
pclass
1    10.526316
2     5.776173
3    72.496474
Name: home.destNull, dtype: float64
```

[8]:
```
# PRZYKŁAD DLA WYBRANYCH KOLUMN:
# mapoowanie kolumny boat względem survived
# połączona funkcja w jedną linijkę wykonana przy pomocy AI
```

```
mean_boat_by_survived = (
    data.assign(BoatNull2=np.where(data['boat'].isnull(), 1, 0))
    .groupby('survived')['BoatNull2']
    .mean() * 100
)
print("Udział % Nan kolumny 'boat' względem 'survived':\n")
print(mean_boat_by_survived)
```

```
Udział % Nan kolumny 'boat' względem 'survived':

survived
0    98.887515
1     4.600000
Name: BoatNull2, dtype: float64
```

**Analiza użyskanych wyników:** Kolumny takie jak pclass, survived, name, sex, sibsp, parch, ticket nie zawierają brakujących danych (w 100 procentach są kompletne).

Kolumny zawierające brakujące dane z uzasadnieniem: - **fare** - kolumna zawiera tylko 1 brakującą wartość ($< 1\%$ ogółu danych w kolumnie), może to być spowodowane błędem w gromadzeniu danych lub wynikać z niedopatrzeń przy rejstrowaniu danych finansowych. Raczej nie jest to powiązane z żadna inną cechą. Po analizie brakujących danych w kontekście przeżywalności, można wywnioskować, że te dwie cechy nie ma między nimi związku, ponieważ udział wynosi $< 1\%$.

- **embarked** - kolumna zawiera tylko 2 wartości NaN ($< 1\%$ ogółu danych w kolumnie), może wynikać z przypadkowych błędów podczas sporządzania dokumentacji. Analiza udziału brakujących danych w kontekście przeżywalności wskazuje, że nie istnieje istotny związek między brakującymi wartościami a przeżywalnością. Wśród pasażerów, którzy nie przeżyli, nie odnotowano brakujących danych, natomiast dla osób, które przeżyły, udział brakujących danych to tylko 0.4%.
- **age** - kolumna zawiera 263 wartości brakujące ($\sim 20\%$ ogółu danych w kolumnie), może to wynikać niekompletnie prowadzonej dokumentacji, a także możliwe, że wiek pasażerów nie był obowiązkowy do podawania przy rejestrowaniu na statku (prawdopodobnie na początku 20 wieku wiek nie zawsze był uznawany za ważną informację do zapisu). Po wyliczeniu udziału brakujących wartości w zależności od klasy, można stwierdzić, że trzecia klasa ma największy odsetek wartości NaN co może sugerować, iż nie przykładano się do dokumentaji pasażerów najniższych klas. Po analizie brakujących danych w kontekście przeżywalności, można wywnioskować, że istnieje mało istotny związek między brakującymi wartościami a przeżywalnością, ponieważ udział wartość NaN wsród osób, które nie przeżyły wynosi ok. 24%.
- **home.dest** - kolumna zawiera 564 wartości NaN ($\sim 43\%$ ogółu danych w kolumnie), co może być spowodowane nieobowiązkowością podawania tych danych przez pasażerów lub ze względu na to, że niektórzy pasażerowie płynęli "przed siebie", szukając pracy/miejsca do osiedlenia się, bez konkretnego celu podróży. Wynikiem tego może być również utrata części dokumentacji podczas zatonięcia statku lub zasada, która nie wymagała zbierania takich informacji od osób podróżujących niższymi klasami — zauważalna jest pewna zależność między niższą klasą a brakiem informacji o miejscu docelowym. Po wyliczeniu udziału brakujących wartości w zależności od klasy, można stwierdzić, że najwyższy odsetek brakujących wartości

występuje w klasie 3, a najniższy w pierwszej klasie, co potwierdza moje wcześniejsze wnioski odnośnie tej cechy. Natomiast po analizie brakujących danych w kontekście przeżywalności, można wywnioskować, że istnieje dość istotny związek między brakującymi wartościami, a przeżywalnością, ponieważ udział wartość NaN wsród osób, które nie przeżyły wynosi ok. 50%.

- **boat** - kolumna zawiera 823 wartości brakujące (~ 63% ogółu danych w kolumnie), może to być następstwem ograniczonej dostępności łodzi ratunkowych - nie każdy pasażer miał przypisane miejsce w łodzi. Analiza brakujących wartości w zależności od klasy pokazuje, że wśród pasażerów trzeciej klasy około 75% nie miało przypisanego miejsca w łodzi ratunkowej, podczas gdy w pierwszej klasie odsetek ten wynosił tylko około 38%. Dodatkowo, po analizie brakujących danych w kontekście przeżywalności, można stwierdzić, że około 98% pasażerów, którzy zginęli, nie miało miejsca w łodziach ratunkowych. Wynika z tego, że posiadanie zagwarantowanego miejsca w łodzi ratunkowej znacząco zwiększało szanse na przeżycie katastrofy. Jest tu wyraźne zauważalna zależność pomiędzy kolumnami boat, a survived.
- **cabin** - kolumna zawiera aż 1014 wartości NaN (~ 77% ogółu danych w kolumnie), może to wynikać z faktu, że pasażerowie z niszych klas (2 i 3 klasa) nie mieli przydzielonych lub rejestrowanych kabin. Analiza brakujących wartości w zależności od klasy pokazuje, że prawie wszyscy pasażerowie drugiej i trzeciej klasy nie mieli zapewnionych miejsc zakwaterowania. Obliczenia pokazują, że w niższych klasach brakujące wartości stanowią kolejno: druga klasa - ok. 92%, trzecia klasa - ok. 98%, co potwierdza moje wcześniejsze wnioski. Natomiast po analizie brakujących danych w kontekście przeżywalności, można wywnioskować, że istnieje istotny związek między brakującymi wartościami a przeżywalnością, ponieważ udział wartość NaN wsród osób, które nie przeżyły wynosi ok. 87%. Te dwie cechy są istotnie powiązane ze sobą.
- **body** - kolumna zawiera aż 1188 wartości brakujące (~ 91% ogółu danych w kolumnie), może to być następstwem nieodnalezienia wielu ciał ofiar katastrofy lub braku identyfikacji ciała (brak możliwości przypisania danych pasażera do znalezionego ciała). Analiza brakujących wartości w zależności od klasy pokazuje w miarę równomierny rozkład udziału wartości NaN pomiędzy klasami. Po analizie brakujących danych w kontekście przeżywalności, można wywnioskować, że ok. 85 procent pasażerów, którzy nie przeżyli nie mają zarejstrowanych numeru ciała, co może być skutkiem nieodnalezienia ciała po katastrofie. Natomiast 100-procentowy wynik wśród osób, które przeżyły, jest zrozumiały, ponieważ żywym pasażerom nie przypisuje się numeru ciała.

**5. Analiza rodzajów brakujących danych w zbiorze**

- całkowicie przypadkowe (MCAR): dane brakujące w kolumnach ‘fare’ oraz ‘embarked’ można klasyfikować jako całkowicie przypadkowe (MCAR), gdyż ich udział procentowy w całości danych dla tych kolumn jest mniejszy niż 1%, nie jest związany z innymi danymi i nie ma istotnego wpływu na analizowane zmienne.
- przypadkowe (MAR): dane brakujące w kolumnie ‘age’ można klasyfikować jako przypadkowe (MAR), gdyż mogą być związane z klasą społeczną (pclass), gdzie braki są częściej obserwowane w niższych klasach, co może świadczyć o mniejszej dokładności, dostępności dokumentacji dla pasażerów z niższych klas społecznych lub braku wymagalności podania wieku.

- nie przypadkowe (MNAR): dane brakujące w kolumnach ‘body’, ‘boat’, ‘cabin’, ‘home.dest’ można klasyfikować jako nie przypadkowe (MNAR), gdyż brak tych danych jest bezpośrednio związany z faktem, czy pasażer przeżył. Przykładowo osoby, które nie przeżyły, rzadziej miały przypisane numery łodzi, a numery ciał są brakujące głównie dla osób, których ciała nie zostały odnalezione lub zidentyfikowane.

**6. Polecenie: po powyższej analizie odpowiedz na pytanie w jaki sposób należy postąpić z brakującymi wartościami.**

- **fare** - jedna brakująca wartość może być łatwo uzupełniona przez medianę lub średnią wartość biletu.
- **embarked** - dwie brakujące wartości można uzupełnić najczęściej występującą wartością (modą) w tej kolumnie.
- **age** - wartości te można imputować przy użyciu zaawansowanych technik - np. imputacja wielowymiarowa.
- **home.dest** - ze względu na niemożliwość odtworzenia celu podróży i dużego odsetku brakujących wartości proponuje pozostawić te wartości jako ‘unknown_dest’.
- **boat** - ze względu na duże powiązania z innymi cechami proponuje pozostawienie tych wartości jako ‘no_boat’.
- **cabin** - wysoki odsetek brakujących danych, szczególnie w niższych klasach, może wskazywać na to, że wielu pasażerów nie miało przydzielonych kabin, dlatego proponuje przypisanie wartości ‘unknown_cabin’.
- **body** ze względu na bardzo wysoki procent brakujących danych i związek tych danych z pasażerami, którzy zginęli i których ciała nie zostały odnalezione lub zidentyfikowane, najlepiej jest pozostawić te wartości jako brakujące albo jako ‘no_body’.