klib library used work little bit fast

August 16, 2024

[28]:
```
!pip install klib
```

```
Requirement already satisfied: klib in c:\users\vikas\anaconda3\lib\site-
packages (1.3.1)
Requirement already satisfied: Jinja2<4.0.0,>=3.1.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (3.1.2)
Requirement already satisfied: seaborn>=0.12.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (0.12.2)
Requirement already satisfied: numpy<2.0.0,>=1.26.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (1.26.4)
Requirement already satisfied: pandas<3.0,>=1.4 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (2.2.2)
Requirement already satisfied: matplotlib<4.0.0,>=3.6.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (3.7.0)
Requirement already satisfied: scipy<2.0.0,>=1.10.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (1.10.0)
Requirement already satisfied: plotly<6.0.0,>=5.11.0 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (5.23.0)
Requirement already satisfied: screeninfo<0.9.0,>=0.8.1 in
c:\users\vikas\anaconda3\lib\site-packages (from klib) (0.8.1)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\users\vikas\anaconda3\lib\site-packages (from Jinja2<4.0.0,>=3.1.0->klib)
(2.1.1)
Requirement already satisfied: cycler>=0.10 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(0.11.0)
Requirement already satisfied: packaging>=20.0 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(22.0)
Requirement already satisfied: pillow>=6.2.0 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(9.4.0)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(2.8.2)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(4.25.0)
```

```
Requirement already satisfied: kiwisolver>=1.0.1 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(1.4.4)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(1.0.5)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\vikas\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.6.0->klib)
(3.0.9)
Requirement already satisfied: tzdata>=2022.7 in
c:\users\vikas\anaconda3\lib\site-packages (from pandas<3.0,>=1.4->klib)
(2024.1)
Requirement already satisfied: pytz>=2020.1 in
c:\users\vikas\anaconda3\lib\site-packages (from pandas<3.0,>=1.4->klib)
(2022.7)
Requirement already satisfied: tenacity>=6.2.0 in
c:\users\vikas\anaconda3\lib\site-packages (from plotly<6.0.0,>=5.11.0->klib)
(8.0.1)
Requirement already satisfied: six>=1.5 in c:\users\vikas\anaconda3\lib\site-
packages (from python-dateutil>=2.7->matplotlib<4.0.0,>=3.6.0->klib) (1.16.0)
```

[ ]:
```
# klib.describe - functions for visualizing datasets
- klib.cat_plot(df) # returns a visualization of the number and frequency of␣
 ↪categorical features
- klib.corr_mat(df) # returns a color-encoded correlation matrix
- klib.corr_plot(df) # returns a color-encoded heatmap, ideal for correlations
- klib.corr_interactive_plot(df, split="neg").show() # returns an interactive␣
 ↪correlation plot using plotly
- klib.dist_plot(df) # returns a distribution plot for every numeric feature
- klib.missingval_plot(df) # returns a figure containing information about␣
 ↪missing values
```

[ ]:
```
 klib.clean - functions for cleaning datasets
- klib.data_cleaning(df) # performs datacleaning (drop duplicates & empty rows/
 ↪cols, adjust dtypes,...)
- klib.clean_column_names(df) # cleans and standardizes column names, also␣
 ↪called inside data_cleaning()
- klib.convert_datatypes(df) # converts existing to more efficient dtypes, also␣
 ↪called inside data_cleaning()
- klib.drop_missing(df) # drops missing values, also called in data_cleaning()
- klib.mv_col_handling(df) # drops features with high ratio of missing vals␣
 ↪based on informational content
- klib.pool_duplicate_subsets(df) # pools subset of cols based on duplicates␣
 ↪with min. loss of information
```

[29]:
```
## klib library  it is used importing, cleaning analysing preprocessing data do␣
 ↪lot of things
```

[30]:
```
##Import necessary library
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

[31]:
```
data=sns.load_dataset('titanic')
```

[32]:
```
data.head()
```

[32]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True

[33]:
```
import klib
```

[39]:
```
klib.cat_plot(data)
```

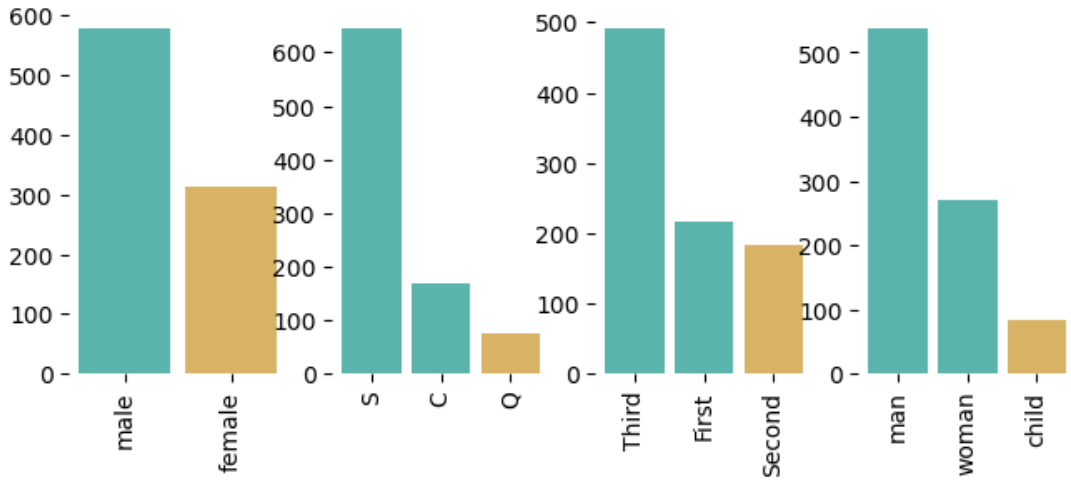
```
C:\Users\Vikas\anaconda3\lib\site-packages\klib\describe.py:122: FutureWarning:
Setting an item of incompatible dtype is deprecated and will raise an error in a
future version of pandas. Value '10' has dtype incompatible with bool, please
explicitly cast to a compatible dtype first.
  data.loc[data[col].isin(value_counts_idx_top), col] = 10
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[39], line 1
----> 1 klib.cat_plot(data)

File ~\anaconda3\lib\site-packages\klib\describe.py:127, in cat_plot(data,␣
 ↪figsize, top, bottom, bar_color_top, bar_color_bottom)
    124 data.loc[((data[col] != 10) & (data[col] != 0)), col] = 5  # noqa:␣
 ↪PLR2004
    125 data[col] = data[col].rolling(2, min_periods=1).mean()
--> 127 value_counts_idx_top = [elem[:20] for elem in value_counts_idx_top]
    128 value_counts_idx_bot = [elem[:20] for elem in value_counts_idx_bot]
    129 sum_top = sum(value_counts_top)
```

```
File ~\anaconda3\lib\site-packages\klib\describe.py:127, in <listcomp>(.0)
    124 data.loc[((data[col] != 10) & (data[col] != 0)), col] = 5  # noqa:␣
 ↪PLR2004
    125 data[col] = data[col].rolling(2, min_periods=1).mean()
--> 127 value_counts_idx_top = [elem[:20] for elem in value_counts_idx_top]
    128 value_counts_idx_bot = [elem[:20] for elem in value_counts_idx_bot]
    129 sum_top = sum(value_counts_top)

TypeError: 'bool' object is not subscriptable
```

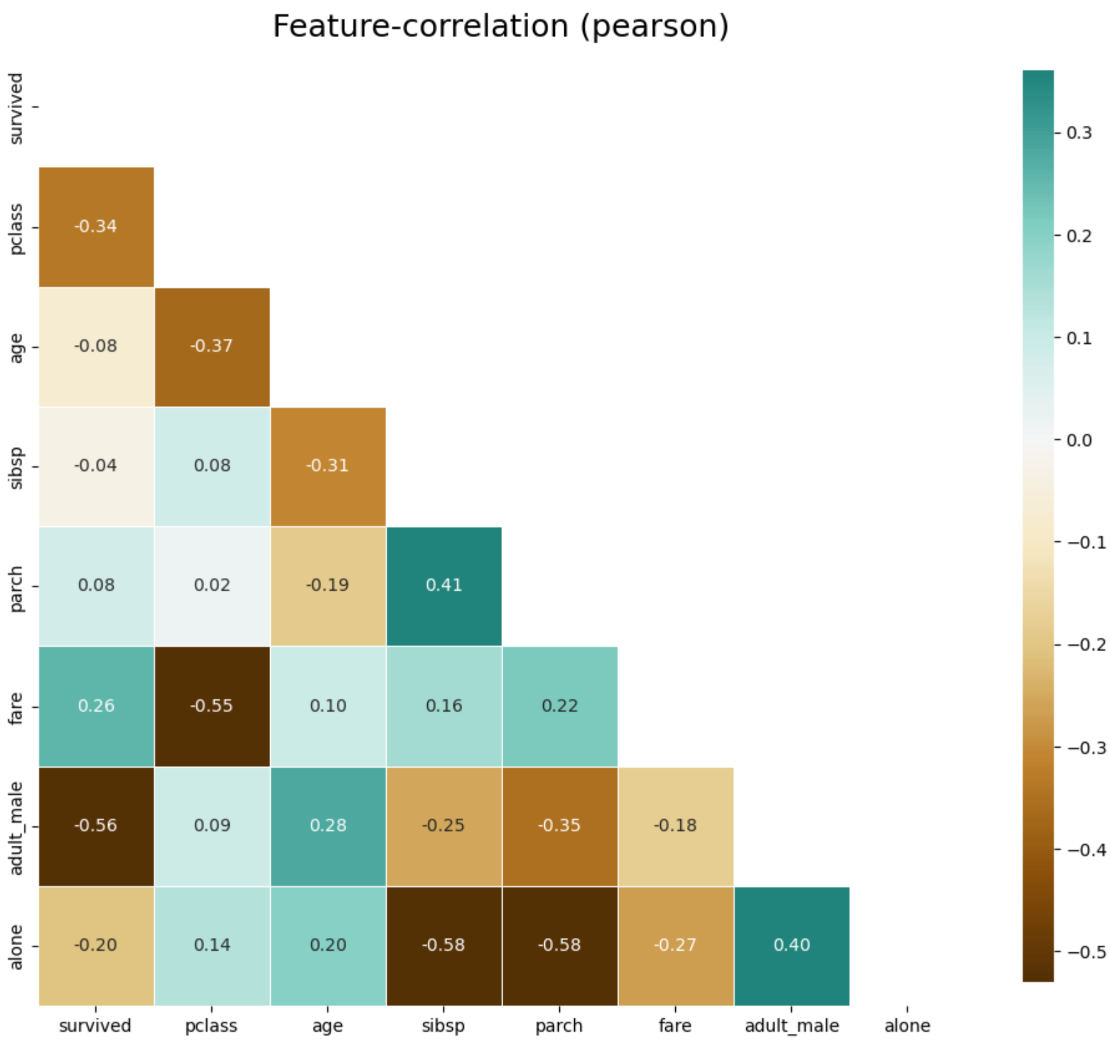


Unique values: 2	Unique values: 3	Unique values: 3	Unique values: 3
Top 1: 577 (64.8%)	Top 2: 812 (91.1%)	Top 2: 707 (79.3%)	Top 2: 808 (90.7%)
Bot 1: 314 (35.2%)	Bot 1: 77 (8.6%)	Bot 1: 184 (20.7%)	Bot 1: 83 (9.3%)

[ ]:
```
klib.corr_mat(data)
```

[40]:
```
klib.corr_plot(data)
```

[40]:
```
<Axes: title={'center': 'Feature-correlation (pearson)'}>
```

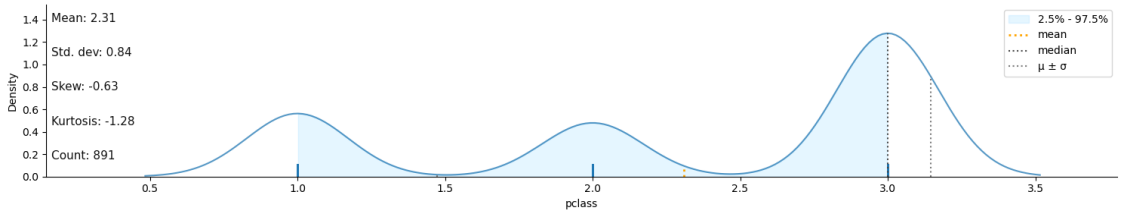


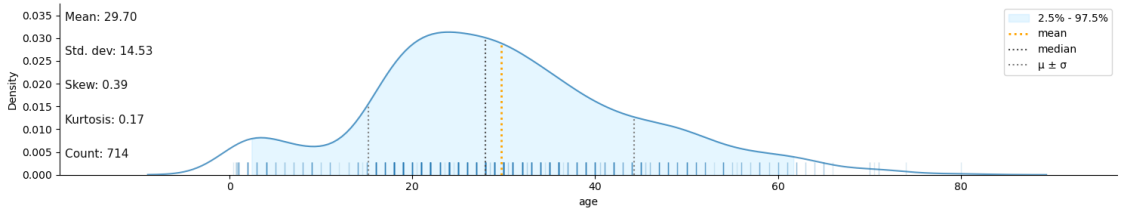
[41]:
```
klib.dist_plot(data)
```

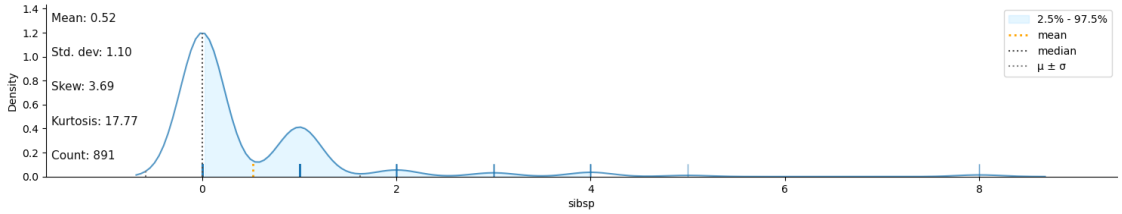
```
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
```

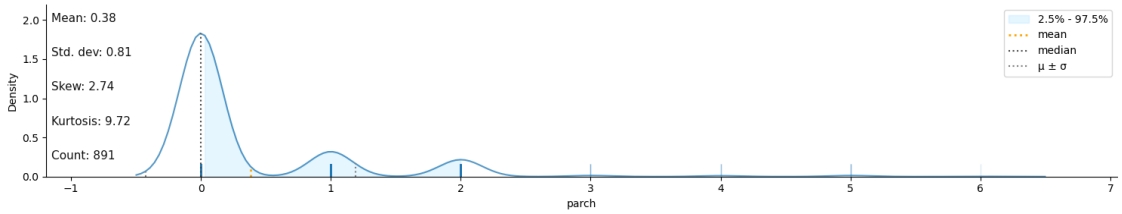
```
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\Vikas\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

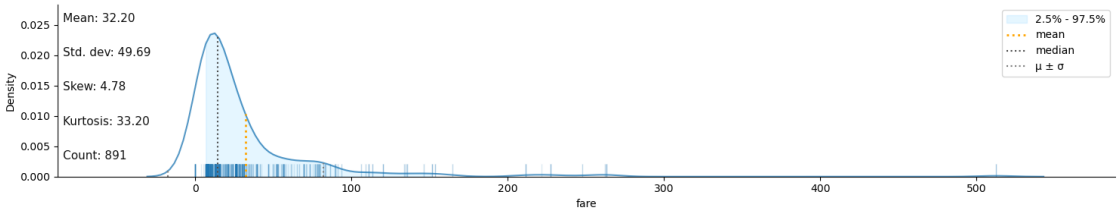
[41]: <Axes: xlabel='fare', ylabel='Density'>





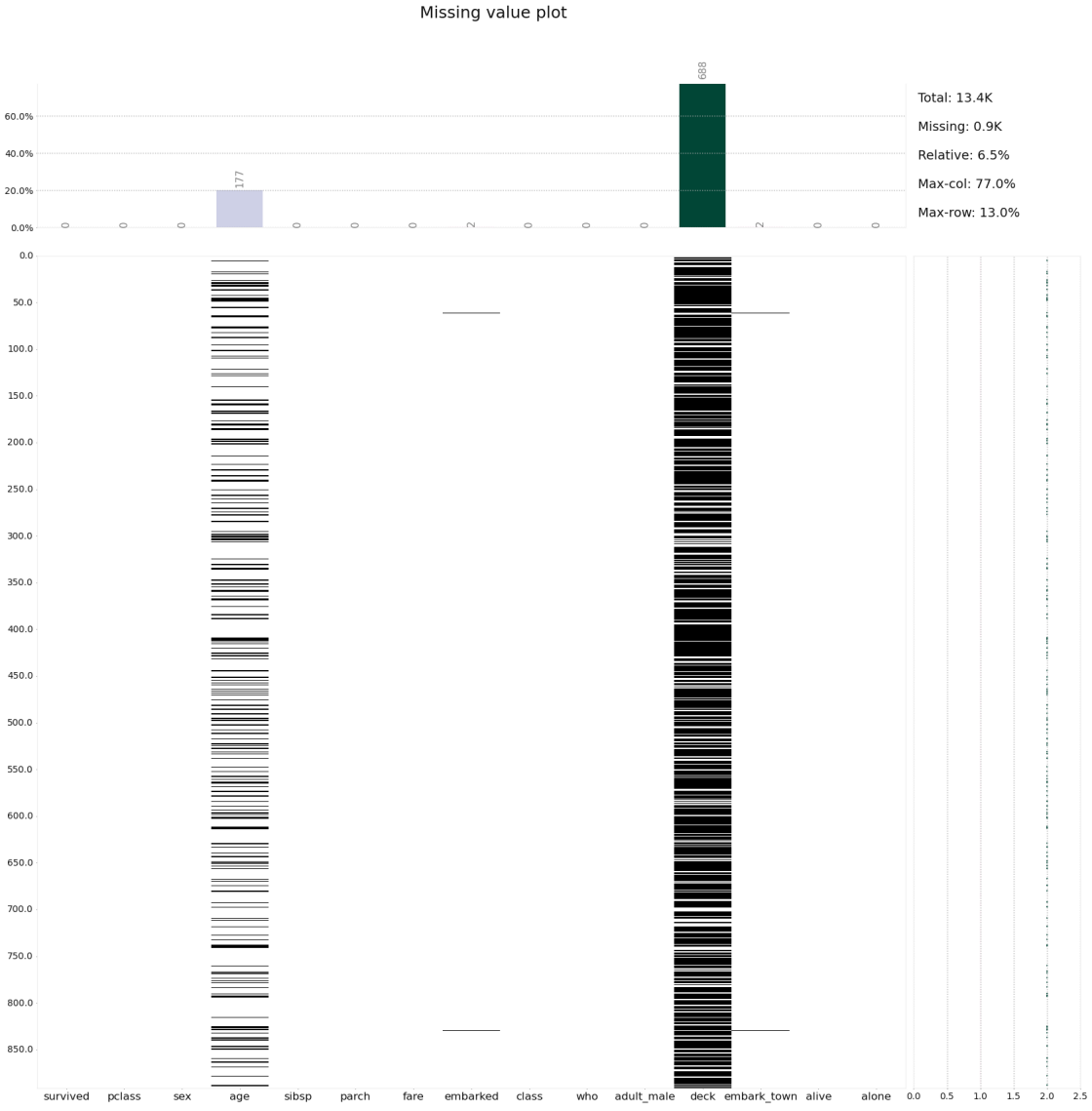






[42]:
```
klib.missingval_plot(data)
```

[42]: GridSpec(6, 6)



[47]:
```
klib.corr_interactive_plot(data, split="neg").show()
```

```
Displaying negative correlations. Specify a negative "threshold" to limit the
results further.
```

[49]:
```

```

[43]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

1 cleaning

[36]:
```
data_clean=klib.data_cleaning(data)
```

```
Shape of cleaned data: (784, 15) - Remaining NAs: 692


Dropped rows: 107
     of which 107 duplicates. (Rows (first 150 shown): [47, 76, 77, 87, 95, 101,
121, 133, 173, 196, 198, 201, 213, 223, 241, 260, 274, 295, 300, 304, 313, 320,
324, 335, 343, 354, 355, 358, 359, 364, 368, 384, 409, 410, 413, 418, 420, 425,
428, 431, 454, 459, 464, 466, 470, 476, 481, 485, 488, 490, 494, 500, 511, 521,
522, 526, 531, 560, 563, 564, 568, 573, 588, 589, 598, 601, 612, 613, 614, 635,
636, 640, 641, 644, 646, 650, 656, 666, 674, 692, 696, 709, 732, 733, 734, 738,
739, 757, 758, 760, 773, 790, 792, 800, 808, 832, 837, 838, 844, 846, 859, 863,
870, 877, 878, 884, 886])
```

```
Dropped columns: 0
     of which 0 single valued.     Columns: []
Dropped missing values: 177
Reduced memory by at least: 0.06 MB (-75.0%)
```

[38]:
```
data_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 784 entries, 0 to 783
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     784 non-null    int8
 1   pclass       784 non-null    int8
 2   sex          784 non-null    category
 3   age          678 non-null    float32
 4   sibsp        784 non-null    int8
 5   parch        784 non-null    int8
 6   fare         784 non-null    float32
 7   embarked     782 non-null    category
 8   class        784 non-null    category
 9   who          784 non-null    category
 10  adult_male   784 non-null    boolean
 11  deck         202 non-null    category
 12  embark_town  782 non-null    category
 13  alive        784 non-null    category
 14  alone        784 non-null    boolean
dtypes: boolean(2), category(7), float32(2), int8(4)
memory usage: 18.8 KB
```

[ ]:
```

```