In [1]:
```
pip install klib
```

```
Requirement already satisfied: klib in c:\users\shree\anaconda3\lib\site-packages (1.1.2)
Requirement already satisfied: numpy<2.0.0,>=1.16.3 in c:\users\shree\anaconda3\lib\site-packages (from klib) (1.23.5)
Requirement already satisfied: screeninfo<0.9.0,>=0.8.1 in c:\users\shree\anaconda3\lib\site-packages (from klib) (0.8.1)
Requirement already satisfied: pandas<3.0,>=1.2 in c:\users\shree\anaconda3\lib\site-packages (from klib) (1.5.3)
Requirement already satisfied: matplotlib<4.0.0,>=3.0.3 in c:\users\shree\anaconda3\lib\site-packages (from klib) (3.7.0)
Requirement already satisfied: seaborn>=0.11.2 in c:\users\shree\anaconda3\lib\site-packages (from klib) (0.12.2)
Requirement already satisfied: scipy<2.0.0,>=1.1.0 in c:\users\shree\anaconda3\lib\site-packages (from klib) (1.10.0)
Requirement already satisfied: Jinja2<4.0.0,>=3.0.3 in c:\users\shree\anaconda3\lib\site-packages (from klib) (3.1.2)
Requirement already satisfied: plotly<6.0.0,>=5.2.2 in c:\users\shree\anaconda3\lib\site-packages (from klib) (5.9.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\shree\anaconda3\lib\site-packages (from Jinja2<4.0.0,>=3.0.3->klib) (2.1.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (4.2
5.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (1.0.
5)
Requirement already satisfied: pillow>=6.2.0 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (3.0.
9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib)
(2.8.2)
Requirement already satisfied: cycler>=0.10 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (0.11.0)
Requirement already satisfied: packaging>=20.0 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (22.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\shree\anaconda3\lib\site-packages (from matplotlib<4.0.0,>=3.0.3->klib) (1.
4.4)
Requirement already satisfied: pytz>=2020.1 in c:\users\shree\anaconda3\lib\site-packages (from pandas<3.0,>=1.2->klib) (2022.7)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\shree\anaconda3\lib\site-packages (from plotly<6.0.0,>=5.2.2->klib) (8.0.1)
Requirement already satisfied: six>=1.5 in c:\users\shree\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib<4.0.0,>=3.0.
3->klib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import seaborn as sns
import pandas as pd
```

In [3]:
```
df = pd.read_csv('Titanic.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	Age	Cabin	Embarked	Fare	Name	Parch	PassengerId	Pclass	Sex	SibSp	Survived	Ticket	Title	Family_Size
0	34.5	NaN	Q	7.8292	Kelly, Mr. James	0	892	3	male	0	NaN	330911	Mr	0
1	47.0	NaN	S	7.0000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	NaN	363272	Mrs	1
2	62.0	NaN	Q	9.6875	Myles, Mr. Thomas Francis	0	894	2	male	0	NaN	240276	Mr	0
3	27.0	NaN	S	8.6625	Wirz, Mr. Albert	0	895	3	male	0	NaN	315154	Mr	0
4	22.0	NaN	S	12.2875	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	NaN	3101298	Mrs	2

In [7]:
```
import klib
```

In [5]:
```
# 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
```

In [8]:
```
#klib.data_cleaning(df) # performs datacleaning (drop duplicates & empty rows/cols, adjust dtypes,...)
klib.data_cleaning(df)
```

```
Shape of cleaned data: (418, 13) - Remaining NAs: 327


Dropped rows: 0
     of which 0 duplicates. (Rows (first 150 shown): [])

Dropped columns: 1
     of which 0 single valued.     Columns: []
Dropped missing values: 418
Reduced memory by at least: 0.02 MB (-50.0%)
```

Out[8]:

	age	cabin	embarked	fare	name	parch	passenger_id	pclass	sex	sib_sp	ticket	title	family_size
0	34.5	<NA>	Q	7.829200	Kelly, Mr. James	0	892	3	male	0	330911	Mr	0
1	47.0	<NA>	S	7.000000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	363272	Mrs	1
2	62.0	<NA>	Q	9.687500	Myles, Mr. Thomas Francis	0	894	2	male	0	240276	Mr	0
3	27.0	<NA>	S	8.662500	Wirz, Mr. Albert	0	895	3	male	0	315154	Mr	0
4	22.0	<NA>	S	12.287500	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	3101298	Mrs	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	30.0	<NA>	S	8.050000	Spector, Mr. Woolf	0	1305	3	male	0	A.5. 3236	Mr	0
414	39.0	C105	C	108.900002	Oliva y Ocana, Dona. Fermina	0	1306	1	female	0	PC 17758	Mrs	0
415	38.5	<NA>	S	7.250000	Saether, Mr. Simon Sivertsen	0	1307	3	male	0	SOTON/O.Q. 3101262	Mr	0
416	30.0	<NA>	S	8.050000	Ware, Mr. Frederick	0	1308	3	male	0	359309	Mr	0
417	4.0	<NA>	C	22.358299	Peter, Master. Michael J	1	1309	3	male	1	2668	Master	2

418 rows × 13 columns

In [9]:
```
#klib.clean_column_names(df) # cleans and standardizes column names, also called inside data_cleaning()
klib.clean_column_names(df)
```

Out[9]:

	age	cabin	embarked	fare	name	parch	passenger_id	pclass	sex	sib_sp	survived	ticket	title	family_size
0	34.5	NaN	Q	7.8292	Kelly, Mr. James	0	892	3	male	0	NaN	330911	Mr	0
1	47.0	NaN	S	7.0000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	NaN	363272	Mrs	1
2	62.0	NaN	Q	9.6875	Myles, Mr. Thomas Francis	0	894	2	male	0	NaN	240276	Mr	0
3	27.0	NaN	S	8.6625	Wirz, Mr. Albert	0	895	3	male	0	NaN	315154	Mr	0
4	22.0	NaN	S	12.2875	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	NaN	3101298	Mrs	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	30.0	NaN	S	8.0500	Spector, Mr. Woolf	0	1305	3	male	0	NaN	A.5. 3236	Mr	0
414	39.0	C105	C	108.9000	Oliva y Ocana, Dona. Fermina	0	1306	1	female	0	NaN	PC 17758	Mrs	0
415	38.5	NaN	S	7.2500	Saether, Mr. Simon Sivertsen	0	1307	3	male	0	NaN	SOTON/O.Q. 3101262	Mr	0
416	30.0	NaN	S	8.0500	Ware, Mr. Frederick	0	1308	3	male	0	NaN	359309	Mr	0
417	4.0	NaN	C	22.3583	Peter, Master. Michael J	1	1309	3	male	1	NaN	2668	Master	2

418 rows × 14 columns

In [10]:
```
#klib.convert_datatypes(df) # converts existing to more efficient dtypes, also called inside data_cleaning()
klib.convert_datatypes(df)
```

Out[10]:

	Age	Cabin	Embarked	Fare	Name	Parch	PassengerId	Pclass	Sex	SibSp	Survived	Ticket	Title	Family_Size
0	34.5	<NA>	Q	7.829200	Kelly, Mr. James	0	892	3	male	0	NaN	330911	Mr	0
1	47.0	<NA>	S	7.000000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	NaN	363272	Mrs	1
2	62.0	<NA>	Q	9.687500	Myles, Mr. Thomas Francis	0	894	2	male	0	NaN	240276	Mr	0
3	27.0	<NA>	S	8.662500	Wirz, Mr. Albert	0	895	3	male	0	NaN	315154	Mr	0
4	22.0	<NA>	S	12.287500	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	NaN	3101298	Mrs	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	30.0	<NA>	S	8.050000	Spector, Mr. Woolf	0	1305	3	male	0	NaN	A.5. 3236	Mr	0
414	39.0	C105	C	108.900002	Oliva y Ocana, Dona. Fermina	0	1306	1	female	0	NaN	PC 17758	Mrs	0
415	38.5	<NA>	S	7.250000	Saether, Mr. Simon Sivertsen	0	1307	3	male	0	NaN	SOTON/O.Q. 3101262	Mr	0
416	30.0	<NA>	S	8.050000	Ware, Mr. Frederick	0	1308	3	male	0	NaN	359309	Mr	0
417	4.0	<NA>	C	22.358299	Peter, Master. Michael J	1	1309	3	male	1	NaN	2668	Master	2

418 rows × 14 columns

In [11]:
```
#klib.drop_missing(df) # drops missing values, also called in data_cleaning()
klib.drop_missing(df)
```

Out[11]:

	Age	Cabin	Embarked	Fare	Name	Parch	PassengerId	Pclass	Sex	SibSp	Ticket	Title	Family_Size
0	34.5	NaN	Q	7.8292	Kelly, Mr. James	0	892	3	male	0	330911	Mr	0
1	47.0	NaN	S	7.0000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	363272	Mrs	1
2	62.0	NaN	Q	9.6875	Myles, Mr. Thomas Francis	0	894	2	male	0	240276	Mr	0
3	27.0	NaN	S	8.6625	Wirz, Mr. Albert	0	895	3	male	0	315154	Mr	0
4	22.0	NaN	S	12.2875	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	3101298	Mrs	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	30.0	NaN	S	8.0500	Spector, Mr. Woolf	0	1305	3	male	0	A.5. 3236	Mr	0
414	39.0	C105	C	108.9000	Oliva y Ocana, Dona. Fermina	0	1306	1	female	0	PC 17758	Mrs	0
415	38.5	NaN	S	7.2500	Saether, Mr. Simon Sivertsen	0	1307	3	male	0	SOTON/O.Q. 3101262	Mr	0
416	30.0	NaN	S	8.0500	Ware, Mr. Frederick	0	1308	3	male	0	359309	Mr	0
417	4.0	NaN	C	22.3583	Peter, Master. Michael J	1	1309	3	male	1	2668	Master	2

418 rows × 13 columns

In [12]:
```
#klib.mv_col_handling(df) # drops features with high ratio of missing vals based on informational content
klib.mv_col_handling(df)
```

Out[12]:

	Age	Embarked	Fare	Name	Parch	PassengerId	Pclass	Sex	SibSp	Survived	Ticket	Title	Family_Size
0	34.5	Q	7.8292	Kelly, Mr. James	0	892	3	male	0	NaN	330911	Mr	0
1	47.0	S	7.0000	Wilkes, Mrs. James (Ellen Needs)	0	893	3	female	1	NaN	363272	Mrs	1
2	62.0	Q	9.6875	Myles, Mr. Thomas Francis	0	894	2	male	0	NaN	240276	Mr	0
3	27.0	S	8.6625	Wirz, Mr. Albert	0	895	3	male	0	NaN	315154	Mr	0
4	22.0	S	12.2875	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	1	896	3	female	1	NaN	3101298	Mrs	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	30.0	S	8.0500	Spector, Mr. Woolf	0	1305	3	male	0	NaN	A.5. 3236	Mr	0
414	39.0	C	108.9000	Oliva y Ocana, Dona. Fermina	0	1306	1	female	0	NaN	PC 17758	Mrs	0
415	38.5	S	7.2500	Saether, Mr. Simon Sivertsen	0	1307	3	male	0	NaN	SOTON/O.Q. 3101262	Mr	0
416	30.0	S	8.0500	Ware, Mr. Frederick	0	1308	3	male	0	NaN	359309	Mr	0
417	4.0	C	22.3583	Peter, Master. Michael J	1	1309	3	male	1	NaN	2668	Master	2

418 rows × 13 columns

In [13]:
```
#klib.pool_duplicate_subsets(df) # pools subset of cols based on duplicates with min. loss of information
klib.pool_duplicate_subsets(df)
```

Out[13]:

	Age	Name	PassengerId	Ticket	pooled_vars
0	34.5	Kelly, Mr. James	892	330911	0
1	47.0	Wilkes, Mrs. James (Ellen Needs)	893	363272	1
2	62.0	Myles, Mr. Thomas Francis	894	240276	2
3	27.0	Wirz, Mr. Albert	895	315154	3
4	22.0	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	896	3101298	4
...	...	...	...	...	...
413	30.0	Spector, Mr. Woolf	1305	A.5. 3236	76
414	39.0	Oliva y Ocana, Dona. Fermina	1306	PC 17758	414
415	38.5	Saether, Mr. Simon Sivertsen	1307	SOTON/O.Q. 3101262	123
416	30.0	Ware, Mr. Frederick	1308	359309	76
417	4.0	Peter, Master. Michael J	1309	2668	417

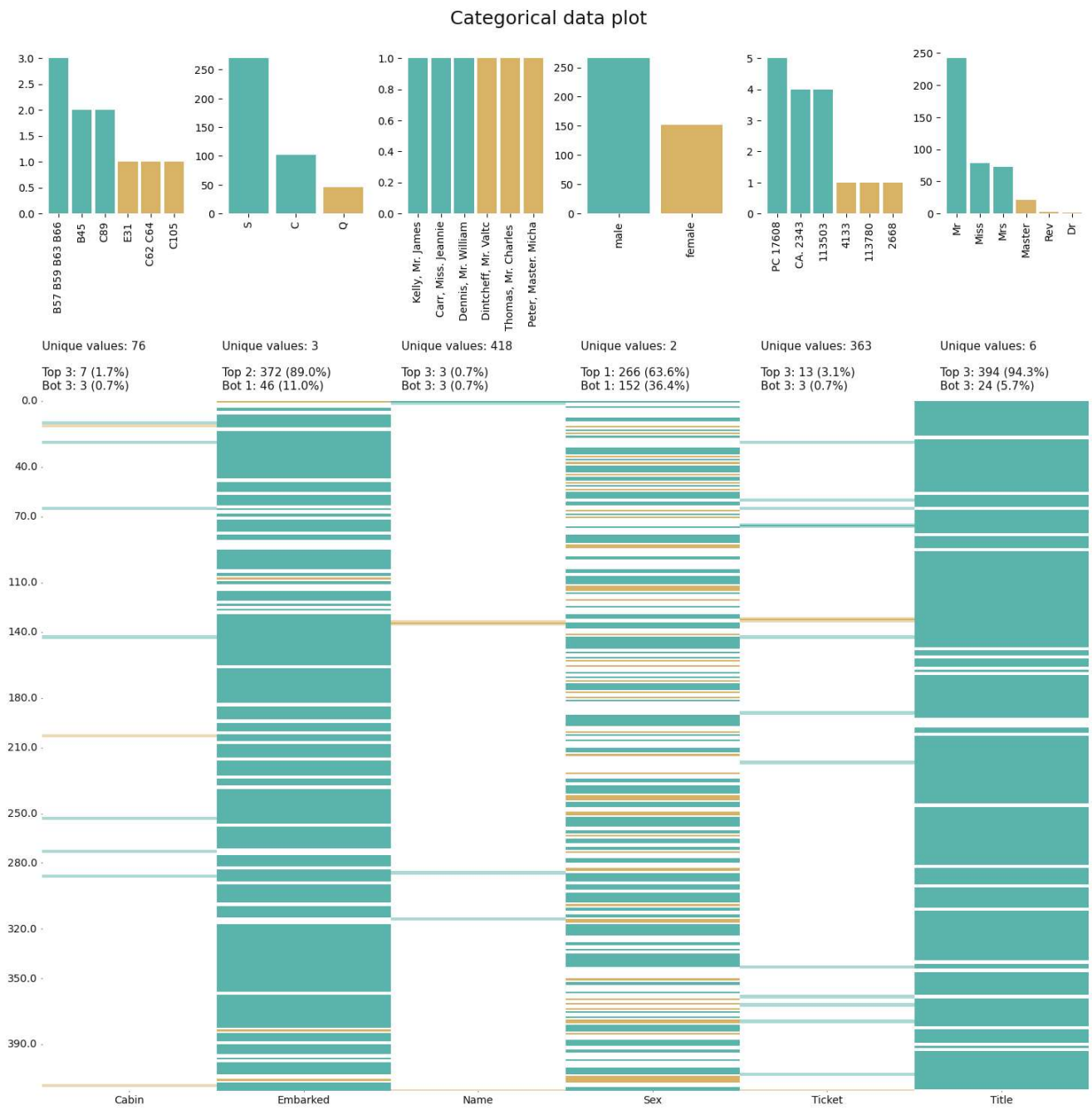
418 rows × 5 columns

In [14]:
```
# 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
```

In [15]:
```
#klib.cat_plot(df) # returns a visualization of the number and frequency of categorical features
klib.cat_plot(df)
```

Out[15]:
```
GridSpec(6, 6)
```



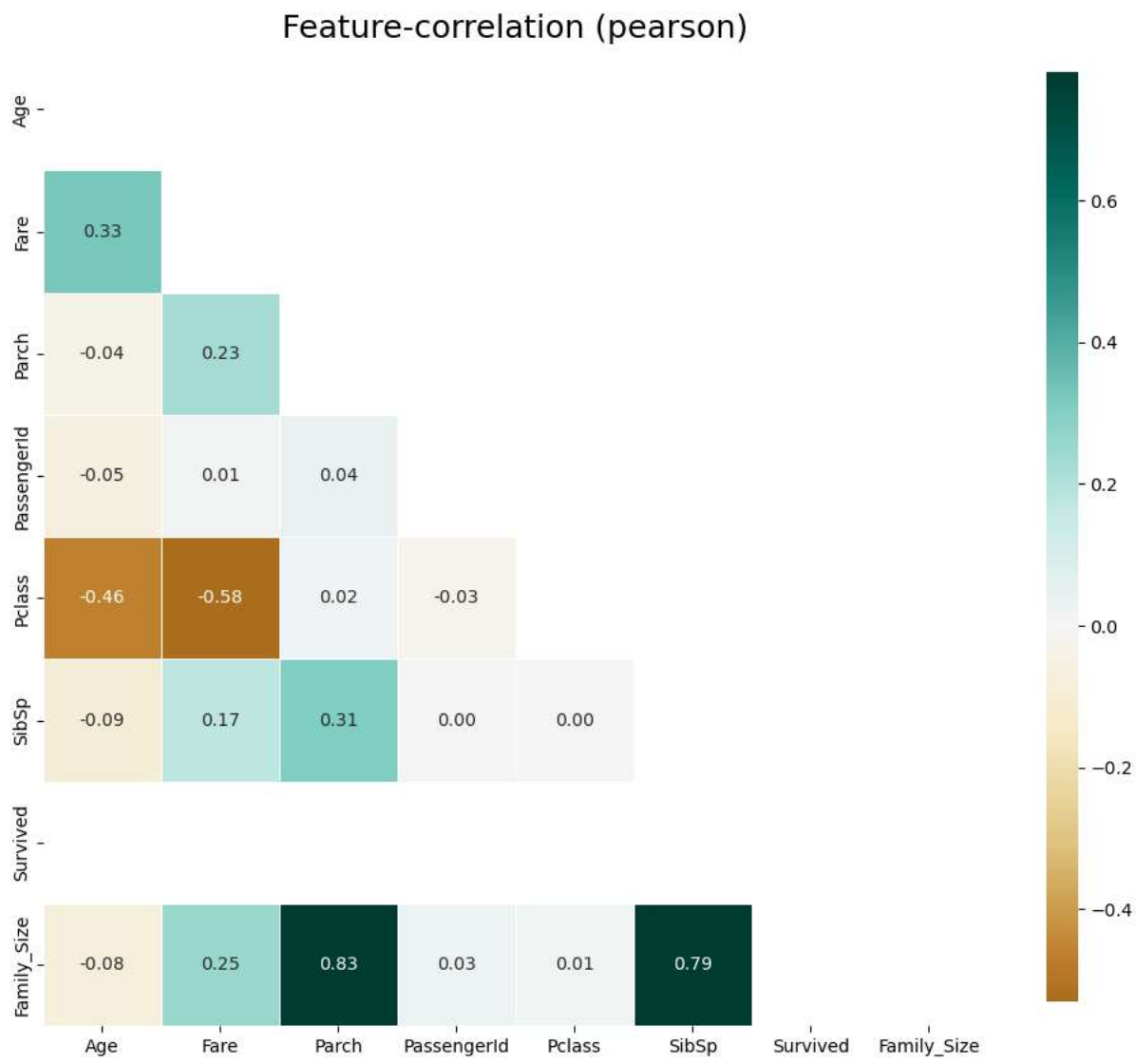
In [16]:
```
#klib.corr_mat(df) # returns a color-encoded correlation matrix
klib.corr_mat(df)
```

Out[16]:

	Age	Fare	Parch	PassengerId	Pclass	SibSp	Survived	Family_Size
Age	1.00	0.33	-0.04	-0.05	-0.46	-0.09	-	-0.08
Fare	0.33	1.00	0.23	0.01	-0.58	0.17	-	0.25
Parch	-0.04	0.23	1.00	0.04	0.02	0.31	-	0.83
PassengerId	-0.05	0.01	0.04	1.00	-0.03	0.00	-	0.03
Pclass	-0.46	-0.58	0.02	-0.03	1.00	0.00	-	0.01
SibSp	-0.09	0.17	0.31	0.00	0.00	1.00	-	0.79
Survived	-	-	-	-	-	-	-	-
Family_Size	-0.08	0.25	0.83	0.03	0.01	0.79	-	1.00

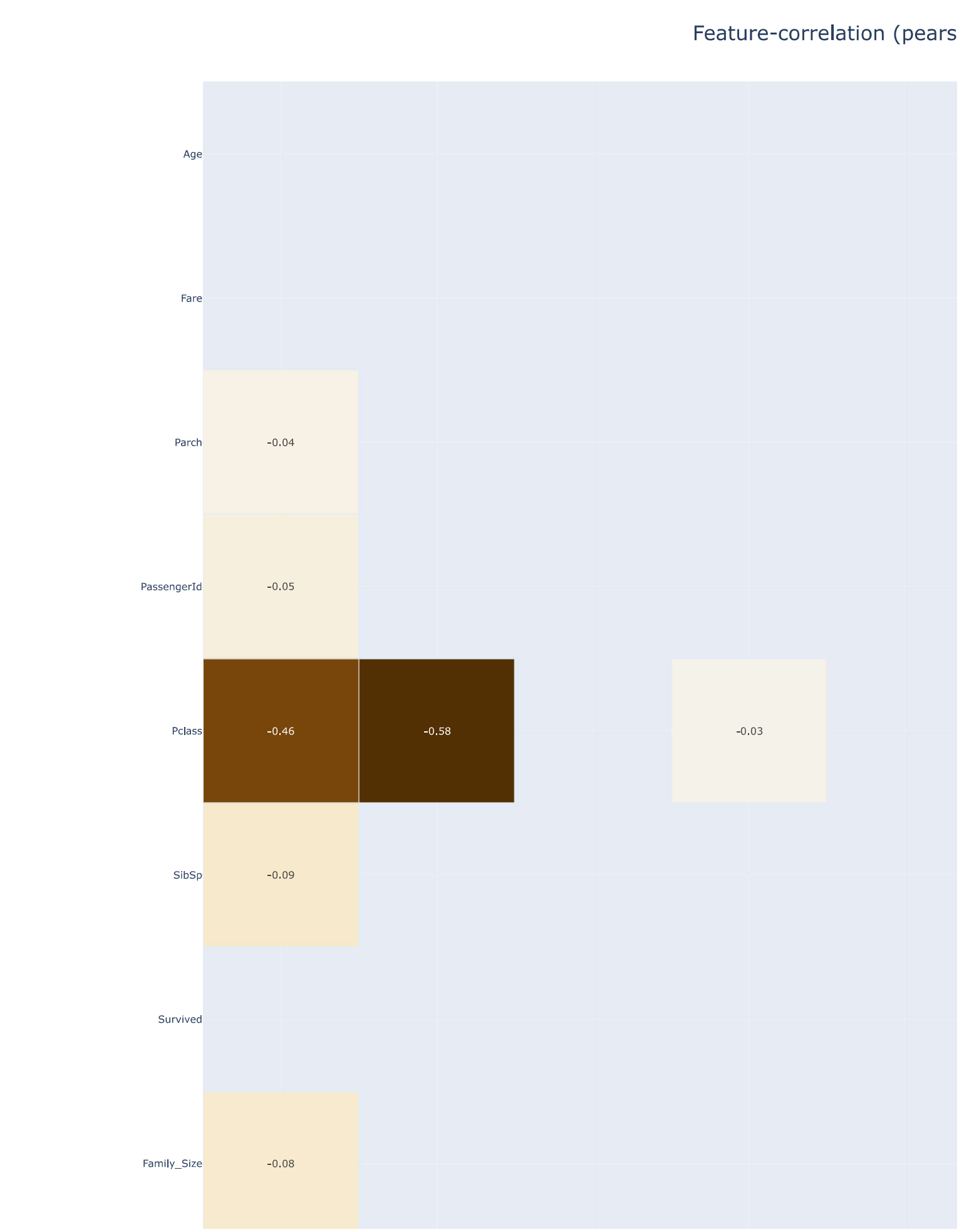
In [17]:
```
#klib.corr_plot(df) # returns a color-encoded heatmap, ideal for correlations
klib.corr_plot(df)
```

Out[17]:
```
<Axes: title={'center': 'Feature-correlation (pearson)'}>
```



In [18]:
```
#klib.corr_interactive_plot(df, split="neg").show() # returns an interactive correlation plot using plotly
klib.corr_interactive_plot(df, split="neg").show()
```

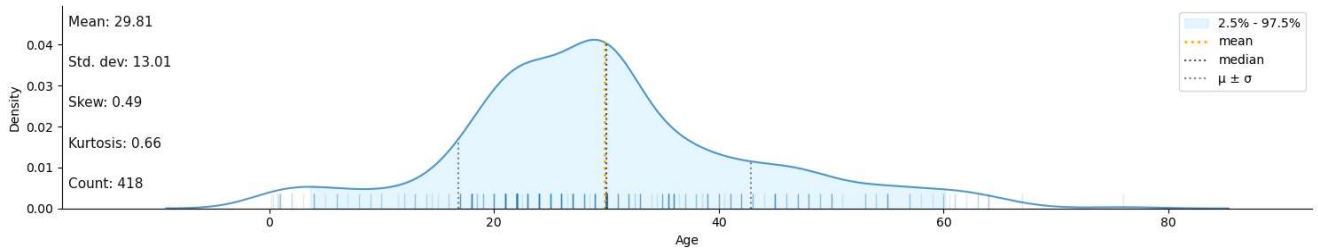
```
Displaying negative correlations. Specify a negative "threshold" to limit the results further.
```


Feature-correlation (pears
Age
Fare
Parch
PassengerId
Pclass
SibSp
Survived
Family_Size
-0.04
-0.05
-0.46
-0.58
-0.03
-0.09
-0.08

Age Fare Parch PassengerId Pclass

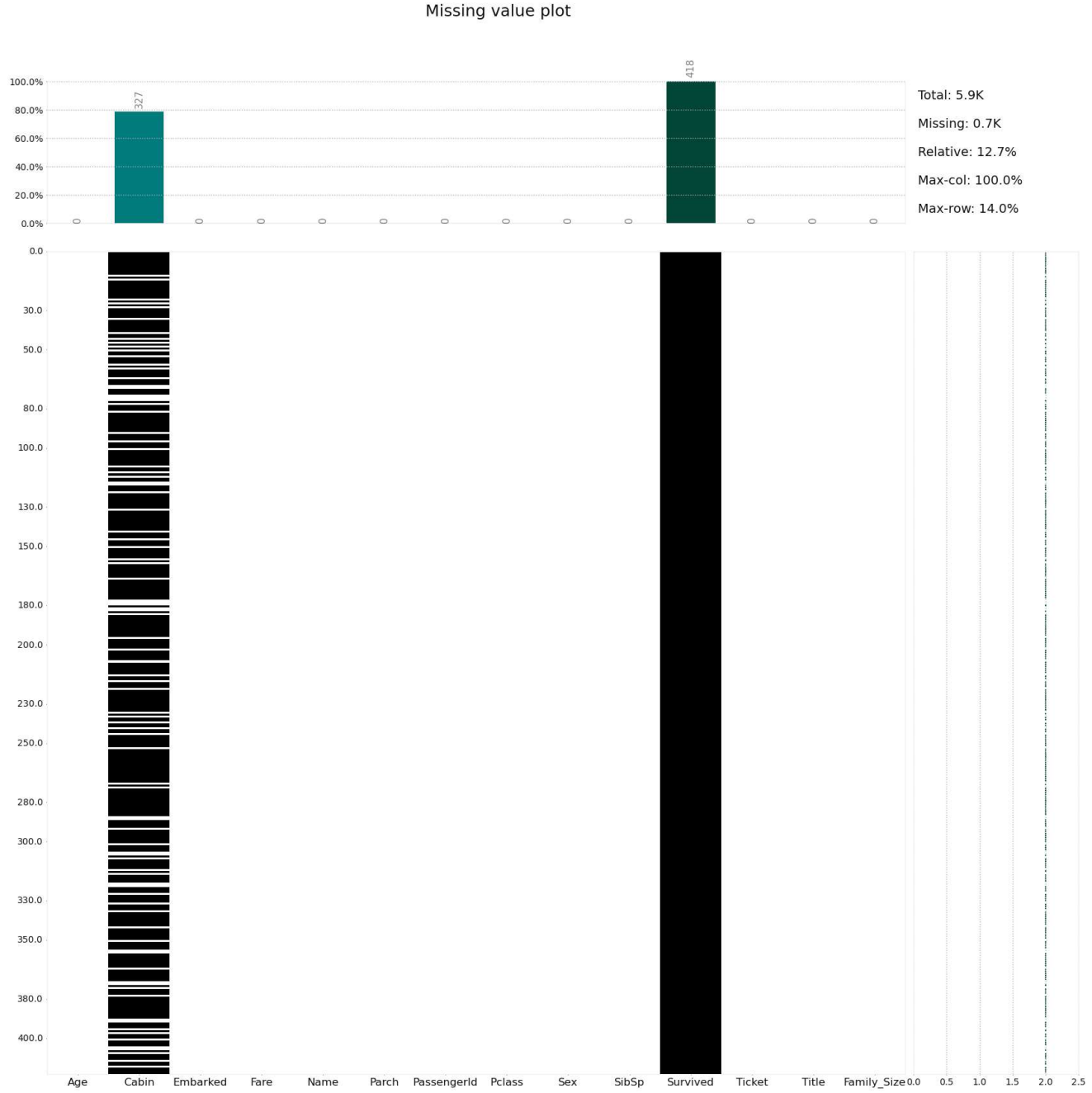
In [19]:
```
#klib.dist_plot(df) # returns a distribution plot for every numeric feature
klib.dist_plot(df)
```

Out[19]:
```
<Axes: xlabel='Age', ylabel='Density'>
```



In [20]:
```
#klib.missingval_plot(df) # returns a figure containing information about missing values
klib.missingval_plot(df)
```

Out[20]:
```
GridSpec(6, 6)
```


Missing value plot
Total: 5.9K
Missing: 0.7K
Relative: 12.7%
Max-col: 100.0%
Max-row: 14.0%
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
327
418
Age
Cabin
Embarked
Fare
Name
Parch
PassengerId
Pclass
Sex
SibSp
Survived
Ticket
Title
Family_Size

In [ ]: