DFutils 主要的功能如下所示。

- count_unique：计算每列唯一值数量，NA 不在统计范围内。
- count_na_col：每列空值的数量。
- count_na_row：每行空值的数量。
- sample_df：DataFrame 随机采样。
- all_na_cols：全是空值的列。
- all_na_rows：全是空值的行。
- is_numeric、is_numeric_bycount：通过数据类型、数量判断是否为数值。
- max_strlen_invalue：最大字符串的长度。

DataMetaInfo 类中实现的功能如下所示。

- report_zero_var：方差接近 0 的列报告。
- report_numeric：数值变量的分析报告，包含百分位、最大与最小、偏度、峰度，以及 0、正数、负数占比等。
- report_cols：各列报告统计，包含数据类型、空值占比、唯一值数量等。
- cal_duplicated：计算重复的行或列。

5.5.2　使用示例

先加载相关数据分析包。

```
from dataeda import *
```

以本章的 titanic_df 数据为例说明该工具包的使用。将原始 DataFrame 输入 DataMetaInfo。

```
dm = DataMetaInfo(titanic_df)
```

查看数据基本统计：

```
dm.psummary()
```

输出结果为：

```
Data shape:
(891, 15)
Data mem size:
0.079M
......
```

运行数据分析：

```
r1,r2,r3 = dm.run()
```

r1、r2、r3 的输出如图 5-23 所示。其中 r1 为空，表示 titanic_df 中没有方差接近 0 的列；r2 显示了数值型变量的数理统计指标；r3 显示了各列的统计信息。这些信息可直接存