1.**题目**：分类数据集的探索式数据分析。

2.**数据集**：从 UCI Machine Learning Repository、Kaggle、Git Hub、Gitee 等数据源，自行选择一个分类数据集。要求特征数量>4，包含数值型特征、类别型特征，有缺失值、有分类标注。

3. **任务描述**：完成如下任务。

（1）装载数据集；
（2）查看部分数据；
（3）描述性统计信息：max/min/4 分位点/中位数/mean/std；
（4）重复值处理；
（5）空值情况、空值处理；
（6）类别型字段信息展示：Row count, null count, distinct value count, distinct value list, histogram；
（7）数值型字段信息展示：Head, row count, null count, distinct value count, density, box plot；
（8）离群值检测、离群值处理（如有必要）；
（9）类别型字段编码；
（10）数值型字段 Binning（如有必要）；
（11）类别型变量和目标变量的关系，数值型变量和目标变量的关系；
（12）单变量和目标变量的关系，双变量和目标变量的关系；
（13）字段相关性、特征选择（相关系数超过阈值）；
（14）数据缩放、数据标准化等；
（15）尝试 KNN、决策树、SVC 等分类模型，结合不进行特征选择/进行特征选择，不进行数据缩放/进行数据缩放，对不同模型效果进行评价；
…其他有必要的探索和预处理操作。

4. **成果提交**: 请于截止时间前，把数据与源代码（*.ipynb）一起压缩为 zip 文件，提交到 obe.ruc.edu.cn 网站。

备注：notebook 的每个代码 Cell，要求加上必要的注释。

5. **参考文献**：无

6.**评价方法**:

（1）本练习为必选练习，根据上述任务完成度评分。
（2）部分完成、晚交适当扣分。
（3）自行完成，不要抄袭，抄袭 0 分。

7.参考数据集：

Titanic, Iris, penguin dataset, Pima Indians Diabetes Database, loan approval, adult income, Breast Cancer……