Python程序设计

案例：用Pandas分析数据

张 华

演员参演电影数量数据分析

需求

数据

- 假设有个Excel文件“电影信息.xlsx”，
- 其中有三列分别为电影名称、导演和演员列表（同一个电影可能会有多个演员，每个演员姓名之间使用逗号分隔）。

要求

- 找到最受欢迎的演员。

电影名称	导演	演员
电影1	导演1	演员1，演员2，演员3
电影2	导演2	演员4，演员5，演员6
电影3	导演3	演员1，演员7，演员8，演员9
电影4	导演1	演员2，演员10，演员5，演员11
电影5	导演2	演员3，演员8，演员12
电影6	导演3	演员4，演员7，演员11，演员9
电影7	导演4	演员5，演员6
电影8	导演5	演员5，演员7，演员8
电影9	导演6	演员6，演员10，演员13
电影10	导演1	演员7，演员9，演员11，演员13
电影11	导演2	演员8，演员9，演员14
电影12	导演3	演员9，演员10，演员11，演员12，演员13
电影13	导演4	演员10，演员1，演员3
电影14	导演5	演员11，演员1，演员2
电影15	导演6	演员12，演员7，演员5，演员11
电影16	导演7	演员13，演员2，演员4，演员5，演员9
电影17	导演8	演员14，演员6，演员8，演员10
电影18	导演9	演员15，演员4
电影19	导演1	演员3，演员6，演员7
电影20	导演2	演员5，演员10
电影21	导演3	演员11，演员4
电影22	导演4	演员10，演员9，演员13
电影23	导演5	演员7，演员6，演员14
电影24	导演6	演员8，演员3，演员15
电影25	导演7	演员9，演员2，演员11

演员参演电影数量数据分析

需求分析

- “找到最受欢迎的演员”
  - 统计每个演员的参演电影数量；
  - 找到参演电影数量最多的3个演员，他/她们是最受欢迎的演员。

技术路线

- 数据分析的一般过程
  - 数据表示：采用合适方式用程序表达数据
  - 数据清理：数据归一化、数据转换、异常值处理等
  - 数据统计：数据的概要理解，数量、分布、中位数等
  - 数据可视化：用图表形式直观展示数据内涵
  - 数据挖掘：从数据分析获得知识，产生数据外的价值
  - 人工智能：数据/语言/图像/视觉等方面深度分析与决策

演员参演电影数量数据分析

分析和结果

- 见下面的文件

  pandas_movies.ipynb