
让数据结构贴合真题的数据关系

第 5 章 | 列表收集、字典映射、嵌套 JSON 与 pandas 接口

真题先问：数据之间是什么关系？

多条同类数据用列表；一条记录的字段和值用字典；嵌套结构逐层判断类型。

- ▶ 名仕：店铺列表、请求字典、嵌套 JSON、分页收集列表
- ▶ 科图：HTML 表格列表、列名列表、聚合规则字典
- ▶ 青娅：目标项目列表、金额列列表、列字典建表

四类组合数据怎样选择？

类型	核心特点	真题或实用场景
list	有序、可变、可重复	列名、目标、记录、分页结果
tuple	有序、不可变	固定分组、解包
set	唯一、无序	去重、成员判断
dict	键映射到值	记录、参数、JSON、规则

空字典可写 `dict()`；空集合必须写 `set()`。

列表索引从 0 开始，切片右边界不包含

```
columns = [" 年度", " 季度", " 网店名称",  
           " 销量", " 退单量", " 售价"]  
columns[0]      # " 年度"  
columns[-1]     # " 售价"  
columns[2:5]    # 三个业务字段
```

列名列表既选择字段，也决定输出顺序。

append 与 extend：整体加入还是逐项加入

```
values = []  
values.append([1, 2])  
# [[1, 2]]
```

```
values = []  
values.extend([1, 2])  
# [1, 2]
```

名仕收集一个 DataFrame 用 **append**；合并 read_html 返回的列表可用 **extend**。

列表在三套真题中的四个角色

1. 列名列表：选择字段并固定顺序；
2. 目标列表：配合 `isin()` 匹配多个项目；
3. 记录列表：JSON 中多条字典记录；
4. DataFrame 列表：收集分页或多个 HTML 表。

先明确列表“装的是什么”，再选择 `append`、`extend` 或 `concat`。

元组与集合：只保留实用边界

```
group_key = (2024, " 华东")  
year, region = group_key
```

```
channels = set([" 直播", " 搜索", " 直播"])  
" 直播" in channels
```

元组用于固定组合；集合用于去重。集合无序，**不能替代输出列名列表**。

字典：通过键读取业务值

```
product = {  
    " 商品名称": " 双肩包",  
    " 销量": 125,  
    " 售价": 199.0,  
}  
product[" 销量"]  
product.get(" 退单量", 0)
```

[] 适合必需字段；get 的默认值必须符合业务口径。

字典可变： 名仕请求参数会被原地更新

```
goods_req["clvalue"] = shop_name  
goods_req["pageNum"] = 1  
goods_req["pageNum"] += 1
```

- ▶ 切换店铺：更新 clvalue
- ▶ 新店铺开始：页码重置为 1
- ▶ 处理下一页：页码加 1

嵌套 JSON：每一层先判断类型



```
data_block = response["data"]  
records = data_block["list"]  
first_name = records[0]["商品名称"]
```

一条字典，多条记录组成列表

```
records = [  
    {"商品": "双肩包", "销量": 125},  
    {"商品": "旅行箱", "销量": 80},  
]  
df = pd.DataFrame(records)
```

字典的键对齐为列；每个字典成为一行。缺少的键会形成缺失值。

items：把规则字典拆成列名和操作

```
rules = {  
    " 实际销量": "sum",  
    " 实际销售额": "sum",  
    " 售价": "mean",  
}  
for column, function_name in rules.items():  
    print(column, function_name)
```

同一字典可直接传给 `groupby().agg(rules)`。

pandas 接收三种组合数据

输入结构	pandas 操作
记录字典列表	<code>pd.DataFrame(records)</code>
列名到列数据的字典	<code>pd.DataFrame(column_dict)</code>
DataFrame 列表	<code>pd.concat(df_list)</code>

concat 前检查列表非空；列字典中的列长度必须一致。

科图与青娅：列表表达“要哪些”

- ▶ 科图：[" 年", " 区域分布", " 付款方式"]
- ▶ 科图：groupby([" 年", " 区域分布"])
- ▶ 青娅：年度金额列列表
- ▶ 青娅：目标项目列表配合 isin()

列表不只是语法容器，它把字段范围、顺序和筛选口径写进代码。

名仕：组合数据类型串成分页流水线

1. 遍历店铺记录列表；
2. 从店铺字典读取网店名称；
3. 更新请求参数字典并重置页码；
4. 从嵌套响应取每页记录列表；
5. 转成 DataFrame 后放入收集列表；
6. 循环结束后统一 concat。

名仕：收集后统一合并

```
dataframe_list = []
for shop in shop_data:
    goods_req["clvalue"] = shop[" 网店名称"]
    goods_req["pageNum"] = 1
    while has_next_page:
        records = response["data"]["list"]
        dataframe_list.append(pd.DataFrame(records))
        goods_req["pageNum"] += 1

df = pd.concat(dataframe_list, ignore_index=True)
```


教材内容怎样取舍？

必讲

- ▶ 列表索引、切片、遍历
- ▶ append 与 extend
- ▶ 列名、目标和记录列表
- ▶ 字典访问、get、修改
- ▶ 嵌套 JSON
- ▶ pandas 三种接口

压缩/按需查阅

- ▶ 复杂推导式
- ▶ 列表删除全套方法
- ▶ 集合高级方法
- ▶ 字典推导式
- ▶ 成语、通讯录案例

离场检验：只问影响真题得分的内容

- ▶ append 与 extend 的结果形状有什么区别？
- ▶ 为什么列名清单不能使用集合？
- ▶ 直接键访问与 get 默认值怎样选择？
- ▶ 嵌套 JSON 的完整访问路径怎样写？
- ▶ 名仕分页为什么先收集、最后 concat？
- ▶ 两种 DataFrame 构造字典有什么不同？