Pandas 备考复习笔记

はなとみず

2025 年 8 月 8 日

提示： 如果你想要完整地学习 Pandas，请**不要**使用这份笔记。这份笔记给浙江省信息技术学考、选考复习用，不适合作为 Pandas 的入门笔记。这里省略了很多考试不会涉及的特性。

目录

1 Series

1.1 简述 Series

定义 1 (Series)**.** *Pandas* 中的一维数据结构，是构建 *DataFrame* 的基础组件。

~~Series 考试考到比较少，了解为主.~~

你可以这样创建一个 Series

```python
import pandas as pd

# 默认索引为0, 1, 2, 3
series = pd.Series([1, 2, 3, 4])
print(series)

# 输出结果为:
# 0    1
# 1    2
# 2    3
# 3    4
# dtype: int64
```

你可以自定义她的下标（index）

```python
import pandas as pd
series = pd.Series([1, 2, 3, 4], index=[1, 2, 3, 4])
print(series)
# 输出结果:
# 1    1
# 2    2
# 3    3
# 4    4
# dtype: int64
```

此外，她的下标（index）和值（value）还可以为浮点型、字符串等。

1.2 常用方法

这里给出几个 Series 的方法，可以自己创建一个 Series 进行尝试

- `index`：获得 Series 的索引
- `values`：获取 Series 的数据部分
- `head(n)`：返回 Series 的前 n 行（默认为 5）
- `tail(n)`：返回 Series 的后 n 行（默认为 5）
- `describe()`：返回 Series 的统计描述（如均值、标准差、最小值等）~~不常用~~
- `sort_values()`：对 Series 中的元素进行排序（给出 `ascending` 参数，`True` 为升序，`False` 为降序）

2 DataFrame

定义 2 (DataFrame)**.** 由多个 *Series* 组成的二维表格结构，是 *Pandas* 的核心对象。

2.1 创建 DataFrame

DataFrame 是一种二维的列表，创建过程如下：

```
import pandas as pd

data = [['Alice', 11], ['Bob', 12], ['Charlie', 13], ["Dick", 14]]

# 创建 DataFrame 通常使用 df 作为变量名
df = pd.DataFrame(data, columns=['Name', 'Age'])

print(df)
# 输出结果：
#       Name  Age
# 0    Alice   11
# 1      Bob   12
# 2  Charlie   13
# 3     Dick   14
```

2.2 数据读取与保存

我们通常使用 `read_csv()` 或者 `read_excel()` 来读入一个 DataFrame

```
import pandas as pd
df1 = pd.read_csv("awa.csv")
df2 = pd.read_excel("uwu.xlsx")

df1.to_csv("new_awa.csv")
df2.to_excel("new_uwu.xlsx") # 这样写不意味着从 Excel 读入的必须写回 Excel 里，这里也可以是 df2.to_csv("filename.csv")
```

2.3 数据选择与筛选

- `df.at[x, y]`：选择第 x 行，第 y 列的元素
- `df.head(n)`：显示前 n 行数据（与 Series 同理）
- `df.tail(n)`：显示后 n 行数据（与 Series 同理）
- `df.drop(x, axis=0/1)`：删除 x 行/列的数据，`axis` 表示行列，0 是行，1 是列
- `df.groupby(x, as_index=False)`：根据 x 列来分组，**后面一定要跟上 `as_index=False`**，否则你的索引就炸了，可以自己试试。分组后需要使用聚合函数（`sum()`, `mean()`, `count()` 等）来处理数据。完整的：`df.groupby(x, as_index=False).sum()` 等。
- `df[df.x > v]`：筛选出 df 中所有 x 列元素大于 v 的行，等效写法有 `df[df["x"] > v]`。你只需要记住在 df 里套一个 df 即可。

- `df.mean()`：计算每列的平均值（**不是 `df.average()`，没有这种说法**）
- `df.median()`：计算每列中位数
- `df.mode()`：计算每列众数
- `df.count()`：数这一列有多少个元素（有些行可能是空的）
- `df.sum()`：计算每列的和
- `df["example"]`：取出标题为 example 的一列，如果这一列不存在，则创建这一列（与 `df.example` 等效，但是如果列不存在，`df.example` 会报错，需要使用 `df["example"]`，标题可以是中文）

  代码示例：

```
import pandas as pd
a = [["Name1", 12, 23], ["Name2", 23, 34], ["Name3", 35, 83]]
df = pd.DataFrame(a, columns=["Names", "Column1", "Column2"])
print(df)
df["Sum"] = df.Column1 + df.Column2
print(df)
```

3 数据可视化

严格来说，数据可视化与 Pandas 无关，是 Matplotlib 的功能

导入 Matplotlib：

```
import matplotlib.pyplot as plt
```

注意的是，包名是 `matplotlib.pyplot`，不要丢了后面的 `.pyplot`，丢了你程序就炸了。当然，考试不会让你填这个的，自己写的时候留意一下即可。

DataFrame 可以直接绘图，考试考的较少，了解即可：

```
df.plot(kind='bar', x='Name', y='Age') # 直接生成条形图
```

一般境况下，我们使用 plt 来绘制图，这里给出一个样例

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

x = [1, 2, 3, 4]
y = [2, 3, 4, 5]
plt.title("a well-designed table")
plt.xlabel("wonderful xlabel")
plt.ylabel("wonderful ylabel")
plt.plot(x, y)
plt.show()
```

然后就可以看到这个窗口了：

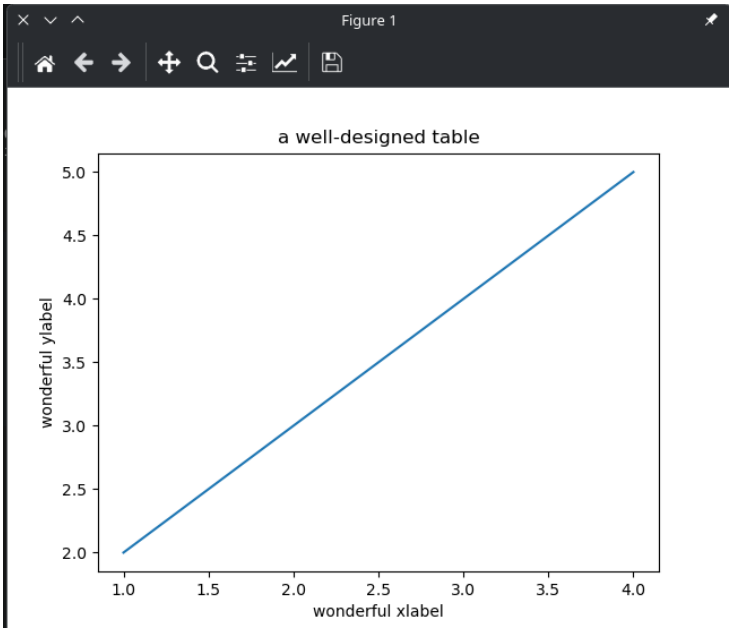

图 1: Matplotlib

再列举几个类型图的画法，仿照样例把 `plt.plot(x, y)` 修改一下即可：

- `plt.plot(x, y)`：用于绘制线图和散点图
- `plt.scatter(x, y)`：用于绘制散点图
- `plt.bar(x, y)`：用于绘制垂直条形图
- `plt.barh(x, y)`：用于绘制水平条形图
- `plt.hist(data)`：用于绘制直方图

其他的杂项：

- `plt.xlabel("x")`：设置横坐标标题为 x
- `plt.ylabel("y")`：设置纵坐标标题为 y