day08笔记

一、加载文件

1、加载 csv 文件

```
# 加载文件 csv
df1 = pd.read_csv(
    './data/2023年北京积分落户数据.csv', # 加载文件的路径
    encoding='utf-8', # 文件编码格式
    index_col='公示编号', # 指定充当行索引的列
    usecols=['公示编号', '姓名', '单位名称'], # 指定需要加载的列（指定行索引是不能缺少的）
    # nrows=3, # 指定加载的行数
    skiprows=np.arange(1, 21) # 跳开文件开头指定的行数
)
df1
```

2、加载 Excel 文件

注意点：加载 Excel 文件，需要先下载一个工具包

安装

```
pip install openpyxl
```

加载文件

```
# 加载 Excel 文件
df2 = pd.read_excel(
    './data/2022年股票数据.xlsx',
    sheet_name='BABA', # 指定 Excel 文件加载哪张表，如果不指定加载默认的第一个
    index_col='Date'
)
df2
```

二、DataFrame 查看方法

```
# 查看所有信息
df1.info()

# 查看前 n 行信息
df1.head(10)

# 查看后 n 行信息
df1.tail(5)
```

三、拼接方法

`concat()` 拼接，可以把多个文件拼接成一个文件

```
# DataFrame 方法
# concat() 拼接，可以把多个文件拼接成一个文件
beijing = pd.read_csv('./data/jobs/beijing_data.csv')
chengdu = pd.read_csv('./data/jobs/chengdu_data.csv')
guangzhou = pd.read_csv('./data/jobs/guangzhou_data.csv')

pd.concat([beijing, chengdu, guangzhou], ignore_index=True) # ignore_index=True 表示把之前的行索引全部废弃掉，重新生成一个行索引
```

四、操作单元格

```
df3 = pd.read_excel('./data/2020年中国大学排名.xlsx')
df3

# 操作单元格
df3.at[1, '人才培养得分']

# 参数1：表示从哪一行开始，默认行索引是从 0 开始的
# 参数2：表示是那一列，默认也是从 0 开始的
df3.iat[1, 9] # 默认从 0 开始的
```

五、切片、花式索引

切片操作

```
# 切片操作
df3.loc[1:3]

# 花式索引
# 表示拿到第一个数据，再哪最后一个数据，在哪第10个数据
df3.iloc[[0, -1, 10]]
```

六、删除行和列

1、删除行

```
# 删除行
df3.drop(index=[0, 1], inplace=True)
```

2、删除列

```
# 删除列
df3.drop(columns=['社会服务得分', '人才培养得分'], inplace=True)
df3
```

七、筛选

```
# 重新读取表格，由于之前做的删除行和列的操作，没有进行操作写入到表格，因此重新读取的数据是完整的
df4 = pd.read_excel('./data/2020年中国大学排名.xlsx')
df4

# 筛选
df4[(df4.省市=='四川') | (df4.省市=='重庆')]

df4[df4.总分 > 500]

# query() 筛选方法
df4.query('总分 > 500')

df4.query('省市=="重庆" or 省市=="四川"')
```