Pandas数据处理

讲师：萨缪尔 Samuel

网易云课堂 | 涨薪计划

萨缪尔老师

网易数据分析教研负责人

前盛大游戏战略规划总监、前腾讯游戏商业智能中心Leader

上海交通大学高金硕士 、《哈佛管理导师》外部导师

- **知乎大V：「萨缪尔」**主要聚集与商业分析、行业与战略分析、Python数据分析等
- **BAT**互联网巨头商业洞察分析 & 咨询公司战略咨询背景
- 擅长行业趋势研究和战略管理咨询工具，**为20多家上市公司提供战略发展决策建议**

课程亮点

1. 理解数据-描述性分析
2. 数据清洗：缺失值、重复值
3. 数据类型的转换
4. 修改列名

理解数据

---描述性分析

1

理解数据 & 描述性分析

```
df.shape # (100, 6) 查看行数和列数

df.info() # 查看索引、数据类型和内存信息

df.describe() # 查看数值型列的汇总统计

df.dtypes # 查看各字段类型

df.axes # 显示数据行和列名

df.columns # 列名
```

数据清洗

——缺失值、重复值

2

缺失值处理

类别	方法名	功能描述
缺失值检测	isnull()	布尔判断，如果存在缺失值，则返回 True
	notnull()	布尔判断，如果没有缺失值，则返回 True
缺失值填充	fillna(0)	如果存在缺失值，则用指定的值进行填充，默认值为 0
缺失值丢弃	dropna()	如果存在缺失值，则无条件将其抛弃
	dropna(how='all')	当前单元格所在的行或列都为缺失值（NaN）时，则抛弃数据
	dropna(axis = 1, how='all')	当列方向（axis = 1）的所有数据都为缺失值时，则抛弃该列
	dropna(axis=1, how='any')	当列方向有任何一个缺失值时，则抛弃该列
	dropna(thresh=5)	当所在行的数据有效值低于 5 个时，抛弃该行，这里的 thresh 是可修改的阈值

缺失值处理

删除空值所在行

dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)

thresh：非空元素最低数量。int型，默认为None。如果该行/列中，非空元素数量小于这个值，就删除该行/列。

how：筛选方式。 ‘any’ ，表示该行/列只要有一个以上的空值，就删除该行/列；‘all’ ，表示该行/列全部都为空值，就删除该行/列。

subset：子集。列表，元素为行或者列的索引。

如果axis=0或者 ‘index’ ，subset中元素为列的索引；如果axis=1或者 ‘column’ ，subset中元素为行的索引。由subset限制的子区域，是判断是否删除该行/列的条件判断区域。

inplace：是否原地替换。布尔值，默认为False。如果为True，则在原DataFrame上进行操作，返回值为None。

重复值处理

删除重复值

```
df.drop_duplicates(subset=['A','B'],keep='first',inplace=True)
```

subset： 输入要进行去重的列名，默认为None

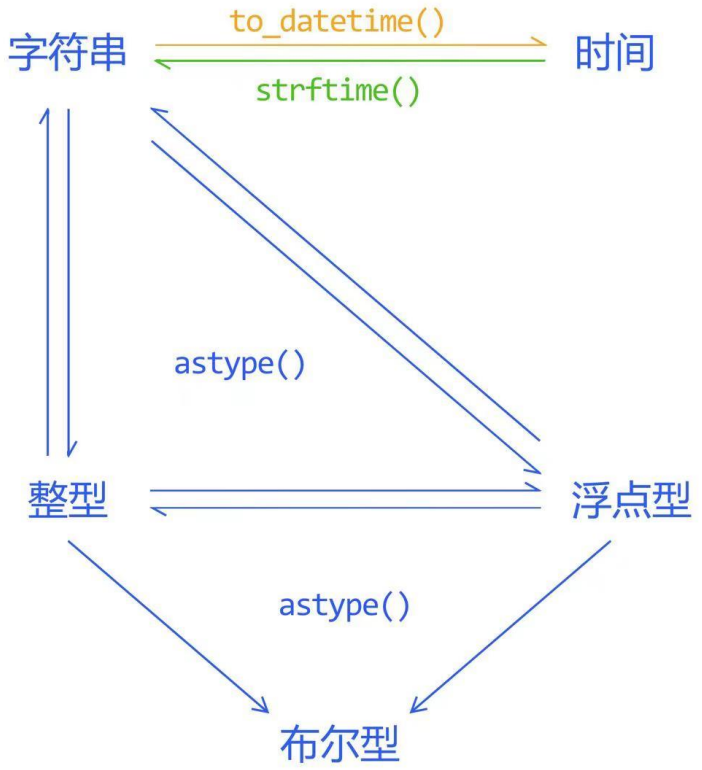
keep： 可选参数有三个：‘**first**’、‘**last**’、**False**，默认值‘first（**first表示：保留第一次出现的重复行，删除后面的重复行；last表示：删除重复项，保留最后一次出现；False表示：删除所有重复项)**

inplace： 布尔值，默认为False，是否直接在原数据上删除重复项或删除重复项后返回副本。

3

数据类型转换

调整数据类型



1. 数据类型之间的转换用astype()
2. 时间转成字符串使用strftime（）
3. 字符串转化为时间可以考虑
   to_datetime（）或strptime（）

4 修改列名

修改列名称

```
#因为原数据字段名是英文，为了便于理解，将字段名修改为中文名。
colNameDict = {
    'season':'季节',
    'holiday':'节假日',
    'workingday':'工作日',
    'weather':'天气',
    'temp':'摄氏温度',
    'atemp':'体感温度',
    'humidity':'湿度',
    'windspeed':'风速',
    'casual':'非注册用户个数',
    'registered':'注册用户个数',
    'count':'租车总人数'
}
df.rename(columns = colNameDict,inplace=True)
df.head()
```

其他根据数据类型的调整

- **逻辑问题筛选**：如商品的价格大于0，数量大于0等
- **格式一致化**：如大小写 / 去除空格，是否有括号等
- **其他**：需要在工作中积累发现的其他问题

案例：爬虫GDP数据并做数据清洗

课程总结

1. Pandas的描述性数据分析，shape，info()，describe()
2. Pandas的数据清洗：缺失值、重复值
3. Pandas数据类型转换
4. Pandas修改列名的方法

谢谢观看