北京工商大学

本科实验报告

课程名称：机器学习

实验名称：空气质量监测数据的预处理和基本分析

专业班级：智能 212　　学号：2006010529

学生姓名：史胤隆

指导教师：高超

2023 年 10 月 8 日

实验目的

掌握 Python 中 Numpy 和 Pandas 进行数据分析的方法。

本案例基于北京市空气质量监测数据，聚焦数据建模中的数据预处理和基本分析环节，说明 Numpy 和 Pandas 的数据读取、数据分组、数据重编码、分类汇总等数据加工处理功能。

实验内容

1. **空气质量监测数据的预处理：** 根据空气质量监测数据的日期，生成对应的季度标志变量；对空气质量指数 AQI 分组，获得对应的空气质量等级。
2. **空气质量监测数据的基本分析：** 计算各季度 AQI 和 PM2.5 的平均值等描述性统计量；找到空气质量较差的若干天数据，以及各季度中空气质量较差的若干天数据；计算季度和空气质量等级的交叉列联表；派生空气质量等级的虚拟变量。

实验程序和结果

空气质量监测数据的预处理程序代码

```
import numpy as np
import pandas as pd
from pathlib import Path

PATH = Path.cwd()  # 数据路径

# 读入数据
df = pd.read_excel(PATH / "北京市空气质量数据.xlsx", header=0)
df = df.replace(0, np.NaN)  # 清理无效数据

# 标注季度
quarter_name = ["一季度", "二季度", "三季度", "四季度"]
df["季度"] = df["日期"].dt.quarter.apply(lambda x: quarter_name[x - 1])

# 根据给出的指标重新划分质量等级
bins = [0, 50, 100, 150, 200, 300, 1000]
labels = ["一级优", "二级良", "三级轻度污染",
          "四级中度污染", "五级重度污染", "六级严重污染"]
df["质量等级"] = pd.cut(df["AQI"], bins=bins, labels=labels)

# 输出数据
print("对AQI的分组结果: ")
df[["日期", "AQI", "质量等级", "季度"]]
```

对AQI的分组结果:

	日期	AQI	质量等级	季度
0	2014-01-01	81.0	二级良	一季度

	日期	AQI	质量等级	季度
1	2014-01-02	145.0	三级轻度污染	一季度
2	2014-01-03	74.0	二级良	一季度
3	2014-01-04	149.0	三级轻度污染	一季度
4	2014-01-05	119.0	三级轻度污染	一季度
...	...	...	...	...
2150	2019-11-22	183.0	四级中度污染	四季度
2151	2019-11-23	175.0	四级中度污染	四季度
2152	2019-11-24	30.0	一级优	四季度
2153	2019-11-25	40.0	一级优	四季度
2154	2019-11-26	73.0	二级良	四季度

2155 rows × 4 columns

空气质量监测数据的基本分析程序代码

各季度 AQI 和 PM2.5 的均值

```
df[["季度", "AQI", "PM2.5"]].groupby("季度").mean()
```

	AQI	PM2.5
季度		
一季度	109.327778	77.225926
三季度	98.911071	49.528131
二季度	109.369004	55.149723
四季度	109.612403	77.195736

各季度 AQI 和 PM2.5 的描述统计量

```
df[["季度", "AQI"]].groupby("季度").describe().drop("count", axis=1, level=1)
```

	AQI						
	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
季度							
一季度	109.327778	80.405408	26.0	48.0	80.0	145.00	470.0

	AQI						
	mean	std	min	25%	50%	75%	max
季度							
三季度	98.911071	45.484516	28.0	60.0	95.0	130.50	252.0
二季度	109.369004	49.608042	35.0	71.0	99.0	140.75	500.0
四季度	109.612403	84.192134	21.0	55.0	78.0	137.25	485.0

```
df[["季度", "PM2.5"]].groupby("季度").describe().drop("count", axis=1, level=1)
```

	PM2.5						
	mean	std	min	25%	50%	75%	max
季度							
一季度	77.225926	73.133857	4.0	24.0	53.0	109.25	454.0
三季度	49.528131	35.394897	3.0	23.0	41.0	67.00	202.0
二季度	55.149723	35.918345	5.0	27.0	47.0	73.00	229.0
四季度	77.195736	76.651794	4.0	25.0	51.0	101.50	477.0

定义 `top()` 函数用于统计

```
def top(df: pd.DataFrame, n: int = 5, column: str = "AQI") -> pd.DataFrame:
    return df.sort_values(by=column, ascending=False).head(n)
```

空气质量最差的 5 天

```
top(df[["日期", "AQI", "PM2.5", "质量等级"]], 5)
```

	日期	AQI	PM2.5	质量等级
1218	2017-05-04	500.0	NaN	六级严重污染
723	2015-12-25	485.0	477.0	六级严重污染
699	2015-12-01	476.0	464.0	六级严重污染
1095	2017-01-01	470.0	454.0	六级严重污染
698	2015-11-30	450.0	343.0	六级严重污染

各季度空气质量最差的 3 天

```
df.groupby("季度").apply(top, n=3)[["日期", "AQI", "PM2.5", "质量等级"]]
```

		日期	AQI	PM2.5	质量等级
季度					
一季度	1095	2017-01-01	470.0	454.0	六级严重污染
	45	2014-02-15	428.0	393.0	六级严重污染
	55	2014-02-25	403.0	354.0	六级严重污染
三季度	186	2014-07-06	252.0	202.0	五级重度污染
	211	2014-07-31	245.0	195.0	五级重度污染
	183	2014-07-03	240.0	190.0	五级重度污染
二季度	1218	2017-05-04	500.0	NaN	六级严重污染
	1219	2017-05-05	342.0	181.0	六级严重污染
	103	2014-04-14	279.0	229.0	五级重度污染
四季度	723	2015-12-25	485.0	477.0	六级严重污染
	699	2015-12-01	476.0	464.0	六级严重污染
	698	2015-11-30	450.0	343.0	六级严重污染

各季度空气质量天数统计

```
pd.crosstab(df["季度"], df["质量等级"], margins=True, margins_name="总计")
```

质量等级	一级优	二级良	三级轻度污染	四级中度污染	五级重度污染	六级严重污染	总计
季度							
一季度	145	170	99	57	48	21	540
三季度	96	209	164	72	10	0	551
二季度	38	240	152	96	14	2	542
四季度	108	230	64	33	58	23	516
总计	387	849	479	258	130	46	2149