2.7.2 可视化探索分析实例

1．背景

数据集是由爬虫从 http://www.tianqihoubao.com/lishi/beijing.html （天气后报网）上获取的包含了北京 2013.12.02-2020.06.05 间每天的天气情况和空气质量。以此数据为基础利用 numpy,pandas,matplotlib 等工具进行可视化探索分析操作。

数据集共 2376 条记录，包含空气质量和天气情况两部分 12 个属性，如下表 2-所示：

表 2-数据属性信息

属性	含义
日期	记录相关日期
AQI	AQI(Air Quality Index),空气质量指数，描述了空气清洁或者污染的程度，以及对健康的影响
质量等级	根据 AQI 将空气质量等级划分为六个等级
PM2.5	直径小于或等于 2.5μm 的尘埃或飘尘在环境空气中的浓度 数值单位： μg/m^3
PM10	直径小于或等于 10.0μm 的尘埃或飘尘在环境空气中的浓度，数值单位：μg/m^3
SO2	二氧化硫，大气的主要污染物之一，数值单位： μg/m^3
CO	一氧化碳，大气的主要污染物之一，数值单位：mg/m^3
NO2	二氧化氮，大气的主要污染物之一，数值单位： μg/m^3
O3_8h	臭氧的 8 小时滑动平均值，数值单位： μg/m^3
天气状况	根据天气情况分为五种
气温	指在野外空气流通、不受太阳直射下测得的空气温度（一般在百叶箱内测定）
风力风向	风吹来的大小和方向

数据示例如下图 2-所示：

```
data.head()
```

	日期	AQI	质量等级	PM2.5	PM10	SO2	CO	NO2	O3_8h	天气状况	气温	风力风向
0	2013-12-02	142	轻度污染	109	138	61	2.6	88	11	多云/多云	11℃/-1℃	无持续风向≤3级/无持续风向≤3级
1	2013-12-03	86	良	64	86	38	1.6	54	45	晴/晴	14℃/-1℃	无持续风向≤3级/无持续风向≤3级
2	2013-12-04	109	轻度污染	82	101	42	2.0	62	23	多云/多云	12℃/0℃	无持续风向≤3级/无持续风向≤3级
3	2013-12-05	56	良	39	56	30	1.2	38	52	晴/晴	12℃/-3℃	无持续风向≤3级/无持续风向≤3级
4	2013-12-06	169	中度污染	128	162	48	2.5	78	15	晴/霾	11℃/-2℃	无持续风向≤3级/无持续风向≤3级

图 2-：天气数据示例

2．可视化探索分析

（1）绘制 AQI, PM2.5, PM10 箱线图

具体代码实现如下：

```
# 绘制 AQI 箱线图
sns.boxplot(y=data["AQI"])
plt.show()
# 绘制 PM2.5 箱线图
sns.boxplot(y=data["PM2.5"])
plt.show()
# 绘制 PM10 箱线图
sns.boxplot(y=data["PM10"])
plt.show()
```

可视化效果如图 2-21 所示。

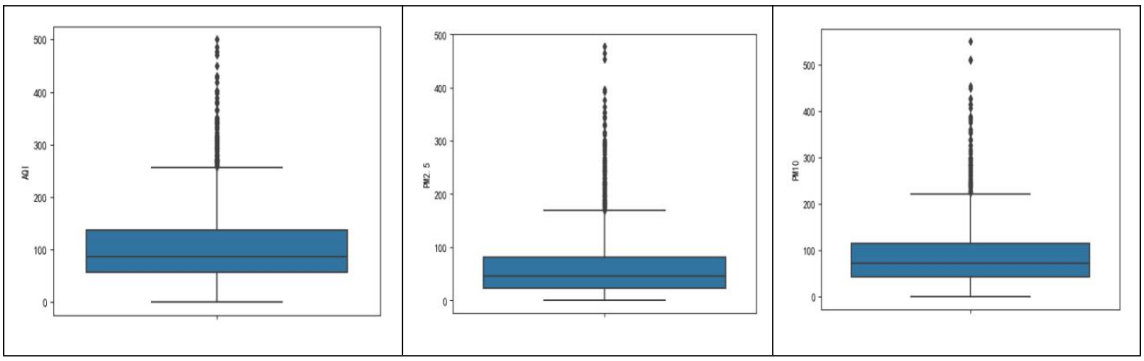

图 2-21 AQI, PM2.5, PM10 箱线图

根据此可视化效果，可以看出 AQI, PM2.5, PM10 属性中超过上限的异常值较多。

（2）绘制 AQI 折线图

```
    fig, axes = plt.subplots(3,3,figsize=(15, 10))
axes_list = []
for i in range(axes.shape[0]):
    for j in range(axes.shape[1]):
        axes_list.append(axes[i, j])
for ax in axes_list:
    ax.set_ylim([0,500])
    ax.set_xticks([])
# bins = np.linspace(-1, 1, 21) #横坐标起始和结束值，分割成 21 份
# plt.xticks(bins) #设置 x 轴
# plt.xlim(-1, 2376) #x 轴开始和结束位置
axes[1,1].set_title("2017AQI",size=10)
axes[1,2].set_title("2018AQI",size=10)

data2017 = data[data['日期'].str.startswith('2017')]
data2018 = data[data['日期'].str.startswith('2018')]
```

```
axes[1,1].plot(data2017['日期'],data2017['AQI'],label ='AQI')
axes[1,2].plot(data2018['日期'],data2018['AQI'],label ='AQI')

plt.figure(figsize=(15, 3))
y = savgol_filter(data['AQI'], 17, 5, mode= 'nearest')
plt.title('2013-2020AQI')
bins = np.linspace(-1, 1, 21) #横坐标起始和结束值，分割成 21 份
plt.xticks(bins) #设置 x 轴
plt.xlim(-1, 2376) #x 轴开始和结束位置
plt.xticks([])
plt.plot(data['日期'],data['AQI'],label ='AQI')
plt.plot(data['日期'],y,'r',label ='AQI')
```

2017 年和 2018 年 AQI 指数可视化效果如图 2-22 所示。

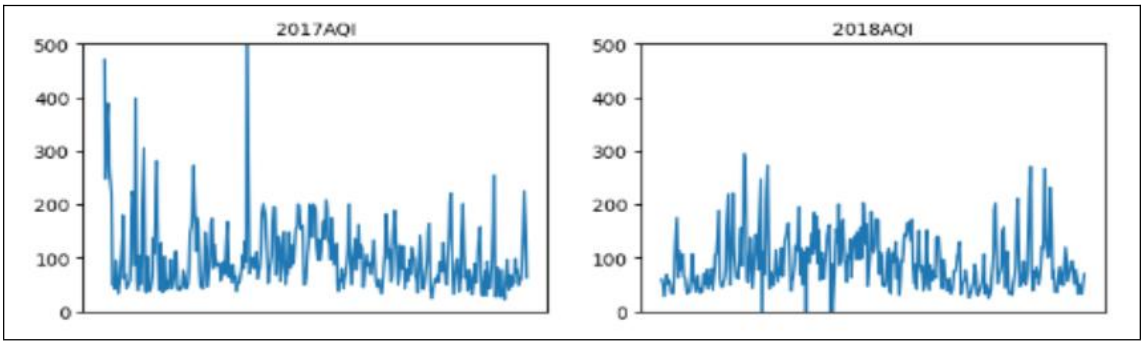

图 2-22 AQI 折线图

根据此可视化效果，可以清晰地看出 2017 年的 AQI 指数整体偏高，说明污染较严重，空气质量较差。这样的情况到 2018 年得到了改善，因此，2018 年是北京市空气污染治理的一个重要节点。

（3）AQI 和 PM2.5 的散点图

```
plt.figure(figsize=(5, 5))
plt.title('AQI 和 PM2.5 的关系')
plt.rcParams['font.sans-serif'] = ['SimHei']     #解决中文显示问题
plt.rcParams['axes.unicode_minus'] = False       # 解决中文显示问题
plt.scatter(data['AQI'],data['PM2.5'],label ='AQI')
```

可视化效果如图 2-23 所示。

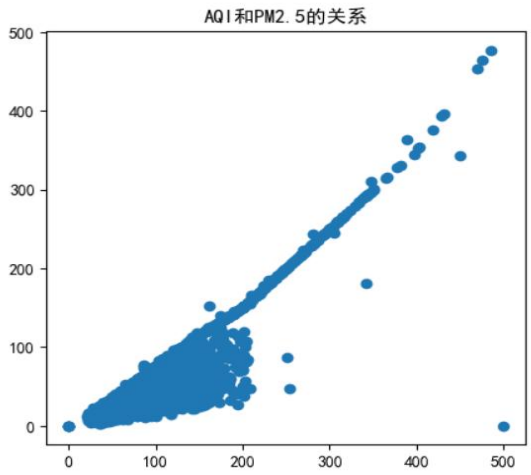

图 2-23 房屋户型分布柱状图

以 AQI 为横坐标，PM2.5 为纵坐标做散点图。可以发现 AQI 和 PM2.5 之间存在一个线性的关系，说明 PM2.5 有可能是空气中的主要污染物。在 AQI0-200 的范围内有大量的点，此时 PM2.5 的值在对应线性关系之下，说明在 AQI 0-200 范围时，还受了其他因素的影响。

（4）空气质量等级的饼状图

```
lt.rcParams['font.sans-serif'] = ['SimHei']     #解决中文显示问题
plt.rcParams['axes.unicode_minus'] = False       # 解决中文显示问题
label = [u'优', u'良', u'轻度污染', u'中度污染', u'重度污染', u'严重污染']
color = ['#32CD32','#FFDAB9','#8A2BE2','#2442aa','#dd5555','#FFFF00']

fig, axes = plt.subplots(3,3,figsize=(15, 10))
axes_list = []
for i in range(axes.shape[0]):
    for j in range(axes.shape[1]):
        axes_list.append(axes[i, j])

for ax in axes_list:
    ax.set_ylim([0,500])
    ax.set_xticks([])
axes[0,0].set_title("2013 空气质量情况分布",size=10)
axes[0,1].set_title("2014 空气质量情况分布",size=10)
axes[0,2].set_title("2015 空气质量情况分布",size=10)
axes[1,0].set_title("2016 空气质量情况分布",size=10)
axes[1,1].set_title("2017 空气质量情况分布",size=10)
axes[1,2].set_title("2018 空气质量情况分布",size=10)
axes[2,0].set_title("2019 空气质量情况分布",size=10)
axes[2,1].set_title("2020 空气质量情况分布",size=10)
```

```
data2013 = data[data['日期'].str.startswith('2013')]
data2014 = data[data['日期'].str.startswith('2014')]
data2015 = data[data['日期'].str.startswith('2015')]
data2016 = data[data['日期'].str.startswith('2016')]
data2017 = data[data['日期'].str.startswith('2017')]
data2018 = data[data['日期'].str.startswith('2018')]
data2019 = data[data['日期'].str.startswith('2019')]
data2020 = data[data['日期'].str.startswith('2020')]
PieData = [data['质量等级'].str.contains('优').sum(),
           data['质量等级'].str.contains('良').sum(),
           data['质量等级'].str.contains('轻度污染').sum(),
           data['质量等级'].str.contains('中度污染').sum(),
           data['质量等级'].str.contains('重度污染').sum(),
           data['质量等级'].str.contains('严重污染').sum()
           ]
PieData2013 = [data2013['质量等级'].str.contains('优').sum(),
           data2013['质量等级'].str.contains('良').sum(),
           data2013['质量等级'].str.contains('轻度污染').sum(),
           data2013['质量等级'].str.contains('中度污染').sum(),
           data2013['质量等级'].str.contains('重度污染').sum(),
           data2013['质量等级'].str.contains('严重污染').sum()
           ]
PieData2014 = [data2014['质量等级'].str.contains('优').sum(),
           data2014['质量等级'].str.contains('良').sum(),
           data2014['质量等级'].str.contains('轻度污染').sum(),
           data2014['质量等级'].str.contains('中度污染').sum(),
           data2014['质量等级'].str.contains('重度污染').sum(),
           data2014['质量等级'].str.contains('严重污染').sum()
           ]
PieData2015 = [data2015['质量等级'].str.contains('优').sum(),
           data2015['质量等级'].str.contains('良').sum(),
           data2015['质量等级'].str.contains('轻度污染').sum(),
           data2015['质量等级'].str.contains('中度污染').sum(),
           data2015['质量等级'].str.contains('重度污染').sum(),
           data2015['质量等级'].str.contains('严重污染').sum()
           ]
```

```
PieData2016 = [data2016['质量等级'].str.contains('优').sum(),
               data2016['质量等级'].str.contains('良').sum(),
               data2016['质量等级'].str.contains('轻度污染').sum(),
               data2016['质量等级'].str.contains('中度污染').sum(),
               data2016['质量等级'].str.contains('重度污染').sum(),
               data2016['质量等级'].str.contains('严重污染').sum()
               ]
PieData2017 = [data2017['质量等级'].str.contains('优').sum(),
               data2017['质量等级'].str.contains('良').sum(),
               data2017['质量等级'].str.contains('轻度污染').sum(),
               data2017['质量等级'].str.contains('中度污染').sum(),
               data2017['质量等级'].str.contains('重度污染').sum(),
               data2017['质量等级'].str.contains('严重污染').sum()
               ]
PieData2018 = [data2018['质量等级'].str.contains('优').sum(),
               data2018['质量等级'].str.contains('良').sum(),
               data2018['质量等级'].str.contains('轻度污染').sum(),
               data2018['质量等级'].str.contains('中度污染').sum(),
               data2018['质量等级'].str.contains('重度污染').sum(),
               data2018['质量等级'].str.contains('严重污染').sum()
               ]

PieData2019 = [data2019['质量等级'].str.contains('优').sum(),
               data2019['质量等级'].str.contains('良').sum(),
               data2019['质量等级'].str.contains('轻度污染').sum(),
               data2019['质量等级'].str.contains('中度污染').sum(),
               data2019['质量等级'].str.contains('重度污染').sum(),
               data2019['质量等级'].str.contains('严重污染').sum()
               ]
PieData2020 = [data2020['质量等级'].str.contains('优').sum(),
               data2020['质量等级'].str.contains('良').sum(),
               data2020['质量等级'].str.contains('轻度污染').sum(),
               data2020['质量等级'].str.contains('中度污染').sum(),
               data2020['质量等级'].str.contains('重度污染').sum(),
               data2020['质量等级'].str.contains('严重污染').sum()
               ]
axes[0,0].pie(PieData2013, labels = label, colors = color, labeldistance = 1.2, startangle = 90,
```

```
shadow = True,autopct='%3.1f%%')
axes[0,1].pie(PieData2014, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[0,2].pie(PieData2015, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[1,0].pie(PieData2016, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[1,1].pie(PieData2017, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[1,2].pie(PieData2018, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[2,0].pie(PieData2019, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
axes[2,1].pie(PieData2020, labels = label, colors = color, labeldistance = 1.2, startangle = 90, shadow = True,autopct='%3.1f%%')
plt.title('2013-2020 空气质量等级分布')
plt.radius = 20
plt.pie(PieData, labels = label, colors = color, startangle = 90, shadow = True,autopct='%3.1f%%')
```

可视化效果如下：

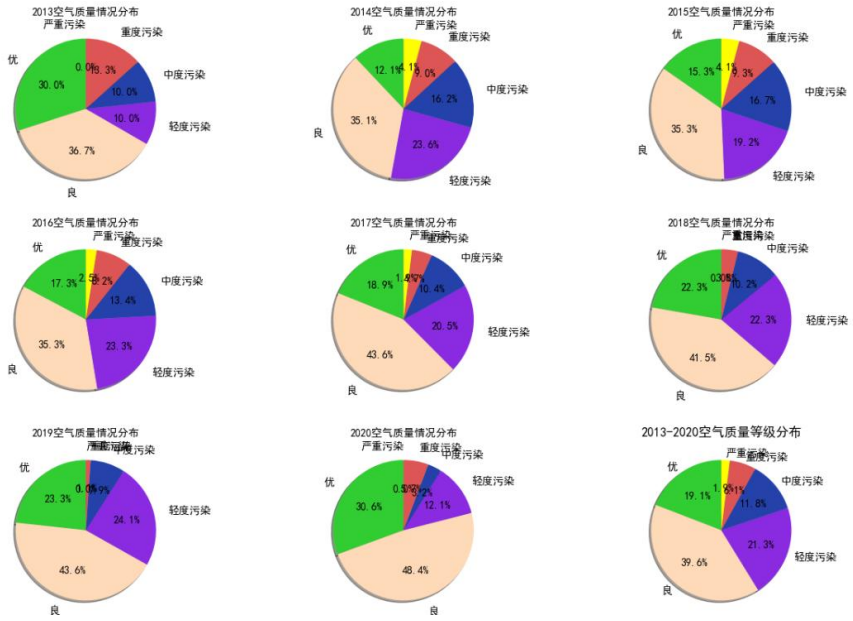

图 2-24 空气质量等级的饼状图