Python数据分析实战

第十六节课 pandas数据统计函数

本节课程目标

- 汇总类统计
- 唯一去重和按值计算
- 相关系数和协方差

0.读取CSV数据

```
import pandas as pd
filePath = './datas/beijing_tianqi_2018.csv'
df = pd.read_csv(filePath)
```

```
df.head()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	ymd	bWendu	yWendu	tianqi	fengxiang	fengli	aqi	aqiInfo	aqiLevel
0	2018-01-01	3℃	-6℃	晴~多云	东北风	1-2级	59	良	2
1	2018-01-02	2℃	-5℃	阴~多云	东北风	1-2级	49	优	1
2	2018-01-03	2℃	-5℃	多云	北风	1-2级	28	优	1
3	2018-01-04	0℃	-8℃	阴	东北风	1-2级	28	优	1
4	2018-01-05	3℃	-6℃	多云~晴	西北风	1-2级	50	优	1

```
#替换掉温度后面的℃
df.loc[:,'bWendu'] = df['bWendu'].str.replace('℃','').astype('int32')
df.loc[:,'yWendu'] = df['yWendu'].str.replace('℃','').astype('int32')
```

```
df.head(20)
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	ymd	bWendu	yWendu	tianqi	fengxiang	fengli	aqi	aqiInfo	aqiLevel
0	2018-01-01	3	-6	晴~多云	东北风	1-2级	59	良	2
1	2018-01-02	2	-5	阴~多云	东北风	1-2级	49	优	1
2	2018-01-03	2	-5	多云	北风	1-2级	28	优	1
3	2018-01-04	0	-8	阴	东北风	1-2级	28	优	1
4	2018-01-05	3	-6	多云~晴	西北风	1-2级	50	优	1
5	2018-01-06	2	-5	多云~阴	西南风	1-2级	32	优	1
6	2018-01-07	2	-4	阴~多云	西南风	1-2级	59	良	2
7	2018-01-08	2	-6	晴	西北风	4-5级	50	优	1
8	2018-01-09	1	-8	晴	西北风	3-4级	34	优	1
9	2018-01-10	-2	-10	晴	西北风	1-2级	26	优	1
10	2018-01-11	-1	-10	晴	北风	1-2级	24	优	1

11	2018-01-12	2	-8	晴	西南风	1-2级	75	良	2
12	2018-01-13	3	-7	多云	南风	1-2级	126	轻度污染	3
13	2018-01-14	6	-5	晴~多云	西北风	1-2级	187	中度污染	4
14	2018-01-15	2	-5	阴	东南风	1-2级	47	优	1
15	2018-01-16	4	-5	多云	南风	1-2级	112	轻度污染	3
16	2018-01-17	6	-7	多云~晴	西北风	1-2级	82	良	2
17	2018-01-18	5	-6	晴	西南风	1-2级	80	良	2
18	2018-01-19	7	-4	晴	南风	1-2级	115	轻度污染	3
19	2018-01-20	3	-6	晴~多云	东风	1-2级	64	良	2

汇总类统计

```
df.describe()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	bWendu	yWendu	aqi	aqiLevel
count	365.000000	365.000000	365.000000	365.000000
mean	18.665753	8.358904	82.183562	2.090411
std	11.858046	11.755053	51.936159	1.029798
min	-5.000000	-12.000000	21.000000	1.000000
25%	8.000000	-3.000000	46.000000	1.000000
50%	21.000000	8.000000	69.000000	2.000000
75%	29.000000	19.000000	104.000000	3.000000
max	38.000000	27.000000	387.000000	6.000000

```
df['bWendu'].mean()
```

```
18.665753424657535
```

```
#最高温
df['bWendu'].max()
```

```
38
```

```
#最低温
df['bWendu'].min()
```

```
-5
```

唯一去重和按值计数

唯一性去重

```
#例如上面的风向列
df['fengxiang'].unique()
```

```
array(['东北风', '北风', '西北风', '西南风', '南风', '东南风', '东风', '西风'], dtype=object)
```

```
df['tianqi'].unique()
```

```
array(['晴~多云', '阴~多云', '多云', '阴', '多云~晴', '多云~阴', '晴', '阴~小雪', '小雪~多云',
       '小雨~阴', '小雨~雨夹雪', '多云~小雨', '小雨~多云', '大雨~小雨', '小雨', '阴~小雨',
       '多云~雷阵雨', '雷阵雨~多云', '阴~雷阵雨', '雷阵雨', '雷阵雨~大雨', '中雨~雷阵雨', '小雨~大
雨',
       '暴雨~雷阵雨', '雷阵雨~中雨', '小雨~雷阵雨', '雷阵雨~阴', '中雨~小雨', '小雨~中雨', '雾~多
云',
       '霾'], dtype=object)
```

```
df['fengli'].unique()
```

```
array(['1-2级', '4-5级', '3-4级', '2级', '1级', '3级'], dtype=object)
```

按值计数

```
df['fengxiang'].value_counts()
```

```
南风      92
西南风     64
北风      54
西北风     51
东南风     46
东北风     38
东风      14
西风       6
Name: fengxiang, dtype: int64
```

```
df['tianqi'].value_counts()
```

```
晴          101
多云           95
多云~晴         40
晴~多云         34
多云~雷阵雨       14
多云~阴         10
小雨~多云         8
雷阵雨           8
阴~多云          8
雷阵雨~多云        7
小雨            6
多云~小雨         5
阴             4
雷阵雨~中雨        4
中雨~小雨         2
霾             2
阴~小雨          2
中雨~雷阵雨        2
小雪~多云         1
小雨~大雨         1
大雨~小雨         1
小雨~雨夹雪        1
雷阵雨~阴         1
雷阵雨~大雨        1
阴~雷阵雨         1
小雨~中雨         1
暴雨~雷阵雨        1
小雨~阴          1
小雨~雷阵雨        1
雾~多云          1
阴~小雪          1
Name: tianqi, dtype: int64
```

```
df['fengli'].value_counts()
```

```
1-2级    236
3-4级     68
1级       21
4-5级     20
2级       13
3级        7
Name: fengli, dtype: int64
```

相关系数和协方差

```
#协方差矩阵
df.cov()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	bWendu	yWendu	aqi	aqiLevel
bWendu	140.613247	135.529633	47.462622	0.879204
yWendu	135.529633	138.181274	16.186685	0.264165
aqi	47.462622	16.186685	2697.364564	50.749842
aqiLevel	0.879204	0.264165	50.749842	1.060485

```
#相关系数矩阵
df.corr()
```

```
.dataframe tbody tr th {
    vertical-align: top;
}

.dataframe thead th {
    text-align: right;
}
```

	bWendu	yWendu	aqi	aqiLevel
bWendu	1.000000	0.972292	0.077067	0.071999
yWendu	0.972292	1.000000	0.026513	0.021822
aqi	0.077067	0.026513	1.000000	0.948883
aqiLevel	0.071999	0.021822	0.948883	1.000000

```
df['aqi'].corr(df['bWendu'])
```

```
0.07706705916811067
```

```
df['aqi'].corr(df['yWendu'])
```

```
0.026513282672968895
```

```
# 空气质量和温差的相关系数
df['aqi'].corr(df['bWendu']-df['yWendu'])
```

```
0.2165225757638205
```

```

```