AI 程序设计@NJU

实验 5 SciPy 生态系统参考答案

1. 假设有 4 个人每周喝奶茶的杯数分别是 1、3、4、7，又假设这 4 个人的体重（kg）分别是 50、65、60、63，请分别用数学、基于 NumPy 数组和基于 pandas 的 corr()方法来计算两者之间的皮尔逊相关系数。
2. 利用 Tushare 包中的接口函数获取某公司 2019 年第一季度的股票数据并完成如下数据处理和分析任务：
(1) 数据只保留 date、open、high、close、low 和 volume 这几个属性，并按时间先后顺序对数据进行排序；
(2) 输出这一季度内成交量最低和最高那两天的日期和分别的成交量；
(3) 列出成交量在 1000000 以上的记录；
(4) 计算这半年中收盘价（close）高于开盘价（open）的天数；
(5) 计算前后两天开盘价的涨跌情况，用两种方式表示，第一种输出每两天之间的差值（后一天减去前一天），第二种输出一个开盘价涨跌列表，涨用 1 表示，跌用-1 表示；[提示：可使用 diff()方法和 sign()函数]
(6) 计算每月收盘价的平均值（提示：可使用 apply()方法）。
[提示：groupby()还常常与 apply()函数连用，apply()函数可将数据分拆、应用和汇总，使用自定义函数更灵活地进行各类数据统计。apply()函数的自由度很高，它的最基本形式为“DataFrame.apply(func, axis = 0)”，func 是函数，可以自己实现，默认 axis 为 0，表示 apply()函数会自动遍历 DataFrame 的每一列数据（一个 Series）按相应函数功能对其进行处理，处理结束后将所有结果组合后返回，若 axis 设为 1 则遍历处理 DataFrame 的每一行数据。]
(7) 绘制 2019 年 1 月该股票最高价 high 和最低价 low 的折线图；
(8) 绘制该股票在此季度内每日收盘价与开盘价之差与当日成交量之间的散点图。

1.
略

2.
```
import tushare as ts
import numpy as np

(1)
df = ts.get_hist_data('600036', start = '2019-01-01', end = '2019-03-31')
df = df.iloc[:, 0:5]
df.sort_index(inplace = True)   # 按 date 列进行排序
(2)
```

```
min_day = df.sort_values('volume').iloc[0,]
min_volume = min_day.volume
min_volume_date = min_day.name
print("the min volume of {} is at {}".format(min_volume, min_volume_date))
max_day = df.sort_values('volume').iloc[-1,]
max_volume = max_day.volume
max_volume_date = max_day.name
print("the max volume of {} is at {}".format(max_volume, max_volume_date))
(3)
print(df[df.volume >= 1000000])
(4)
print(len(df[df.close > df.open]))
(5)
print(df.open.diff())
print(np.sign(np.diff(df.open)))
(6)
month = [item[5:7] for item in df.index]
print(df.groupby(month).close.apply(np.mean))
(7)
df_new = df.loc[:,['high', 'low']]
df_new.sort_index().plot()
(8)
plt.scatter(df.close-df.open, df.volume)
```