# #补全代码，输出对应结果 （86题共86分）

import pandas as pd

import numpy as np

#1.将下面的字典创建为DataFrame

data = {"grammer":["Python","C","Java","GO",np.nan,"SQL","PHP","Python"],

"score":[1,2,np.nan,4,5,6,7,10]}

#2.提取含有字符串"Python"的行

#3.打印输出df的所有列名

#4.修改第二列列名为’popularity’

#5.统计grammer列中每种编程语言出现的次数

#6.将空值用上下值的平均值填充

#7.提取popularity列中值大于3的行

#8.按照grammer列进行去除重复值

#9.计算popularity列平均值

#10.将grammer列转换为list

#11.将DataFrame保存为EXCEL

#12.查看数据行列数

#13.提取popularity列值大于3小于7的行

#14.交换两列位置

#15.提取popularity列最大值所在行

#16.查看最后5行数据

#17.删除最后一行数据

#18.添加一行数据[‘Perl’,6.6]

#19.对数据按照"popularity"列值的大小进行排序

#20.统计grammer列每个字符串的长度

#################################################################

#21.读取本地EXCEL数据pandas120.xlsx

#22.查看df数据前5行

#23.将salary列数据转换为最大值与最小值的平均值

#24.将数据根据学历进行分组并计算平均薪资

#25.将createTime列时间转换为月-日

#26.查看索引、数据类型和内存信息

#27.查看数值型列的汇总统计

#28.新增一列categories根据salary将数据分为三组

bins = [0,5000, 20000, 50000]

group\_names = ['低', '中', '高']

#29.按照salary列对数据降序排列

#30.取出第33行数据

#31.计算salary列的中位数

#32.绘制薪资水平频率分布直方图

#33.绘制薪资水平密度曲线

#34.删除最后一列categories

#35.将df的第一列与第二列合并为新的一列test

#36.将education列与salary列合并为新的一列test1

#37.计算salary最大值与最小值之差

#38.将第一行与最后一行拼接

#39.将第8行数据添加至末尾

#40.查看每列的数据类型

#41.将createTime列设置为索引

#42.生成一个和df长度相同的随机数dataframe

#43.将上一题生成的dataframe与df合并

#44.生成新的一列new为salary列减去之前生成随机数列

#45.检查数据中是否含有任何缺失值

#46.将salary列类型转换为浮点数

#47.计算salary大于10000的次数

#48.查看每种学历出现的次数

#49.查看education列共有几种学历

#50.提取salary与new列的和大于40000的最后3行

###########################################################

51.# 创建serA数据：1，2，3；指定索引为a,b,c

# 创建serB数据：4，5，6；指定索引为b,c,d

# 创建serC为serA和serB相加的值

#52.通过pd读取http://192.168.105.242/test.php数据创建dataframe,命名df2

#53.查看每一列的均值

#54.查看回每一列的中位数

#55.获取行索引

#56.获取列索引

#57.排序，index降序排

#58.排序, column升序排

#59.用for循环逐行读取数据，打印输出year

#60.输出数据到df3.xlsx（输出包括标题和索引）

##########################################

#61.创建dataframe数据df3读取自sample.csv

#62筛选显示前4行，StockCode列DataFrame数据

#63.筛选数据,条件: StockCode 为 71053

#64.筛选索引偶数行

df = pd.read\_excel('sales-funnel.xlsx')

#65建立数据透视表，索引index：Manager,Rep ，指定 统计value为Price,Quantity，统计方式Price为平均，Quantity为求各，统计columns为Product,显示相关产品销售情况对缺失值填充0

df1 = pd.DataFrame({'lkey': ['foo', 'bar', 'baz', 'lemon'],

'value': [1, 2, 3, 5]})

df2 = pd.DataFrame({'rkey': ['foo', 'bar', 'baz', 'jack'],

'value': [5, 6, 7, 8]})

#66.DataFrame内连接操作，left\_on='lkey',right\_on='rkey'

#67.DataFrame左连接操作，left\_on='lkey',right\_on='rkey'

#68.DataFrame右连接操作，left\_on='lkey',right\_on='rkey'

data = pd.read\_csv('groupby.csv')

#69.计算不同公司员工的平均年龄和平均薪水

#70.计算不同公司员工的平均年龄和平均薪水

#71. 计算不同公司员工的平均年龄以及薪水的中位数

#72.计算不同公司员工的平均年龄以及薪水的中位数，均值，总人数

df = pd.DataFrame({"Name":["Alice", "Bob", "Mallory", "Mallory", "Bob" , "Mallory"],

"City":["Seattle", "Seattle", "Portland", "Seattle", "Seattle", "Portland"],

"Val":[4,3,3,np.nan,np.nan,4]})

#73. 对city分组，对val进行不包含NaN值和包含NaN值分别进行计数，并统计人员数量（name不重复计数）

data = pd.read\_csv('map.csv')

#74.把data 的 gender做替换， 男 --> 1，女 --> 0，其它值NaN

df = pd.DataFrame(

{

'名称':['产品1','产品2','产品3','产品4','产品5'],

'金额':[0,0.48,'F',0.74,0],

'合计':[0.37,0.28,'C',0.57,0.06],

}

)

#75.把df中的F替换为0,其它值保持不变

score = np.random.randint(25, 100, size=20)

s = pd.DataFrame(score,columns=['score'])

#76.对s成绩进行分箱并打标，分箱中0-60（不含）打标为不合格，60-70（不含）为合格，70-80良好，80-90为优秀，打标结果生成一列range

#77.对s各成绩段数量进行统计

df = pd.read\_csv('apply.csv')

#78.增加一列total\_score，如果民族不是汉族，则在score分数上加10分，不然值就是score

#79.增加一列interval，值为date\_from到date\_to的间隔天数

#80.增加一列label,如果score >=600 且 interval<=15 则打个标签 --> 1,否则为 0

data = {'company': ['B', 'A', 'B', 'B'],

'gender': ['female', 'female', 'male','male'],

'num': [40, 31, 28,28]}

df = pd.DataFrame(data)

#81.将列名 num --> age将索引 0-3 --> A-D

#82.清除company重复值, 保留最后那条记录。

#83.删除第一行和age列

#84.筛选gender为female的记录。

data = {'company': ['B', 'A', 'B', np.nan],

'gender': ['female', 'female', 'male','male'],

'age': [40, 31, np.nan,28]}

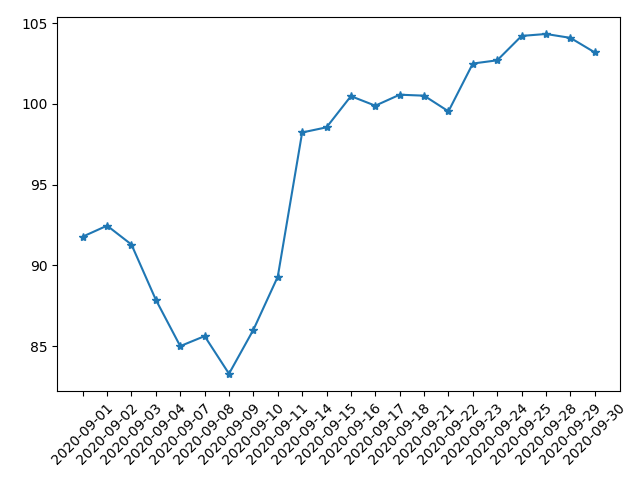
#85.筛选出company为NaN的记录

#86.填充age缺失的数据， 值采用age前后年龄记录的平均值

### 金融数据可视化 共14分

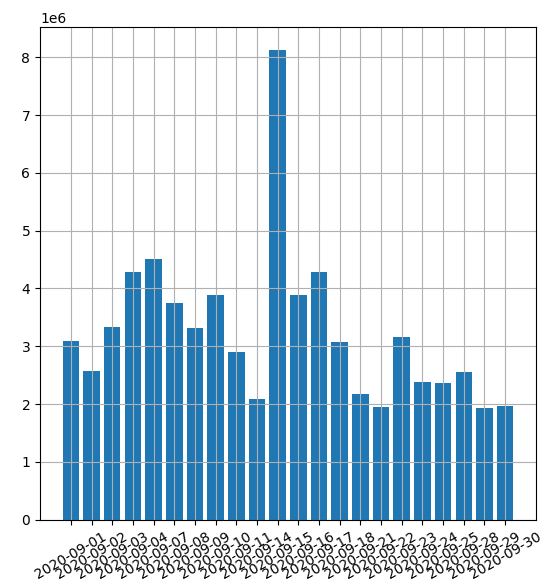
#### A. 利用Matplotlib对金融数据进行可视化4分

1. 读文件“600132.csv”中的股票数据，利用Python和Matplotlib绘制2020年9月收盘价线型图，为每个数据点加标识“\*”。



代码截图：

2. 绘制每天成交量的柱形图。4分



代码截图：

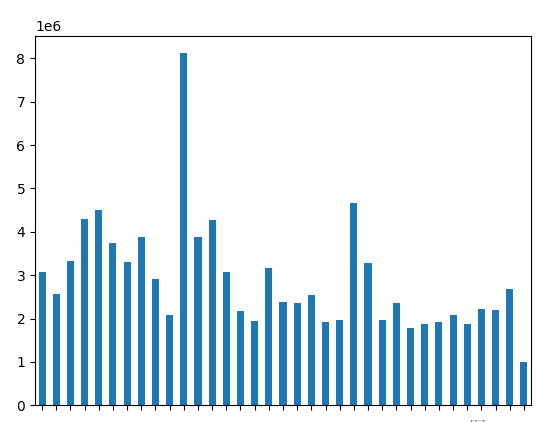
#### B. 利用Pandas对金融数据进行可视化6分

Pandas是专门用于数据分析的库，其提供的read\_csv()方法可以方便的读取csv格式的文件中的数据到Pandas中的DataFrame中。

DataFrame.plot(kind='line')

可以通过修改kind参数值为“line”、“bar”、“barh”、“hist”、“box”、“kde”、“density”、 “area”、“pie”、“scatter”或“hexbin”绘制线型图、柱型图、直方图等不同类型的图。

绘制日成交量的柱型图：



代码截图：