高三上信息专项练习——pandas(作业 35)

1. 第 19 届杭州亚运会已完美落幕，亚运会赛事以"杭州为主，全省共享"的原则分布在杭州、宁波、温州、湖州、绍兴、金华各地。大会共有 40 个大项，61 个分项，最终诞生了 481 块金牌。小李作为一名体育爱好者想重温大会赛程安排，他从杭州亚运会的官网上采集了相关数据，整理后存储在"杭州第 19 届亚运会总赛程.xlsx"文件中（0 表示有赛事但不产生金牌，其他数字表示当天产生的金牌数）。

	A	B	C	D	E	F	G	T	U	V	W	X	Y
1	大项	比赛项目	9/19	9/20	9/21	9/22	……	10/6	10/7	10/8	金牌总数	竞赛场馆	城市
2	游泳	游泳-花样游泳						0	1	1	2	杭州奥体中心游泳馆	杭州
3	游泳	游泳-跳水									10	杭州奥体中心游泳馆	杭州
4	游泳	游泳-马拉松游泳						1	1		2	千岛湖体育中心游泳馆	杭州
26	马术	马术						1			6	桐庐马术中心	杭州
27	击剑	击剑									12	杭州电子科技大学体育馆	杭州
28	足球	足球	0		0			1	1		2	黄龙体育中心体育场	杭州
29	足球	足球	0		0	0						临平体育中心体育场	杭州
30	足球	足球		0		0			0			上城体育中心体育场	杭州
31	足球	足球	0	0	0	0						萧山体育中心体育场	杭州
32	足球	足球	0		0							金华体育中心体育场	金华
33	足球	足球	0		0							浙江师范大学东体育场	金华
34	足球	足球			0	0						温州奥体中心体育场	温州
35	足球	足球				0						温州体育中心体育场	温州
36	高尔夫球	高尔夫球									4	千岛湖高尔夫球俱乐部	杭州
76	武术	武术									15	萧山瓜沥文化体育中心	杭州

第 1 题图 a

为了更清楚地了解相关赛事信息，小李编写了 Python 程序，请回答以下问题。

(1)足球是小李最关注的大项，为了解足球的赛程安排，划线处应填入的代码为▲_____

```
import pandas as pd
df=pd.read_excel("杭州第 19 届亚运会总赛程.xlsx")
df2=▲_____
print(df2)
```

df[df.大项=='足球']　　df[df['大项']=='足球']

df[df.比赛项目=='足球']　　df[df['比赛项目']=='足球']

(2)足球项目的比赛分布在杭州等城市的八个场馆，了解各个场馆举办足球赛事的具体场次，找到连续举办足球赛事最多的场馆，如第 1 题图 b 所示。程序代码如下，请在划线处填入合适的代码。

```
黄龙体育中心体育场[9/19,9/21,9/24,9/27,9/28,10/1,10/4,10/6,10/7]
临平体育中心体育场[9/19,9/21,9/22,9/24,9/25,9/28,9/30,10/3]
上城体育中心体育场[9/20,9/22,9/24,9/25,9/27,9/28,10/1,10/3,10/7]
萧山体育中心体育场[9/19,9/20,9/21,9/22,9/24,9/25,9/27,9/28,10/1,10/4]
金华体育中心体育场[9/19,9/21,9/24,9/27]
浙江师范大学东体育场[9/19,9/21,9/24,9/27]
温州奥体中心体育场[9/21,9/22,9/24,9/25,9/27,9/28,9/30]
温州体育中心体育场[9/22,9/24,9/25,9/27,9/28,9/30]
连续举办足球赛事最多的场馆：萧山体育中心体育场
```

第 1 题图 b

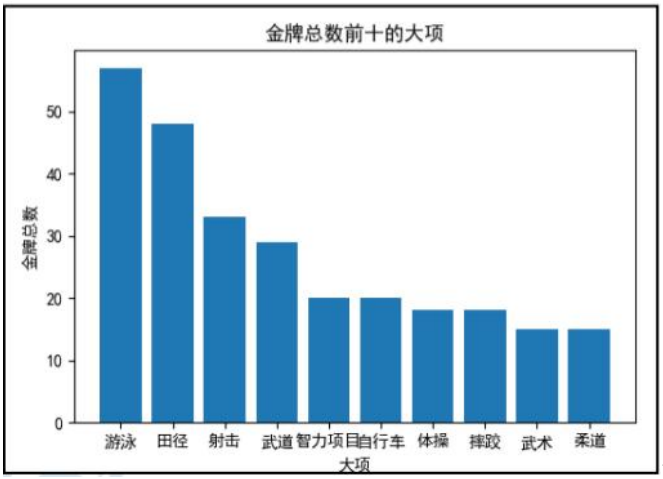
```
dic={};max=0;maxsta=""
lst=df2.竞赛场馆.tolist()        #将数据转换为列表
for i in lst:
    dic[i]=[]
for i in df2.index:
    c=0
    sta=df2["竞赛场馆"][i]
    for j in df2.columns[2:-3]:
        if df2.at[i,j]==0 or df2.at[i,j]==1:
            ①_____
            c=c+1
            if c>max:
                max=c;maxsta=sta
        else:
            ②_____
for i in dic:
    print(i,dic[i])
print(" 连续举办足球赛事最多的场馆：",③_____)
```

① dic[sta].append(j)或 dic[sta]+=[j]

②c=0

③ maxsta

(3)足球在亚运会期间总共产生两枚金牌，统计分析其他大项产生的金牌总数，找出产生金牌总数最多的十个大项，并绘制图形如第1题图c所示。程序代码如下，请在划线处填入合适的代码。



第1题 图c

总赛程.xlsx" 文件中（0 表示有赛事但不产生金牌，其他数字表示当天产生

	A	B	C	D	E	F	G	T	U	V	W	X	Y
1	大项	比赛项目	9/19	9/20	9/21	9/22	……	10/6	10/7	10/8	金牌总数	竞赛场馆	城市
2	游泳	游泳-花样游泳						0	1	1	2	杭州奥体中心游泳馆	杭州
3	游泳	游泳-跳水									10	杭州奥体中心游泳馆	杭州
4	游泳	游泳-马拉松游泳									2	钱	杭
26	马术	马术						1			6	桐庐马术中心	杭州
27	击剑	击剑									12	杭州电子科技大学体育馆	杭州
28	足球	足球	0		0			1	1		2	黄龙体育中心体育场	杭州
29	足球	足球	0		0	0						临平体育中心体育场	杭州
30	足球	足球		0		0			0			上城体育中心体育场	杭州

```
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['simhei']#图表中文标签显示为黑体
grp=____①____
grp=____②____
x=grp.大项
y=grp.金牌总数
plt.bar(x,y)
#设置绘图参数，代码略
```

df.groupby('大项', as_index=False).sum()
df.groupby('大项', as_index=False)['金牌总数'].sum()
①df.groupby('大项', as_index=False).金牌总数.sum()
②grp.sort_values("金牌总数",ascending=False).head(10)

2. 小林要使用python程序对无人超市一段时间内的经营数据进行分析。导出的销售数据“销售清单.xlsx”如第2题图a所示。请回答下列问题。

(1)要分析“生鲜”类商品单次销售额最高的数据(如图a所示的部分商品，单次销售最高是145.6所在的数据行)，假设数据保存在DataFrame对象df中，下列方法可以实现的是_▲C_(单选，填字母)。

A. 对df中数据按金额降序排序，取排在最前面的数据 最高的不一定是生鲜

B. 对df中数据按品类升序排序并保存到dfs，再对dfs按金额降序排序，取排序后dfs中最前面的数据

C. 从df中筛选出品类为“生鲜”的数据dfs，再对dfs按金额降序排序，取排序后dfs中最前面的数据

第一次排序结果会被第二次覆盖

	A	B	C	D	E	F	G	H	I
1	序号	品类	名称	规格	数量	单价	金额	支付	支付时间
2	1	生鲜	猪排骨	斤	1.4	30	42	支付宝	2024-01-13 00:03:41
3	2	生鲜	猪肉	斤	1.1	20	22	微信	2024-03-25 09:10:31
4	3	生鲜	牛肉	斤	1.6	60	96	微信	2024-01-17 15:31:55
5	4	生鲜	牛排骨	斤	5.2	28	145.6	余额	2024-03-05 13:29:15
6	5	食品	唐师傅方便面	袋	2	19.8	39.6	支付宝	2024-01-15 01:21:23
2134	2133	水果	巴西香蕉	斤	2.8	12.5	35	余额	2023-10-27 06:27:11
2135	2134	饮料	哇哈哈纯净水	瓶	5	2	10	余额	2023-11-14 07:18:26
2136	2135	食品	盼盼面包	袋	1	18.9	18.9	余额	2023-04-27 17:03:36

第2题 图a

(2)要统计这段经营时间内每月的销售总额最大是多少，小林设计了如下Python程序，请在划线处填入合适的代码。

相同的情况不考虑

```
import pandas as pd
df=pd.read_excel("销售清单.xlsx")    #读入原始数据
df=df.sort_values("支付时间")    #按支付时间升序排序
```

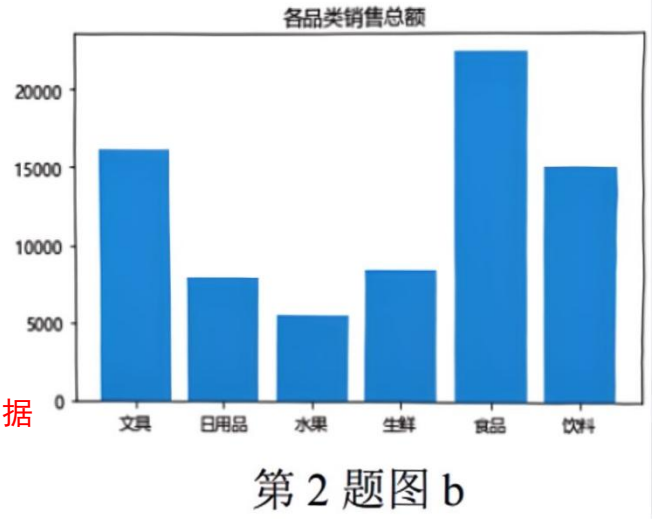
金额	支付时间
xxx	240101
xxx	240102
xxx	240201
xxx	240202
xxx	240301
xxx	240302

```
df=df.sort_values("支付时间") #按支付时间升序排序
f={}
prevm=maxe=0
for i in df.index:
    t=str(df.at[i,"支付时间"])          #通过行、列标签取得单个值
    m=t[0:7]
    if m in f:
        f[m]+=df.at[i,"金额"]
    else:  月首
        f[m]=df.at[i,"金额"]
        if prevm!=0 and ____①____:
            ____②____
更正缩进 prevm=m
    if maxe<f[prevm]:
        maxe=f[prevm]
print(maxe)
```

maxe<f[prevm]

maxe=f[prevm]

每月月初 统计上月数据
最后1个月通过 统计



第 2 题图 b

(3)要统计这段经营时间内每个品类总销售额，绘制如第14题图b 所示的柱形图，实现该功能的部分Python程序如下，请在划线处填入合适的代码。

```
dfg=____③____
x=dfg.index
y=dfg.values
plt.bar(x,y)
#设置绘图参数，代码略
```

df.grouby("品类", as_index=True).金额.sum()

df.grouby("品类", as_index=True)['金额'].sum()

3. 小明收集了浙江省某地 3 月各类共享单车的部分骑行数据记录，每天的用户数据存储在 “bike.xlsx” 文件中（不考虑跨天数据）。部分数据格式如图 a 所示，请回答下列问题：

(1)trans 函数功能为：读取骑行开始时间或结束时间，获取小时和分钟部分，转换为分钟格式并返回，如“2024/03/01 21:12”获取“21:12”转换为 1272(21*60+12=1272),代码如下。请在划线处填入合适的代码。

```
def trans(t):
    n=len(t)
    for i in range(n):
        if t[i]==" ":
            p=i
        ift[i]==":"
            q=i
    time=____________
    return time
```

int(t[p+1:q])*60+int(t[q+1:])

int(t[p+1:p+3])*60+int(t[q+1:q+3])

用户编号	单车类型	开始时间	结束时间
81	美团单车	2024/03/01 21:12	2024/03/01 22:14
6	摩拜单车	2024/03/01 05:45	2024/03/01 07:38
39	青桔单车	2024/03/01 19:06	2024/03/01 21:03
54	美团单车	2024/03/01 03:05	2024/03/01 04:37
15	OFO单车	2024/03/02 02:35	2024/03/02 04:11
66	青桔单车	2024/03/02 05:59	2024/03/02 07:31
16	摩拜单车	2024/03/02 03:06	2024/03/02 04:48
28	摩拜单车	2024/03/02 16:15	2024/03/02 18:05

图 a

(2)共享单车计费规则如下：起步价 1.5 元(含 15 分钟),超出 15 分钟，时长费 0.5 元/15 分钟，不足 15 分 钟以 15 分钟计算。考虑到车辆坏损等情况，2 分钟内(不含 2 分钟)的骑行数据作废，因此本程序实现过滤骑行时间在 2 分钟内的数据行，统计各条记录的骑行时间及本月各类单车的收益。请在划线处填入合适的代码。

```
import pandas as pd
import math
df.insert(4,"骑行时长","")      #插入一列
dic={'青桔单车': 0, 'OFO 单车': 0, '美团单车' : 0 , '摩拜单车': 0 }
for i in df.index:
    c=trans(df.at[i,"结束时间"])-trans(df.at[i,"开始时间"])
    df.at[i,"骑行时长"]=c
    if c<2:
        fee=0
    elif 2<=c<=15:
        fee=1.5
    elif c>15:
        fee=1.5+math.ceil((c-15)//15)*0.5       # 函数 ceil()实现向上取整
    ① dic[df.at[i,'单车类型']]+=fee     #统计各类型单车的总收益存入 dic 中
df=   ②df[df.骑行时长>=2]              #过滤骑行 2 分钟内的数据行
print(df)
print(dic)
```

(3)统计各类型单车的骑行次数（不包含 2 分钟内的）并实现降序排序，绘制柱形图，代码如下，绘制的图表如图 b 所示，请在划线处填入合适的代码。

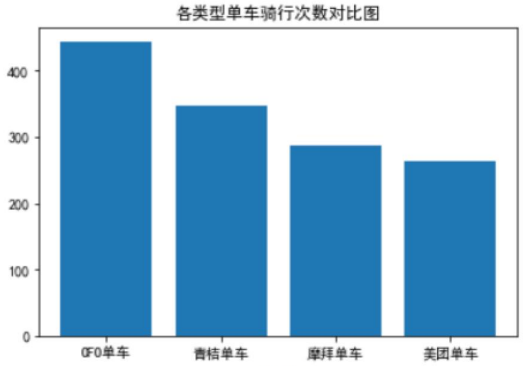


图 b

用户编号	单车类型	开始时间	结束时间
81	美团单车	2024/03/01 21:12	2024/03/01 22:14
6	摩拜单车	2024/03/01 05:45	2024/03/01 07:38
39	青桔单车	2024/03/01 19:06	2024/03/01 21:03
54	美团单车	2024/03/01 03:05	2024/03/01 04:37
15	OFO单车	2024/03/02 02:35	2024/03/02 04:11
66	青桔单车	2024/03/02 05:59	2024/03/02 07:31
16	摩拜单车	2024/03/02 03:06	2024/03/02 04:48
28	摩拜单车	2024/03/02 16:15	2024/03/02 18:05

"骑行时长"
也可以是"开始时间"、"结束时间"

```
import matplotlib.pyplot as plt
# 显示中文标签,代码略
df1=    ① df.groupby('单车类型', as_index=False)['骑行时长'].count()
                或df.groupby('单车类型', as_index=False).count()
df2=    ②df1.sort_values("骑行时长",ascending=False)
plt.title("各类型单车骑行次数对比图")
plt.bar(    ③df2.单车类型,df2.骑行时长    )
plt.show()
```

关联答案

```
df1= df.groupby('单车类型', as_index=True).count()

plt.bar(df2.index,df2.骑行时长)
```