# SAS BASE LAB 每日一练-Update by 4/11/2021

本lab机经使用说明：将文件夹cert拷贝至C盘根目录。

（1）在考试时：建立名为cert的library，c:\cert\input中有各题的sas原始数据inputXX。

libname cert ‘c:\cert\input’; 建立名为results的library，地址为c:\cert\output，所有大题均要求的过程数据集都要写为results.outputXX的形式。libname results ‘c:\cert\output’; 全部代码写完后，将program保存至c:\cert\programs。

（2）在使用本lab机经练习时，把Word格式代码直接复制进去当作题干即可，SAS格式代码为答案

第一题：

用libname方法引入c:\cert\input中名为input01的xlsx文件，要求library的名称叫certXL。此文件中都有哪些sheet名称？其中对第三个sheet中的height变量求平均数、求和。再查看第二个sheet中的第二条观测是什么。

libname certXL xlsx 'C:\cert\input\input01.xlsx';

**proc** **contents** data=certXL.\_all\_;

**run**;

**proc** **means** data=certXL.sheetC mean sum;

var height;

**run**;

**proc** **print** data=certXL.'sheet BB'n (obs=**2** firstobs=**2**);

**run**;

第二题：

用proc import方法引入c:\cert\input中名为input02的xlsx文件的sheetA，且要求程序自动识别变量名称。如果变量name的值是jordan则输出到results.output02\_1，如果值是bryant则输出到results.output02\_2。如果值是james则输出到临时results.output02\_3。自定义一个format，让bryant显示为kobebryant，让james显示为lebronjames。

libname cert 'C:\cert\input';

libname results 'C:\cert\output';

**proc** **import**

datafile='C:\cert\input\input02.xlsx'

dbms=xlsx

out=results.output02

replace;

sheet='sheetA';

getnames=yes;

**run**;

**proc** **format**;

value $fmt

'bryant' = 'kebobryant'

'james' = 'lebronjames';

**run**;

**data** results.output02\_1 results.output02\_2 results.output02\_3;

set results.output02;

if name='jordan' then output results.output02\_1;

if name='bryant' then output results.output02\_2;

if name='james' then output results.output02\_3;

**run**;

**proc** **print** data=results.output02\_1;

**run**;

**proc** **print** data=results.output02\_2;

format name $fmt.;

**run**;

**proc** **print** data=results.output02\_3;

format name $fmt.;

**run**;

第三题：

把major变量中含有and的都挑出来，and前面的专业重新命名为major变量，and后面的专业命名为second\_major变量。（1）主副专业里分别是哪个专业出现最多？（2）新建一个变量grade，根据学生的test1~test4成绩进行评价，平均成绩低于3.5分的为bad，大于等于3.5分的为good，然后筛选出来自senior，Beijing或Tianjin，good且没有second\_major的学生，输出结果要略去所有以test开头的变量。

libname cert 'c:\cert\input';

Data cert.input03;

Input name $ city $ 3-10 major $ 12-27 class $ test1 test2 test3 test4;

Datalines;

A Beijing math senior 5 4 3 4

B Beijing math senior 2 1 1 5

C Tianjin history senior 5 5 5 5

D Shanghai history and ball junior 4 4 4 4

E Hebei ball and reading junior 1 1 2 3

F Beijing run junior 2 2 1 2

G Beijing run and ball senior 5 4 5 1

;

Run;

libname results 'C:\cert\output';

**data** results.output03\_1;

set cert.input03;

n=find(major, 'and', 'i');

if n ne **0** then do;

second\_major=substr(major, n+**4**);

major=substr(major, **1**, n-**2**);

end;

**run**;

**proc** **freq** data=results.output03\_1 order=freq;

table major second\_major;

**run**;

**data** results.output03\_2(drop=n avg test:);

set cert.input03;

n=find(major, 'and', 'i');

if n ne **0** then do;

second\_major=substr(major, n+**4**);

major=substr(major, **1**, n-**2**);

end;

else if n=**0** then second\_major='none';

avg=mean(of test1 - test4);

if avg >= **3.5** then grade='good';

else if avg < **3.5** then grade='bad';

if class='senior' and (city='Beijing' or city='Tianjin') and n=**0** and grade='good';

**run**;

**proc** **print** data=results.output03\_2;

**run**;

考点：（1）查询字段要用什么function合适。（2）因为major涉及重定义，所以要借助原始major定义second\_major。（3）新建字符型变量的时候，先定义其长度。

第四题（我考试的时候没有遇到此题，前辈们的机经关于此考点也语焉不详，本题和第八题是两种可能的考查方式）：

将word变量中是两个字符长度的值大写，输出的时候把Word变量重命名为Wo。

libname cert 'c:\cert\input';

Data cert.input04;

Input word $;

Datalines;

she

it

He

And

WHO

;

Run;

libname results 'C:\cert\output';

**data** results.output04(drop=n rename=(word=wo));

set cert.input04;

n=length(word);

if n=**2** then word=upcase(word);

**run**;

**proc** **print** data=results.output04;

**run**;

第五题（别笑，这是一道真实的大题，简单吧？）：

数据集sashelp.class的属性里，encoding和label的值分别是什么。

**proc** **contents** data=sashelp.class;

**run**;

第六题：

在每个department中，挑出含有最高income的观测。

libname cert 'c:\cert\input';

Data cert.input06;

Input department $ income dollar6.;

Datalines;

Hr $1000

Eng $500

Hr $10000

Hr $7500

Eng $200

Clean $20

Eng $3000

Clean $10

;

Run;

libname results 'C:\cert\output';

**proc** **sort** data=cert.input06 out=sort\_input06;

by department descending income;

**run**;

**data** results.output06;

set sort\_input06;

by department descending income;

if first.department;

**run**;

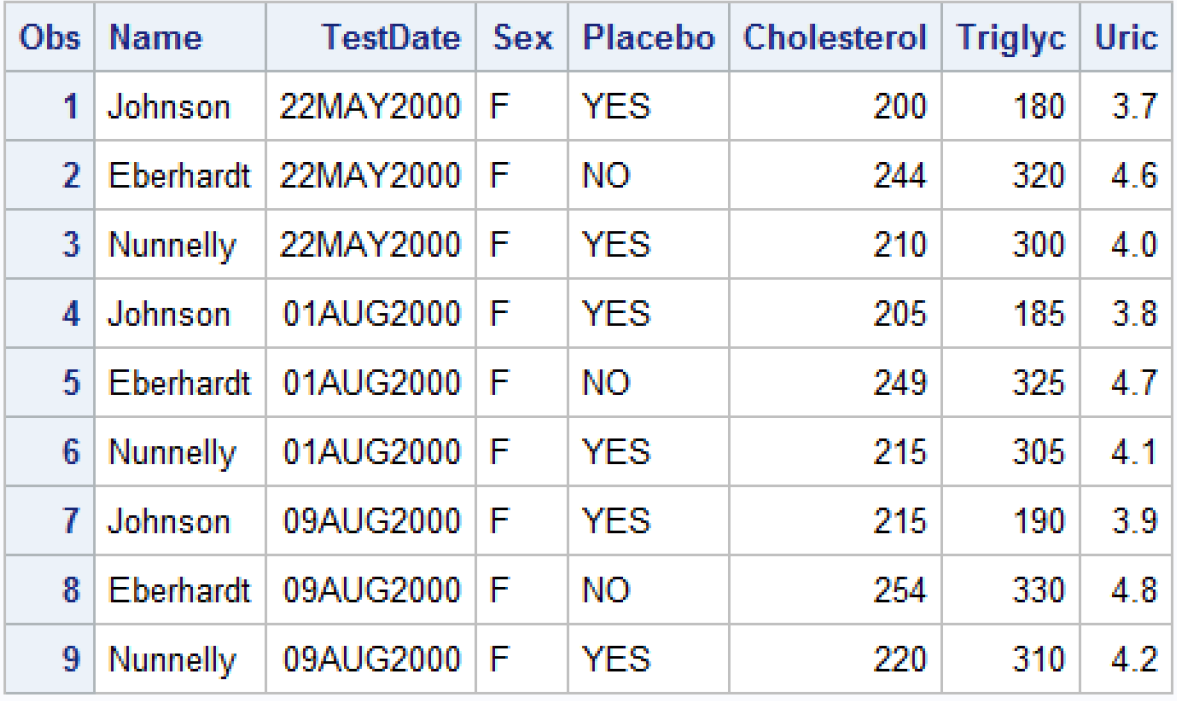
**proc** **print** data=results.output06;

**run**;

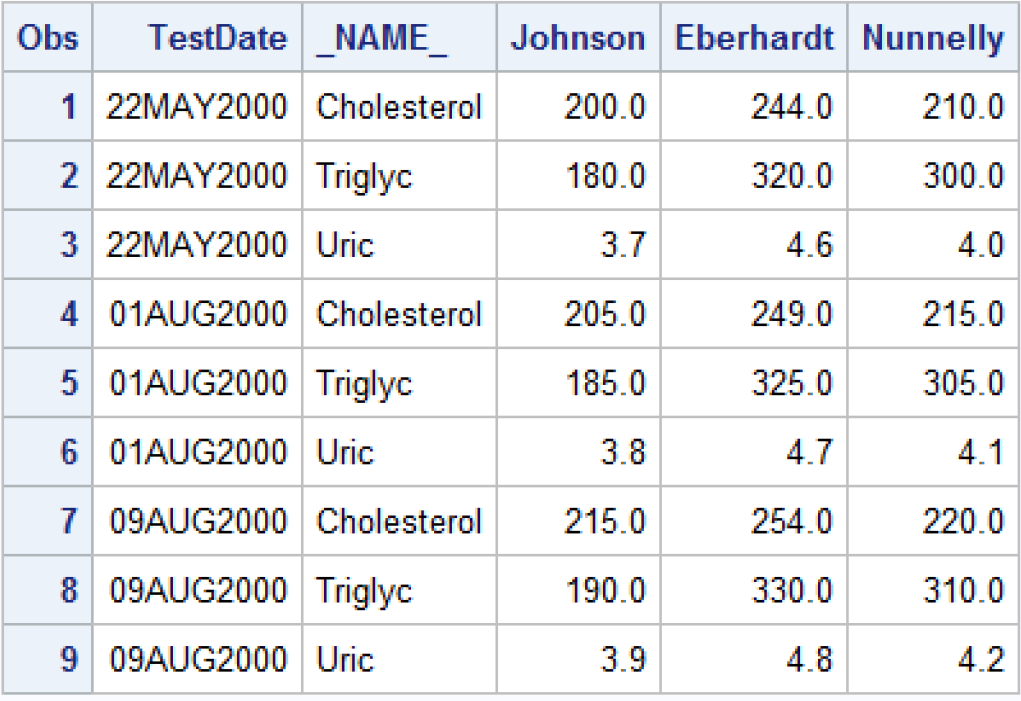
第七题：（正常来说这是一道填空题，只有一位前辈说见过lab题考transpose的，但是考的很基础，可能还不如填空题难）下面仅仅是个例子，并不是真题。（4/11/2021-update）

已知初始数据的模样和最终数据的模样，问transpose中的var、id、by都分别填什么。

The original data set



The transposed data set



**proc** **sort** data=cert.cltrials out=sort\_trials;

by testdate;

**run**;

**proc** **print** data=sort\_trials;

**run**;

**proc** **transpose** data=sort\_trials out=results.trans\_trials;

var cholesterol triglyc uric;

id name;

by testdate;

**run**;

**proc** **print** data=results.trans\_trials;

**run**;

第八题（我考试的时候没有遇到此题，前辈们的机经关于此考点也语焉不详，本题和第四题是两种可能的考查方式）：

引入c:\cert\input中名为input08的xlsx文件。在此文件的一个名叫state的sheet中有一些数据，请将变量location的值中的省份部分大写（提示：location中的值的样式为‘城市 省份’，比如Chengdu sichuan）。

libname cert08 xlsx 'C:\cert\input\input08.xlsx';

libname results 'C:\cert\output';

**data** results.output08;

set cert08.'state'n;

provice=scan(Location, **2**, ' ');

provice=upcase(provice);

**run**;

**proc** **print** data=results.output08;

**run**;

**OR:**

libname cert08 'C:\cert\input\input08.xlsx';

libname results 'C:\cert\output';

**data** results.output08;

set cert08.'state$'n;

province=scan(Location, **2**, '');

province=upcase(province);

**run**;

**proc** **print** data=results.output08;

**run**;

第九题：

提取custID中的数字部分，设为新变量a。再设立一个新变量b，b=a\*1000。最后把custID中的数字部分统一替换为9999。

libname cert 'c:\cert\input';

Data cert.input09;

Input custID $;

Datalines;

ID1573

ID2369

ID5361

ID8902

;

Run;

libname results 'C:\cert\output';

**data** results.output09;

set cert.input09;

a=substr(custID, **3**, **6**);

b=a\***1000**;

substr(custID, **3**, **6**)='9999';

**run**;

**proc** **print** data=results.output09;

**run**;

**OR**

**data** results.output09;

set cert.input09;

a=input(substr(custID, **3**, **6**), **4.**);

b=a\***1000**;

substr(custID, **3**, **6**)='9999';

**run**;

**proc** **print** data=results.output09;

**run**;

考点：（1）如果拆开了写，即先写a=sbustr(custID,3,4)再写a=input(a,4.)，会发现a的属性转不过来，还是字符型的。（2）9999要用引号括起来，即以字符型的属性反代回去，如果以数字型的属性，substr会自动将其用best12.的格式转化为字符型，9999的前面就会有空格，反代回去的结果就成了ID 9999而非正确答案ID9999。

第十题：

建立新的变量fullname，其值为原数据集中firstname和lastname的拼接。新建两个变量zflag 和nozflag，记录下fullname中有z（或Z）的有多少（有z/Z时，zflag = 1，nozflag = 0），没有z（或Z）的有多少（无z/Z时，zflag = 0，nozflag = 1），结果不要显示原始数据集的firstname和lastname变量。

libname cert 'c:\cert\input';

Data cert.input10;

Input firstname $ lastname $;

Datalines;

Lazrry brown

Kobe Bryant

Zach oneal

Kevin Durant

Lebron jamesz

chrisZ pual

chris Zwebber

;

Run;

libname results 'C:\cert\output';

**data** results.output10(drop=firstname lastname);

set cert.input10;

fullname=catx(' ', firstname, lastname);

z=find(fullname, 'z', 'i');

if z ne **0** then

do;

zflag=**1**;

nozflag=**0**;

end;

else if z=**0** then

do;

zflag=**0**;

nozflag=**1**;

end;

**run**;

**proc** **print** data=results.output10;

sum zflag nozflag;

**run**;

第十一题（改错题）：

原始数据集name里有叫Amanda、Tao或Chen的，挑出来。建立一个计数变量count。上述名字分别出现了几次？（港真，这题直接proc freq data = XXX; tables name; run;不就vans了吗，真不知道改成这么写算不算正确）

libname cert 'c:\cert\input';

Data cert.input11;

Input name $;

Datalines;

Amanda

Tao

Tao

Chen

Kobe

Chen

Amanda

Tao

Chen

Jordan

Tao

;

Run;

data output11;

set test11std;

by name;

if name in ('AMANDA' 'TAO' 'CHEN') then do;

if first.name then count=0;

count = count + 1;

else count = 0;

end;

run;

proc print data = output11;

run;

正确写法：

Method One:

libname results 'C:\cert\output';

**proc** **sort** data=cert.input11 out=test11std;

by name;

**run**;

**data** output11;

set test11std;

by name;

if name in ('Amanda', 'Tao', 'Chen') then

do;

if first.name then count=**0**;

count+**1**;

end;

else count=**0**;

**run**;

**proc** **print** data=output11;

**run**;

第十二题：

在每个department内，求income的总和。

libname cert 'c:\cert\input';

Data cert.input12;

Input department $ income dollar6.;

Datalines;

Hr $1000

Eng $500

Hr $10000

Hr $7500

Eng $200

Clean $20

Eng $3000

Clean $10

;

Run;

libname results 'C:\ Base\cert\_review\output';

**proc** **sort** data=cert.input12 out=sort\_input12;

by department income;

**run**;

title 'Calculate the total income';

**data** results.output12(drop=income);

set sort\_input12;

by department;

if first.department then total=**0**;

total+income;

if last.department;

**run**;

**proc** **print** data=results.output12;

**run**;

第十三题：

假设初始存款为50元，月利率为5%，第二年起每年年初再额外存入50元。（1）问3年后是多少元（答案：540.64元），请让程序显示每个月份的金额结果。（2）请让程序仅仅显示最终一条结果，在这条结果中，year和month的值是多少（答案：删掉上一问答案代码中的output;可看到本题答案分别为4和13）？

libname results 'C:\cert\output';

title 'Q13-1';

**data** results.output13\_1;

do year=**1** to **3**;

total+**50**;

do month=**1** to **12**;

total+total\***0.05**;

output;

end;

end;

**run**;

**proc** **print** data=results.output13\_1;

**run**;

title 'Q13-2';

**data** results.output13\_2;

do year=**1** to **3**;

total+**50**;

do month=**1** to **12**;

total+total\***0.05**;

end;

end;

**run**;

**proc** **print** data=results.output13\_2;

**run**;

第十三-2题：

假设初始存款为100元，月利率为5%，第二年起每年年初再额外存入50元。问3年后是多少元（答案：830.23元）？

libname results 'C:\cert\output';

**data** results.output13\_b;

total=**100**;

do year=**1** to **3**;

do month=**1** to **12**;

total+total\***0.05**;

output;

end;

total+**50**;

end;

**run**;

**proc** **print** data=results.output13\_b;

**run**;

第十四-1题（第十三题及其变种和第十四题及其变种中，本题是真题）：

假设初始存款为50元，年利率为60%，第二年起每年年初再额外存入50元。问多少年超过10000元（答案：第10年，14526.82元）？

libname results 'C:\cert\output';

**data** results.output14a;

do until(total>**10000**);

total+**50**;

total+total\***0.6**;

year+**1**;

output;

end;

**run**;

**proc** **print** data=results.output14a;

**run**;

第十四-2题：

假设初始存款为100元，年利率为60%，第二年起每年年初再额外存入50元。问多少年超过10000元（答案：第9年，12465.24元）？（变种拓展：如果每年续投款50元这个动作，从第一年起固定发生在每年的最后一天，代码应该怎么写？）

libname results 'C:\cert\output';

**data** results.output14b;

total=**100**;

do until(total>**10000**);

total+total\***0.6**;

year+**1**;

output;

total+**50**;

end;

**run**;

**proc** **print** data=results.output14b;

**run**;

变种拓展说明：每年续投款这个动作是发生在每年的最后一天还是次年的第一天，这对于最最后一条观测来说，是有很大区别的。大家可以试着把本代码中的output;挪到total + 50;后面去，对比一下两种结果，很容易就明白了。当然，考试不会考这么变态的。

第十五题：(4/11/2021 update)

One to one merge，在merge data1 data2; by XXX;之前，应当先进行什么操作（分别sort它们）？

libname cert 'C:\cert\input';

**data** cert.input15a;

Input product $ qty ;

Datalines;

cake 2

egg 5

water 10

apple 7

banana 2

meat 2

beef 3

pork 1

;

**Run**;

**data** cert.input15b;

Input product $ price;

Datalines;

cake 10

egg 5

water 1

apple 7

banana 2

meat 20

beef 30

pork 10

;

**Run**;

libname results 'C:\cert\output';

**proc** **sort** data=cert.input15a out=sort\_input15a;

by product;

**run**;

**proc** **sort** data=cert.input15b out=sort\_input15b;

by product;

**run**;

**data** results.output15;

merge sort\_input15a sort\_input15b;

by product;

total=price\*qty;

**run**;

**proc** **print** data=results.output15(obs=**3** firstobs=**3**);

var total;

**run**;

**proc** **means** data=results.output15(obs=**6** firstobs=**6**) mean;

var total;

**run**;

第十六题（可能会出改错题，也可能是选择题）：

If … then …; If … then …; Else …; 和 If … then …; Else if … then …;Else …;的区别是？

第十七题（这个挺容易弄混的，可能在选择题里考，大题不会考的）：

Id by class var tables都能用在下列哪种procedure里？独立或组合的使用效果是什么？

Proc print能用：id，by，var; Proc means能用：id，by，class，var ; Proc freq能用：tables

第十九题（Macro）：考试的时候一般是填空题, 涂黑部分为考点, 如果是character，需要双引号（4/11/2021 update）

%let Cyl\_Count=5;

**proc** **print** data=sashelp.cars;

where Cylinders=&Cyl\_Count;

var Type Make Model Cylinders MSRP;

**run**;

%let CarType=Wagon;

**proc** **print** data=sashelp.cars;

where Type="&CarType";

var Type Make Model MSRP;

**run**;