# STA 402/502 Homework 9

Due: November 5 (Monday), before class

**Please read the homework guidelines before working on the homework. Homework that does not follow the guidelines will be deducted points.** You are to complete this assignment on your own. Remember to include an intro comment block on all programs written. Each problem should be attempted as its own program.

**Do the following exercise and make sure you do not edit the given permanent dataset. Use ODS RTF to save the created tables (as well as corresponding titles) and include them in your homework.**

1. Do this problem using PROC SQL. Suppose that you have mortgage application data from a national bank with five branches in California contained in the SAS data set called LOANAPP. The Board of Directors would like an executive summary to assess their fixed-rate, home loan applicants with a comparison of the five branches. They are particularly concerned about loans that have less than 5% on the down payment. Variables in this data set include the identification information for the applicant, credit score, loan information, home price, and the down payment as a percentage of the sales price.

   (a) Examine this SAS data set including the variable labels and attributes. Create a table that provides information about the number of loan approvals, the total number of loan applications and the percent of loan approvals within each branch for loans with less than 5% down. Be sure to format the table and label variables according to the following requirements:
   - The branch names are "LIV925", $\cdots$, "COR760" instead of "1,2,3,4,5";
   - The number of loans approvals and the number of total loan applications are integers;
   - The percentage of loan approvals have a percentage sign after the number and has three decimal points. (e.g. xx.xxx%)

- Add a title called "Summary for Loan Applications with Less Than 5% Down" above the table;
  (hint: to add a title in PROC SQL, you may just use a statement:
  Title "Insert your title here";
  proc sql;
  ⋮
  )
- The table should have the following columns:

| Branch name | Number of loans approvals | Number of total loan applications | Percent of loan approvals |
|---|---|---|---|

(b) Create a table that calculates the count, mean loan amount, mean home price, and median credit score, by branch for all approved loans. Format the loan and price data to include dollar signs, commas, and three decimal places(e.g. $xxx,xxx,xxx.xx), and format the count and credit score to be whole numbers.

The branch name requirement is the same as in part (a), include an appropriate title and column label, so that the table is presentation-ready.

2. Do this problem using PROC SQL with the same dataset we have seen in Homework 7, Question 2. The World Health Organization (WHO) collected data in countries across the world regarding the outbreak of swine flu cases and deaths in 2009. The data in the SAS data set called SFF includes information on cases and deaths per country by month during the epidemic.

   (a) Create a dataset called "casedata" that contains the following variables, and use PROC PRINT with appropriate options to show the dataset with correct labels (Hint: use "proc print data=datasetname label;" to print the dataset with labels as header.):

   - A variable with label "Continent", which records the continent name.
   - A variable with label "Number of countries", which counts the number of countries within each continent.
   - A variable with label "Sum of cumulative cases reported on the first day of the month for July", which counts the number

of cumulative cases reported on the first day of the month for July.

(b) Generate a table containing the following variables (this dataset only has 6 observations) with specific labels. (Hint: you may use sum(some logic statement) to calculate the number of observations satisfying a specific logic condition.)

- A variable with label "Continent", which records the continent name.
- A variable that counts the number of countries per continent that reported no cases for the number of cumulative cases reported on the first day of August. Label this variable as "Number of countries reported no cases".
- A variable that counts the number of countries per continent that reported at least one case for the number of cumulative cases reported on the first day of August. Label this variable as "Number of countries reported cases".
- A variable that calculates the summation of cumulative cases per continent for the number of cumulative cases reported on the first day of August. Label this variable as "Sum of cumulative cases reported on the first day of August".

3. (**STA 502 only**) The local school district has a SAS data set called DISTRICT that contains the rubric used for classifying teachers annual evaluations. The rubric is based on two components: a teacher score and a curriculum grade. Using these components, an overall rating for the teacher can be determined. The data for 10 elementary school teachers who are due for evaluation are stored in the SAS data set called TEACHERS. Before working on the exercises below, examine these two SAS data sets including the variable name, labels and other attributes.

   Combine the two SAS data sets so that the district rating is properly assigned to each teacher. Sort the observations by teacher name so that administrators can easily locate each teachers rating. Make sure no missing value is included in the table. The table has the following columns:

| Teacher name | Teacher score | Curriculum grade | District rating |
|---|---|---|---|

4. Multiple Choice Questions (2pt for each question, you may just provide the answer.)

   (a) Given the SAS dataset TEST:

   ```
   Salary
   ------
   200
   205
   .
   523
   ```

   The following SAS program is submitted:

   ```
   proc sql;
   select *
   from TEST
   [_insert_where_clause_]
   ;
   quit;
   ```

   The following output is desired:

   ```
   Salary
   ------
   200
   205
   523
   ```

   Which WHERE expression completes the program and generates the desired output?

   A. where Salary is not .
   B. where Salary ne missing
   C. where Salary is not missing
   D. where Salary ne null

   (b) Given the SAS data set WORK.ONE:

   ```
   Rep Cost
   --------- ----
   SMITH 200
   SMITH 400
   JONES 100
   SMITH 600
   JONES 100
   ```

The following SAS program is submitted;

```
proc sql;
select
Rep,
avg(Cost)
from WORK.ONE
order by Rep
;
quit;
```

Which result set would be generated?

A.
```
JONES 100
JONES 100
SMITH 600
SMITH 600
SMITH 600
```

B.
```
JONES 600
SMITH 100
```

C.
```
JONES 280
SMITH 280
```

D.
```
JONES 280
JONES 280
SMITH 280
SMITH 280
SMITH 280
```

(c) Given the SAS dataset WORK.TWO

```
Name Salary
----- ------
Hans 200
Maria 205
Jose 310
Ariel 523
```

The following SAS program is submitted:

```
proc sql;
[_insert_select_clause_]
from WORK.TWO
;
quit;
```

The following output is desired:

```
Salary Bonus
------ -----
200 20
205 20.5
310 31
523 52.3
```

Which SQL procedure clause completes the program and generates the desired output?

A. `select Salary Bonus as Salary*.10 as Bonus`

B. `select Salary, Salary*.10 label="Bonus"`

C. `select Salary Bonus=Salary*.10 "Bonus"`

D. `select Salary, Salary*.10 column="Bonus"`

(d) To create a list of unique Customer_Id values from the customer data set, which of the following techniques can be used?

```
technique 1: proc SORT with NODUPKEY and OUT=
technique 2: data step with IF FIRST.Customer_Id=1
technique 3: proc SQL with the SELECT DISTINCT statement
```

A. only technique 1

B. techniques 1 and 2

C. techniques 1 and 3

D. techniques 1, 2, or 3

(e) Given the SAS data sets:

```
WORK.ONE            WORK.TWO
Year Qtr Budget     Year Qtr Sales
---- --- ------     ---- --- -----
2001 3 500          2001 4 300
2001 4 400          2002 1 600
2003 1 350
```

The following SAS program is submitted:

```
proc sql;
select
TWO.*, budget
from
WORK.ONE
```

```
[_insert_join_operator_]
WORK.TWO
on ONE.Year=TWO.Year
;
quit;
```

The following output is desired:

```
Year Qtr Sales Budget
---- --- ----- ------
2001 4 300 500
2001 4 300 400
2002 1 600 .
. . . 350
```

Which join operator completes the program and generates the desired output?

A. full join
B. left join
C. right join
D. outer join