Lab 3.

libname mylib 'your folder';
*options mergenoby=nowarn;
* type in HELP nomergerby in the command box to get information on how sas is handling merges without a by statement on your system;
*Part A.1;
data one_two ; set mylib.data1 mylib.data2;
run;
* 30 obs in data1 42 obs in one_two 13 variables in one_two;

*Part A.2;
proc sort ; by id;
data alldata; merge one_two mylib.data3;
proc print data=alldata;
* There are 42 obs in alldata without using a by statement on the merge. Records were matched sequentially NOT by matching ids. After adding the by statement, records were matched by id. There are 43 obs because 3 records did not match (see next data step);
run;

*Part A.3;
data alldata; merge one_two (in=count) mylib.data3 (in=count2); by id;
if count=0 then put id= count= count2=;
if count2=0 then put id= count= count2=;
title 'with by statement';
run;
* 2 records in one_two that do not have a match to data3
  1 record in data3 that does not have a match in one_two;

*Part A.4;
data alldata; merge one_two (in=count) mylib.data3 (in=count2); by id;
if count=0 or count2=0 then delete;
run;
* 40 observations in alldata;
* submit the log from this data step and the answer to the question;

26   * 2 records in one_two that do not have a match to data3
27     1 record in data3 that does not have a match in one_two;
28
29   *Part A.4;
30   data alldata; merge one_two (in=count) mylib.data3 (in=count2); by id;
31   if count=0 or count2=0 then delete;
32   run;

NOTE: There were 42 observations read from the data set WORK.ONE_TWO.
NOTE: There were 41 observations read from the data set MYLIB.DATA3.
NOTE: The data set WORK.ALLDATA has 40 observations and 17 variables.
NOTE: DATA statement used (Total process time):
      real time      0.01 seconds
      cpu time      0.01 seconds
*Part A.Q5;
data b ;set alldata;
c=0;
array dxall(6)$ dx1-dx6;
do i=1 to 6 ;
if dxall(i)^="" then c+1;
end;
proc freq data=b; tables c;
title 'Part A. Q5';
run;
*c is a new variable created in this data step in a sum statement. By default it is
set to zero
and value is retained from one iteration to the next. If you do not set c =0,
the result will be the cumulative number of unknowns for each record.
* submit output window and the answer to the question;

<table>
<thead>
<tr>
<th>c</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>25.00</td>
<td>10</td>
<td>25.00</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2.50</td>
<td>11</td>
<td>27.50</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5.00</td>
<td>13</td>
<td>32.50</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>17.50</td>
<td>20</td>
<td>50.00</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7.50</td>
<td>23</td>
<td>57.50</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7.50</td>
<td>26</td>
<td>65.00</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>35.00</td>
<td>40</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*Part B.1;
data employee1_2; set mylib.employee1 mylib.employee2;
label name='employee name'
    hire='hire date'
    phone='office phone';
proc contents data=employee1_2;
* the data set has 10 records;
run;

*Part B.Q2;
data employee123; set employee1_2 mylib.employee3;
drop name;
proc print data=employee123;
title 'Employee123';
run;
*15 observations in employee123
gender is blank for SSN=244967839
phone is blank for SSN=933476520;
* submit output window and the answers to the questions;

<table>
<thead>
<tr>
<th>Obs</th>
<th>ssn</th>
<th>hire</th>
<th>salary</th>
<th>phone</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1    429685482     7526    34800    493-0824
2    244967839     9245    27900    929-2623
3    996740216     8750    32600    933-6908
4    675443925    10272    29900    493-3993
5    845729308     8327    39800    286-0519
6    324987451    10405    41400    933-1472
7    596771321    10304    41200    929-4800
8    477562122     7702    47400    286-2816
9    894724859     8575    41400    493-1472
10    988427431     9319    43700    929-3885
11    744289612    10356    33400    F
12    824904032     8039    38200    M
13    242779184     9402    37500    M
14    544382887     8971    45000    F
15    933476520     8144    39900    F

*Part B.3;
proc contents data=mylib.employee4; run;

proc sort data=employee123; by ssn;

data employee_total;
merge employee123 (in=in3) mylib.employee4(rename=(phone=home) in=in4); by ssn;
if in3^=in4 then put ssn= in3= in4=;

proc print data=employee_total;
run;

PART B Q3.
* Each file on merge statement has 15 records
  employee_total has 16 records
  the following will be listed in log window:
  ssn=112234567 in3=0 in4=1
  ssn=324987451 in3=1 in4=0
  employee_total has 7 variables;
*submit log and output windoes and the answers to the questions;

76   data employee_total;
77   merge employee123 (in=in3) mylib.employee4(rename=(phone=home) in=in4); by ssn;
78   if in3^=in4 then put ssn= in3= in4=;
79

  ssn=112234567 in3=0 in4=1
  ssn=324987451 in3=1 in4=0

NOTE: There were 15 observations read from the data set WORK.EMPLOYEE123.
NOTE: There were 15 observations read from the data set MYLIB.EMPLOYEE4.
NOTE: The data set WORK.EMPLOYEE_Total has 16 observations and 7 variables.
NOTE: DATA statement used (Total process time):
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>real time</td>
<td>0.01</td>
</tr>
<tr>
<td>cpu time</td>
<td>0.01</td>
</tr>
</tbody>
</table>

80   proc print data=employee_total;
81   run;

NOTE: There were 16 observations read from the data set WORK.EMPLOYEE_TOTAL.
*Part B.Q4;
data employee_total;
merge employee123 (in=in3) mylib.employee4(rename=(phone=home) in=in4); by ssn;
if in3^=in4 then delete;
proc print data=employee_total;
run;
*submit a copy of the log window;

*Part B.Q5;
data females males; set employee_total;
if gender="F" then output females;
else if gender="M" then output males;
run;
*submit a copy of the log window;
*be careful there are observations with unknown gender;
if gender="F" then output females;
else if gender="M" then output males;
run;

NOTE: There were 14 observations read from the data set WORK.EMPLOYEE_TOTAL.
NOTE: The data set WORK.FEMALES has 3 observations and 7 variables.
NOTE: The data set WORK.MALES has 2 observations and 7 variables.
NOTE: DATA statement used (Total process time):
  real time       0.01 seconds
  cpu time        0.01 seconds