Combining Multiple Tables with Different Matching Columns

Merge the **pg2.np_codelookup**, **pg2.np_final**, and **pg2.np_species** tables to create a table that contains information about the common birds found at locations that have more than 5,000,000 visitors a year.

1. Open **p205p05.sas** from the **practices** folder. The first three steps sort and merge the **pg2.np_codelookup** and **pg2.np_final** tables. Submit the first two PROC SORT steps and the DATA step and examine the **highuse** table.

2. How many rows and columns are in the **highuse** table? What are the column names?

   There are 713 rows and two columns in the **highuse** table. The column names are **ParkName** and **ParkCode**.

3. Add a subsetting IF statement in the DATA step to output only the rows in which **DayVisits** is greater than or equal to 5,000,000. Submit the DATA step. Why must you use IF instead of a WHERE statement? How many rows are in the **highuse** table now?

```
proc sort data=pg2.np_CodeLookup out=sortnames(keep=ParkName ParkCode);
    by ParkName;
run;

proc sort data=pg2.np_final out=sortfinal;
    by ParkName;
run;

data highuse(keep=ParkCode ParkName);
    merge sortfinal sortnames;
    by ParkName;
    if DayVisits ge 5000000;
run;
```

   You must use a subsetting IF statement because the **DayVisits** column is in only one of the tables in the MERGE statement. Now the **highuse** table has three rows.

4. Submit the final PROC SORT step to sort and subset the **pg2.np_species** table. Compare the columns in the output **birds** table with the **highuse** table to determine the matching column. Which column is in both tables?

   The **ParkCode** column is in both tables.

5. Add a PROC SORT step to sort the **highuse** table by the matching column in the **birds** table. What is the value of **ParkCode** in the first row of the **highuse** table?

```
proc sort data=highuse;
    by ParkCode;
run;
```

ParkCode is *GRCA* in the first row.

6. Add a DATA step to merge the **highuse** and **birds** tables and create a table named **birds_largepark**. Include in the output table only **ParkCode** values that are in the **highuse** table.

```
data work.birds_largepark;
    merge birds highuse(in=inPark);
    by ParkCode;
    if inPark=1;
run;
```

7. How many rows and columns are in the **birds_largepark** table?

There are 274 rows and five columns in the **birds_largepark** table.