```
/*****************************************************************/
/* Program Name: Homework 5                                      */
/* Program Location: C:\Users\dsingh\Dropbox\Stat_604\Homework\HW5 */
/* Date Created: 6/8/2019                                        */
/* Author: Dhruv Singh                                           */
/* Purpose: Oklahoma Schools data cleaning                       */

libname ok_sch 'C:\Users\dsingh\Dropbox\Tamu\Stat_604\Homework\hwdata'
access = readonly;
libname ok_cln 'C:\Users\dsingh\Dropbox\Tamu\Stat_604\Homework\HW5'
access = readonly;
filename report
'C:\Users\dsingh\Dropbox\Tamu\Stat_604\Homework\HW5\DSingh_HW05_output.pd
f';


/* Step 1: reading in dataset */

/* Step 2a: Replacing county with blanks */
data schools;
      set ok_sch.ok_schools;
      newvar = tranwrd(county, 'COUNTY', '');
run;

* dropping original county var;
data schools2;
      set schools;
      drop county;
run;

* renaming new couunty var;
data schools2;
      set schools2 (rename = (newvar=County));
run;

* the following temp data creates new vars for numeric grade vars;
data schools3;
      set schools2;
      newvar = 0;
      select;
            when (grade8 not in ('n/a', '*'))
            newvar=input(grade8, comma6.);
            otherwise newvar=.;
      end;

      newvar1 = 0;
      select;
            when (grade9 not in ('n/a', '*'))
            newvar1=input(grade9, comma6.);
            otherwise newvar1=.;
      end;

      newvar2 = 0;
      select;
            when (grade10 not in ('n/a', '*'))
            newvar2=input(grade10, comma6.);
            otherwise newvar2=.;
      end;

      newvar3 = 0;
```

```
      select;
            when (grade11 not in ('n/a', '*'))
            newvar3=input(grade11, comma6.);
            otherwise newvar3=.;
      end;

      newvar4 = 0;
      select;
            when (grade12 not in ('n/a', '*'))
            newvar4=input(grade12, comma6.);
            otherwise newvar4=.;
      end;

run;


* here the following temp data step renames created vars, and drop old
char types;
data schools4;
      set schools3;
      drop grade8 grade9 grade10 grade11 grade12;
run;

data schools4;
      set schools4 (rename=(newvar=Grade8 newvar1=Grade9 newvar2=Grade10
newvar3=Grade11 newvar4=Grade12));
run;

* search and replace in string variables ;
data schools5;
      set schools4;

      /*Step 2c: renaming city variables to correct for mispellings */
      select;
            when (city = 'CHUOTEAU')
            city=tranwrd(city, 'CHUOTEAU', 'CHOUTEAU');

            when (city = 'OKC')
            city=tranwrd(city, 'OKC', 'OKLAHOMA CITY');

            when (city='JENKS')
            city=tranwrd(city, 'JENKS', 'TULSA');

            when (city='MUSKOGE')
            city=tranwrd(city, 'MUSKOGE', 'MUSKOGEE');

            when (city='RUSHSPRINGS')
            city=tranwrd(city, 'RUSHSPRINGS', 'RUSH SPRINGS');

            when (city='SEMIONOLE')
            city=tranwrd(city, 'SEMIONOLE', 'SEMINOLE');

            when (city='SO. COFFEEVILLE')
            city=tranwrd(city, 'SO. COFFEEVILLE', 'SOUTH COFFEEVILLE');

            when (city='WOOWARD')
            city=tranwrd(city, 'WOOWARD', 'WOODWARD');

            otherwise;
      end;
```

```
      /* Step 2d: renaming one county to combine it with another */
      select;
            when (county='ALFALFA')
            county=tranwrd(county, 'ALFALFA', 'CHEROKEE');

            otherwise;
      end;
run;


/*Step 3a: Sorting data set for by group processing*/
proc sort data=schools5 out=summary;
      by City;
run;


/*Summarize Class Size by City*/
data summary (keep = City Grade8Sum Grade9Sum Grade10Sum Grade11Sum
Grade12Sum);
      set summary;
      by City;
      if First.City then Grade8Sum=0;
      Grade8Sum+Grade8;

      if First.City then Grade9Sum=0;
      Grade9Sum+Grade9;

      if First.City then Grade10Sum=0;
      Grade10Sum+Grade10;

      if First.City then Grade11Sum=0;
      Grade11Sum+Grade11;

      if First.City then Grade12Sum=0;
      Grade12Sum+Grade12;
      *step 3b: removing redundant observations;
      if Last.City;
run;


data summary2;
      set summary;
      *step 3c: labelling summary variables;
      label Grade8Sum = 'Eighth Graders'
            Grade9Sum = 'Ninth Graders'
            Grade10Sum = 'Tenth Graders'
            Grade11Sum = 'Eleventh Graders'
            Grade12Sum = 'Twelfth Graders';

      * step 3d: creating new vars for current and projected enrollment;
      CurrentEnrollment = Sum(Grade9Sum, Grade10Sum, Grade11Sum,
Grade12Sum);
      ProjectedEnrollment = Sum(Grade8Sum, Grade9Sum, Grade10Sum,
Grade11Sum);

      PercentChange = ((ProjectedEnrollment -
CurrentEnrollment)/CurrentEnrollment)*100;
run;
```

```
/* Step 4: Creating output delivery system to pdf */
ods pdf file = report bookmarkgen = no;

/* Step 5: Printing relevant data with labels created */
proc print data = schools5;
run;

proc print data = summary2 label;
run;

ods pdf close;
```