```
/*
Program: ST662 - Assignment 3 - Version 5.sas
Author : Sean O'Riogain (18145426)
Date   : 23rd March 2019

Part A

In your last assignment, you examined data from a grassland biodiversity experiment that was conducted at many sites across Europe and one in Canada.

Here, we will focus on the weather dataset from the experiment.

Information on the experiment is available at
Abstract: http://onlinelibrary.wiley.com/doi/10.1890/14-0170.1/abstract.
Datasets for download: http://www.esapubs.org/archive/ecol/E095/232/.
Datasets' descriptions: http://www.esapubs.org/archive/ecol/E095/232/metadata.php.

Download the climate.csv dataset and read it into SAS. We will concentrate on the variable air mean.
*/

PROC IMPORT DATAFILE="/home/seanoriogain200/ST662/climate_for_assignment_3.csv"
    DBMS=CSV
    OUT=work.clim1 REPLACE;
    GETNAMES=YES;
RUN;

proc print data=clim1 (obs=10);
run;

/* Question 1 */

/*
Explore air mean to identify missing values.
*/

proc means data=clim1 n nmiss;
    by site year;
    var air_mean;
    OUTPUT OUT=summ1(drop=_freq_ _type_) n=N nmiss=Missing;
quit;

/*
Create a new dataset to summarise the total number of observations and the number missing for each site and
year. Use the following format for the table:
    Site Year N Missing % Missing
    ...
    ...
    ...
    ...
    ...
Restrict this new dataset to only values that have > 0 missing values.
*/

data summ1;
    set summ1;
    where Missing > 0;              /* Restrict the dataset as required. */
    N = N + Missing;                /* Ensure that N represents all observations - both non-null and null */
    Pct_Missing = (Missing/N)*100;  /* The missing observations are expressed as a percentage of all obs. */
    label Site='Site', Year='Year' Pct_Missing='% Missing'; /* Label the relevant columns to match spec.  */
    format Pct_Missing 10.2;
run;

/*
Provide a printout of the reduced dataset only.
*/

title 'Question 1 - Summary Dataset (summ1)';

proc print data=summ1 label;        /* Use label names instead of variable names as column headings */
run;

title;

/* Verifying results using SQL.... */

proc sql number;
    select site, year, count(*) as N, count(*) - count(air_mean) as Missing,
            (((count(*) - count(air_mean)) / count(*)) * 100) as Pct_Missing format=10.2
    from clim1
    group by site, year
    having count(*) - count(air_mean) > 0
    order by site, year;
quit;

proc sql;
    select count(*) as Missing
    from clim1
    where air_mean = .;
quit;

/* 1306 obs returned */

proc sql;
    select sum(Missing) as Missing
    from summ1;
quit;
```

```
/* 1306 rows returned  (as expected) */

/* Question 2 */

/* Creating a new version of the clim1 dataset after merging the previous dataset with it (using inner join)... */

/* Appending the overall mean values for air_min and air_max to the clim2 dataset..... */

/*
Impute missing air mean values using the following guidelines

- You do not need to provide any output here.

- If a lot more than 5% of observations for a site in a year are missing, use the average of
air min and air max.
*/

/* Let's see what "a lot more than 5%" might mean....*/

proc sql;
    select pct_missing, site, year
    from summ1
    order by 1 desc, 2, 3;
quit;

/* From the results of the above query we can see that the Site-Year combinations with 100% missing
      observations are the only ones to qualify as being a lot greater than 5%, that all of them
      pertain to Site 36, and that all of Site 36's observations are missing their air_mean values.
   Therefore, we can use Site=36 as the selection criterion for imputation treatment 1.
*/

/* Counting the observations for imputation treatment 1 */

proc sql;
    select count(*) as Missing_36
    from clim1
    where air_mean is null
      and site = 36;
quit;

/* 1244 rows returned */

/* Counting the observations for imputation treatment 2 */

proc sql;
    select count(*) as Missing_Other
    from clim1
    where air_mean is null
      and site <> 36;
quit;

/* 62 rows returned. */

/* 1244 + 62 = 1306 (which is expected number of total imputations calculated in Q1 above) */

/* As air temperatures are related to the location (i.e. site) and time of their measurement, the
      best average of air_min and air_max would be at the observatvation level (in preference to
      group (Site-Year) level or overall level) - provide the level of missing air_min and air_max
      values for Site 36 is zero or very small in number, at least. */

/* The results of the previous query show that Site 36 has 1244 observations in total */
```

```
proc sql;
    select 'Site 36 Total', count(*) as count
    from clim1
    where site=36
    UNION
    select 'Site 36 Missing OR', count(*) as count
    from clim1
    where site=36
      and (air_min is null or air_max is null)
    UNION
    select 'Site 36 Missing AND', count(*) as count
    from clim1
    where site=36
      and (air_min is null and air_max is null);
quit;

/* The results of the above query show that 2 of Site 36's  1244 observations have missing values for both air_min
      and air_max, while 59 have a missing values for either air_min or air_max. Even though imputation of
      air_mean values will be suboptimal for those 59 observations, we will proceed with imputing using an
      observation-level mean....*/

/* Performing imputation treatment 1 (for Site 36).... */
```

```
data clim2;
    set clim1;
    if site = 36 & air_mean = . then air_mean = (air_min + air_max)/2;
run;

/* Validating results using SQL (expected row count = 62) */
```

```
proc sql;
    select *
    from clim2
    where site=36
      and air_mean is null;
quit;

/* 121 rows returned (expected 62  + suboptimal 59 = 121) */

/* This means that the 59 observations for Site 36 with missing air_min and/or air_max values do not have an
      imputed air_mean value. However, we will allow them to be processed by imputation 2 below. */

/* - Otherwise, use the average of all other air mean values for that site and year. */

/* Performing imputation treatment 2.... */

proc stdize data=clim2 out=clim3 method=mean missing=mean reponly;
    by site year;
    var air_mean;
run;

/* Validating results using SQL (expected row count = 0) */

proc sql;
    select count(*) as Missing
    from clim3
    where air_mean is null;
quit;

/* 0 rows returned (as expected) */

/* Summarising the site-year group air_mean averages for sites other than Site 36 in the STDIZE-imputed' dataset
      (clim3)..... */

proc sql number;
    select site, year, round(avg(air_mean), 0.01) as avg_air_mean
    from clim3 c
    where site <> 36
      and exists(select 1
                    from summ1 s
                    where s.site = c.site
                      and s.year = c.year)
    group by site, year
    order by site, year;
quit;

/* Verifying the site-year group air_mean averages for Site 36 using the original dataset (clim1)...... */

proc sql number;
    select site, year, round(avg(air_mean), 0.01) as avg_air_mean
    from clim1 c
    where site <> 36
      and exists(select 1
                    from summ1 s
                    where s.site = c.site
                      and s.year = c.year)
    group by site, year
    order by site, year;
quit;

/* The output of both of the previous queries is identical which indictaes that STDIZE worked as expected
      (for sites other than Site 36). */

/* Question 3 */

/*
For those sites and years that had some missing values, create a new dataset with the average of
air mean pre- and post- imputation.
*/

/* Merging pre- and post-imputation datasets with missing air_mean summary (summ1) */

data merge_pre;
    merge clim1 summ1;
    by site year;
run;

data merge_post;
    merge clim3 summ1;
    by site year;
run;

/* Getting site-year group averages for the previous 2 datasets (for groups with missig values only)......*/

proc means data=merge_pre mean maxdec=2;
    by site year;
    var air_mean;
    where Missing is not null;
    output out=avgs_pre(drop=_freq_ _type_) mean=avg_pre;
    format avg_pre 10.2;
run;

proc means data=merge_post mean maxdec=2;
    by site year;
    var air_mean;
    where Missing is not null;
    output out=avgs_post(drop=_freq_ _type_) mean=avg_post;
    format avg_post 10.2;
run;
```

```
data avgs_pre_post;
    merge avgs_pre avgs_post;
    by site year;
run;

/* Provide a printout of the dataset. */

title 'Question 3 - Averages Comparison Dataset (avgs_pre_post)';
```

```
proc print data=avgs_pre_post;
run;

title;

/* Comment on what you have found.*/

/*
We can see that the pre- and post-imputation averages are the same for the Site-Year groupings for all sites
except Site 36. This is to be expected because the missing air_mean values for those sites were imputed using
the group air_mean average and, therefore, that group average should remain unchanged subsequent to imputation.

In the case of Site 36's Site-Year groupings, the pre- and post-impution values differ in line with expectations -
as those group averages were null (effectively zero) before imputation and were set to the observation-level
average of air_min and air_max for the vast majority of its 1244 observations while the remaining 59 observations
were assigned Site-Year group-level air_mean average values.
```