```
/*
Program: ST662 - Assignment 2 - Version 3.sas
Author : Sean O'Riogain (18145426)
Date   : 3rd March 2019

A grassland biodiversity experiment was conducted at many sites across Europe and one in Canada. The
data from this experiment was published in the journal called Ecology.

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.

Write a SAS programme to do the following data manipulation exercises.

1. (a) Download the biomass.csv dataset and read it into SAS.
*/

PROC IMPORT DATAFILE="/home/seanoriogain200/ST662/biomass.csv"
      OUT=bio
      DBMS=csv REPLACE;
      getnames=yes;
run;

/* Get the total number of observations and variables */
proc contents data=bio;
run;
/* 15372 observations found with 22 variables */

/* Get the missing data count and minimum value by variable */
proc means data=bio maxdec=0 n nmiss min;
run;
/* Missing data is flagged for 7 of the 20 variables,
      with almost 2500 missing for each of 5 of them.
   Note first year (min) = 2003 for later usepossible future reference. */
```

```
/*
(b) Restrict the dataset to only sites 13, 14, 23, 25, 33 and 52, to only the first year of experimental
data, and to only treatment 1.
*/

/* Create the required dataset (BIO1) */
DATA bio1;
    set bio;
    where site in(13, 14, 23, 25, 33, 52) & yearn = 1 & treat = 1;
run;
/* 360 rows output */

/* Get the number of restricted observations with miss count */
proc means data=bio1 n nmiss;
run;
/* n = 360 and note that there are still 3 missing values each for the final 6 variables */

/*
(c) Create a new dataset that provides the annual yield for each plot at each site.
*/

/* It is not clear whether this new dataset is to be constructed from the restricted dataset or from
      the original one, so I'm going to assume that it's from the restricted one. */

/* Sort the dataset by site, plot and year */
proc sort data=bio1 out=bio2;
   by site plot year;
run;

/* Create the required dataset (BIO2_SAS) using SAS (the hard way) */
data bio2_sas;
    set bio2;
    by site plot year;

    if first.year then tot_yield = 0;

    tot_yield + harv_yield;
```

```
    if last.year then output bio2_sas;

    keep site plot year tot_yield;
run;
/* 180 rows output */

/* Create the required dataset (BIO2_SQL) using SQL (the hard way) */
proc sql;
    create table bio2_sql
    as
    select site, plot, year, sum(harv_yield) as tot_yield
    from bio2
    group by site, plot, year;
    order by site, plot, year;
quit;
/* 180 rows output (as per the SAS method above) */

*/
(d) Create a new dataset that provides the average annual yield for each site (i.e. averaged across
all plots).
*/

/* Sort the dataset by site and year */
proc sort data=bio1 out=bio3;
   by site year;
run;

/* Create the required dataset (BIO3_SAS) using SAS (the hard way) */
data bio3_sas;
    set bio3;
    by site year;

    if first.year then do
        count = 0;
        tot_yield = 0;
        avg_yield = 0;
```

```
        end;

    count + 1;
    tot_yield + harv_yield;

    if last.year then do
        avg_yield = tot_yield / count;
        output bio3_sas;
        end;

    keep site year avg_yield;
run;
/* 6 rows output */

/* Create the required dataset (BIO3_SQL) using SQL (the hard way) */
proc sql;
    create table bio3_sql
    as
    select site, year, avg(harv_yield) as avg_yield
    from bio3
    group by site, year
    order by site, year;
quit;
/* 6 rows output (as per the SAS method above) */

/*
2. (a) Download the climate.csv dataset and read it into SAS.
*/

proc import DATAFILE="/home/seanoriogain200/ST662/climate.csv"
      OUT=cli;
      DBMS=csv REPLACE;
      getnames=yes;
run;

/* Get the total number of observations and variables */
proc contents data=cli;
```

```
run;
/* 39712 observations found with 12 variables */

/* Get the missing data count and minimum value by variable */
proc means data=cli maxdec=0 n nmiss min;
run;
/* Missing data is flagged for 4 of the 12 variables,
      with 80-100 missing for 3 of them and just over 1300 for the other (AIR_MEAN).
   Other potential data quality issues include: Zero values for the AIR and MONTH variables;
      and negative values (which may be OK) for the AIR_MIN, AIR_MAX and AIR_MEAN variables.
   Note first year (min) = 2002 for possible future reference. */

/*
(b) Restrict the dataset to only sites 13, 14, 23, 25, 33 and 52.
*/

DATA cli1;
    set cli;
    where site in(13, 14, 23, 25, 33, 52);
run;
/* 7667 rows extracted */

/* Get the number of restricted observations with miss count */
proc means data=cli1 n nmiss;
run;
/* n = 7667 and note that there are still 24 missing values in total for the final 6 variables */

/*
(c) Create a new dataset that provides the average `air mean' for each site and each year.
*/

/* Again, I'm going to assume that the new dataset is to be constructed from the restricted dataset
      instead of the original one. */

/* Sort the original dataset by site and year */
proc sort data=cli1 out=cli2;
    by site year;
```

```
run;

/* Create the new dataset (CLI2_SAS) using SAS (the hard way) */
data cli2_sas;
    set cli2;
    by site year;

    if first.year then do
        count = 0;
        tot_air = 0;
        end;

    count + 1;
    tot_air + air_mean;

    if last.year then do
        avg_air = tot_air / count;
        output cli2_sas;
        end;

    keep site year avg_air;
run;
/* 23 rows output */

/* Create the required dataset (CLI2_SQL) using SQL (the easy way) */
proc sql;
    create table cli2_sql
    as
    select site, year, avg(air_mean) as avg_air
    from cli2
    group by site, year
    order by site, year;
quit;
/* 23 rows output (as per the SAS method above) */

/*
3. (a) Merge the biomass dataset created in Qu 1d with the relevant year of the climate dataset
```

```
created in Qu 2c.
*/

/* Create the required dataset (BIOCLI_SAS) using SAS (the hard way) */
data biocli_sas;
    merge bio3_sas cli2_sas;
    by site year;
run;
/* 23 rows output (6 of which are complete) */

/* Create the required dataset (BIOCLI_SQL) using SQL (the easy way) */
proc sql;
    create table biocli_sql
    as
    select c.site, c.year, b.avg_yield, c.avg_air
    from cli2_sql as c left join bio3_sql as b
      on c.site = b.site and c.year = b.year
    order by 1, 2;
quit;
/* 23 rows output (6 of which are complete - as per the SAS method above) */

/*
(b) Create a scatter plot of average annual yield versus average annual temperature. Ensure the
quality of the scatterplot is suitable for including in a presentation or report (e.g. put a title
on it, check the font sizes of labels, perhaps label points within the graph etc).
*/

%macro scatter;
    proc sgplot data=biocli_sas;
        title height=12pt color=CX0000FF 'Average Mean Air Temperature versus Annual Yield (Label=Site)';
        scatter x=avg_air y=avg_yield
            / markerattrs=(symbol=StarFilled size=2pct color=CXFF0000) datalabel=site;
        styleattrs datacolors=(red);
        xaxis label='Temperature' labelattrs=(size=12pt) grid;
        yaxis label='Yield' labelattrs=(size=12pt) grid;
    run;
%mend;
```

```
%scatter;
/* Only 6 points output (for site 13, 14, 23, 25, 33 and 52, respectively). */

/* These should both fit on one page. */
options orientation=landscape;

ODS printer PDF file="/home/seanoriogain200/ST662Lib/ST662 - Assignment 2 - Q3 Output Merge.PDF"
    author = 'Sean O''''Riogain'
    keywords = 'Annual Yield Temperature Yield Site'
    subject = 'Air Temperature versus Annual Yield'
    title = 'Average Mean Air Temperature versus Annual Yield';

     /* Define the columns of output */
    ODS layout start columns=2;

    /* Print the output for question 3a in column 1 */
    ODS region;
        proc print data=biocli_sas;
        run;
        quit;

    /* Print the output for question 3b in column 2 */
    ODS region;
        %scatter;
        quit;

    ODS layout end;
```