```
/*************************************************************************************************************/
/* ST662 Topics in Data Analytics
/* Student: Paul Williamson
/* Student ID: 18145469

/* Assignment Sheet 2:

/* 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.
/*************************************************************************************************************/


/*****************************************************************************/
/* Question 1
/*****************************************************************************/

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

proc sort OUT=ST662LIB.biomass1 /* permanent copy in ST662 library */
    datafile="/home/polmacuilliam10/ST662/Datasets/biomass.csv"
    dbms=CSV replace;
    getnames=YES;
run;

/* (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. */

data ST662LIB.biomass2;                          /* create copy of the original dataset to work with */
    set ST662LIB.biomass1;                       /* original dataset */
    if ((SITE = 13 or SITE = 14 or SITE = 23 or SITE = 25 or SITE = 33 or SITE = 52)
        and (YEARN = 1)
        and (TREAT = 1))
    then output;
run;

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

data sample_biomass1; /* create a temp subset of dataset in the work library folder */
   set ST662LIB.biomass2;
run;

proc sort data=sample_biomass1; /* sort class variables before proc means used - may not be necessary if using class keyword */
    by YEAR SITE PLOT;
run;

proc means data=sample_biomass1 sum;
    title 'Annual yield per site and per plot';
    var HARV_YIELD;
    by YEAR SITE PLOT;
    output out=sample_biomass2 sum=annual_YIELD;
run;

data sample_biomass2;
   set sample_biomass2;
   keep YEAR SITE PLOT annual_YIELD; /* drop-delete _type_ and _freq_ columns from output dataset*/
run;

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

proc means data=sample_biomass1 mean;
    title 'Average yield per site';
    var HARV_YIELD;
    by YEAR SITE;
    output out=sample_biomass3 mean=avg_YIELD;
run;

data sample_biomass3;
   set sample_biomass3;
   keep YEAR SITE avg_YIELD; /* drop-delete _type_ and _freq_ columns from output dataset*/
run; quit;
```

```
/*************************************************************************************/
/* Question 2
/*************************************************************************************/

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

proc sort out=ST662LIB.climate1 /* permanent copy in ST662 library */
    datafile="/home/polmacuilliam10/ST662/Datasets/climate.csv"
    dbms=CSV replace;
    getnames=YES;
run;

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

data ST662LIB.climate2; /* create copy of the original dataset to work with */
    set ST662LIB.climate1; /* original dataset */
    if (SITE = 13 OR SITE = 14 OR SITE = 23 OR SITE = 25 OR SITE = 33 OR SITE = 52)
    then output;
run;

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

data sample_climate1; /* create a temp subset of dataset in the work library folder */
   set ST662LIB.climate2;
run;

 /* sort class variables before proc means used - may not be necessary if using class keyword */
proc sort data=sample_climate1;
    by YEAR SITE;
run;

proc means data=sample_climate1 mean;
    title 'Average air_mean per site per year';
    var AIR_MEAN;
    by YEAR SITE;
    output out=sample_climate2 mean=avg_AIR_MEAN;
run;

data sample_climate2;
   set sample_climate2;
   keep YEAR SITE avg_AIR_MEAN;  /* drop-delete _type_ and _freq_ columns from output dataset*/
run; quit;


/*************************************************************************************/
/* Question 3
/*************************************************************************************/

/* (a) Merge the biomass dataset created in Qu 1d with the relevant year of the climate dataset created in Qu 2c. */

data biomass_climate_combined;
    merge sample_biomass3 sample_climate2;
    by YEAR SITE;
    drop _type_ _freq_;
run;

/* (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). */

/*avg_YIELD values truncated to 3 decimal places just for display - actual values not changed*/
data biomass_climate_combined;
    set biomass_climate_combined;
    format avg_YIELD 6.3;  /*truncate for display purposes only*/
run;
title "ST662 Assignment1 Q3(b) Scatter plot:"; title2 " "; /* added extra title to create whitespace */
title3 "AVG_YIELD (Y-axis) Vs AVG_AIR_MEAN (X-axis)"; title4 " ";
proc sgplot data=biomass_climate_combined NOAUTOLEGEND ;
    scatter x=avg_AIR_MEAN y=avg_YIELD / DATALABEL = avg_YIELD
    markerattrs=(symbol=circlefilled size=2mm) datalabelattrs=(family='Times New Roman' size=12pt);
    YAXIS LABEL = 'Average Annual Yield (DM/m²)'; XAXIS LABEL = 'Average Annual Temperature';
run; quit;

/* add regression line to the scatterplot */
title "ST662 Assignment1 Q3(b) Scatter plot w/Regression line:"; title2 " "; /* added extra title to create whitespace */
title3 "AVG_YIELD (Y-axis) Vs AVG_AIR_MEAN (X-axis)"; title4 " ";
proc sgplot data=biomass_climate_combined NOAUTOLEGEND ;
    reg x=avg_AIR_MEAN y=avg_YIELD / DATALABEL = avg_YIELD lineattrs=(color=red thickness=0.5)
    markerattrs=(symbol=circlefilled size=2mm) datalabelattrs=(family='Times New Roman' size=12pt);
    YAXIS LABEL = 'Average Annual Yield (DM/m²)'; XAXIS LABEL = 'Average Annual Temperature';
run; quit;
```